Hybrid Vision Transformer and MLP-Mixer for Epileptic Seizure Detection in Intracranial EEG

Thouraya Guesmi[1], Abir Hadriche[2], Nawel Jmail[3]
Miracl Lab, Sfax University, Tunisia[1, 3]
Gabes National Engineering School, Gabes University, Tunisia[1]
Digital Research Center of Sfax, Tunisia[2, 3]
Regim Lab-ENIS, Sfax University, Tunisia[2]

***Abstract*—Accurate and timely seizure detection is essential for effective epilepsy management, and automated systems can play a valuable role in supporting clinical practice. In this study, we introduce a hybrid approach that uses time-frequency representations of Intracranial electroencephalography (iEEG) signals filtered at High-Frequency Oscillations (HFOs) bands as input to different convolutional neural network (CNN) backbones for feature extraction, followed by classification with either a Vision Transformer (ViT) or MLP-Mixer. This work establishes a systematic, comparative framework for benchmarking hybrid CNN-ViT against CNN-MLP-Mixer, providing a critical new reference for automated epileptic seizure detection within HFOs filtered iEEG signals. Extensive evaluation demonstrates that the ViT consistently achieves superior performance, with an EfficientNetB0-ViT model attaining remarkable accuracy (97.85%) and specificity (98.92%). Crucially, the MLP-Mixer emerges as a highly competitive alternative, exhibiting strong recall capabilities that make it suitable for applications where missing a seizure is not an option. Overall, our findings suggest that self-attention mechanisms in ViTs provide a distinct advantage for capturing complex seizure dynamics, yet MLP-based models present a powerful, efficient option.**

Keywords*—Vision transformer; MLP-Mixer; iEEG; HFOs; ResNet; GoogleNet; EfficientNetB0

I. Introduction

Epilepsy represents a major global health challenge, characterized by recurrent, unprovoked seizures that disrupt neural function and impose a significant burden on the quality of life for over fifty million individuals worldwide [1]. The clinical management of epilepsy, particularly drug-resistant forms, relies heavily on the precise identification and localization of seizure events [2]. Consequently, the development of accurate and reliable seizure detection systems is paramount for enabling timely intervention and improving patient outcomes [3][4].

Intracranial electroencephalography (iEEG) is commonly used to evaluate patients with refractory epilepsy before surgery. It offers detailed recordings from the brain's surface or deeper areas. Analyzing iEEG visually requires a significant amount of time for neurologists, which slows down clinical workflow. Automated detection algorithms that can accurately review long recordings are becoming increasingly popular due to the need for efficiency and precision.

The evolution of these automated systems has progressed from traditional machine learning techniques, which depend on manually engineered features, to deep learning systems that learn discriminative patterns from raw or preprocessed data [5]. Among these, the ability to analyze time-frequency representations of iEEG signals has been demonstrated by Convolutional Neural Networks (CNNs). They demonstrated exceptional proficiency in analyzing time-frequency representations of iEEG signals, effectively identifying localized morphological patterns associated with seizure onset[6]. Despite their strengths, a fundamental constraint of CNNs is their limited receptive field, which can hinder their ability to model the long-range spatiotemporal dependencies that characterize the dynamic propagation of seizures through neural networks [7].

To address this limitation, Vision Transformers (ViT) have the ability to capture global contextual relationships across an entire input sequence through their self-attention mechanism, potentially offering a more nuanced understanding of complex ictal dynamics [8] [9].

In recent research, hybrid architectures have been explored that combine the local feature extraction capabilities of CNNs with the global contextual modeling of ViTs, which have resulted in significant gains in both accuracy and generalizability for seizure detection tasks.[10].

Parallel to these developments, the MLP-Mixer architecture presents a notably different approach. As articulated by Tolstikhin et al. (2021), "In contrast to ViT, which relies on self-attention, the MLP-Mixer proposes an architecture based solely on Multi-Layer Perceptrons (MLPs) for spatial and channel mixing, offering a competitive accuracy-efficiency trade-off[11]". While MLP-Mixers have shown promise in other domains, their application to the specific challenge of detecting epileptic seizures from iEEG signals filtered in the high-frequency oscillation (HFO) bands remains a critical, largely unexplored research gap.

This study addresses this gap by investigating the fundamental trade-offs between these novel sequence modeling approaches when applied to a highly sensitive biomarker: HFO-filtered iEEG data. This leads to our core research question: What are the distinct performance, efficiency, and mechanism trade-offs between hybrid CNN-ViT and CNN-MLP-Mixer systematically benchmarked for automated epileptic seizure detection in the HFOs domain?

In summary, our contributions are as follows: We introduce a systematic and novel comparative framework for benchmarking the performance and mechanism trade-offs of hybrid CNN-Vision Transformer and CNN-MLP-Mixer architectures.

We establish a novel deep learning reference for the automated detection of epileptic seizures in the High-Frequency Oscillation (HFO) domain of iEEG, addressing a major gap in the current literature.

We provide a detailed performance analysis that contrasts the advantages of the self-attention mechanism versus the purely algebraic mixing of MLP-Mixer, highlighting critical considerations for clinical deployment.

We systematically evaluate six widely used CNN backbones (VGG19, ResNet18, ResNet50, ResNet101, GoogleNet, EfficientNet-B0) to provide a comprehensive baseline for feature extraction in HFO analysis.

This study is organized as follows: Following the introduction, a theoretical framework reviews relevant literature (Section II). Details on materials and methods are provided in Section III. The results are presented in Section IV, followed by a comprehensive discussion of the findings, their implications, and the limitations of the study in Section V. The study concludes in Section VI.

II. Related Works

The pursuit of automated seizure detection has progressively shifted from reliance on manually crafted features to the use of deep learning models. Within this evolution, [12] showed that HFOs detected by neuromorphic neural networks are strong predictors of seizure prognosis across various recording modalities, including iEEG. This work demonstrates the clinical relevance of HFO-based detection but relies on hand-crafted feature engineering rather than end-to-end deep learning. Other research works [13][14] substantiate the strong correlation between HFOs and epileptogenic tissue, establishing a solid foundation for their use in automated prognosis systems. The strength of these studies lies in validating HFOs as critical biomarkers for seizure detection. However, a key weakness is the limited exploration of modern sequence modeling architectures (Transformers, MLP-Mixers) that could better capture the temporal dynamics of HFO patterns. The undesirable assumption often made is that traditional feature engineering is sufficient to capture the intricate, non-linear dynamics of HFOs.

A critical step in applying modern deep learning architectures to signal data involves transforming one-dimensional time series into a format that can be used with image-based models. In this context, Zhuohan Wang et al [15] demonstrated that transforming iEEG signals into scalograms facilitates the use of image-based deep learning models. This approach enables the model to simultaneously leverage temporal and spectral information. We build upon this foundation, using time-frequency transformation as the essential input for our Vision Transformer and MLP-Mixer models.

In the context of iEEG analysis, CNNs have demonstrated superior performance in detecting epileptic seizures by learning local discriminative patterns in spectrograms, with variants such as ResNet, GoogLeNet and VGG being widely used to extract meaningful features from biomedical signals [16]. While excellent at capturing features within a limited spatial context, they often struggle to integrate information across distant regions of a scalogram, which is crucial for understanding the full spatial-temporal propagation of a seizure. This limited and fixed receptive field is the core weakness of CNN-only approaches, limiting their ability to model global seizure connectivity. This inadequacy is precisely why we propose a hybrid framework, pairing the local robustness of CNNs with sequence models capable of global context aggregation.

The MLP-Mixer architecture challenges the dominance of convolutions and self-attention by relying solely on a multilayer perceptron, applied separately to spatial (token-mixing) and channel (feature-mixing) dimensions [11]. Its simplicity and computational efficiency make it attractive for EEG-based applications. Recent studies have shown that MLP-Mixers, enhanced with attention mechanisms, can effectively decode EEG motor imagery tasks [17]. Despite this, the specific efficacy of the MLP-Mixer for the nuanced task of iEEG-based seizure detection, particularly when operating on HFOs, remains an open question. The current literature lacks a systematic comparison between the self-attention mechanism (ViT) and the purely algebraic mixing (MLP-Mixer) to determine which is inherently superior for modeling HFO dynamics. This forms the central motivation and unique contribution of our study.

III. Materials and Methods

This section details the methodology adopted, from the preparation of raw data to the configuration of hybrid models. Our study proposes a systematic comparative framework for hybrid deep learning CNN-VIT and CNN-MLP-Mixer applied to epileptic seizure detection in iEEG filtered in HFO bands. This study proposes a hybrid pipeline for binary classification between ictal and interictal states. The core approach involves using Convolutional Neural Networks (CNNs) backbones (EfficientNetB0, ResNet18, ResNet50, ResNet101, GoogleNet and VGG19) to extract features from time-frequency images of iEEG signals, which are then classified by either a Vision Transformer (ViT) or an MLP-Mixer. The overall workflow is depicted in Fig. 1.

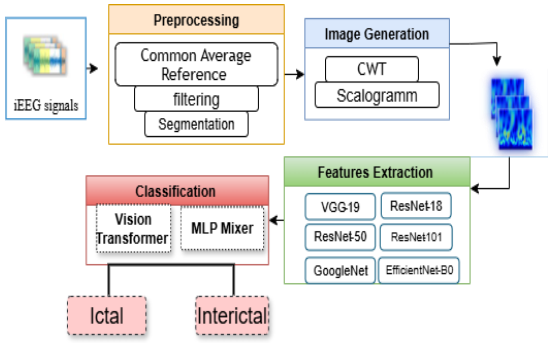

Fig. 1. Workflow of the proposed method.

A. Data Description

This study utilized a subset of the publicly available HUP iEEG dataset [18], which comprises intracranial EEG recordings from patients undergoing evaluation for drug-resistant epilepsy at the Hospital of the University of Pennsylvania. From the original cohort of 58 subjects, we selected 21 patients whose recordings were sampled at 1024 Hz, a frequency sufficient for the reliable analysis of High-Frequency Oscillations (HFOs). The dataset includes ictal and interictal recordings, to precise annotations marking the beginning (onset) and end (offset) of seizure events, allowing for clear identification of ictal periods.

B. Data Processing

The iEEG data preprocessing pipeline consisted of three main stages: segmentation, filtering/re-referencing, and time-frequency transformation.

Segmentation and labeling were performed for the binary classification task of distinguishing ictal from interictal states. The continuous iEEG recordings were segmented into 1-second epochs. Ictal segments were extracted from annotated seizure intervals (onset to offset). Each segment was assigned a binary label (ictal=1, interictal=0), defining the classification target as $y \in \{0,1\}$.

Channels marked as artifactual in the original dataset metadata were discarded. The signals were then re-referenced using a Common Average Reference (CAR) filter computed from all remaining valid channels, a crucial step for reducing common-mode noise. Subsequently, a zero-phase finite impulse response (FIR) filter was applied to isolate the High-Frequency Oscillation (HFO) band of interest (80-250 Hz).

To leverage image-based deep learning models, each 1-second HFO-filtered segment was converted into a 2D time-frequency (TF) (see Fig. 2) representation using the Continuous Wavelet Transform (CWT) with a complex Morlet wavelet. The wavelet function is defined as:

$$\varphi(t) = C_\sigma \pi^{-1/4} e^{-t^2/2} e^{i\sigma t} \quad (1)$$

where, C_σ is the normalization constant and σ is the central frequency parameter. The 2D representation results were scaled to dimensions of 224 × 224 × 3 pixels as input images for the proposed model. The final dataset comprised 4092 samples, which were partitioned into training (80%), validation (10%), and test (10%) sets.

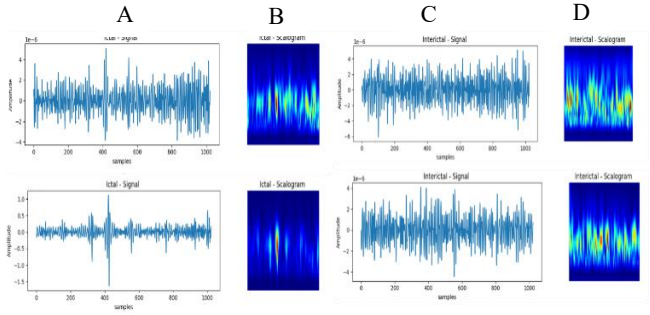

Fig. 2. Transformation of signal into 2D representation: (A) ictal representation in time, (B) ictal representation in TF, (C) interictal representation in time, (D) interictal representation in TF.

C. Model Development

1) CNN backbones: Resnet (Residual Network): Resnet is a fundamental contribution to deep learning, specifically addressing the problem of performance degradation when training extremely deep convolutional neural networks (CNNs). Its main innovation lies in the introduction of residual blocks, which enable the learning of residual mapping functions via skip connections. In the ResNet architecture, the fundamental residual connection in ResNet consists of summing the original input x with the learned transformation F(x) to produce the block's output, according to the following equation:

$$y = F(xw_i) + x \quad (2)$$

ResNet50: This architecture comprises 50 layers, and its structure incorporates batch normalization layers and ReLU activations to stabilize training. Features are extracted from the avgpool layer, resulting in a 2048-dimensional vector.

ResNet18: With 18 layers, its skip connections between convolutional layers, effectively solve the vanishing gradient problem. Although less deep.

ResNet101: This deep architecture has 101 layers and optimizes residual learning between the inputs and outputs of convolutional blocks, enabling richer feature extraction.

VGG19: It is characterized by its fixed depth of 19 layers and its uniform architecture using exclusively 3×3 convolutions.

GoogleNet (Inception): It has a 22-layer architecture optimized for multi-scale feature extraction via its Inception modules, which incorporate Inception blocks that use parallel multi-kernel convolutions. This design enables the model to identify patterns across various spatial scales.

EfficientNet-B0: It is an architecture optimized via compound scaling that balances depth, width, and resolution. The model uses MBConv blocks with expansion-reduction mechanisms and channel attention (SE modules).

For each CNN, the fully connected classification layer is removed. Consequently, the feature extraction configuration for each model is detailed in Table I, which shows the dimensions of the final feature maps extracted from the CNN backbones for an input image size of 224×224×3.

TABLE I. DIMENSIONS OF THE FINAL FEATURE MAPS OF THE CNN BACKBONES

Backbone	Feature Map [B, C, H, W]
ResNet18	[B, 512, 7, 7]
ResNet50	[B, 2048, 7, 7]
ResNet101	[B, 2048, 7, 7]
VGG19	[B, 512, 7, 7]
GoogleNet	[B, 1024, 7, 7]
EffecientNet-B0	[B, 1280, 7, 7]

B = batch size.

C = number of channels (feature dimension).

H, W = spatial dimensions (here 7×7 for input 224×224).

2) Vision transformer: Transformers are an important advancement in deep learning. They were first created for natural language processing [19] but now help solve many sequence modeling problems. Because they can handle complex, long-range relationships, they are useful in areas like computer vision [19] and biomedical data analysis [20]. Specifically, The Vision Transformer (ViT) is a transformer that has been adapted for image inputs. It works by dividing images into patches, which are then embedded linearly. Each patch vector is then enriched with a position embedding to preserve spatial information. The embeddings are then subjected to a series of Transformer blocks. The blocks that make it up include Multi-Head Self-Attention (MHSA), Layer Normalization (LN), and Feed-Forward Network (FFN). The Multi-Head Self-Attention (MSA) calculates the interactions between all tokens in the sequence. For each head, linear projections generate the matrices query (Q), key (K), and value (V) vectors, as shown in the following equations:

$$\mathrm{Q}=\mathrm{XW}^{\mathrm{Q}}\ ,\quad K=XW^{K}\ ,\ V=XW^{V} \tag{3}$$

Let W^{Q}, W^{K}, and W^{V} denote the parameter weight matrices. Then, the attention mechanism computes a weighted sum of the values, where attention coefficients are derived from a normalized query-key similarity as represented in the following equation:

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{4}$$

in which, normalization by $\sqrt{d_k}$ aims to stabilize gradients during learning. Subsequently, the output of the attention layer was subjected to Layer Normalization (LN), a process that has been demonstrated to stabilize the training process (see equation below).

$$LN(z)=\frac{z-\mu}{\sqrt{\sigma^{2}+\epsilon}}.\gamma+\beta \tag{5}$$

The input's meaning and standard deviation are μ and , while $\gamma+\beta$ are acquired parameters. A non-linear GELU activation separates two linear layers in a Feed-Forward Network (FFN):

$$FFN(z)=\mathrm{W}_{2}.GELU(\mathrm{W}_{1}.z+\mathrm{b}_{1})+\mathrm{b}_{2} \tag{6}$$

With the b being biases and W representing weight matrices.

3) Proposed CNN-Vision Transformer hybrid model: This proposed hybrid model combines the strength of CNNs in extracting features, which are subsequently converted into tokens and processed by a Transformer encoder. This fusion skillfully leverages CNNs for capturing detailed local features and the Transformer for global context modeling via the self-attention mechanism, the architecture proceeds as follows:

Feature extraction by CNN: Let an input image (spectrogram) be denoted as $I\in\mathbb{R}^{224\times224\times3}$ which is processed by a pre-trained CNN network(ReNet18, ResNet50, ResNet101, VGG19, GoogleNet, EfficientNet-B0), producing a feature map $F\in\mathbb{R}^{7\times7\times C_{nn}}$ whose final convolutional outputs are 7×7 spatial maps for 224×224 inputs, the output of each CNN is detailed in Table I where C_{nn} represents the number of output channels.

The transition to the ViT: The resulting feature maps for each CNN are linearly projected to embedding dimension $D=768$ using 2D (1×1) conventional with (padding='same') followed by reshaping to obtain $(7\times7,768)$. Consequently, $N=7\times7=49$ tokens. This embedding dimension (D=768) is adopted from the ViT-Base specification, which balances model capacity with computational efficiency for sequences of moderate length. Then, a learnable [CLS] token is prepended to the token sequence. Learnable positional embedding $E=49+1=50$ is added to preserve spatial information.

Transformer Encoder: processes the input sequence through twelve encoder layers, where each layer helps the model understand context and share information between tokens. The model can focus on multiple parts of the input sequence simultaneously by using a multi-head self-attention mechanism in each encoder layer that has twelve attention heads.

Classification: Each encoder layer also has an MLP block with 3072 dimensions (3072 = 4×768 to provide sufficient non-linear transformation capacity) and uses GELU activation. Following the encoder layers, the final state of the classification token is extracted, layer-normalized, and processed by a multilayer perceptron (MLP) classification head to produce the final class predictions, where the classification token gathers information from all tokens using self-attention. The final prediction is obtained through a sigmoid activation function:

$$MLP-Head(z_{cls})=W_{2}.GELU(W_{1}.LayerNorm(z_{cls})+b_{1})+b_{2} \tag{7}$$

$$\hat{y}=\sigma(W_{class}.MLP-Head(z_{cls})+b_{class}) \tag{8}$$

where, $z_{cls}\in\mathbb{R}^{768}$ is the final cls token, $W_{1}\in\mathbb{R}^{3072\times768}$, $W_{1}\in\mathbb{R}^{768\times3072}$ and σ is the sigmoid function for binary classification. We use Transformer-related parameters including an embedding dimension of 768, 12 encoder layers, 12 attention heads, and an MLP hidden size of 3072 following the standard ViT-Base configuration. This ensures consistency with well-established architecture and allows our hybrid model to benefit from proven practices for effective global context modeling. The complete architecture is illustrated in Fig. 3 and the detailed algorithm for the model is given below (see Algorithm 1).

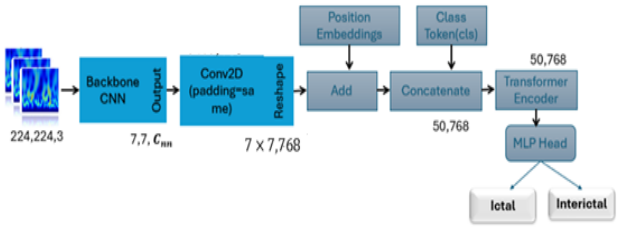

Fig. 3. Architecture of CNN vision transformer.

Algorithm 1 : Hybrid CNN-Vision Transformer

Input:
Time frequency (TF) image $I \in \mathbb{R}^{224\times224\times3}$

Target $y_i \in$ [0 1] 1=ictal; 0=interictal

Output:

Predicted class probability $\hat{y} \in$ [0 1]

Begin:
Data selection

IEEG data

Pre-processing

filtering/re-referencing

Segmentation

2D transformation using CWT

Resize image to 224×224

Labelling $y = \begin{cases} 1: ictal \\ 0: interictal \end{cases}$

Step1: CNN Features Extraction

F_cnn = CNN(I) $F \in \mathbb{R}^{7\times7\times C_{nn}}$

Step2 Token Projection

Conv2D (1 × 1) (padding= same); (7,7,768)

Reshape (7*7,768)

Z_flat = Reshape (Conv2D(F_cnn)); Z_flat $\in \mathbb{R}^{49\times768}$

Z = Concat([CLS], Z_flat) + Positional Embedding

$Z \in \mathbb{R}^{50\times768}$

Step3: Transformer Encoder
for layer = 1 to 12:

-Head Self-Attention

Z_norm = Layer Norm(Z)

$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Z_att = MultiHead (Attention) + Z

-Forward Network

$$FFN(z) = W_2.GELU(W_1.z + b_1) + b_2$$

Step4: MLP Classification Head

$MLP - Head(z_{cls}) = W_2.GELU\,(W_1.LayerNorm(z_{cls}b_1) + b_2$

$\hat{y} = \sigma(W_{class}.MLP - Head(z_{cls}) + b_{class})$

Trainnig

Loss= binary cross entropy

Train model model.fit(X_train, y_train)

Evaluation

model.evaluate(X_test, y_test)

End

4) Proposed CNN MLP-Mixer hybrid model: The MLP-Mixer approach eliminates both convolution and attention operations, in contrast to the transformer architecture, by employing multilayer perception (MLPs) organized into token-mixing and channel-mixing layers. This architecture offers a conceptually simpler alternative to self-attention while maintaining competitive performance through pure algebraic mixing operations [11].

For an input matrix $X \in \mathbb{R}^{S\times D}$, where S is the number of tokens and D is the number of feature dimensions; the fundamental operations of the MLP-Mixer are defined as follows:

Token mixing (modelling spatial relationships):

$$U_{*,i} = X_{*,i} + W_2\sigma(W_1 LayerNorm(X)_{*,i}) \quad (9)$$

for $i = 1,..D$

σ is the GELU activation function.

$X_{*,i}$: all tokens for a channel i (the i-th column of X).

W : Weight of MLP mixing tokens (for $W_1 \in \mathbb{R}^{S'\times S}$, $W_2 \in \mathbb{R}^{S\times S'}$ with S' being the intermediate dimension.

Channel mixing (modelling relationships between characteristics):

$$Y_{j,*} = U_{j,*} + W_4\sigma(W_3 LayerNorm(U)_{j,*}) \quad (10)$$

for $j = 1, \ldots., S$

$U_{*,i}$: all tokens for channel j (the jth column of U).

W : Weights of the MLPs of channel mixing (for $W_3 \in \mathbb{R}^{D'\times D}$, $W_4 \in \mathbb{R}^{D\times D'}$ with D' being the intermediate dimension).LayerNorm refers to layer normalization.

For the final classification a Global average pooling is applied followed by a linear classifier:

$$\hat{y} = \sigma(W_{class}.GlobalAvreagePooling(Y) + b_{class}) \quad (11)$$

Architectural Configuration: (The feature extraction step is described in the previous section), to configure the model architecture, the following approach was adopted a hidden dimension of D= 192, which aligns with Base-type models and strikes a balance between expressiveness and computational limits [11]. Eight layers (L=8) are stacked to reduce overfitting and handle dependencies in sequences of moderate length.

The architecture features an intentional imbalance in its MLP dimensions. For processing spatial information (token-mixing), a bottleneck of 96 units is employed to promote efficiency. For transforming feature information (channel-mixing), a wide network of 768 units is used to facilitate richer representations. This asymmetric design reflects the observation that channel-wise feature interactions typically require higher capacity than spatial token interactions for time-frequency representations. The detailed algorithm for the model is given in Algorithm 2, and the architecture is illustrated in Fig. 4.

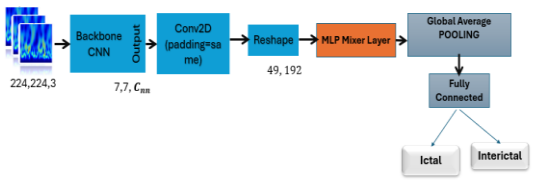

Fig. 4. Architecture of CNN MLP-Mixer.

Algorithm 2: Hybrid CNN-MLP-Mixer

Input:

Time frequency (TF) image $I \in \mathbb{R}^{224\times224\times3}$

Target $y_i \in$ [0 1] 1=ictal; 0=interictal

Output:

Predicted class probability $\hat{y} \in$ [0 1]

Begin:

Data selection

IEEG data

Pre-processing

filtering/re-referencing

Segmentation

2D transformation using CWT

Resize image to 224×224

Labelling $y = \begin{cases} 1: ictal \\ 0: interictal \end{cases}$

Step1: CNN Features Extraction

F_cnn = CNN(I) $F \in \mathbb{R}^{7\times7\times C_{nn}}$

Step2 Features Projection

Conv2D (1 × 1) (padding= same); (7,7,192)

Reshape (7*7,192)

Z = Reshape (Conv2D(F_cnn)); $Z \in \mathbb{R}^{49\times192}$

Step 3: MLP-Mixer Layers (8 blocks)

Token-Mixing MLP (spatial)

U_norm = LayerNorm(Z)

for i = 1 to 192: Per-channel processing

$$U_{*,i} = X_{*,i} + W_{2^\sigma}(W_1 LayerNorm(X)_{*,i})$$

$W_1 \in \mathbb{R}^{96\times49}$ $W_3 \in \mathbb{R}^{49\times96}$

for j= 1 to 49: Per-token processing

$$Y_{j,*} = U_{j,*} + W_{4^\sigma}(W_3 LayerNorm(U)_{j,*})$$

$W_3 \in \mathbb{R}^{768\times192}$ $W_4 \in \mathbb{R}^{192\times768}$

Step 4: Classification

x_pool = GlobalAveragePooling(Y) ; x_pool $\in \mathbb{R}^{192}$

$\hat{y} = \sigma(\text{x\_pool} \times W_{class}. + b_{class})$ $W_{class} \in \mathbb{R}^{(1,192)}$

Trainnig

Loss= binary cross entropy

Train model model.fit(X_train, y_train)

Evaluation

model.evaluate(X_test, y_test)

End

5) Simulation Map: All models were implemented in Python 3.9 using PyTorch 2.0.1. Experiments were conducted on Google Colab using GPU acceleration. The following Table II shows the different hyperparameters used in the study.

TABLE II. CONFIGURATION OF TRAINING HYPERPARAMETERS

Hyperparameters	Value
Optimizer	ADAM
Loss function	Binary cross entropy (BCE)
Batch Size	32
Epochs	80(early stopping with patience=10
Learning Rate	1e-4
Dropout	0.1

IV. RESULTS

The models were evaluated using standard binary classification metrics: accuracy, precision, recall (sensitivity), specificity, F1-score (see Table III) and the area under the receiver operating characteristic curve (AUC-ROC). These metrics were derived from the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the predictions.

TABLE III. EVALUATION METRICS

Metrics	Equations
Accuracy(ACC)	$\frac{TP+TN}{TP+FP+FN+TN}$
Precision (Pr)	$\frac{TP}{TP+FP}$
Recall (Se)	$E = \frac{TP}{TP+FN}$
Specificity (Sp)	$1 - \frac{FP}{FP+TN}$
F1-score	$F1-score = \frac{2\times TP}{(2\times TP+FP+FN)}$

Table IV and Table V present the comprehensive performance evaluation of Vision Transformer (ViT) and MLP-Mixer classifiers across six CNN backbones. The results reveal a consistent performance hierarchy, with the EfficientNetB0-ViT configuration achieving the highest overall scores (Accuracy: 97.85%, Specificity: 98.92%). GoogleNet was also consistent, performing well with both classifiers (96.45% with ViT and 95.20% with MLP-Mixer) and stood out in recall metrics. However, the results obtained with VGG19 are the least optimal.

TABLE IV. PERFORMANCE COMPARISON OF CNN-VISION TRANSFORMER

Model	ACC%	Pr %	Se%	Sp%	F1-score %
EfficientNETB0-ViT	**97.85**	98.32	**96.88**	**98.92**	97.59
ResNet18-ViT	95.35	94.90	92.70	95.95	93.78
ResNet50-ViT	94.80	95.42	93.25	96.35	94.32
ResNet101 ViT	95.25	95.05	95.50	95.00	95.27
GoogleNet VIT	96.45	97.15	95.20	97.85	96.16
VGG19 VIT	93.20	93.65	91.85	94.70	92.74

TABLE V. PERFORMANCE COMPARISON OF CNN MLP-MIXER

Model	ACC%	Pr %	Se%	Sp%	F1-score %
EfficientNETB0MLP-Mixer	97.08	96.55	94.45	97.35	95.48
ResNet18 MLP-Mixer	92.90	91.35	94.25	91.55	92.78
ResNet50- MLP-Mixer	93.25	93.85	91.55	95.05	92.68
ResNet101 MLP-Mixer	93.80	92.45	95.10	92.50	93.76
GoogleNet MLP-Mixer	95.20	94.80	96.75	93.65	95.76
VGG19 MLP-Mixer	91.75	92.20	90.15	93.55	91.16

This comparative analysis provides a foundation for further investigation of operational characteristics, such as ROC curves. As illustrated in Fig. 5, the comparative ROC curves demonstrate the discriminatory power of the various models. The EfficientB0-ViT model demonstrates a remarkable performance with an Area Under the Curve (AUC) of 0.979, the subsequent models are GoogleNet-ViT (AUC = 0.968) and EfficientB0-MLP (AUC = 0.963). The ViT curves have been shown to demonstrate a systematic superiority that confirms and substantiates its efficacy in differentiating between ictal and interictal states.

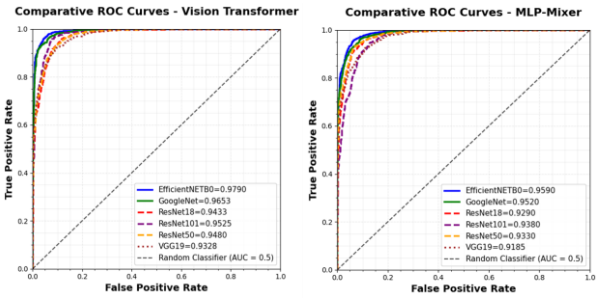

Fig. 5. Comparative ROC Curve: (A) Curve Roc VIT, (B) Curve ROC MLP-mixer.

To understand the impact of the model's performance, we analyzed normalized confusion matrices. These matrices, shown in Fig. 6, reveal each architecture's specific error profile, particularly their trade-offs between false positives and false negatives in distinguishing ictal from interictal states.

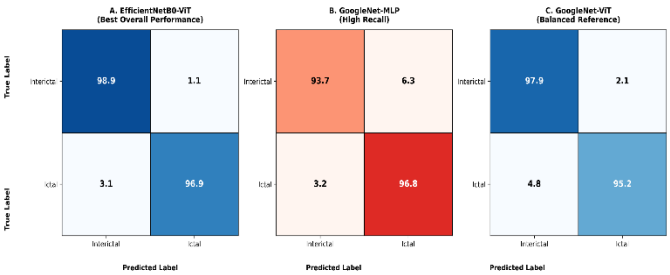

Fig. 6. Normalized confusion matrices for the three best-performing models: (A) EfficientNetB0-ViT, (B) GoogleNet-MLP, (C) GoogleNet-VIT.

A detailed analysis of the error profiles, provided by the confusion matrices in Fig. 6, offers crucial insights for clinical application. The primary strength of EfficientB0-ViT [see Fig. 6(A)] shows only 3.1% false negatives and 1.1% false positives. This reliability makes it ideal for clinical deployment, where the cost of a false alarm (false positive). Interestingly, the GoogleNet-MLP Mixer [see Fig. 6(B)] model has the best recall rate (96.8%). However, it has a higher rate of false positives (6.3%) compared to false negatives (3.2%), which is important for minimizing false negatives (missed seizures). Meanwhile, GoogleNet-ViT [see Fig. 6(C)] offers balanced performance with moderate error rates (false negatives at 4.8% and false positives at 2.1%).

Fig. 7 provides a direct comparative analysis of the performance differential between Vision Transformer and MLP-Mixer architectures. The histogram quantifies the percentage point advantage of ViT over MLP-Mixer for the EfficientNetB0 backbone.

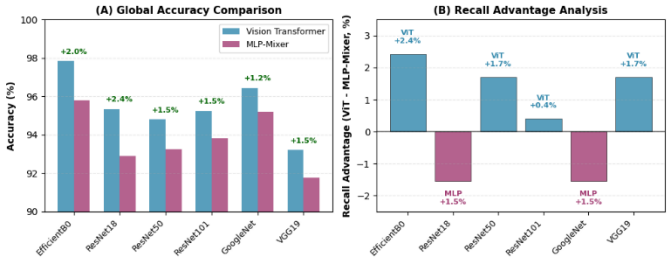

Fig. 7. Performance gain.

Fig. 7 presents the performance differential between Vision Transformer and MLP-Mixer architectures when paired with the EfficientNetB0 backbone. The results demonstrate a consistent advantage for the ViT model across all evaluation metrics, with particularly notable margins in recall and accuracy. However, the MLP-Mixer maintains competitive performance, trailing by less than 2.5 % points in all categories. This narrow performance gap, especially in specificity where both architectures excel, underscores the MLP-Mixer's viability as an efficient alternative to attention-based models for seizure detection tasks.

The findings confirm that while the self-attention mechanism in ViT yields superior overall performance, the MLP-Mixer achieves remarkably close results. This suggests context-dependent applicability: ViT for maximum diagnostic precision versus MLP-Mixer for scenarios prioritizing computational efficiency alongside strong detection capability.

The superior performance of the proposed EfficientNetB0-ViT model is evident when benchmarked against recent literature, as summarized in Table VI. Our method establishes a new accuracy of 97.9%, representing a significant improvement over previous approaches for similar seizure detection challenges.

TABLE VI. COMPARISON OF PREVIOUS WORKS

Authors	Year	Methods	Accuracy%
S.Roy et al.[21]	2019	ChronoNet	90.60
RaviPrakash et al.[22]	2020	1D-CNN-LTSM	89.73
Priya et al.[23]	2022	ResNet50	82.80
Rukshar et al. [24]	2023	Vision Transformer	89.07
Gupta et al.[25]	2024	FUPTBSVM(hybrid model)	88.66
Qi et al.[26]	2025	Vision Transformer	93.65
OURS		**EfficientNet-B0 ViT**	**97.9**

V. DISCUSSION

The experimental results provide a clear answer to the primary research question: Vision Transformer (ViT) classifiers consistently outperform MLP-Mixer counterparts when integrated with CNN backbones for iEEG-based seizure detection. The superior performance of the EfficientNetB0-ViT model (97.85% accuracy, 98.92% specificity) establishes the effectiveness of self-attention mechanisms for capturing the spatiotemporal dynamics of epileptic seizures. This advantage can be theoretically explained by the ViT's ability to model global dependencies across the full time–frequency representation, a capability that aligns well with the distributed neural networks underlying seizure propagation. The observed performance gap between ViT and MLP-Mixer models provides valuable insight into the architectural requirements for effective seizure detection. The consistent advantage of ViT across multiple metrics suggests that the ability to dynamically weight different regions of the time-frequency representation (a core property of self-attention mechanisms) provides significant value for this task. This is particularly relevant for seizures that manifest as evolving patterns across both time and frequency domains. Our results for the MLP-Mixer architecture align with findings from computer vision (Tolstikhin et al., 2021), where MLP-based models have demonstrated competitive performance despite their conceptual simplicity. Extending this observation to biomedical signal analysis, our findings demonstrate that MLP-Mixers constitute a viable alternative to attention-based architectures for time-series classification, offering a balance between interpretability and efficiency. The strong correlation between CNN backbone quality and final performance (evidenced by EfficientNetB0's superior results across both classifiers) underscores the continued importance of effective feature extraction. Indeed, the identification of dominant patterns is critical for characterizing neurological disorders [27]. Even highly expressive classification heads such as ViT cannot compensate for suboptimal convolutional representations, confirming the hierarchical dependency between feature extraction and global context modeling in hybrid deep-learning architectures.

VI. CONCLUSION

This study presented a systematic and novel comparative framework for the automated detection of epileptic seizures from intracranial EEG (iEEG) signals filtered in the High-Frequency Oscillation (HFO) band. By integrating convolutional feature extractors with advanced sequence-modeling architectures Vision Transformer (ViT) and MLP-Mixer we established a comprehensive evaluation of hybrid deep-learning approaches for seizure detection.

The proposed benchmarking protocol represents a novel and systematic comparison between CNN-ViT and CNN-MLP-Mixer models specifically tailored to HFO-based iEEG analysis. Across multiple CNN backbones, the ViT consistently outperformed the Mixer counterpart, with the EfficientNetB0-ViT achieving the highest overall accuracy of 97.85 % and specificity of 98.92 %. This performance highlights the ability of self-attention mechanisms to capture the complex spatiotemporal dynamics of seizure propagation. The MLP-Mixer, on the other hand, demonstrated strong recall (up to 96.75 % with GoogleNet-MLP-Mixer), confirming its suitability for scenarios where sensitivity is clinically prioritized, such as continuous monitoring or early-warning systems.

The comparative trends observed in this work have direct practical value: ViT-based hybrids are better suited for high-specificity diagnostic contexts, such as pre-surgical assessment where false positives must be minimized, whereas Mixer-based hybrids provide a high-recall solution ideal for long-term monitoring. These insights contribute to a deeper understanding of how architectural design choices in hybrid deep learning can be aligned with distinct clinical objectives. Moreover, the consistent superiority of ViT across architectures suggests that self-attention represents a robust modeling paradigm for the non-linear, high-frequency dynamics of epileptic activity extending its relevance beyond computer vision to neurophysiological signal analysis.

While the proposed framework demonstrates strong performance, its generalizability should be validated on larger, multi-center datasets including diverse patient populations. In addition, the focus on high-frequency oscillations, though biologically motivated, may have excluded complementary low-frequency information (e.g., delta or theta bands) relevant to seizure onset and propagation. Addressing these limitations will strengthen the translational impact of the approach.

Future research will extend the current binary framework to multi-class classification, encompassing different HFO subtypes and pre-ictal phases. Another direction will involve integrating explainability techniques such as Grad-CAM or LIME to visualize discriminative regions in the time-frequency maps, thereby enhancing clinical interpretability and trust. Finally, incorporating multimodal spectral inputs and transfer-learning strategies will further improve robustness and pave the way for deployment in real-world clinical environments.

ACKNOWLEDGMENT

The authors are grateful to the investigators who shared their iEEG data.

REFERENCES

[1] N. Jmail et al., "Separation between spikes and oscillation by stationary wavelet transform implemented on an embedded architecture," Journal of the Neurological Sciences, Volume 381, 542.2017. doi: 10.1016/j.jns.2017.08.3735.

[2] A. Hadriche, I. Behy, A. Necibi, A. Kachouri, C. Ben Amar, and N. Jmail, "Assessment of Effective Network Connectivity among MEG None Contaminated Epileptic Transitory Events," Comput. Math. Methods Med., vol. 2021, 2021, doi: 10.1155/2021/6406362.

[3] Hadriche, I. ElBehy, A. Hajjej, and N. Jmail, "Evaluation of Techniques for Predicting a Build Up of a Seizure," Lect. Notes Networks Syst., vol. 418 LNNS, no. 1, pp. 816–827, 2022, doi: 10.1007/978-3-030-96308-8_76.

[4] Abir Hadriche,Nawel Jmail "Clinical Images and Medical Case Reports A build up of seizure prediction and detection Software : A review, J Clin Images Med Case Rep. 2021, 2 (2):1087.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review.," J. Neural Eng., vol. 16, no. 5, p. 51001, Aug. 2019, doi: 10.1088/1741-2552/ab260c.

[6] R. Hussein, H. Palangi, R. Ward, and Z. J. Wang, "Epileptic Seizure Detection: A Deep Learning Approach," pp. 1–12, 2018, [Online]. Available: http://arxiv.org/abs/1803.09848

[7] X. Wu, Z. Yang, T. Zhang, L. Zhang, and L. Qiao, "An end-to-end seizure prediction approach using long short-term memory network," Front. Hum. Neurosci., vol. 17, no. May, pp. 1–10, 2023, doi: 10.3389/fnhum.2023.1187794.

[8] K. Mohiuddin et al., "Retention Is All You Need," Int. Conf. Inf. Knowl. Manag. Proc., no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

[9] S. Yuan, K. Yan, S. Wang, J. X. Liu, and J. Wang, "EEG-Based Seizure Prediction Using Hybrid DenseNet–ViT Network with Attention Fusion," Brain Sci., vol. 14, no. 8, 2024, doi: 10.3390/brainsci14080839.

[10] Q. Li, W. Cao, and A. Zhang, "Multi-stream feature fusion of vision transformer and CNN for precise epileptic seizure detection from EEG signals," J. Transl. Med., vol. 23, no. 1, 2025, doi: 10.1186/s12967-025-06862-z.

[11] Tolstikhin et al., "MLP-Mixer: An all-MLP Architecture for Vision," Adv. Neural Inf. Process. Syst., vol. 29, pp. 24261–24272, 2021.

[12] Burelo, M. Sharifshazileh, G. Indiveri, and J. Sarnthein, "Automatic Detection of High-Frequency Oscillations With Neuromorphic Spiking Neural Networks," Front. Neurosci., vol. 16, no. June, pp. 1–17, 2022, doi: 10.3389/fnins.2022.861480.

[13] T. Guesmi, A. Hadriche, and N. Jmail, "Effective connectivity of high-frequency oscillations (HFOs) using different source localization techniques," ISDA 2022, doi: 10.1007/978-3-031-35507-3_36.

[14] T. Guesmi, A. Hadriche, N. Jmail, and C. Ben Amar, "Evaluation of Stationary Wavelet Transforms in Reconstruction of Pure High Frequency Oscillations (HFOs)," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12157 LNCS, no. 80200, pp. 357–363, 2020, doi: 10.1007/978-3-030-51517-1_32.

[15] Z. Wang et al., "EEG-Based Seizure Detection Using Dual-Branch CNN-ViT Network Integrating Phase and Power Spectrograms," Brain Sci., vol. 15, no. 5, pp. 1–24, 2025, doi: 10.3390/brainsci15050509.

[16] Zayneb Sadek, "An Efficient CNN and RNN Hybrid Model for the Detection of Epileptic Seizures in EEG Signals," in IEEE Symposium on Bioinformatics and Bioengineering (BIBE), 2024.

[17] M. Bashar, O. Monjur, S. Islam, M. G. Shams, and N. Quader, "Exploring Synergistic Ensemble Learning: Uniting CNNs, MLP-Mixers, and Vision Transformers to Enhance Image Classification," 2025, [Online]. Available: http://arxiv.org/abs/2504.09076.

[18] and B. L. John M. Bernabei, Adam Li, Andrew Y. Revell, Rachel J. Smith, Kristin M. Gunnarsdottir, Ian Z. Ong, Kathryn A. Davis, Nishant Sinha, Sridevi Sarma, "HUP iEEG Epilepsy Dataset. OpenNeuro. [Dataset]," 2023, doi: doi: doi:10.18112/openneuro.ds004100.v1.1.3.

[19] A. Dosovitskiy et al., "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," ICLR 2021 - 9th Int. Conf. Learn. Represent., 2021.

[20] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling Neural Machine Translation," WMT 2018 - 3rd Conf. Mach. Transl. Proc. Conf., vol. 1, pp. 1–9, 2018, doi: 10.18653/v1/w18-6301.

[21] S. Roy, I. Kiral-Kornek, and S. Harrer, "Chrononet: A deep recurrent neural network for abnormal EEG identification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11526 LNAI, pp. 47–56, 2019, doi: 10.1007/978-3-030-21642-9_8.

[22] H. RaviPrakash et al., "Deep Learning Provides Exceptional Accuracy to ECoG-Based Functional Language Mapping for Epilepsy Surgery," Front. Neurosci., vol. 14, no. May, pp. 1–14, 2020, doi: 10.3389/fnins.2020.00409.

[23] S. S. Priya and N. J. Nalini, "Eeg Signal Analysis for Epileptic Seizure Using Scaleogram Based Transfer Learning," J. Pharm. Negat. Results, vol. 13, no. 7, pp. 2825–2834, 2022, doi: 10.47750/pnr.2022.13.S07.376.

[24] S. Rukhsar and A. K. Tiwari, "Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals," Comput. Methods Programs Biomed., vol. 242, pp. 1–13, 2023, doi: 10.1016/j.cmpb.2023.107856.

[25] D. Gupta, "Functional iterative approach for universum-based primal twin bounded support vector machine to eeg classification (fuptbsvm)," Multimed. Tools Appl., vol. 83(8), 221, 2024.

[26] Q. Li, T. Zhang, Y. Song, and M. Sun, "Transformer-based spatial-temporal feature learning for P300," 2022 16th ICME Int. Conf. Complex Med. Eng. C. 2022, pp. 310–313, 2022, doi: 10.1109/CME55444.2022.10063297.

[27] Abir Hadriche, Nawel Jmail, Jean-Luc Blanc, Laurent Pezard. "Using centrality measures to extract core pattern of brain dynamics during the resting state" Computer Methods and Programs in Biomedicine, vol. 179. 2019,104985.doi:10.1016/j.cmpb.2019.104985.