DOI: 10.2478/ias-2026-0014 • JIAS • 21(5) • 2026 • 274–293

Article

Attention-Based Fusion of EEG Spectrograms and Scalograms for Schizophrenia Detection Using Vision Transformers

Farah Shan* **and Shalini Z. Ninoria**

Computer Science and Engineering, Teerthankar Mahaveer University, Moradabad, 244102, India; shalinininoria@gmail.com

* Correspondence author: farahshanphd@gmail.com

Abstract: Schizophrenia is a neuropsychiatric disorder that affects brain activity, making timely and reliable diagnosis challenging. Although electroencephalography (EEG) provides a non-invasive and cost-effective diagnostic modality, its non-stationary nature and complex temporal–spectral characteristics limit the effectiveness of conventional machine learning and deep learning approaches, which often rely on a single time–frequency representation or simple feature fusion strategies. To address these limitations, this study proposes an attention-based fusion framework that integrates complementary time–frequency representations for EEG-based schizophrenia detection. Spectrograms and scalograms generated using the Short-Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT) capture global spectral information and localized temporal–frequency characteristics, respectively. These representations are processed by a Vision Transformer (ViT), while a bidirectional cross-attention module enables effective interaction and fusion of complementary features before global feature learning. The proposed framework achieved average classification accuracies of 98.34 ± 0.21% and 98.68 ± 0.15% on Dataset 1 and Dataset 2, respectively. Comparative experiments demonstrate that the proposed framework consistently improves classification performance over methods based on individual time–frequency representations and conventional feature fusion approaches, indicating that the proposed attention-guided fusion strategy effectively exploits complementary EEG information for schizophrenia detection.

Keywords: Schizophrenia Detection; EEG Signals; Spectrogram; Scalogram; Time–Frequency Analysis; Attention-Based Fusion; Cross-Attention; Vision Transformer (ViT)

1. Introduction

Schizophrenia (SZ) is one of the most severe neuropsychiatric disorders, affecting millions of people worldwide. It significantly impairs perception, cognition, and behaviour, leading to substantial functional and social disabilities [1]. Early and accurate diagnosis of SZ remains challenging because of its heterogeneous symptomatology and the subjective nature of clinical assessment. Conventional diagnostic approaches primarily rely on behavioural observations and patient self-reports, which may result in delayed or inconsistent diagnoses. Therefore, objective and evidence-based diagnostic methods are essential for improving the early and reliable detection of schizophrenia.

Electroencephalography (EEG), owing to its high temporal resolution, non-invasive nature, and relatively low cost, has emerged as a promising technique for assessing brain activity. EEG signals reflect the underlying neural dynamics associated with the cognitive and functional abnormalities observed in schizophrenia [2]. However, EEG signals are inherently non-stationary and highly complex. Direct analysis in the time domain often fails to capture their discriminative temporal and spectral characteristics, resulting in reduced classification performance. To address this limitation, EEG signals are commonly transformed into informative visual patterns known as time–frequency representations, which simultaneously preserve temporal and spectral information.

A spectrogram is typically generated using the Short-Time Fourier Transform (STFT) [3], providing a time-varying representation of the signal's frequency content. Although spectrograms effectively capture global spectral characteristics, they are less suitable for identifying transient and multi-scale patterns in EEG signals. In contrast, scalograms**,** generated using the Continuous Wavelet Transform (CWT) [4], provide multi-resolution analysis that effectively captures transient features and localized frequency variations. Consequently, combining these complementary representations can provide richer and more discriminative information for schizophrenia classification.

Recent advances in machine learning (ML) [5,6] and deep learning (DL) techniques [7] have significantly improved EEG-based image analysis. By converting EEG signals into image representations, deep learning models can automatically learn discriminative features for neurological disorder classification. Among these approaches, Convolutional Neural Networks (CNNs) have demonstrated promising performance in EEG image classification tasks [1]. However, CNNs have inherent limitations in modelling long-range dependencies and capturing global contextual information because of their localized receptive fields. To overcome these limitations, Vision Transformers (ViTs) [8] have emerged as a powerful alternative by employing self-attention mechanisms to model global relationships among image patches. ViTs have achieved remarkable success in computer vision applications and have recently demonstrated considerable potential in biomedical signal analysis.

Despite these advances, effectively integrating multiple EEG representations remains a challenging problem. Conventional feature fusion methods, such as feature concatenation and early fusion, often fail to fully exploit the complementary information contained in spectrograms and scalograms. Attention-based fusion mechanisms provide a more effective alternative by adaptively selecting and integrating informative features from different representations. In particular, cross-attention enables one representation to selectively focus on the most relevant information from another representation, thereby facilitating richer feature interactions and improving discriminative capability.

To address these challenges, this study proposes an attention-based fusion framework for EEG-based schizophrenia detection using a Vision Transformer architecture. The proposed framework integrates spectrograms and scalograms through a bidirectional cross-attention mechanism to effectively exploit complementary multi-resolution time–frequency information. The fused feature representations are subsequently processed by a Vision Transformer to learn high-level contextual features for accurate schizophrenia classification.

Some of the important findings of this work are as follows:

1. The complementary integration of two heterogeneous time–frequency representations (STFT and CWT), which capture distinct spectral and temporal characteristics of EEG signals.
2. A bidirectional feature interaction strategy that enables mutual information exchange between the STFT and CWT feature streams before feature fusion, allowing each representation to refine the other and produce more discriminative EEG embeddings.
3. An end-to-end Vision Transformer framework that jointly learns cross-representation dependencies instead of treating the two representations independently or simply concatenating their features.
4. A comprehensive experimental validation on two benchmark schizophrenia EEG datasets, together with extensive ablation studies demonstrating that the proposed bidirectional feature interaction consistently outperforms single-representation models and conventional feature fusion approaches.

The rest of this paper is organized as follows: In section II, the related research works conducted for the problem detection for EEG related to SZ and DL methods are discussed. In Section III, the proposed methodology is described, starting with data pre-processing, followed by features extraction and then attention-based fusion framework. Section IV gives details of the experimental setup and results and Section V provides the discussion. The last part (VI) summarizes the paper by giving directions for further research.

2. Related Works

The existing literature demonstrates significant progress in automated schizophrenia detection from EEG signals through the application of machine learning and deep learning techniques. Dakka et al. [9] introduced recurrent neural networks (RNNs) to learn temporal EEG representations for schizophrenia classification, demonstrating the effectiveness of sequential deep learning models for EEG analysis. Phang et al. [10] proposed a multi-domain connectome convolutional neural network (MDC-CNN) that exploited EEG connectivity patterns to improve classification performance by capturing functional brain connectivity. Similarly, Palaniyappan et al. [11] employed effective connectivity within the triple-network brain system to discriminate schizophrenia spectrum disorders at the individual level, highlighting the importance of connectivity-based biomarkers. Despite their promising performance, these approaches rely heavily on accurate connectivity estimation and computationally intensive preprocessing, which may limit their practical applicability.

To better capture temporal EEG characteristics, Chandran et al. [12] developed an LSTM-based framework for automated schizophrenia detection, while Sun et al. [13] proposed a hybrid deep neural network that combined multiple architectures to learn discriminative EEG features. Similarly, Shoeibi et al. [14] introduced a CNN–LSTM framework that jointly learned spatial and temporal information from EEG recordings. Although these hybrid models improved feature representation, they generally require higher computational complexity, longer training times, and large annotated datasets for effective training.

Recent research has increasingly focused on time–frequency representations of EEG signals. Aslan and Akin [15] transformed EEG signals into scalogram images and employed deep learning for schizophrenia detection, demonstrating the effectiveness of time–frequency analysis. Gosala et al. [16] utilized wavelet transforms for feature extraction, whereas Siuly et al. [17] adopted deep residual networks to automatically learn robust EEG features. The deep residual learning framework introduced by He et al. [18] has significantly influenced modern deep learning architectures by enabling the effective training of very deep networks. Furthermore, Gou et al. [19] reviewed knowledge distillation techniques and highlighted their potential for reducing model complexity while maintaining classification performance. Nevertheless, handcrafted feature engineering often lacks adaptability, whereas deeper neural networks generally require substantial computational resources.

Several studies have also explored alternative feature extraction strategies for schizophrenia diagnosis. Shim et al. [20] combined sensor-level and source-level EEG features to improve classification performance, while Jahmunah et al. [21] investigated nonlinear signal processing techniques to characterize the complex dynamics of EEG signals. Bagherzadeh et al. [22] integrated deep learning with effective connectivity images derived from EEG recordings, and Sharma and Acharya [23] proposed wavelet-based feature extraction from single-channel EEG signals. Although these methods produced encouraging results, they either require computationally expensive preprocessing or do not fully exploit the complementary spatial information available in multichannel EEG recordings.

More recently, Transformer-based architectures have demonstrated promising performance in EEG analysis. Shoeibi et al. [24] introduced a one-dimensional Transformer model for early schizophrenia diagnosis, while Li et al. [25] proposed a spatial–temporal feature mapping framework integrated with LeViT to jointly capture local and global EEG characteristics. These studies demonstrate the capability of attention mechanisms to model long-range dependencies in EEG signals. However, most existing approaches rely on a single time–frequency representation, such as either spectrograms or scalograms, without effectively exploiting the complementary information contained in multiple representations. Furthermore, adaptive attention-based feature fusion has received relatively limited attention in EEG-based schizophrenia detection.

Overall, existing studies demonstrate that machine learning, convolutional neural networks, recurrent neural networks, and Transformer-based architectures have significantly advanced automated schizophrenia detection from EEG signals. Nevertheless, important challenges remain, including dependence on computationally intensive preprocessing, reliance on a single time–frequency representation, and limited exploitation of complementary information through adaptive feature fusion. These limitations motivate the proposed framework, which integrates complementary spectrogram and scalogram representations using an attention-guided feature fusion mechanism to enable more comprehensive feature learning and improve the robustness of schizophrenia classification. A summary of the related studies is presented in Table 1.

Table 1. Comparison of Recent EEG-Based Schizophrenia Detection Methods.

Reference	Method	Representation	Strengths	Limitations
[9]	RNN	Raw EEG	Learns temporal dependencies directly from EEG sequences	Limited spatial feature extraction and weak discrimination capability
[10]	Multi-domain Connectome CNN	EEG connectivity matrices	Captures functional brain connectivity across multiple domains	High computational cost due to connectivity estimation
[11]	Effective Connectivity Analysis	Effective connectivity features	Provides interpretable neurophysiological insights	Complex preprocessing and limited real-time applicability
[12]	LSTM	Raw EEG	Captures long-term temporal dependencies effectively	Ignores spatial relationships among EEG channels
[13]	Hybrid Deep Neural Network	Raw EEG	Learns complementary deep feature representations	High computational complexity and training cost
[14]	CNN–LSTM	Raw EEG	Extracts both spatial and temporal features	Large model size and longer training time
[15]	CNN (Scalogram-based)	Scalogram	Effectively captures time-frequency information	Uses only a single representation
[16]	Wavelet Feature Engineering	Wavelet coefficients	Good for non-stationary EEG signal representation	Depends on handcrafted features

Reference	Method	Representation	Strengths	Limitations
[17]	Deep Residual Network (ResNet)	EEG feature maps	Strong hierarchical feature learning capability	High computational and training cost
[20]	Machine Learning with Sensor + Source Features	Sensor & source EEG	Combines complementary EEG information	Requires source localization and preprocessing complexity
[21]	Nonlinear Signal Processing	Nonlinear EEG features	Captures nonlinear brain dynamics	Feature-engineering dependent approach
[22]	Deep Learning with Connectivity Images	Effective connectivity images	Integrates connectivity with deep learning	Computationally expensive preprocessing pipeline
[23]	Wavelet-Based Single-Channel EEG	Single-channel EEG	Low hardware complexity	Limited spatial information
[24]	1D Transformer	Raw EEG	Captures long-range temporal dependencies	Requires large datasets and high computation
[25]	LeViT-based Spatial–Temporal Mapping	EEG feature maps	Efficient local + global feature extraction	Does not exploit multimodal EEG representations

3. Methodology

Recent advances in deep learning have significantly improved EEG signal analysis by enabling automatic extraction of hierarchical representations from complex neural signals. Hybrid deep learning architectures have demonstrated enhanced classification performance by combining the complementary strengths of multiple network components for robust feature learning. Similarly, explainable deep learning frameworks have shown the potential to improve cross-subject EEG analysis while providing interpretable decision-making capabilities for neurological disorder detection. Despite these advances, many existing approaches primarily operate on one-dimensional EEG signals or a single feature representation, limiting their ability to exploit complementary temporal and spectral information embedded in EEG data. Furthermore, the effective integration of multiple time–frequency representations remains an open challenge for improving the discrimination of subtle neural patterns associated with schizophrenia. Recent studies, including the layered cascade deep learning model proposed by Liu et al. [26] and the explainable cross-subject EEG framework (SEEG-Net) developed by Wang et al. [27], demonstrate the effectiveness of advanced deep learning architectures for EEG-based neurological disorder analysis and further motivate the development of more comprehensive multi-representation learning strategies.

The proposed framework introduces a unified multi-representation learning strategy for EEG-based schizophrenia detection by integrating complementary time–frequency representations within an attention-guided Vision Transformer architecture (Figure 1). Unlike conventional approaches that rely on a single image representation, the proposed framework jointly exploits the high temporal localization capability of spectrograms and the multi-resolution characteristics of scalograms to capture complementary neural patterns associated with schizophrenia. To effectively integrate these heterogeneous representations, an attention-guided cross-modal feature interaction mechanism is employed to enable adaptive information exchange before transformer-based global feature learning and final classification. This design facilitates more comprehensive feature representation, allowing the model to capture both local and global EEG characteristics for improved schizophrenia detection. The overall processing pipeline consists of the following stages:

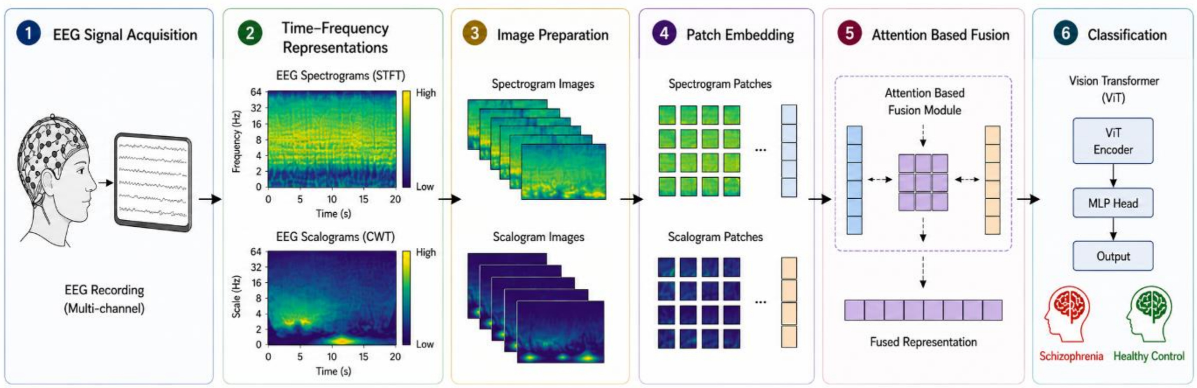


Figure 1. Schematic diagram of the proposed attention-based fusion framework for schizophrenia detection using EEG spectrograms and scalograms with a Vision Transformer architecture.

3.1 EEG Signal Preprocessing

The EEG signals are first pre-processed to improve signal quality by removing noise and physiological artifacts, including baseline drift, muscle activity, and power-line interference. This is accomplished by applying a band-pass filter that preserves the clinically relevant frequency components of the EEG signals, typically within the range of 0.5–50 Hz. The resulting filtered signals provide a cleaner and more reliable representation for subsequent time–frequency transformation and feature extraction. The filtered signal can be expressed as (Equation (1)):

$$x_f(t) = \mathcal{F}_{bp}(x(t)) \tag{1}$$

where $\mathcal{F}_{bp}$ denotes the band-pass filtering operation and $x(t), t = 1,2, \ldots, T$ is raw EEG signal. This step ensures that irrelevant low-frequency drifts and high-frequency noise are effectively suppressed.

Following filtering, the continuous EEG signal is segmented into smaller, fixed-length epochs to facilitate efficient analysis and model training. Each segmented epoch is represented as (Equation (2)):

$$x_e^{(i)}(t), i = 1,2, \ldots, N \tag{2}$$

where N represents the total number of epochs. Such segmentation helps the model learn localized temporal patterns while boosting the number of training examples.

In order to make the data more consistent, each epoch is normalized so as to limit amplitude differences among various recordings and individuals. This is achieved through z-score normalization as (Equation (3)):

$$\tilde{x}_e^{(i)} = \frac{x_e^{(i)} - \mu}{\sigma} \tag{3}$$

where μ and σrepresent the mean and standard deviation of the epoch, respectively. This normalization step stabilizes the training process, improves convergence, and enhances the generalization capability of the proposed model.

3.2 Time–Frequency Representations

The spectrogram is a widely used time–frequency representation that captures how the spectral content of a signal evolves over time. It is computed using the Short-Time Fourier Transform (STFT), which applies the classical Fourier Transform over short, overlapping time windows to handle the non-stationary nature of EEG signals. The STFT is defined as (Equation (4)):

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t)\, e^{-j2\pi f\tau} d\tau \tag{4}$$

Here, $x(\tau)$represents the input EEG signal, $w(\tau - t)$is a window function (such as Hamming or Gaussian) centered at time t, and $f$denotes frequency. The windowing operation localizes the signal in time, allowing the Fourier Transform to analyze only a small portion of the signal at a time. By sliding the window across the signal, a time-dependent frequency representation is obtained.

The spectrogram is then computed as the squared magnitude of the STFT as (Equation (5)):

$$(t, f) = | X(t, f) |^2 \tag{5}$$

The spectrogram represents the distribution of signal power across the time and frequency domains. It is generated using the Short-Time Fourier Transform (STFT), which assumes that the signal is locally stationary over short time intervals. Under this assumption, STFT effectively analyzes slowly varying signals whose frequency content changes gradually over time.

A key limitation of STFT is its fixed time–frequency resolution, which is determined by the size of the analysis window. A shorter analysis window provides higher temporal resolution but lower frequency resolution, whereas a longer window improves frequency resolution at the expense of temporal resolution. This trade-off arises from the Heisenberg uncertainty principle**,** which states that high resolution cannot be achieved simultaneously in both the time and frequency domains.

Despite this limitation, spectrograms provide an effective visualization of the global spectral characteristics and oscillatory patterns present in EEG signals. For schizophrenia detection, spectrograms facilitate the identification of abnormalities in neural oscillatory activity across different frequency bands, thereby providing informative features for subsequent deep learning-based classification.

3.3 Scalogram (CWT)

The scalogram is a time–frequency representation derived from the Continuous Wavelet Transform (CWT), which is particularly well-suited for analysing non-stationary signals such as EEG. Unlike the STFT, which uses a fixed window, the CWT employs scalable and translatable wavelets to adaptively analyse the signal at different resolutions. The CWT is defined as (Equation (6)):

$$(a,b) = \frac{1}{\sqrt{\mid a \mid}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a}\right) dt \tag{6}$$

where $x(t)$is the EEG signal, $\psi$is the mother wavelet, $a$represents the scale parameter (inversely related to frequency), and $b$denotes the time shift. The term ψ^*indicates the complex conjugate of the wavelet function. By varying the scale a, the wavelet effectively stretches or compresses, allowing the analysis of both low-frequency (large-scale) and high-frequency (small-scale) components of the signal.

The scalogram is obtained as the squared magnitude of the wavelet coefficients as (Equation (7)):

$$C(a,b) =\mid W(a,b) \mid^2 \tag{7}$$

The scalogram provides a visual representation of the distribution of signal energy across multiple time and frequency scales. Unlike the fixed time–frequency resolution of the Short-Time Fourier Transform (STFT), the Continuous Wavelet Transform (CWT) performs multi-resolution analysis, providing high temporal resolution and low frequency resolution at high frequencies, while offering high frequency resolution and low temporal resolution at low frequencies. This adaptive resolution enables CWT to effectively characterize the transient and non-stationary nature of EEG signals.

For schizophrenia detection, scalograms are particularly valuable because they can reveal transient neural activities and subtle irregularities in brain dynamics that may not be adequately captured by fixed-resolution methods. Furthermore, the selection of an appropriate mother wavelet, such as the Morlet or Mexican Hat wavelet, enhances the ability of the CWT to capture discriminative EEG signal characteristics, resulting in more informative time–frequency representations for subsequent feature learning and classification.

3.4 Image Representation and Patch Embedding

Both spectrogram and scalogram representations are first converted into image-like formats to make them compatible with deep learning architectures. Specifically, each representation is transformed into a three-channel RGB image, denoted as (Equation (8)):

$$I_s \in \mathbb{R}^{H \times W \times 3}, I_c \in \mathbb{R}^{H \times W \times 3} \tag{8}$$

where $H$and Wrepresent the height and width of the image, respectively. This conversion allows the model to leverage spatial feature extraction techniques similar to those used in computer vision tasks.

Each image is then partitioned into a set of non-overlapping patches of size $P \times P$. The total number of patches is given by (Equation (9)):

$$N = \frac{HW}{P^2} \tag{9}$$

This patching strategy enables the transformation of the 2D image into a sequence of smaller regions, which can be processed as tokens by the Vision Transformer.

Subsequently, each patch is flattened into a one-dimensional vector and projected into a latent embedding space using a learnable linear transformation. The embedding process is defined as (Equation (10)):

$$z_i = E \cdot \text{vec}(I_i) + p_i \tag{10}$$

where E is the embedding matrix that transforms the flattened patch into a d-dimensional feature vector, while $\text{vec}(I_i)$ is the vector representation of the i-th patch. The p_i in this equation refers to the positional encoding, which is incorporated to maintain the spatial position information of the patches in the original image.

In essence, the input images are transformed into sequences of embedded tokens, which include information on features and positions, making it possible to further process them using the transformer model.

3.5 Attention-Based Cross-Representation Fusion Mechanism

Schizophrenia-related EEG abnormalities are distributed across multiple temporal and spectral scales, making a single time–frequency representation insufficient to capture the underlying neural dynamics comprehensively. Spectrograms effectively represent localized spectral energy variations over time, whereas scalograms provide multi-resolution analysis that preserves both low-frequency oscillatory patterns and transient high-frequency components. To effectively exploit the complementary information contained in these representations, the proposed framework incorporates an attention-guided cross-representation fusion mechanism (Figure 2). This mechanism enables adaptive feature interaction, allowing discriminative information learned from one representation to refine the feature representation of the other before transformer-based global feature learning. Unlike conventional feature concatenation, which treats all extracted features with equal importance, the proposed attention-guided fusion adaptively emphasizes the most informative complementary features while suppressing redundant information. As a result, the fused representation captures richer temporal–spectral characteristics of EEG signals, leading to more discriminative feature learning and improved schizophrenia classification performance.

Let the embedded patch tokens extracted from the spectrogram and scalogram branches be represented as (Equation (11)):

$$X_s \in \mathbb{R}^{N\times d}, X_c \in \mathbb{R}^{N\times d} \tag{11}$$

where Ndenotes the number of image patches and $d$represents the embedding dimension.

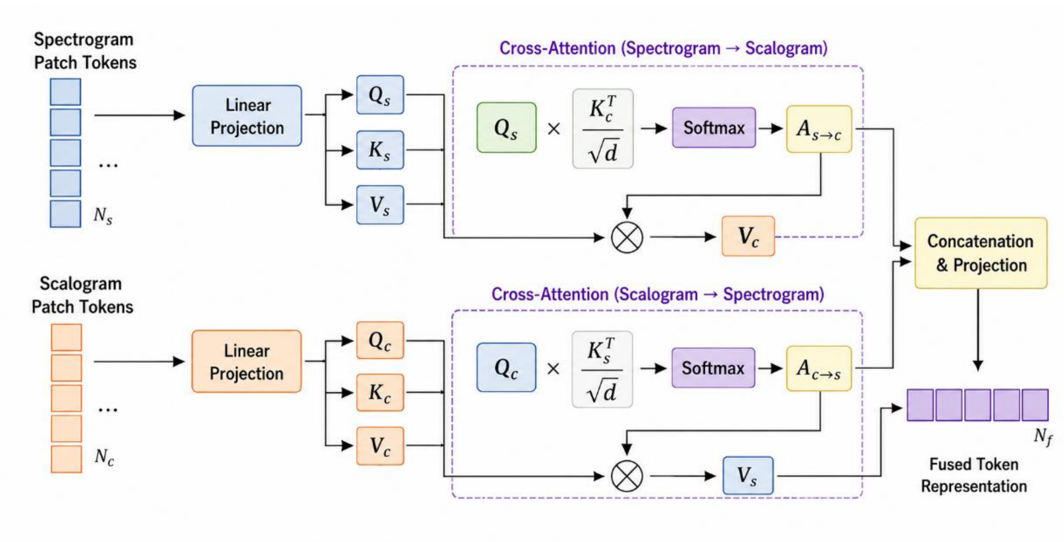


Figure 2. Illustration of the proposed attention-based cross-representation fusion framework.

3.5.1 Spectrogram-Guided Feature Refinement Using Scalogram Information

In the first cross-attention stage, the spectrogram embeddings selectively incorporate complementary information from the scalogram representation. Specifically, the spectrogram tokens are used as queries, while the scalogram tokens serve as keys and values. This mechanism enables each spectrogram patch to attend to the most relevant multi-resolution features in the scalogram representation, thereby enriching its feature representation with complementary temporal–frequency information before subsequent transformer-based feature learning. The scaled dot-product attention is defined as (Equation (12)):

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{12}$$

where the query, key, and value matrices are computed as (Equation (13)):

$$Q_s = X_s W_Q, K_c = X_c W_K, V_c = X_c W_V \tag{13}$$

with W_Q, W_K, and $W_V$denoting learnable projection matrices.

The attention coefficients are calculated as (Equation (14)):

$$A_{s\to c} = \text{Softmax}\left(\frac{Q_s K_c^T}{\sqrt{d}}\right) \tag{14}$$

and the refined spectrogram features are obtained by (Equation (15)):

$$Z_{s\to c} = A_{s\to c}V_c \quad (15)$$

This operation enables the spectrogram branch to emphasize frequency regions that are reinforced by the complementary multi-resolution information contained in the scalogram representation. Consequently, both transient spectral variations and long-duration oscillatory patterns are effectively preserved within the learned feature space, resulting in a more informative representation for subsequent classification.

3.5.2 Scalogram-Guided Feature Refinement Using Spectrogram Information

To achieve reciprocal feature interaction, the second attention stage performs the reverse operation, allowing scalogram embeddings to selectively utilize contextual information extracted from the spectrogram representation. In this direction, (Equation (16)):

$$Q_c = X_c W_Q, K_s = X_s W_K, V_s = X_s W_V \quad (16)$$

The corresponding refined scalogram representation is computed as (Equation (17)):

$$Z_{c\to s} = \mathrm{Softmax}\left(\frac{Q_c K_s^T}{\sqrt{d}}\right) V_s \quad (17)$$

This reciprocal interaction enables the scalogram branch to enhance its multi-resolution feature representation by incorporating complementary localized spectral information from the spectrogram representation. The resulting bidirectional information exchange promotes consistency between the two representations while reducing representation-specific redundancy. Consequently, the fused features provide a more comprehensive characterization of the underlying EEG patterns for subsequent transformer-based learning and schizophrenia classification. Finally, the outputs obtained from both attention directions are concatenated and projected into a unified latent representation as (Equation (18)):

$$Z_f = \mathrm{Concat}(Z_{s\to c}, Z_{c\to s}) W_f \quad (18)$$

where W_f denotes a learnable projection matrix responsible for integrating the bidirectional features into a compact embedding.

Rather than processing the two time–frequency representations independently, the proposed cross-representation interaction enables each branch to adaptively refine its features by incorporating complementary information from the other branch. This bidirectional feature exchange is particularly well suited to EEG analysis because schizophrenia-related abnormalities are distributed across multiple frequency bands and temporal resolutions rather than being confined to a single representation. By jointly exploiting the localized spectral characteristics captured by spectrograms and the multi-resolution time–frequency information preserved by scalograms, the proposed framework generates a richer and more discriminative feature representation. The fused features are subsequently provided to the Vision Transformer for global contextual modelling and schizophrenia classification.

3.6 Vision Transformer (ViT)

The fused feature tokens, denoted as (Z_f), obtained through the attention-based fusion mechanism, are subsequently passed to a ViT encoder [6] for global feature learning (Figure 3). Unlike convolutional neural networks, the ViT processes the input as a sequence of embedded tokens and employs self-attention mechanisms to model long-range dependencies and global contextual relationships among all tokens. Each Transformer encoder layer consists of two primary components: a Multi-Head Self-Attention (MHSA) module, which captures interactions among token embeddings, and a Feed-Forward Network (FFN), which further refines the learned feature representations before passing them to the next encoder layer.

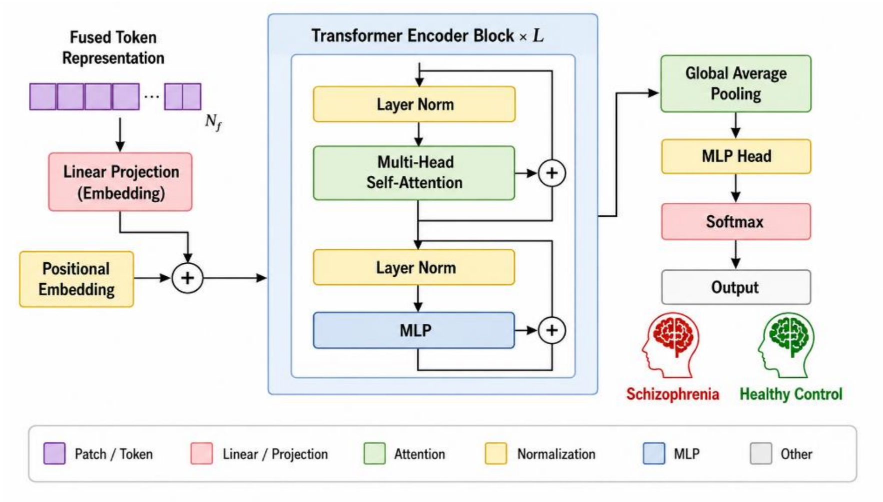


Figure 3. Schematic diagram of the ViT.

3.6.1 Multi-Head Self-Attention (MHSA)

The MHSA mechanism enables the model to attend to information from different representation subspaces simultaneously. It is defined as (Equation (19)):

$$\mathrm{MHSA}(Z) = \mathrm{Concat}(head_1, \dots, head_h)W_O \tag{19}$$

where each attention head is computed as (Equation (20)):

$$head_i = \mathrm{Attention}(Q_i, K_i, V_i) \tag{20}$$

Here, the query, key, and value matrices are obtained through linear projections of the input tokens. Multiple attention heads allow the model to learn diverse feature interactions, capturing both local and global dependencies within the EEG-derived representations.

3.6.2 Feed-Forward Network (FFN)

Following the attention mechanism, each token is passed through a position-wise feed-forward network (Equation (21)):

$$\mathrm{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{21}$$

where σdenotes a non-linear activation function (such as GELU or ReLU). This component enhances the representational capacity of the model by introducing non-linearity and enabling complex feature transformations.

3.6.3 Transformer Block

Each transformer encoder block combines MHSA and FFN with residual connections and normalization (Equation (22)), (Equation (23)):

$$Z' = \mathrm{LayerNorm}(Z + \mathrm{MHSA}(Z)) \tag{22}$$

$$Z'' = \mathrm{LayerNorm}(Z' + \mathrm{FFN}(Z')) \tag{23}$$

These residual connections help preserve gradient flow during training, while layer normalization stabilizes learning and accelerates convergence.

3.6.4 Classification Head

After passing through multiple transformer layers, a global feature representation is extracted using a pooling operation (Equation (24)):

$$z_{cls} = \mathrm{GAP}(Z'') \tag{24}$$

where GAP (Global Average Pooling) aggregates information across all tokens. The final classification is obtained using a fully connected layer followed by a softmax function as (Equation (25)):

$$\hat{y} = \text{softmax}(Wz_{cls} + b) \tag{25}$$

This produces class probabilities for schizophrenia detection. To train the model, categorical cross-entropy loss is employed as (Equation (26)):

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{26}$$

where C is the number of classes, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability. The cross-entropy loss aims at giving a larger probability score to the actual class while reducing the misclassification error.

The Vision Transformer provides global modelling capabilities through the modelling of features between all patches concurrently, making it very useful in EEG analysis, where discriminative features can be dispersed into various time-frequency bins. Through attention-based feature fusion, the ViT creates an effective feature representation of a subject.

4. Simulation and Results

The first data set (Dataset:1) used in this study is taken from the Repository for Open Data (RepOD, 2017) [28] which was first published by Olejarczyk and Jernajczyk [29]. The data set comprises of EEG from clinically diagnosed persons suffering from paranoid SZ (F20.0 ICD-10) alongside healthy persons matched in both age and gender. The EEG data for each person was collected over a period of 15 minutes using the internationally standardized 10/20 electrode configuration. A total of 19 channels were employed (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, and Pz), ensuring coverage of major brain regions. The sampling of the recording was done at a rate of 250Hz and filtered using a band-pass filter of 2 to 45 Hz, where only the most significant frequency information of the EEG signal can be retained. For preparation of the inputs in the model, the signals were split into segments of 2 seconds. The total number of persons used in the dataset is 28, including 14 SZ patients and 14 healthy persons. Each subject has 50 epochs in total, making total of 1400 EEG samples.

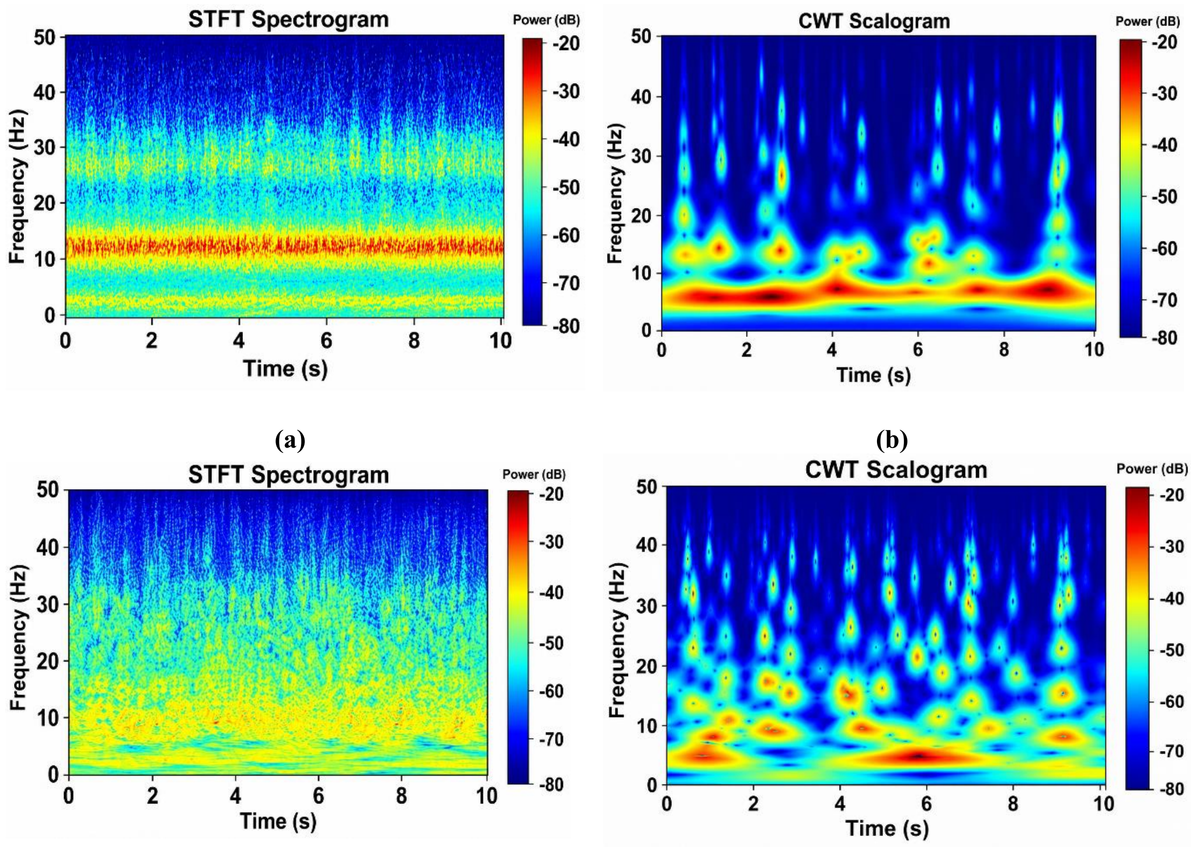
The second EEG dataset (Dataset:2) used in this study was acquired from Huilongguan Hospital, Beijing, China [27], and consists of recordings from 109 subjects**,** including 55 healthy controls (HC) and 54 patients diagnosed with schizophrenia (SZ)**.** Resting-state EEG signals were recorded under an eyes-closed condition using a 64-channel NeuroScan EEG acquisition system**,** with 60 EEG channels utilized for subsequent analysis. The signals were sampled at 500 Hz following the international 10–20 electrode placement system**,** providing high-resolution temporal information for neurological assessment. All schizophrenia participants were clinically diagnosed according to established diagnostic criteria and were under stable medication, while healthy controls had no history of neurological or psychiatric disorders.

To develop a balanced dataset suitable for deep learning, the continuous EEG recordings from each subject were segmented into multiple fixed-length epochs after preprocessing. This segmentation process generated a total of 1,600 EEG samples**,** comprising 800 healthy control samples and 800 schizophrenia samples**.** Each segmented EEG sample was subsequently transformed into the required time-frequency representations (spectrograms and scalograms) before being used for model training and evaluation. The equal number of samples from both classes eliminates class imbalance, ensuring unbiased learning and improving the robustness and generalization capability of the proposed classification framework. The list of simulation parameters is detailed in Table 2.

The comparative time–frequency analysis of the signals from the normal class and the SZ class is shown in Figure 4 using two commonly used signal processing methods, namely spectrograms and scalograms. The spectrogram and scalogram representations of the normal class are shown in subfigures (a) and (b), respectively, depicting the distribution of the energy of the signal over time and frequency domains with relatively stable and structured patterns. On the other side, the representations for the SZ class show identifiable irregularities, changes in frequency components and in the temporal dynamics (subfigures (c) and (d)). The spectrogram from STFT gives a linear frequency resolution and the scalogram obtained from CWT gives a multi-resolution analysis capturing both high- and low-frequency features in a proper manner. This comparative study highlights the difference between the two classes in terms of their time-frequency behaviour, a feature that can play a significant role in feature extraction, signal classification, and other applications in neurological signal processing.

Table 2. List of Simulation Parameters.

Category	Parameter	Value / Description
EEG Acquisition	Sampling Frequency	256 Hz
	Number of Channels	20
	Epoch Length	2–5 seconds
	Frequency Range	0.5 – 50 Hz
	Dataset Split	80% Train / 20% Validation
Preprocessing	Band-pass Filter	0.5 – 50 Hz
	Normalization	Z-score
	Artifact Removal	Optional (ICA / Filtering)
Spectrogram (STFT)	Window Type	Hamming
	Window Length	256 samples
	Overlap	50%
	FFT Points	256
	Output Size	224 × 224
Scalogram (CWT)	Mother Wavelet	Morlet
	Scales	1 – 128
	Frequency Resolution	Adaptive
	Output Size	224 × 224
Image Processing	Image Type	RGB
	Image Size	224 × 224 × 3
	Data Augmentation	Rotation, Flipping, Scaling
Patch Embedding	Patch Size	16 × 16
	Number of Patches (N)	196
	Embedding Dimension (d)	768
Attention Fusion	Fusion Type	Cross-Attention
	Attention Heads	8
	Fusion Strategy	Bidirectional (S↔C)
	Projection Matrix	Learnable
Vision Transformer	Model Type	ViT-Base
	Number of Layers	12
	Hidden Dimension	768
	MLP Size	3072
	Dropout	0.1
Training	Optimizer	Adam
	Learning Rate	1e-4
	Batch Size	16 / 32
	Epochs	50 – 100
	Weight Decay	1e-5
Classification	Activation	Softmax
	Loss Function	Categorical Cross-Entropy



(a) (b) (c) (d)

Figure 4. Time–Frequency Representations of EEG Signals for Normal and Schizophrenia Classes: (a) Spectrogram of the Normal class (b) Scalogram of the Normal class (c) Spectrogram of the Schizophrenia class (d) Scalogram of the Schizophrenia class.

4.1 Random Splitting Results

The epoch-level random split (80/10/10) is included solely to enable direct comparison with previous studies employing the same evaluation protocol. Since EEG epochs from the same subject may appear in both the training and test sets, this protocol can introduce data leakage and overestimate classification performance. Accordingly, the subject-level cross-validation results reported in Section 4.2 are regarded as the primary and more reliable measure of model generalization.

Figure 5 presents the variation of training and validation loss over successive epochs, providing insight into the learning behaviour and generalization capability of the proposed model. The training loss decreases smoothly throughout the training process, indicating stable optimization and effective learning. Although the validation loss exhibits minor fluctuations, its overall downward trend and convergence demonstrate consistent generalization performance. The absence of a significant gap between the training and validation loss curves suggests that the model does not suffer from severe overfitting and is able to learn robust feature representations.

Figure 6 illustrates the training and validation accuracy curves over successive epochs, providing insight into the learning progression and generalization capability of the proposed model. Both curves exhibit a consistent upward trend throughout the training process, indicating continuous improvement in classification performance. The training accuracy increases smoothly and gradually converges to approximately 99.2%, while the validation accuracy follows a similar trend with only minor fluctuations, reaching nearly 98.6% by the final epochs. The small gap between the training and validation curves suggests good generalization and demonstrates that the model does not suffer from significant overfitting. Furthermore, the stable convergence of both curves indicates effective optimization and robust learning of discriminative EEG features.

Figure 7 presents the attention maps generated by the proposed model for (a) a healthy (Normal) subject and (b) a schizophrenia (SZ) subject. The visualizations illustrate the regions within the time–frequency representations that contribute most significantly to the model's classification decision. For the normal sample, the attention is relatively well-distributed and concentrated over consistent signal patterns, reflecting stable neural activity. In contrast, the schizophrenia sample exhibits attention focused on distinct discriminative regions, indicating the presence of abnormal time–frequency characteristics associated with the disorder. These attention maps demonstrate that the proposed attention-based fusion mechanism effectively identifies informative EEG regions while suppressing less relevant information, thereby enhancing both the classification performance and the interpretability of the proposed framework.

The confusion matrices for the training set (a) and the validation set (b) are shown in Figure 8. This figure gives proper detail about the accuracy of the classification model. During the training phase, the model does not have a significant number of False Positives (FP) or False Negatives (FN); indeed, it records only 4 FP cases and just 5 FN cases, while having 555 true positive (TP) cases and 556 true negative (TN) cases, so there is a high level of accuracy and balance between both classes. During the validation, the model shows good generalization capability and results in 137 numbers of TP and 139 numbers of TN, only 3 numbers of FN and 1 number of FP. The small number of false predictions in both phases shows the robustness and reliability of the model, as well as indicating its effectiveness in correctly differentiating between the two classes being distinguished and its ability to be consistent across the training data and unseen data.

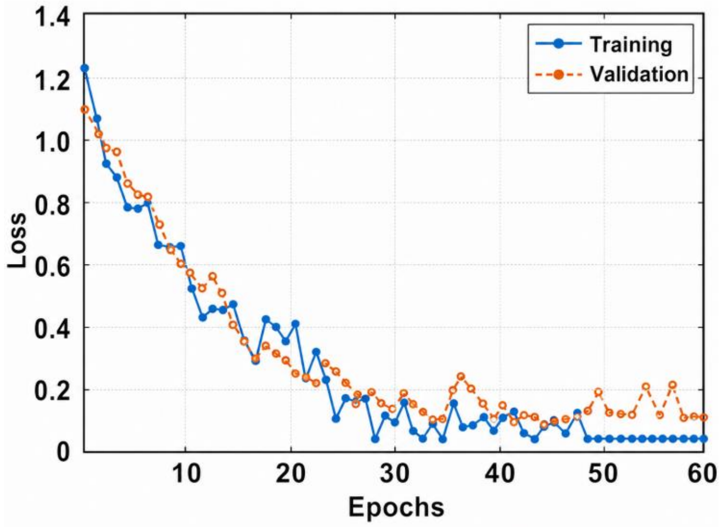


Figure 5. Training and Validation Loss Curves of the Proposed Model During Learning.

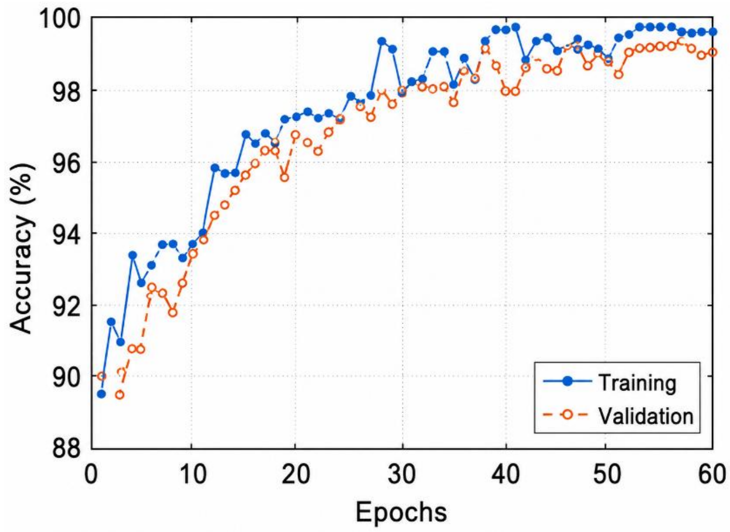
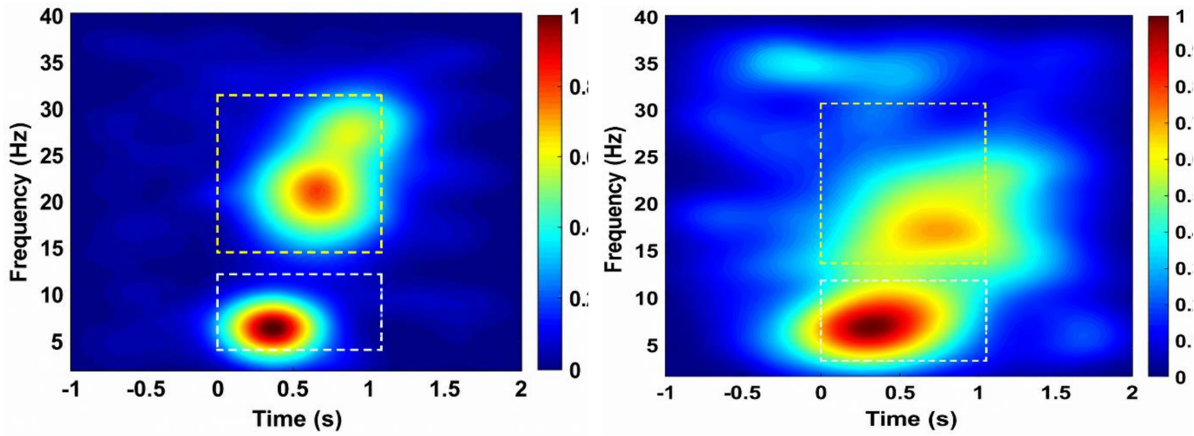
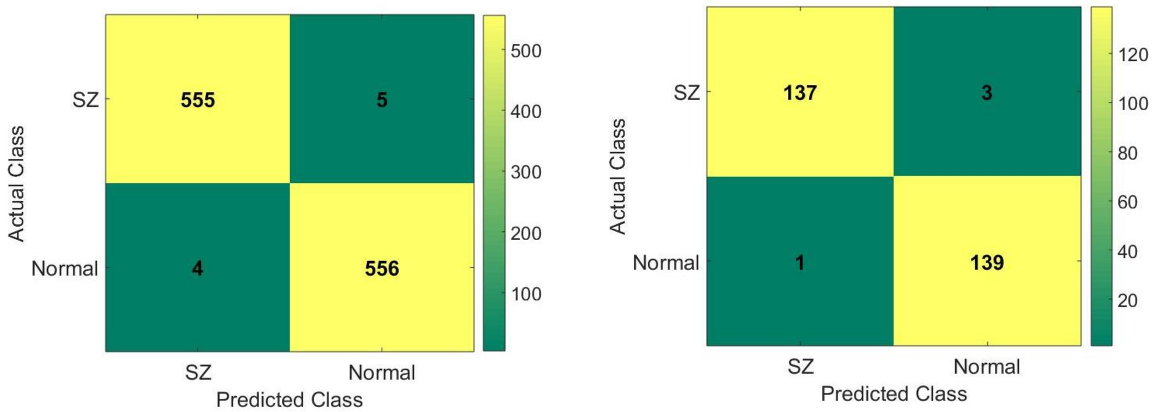


Figure 6. Training and Validation Accuracy Curves of the Proposed Model During Learning.



(a) (b)

Figure 7. Attention Map Visualization of the Proposed Model for EEG Signal Classification (a) Normal (b) SZ.



(a)Training **(b)Validation**

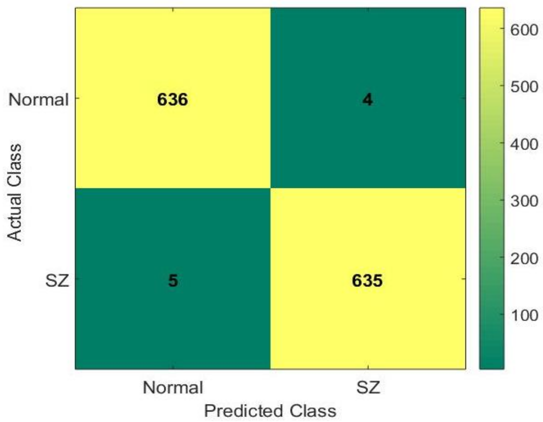
Figure 8. Confusion Matrices of the Proposed Model for Schizophrenia Classification on Dataset 1(a) Training Confusion Matrix (b) Validation Confusion Matrix.

Table 3 presents the performance of the proposed model on Dataset 1 using a random train–validation data split. The model achieved excellent performance on the training set, with a precision of 0.9928 ± 0.0012, recall of 0.9911 ± 0.0015, F1-score of 0.9920 ± 0.0013, and an accuracy of 0.9920 ± 0.0012, indicating effective learning of the training data. On the validation set, the proposed framework maintained strong generalization capability, achieving a precision of 0.9928 ± 0.0024, recall of 0.9786 ± 0.0041, F1-score of 0.9857 ± 0.0033, and an overall accuracy of 0.9857 ± 0.0032. The small standard deviations across all evaluation metrics demonstrate the stability and consistency of the proposed model under repeated experiments. Moreover, the close agreement between the training and validation performance indicates that the model generalizes well while maintaining high classification accuracy.

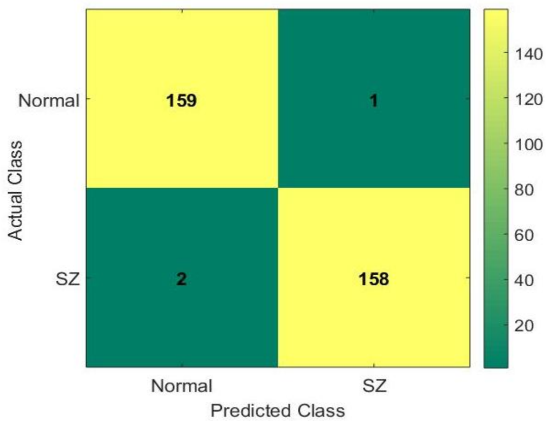
Table 3. Performance Metrics Comparison Using Random Data Splitting (Dataset: 1).

Dataset	Precision	Recall	F1-score	Accuracy
Training	0.9928 ± 0.0012	0.9911 ± 0.0015	0.9920 ± 0.0013	0.9920 ± 0.0012
Validation	0.9928 ± 0.0024	0.9786 ± 0.0041	0.9857 ± 0.0033	0.9857 ± 0.0032

The confusion matrix results demonstrate the strong classification performance of the proposed model on Dataset 2 under both training and validation settings (Figure 9). In the training phase, the model correctly classifies 636 samples as true negatives and 635 samples as true positives, with only 4 false positives and 5 false negatives, indicating very high learning accuracy and minimal misclassification. Similarly, in the validation phase, the model maintains stable performance, correctly classifying the majority of samples with only a very small number of misclassifications (1–2 errors per class). This consistency between training and validation confusion matrices confirms that the proposed model generalizes well and does not suffer from overfitting. Overall, the results clearly indicate strong discriminative capability and robust performance of the proposed framework for schizophrenia detection.



(a)Training



(b)Validation

Figure 9. Confusion Matrices of the Proposed Model for Schizophrenia Classification on Dataset 2 (a) Training Confusion Matrix (b) Validation Confusion Matrix.

Table 4 summarizes the performance of the proposed framework on Dataset 2 using a random train–validation data split. The model achieved outstanding performance on the training set, recording a precision of 0.9937 ± 0.0011, recall of 0.9922 ± 0.0014, F1-score of 0.9930 ± 0.0012, and an accuracy of 0.9930 ± 0.0011. On the validation set, the proposed model continued to demonstrate excellent generalization capability, achieving a precision of 0.9937 ± 0.0023, recall of 0.9875 ± 0.0038, F1-score of 0.9906 ± 0.0030, and an overall accuracy of 0.9906 ± 0.0028. The low standard deviation values across all performance metrics indicate that the model produces stable and consistent results over multiple runs. Furthermore, the close correspondence between the training and validation metrics suggests that the proposed framework generalizes effectively while maintaining high classification performance on unseen data.

Table 4. Performance Metrics Comparison Using Random Data Splitting (Dataset: 2).

Dataset	Precision	Recall	F1-score	Accuracy
Training	0.9937 ± 0.0011	0.9922 ± 0.0014	0.9930 ± 0.0012	0.9930 ± 0.0011
Validation	0.9937 ± 0.0023	0.9875 ± 0.0038	0.9906 ± 0.0030	0.9906 ± 0.0028

Figure 10 presents the Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) values for the proposed model on both datasets. Figure 10(a) shows the ROC curves for Dataset 1, where the model achieved an AUC of 0.988 on the training set and 0.977 on the validation set. These high AUC values indicate that the proposed framework can effectively discriminate between schizophrenia and healthy control classes across different classification thresholds. The small difference between the training and validation AUC values further suggests that the model generalizes well to unseen data without exhibiting significant overfitting.

Similarly, Figure 10(b) illustrates the ROC curves for Dataset 2. The proposed model achieved an AUC of 0.993 on the training set and 0.985 on the validation set, indicating a high level of classification accuracy and discriminative ability. The consistently high AUC values on both datasets confirm the robustness and reliability of the proposed framework in distinguishing between the target classes. Furthermore, the comparable training and validation performances indicate stable learning and strong generalization across different EEG datasets, highlighting the effectiveness of the proposed approach for automated schizophrenia detection.

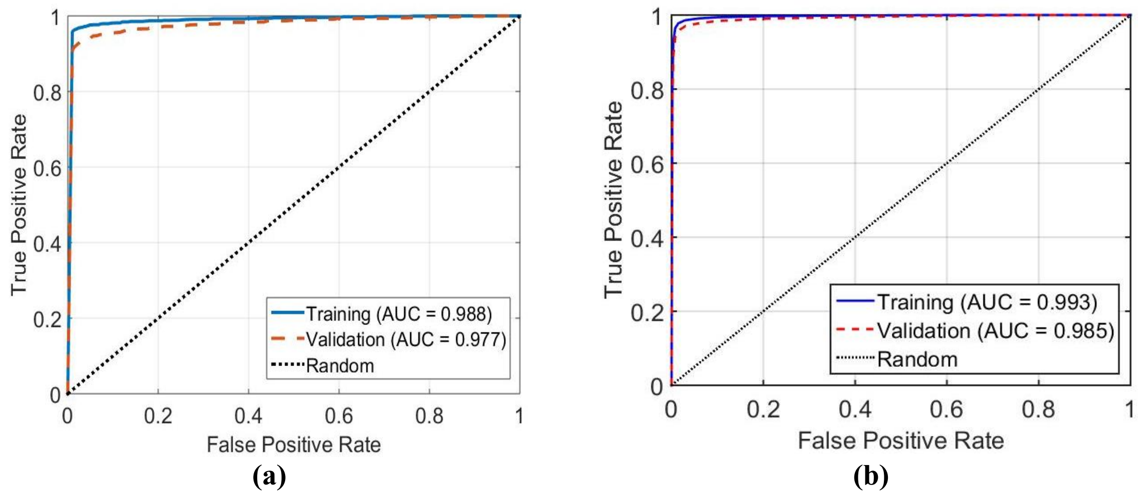


Figure 10. ROC and AUC curves (a) Dataset 1 (b) Dataset 2.

4.2 Subject-Level Cross-Validation Results

In order to perform a proper test on the generalization ability of the proposed attention-based fusion method, a 5-fold cross-validation at subject level was used. This would enable us in ruling out the probability of having any single EEG data from the same person in both training and test set to avoid any kind of data leakage issues. The current data set consists of 28 subjects, which comprise of 14 SZ patients and 14 healthy subjects. These 28 subjects have been distributed into five folds such that class balancing exists. In each fold, four folds comprising of 80 percent of the subjects were selected for training and validation. One-fold containing 20 percent of subjects was considered for testing. This process is repeated five times such that every subject is tested once.

Table 5 presents the subject-level 5-fold cross-validation results of the proposed method on Dataset 1. The model consistently achieved high classification performance across all five folds, demonstrating its robustness and generalization capability. The classification accuracy ranged from 98.05% to 98.63%, with the highest performance observed in Fold 4 and the lowest in Fold 3. Similarly, precision varied between 98.19% and 98.74%, while recall ranged from 97.91% to 98.45%. The F1-score remained consistently high across all folds, varying from 98.03% to 98.59%. Overall, the proposed method attained a mean accuracy of 98.34% ± 0.21, mean precision of 98.48% ± 0.20, mean recall of 98.15% ± 0.19, and mean F1-score of 98.31% ± 0.20. The low standard deviation values across all evaluation metrics indicate stable and reliable performance with minimal variation between folds, confirming the effectiveness of the proposed framework for subject-level schizophrenia classification.

Table 6 summarizes the subject-level 5-fold cross-validation performance of the proposed method on Dataset 2. The results demonstrate consistently high classification performance across all folds, indicating the robustness and reliability of the proposed framework. The classification accuracy varied from 98.49% to 98.91%, with the best performance achieved in Fold 4 and the lowest in Fold 3. Precision ranged between 98.61% and 99.02%, while recall varied from 98.33% to 98.76%. Likewise, the F1-score remained consistently high, ranging from 98.47% to 98.89% across the five folds. Overall, the proposed model achieved a mean accuracy of 98.68% ± 0.15, mean precision of 98.80% ± 0.16, mean recall of 98.52% ± 0.17, and mean F1-score of 98.66% ± 0.16. The low standard deviation values across all evaluation metrics indicate stable and consistent performance with minimal variability among the folds, demonstrating the strong generalization capability of the proposed method for subject-level schizophrenia classification on Dataset 2.

Table 7 presents the statistical significance analysis of the proposed method using both the paired *t*-test [30] and the Wilcoxon signed-rank test [31] across the four-evaluation metrics. The paired *t*-test yielded statistically significant results for all metrics, with *t*-values of 7.66, 8.00, 8.72, and 7.82 for accuracy, precision, recall, and F1-score, respectively, corresponding to *p*-values ranging from 0.0010 to 0.0016, all well below the significance threshold of 0.05. Similarly, the non-parametric Wilcoxon signed-rank test produced a test statistic (W) of 15 and a *p*-value of 0.0313 for all evaluation metrics, further confirming the statistical significance of the observed performance improvements. Since the *p*-values obtained from both statistical tests are less than 0.05, the null hypothesis is rejected, indicating that the performance gains achieved by the proposed method are statistically significant rather than resulting from random variation. These findings validate the effectiveness and reliability of the proposed schizophrenia classification framework across all considered performance metrics.

Table 5. Subject-Level 5-Fold Cross-Validation Results (Proposed Method: Dataset:1).

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold 1	98.21	98.35	98.02	98.18
Fold 2	98.47	98.62	98.28	98.44
Fold 3	98.05	98.19	97.91	98.03
Fold 4	98.63	98.74	98.45	98.59
Fold 5	98.32	98.48	98.11	98.29
Mean ± SD	98.34 ± 0.21	98.48 ± 0.20	98.15 ± 0.19	98.31 ± 0.20

Table 6. Subject-Level 5-Fold Cross-Validation Results (Proposed Method: Dataset:2).

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold 1	98.58	98.69	98.41	98.55
Fold 2	98.76	98.88	98.60	98.74
Fold 3	98.49	98.61	98.33	98.47
Fold 4	98.91	99.02	98.76	98.89
Fold 5	98.67	98.80	98.52	98.65
Mean ± SD	98.68 ± 0.15	98.80 ± 0.16	98.52 ± 0.17	98.66 ± 0.16

Table 7. Statistical Significance Analysis of the Proposed Method Using Paired *t*-Test and Wilcoxon Signed-Rank Test.

Metric	Paired *t*-test (*t*)	Paired *t*-test (*p*-value)	Wilcoxon *W*	Wilcoxon *p*-value	Significant ($p < 0.05$)
Accuracy	7.66	0.0016	15	0.0313	Yes
Precision	8.00	0.0013	15	0.0313	Yes
Recall	8.72	0.0010	15	0.0313	Yes
F1-score	7.82	0.0015	15	0.0313	Yes

4.3 Ablation Study

Tables 8 and 9 present the ablation study of the proposed framework on Dataset 1 and Dataset 2, respectively, using subject-level 5-fold cross-validation. The objective of this analysis is to quantify the contribution of each component of the proposed architecture, including the individual spectrogram and scalogram branches, feature fusion, and the attention mechanism.

Table 8. Ablation Study of Different Model Variants on Dataset 1 Using Subject-Level 5-Fold Cross-Validation.

Model Variant	Precision	Recall	F1-score	Accuracy
ViT (Spectrogram)	97.21 ± 0.29	96.93 ± 0.31	97.07 ± 0.30	97.10 ± 0.30
ViT (Scalogram)	97.74 ± 0.25	97.46 ± 0.27	97.60 ± 0.26	97.63 ± 0.26
Feature Fusion (Without Attention)	98.13 ± 0.22	97.86 ± 0.23	97.99 ± 0.22	98.02 ± 0.22
Feature Fusion + Attention (Proposed)	98.48 ± 0.20	98.15 ± 0.19	98.31 ± 0.20	98.34 ± 0.21

Table 9. Ablation Study of Different Model Variants on Dataset 2 Using Subject-Level 5-Fold Cross-Validation.

Model Variant	Precision	Recall	F1-score	Accuracy
ViT (Spectrogram)	97.56 ± 0.24	97.28 ± 0.25	97.42 ± 0.24	97.45 ± 0.24
ViT (Scalogram)	98.04 ± 0.20	97.81 ± 0.21	97.92 ± 0.20	97.96 ± 0.21
Feature Fusion (Without Attention)	98.43 ± 0.18	98.16 ± 0.19	98.29 ± 0.18	98.31 ± 0.18
Feature Fusion + Attention (Proposed)	98.80 ± 0.16	98.52 ± 0.17	98.66 ± 0.16	98.68 ± 0.15

For both datasets, the ViT (Spectrogram) model achieved the lowest performance, obtaining accuracies of 97.10% ± 0.30 and 97.45% ± 0.24 on Dataset 1 and Dataset 2, respectively. Replacing the spectrogram input with the ViT (Scalogram) branch improved the accuracy to 97.63% ± 0.26 on Dataset 1 and 97.96% ± 0.21 on Dataset

2, indicating that scalogram representations provide more discriminative time-frequency features for schizophrenia classification. Integrating both branches through feature fusion without attention further enhanced the classification performance, increasing the accuracy to 98.02% ± 0.22 on Dataset 1 and 98.31% ± 0.18 on Dataset 2, while also yielding consistent improvements in precision, recall, and F1-score. Finally, incorporating the proposed attention-based feature fusion produced the best overall performance, achieving accuracies of 98.34% ± 0.21 and 98.68% ± 0.15 on Dataset 1 and Dataset 2, respectively, together with the highest precision, recall, and F1-score values.

A comparison of the results demonstrates that each architectural enhancement contributes positively to the overall performance. On Dataset 1, the proposed model improves the classification accuracy by 1.24% over the ViT (Spectrogram) model, 0.71% over the ViT (Scalogram) model, and 0.32% over feature fusion without attention. Similarly, on Dataset 2, the proposed framework achieves accuracy improvements of 1.23%, 0.72%, and 0.37%, respectively. The consistent gains observed across all evaluation metrics confirm that both multimodal feature fusion and the attention mechanism effectively enhance feature representation by emphasizing the most informative time-frequency characteristics while suppressing redundant information. Furthermore, the lower standard deviation values obtained by the proposed model indicate more stable and consistent performance across the five subject-level folds, suggesting good generalization capability on both schizophrenia EEG datasets.

4.4 Comparison with the state-of-the-art method

Table 10 presents a comparative analysis of the proposed method with several state-of-the-art approaches using K-fold cross-validation. Bagherzadeh et al. [22] employed a DenseNet121-based deep CNN model on a dataset consisting of 14 schizophrenia (SZ) and 14 healthy control subjects, achieving an accuracy of 96.26%. Sharma et al. [23] applied a classical machine learning approach using the KNN classifier on the same dataset configuration and reported a higher accuracy of 97.20%, indicating that well-designed traditional methods can still perform competitively in EEG-based schizophrenia detection. Transformer-based methods further improved performance, where Shoeibi et al. [24] achieved an accuracy of 97.62% using a Transformer model on a larger dataset of 26 SZ and 30 healthy subjects, demonstrating the effectiveness of long-range temporal dependency modeling in EEG signals. In contrast, the LeViT-based approach reported by Beilin et al. [25] achieved an accuracy of 85.04%, indicating relatively lower generalization capability for this dataset.

In comparison, the proposed framework is evaluated using multiple model variants to demonstrate incremental improvements. The ViT-based spectrogram model achieves an accuracy of 96.80%, while the ViT-based scalogram model improves performance to 97.27%, highlighting the importance of time-frequency representation in EEG classification. Further improvement is observed with feature fusion without attention, which achieves an accuracy of 97.80%, confirming that combining complementary representations enhances discriminative capability. Finally, the proposed fusion model with attention mechanism achieves the highest accuracy of 98.13%, outperforming all compared methods. These results clearly demonstrate that adaptive attention-based multimodal fusion provides a consistent and significant improvement over both classical machine learning and advanced deep learning and Transformer-based methods for EEG-based schizophrenia detection.

Table 10. Comparison with the state-of-the-art methods (K-fold Cross Validation).

Category	Study	Model Description	Dataset Composition	Accuracy (%)
Deep CNN	Bagherzadeh et al. [22]	DenseNet121	14 SZ – 14 Healthy	96.26
Classical ML	Sharma et al. [23]	KNN	14 SZ – 14 Healthy	97.20
Transformer	Shoeibi et.al [24]	Transformer	26 SZ – 30 Healthy	97.62
LeViT	Beilin et.al [25]	Transformer	-	85.04
Proposed Variants	This Work	ViT (Spectrogram)	14 SZ – 14 Healthy	96.80
Proposed Variants	This Work	ViT (Scalogram)	14 SZ – 14 Healthy	97.27
Proposed Variants	This Work	Fusion (Without Attention)	14 SZ – 14 Healthy	97.80
Proposed Method	This Work	Fusion with Attention Mechanism	14 SZ – 14 Healthy	98.13

5. Conclusion

In this study, an efficient and high-performance DL framework for automatic SZ detection is presented using TFR of EEG signals. The proposed method effectively exploits discriminative information in both time and frequency domains through spectrogram and scalogram transformations. These representations are further fused into a unified feature space using a ViT model, enhanced with an attention mechanism, enabling more informative feature learning and improved classification performance. To ensure robustness and generalization, both random train–test splits and subject-level five-fold cross-validation are employed. The obtained results demonstrate strong and consistent performance across evaluation metrics, including accuracy, precision, recall, and F1-score, confirming the reliability of the proposed framework. Furthermore, an ablation analysis highlights the contribution of each component, showing that both feature fusion and the attention mechanism significantly improve classification performance. The proposed model is also extensively benchmarked against state-of-the-art methods in the literature. The comparison results indicate that the proposed approach achieves a maximum accuracy of approximately 98.1% outperforming traditional machine learning models, hybrid architectures, and several recent deep learning and Transformer-based methods. These results demonstrate the effectiveness of the proposed multimodal attention-guided framework for SZ detection. Overall, the proposed method offers a promising and reliable solution for automated schizophrenia diagnosis. Future work may focus on extending the model to large-scale and heterogeneous datasets, real-time clinical deployment, and adaptation to other neurological and psychiatric disorder classification tasks.

Author Contributions

Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing—original draft preparation, F. S.; Validation, supervision, writing—review and editing, and project administration, S. Z. N. Both authors contributed to the interpretation of the results and critically reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Data Availability Statement
No new data were generated during the preparation of this manuscript.

References

1. Khan, Farah Shan, and Shalini Z. Ninoria. "Predicting Cognitive Disorder Progression Using EEG Signals and Convolutional Neural Networks", In *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*, pp. 102–107, 2024.
2. Khan, Farah Shan, and Shalini Z. Ninoria. "Predicting Schizophrenia Using EEG Signals and Convolutional Long Short-Term Memory", In *2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, pp. 1-6, 2025.
3. Griffin, Daniel, and Jae Lim. "Signal estimation from modified short-time Fourier transform", *IEEE Transactions on acoustics, speech, and signal processing* 32(2), pp. 236–243, 1984.
4. Rioul, Olivier, and Pierre Duhamel. "Fast algorithms for discrete and continuous wavelet transforms", *IEEE transactions on information theory* 38(2), pp. 569–586, 2002.
5. Hosseini, Mohammad-Parsa, Amin Hosseini, and Kiarash Ahi. "A review on machine learning for EEG signal processing in bioengineering." *IEEE reviews in biomedical engineering* 14, pp. 204–218, 2020.
6. Wu, Youpeng, Lun Lu, Ao Xu, Yinan Wang, Zhiwei Li, Zhuanyi Yang, Lingli Zeng, and Qingjiang Li. "Neural networks for epilepsy detection and prediction with EEG signals: a systematic review." *Artificial Intelligence Review* 59(1), p. 30, 2026.
7. Khan, Farah Shan, and Shalini Ninoria. "Predictive Modelling of Mental Disorders Using EEG Signal Analysis: A Review of DNN Approaches", *Int. J. Intell. Commun. Comput. Sci.* 1(2), pp. 37–53, 2023.
8. Han, Kai, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang et al. "A survey on vision transformer", *IEEE transactions on pattern analysis and machine intelligence* 45(1), pp. 87–110,2022.
9. Dakka, Jumana, Pouya Bashivan, Mina Gheiratmand, Irina Rish, Shantenu Jha, and Russell Greiner. "Learning neural markers of schizophrenia disorder using recurrent neural networks", *arXiv preprint arXiv*:1712.00512 (2017).
10. Phang, Chun-Ren, Fuad Noman, Hadri Hussain, Chee-Ming Ting, and Hernando Ombao. "A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns", IEEE journal of biomedical and health informatics 24(5), pp. 1333–1343, 2019.
11. Palaniyappan, Lena, Gopikrishna Deshpande, Pradyumna Lanka, D. Rangaprakash, Sarina Iwabuchi, Susan Francis, and Peter F. Liddle. "Effective connectivity within a triple network brain system discriminates schizophrenia spectrum disorders from psychotic bipolar disorder at the single-subject level", Schizophrenia research 214, pp. 24-33, 2019.
12. A. Nikhil Chandran, K. Sreekumar, and D. Subha, "EEG-based automated detection of schizophrenia using long short-term memory (LSTM) network", in Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019, pp. 229–236, 2021.
13. J. Sun et al., "A hybrid deep neural network for classification of schizophrenia using EEG Data", Scientific Reports, 11(1), pp. 1–16, 2021.
14. A. Shoeibi et al., "Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models", Frontiers in neuroinformatics, p. 58, 2021.

15. Z. Aslan and M. Akin, "A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals", Physical and Engineering Sciences in Medicine, 45(1), pp. 83–96, 2022.
16. B. Gosala, P. D. Kapgate, P. Jain, R. N. Chaurasia, and M. Gupta, "Wavelet transforms for feature engineering in EEG data processing: An application on Schizophrenia", Biomedical Signal Processing and Control, 85, p. 104811, 2023.
17. S. Siuly, Y. Guo, O. F. Alcin, Y. Li, P. Wen, and H. Wang, "Exploring deep residual network based features for automatic schizophrenia detection from EEG", Physical and Engineering Sciences in Medicine, 46(2), pp. 561–574, 2023.
18. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016.
19. J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey", International Journal of Computer Vision, 129, pp. 1789-1819, 2021.
20. Shim, Miseon, Han-Jeong Hwang, Do-Won Kim, Seung-Hwan Lee, and Chang-Hwan Im. "Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features", *Schizophrenia research* 176(2-3), pp.314-319, 2016.
21. V. Jahmunah et al., "Automated detection of schizophrenia using nonlinear signal processing methods", Artificial intelligence in medicine, 100, p. 101698, 2019.
22. S. Bagherzadeh, M. S. Shahabi, and A. Shalbaf, "Detection of schizophrenia using hybrid of deep learning and brain effective connectivity image from electroencephalogram signal", Computers in Biology and Medicine, 146, p. 105570, 2022.
23. M. Sharma and U. R. Acharya, "Automated detection of schizophrenia using optimal wavelet-based l 1 norm features extracted from single-channel EEG", Cognitive Neurodynamics, 15(4), pp. 661–674, 2021.
24. Shoeibi, Afshin, Mahboobeh Jafari, Delaram Sadeghi, Roohallah Alizadehsani, Hamid Alinejad-Rokny, Amin Beheshti, and Juan M. Gorriz. "Early diagnosis of schizophrenia in EEG signals using one-dimensional transformer model", In *International work-conference on the interplay between natural and artificial computation*, pp. 139–149. Cham: Springer Nature Switzerland, 2024.
25. Li, Beilin, Jiao Wang, Zhifen Guo, and Yue Li. "Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals", *Expert Systems with Applications*, 224, p.119969,2023.
26. Liu, Chang, Wanzhong Chen, and Mingyang Li. "A hybrid EEG classification model using layered cascade deep learning architecture", *Medical & Biological Engineering & Computing* 62(7), pp. 2213–2229, 2024.
27. Wang, Yiping, Yanfeng Yang, Gongpeng Cao, Jinjie Guo, Penghu Wei, Tao Feng, Yang Dai, Jinguo Huang, Guixia Kang, and Guoguang Zhao. "SEEG-Net: An explainable and deep learning-based cross-subject pathological activity detection method for drug-resistant epilepsy", *Computers in Biology and Medicine*, 148, p.105703, 2022.
28. “EEG in schizophrenia - IBIB PAN”, Available at: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441 (Accessed on date: 20 Feb 2020).
29. Olejarczyk, Elzbieta; Jernajczyk, Wojciech, 2017, "EEG in schizophrenia", Available at: https://doi.org/10.18150/repod.0107441, RepOD, V1. (Accessed on date: 20 Feb 2020).
30. Hsu, Henry, and Peter A. Lachenbruch. "Paired t test." *Wiley StatsRef: statistics reference online* (2014).
31. Woolson, Robert F. "Wilcoxon signed-rank test." *Wiley encyclopedia of clinical trials*, pp.1–3, 2007.