Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Brain Activity Classification using Retentive Networks with Explainable AI

Adeel Hashmi*[1], Kunsh Sabharwal[2], Goyam Jain[2]

[1]Department of Artificial Intelligence and Data Science, School of Engineering & Technology, Vivekananda Institute of Professional Studies-Technical Campus, Pitampura, New Delhi, India
[2]Department of Artificial Intelligence and Machine Learning, School of Engineering & Technology, Vivekananda Institute of Professional Studies-Technical Campus, Pitampura, New Delhi, India

ARTICLE INFO

Article history:
Received: 08/01/2026.
Revised: 30/04/2026,
Accepted: 10/06/2026,
Available online: 15/06/2026
Keywords:
Brain Activity Classification
Convolutional Neural Networks
Vision Transformer
RetNet
Explainable AI (XAI)

ABSTRACT

Harmful brain activity can significantly impact an individual's life, leading to debilitating seizures, epileptic episodes, and long-term cognitive impairments that affect daily functioning and overall quality of life. These abnormalities can be effectively detected using electroencephalography (EEG) image data, which captures the underlying neural activity of the brain in a non-invasive manner. In this study, Retentive Networks (RetNets) are employed to classify EEG spectrogram images for identifying different types of harmful brain activity. While deep learning architectures such as EfficientNet and Vision Transformers have demonstrated strong performance in image-based classification tasks, RetNets offer a compelling advantage in terms of reduced computational complexity and efficient sequence modelling. The dataset used in this work is a publicly available EEG dataset comprising approximately 17,000 spectrogram images derived from EEG recordings, ensuring sufficient diversity and representation for robust model training and evaluation. Experimental evaluation demonstrates that the proposed RetNet model achieved the highest overall performance, with a precision of 94%, recall of 100%, and an F1-score of 97%, indicating balanced and highly reliable classification capability. In comparison, EfficientNet achieved a precision of 91%, recall of 90%, and an F1-score of 96%, while the Vision Transformer achieved precision, recall, and F1-scores of 94%, 86%, and 90%, respectively. Furthermore, the proposed approach integrates model interpretability through Explainable Artificial Intelligence (XAI) techniques. The novelty of the proposed work lies in combining RetNet-based EEG spectrogram classification with XAI-driven interpretability for harmful brain activity detection.

1. INTRODUCTION

Harmful brain activity, manifested through conditions such as seizures, epilepsy, and other neurological disorders, poses a significant challenge to global healthcare systems due to its complex diagnosis and long-term impact on patients' quality of life. Electroencephalography (EEG) is a widely used, non-invasive technique for monitoring brain activity and plays a crucial role in the detection and analysis of abnormal neural patterns. However, traditional EEG interpretation relies heavily on expert knowledge and manual analysis, which can be time-consuming, subjective, and prone to inter-observer variability. According to LeCun et al. [1], deep learning has significantly improved the capability of automated systems to analyze complex biomedical data and extract meaningful patterns from EEG signals. As per Mathew et al. [2], various deep learning techniques and architectures have been explored for improving classification accuracy and feature learning in healthcare-related applications. Recent studies like Razavi [3] have also emphasized the interpretability and modelling capabilities of deep learning approaches for scientific and signal-processing applications. As per the study of Wang et al. [4], transformer-based architectures have demonstrated strong performance in EEG-related tasks by effectively capturing hierarchical spatial information and long-range dependencies within EEG signals. According to Xu et al. [5], convolutional neural network-based frameworks have further enhanced EEG signal classification through deep transfer learning and automated feature extraction techniques. In the study by Habijan et al. [6], ensemble deep learning approaches have shown promising results

* Corresponding author's E-mail: adeel.hashmi@vips.edu
DOI: 10.24237/djes.2026.19208

for harmful brain activity classification, improving robustness and predictive performance in EEG-based diagnosis systems.

The systematic review by Rajwal and Aggarwal [7] and study of Chakravarthi et al. [8] discuss EEG signal analysis. 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.

While AI models can generate results and help doctors to understand the patient's disorder better, they still fail to provide confidence in their suggested course of treatment or action. Hence, while high accuracy is paramount in such cases, the clinical deployment of AI algorithms needs something beyond performance in the form of interpretability and trustworthiness. XAI aims at demystifying the internal decision-making within an AI model and making it 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 rather than spuriously occurring patterns. With XAI, we aim to deliver not only precise but also reliable insights about model behaviour. This improves prospects for clinical uptake, helps support expert verification, and aids in improving diagnostic decision-making. In total, the integration of XAI closes the gap between black-box AI systems and practical medical utility. Gagliardi et al. [9] proposed an XAI framework for brain MRI analysis.

Despite significant advancements in deep learning for EEG-based brain activity classification, several limitations still exist in current approaches. Existing studies primarily focus on conventional convolutional neural networks or transformer-based architectures, with limited exploration of Retentive Networks (RetNet) for EEG spectrogram classification. Furthermore, many existing methods rely solely on image-based EEG representations and often ignore complementary metadata information that may improve classification performance. Another major challenge is the limited interpretability of deep learning models, which reduces their reliability and trustworthiness in clinical applications.

Motivated by these research gaps, the main objective of this study is to develop a RetNet-based hybrid framework for EEG spectrogram classification and compare its performance with state-of-the-art architectures including VGG16, EfficientNet, and Vision Transformer. The proposed work also aims to integrate EEG spectrogram features with metadata information for enhanced multi-modal learning and improved classification performance. In addition, Explainable Artificial Intelligence (XAI) techniques, specifically LIME, are incorporated to improve transparency and interpretability of model predictions.

The key contributions of this research are as follows: (i) implementation and comparative evaluation of RetNet, VGG16, EfficientNet, and Vision Transformer for EEG-based harmful brain activity classification; (ii) integration of EEG spectrogram images and metadata within a unified hybrid learning framework; (iii) incorporation of LIME-based explanations for interpretable and trustworthy predictions; and (iv) achievement of superior classification performance using the proposed RetNet-based approach, which obtained 94% precision, 100% recall, and a 97% F1-score, demonstrating its effectiveness for real-time neurological monitoring and clinical decision-support applications.

This paper is further organized as follows: section-2 covers the literature review, providing related existing work, section-3 provides the methodology used for this work, section-4 presents the results with discussion, followed by the conclusion in section-5.

2. LITERATURE REVIEW

Electroencephalography (EEG) signals are widely used for brain activity analysis in applications such as motor imagery classification, seizure prediction, drowsiness detection, and harmful brain activity monitoring. However, EEG signals are highly non-stationary, noisy, and sensitive to external interference, which makes accurate classification a challenging task. In the work by Wang et al. [10], CNN-based visualization studies have shown that convolutional neural networks can effectively learn hierarchical image features for classification tasks. The research by Chauhan et al. [11] on convolutional neural network architectures for image recognition further demonstrated the effectiveness of CNNs in extracting discriminative visual patterns. Transfer learning approaches using VGG16 have also improved feature classification and representation learning in image-based datasets, as shown in the work by Tao et al. [12]. The work by Kaur and Gandhi [13], automated brain image classification systems using VGG-16 and transfer learning, showing strong performance in biomedical image analysis tasks. Rashid et al. [14], developed EEG and EMG-based multimodal driver drowsiness detection systems using continuous wavelet transforms and improved VGG-16 architectures. In the work by Li [15], deep learning-based EEG identity recognition using VGGNet has also shown promising results for biometric authentication applications. Study on EEG source imaging by Michel et al. [16] emphasized the complexity of EEG signal

generation and interpretation. Furthermore, convolutional retentive network architectures have recently been introduced for EEG decoding by Wang et al. [17], demonstrating the growing integration of retention-based learning mechanisms in EEG analysis. Research on neural signal interpretation by Cohen [18], highlighted the complexity and variability of EEG recordings.

A comprehensive review of deep learning methods for EEG motor imagery classification by Altaheri et al. [19] highlighted the success of CNNs, recurrent neural networks, and transformer-based approaches in capturing EEG features. In the work by Deng et al. [20], CNN and Swin Transformer hybrid models demonstrated strong performance in EEG-based motor imagery classification tasks. The work done by Hallal et al. [21] on EEG classification further showed that optimized architectures and feature extraction strategies significantly improve classification accuracy. The research by Gao et al. [22] combining complex network analysis with deep learning also demonstrated improved EEG signal interpretation and classification performance. Comparative analysis of deep learning approaches for harmful brain activity detection by Bhatti et al. [23] revealed that optimized architectures and feature extraction techniques substantially improve EEG classification performance. Spectrogram-based transfer deep learning frameworks have also shown competitive performance in harmful brain activity classification as shown by Ganesan et al. [24]. In addition, transfer learning methods based on VGG-16 convolutional neural networks by Li and Xu [25] demonstrated improved motor imagery classification performance.

Transformer-based approaches have recently gained attention in EEG and image classification tasks because of their ability to model long-range dependencies. A survey of transformer architectures by Lin et al. [26] discussed the effectiveness of transformers in sequence modelling and computer vision applications while also highlighting their computational complexity and dependence on large datasets. Study by Chen et al. [27] focusing on transformer applications in brain sciences emphasized their ability to capture attention-based representations from EEG and neural data. Research on transformer-based brain activity prediction by Adeli et al. [28] further demonstrated the capability of transformers to model complex neural representations effectively. Transformer-based EEG classification framework integrating spatial-temporal information by Xie et al. [29] achieved improved classification performance on raw EEG signals. Comparative survey by Wang et al. [30] on Vision Transformers (ViTs) reported that advanced ViT models achieve competitive or superior performance compared to conventional CNNs in many image classification benchmarks. Dongre and Mehta [31] utilized RetViT models, integrating retentive mechanisms with vision transformers to demonstrate improved efficiency and representation learning performance.

An optimized EfficientNet model by Islam et al. [32] achieved high precision in brain tumor classification while maintaining computational efficiency. Sadoon et al. [33] combined EfficientNet-B0 with support vector machines and achieved promising performance in epileptic seizure prediction from EEG signals. Furthermore, wavelet-based autoencoder frameworks integrated with EfficientNet by Naik and Ahamed [34] demonstrated effective schizophrenia detection from EEG data.

The RetNet architecture by Sun et al. [35] demonstrated that retention-based mechanisms can achieve transformer-level performance while reducing inference complexity and memory requirements. Retformer architecture by Erabati and Araujo [36] further extended retentive mechanisms to point cloud transformer models, improving long-range feature learning efficiency. Survey study on RetNet architectures by Yang et al. [37] highlighted the advantages of multi-scale retention mechanisms, and recurrent inference formulations. Retentive architectures have also been successfully applied in medical imaging applications, including diabetic retinopathy diagnosis by Sebasthiyar [38] and echocardiography image segmentation by Lin et al. [39].

RetNet30, a stacked convolutional neural network architecture, demonstrated strong performance in automated retinal disease diagnosis systems in the work by Subramaniam and Naganathan [40]. By combining a dedicated 30-layer CNN with a fine-tuned Inception V3 network, RetNet30 achieved highly accurate retinal image classification and produced an AUROC value of 0.98, demonstrating excellent discrimination between healthy and diseased retinal images. However, RetNet30 is not a retentive network, instead RetNet here stands for “RETinal disease NETwork”.

The major limitation of traditional CNN-based EEG models is their inability to effectively capture long-term temporal dependencies across EEG sequences. Transformers improve long-context learning but suffer from high computational complexity. RetNet provides parallel training, recurrent inference, lower memory usage, reduced inference latency. Its near-linear or efficient recurrent inference mechanism makes it more practical for portable EEG devices, real-time BCI systems, low-resource clinical environments. RetNet supports chunk-wise recurrent inference, streaming sequence processing, low-latency decoding. Hence, RetNet is highly beneficial for classification of EEG images compared to CNNs and other transformer models.

3. Methodology

The methodology followed in this work is presented in Figure 1. The dataset contains six target classes (seizure, LPD, GPD, LRDA, GRDA, and other). The class distribution is balanced, with each category containing 15,000–20,000 EEG segments. While minor variations exist across classes, no category is significantly underrepresented, reducing the risk of severe class imbalance during model training.

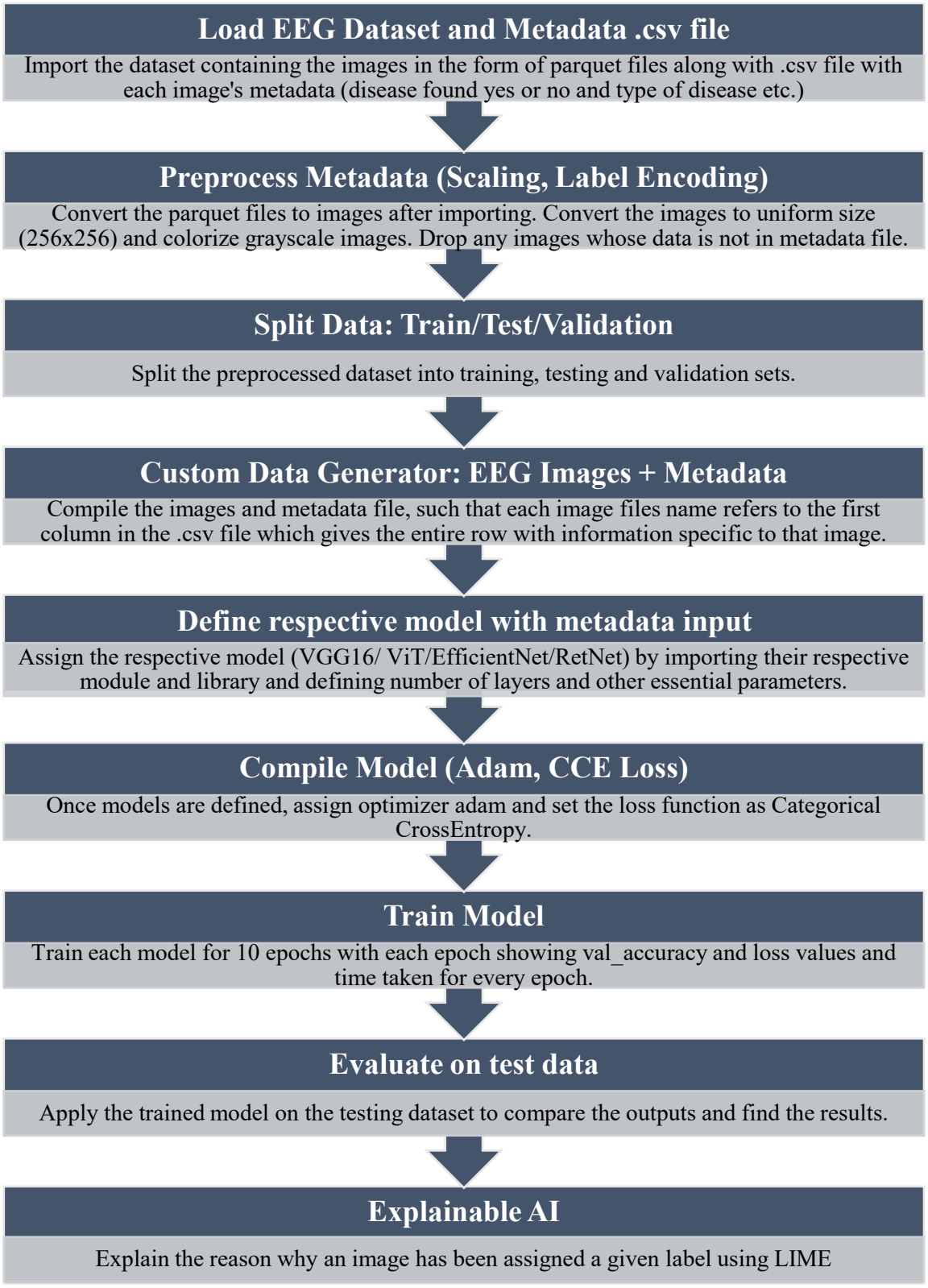


Figure 1. Project Lifecycle Flow Chart

To prevent data leakage, the dataset was partitioned using recording-level separation, ensuring that spectrograms derived from the same EEG recording were not distributed across the training, validation, and testing sets. This approach preserves strict independence between evaluation splits and enables a more reliable assessment of model performance.

The data files are initially stored in Parquet format and are linked with corresponding metadata files in CSV format. These Parquet files are then processed to generate spectrogram images, after which appropriate pre-processing steps are applied to enhance data quality and consistency. The dataset is split into training and testing sets using an 80:20 ratio, ensuring sufficient data for both model learning and evaluation.

The models used for training and comparison are VGG16, Vision Transformer, EfficientNet, and RetNet. This work focuses on identifying the model that achieves the highest accuracy in detecting harmful brain activity from EEG images. To evaluate the results, standard metrics such as precision, recall, F1-score, and support are used.

Data Pre-processing

Pre-processing is an important step before applying any deep learning model to EEG datasets, as it ensures that the data are in a suitable format for training. EEG signals are typically raw and noisy. Therefore, pre-processing consists of multiple stages, such as converting raw EEG signals into spectrograms, which serve as input images for deep learning models.

i. Noise Removal from EEG signals

The process begins by filtering out noise and artifacts from the EEG data, retaining only the relevant patterns of brain activity. This is generally achieved using band-pass filters that exclude frequency bands of no interest, such as those unrelated to seizures or other abnormal brain activity.

ii. Conversion to spectrograms

Once the EEG signal has been cleaned, it is converted into a spectrogram, a time-frequency representation of the signal. This process transforms the raw data into 2D images that can be used as input to convolutional neural networks (CNNs) and other deep learning models. Short-Time Fourier Transform (STFT) is commonly used to generate spectrograms by decomposing the signal into its frequency components over time.

iii. Resizing

After this step, the spectrograms are resized to standard input dimensions (e.g., 224 × 224 pixels) to ensure uniformity across all images.

iv. Normalization

Finally, normalization techniques are applied to scale pixel values, typically to a range between 0 and 1 or −1 and 1, depending on the model requirements.

VGG-16

The VGG16 model consists of 16 layers, comprising 13 convolutional layers, 3 fully connected layers, and 5 max-pooling layers. The core idea of VGG16 is to use very small 3 × 3 filters throughout the network to gradually extract features from the input image, which in this case is an EEG spectrogram.

- Input Layer: The input EEG spectrogram image is fed into the model, and each spectrogram is pre-processed to a fixed dimension (e.g., 224 × 224 pixels).
- Convolutional Layers: The initial layers of the network consist of pairs of convolutional layers with small filters of size 3 × 3. These filters move across the input image to detect low-level features such as edges, lines, and simple textures. Each convolution is followed by activation functions such as ReLU (Rectified Linear Unit), which introduce non-linearity and enable the model to learn complex patterns.
- Max-Pooling Layers: Max-pooling is applied after each set of convolutional layers to reduce the spatial dimensions of the feature maps while retaining important information.
- Fully Connected Layers: After feature extraction, the output of the final convolutional layer is a set of feature maps. These are flattened and used as input to the fully connected layers.
- Softmax Output Layer: This layer is applied after the final fully connected layer and produces a probability distribution over the different classes.

EfficientNet

EfficientNet is designed to be highly efficient and can achieve strong performance using fewer parameters than traditional CNN architectures such as VGG16.

- Input Layer: The resized EEG spectrogram is fed into the model after being adjusted to a fixed input size. EfficientNet processes these spectrograms through a sequence of mobile inverted bottleneck convolution (MBConv) blocks.
- MBConv Blocks: Each MBConv block is computationally efficient and utilizes depthwise separable convolutions to reduce the number of parameters while maintaining performance. These blocks also incorporate squeeze-and-excitation layers, which dynamically adjust the importance of different channels.
- Depth-wise Separable Convolutions: Unlike standard convolutions, depthwise separable convolutions operate on each channel independently, significantly reducing computational cost while still capturing important features from the spectrogram.
- Pooling and Bottleneck Layers: After feature extraction, the model down-samples feature maps

using global average pooling and bottleneck layers. This helps compress the learned representations into a dense feature vector.
- Classification Head: The pooled output is passed through a fully connected classification layer. The classification head employs a softmax activation function to predict the probability of different brain activity classes.

Vision Transformer (ViT)

Vision Transformer (ViT) is a novel deep learning architecture that diverges from standard CNNs by processing an image as a sequence of patches, which are fed into transformer layers. The design of ViT enables it to learn long-range dependencies in spectrograms and recognize complex and nuanced patterns of abnormal brain activity.

- Input Layer: The EEG spectrogram is divided into patches, typically of size 16 × 16 pixels. Each patch is flattened into a one-dimensional vector and projected into a higher-dimensional embedding space. These embeddings are then passed through transformer layers.
- Patch Embedding: The image (spectrogram) is split into patches, and each flattened patch is mapped to a fixed-size vector using a learnable linear transformation. These vectors are combined with positional embeddings to preserve the spatial information of the patches.
- Transformer Layers: The core of the ViT model consists of multiple transformer layers that utilize self-attention mechanisms. Each patch attends to other patches in the spectrogram based on their relevance to the task. Self-attention enables the model to capture long-range relationships and focus on important patterns, such as frequency changes that may indicate seizures or other abnormal brain activity.
- Feed-Forward Layers: The output of the attention mechanism is passed through feed-forward layers, which are fully connected networks using ReLU activation. These layers help the model learn complex relationships and refine the representations obtained from the attention mechanism.
- Classification Head: The final output of the transformer layers is pooled (typically using global average pooling) and fed into a fully connected classification head. This component produces the final predictions of brain activity classes.
- Training: Similar to VGG16, the Vision Transformer is trained using categorical cross-entropy loss and the Adam optimizer. Self-attention mechanisms help ViT learn subtle relationships within the spectrogram, improving its ability to detect fine-grained patterns associated with abnormal brain activity.

Despite their advantages, transformer-based models have certain limitations. One major drawback is their quadratic computational complexity, as each token attends to every other token in the sequence. This becomes computationally expensive for longer sequences, increasing both processing time and resource requirements. Additionally, transformers have high memory demands. Models such as BERT process up to 512 tokens simultaneously, and as sequence length increases, memory consumption grows significantly. Furthermore, transformers may sometimes focus on global patterns at the expense of capturing finer local details, which can reduce their effectiveness in certain tasks.

RetNet (Retentive Networks)

The RetNet model [35] is a deep residual network that uses skip (residual) connections between layers to enable the learning of deeper representations without suffering from the vanishing gradient problem. The retention architecture supports three computation paradigms: parallel, recurrent, and chunk-wise recurrent. The recurrent formulation enables low-cost O(1) inference, improving decoding throughput, reducing latency, and lowering GPU memory usage without sacrificing performance. The chunk-wise recurrent formulation facilitates efficient long-sequence modeling with linear complexity, where each chunk is encoded in parallel while being recurrently summarized. This capability is particularly useful for EEG spectrogram analysis, where complex patterns of brain activity require high-level abstraction to detect abnormal or dangerous conditions.

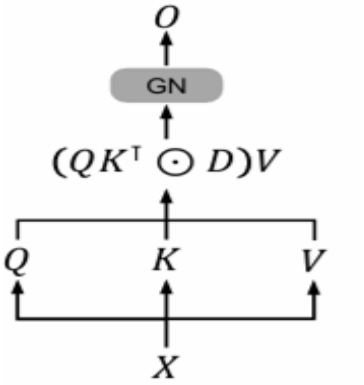
RetNets have following layers:

- Input Layer: Images fed into the RetNet model are passed through an initial sequence of convolutional layers, which extract low-level features such as edges and textures, similar to conventional CNNs.
- Residual Blocks: Residual blocks form the core of the RetNet architecture. Within each block, the input is added to the output via a skip connection. This preserves important information across layers and enables effective training of deeper networks.
- Convolutional Layers: Due to the presence of residual connections, these layers can learn high-level representations without encountering vanishing gradient issues, even in deep networks.
- Global Pooling: After a series of residual blocks, global average pooling is applied to the feature maps to down-sample them and produce a compact feature representation.

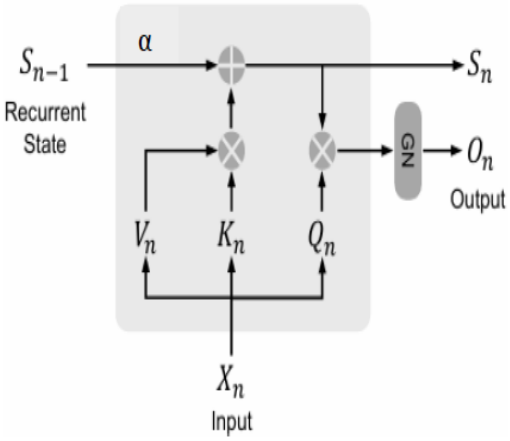
- Fully Connected Layer: The resulting feature vector is then passed to a fully connected layer, which performs the final classification.

Similar to other models, RetNet is trained using categorical cross-entropy loss and the Adam optimizer. A key strength of RetNet lies in its ability to capture both global abstract features and fine-grained local patterns through residual connections across layers.

RetNet uses a "parallelized (vectorized) retention" computation, implemented with convolution or cumulative sum operations, to accelerate training on GPUs. A typical retention block includes input projections (to generate queries, keys, and values), retention computation over past values, and a residual connection to produce the final output. Multiple retention heads with distinct decay parameters and weights enable the model to learn diverse memory patterns, similar to multi-head attention. Their outputs are concatenated and projected to form the final representation (see Figure 2).



(a) Parallel representation



(b) Recurrent mechanism

Figure 2. Dual Nature of RetNet [35]

i. Parallel formulation

During training, RetNet can process the entire sequence in parallel, similar to transformers. The parallel formulation of the retention mechanism computes outputs for all time steps simultaneously. Let W_Q, W_K and W_V represent learnable weight matrices for queries (Q), keys (K), and values (V), respectively, as in transformers. Let "D" represent a decay matrix that applies the decay factor "α" to weight the contributions of past tokens. It is typically a lower-triangular matrix with entries that decay exponentially based on the distance between time steps.

For a sequence X, the retention output is computed as shown in Equation-1, where ⊙ represents Hadamard (element-wise) product of matrices.

$$Y = Retention(X) = [(XW_Q)(XW_k)^T \odot D](XW_V) \quad (1)$$

Group Normalization (GN), is a normalization method that divides the channels of a layer's activations into smaller groups and normalizes the activations within each group.

The final output is obtained by applying the output projection as given in Equation-2.

$$O = YW_O \quad (2)$$

ii. Recurrent formulation

The retention mechanism is analogous to the attention mechanism in transformers but is designed to be more efficient. It computes a weighted combination of past inputs while maintaining a state that summarizes the sequence. For a sequence of input vectors X_n, the retention mechanism updates a state S_n, which represents a compressed history of the sequence up to time "n". Let α denote the decay factor that controls how much of the previous state is retained (a larger value assigns more weight to past information).

The retention mechanism can be expressed as given in Equation-3.

$$S_n = \alpha S_{n-1} + (1-\alpha) W_k X_n \cdot (W_Q X_n)^T (W_V X_n) \quad (3)$$

The output of the retention layer at time "n" is typically a combination of the state and the current input as described in Equation-4.

$$O_n = W_O S_n \quad (4)$$

iii. Multi-Head Retention

Similar to multi-head attention in transformers, RetNets can use multiple retention heads to capture different aspects of the sequence. For H heads, the input is split into H subspaces, and the retention mechanism is applied independently to each subspace.

3. RESULTS AND DISCUSSIONS

In this section, the results are presented and discussed. All models were trained for 10 epochs on an NVIDIA

DGX H100. The results include evaluation metrics such as precision, recall, accuracy, F1 score, and support. Multiple-run averaging was performed to obtain the reported results for each model, wherein 10 runs of each model were conducted and averaged to compute the final accuracy.

The same hyper-parameters and pre-processing strategies were used across all models to ensure experimental consistency. Table 1 provides a summary of the corresponding training configurations.

Table 1. Hyper-parameters

Model Name	Learning Rate	Batch Size
VGG16	1e-4 (Adam)	32
Vision Transformer (ViT)	1e-4 (Adam)	32
EfficientNetB0	1e-4 (Adam)	32
RetNet	1e-4 (Adam)	32

Among the four models, RetNet achieved the highest validation accuracy 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.

These results, visualized in Figure 3, highlight that RetNet achieved the highest accuracy, confirming its strong capability to identify complex features in EEG spectrograms. Each model's performance during training was analyzed using line graphs (Figures 4 and 5), which display the loss and validation accuracy over epochs.

- VGG16: The VGG16 model exhibited relatively slow convergence, with a steady decrease in loss and a moderate increase in validation accuracy. However, its validation accuracy plateaued at approximately 85% by the tenth epoch, suggesting that the model may have reached its learning capacity for this dataset.
- Vision Transformer (ViT): The ViT model showed a steeper decline in loss and a significant increase in validation accuracy, reaching approximately 91% in the later stages of training. This suggests that ViT's attention mechanism effectively focused on critical features in the EEG data, enabling more efficient learning.
- EfficientNet: EfficientNet demonstrated a highly effective learning curve, with a sharp reduction in loss and validation accuracy reaching around 96% within a few epochs. This indicates that EfficientNet efficiently utilized computational resources and had strong generalization capability, resulting in high performance.
- RetNet: The RetNet model exhibited the best training dynamics, with a consistent decrease in loss and an increase in validation accuracy to approximately 96%. The model's strong performance can be attributed to its design for temporal data processing, which is well suited to EEG spectrogram classification.

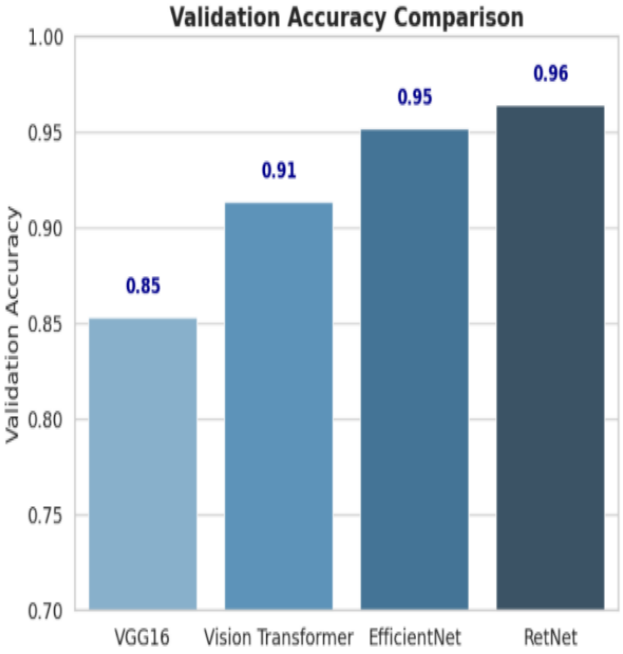


Figure 3. Comparison of Loss and Validation Accuracy across the models

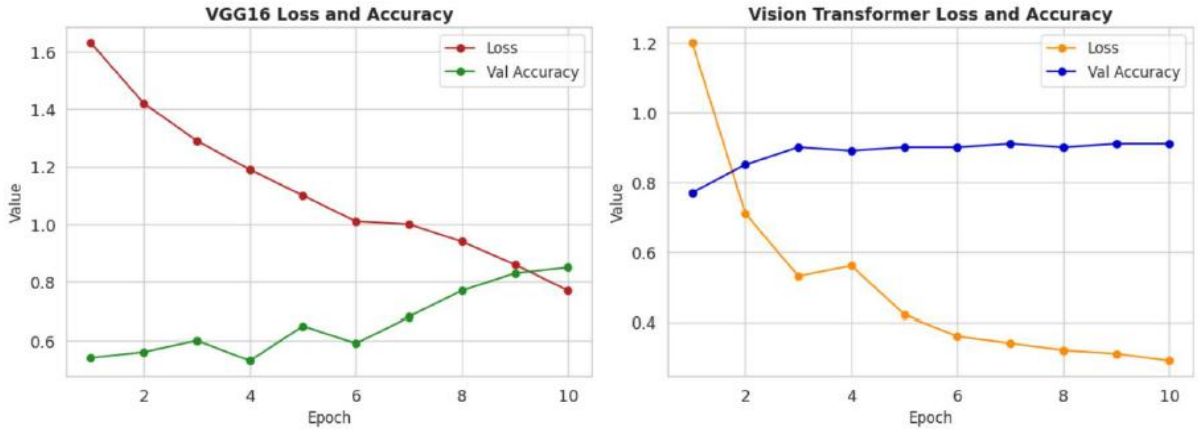


Figure 4. Loss and Validation Accuracy of VGG16 and ViT

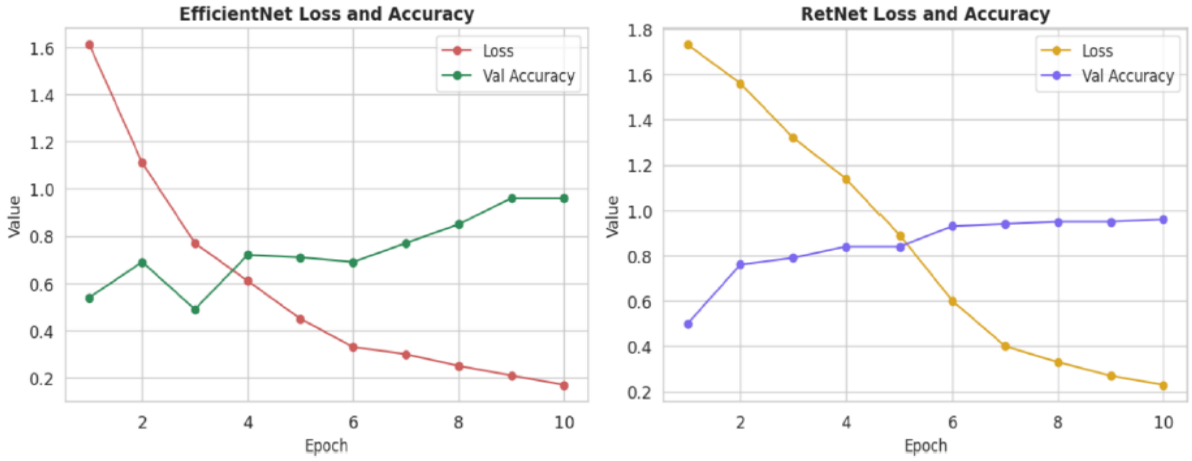


Figure 5. Loss and Validation Accuracy of EfficientNet and RetNet

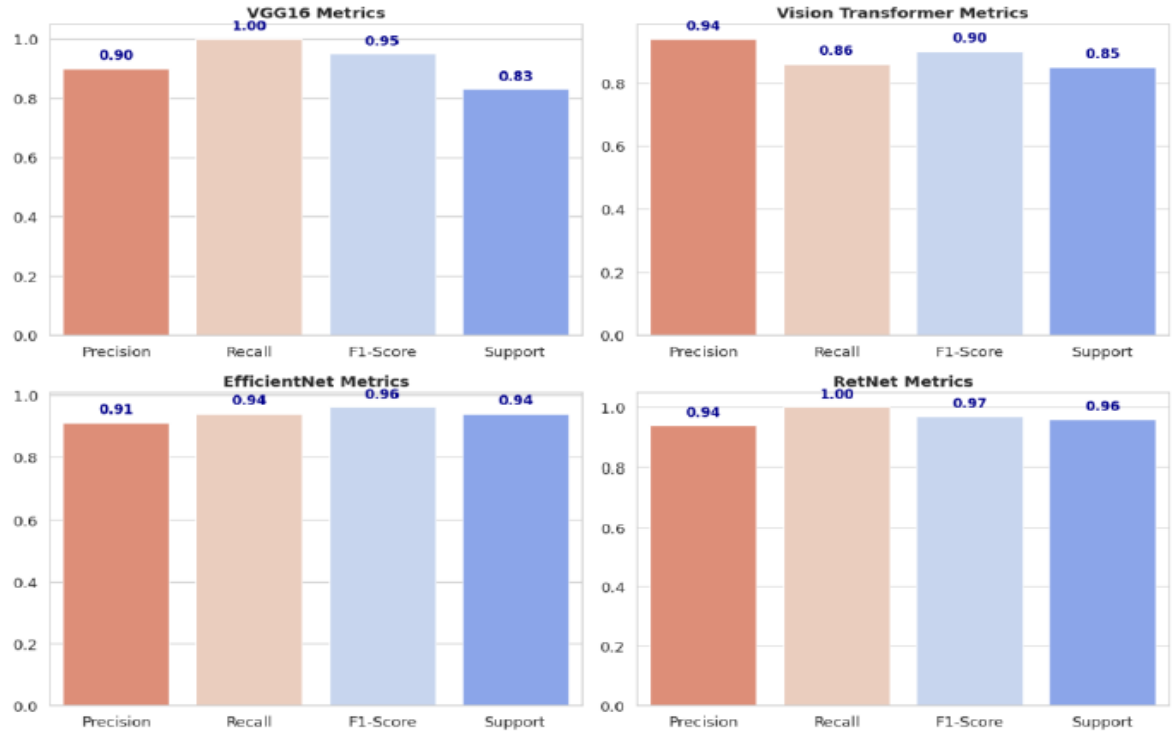


Figure 6. Precision, Recall, F1-score and Support comparison across all models

Apart from validation accuracy, the models were also evaluated using precision, recall, F1-score, and support. Figure 6 presents a comparative analysis in the form of bar charts.

- RetNet: The RetNet model achieved the highest performance across all metrics, with a precision of 94%, recall of 100%, and an F1-score of 97%. This indicates that its high accuracy is supported by a well-balanced performance.
- EfficientNet: EfficientNet performed comparably, with a precision of 91%, recall of 90%, and an F1-score of 96%, demonstrating strong classification capability, although slightly lower than RetNet.
- Vision Transformer (ViT): The ViT model performed well, achieving precision, recall, and F1-scores of 94%, 86%, and 90%, respectively. However, it was outperformed by the other models.
- VGG16: Despite lower validation accuracy, VGG16 reported strong performance metrics, with 90% precision and 100% recall. Its F1-score of 95% suggests a trade-off affecting overall generalizability across the dataset.

In addition to quantitative performance measures, the inclusion of explainable artificial intelligence (XAI) provides valuable insight into the model's decision-making process. Clinical validity in EEG-based brain activity classification depends not only on accuracy but also on interpretability, with predictions supported by meaningful biomedical explanations. Using the LIME algorithm, post hoc visual explanations were generated

for each class, highlighting the time–frequency regions most significant to the model's predictions (Figure 7). The LIME visualizations highlight specific time–frequency regions of the EEG spectrogram that have the greatest influence on the model's predictions. These regions often correspond to areas with increased spectral energy and persistent frequency patterns, which data-driven models use to identify abnormal brain activity. This qualitative analysis provides insight into how the model makes decisions, even without explicit neurophysiological knowledge.

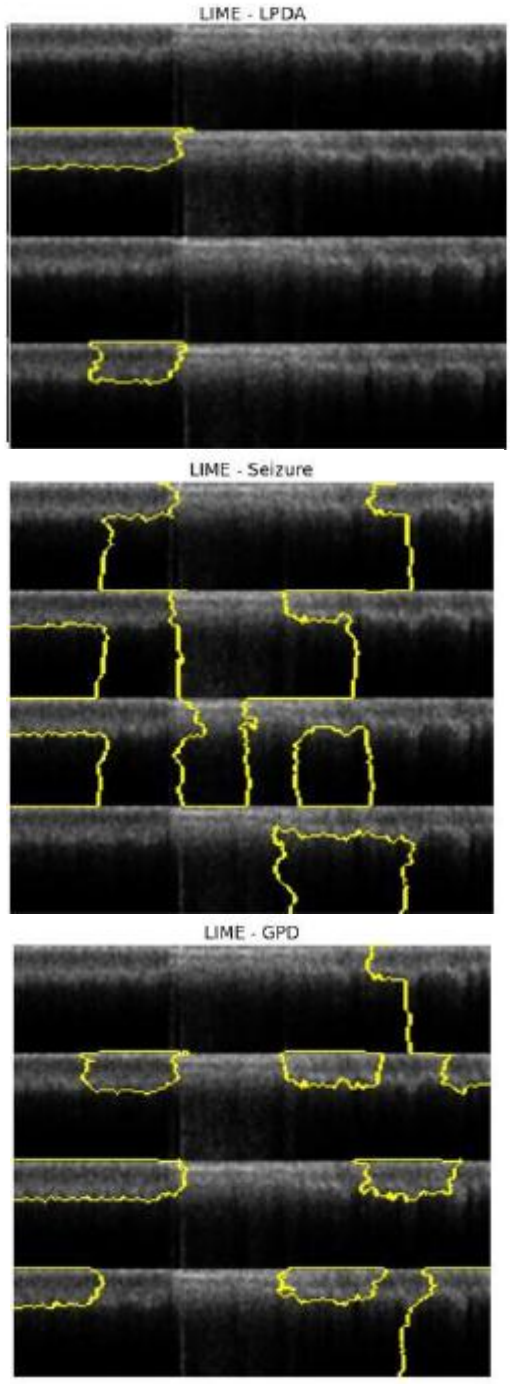


Figure 7. LIME representations of different classes

5. CONCLUSION

In conclusion, this study presented an efficient and explainable deep learning framework for harmful brain activity classification using EEG spectrogram images. The RetNet-based approach showed clear advantages in classification, multimodal learning enhancement and interpretively, marking it as a potential candidate for real-time monitoring of brain activity and decision-support system.

- A RetNet based framework which helps to effectively classify the harmful brain activity from the EEG spectrogram images has a lower computational complexity than transformer-based.
- Among the architectures tested in this study, RetNet model showed the highest overall performance with 94% precision, 100% recall and 97% F1-score, which is better than EfficientNet models and Vision Transformer models.
- EEG spectrogram images combined with metadata demonstrated a multi-modal learning strategy for closer view of the brain by improved model capturing complex and meaningful patterns related to brain activity.
- Use of Explainable Artificial Intelligence (XAI) techniques, specifically LIME, facilitated transparency and trustworthiness through visual explanations and identification of physiologically meaningful areas that influenced the model's classification decisions.
- The suggested framework shows significant promise for real-time harmful brain activities detection, early diagnosis of neurological disorders and support to clinical decision making due to its high performance, interpretability and computational efficiency.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[2] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: An overview," in Advanced Machine Learning Technologies and Applications (AMLTA 2020), A. Hassanien, R. Bhatnagar, and A. Darwish, Eds. Singapore: Springer, 2021, pp. 599–608, doi: 10.1007/978-981-15-3383-9_55.

[3] S. Razavi, "Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling," Environmental Modelling & Software, vol. 144, p. 105159, Oct. 2021, doi: 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," IEEE Sensors Journal, vol. 22, no. 5, pp. 4359–4368, Mar. 2022, doi: 10.1109/JSEN.2021.3133313.

[5] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, et al., "A deep transfer convolutional neural network framework for EEG signal classification," IEEE Access, vol. 7, pp. 112767–112776, Aug. 2019, doi: 10.1109/ACCESS.2019.2935945.

[6] M. Habijan, R. Šojo, I. H. Tolić, and I. Galić, "Harmful brain activity classification using ensemble deep learning," in Proc. 2024 International Symposium ELMAR, Zadar, Croatia, Sept. 16–19, 2024. Piscataway, NJ, USA: IEEE, 2024, pp. 109–112, doi: 10.1109/ELMAR62776.2024.10722449.

[7] S. Rajwal and S. Aggarwal, "Convolutional neural network-based EEG signal analysis: A systematic review," Archives of Computational Methods in Engineering, vol. 30, no. 6, pp. 3585–3615, Jun. 2023, doi: 10.1007/s11831-023-09915-z.

[8] B. Chakravarthi, S. C. Ng, M. R. Ezilarasan, and M. F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," Frontiers in Computational Neuroscience, vol. 16, p. 1019776, Nov. 2022, doi: 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, vol. 161, p. 105629, Feb. 2025, doi: 10.1016/j.imavis.2025.105629.

[10] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, et al., "CNN Explainer: Learning convolutional neural networks with interactive visualization," IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1396–1406, Feb. 2021, doi: 10.1109/TVCG.2020.3030418.

[11] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional neural network (CNN) for image detection and recognition," in Proc. 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, Dec. 15–17, 2018, pp. 278–282, doi: 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," in Proc. 2021 2nd China International SAR Symposium (CISS), Shanghai, China, Nov. 3–5, 2021, pp. 1–3, doi: 10.1109/CISS52396.2021.9741235.

[13] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in Proc. 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, Dec. 19–21, 2019, pp. 94–98, doi: 10.1109/ICIT48102.2019.00023.

[14] M. Rashid, M. Mustafa, N. Sulaiman, and M. N. Islam, "EEG and EMG-based multimodal driver drowsiness detection: A CWT and improved VGG-16 pipeline," in Proceedings of the 2nd Human Engineering Symposium (HUMENS 2023), Singapore: Springer, 2024, pp. 337–349, doi: 10.1007/978-981-99-8259-9_28.

[15] L. Li, "Deep learning-based EEG signal identity recognition using VGGNet," in Proc. 2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, Jan. 19–21, 2024, pp. 1092–1095, doi: 10.1109/NNICE60191.2024.10467718.

[16] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. G. de Peralta, "EEG source imaging," Clinical Neurophysiology, vol. 115, no. 10, pp. 2195–2222, Oct. 2004, doi: 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," in Proc. 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, Apr. 6–11, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889246.

[18] M. X. Cohen, "Where does EEG come from and what does it mean?" Trends in Neurosciences, vol. 40, no. 4, pp. 208–218, Apr. 2017, doi: 10.1016/j.tins.2017.02.004.

[19] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review," Neural Computing and Applications, vol. 35, no. 20, pp. 14681–14722, Oct. 2023, doi: 10.1007/s00521-022-07959-3.

[20] X. Deng, H. Huo, L. Ai, D. Xu, and C. Li, "A novel 3D approach with a CNN and Swin Transformer for decoding EEG-based motor imagery classification," Sensors, vol. 25, no. 9, p. 2922, May 2025, doi: 10.3390/s25092922.

[21] L. Hallal, J. Rhinelander, R. Venkat, and A. Newman, "Efficient feature extraction for EEG-based classification: A comparative review of deep learning models," AI, vol. 7, no. 2, p. 50, Feb. 2026, doi: 10.3390/ai7020050

[22] Z. Gao, W. Dang, X. Wang, X. Hong, L. Hou, K. Ma, et al., "Complex networks and deep learning for EEG signal analysis," Cognitive Neurodynamics, vol. 15, no. 3, pp. 369–388, Jun. 2021, doi: 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," in Proc. 2024 IEEE 8th International Conference on Information and Communication Technology (CICT), Prayagraj, UP, India, Feb. 16–18, 2024, pp. 1–6, doi: 10.1109/CICT62185.2024.10544356.

[24] S. Ganesan, Y. N. Kiran, and S. Ram, "Harmful brain activity classification of spectrograms with transfer deep learning," Research Square, preprint, 2024, doi: 10.21203/rs.3.rs-5507813/v1.

[25] M. A. Li and D. Q. Xu, "A transfer learning method based on VGG-16 convolutional neural network for MI classification," in Proc. 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, May 22–24, 2021, pp. 5430–5435, doi: 10.1109/CCDC52312.2021.9602105.

[26] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," AI Open, vol. 3, pp. 111–132, Dec. 2022, doi: 10.1016/j.aiopen.2022.10.001.

[27] C. Chen, H. Wang, Y. Chen, Z. Yin, X. Yang, H. Ning, et al., "Understanding the brain with attention: A survey of transformers in brain sciences," Brain-X, vol. 2, no. 2, p. e29, Jun. 2024, doi: 10.1002/brx2.29.

[28] H. Adeli, S. Minni, and N. Kriegeskorte, "Predicting brain activity using transformers," bioRxiv, preprint, 2023, doi: 10.1101/2023.09.14.557711.

[29] J. Xie, J. Zhang, J. Sun, Z. Chen, Y. Yang, Y. Zhang, et al., "A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2126–2136, Oct. 2022, doi: 10.1109/TNSRE.2022.3201937.

[30] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," Technologies, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/technologies13010032.

[31] S. Dongre and S. Mehta, "RetViT: Retentive vision transformers," in Proc. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jun. 18–22, 2024, pp. 1–8, doi: 10.1109/ICCCNT61001.2024.10724924.

[32] M. M. Islam, M. A. Talukder, M. A. Uddin, A. Akhter, and M. Khalid, "BrainNet: Precision brain tumor classification with optimized EfficientNet architecture," BioMed Research International, vol. 2024, p. 3583612, Feb. 2024, doi: 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," Bioengineering, vol. 12, no. 2, p. 109, Feb. 2025, doi: 10.3390/bioengineering12020109.

[34] M. U. K. Naik and S. R. Ahamed, "Wavelet-based autoencoder and EfficientNet for schizophrenia detection from EEG signals," arXiv preprint, 2024, doi: 10.48550/arXiv.2405.15463.

[35] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, et al., "Retentive network: A successor to transformer for large language models," arXiv preprint, 2023, doi: 10.48550/arXiv.2307.08621.

[36] G. K. Erabati and H. Araujo, “Retformer: Embracing point cloud transformer with retentive network,” IEEE Transactions on Intelligent Vehicles, 2024, doi: 10.1109/TIV.2024.3448057.
[37] H. Yang, Z. Li, Y. Chang, and Y. Wu, “A survey of retentive network,” arXiv preprint, 2024, doi: 10.48550/arXiv.2410.01201.
[38] D. Umamaheswari, N. Nachammai, and S. Anita, “Early diagnosis of diabetic retinopathy using retinal network,” Multimedia Tools and Applications, 2025, doi: 10.1007/s11042-025-20569-7.
[39] Z. Lin, R. Cui, L. Ning, and J. Peng, “Temporal features-fused vision retentive network for echocardiography image segmentation,” Sensors, vol. 25, no. 6, p. 1909, Mar. 2025, doi: 10.3390/s25061909.
[40] K. Subramaniam and A. Naganathan, “RetNet30: A novel stacked convolution neural network model for automated retinal disease diagnosis,” International Journal of Imaging Systems and Technology, vol. 34, no. 5, p. e23187, Sept. 2024, doi: 10.1002/ima.23187.