

DGCSA: Dynamic graph convolution with self-attention for EEG-based attention recognition in VR and 2D environment

Wenyang Yang

Xi'an Shiyou University

Jingrui Yuan

Xi'an Shiyou University

Kexin Xu

Xi'an Shiyou University

Steven Kwok Keung Chow

`steven.chow@monash.edu`

Monash University

Research Article

Keywords: Electroencephalogram (EEG), Attention level recognition, Virtual learning environment, Dynamic Graph Convolutional Neural Network (DGCNN)

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10483588/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

DGCSA: Dynamic graph convolution with self-attention for EEG-based attention recognition in VR and 2D environment

Wenyang Yang^{1,2}, Jingrui Yuan¹, Kexin Xu¹, Steven Kwok Keung Chow³

¹ School of Computer Science, Xi'an Shiyou University, Xi'an, 710065, P.R. China

² Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an Shaanxi, P.R. China

³ Department of Medical Imaging and Radiation Sciences, School of Primary and Allied Health Care, Faculty of Medicine Nursing and Health Sciences, Monash University, Wellington Road, Clayton, VIC 3800, Australia

Corresponding authors: Steven Kwok Keung Chow (e-mail: steven.chow@monash.edu)

Steven's ORCID: <https://orcid.org/0000-0002-7201-4888>

Highlights (for review)

- We proposed DGCSA: dynamic graph convolution and self-attention for EEG attention.
- We used the Bayesian-optimized band weights to fuse PSD and DE while controlling dimension.
- The DGCSA model reached 91.67% accuracy for three-class attention, beating baselines.
- Collected a 3-level Schulte Grid EEG dataset with 873 segments from 20 participants
- VR learning showed higher and more stable attention than 2D, with $p < 0.001$
-

Abstract: Attention is a fundamental cognitive function in learning, and its precise recognition is essential for advancing personalized educational interventions. Conventional EEG-based methods are constrained by limited feature integration and inadequate modeling of dynamic spatial relationships, thereby restricting recognition accuracy. To address these challenges, a framework termed DGCSA (Dynamic Graph Convolutional Network with Self-Attention mechanism) is proposed for EEG-based attention recognition, particularly in virtual environment. The framework employs a Bayesian-optimized frequency-band weighting strategy to integrate power spectral density and differential entropy features, while combining dynamic graph convolution with self-attention to capture both spatial dependencies and global contextual information. Evaluation on EEG data from 20 participants performing standardized and virtual/2D learning tasks demonstrated that DGCSA achieved 91.67% accuracy in three-class classification, outperforming SVM, CNN, and LSTM baselines. Statistical analyses further indicated significantly higher and more stable attention levels in virtual learning compared to 2D scenarios ($p < 0.001$). These findings establish DGCSA as a robust solution for EEG-based attention decoding and provide neurophysiological

evidence that immersive virtual environments could effectively enhance cognitive engagement.

Keywords Electroencephalogram (EEG); Attention level recognition; Virtual learning environment; Dynamic Graph Convolutional Neural Network (DGCNN)

1 Introduction

Attention is a crucial psychological process in learning and cognitive activities, directly influencing task performance efficiency and learning outcomes [1]. Objective and real-time evaluation of individual attention states has long been a central research topic in cognitive neuroscience and educational technology. Electroencephalography (EEG), with its high temporal resolution, low cost, and noninvasiveness, has been widely applied in attention research and brain–computer interface (BCI) systems, providing reliable neurophysiological indicators for attention recognition [2].

In recent years, substantial progress has been achieved in EEG-based attention recognition. Traditional approaches typically employ frequency-domain features such as power spectral density (PSD) and differential entropy (DE), combined with machine learning classifiers including support vector machines (SVM), k-nearest neighbors (KNN), and linear discriminant analysis (LDA)[3, 4]. However, these approaches often rely on handcrafted features and struggle to fully capture the spatiotemporal dynamics of EEG signals. With the advancement of deep learning, models such as convolutional neural networks (CNN), recurrent neural networks (RNN), and Transformer architectures have been introduced into EEG analysis, leading to notable improvements in automated feature extraction and classification performance. Nevertheless, most existing studies simply concatenate multi-band features without incorporating effective adaptive fusion mechanisms. In addition, current research has primarily focused on 2D screen-based paradigms, while systematic investigations of EEG-based attention recognition in immersive virtual reality (VR) learning scenarios remain scarce.

VR technology, characterized by immersion and interactivity, has demonstrated unique advantages in education and training [5]. Behavioral studies have shown that VR environments could enhance learning concentration and task engagement [6]. However, neurophysiological evidence based on EEG remains lacking regarding the impact of VR on attention levels. The limitation not only restricts objective evaluation of VR-based learning effectiveness but also constrains the development of intelligent learning systems and educational BCIs. Consequently, exploring the feasibility and mechanisms of EEG-based attention recognition in VR learning scenarios holds both theoretical significance and practical value.

2 Related Work

Attention is a key cognitive function that enables individuals to selectively process information in complex environments. It encompasses fundamental characteristics such as concentration, stability, allocation, and shifting, and is closely associated with dynamic interactions among the

prefrontal cortex, parietal regions, and related neural networks [7-9] . In virtual learning environments, instructors face challenges in evaluating students' attentional states in real time through traditional methods, underscoring the need for objective and automated approaches to attention recognition.

Existing research on attention recognition could be broadly categorized into two types, with one focusing on behavior-based methods and the other on physiology-based methods. The former utilize facial expressions, body posture, and eye movements for modeling, offering the advantage of being noninvasive but suffering from susceptibility to external noise and deliberate disguise, which limits recognition accuracy [10, 11] . In contrast, EEG directly reflects neural activity of the brain and provides greater objectivity and temporal resolution, thereby demonstrating considerable potential for attention recognition in learning contexts [12, 13]. For example, Gao et al. [14] compared online instructional videos with different presentation styles, finding that attention levels were significantly enhanced when the instructor appeared alongside the teaching content and pointed to the corresponding information.

Traditional EEG-based approaches primarily rely on handcrafted features such as PSD, DE, and wavelet transforms, combined with classifiers including support vector machines (SVM), k-nearest neighbors (KNN), and random forests (RF). Zhang et al. [3] employed a fourth-order Butterworth filter to extract tri-band ratio features and achieved favorable classification performance using SVM and linear discriminant analysis (LDA). Peng et al. [4] adopted a DE-based strategy that incorporated both spectral features and spatial channel distributions. Zuo et al. [15] compared classifier performance in 2D and 3D VR environments, reporting that RF was more suitable for handling high-dimensional EEG data. Although these methods have demonstrated the capacity to reflect frequency-band differences associated with attention, they remain limited in addressing the high-dimensional and nonlinear characteristics of EEG signals.

With the development of deep learning, models such as convolutional neural networks (CNN) and long short-term memory networks (LSTM) have been introduced into EEG analysis. Pathak et al. [16] classified real-time EEG signals recorded from learners during online courses using CNN and LSTM, achieving classification accuracies of 68% and 97%, respectively. Rivas et al. [17] exploited the temporal sequence modeling capability of recurrent neural networks (RNN) to evaluate the effectiveness of LSTM and gated recurrent units (GRU) in predicting future cognitive states, while struggling to capture cross-channel spatial dependencies. To overcome such limitations, graph neural networks (GNN) have been increasingly adopted to model topological relationships among EEG channels, while attention mechanisms have been employed to weight salient features and thereby enhance both accuracy and robustness of classification models [18, 19]. Li et al. [20] proposed a hybrid model combining graph attention mechanisms with a Transformer encoder, which improved performance in emotion recognition. Yang et al. [21] developed an attention-based GNN algorithm, where empirical mode decomposition was used to construct graphs and optimize adjacency matrices, achieving 99.83% accuracy in EEG classification tasks. The integration of GNN and self-attention offers a promising research direction for attention recognition based on EEG signals.

There still remain two major limitations despite the technological advancement . First, multi-band features are often fused through simple concatenation, which fails to fully exploit the

discriminative information across frequency bands. Second, most existing graph convolution approaches rely on fixed adjacency structures, limiting their ability to represent dynamic spatial relationships among EEG channels, while integration with attention mechanisms remains insufficient. Future research therefore requires further exploration of adaptive multi-band feature fusion combined with dynamic graph modeling and global attention mechanisms. Building upon this perspective, the present study proposes a frequency-band weighted representation that integrates PSD and DE features could achieve more accurate attention recognition from EEG signals in virtual learning scenarios.

To address the aforementioned limitations, we develop an EEG-based attention recognition approach that integrates multi-band feature fusion, a dynamic graph convolutional neural network (DGCNN), and a self-attention (SA) mechanism. The SA module learns band-specific attention weights across frequency bands, enabling adaptive emphasis of the most informative spectral components. In contrast to conventional methods that treat EEG spatial relationships uniformly, the DGCNN explicitly captures the directional organization of brain activity by modeling interactions along the different axes. At the feature level, weighted fusion of power spectral density (PSD) and differential entropy (DE) is employed to capture discriminative information across different frequency bands. At the modeling level, DGCNN is introduced to characterize dynamic spatial relationships among EEG channels, where adjacency matrices are adaptively updated based on feature similarity to better capture spatiotemporal dependencies. The incorporation of the SA module further enhances global dependency modeling and strengthens feature discriminability.

Finally, experiments are conducted in both VR and conventional 2D learning scenarios to examine the impact of environmental settings on attention levels. Experimental results demonstrate that the proposed approach significantly outperforms multiple baseline models, achieving superior classification performance and confirming that VR environments effectively enhance learning-related attention. The study introduces a novel methodological framework for EEG-based attention recognition, which provides empirical evidence for the objective evaluation of VR-based educational effectiveness.

The main contributions of this study can be summarized as follows:

1. A controlled VR and 2D Schulte Grid protocol with three attention levels is introduced, and EEG is labeled using behavioral indicators to produce a curated dataset for classification.
2. A PSD and DE fusion method with Bayesian optimized band weights is proposed, improving feature discriminability and adaptability while controlling dimensionality.
3. The DGCSA model combines dynamic graph convolution and self attention, outperforming MLP, CNN, and LSTM on multi class attention recognition; VR shows higher accuracy and stability than 2D, and band level topographies (Theta, Alpha, Beta) illuminate the neural mechanisms of attention in immersive learning.

2.1 DGCSA Model Architecture

To overcome limitations of conventional classifiers in EEG-based attention recognition, such as strong reliance on handcrafted features, weak generalization, and inability to model complex

spatiotemporal relationships, a hybrid deep model integrating DGCNN and SA, referred to as DGCSA, was developed to achieve high-precision classification of multichannel EEG signals.

The model consists of three primary modules. The DGCNN module dynamically models the spatial topology of EEG signals, while the SA module captures global temporal dependencies across channels. Outputs from both modules are combined through a weighted fusion mechanism and subsequently fed into a fully connected layer for classification. The overall architecture of the DGCSA model is depicted in Fig.1.

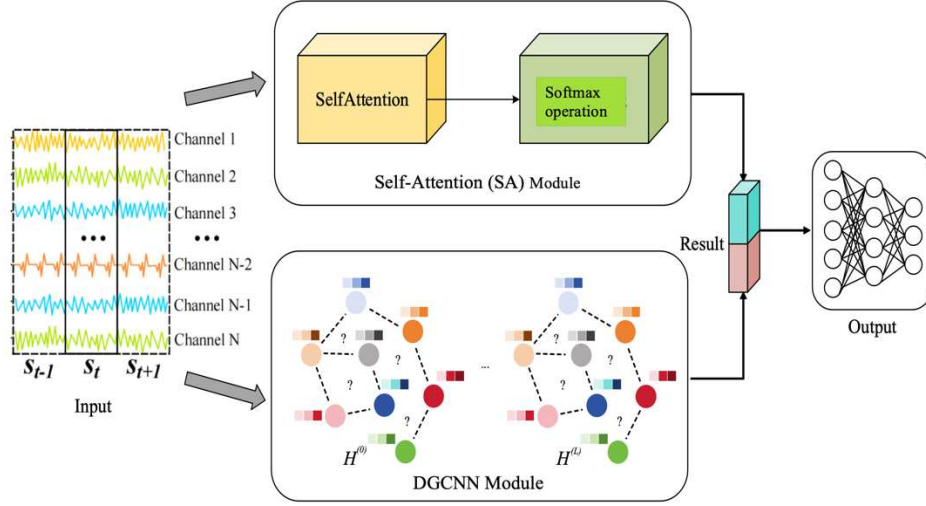


Fig. 1. Overall architecture of the DGCSA model

2.1.1 DGCNN Module

In the DGCNN module, EEG channels are modeled as nodes in a graph, and a learnable adjacency matrix is employed to dynamically characterize the latent functional connectivity between channels. Let the input feature be $X \in \mathbb{R}^{b \times n \times f}$, where b denotes the batch size, n is the number of EEG channels (i.e., graph nodes), and f represents the feature dimension of each channel corresponding to the fused sub-band features. A learnable adjacency matrix $A \in \mathbb{R}^{n \times n}$ is first initialized. To ensure numerical stability and incorporate self-connections, the adjacency matrix is symmetrically normalized as Eq. (1) :

$$\tilde{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \quad \text{Eq. (1)}$$

Here, $\tilde{A}$ denotes the symmetrically normalized adjacency matrix; $I \in \mathbb{R}^{n \times n}$ denotes the identity matrix, and D is the degree matrix with diagonal elements defined as $D_{ii} = \sum_j A_{ij}$.

To enhance the modeling capacity for high-order graph structural information, Chebyshev polynomial-based graph convolution is adopted. The Chebyshev polynomials $T_k(\tilde{A})$ are recursively defined as Eq. (2):

$$T_0 = I, T_1 = \tilde{A}, T_k = 2\tilde{A}T_{k-1} - T_{k-2} \quad \text{Eq. (2)}$$

where k denotes the polynomial order.

The node representation at the l -th graph convolution layer is updated according to Eq. (3):

$$H^{(l)} = \sigma(\sum_{k=0}^K \theta_k T_k(\tilde{A}) H^{(l-1)} W^{(l)}) \quad \text{Eq. (3)}$$

where, $H^{(l)}$ and $H^{(l-1)}$ denote the output and input node representations of the l -th layer (with $H^{(0)} = X$), $W^{(l)}$ is the learnable weight matrix, θ_k represents the Chebyshev coefficients, K is the maximum polynomial order, and $\sigma(\cdot)$ denotes the nonlinear activation function. ReLU is used in this work. By setting $K=40$, the DGCNN is able to capture high-order spatial dependencies among EEG channels over a large receptive field, producing 64-dimensional topology-aware feature representations.

2.1.2 Self-Attention (SA) Module

The SA module is introduced to explicitly model global dependencies among EEG channel nodes, thereby enhancing the representation of node features with respect to overall inter-channel interaction patterns. Linear projections are first applied to generate the query (Q), key (K), and value (V) matrices in Eq. (4). In addition, the input feature be $X \in R^{b \times n \times f}$, where b , n , and f denote the batch size, number of EEG channels, and feature dimension per channel, respectively.

$$Q = XW_Q, K = XW_K, V = XW_V \quad \text{Eq. (4)}$$

where $Q, K, V \in R^{b \times n \times d}$, W_Q, W_K, W_V are learnable weight matrices, and d denotes the dimensionality of the query and key feature space. The attention weight matrix is then computed as Eq. (5):

$$S = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \quad \text{Eq. (5)}$$

where $S \in R^{b \times n \times n}$. The scaling factor $\sqrt{d}$ is used to alleviate gradient instability caused by large dot-product values, and the Softmax operation is applied along the node dimension to ensure that the attention weights for each node sum to one.

The output of the SA module is obtained as $Z = SV, Z \in R^{b \times n \times d}$, representing globally attention-weighted node features. To facilitate feature fusion with the DGCNN module, the SA output is further projected to a 64-dimensional feature space via a linear transformation.

2.1.3 Fusion Mechanism and Classifier Design

The outputs of the DGCNN module and the SA module are denoted as $H \in R^{b \times n \times 64}$ and $Z \in R^{b \times n \times 64}$, respectively, which share the same dimensionality. A weighted fusion strategy is adopted to construct the final feature representation, defined as Eq (6):

$$F = \alpha \cdot H + (1 - \alpha) \cdot Z, \alpha = 0.5 \quad \text{Eq. (6)}$$

where $F \in R^{b \times n \times 64}$ denotes the fused feature, and α controls the relative contribution of graph-based spatial features and self-attention-based global features. In this study, equal weights are assigned to both branches to balance local topological information and global dependency modeling.

Following graph structure modeling and attention feature extraction, the fused features were flattened and fed into three successive fully connected layers for dimensionality reduction and classification. The first and second fully connected layers produced 32-dimensional and 8-dimensional feature representations, respectively, each followed by a ReLU activation function to enhance nonlinear modeling capacity. Dropout was employed during training to mitigate overfitting.

The final output layer applied a Softmax activation function to map the low-dimensional

features to a three-class attention-level label space, enabling multi-class classification of low, medium, and high attention states. Model parameters were optimized via backpropagation, with cross-entropy loss serving as the objective function and the Adam optimizer selected to improve convergence speed and generalization capability. The configuration ensured robust classification performance and stability of the model in virtual learning scenarios.

3 Methodology

The complete pipeline for developing and validating DGCSA, a dynamic graph convolutional network with self-attention, for EEG-based attention recognition in immersive learning is presented. Participant recruitment and task design are described first, followed by EEG acquisition and preprocessing. Feature extraction using power spectral density and differential entropy is then presented, together with a frequency band weighted fusion tuned by Bayesian optimization. Finally, the model architecture, training protocol, and evaluation metrics used to assess performance are outlined.

3.1 Data Acquisition and Preprocessing

3.1.1 Participants

A total of 20 university students (aged 18 to 23 years; mean 21.7; equal gender distribution) participated in the study. All were neurologically healthy, had no history of psychiatric disorders, had normal or corrected-to-normal vision and hearing, were right-handed, and reported no prior attention-related training. Written informed consent was obtained from all participants, who were advised of their right to withdraw at any time.

3.1.2 Experimental Design

In the Schulte Grid study, a completion time of 25 seconds is considered the passing threshold for adults, while 8 seconds indicates a higher level of attention [22]. A completion time exceeding 25 seconds may suggest attention deficits, further confirming the sensitivity and discriminative capacity for attention assessment. In related research, Wang et al. employed a 5×5 Schulte Grid in an attention induction experiment, classifying attention levels based on task completion time and accuracy [23]. Drawing on a large body of prior studies, the present research adopted a scientifically grounded classification of attention levels. Specifically, high attention was defined as a completion time of 10–18 seconds with 100% accuracy, medium attention as 18–26 seconds with at least 95% accuracy, and low attention as an eyes-closed resting state with no clicks. These criteria considered both the statistical distribution of completion times and the physiological characteristics of EEG signals, ensuring the validity and reliability of attention-state labeling.

The experimental paradigm of the Schulte Grid task is illustrated in Fig. 2B and comprises both practice and formal testing phases. At the beginning, participants received standardised instructions and completed training trials to familiarise themselves with the procedure. In the formal phase, participants were required to sequentially click numbers from 1 to 25. Any incorrect click triggered an immediate red flash on the grid as an error prompt. Upon task completion, the

experimenter securely stored the data and documented both task completion time and accuracy, providing essential metrics for subsequent analysis.

Two complementary experimental paradigms were designed to capture EEG signals across different attention levels. In the first paradigm, a dynamic Schulte Grid task developed in Unity3D was used to construct a standardized attention-level dataset. Participants sequentially clicked numbers from 1 to 25 in a randomly arranged 5×5 matrix, with task completion time and accuracy jointly determining attention levels. Based on previous studies and the present data, attention was categorized as high (10–18 s, 100% accuracy), medium (18–26 s, $\geq 95\%$ accuracy), or low (eyes closed, no clicks). Each participant completed three rounds of the task with 15-minute breaks between rounds to minimize fatigue. The dataset comprised 873 EEG segments (3 seconds each), including 242 high-attention, 266 medium-attention, and 365 low-attention segments, recorded across 64 channels with 1,500 sampling points per segment.

The second paradigm involved a comparative experiment between virtual and 2D learning scenarios to examine the influence of environmental context on attention. Both scenarios were based on the theme of petroleum extraction and are shown in Fig. 2E. The 2D scenario was created in Unity 3D and Photoshop, incorporating voice narration and dynamic images. The virtual scenario, on the other hand, integrated SteamVR and the VRTK plugin, delivering an immersive experience via an HTC Vive headset, as shown in Fig. 2D. Participants were randomly assigned to complete the two tasks in either virtual–2D or 2D–virtual order. Fig. 2F. shows set up the 2D environment and Fig. 2G. shows set up the VR environment to collect the attention EEG signal. Each task lasted 90 seconds, with a one-day interval between sessions. EEG signals were continuously recorded throughout both tasks for subsequent analysis of attention differences.

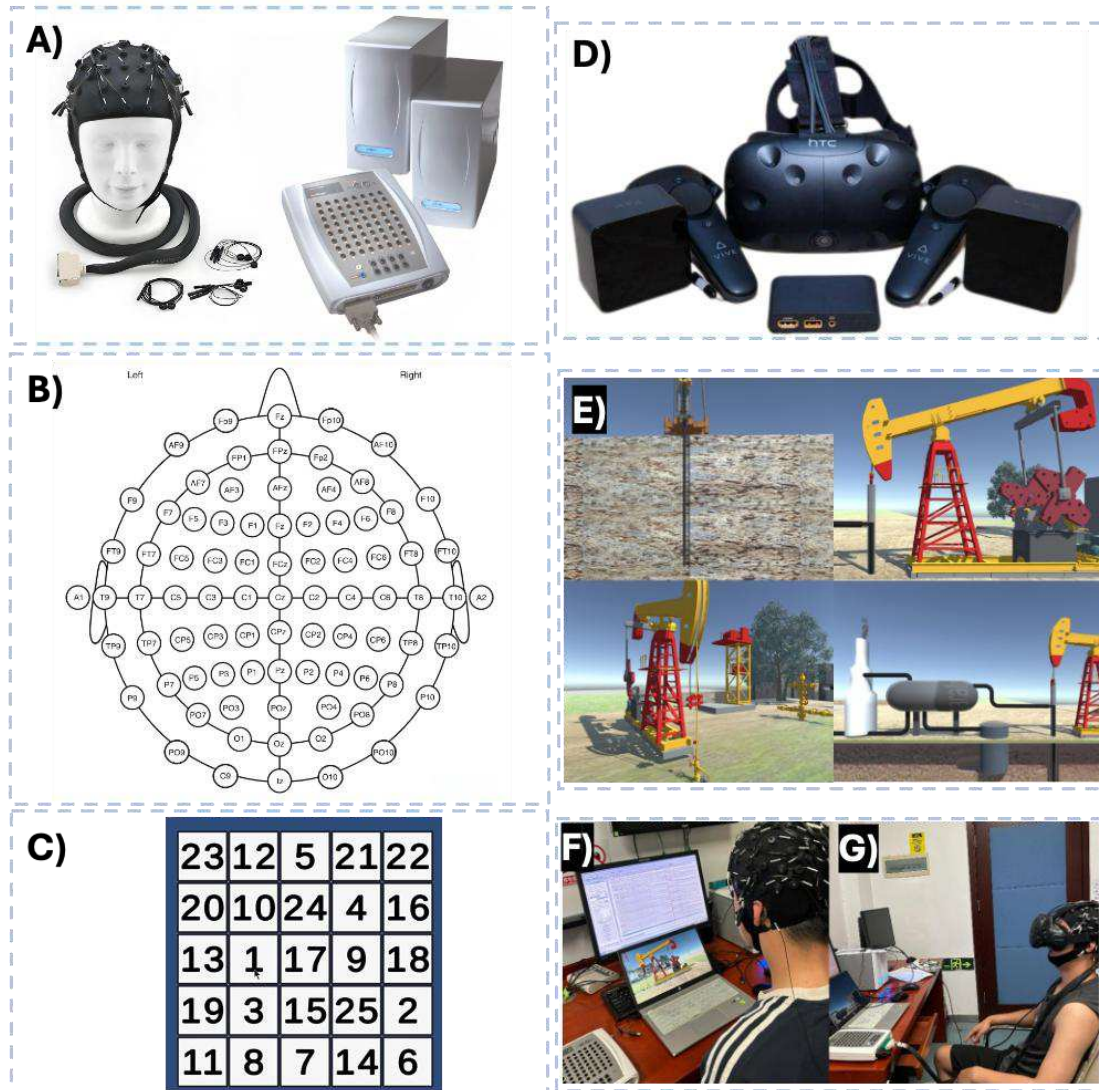


Fig. 2. Experimental setup and task materials. (A) Neuroscan SynAmps² amplifier with 64-channel Quik-Cap used for EEG acquisition. (B) Electrode layout following the international 10–20 system. (C) Sample Schulte Grid (5×5) used for attention induction. (D) HTC Vive headset for immersive VR. (E) Learning scenario theme: petroleum extraction. (F) 2D learning environment created in Unity3D with narration and dynamic images. (G) VR learning environment implemented with SteamVR and VRTK on an HTC Vive headset.

3.1.3 EEG Acquisition

To ensure the quality of EEG data acquisition, all experiments were conducted in an electromagnetically shielded room that met the standards for electrophysiological recordings. This setting effectively suppressed 50 Hz power-line interference and minimized electromagnetic noise from the external environment. The laboratory temperature was maintained at 22 to 25 °C, and the relative humidity was controlled at 40% to 60% to prevent fluctuations in electrode impedance caused by sweating, thereby ensuring stable electrode–scalp contact.

Participants underwent standardised preparation procedure for EEG data collection and informed consent for use of data and EEG experiment regulation by the university. Neuroscan

SynAmps 2 were used in this study and are shown in Fig. 2A. The scalp was cleaned to improve conductivity, and a 64-channel Quik-Cap was fitted according to the international 10-20 system. The outline of the scalp is shown in Fig. 2B. The cap was connected to a Neuroscan SynAmps 2 EEG amplifier via an interface box, with electrode positions aligned to anatomical landmarks (nasion and inion) to ensure full cortical coverage, including frontal (Fp1/Fp2/Fz), central (C3/C4/Cz), parietal (P3/P4/Pz), occipital (O1/O2/Oz), and temporal (T7/T8) regions. Participants received detailed instructions before testing to minimize anxiety and non-neural artifacts, supporting stable and high-quality data acquisition.

3.2 EEG Data Preprocessing

EEG was processed in MATLAB R2022b using EEGLAB. Data were average-referenced, band-pass filtered (0.05–50 Hz), and notch-filtered at 48 to 52 Hz to remove power-line noise. Ocular channels (HEO/VEO) were excluded; noisy electrodes were corrected by interpolation. To facilitate artefact inspection, recordings were initially segmented into 2s epochs with baseline correction, and gross-motion artefacts were removed following visual review. Independent component analysis (ICA) was then used to identify and reject ocular/muscle/environmental components (semi-automatic labeling with manual confirmation). For downstream modeling and feature extraction, the cleaned signals were re-segmented into non-overlapping 3s windows, aligning with the dataset specification in Section 3.1.2. Cleaned data were stored in MATLAB format for subsequent analysis.

EEG data were processed in MATLAB R2022b using EEGLAB. Signals were re-referenced to the common average, band-pass filtered between 0.05 and 50 Hz, and notch-filtered at 48–52 Hz to suppress power-line interference. Ocular channels (HEO/VEO) were excluded, and channels exhibiting abnormal noise were corrected via spherical interpolation based on neighboring electrodes. To facilitate artefact inspection, continuous recordings were initially segmented into 2-s epochs with baseline correction. Segments containing gross motion artefacts were identified and removed through visual inspection. Independent component analysis (ICA) was subsequently applied to decompose the EEG into temporally independent components. Artefact-related components (e.g., ocular, muscular, and environmental noise) were identified using EEGLAB's semi-automatic component labeling and subsequently verified through manual inspection of scalp topographies, time-domain activations, and power spectral characteristics, following standard EEGLAB practice (see Fig. 3 for an illustrative example). Only components confirmed as non-neural were rejected. For downstream modeling and feature extraction, the cleaned EEG signals were re-segmented into non-overlapping 3-s windows, consistent with the dataset specification described in Section 3.1.2. The final preprocessed data were stored in MATLAB format for subsequent analysis.

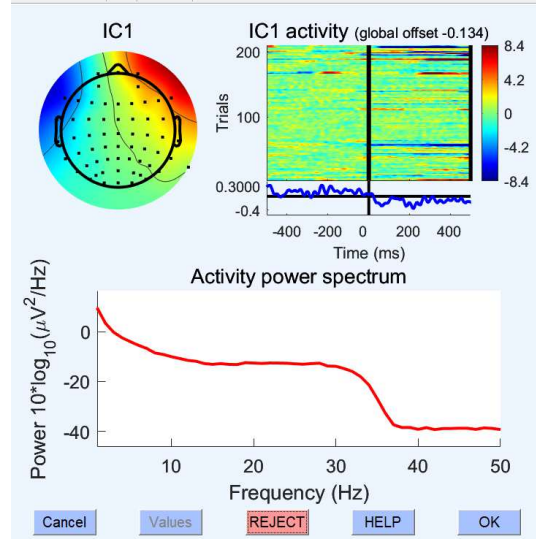


Fig. 3. An illustrative example of manual ICA artefact inspection using EEGLAB

3.3 Feature Extraction and Fusion for Attention Level Modeling

Two complementary feature families were extracted from each 3-s window: power spectral density (PSD) and differential entropy (DE). PSD characterizes frequency-domain energy, whereas DE quantifies signal complexity under an approximate Gaussian assumption after band-limited transformation. Signals were partitioned into eight frequency bands (Delta 0.5 to 4 Hz, Theta 4 to 8 Hz, Alpha1 8 to 10.5 Hz, Alpha2 10.5 to 13 Hz, Beta1 13 to 21.5 Hz, Beta2 21.5 to 30 Hz, Gamma1 30 to 40 Hz, Gamma2 40 to 50 Hz). For each 3s window, PSD and DE matrices of size 62×8 (channels × bands) were computed.

To leverage complementary information while avoiding dimensional inflation from simple concatenation, we employed a Bayesian-optimized frequency-band weighting strategy that assigns band-specific weights to PSD and DE and fuses them into a single representation. The optimization used a Gaussian-process surrogate with expected improvement, exploring discretized weight grids to maximize validation accuracy efficiently. The resulting fused features served as input to the downstream DGCSA classifier.

3.2.1 PSD-Based Feature Extraction

PSD characterizes the energy distribution of EEG signals across frequency components and is one of the most informative frequency-domain features for attention level recognition. In this study, the Welch method was employed to estimate the PSD of EEG signals, as it produces smooth and robust spectral representations through signal segmentation, windowing, Fourier transformation, and power averaging.

Specifically, given an EEG signal x_n of length N , the Welch method divides the signal into K segments, each with a length of L , and with an overlap of D samples between adjacent segments (typically $D = L/2$). The number of segments K is calculated as Eq. (7):

$$K = \left\lfloor \frac{N-D}{L-D} \right\rfloor \quad \text{Eq. (7)}$$

where N denotes the length of the original signal.

After segmentation, a window function is applied to each segment to reduce spectral leakage. In this study, the Hanning window was adopted, which is mathematically defined as Eq. (8):

$$w(n) = 0.5 \left(1 - \cos \left(\frac{2\pi n}{L-1} \right) \right), n = 0, 1, \dots, L-1 \quad \text{Eq. (8)}$$

The window length was set to 1,000 samples, corresponding to 2 seconds of EEG data. The Hanning window attenuates signal edges and effectively suppresses spectral leakage, thereby improving spectral estimation accuracy.

Let $x_i(n)$ denote the i -th segmented signal ($1 \leq i \leq K$). The windowed signal $y_i(n)$ is obtained by multiplying $x_i(n)$ with the Hanning window. The windowed time-domain signal is then transformed into the frequency domain using the fast Fourier transform (FFT). The spectrum of the i -th segment, $Y_i(k)$, is computed as Eq. (9):

$$Y_i(k) = \sum_{n=0}^{L-1} y_i(n) e^{-j \frac{2\pi}{L} kn}, k = 0, 1, \dots, L-1 \quad \text{Eq. (9)}$$

where L denotes the window length, j is the imaginary unit, and k represents the frequency index. Through the FFT, each windowed EEG segment is converted from the time domain to the frequency domain, yielding complex-valued spectral coefficients containing both amplitude and phase information.

The power spectrum of each segment is obtained by computing the squared magnitude of its spectrum. The PSD estimate is then calculated by averaging the power spectra across all segments, as expressed by Eq. (10):

$$P(k) = \frac{1}{K} \sum_{i=1}^K |Y_i(k)|^2 \quad \text{Eq. (10)}$$

To further enhance frequency resolution and align with the neurophysiological characteristics of brain rhythms, the EEG signals were divided into eight sub-bands (see Section 3.3.3). For each channel, the average PSD value within each frequency band was computed. The resulting PSD feature matrix has a dimensionality of 62×8 , capturing the spatial-spectral energy distribution across EEG channels.

3.2.2 DE-Based Feature Extraction

DE is a continuous extension of Shannon entropy and is commonly used to quantify the uncertainty and complexity of signals. It is particularly effective in characterizing nonlinear and diverse frequency-domain properties of EEG signals. Previous studies have shown that EEG signals, after band-pass filtering and frequency-domain transformation, exhibit sub-band features that approximately follow a Gaussian distribution. Under this assumption, DE can be employed to quantify the energy complexity of different frequency bands. For an EEG signal segment that approximately follows a Gaussian distribution $N(\mu, \sigma_i^2)$, the DE is defined as Eq. (11):

$$DE = - \int_a^b f(x) \log(f(x)) dx = \frac{1}{2} \log(2\pi e \sigma_i^2) \quad \text{Eq. (11)}$$

where, $f(x)$ denotes the probability density function, $[a, b]$ represents the range of signal values, and σ_i^2 is the signal variance. This formulation indicates that the DE corresponds to the logarithm of the energy spectrum within the target frequency band.

In practice, the time-domain EEG signals were first transformed into the frequency domain using the FFT. The variance and DE values were then computed for each frequency band. Finally,

the DE features from all EEG channels were stacked row-wise to form a feature matrix of size 62×8 for each sample, which was subsequently used as input for downstream modeling and classification.

3.2.3 Frequency-Band Weighted Fusion of PSD and DE Features

To fully leverage the complementary strengths of different features in attention-level recognition, a frequency-band-weighted fusion strategy based on PSD and DE was proposed. Specifically, preprocessed EEG signals were segmented into non-overlapping 3-second epochs and divided into eight frequency bands: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha1 (8–10.5 Hz), Alpha2 (10.5–13 Hz), Beta1 (13–21.5 Hz), Beta2 (21.5–30 Hz), Gamma1 (30–40 Hz), and Gamma2 (40–50 Hz). PSD and DE features were subsequently extracted from each segment and frequency band, resulting in two feature matrices of dimension 62×8 for each sample.

During the feature fusion stage, distinct fusion weights were assigned to PSD and DE within each frequency band to reflect the relative contribution of each feature. Given the large number of frequency bands and two tunable weight parameters per segment (PSD and DE), exhaustive grid search would result in 11^{16} possible combinations, leading to high computational cost and inefficiency. To address the issue, a Bayesian optimization algorithm was employed to efficiently explore the weight parameter space and identify optimal fusion weights.

Bayesian optimization constructed a Gaussian process model to simulate the distribution of the objective function, defined as classification accuracy, within the parameter space, and employed an expected improvement (EI) strategy to dynamically balance exploration and exploitation. In implementation, fusion weights were discretized into grid points with a step size of 0.1, which substantially reduced the search space complexity. The optimization procedure followed a two-stage strategy. In the initial stage, ten candidate weight sets were generated through random sampling to cover the global parameter space. Subsequently, fifty iterative rounds were conducted, guided by the model to focus on regions with potential high performance, resulting in a total of sixty evaluations, which offered significant efficiency advantages over exhaustive search.

In the training iteration, the candidate weight combination was applied to fuse PSD and DE features. Classifiers including Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Trees (GBDT) were trained on the training set and evaluated on the validation set, with classification accuracy used as the objective function. The weight configuration that achieved the highest average performance was selected as the final feature fusion strategy.

The optimized weights are summarized in Table 1. Low-frequency bands were predominantly weighted toward PSD features, high-frequency bands were dominated by DE features, and mid-frequency bands adopted a balanced weighting strategy. This frequency-specific fusion reflects distinct neural mechanisms underlying attention states. Energy changes in low-frequency signals largely corresponded to cortical inhibition associated with dispersed attention, whereas complexity in high-frequency signals was more sensitive to synchronized neuronal firing linked to focused attention. All preprocessed and feature-extracted data were subsequently used for constructing and comparing attention-level classification models.

Table 1. Frequency Band Definitions and Optimal Feature Weights

Frequency Band	Range (Hz)	PSD Weight	DE Weight
----------------	------------	------------	-----------

Delta	0.5–4	0.7	0.3
Theta	4–8	0.6	0.4
Alpha1	8–10.5	0.5	0.5
Alpha2	10.5–13	0.5	0.5
Beta1	13–21.5	0.4	0.6
Beta2	21.5–30	0.4	0.6
Gamma1	30–40	0.3	0.7
Gamma2	40–50	0.3	0.7

4 Results

4.1 Effectiveness of the Frequency-Band Weighted Feature Fusion Method

To assess the effectiveness of the proposed frequency-band-weighted fusion of PSD and DE for attention recognition, we evaluated with support vector machine (SVM), random forest (RF), and gradient boosting decision tree (GBDT). The experimental configuration and detailed results are reported below.

4.1.1 Experimental Settings and Classifier Configuration

All classifiers were implemented in scikit-learn and evaluated using five-fold cross-validation to ensure stable estimates. For SVM, we used an RBF kernel and tuned the regularization parameter C via grid search to balance bias–variance. The RF model employed 100 trees with bootstrap sampling and randomized feature selection at each split (`random_state = 42`) to enhance ensemble diversity. The GBDT model used 100 estimators, learning rate = 0.1, and `random_state = 42`. Each classifier was trained on the same dataset under four feature configurations on PSD only, DE only, PSD+DE concatenation, and the proposed frequency-band-weighted fusion—and validation accuracy was recorded for comparison.

4.1.2 Classification Performance Comparison and Analysis

To validate the effectiveness of the proposed frequency band weighted fusion strategy, we evaluated four feature representations across three classifiers (SVM, RF, and GBDT). The feature sets were PSD only, DE only, PSD plus DE concatenation, and the frequency band weighted fusion. Fig. 4 presents the classification accuracy of these feature sets for each classifier.

The results demonstrate that frequency-band weighted fusion features achieved the highest performance in all classifiers, with accuracies of 84.8% (SVM), 87.9% (RF), and 81.8% (GBDT), which were markedly higher than those of the other feature combinations. By comparison, PSD only performed relatively well for RF and GBDT but was substantially weaker for SVM, whereas DE only produced the lowest accuracy across classifiers, indicating limited discriminative power

for attention recognition when used in isolation.

Concatenated PSD and DE features showed a slight improvement over PSD-only features with SVM, but resulted in marginally reduced performance with RF and GBDT. This reduction could be attributed to the doubled feature dimensionality, which introduced redundant information, thereby increasing model complexity and negatively affecting recognition performance. In comparison, the frequency-band weighted fusion approach enhanced feature discriminability by automatically learning the optimal combination weights of PSD and DE in each frequency band through Bayesian optimization. The method further improved classification performance while maintaining controlled dimensionality.

An additional observation is that the performance gain from the weighted fusion was more pronounced for SVM and RF, whereas GBDT showed relatively limited improvement. This likely reflects different sensitivities to feature representations. RF, as an ensemble with subspace sampling and majority voting, better exploits the complementarity between PSD and DE. In contrast, GBDT follows a residual driven, sequential learning strategy in which the dominance of PSD within the fusion can suppress contributions from auxiliary features such as DE, limiting the gain. Even so, the weighted fusion still outperformed both concatenated and single feature inputs under GBDT, indicating robust generalizability. Overall, the frequency band weighted fusion consistently demonstrated superior adaptability and performance across classifiers, confirming its effectiveness for attention level recognition.

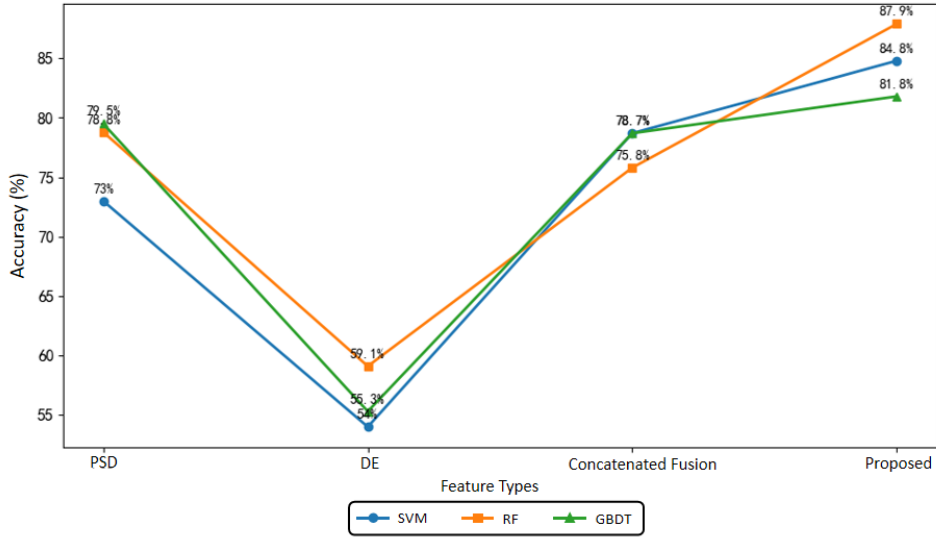


Fig. 4 Comparison of classification performance across classifiers

4.2 Performance Evaluation of the DGCSA Model

The section systematically evaluates the attention recognition performance of the proposed DGCSA model on multi-channel EEG signals in virtual learning scenarios, based on the previously constructed sub-band weighted fusion feature set. The experimental design includes training convergence analysis, confusion matrix visualization, ablation experiments of model modules, and comparative analysis with typical deep learning models (such as MLP, CNN, and LSTM), in order to comprehensively validate the effectiveness and robustness of the model.

4.2.1 Model Performance Evaluation and Confusion Matrix Analysis

To evaluate DGCSA for attention-level recognition, we trained on the frequency-band-weighted feature set using an 85:15 train-test split. Optimization used AdamW with cross-entropy loss (maximum 100 epochs, batch size = 8, dropout = 0.2, learning rate = 5×10^{-4}). During training, performance was tracked on a held-out validation subset drawn from the training split, and the checkpoint with the highest validation accuracy was used for final test evaluation.

Fig. 5 shows the Training and validation curves and test-set confusion matrix for the DGCSA model. Training accuracy rose rapidly within the first ~20 epochs and stabilized near 95% by about epoch 32; test accuracy followed a similar trajectory and stabilized around 92%. Training and validation losses decreased markedly over the first ~30 epochs and then plateaued with a small, stable gap. Minor late-stage fluctuations in validation loss were observed without evidence of overfitting, indicating stable optimization and good generalization.

Applying the trained DGCSA to the held-out test set yielded a three-class confusion matrix that illustrates discrimination across attention levels (Fig. 5c). Class labels are SL (low), SM (medium), and SH (high). DGCSA achieved its best performance on SH, with all 37 samples correctly classified (100%). Performance on SM was slightly weaker, with 35 correct and 5 misclassified samples. For SL, overall accuracy was high, although 7 instances were misclassified as SH, suggesting partial feature overlap between low and high attention states. Together with the learning curves, these results indicate that DGCSA provides strong discriminability and generalization for attention recognition in virtual learning scenarios.

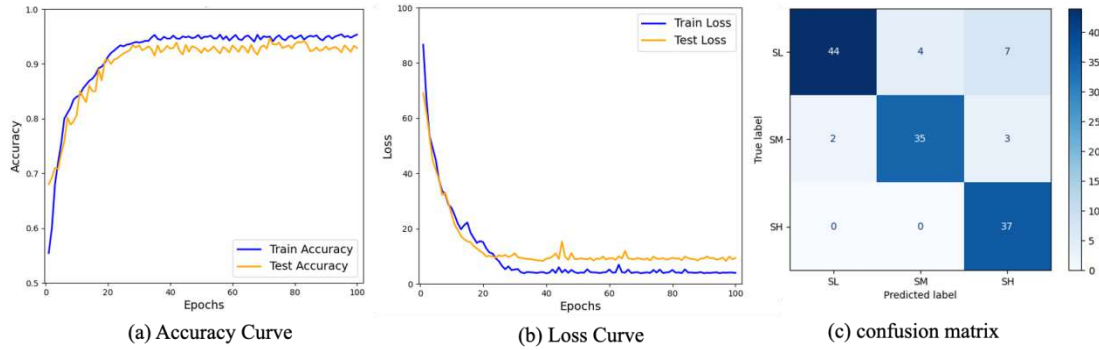


Fig. 5 Training and validation curves and test-set confusion matrix for the DGCSA model: (a) accuracy vs. epoch, (b) cross-entropy loss vs. epoch, and (c) confusion matrix on the held-out test set.

4.2.2 Model Performance and Module Contribution Analysis

To rigorously assess the advantages of DGCSA for EEG-based attention recognition and quantify the contribution of its key modules, we conducted two complementary experiments. First, in ablation studies, DGCSA was decomposed to isolate the effects of dynamic graph convolution DGCNN and SA modules. Second, in comparative evaluations, DGCSA was benchmarked against three representative MLP, CNN, and LSTM deep models analyze recognition accuracy and

robustness. All models were trained with the same sub-band-weighted fusion features, identical data partitions, and matched hyperparameters. Performance was reported using accuracy, precision, recall, and F1-score.

In the ablation studies, DGCNN alone achieved 87.9% accuracy, 3.8 percentage points below the full DGCSA, highlighting its central role in dynamically modelling inter-channel topology. The self-attention (SA) module provided additional gains by capturing cross-channel dependencies, but its impact was limited without a graph structure. A static-topology GCNN reached 84.8%, which is 6.9 % less than DGCSA, indicating that fixed adjacency is insufficient to represent individualised EEG connectivity patterns. Taken together, these results show that coupling dynamic graph modelling with self-attention enables DGCSA to encode both spatial topology and channel salience, substantially enhancing discriminative capacity.

In comparative evaluations, DGCSA consistently outperformed on the MLP (85.61%), CNN (84.8%), and LSTM (87.88%), baseline models across accuracy, precision, recall, and F1-score, with the largest gains in accuracy and F1. Fig. 6 presents a radar-chart comparison of the four performance metrics. The CNN underperformed due to limited capacity to model higher-order spatial dependencies; the LSTM leveraged temporal cues but insufficiently captured cross-channel structure; and the MLP, lacking explicit structural modeling, performed worst. By contrast, DGCSA's synergistic combination of dynamic graph modeling and self-attention yielded markedly higher overall recognition performance and demonstrated stronger generalization and robustness on complex EEG data.

In summary, DGCSA not only delivers substantial gains in attention recognition under virtual learning conditions but also offers a principled, structured and interpretable EEG modeling framework by jointly leveraging dynamic graph topology and self-attention within a multi-module fusion architecture.

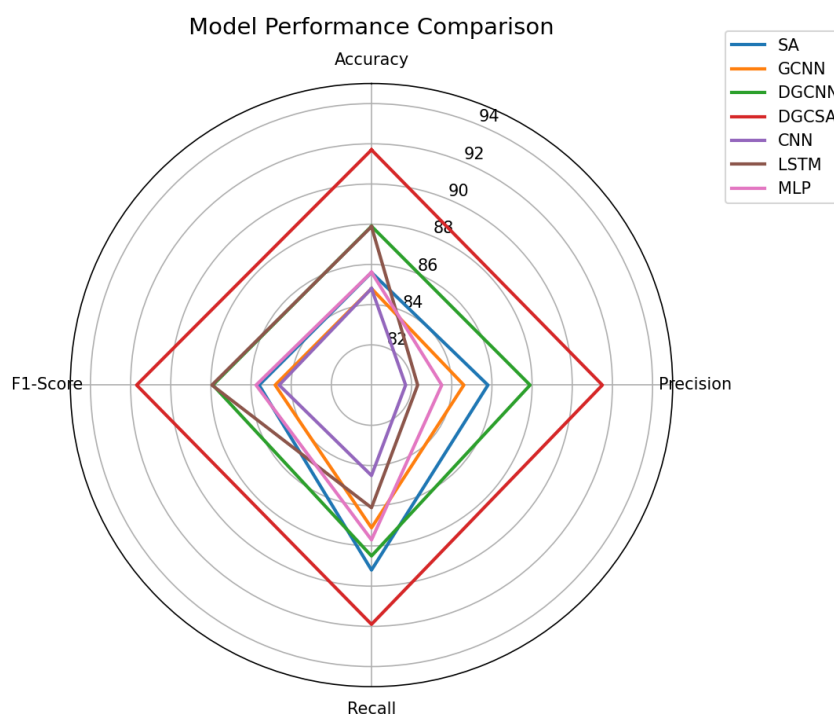


Fig. 6. The chart comparison of different models across four performance metrics across

different models.

4.3 Cross-Scenario Analysis of Attention Differences

To comprehensively compare attention between VR and 2D learning scenarios, we examined not only subjective attention ratings and global EEG measures but also the topographical distribution of band-limited activity. Integrating behavioral outcomes with frequency-specific EEG maps enabled a more systematic characterization of scenario-dependent attention differences and clarified the neurophysiological mechanisms underlying these variations.

4.3.1 Attention Score Statistics

According to the experimental setup in Section 3.1, all participants completed the same task in both 2D and VR conditions, with 90 s of EEG recorded in each scenario. For scoring, the 90-s sequence was partitioned into nine 10-s scoring segments, each composed of ten non-overlapping 1-s windows. From each 10-s segment, EEG features were extracted using the sub-band-weighted fusion method (Section 3), and the trained DGCSA classifier produced a three-class attention label, low (L), medium (M), or high (H) mapped to scores of 10, 20, and 30, respectively.

Based on this procedure, mean and standard deviation of attention scores over the 90 s were computed for each participant. As shown in Fig. 7, most participants achieved higher mean attention with smaller variability in the VR condition than in 2D, indicating that immersion not only elevates overall attention but also stabilizes attentional performance. These results provide initial evidence that VR confers an advantage for maintaining high and stable attention during learning.

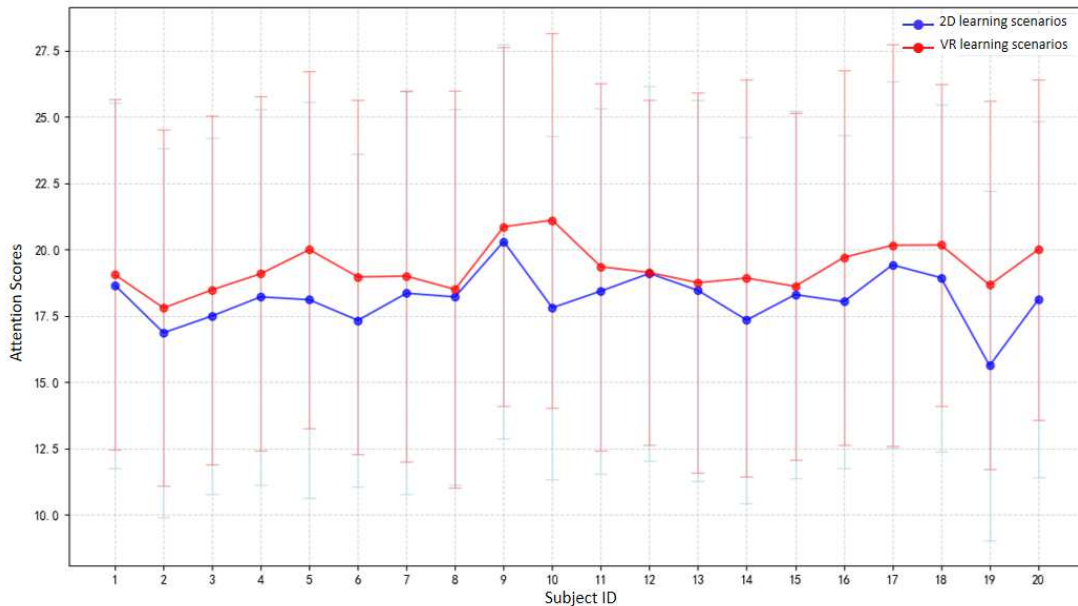


Fig. 7 Mean and standard deviation of attention scores in 2D and VR learning scenarios

4.3.2 Paired t-Test and Statistical Significance Analysis

To further validate the statistical reliability of the observations, paired-sample analyses were conducted from two perspectives, including correlation analysis and significance testing.

Firstly, Pearson correlation analysis was performed to assess the synchronous trend in attention means under the two learning conditions. As shown in Table 2, the correlation coefficient between the attention means of the virtual and 2D learning scenarios was 0.5335, with a significance level of $p=0.0154$ (< 0.05). This indicates a statistically significant positive correlation between the two conditions, suggesting that participants' attention trends were consistent across scenarios.

Table 2. Paired sample correlation

	N	Correlation	Sig. (2-tailed)
Mean Attention Scores in VR and 2D Learning Scenarios	20	0.5335	1.5424e-2

Subsequently, paired-sample t-tests were performed to evaluate the significance of differences in mean attention scores for the 20 participants between the two scenarios. As shown in Table 3, the mean score in the VR scenario was 19.32 ± 0.82 , significantly higher than 18.16 ± 0.95 in the 2D scenario, with a mean difference of 1.16. The paired-sample t-test yielded $t = 5.84$, $p = 1.27 \times 10^{-5}$, and a 95% confidence interval for the difference of [0.78, 1.54], which does not include 0. These results further confirm that the VR scenario has a significant effect on enhancing attention levels.

Table 3. Statistical comparison of attention scores between 2D and virtual learning scenarios using a paired-sample t-test

Condition	Mean $\pm$ SD	Mean Diff	95% CI of Diff	t	p-value
Virtual	19.32 ± 0.82	—	—	—	—
2D	18.16 ± 0.95	—	—	—	—
Virtual - 2D	—	1.16	[0.78, 1.54]	5.84	1.27×10^{-5}

In conclusion, the virtual learning environment demonstrated a statistically significant advantage in improving average attention levels and maintaining stability. Moreover, the strong correlation with 2D performance indicates consistent behavioral trends. These results provide empirical support for the effectiveness of virtual environments in cognitive regulation and EEG-guided attention enhancement.

4.3.3 EEG Frequency Band Differences

Beyond behavioral performance and overall attention level differences, the study further explored the neural mechanisms underlying attention variations between VR and 2D learning scenarios by comparing the topographical distributions of EEG frequency bands. Considering prior evidence indicating that Theta, Alpha, and Beta rhythms are closely related to attention processes, the following focuses on the results of these three bands. Fig. 8 shows the Topographical graph on VR-2D, VR and 2D differences in Bayesian-weighted fused EEG features across Delta (0.5–4 Hz),

Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–50 Hz) using EEGLab.

Previous studies have indicated that frontal Theta activity is closely related to attentional state and arousal level. Specifically, when individuals are in an externally oriented attentional state with heightened focus and alertness, frontal Theta power tends to exhibit a relative decrease [24, 25]. In the present study, compared with the 2D condition, VR elicited a significant decrease of Theta in the fronto-central regions, whereas occipital Theta was notably enhanced Fig. 8(a). These findings suggest that VR environments may help learners reduce “endogenous attention” or mind-wandering, leading to greater focus on external stimuli. Meanwhile, the increase in occipital Theta may reflect enhanced visual-spatial processing and the ongoing integration of spatial information during immersive virtual exploration.

Alpha rhythms are widely regarded as indicators of inattentive states and relaxation, and decreases in occipital Alpha, known as Alpha suppression, are classic markers of visual attention [26, 27]. In this study, VR elicited a pronounced reduction of occipital Alpha relative to the 2D condition, demonstrating typical Alpha suppression and suggesting stronger visual cortical activation and external attentional engagement in the immersive environment. Additionally, the increase of Alpha in left temporal and parietal regions may reflect the higher demands for semantic processing and information integration under VR conditions. Overall, the results align closely with classical models and further indicate that immersive environments can substantially enhance visual attention.

Beta rhythms are closely associated with cortical excitability, active cognitive processing, and sensorimotor integration [25, 28]. In this study, VR elicited significant increases of Beta activity in parietal and occipital regions, while frontal Beta was relatively reduced. This pattern suggests that learners in virtual environments face higher demands for multimodal information integration, particularly in parieto-occipital regions where visual, spatial, and proprioceptive inputs are simultaneously processed. Meanwhile, the relative decrease of frontal Beta may reflect a shift of attentional resources toward external sensory signals rather than internal abstract reasoning or planning. The results provide a more spatially detailed complement to existing research on Beta rhythms and attention concentration.

Integrating subjective evaluations, statistical analyses, and frequency-band topographical results, the study demonstrates that VR environments are more effective than 2D environments in maintaining learners' attention. On one hand, attention scores in VR were generally higher and showed significant differences in statistical tests. On the other hand, EEG band analysis revealed decreases in Theta, suppression of Alpha, and enhancement of Beta under VR conditions, collectively reflecting increased external attention, visual cortical activation, and multimodal information integration. These converging results from behavioral and neurophysiological levels indicate that immersive virtual environments provide a significant advantage in promoting sustained attention.

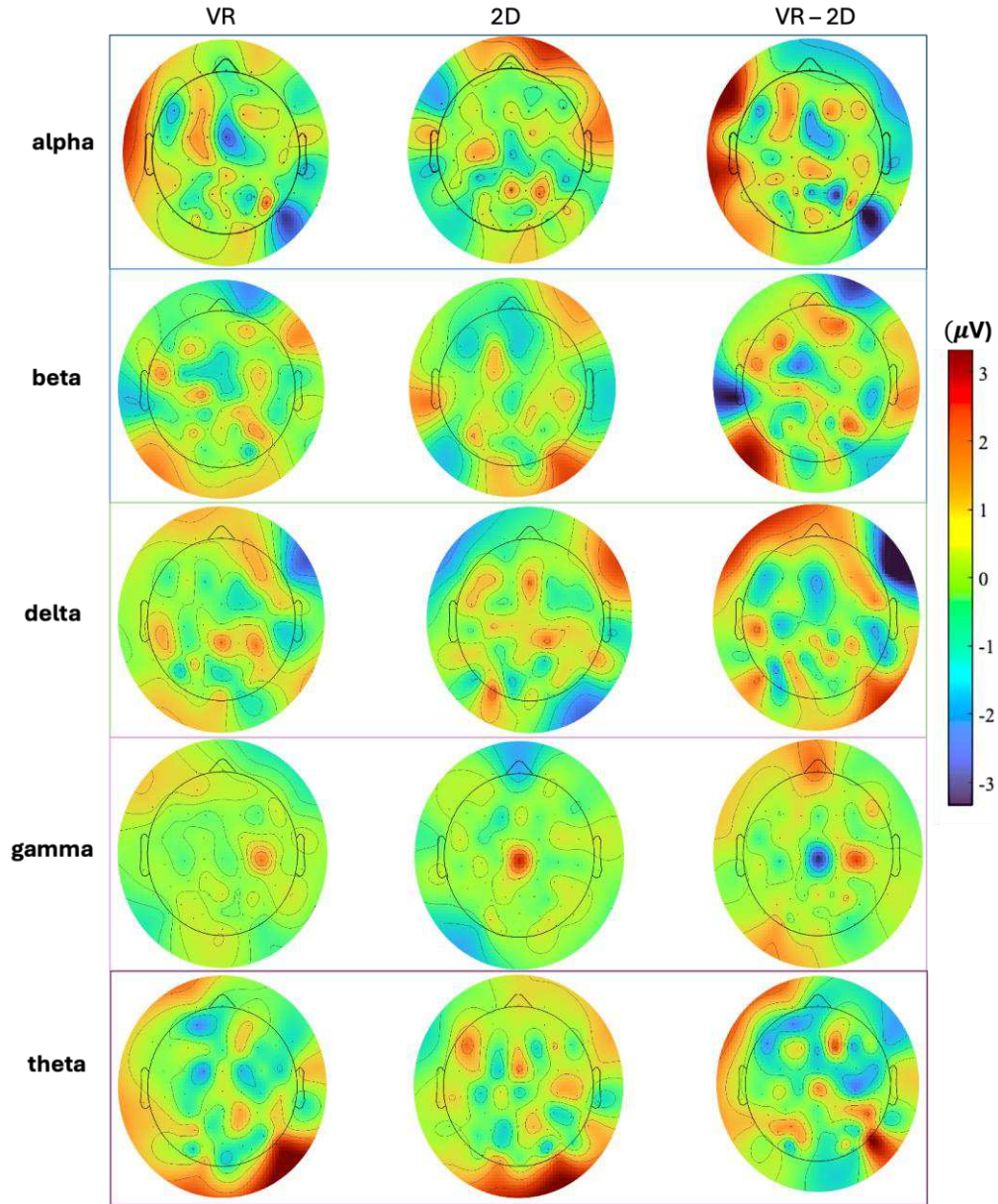


Fig. 8. Topographical graph on VR-2D, VR and 2D differences in Bayesian-weighted fused EEG features across Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–50 Hz) using EEGLab. Warm colors indicate higher values in VR; cool colors indicate higher values in 2D.

5 Discussion

This study introduces DGCSA, a dynamic graph convolutional network with self-attention mechanisms, for EEG-based attention recognition in immersive learning environments. The research evaluates differences in attention between VR and 2D learning scenarios. Using a

standardised Schulte Grid paradigm, we assembled a three-tiered attention dataset (SL/SM/SH) comprising 873 EEG segments, each lasting 3 seconds, recorded across 64 channels. The DGCSA model demonstrated robust three-class discriminability on the reserved test set, achieving optimal performance on high attention (SH), with all 37 SH samples accurately classified, yielding a 100% accuracy. Furthermore, cross-scenario analysis indicates that VR is associated with higher mean attention levels and lower variability than 2D learning, suggesting that immersive environments may enhance and stabilise learners' attention.

The primary motivation of DGCSA is that attention is not solely indicated by local spectral alterations but also by temporal variations in coordination across distributed neural networks, which manifest as dynamic dependencies among EEG channels. A static adjacency matrix enforces a fixed connectivity prior that may not accurately represent transient, task-dependent coupling patterns. DGCSA mitigates this limitation by learning a dynamic inter-channel topology, while the self-attention (SA) module captures global dependencies and emphasises channel-wise salient information. Empirical evidence from module contribution analysis supports this modelling approach: a static-topology GCNN demonstrated inferior performance (84.8%), whereas DGCSA attained higher accuracy; further ablation studies indicate that dynamic graph modeling is a crucial factor for the observed improvement (DGCNN alone 87.9%, 3.8 percentage points below the full DGCSA). Collectively, these findings suggest that integrating dynamic topology learning with self-attention enhances the representation of both spatial structure and global channel interactions relevant to attentional-state decoding.

Although the comprehensive confusion matrix demonstrates robust discrimination across attention levels, a prominent error pattern involves the misclassification of certain low-attention samples (SL) as high-attention (SH). This pattern likely reflects partial feature overlap between segments labelled “low” and brief intervals characterised by transient arousal or alertness. It may also originate from a labeling asymmetry: in the current protocol, SL corresponds to an eyes-closed resting state with no clicks, whereas SM/SH are defined using task completion time and accuracy thresholds. This discrepancy could introduce distributional differences between SL and task-engaged states, thereby complicating the decision boundaries for short windows. Practically, this analysis suggests several avenues for improvement. Firstly, incorporating sequence-level smoothing (e.g., aggregation across consecutive windows) to mitigate isolated fluctuations. Secondly, enriching the “low attention” definition with low-engagement periods during task segments, which would be more ecologically valid. Finally, evaluating ordinal or cost-sensitive objectives to reduce confusion between extreme classes.

The comparison between VR and 2D was conducted with a counterbalanced order and a one-day interval, incorporating 90 seconds of EEG recording in each condition. The analysis of attention scores reveals that the majority of participants exhibited higher average attention levels and lower variability in the VR environment than in the 2D environment. One plausible interpretation is that immersive interaction enhances sustained engagement and minimises attentional fluctuations during learning. However, it is worth considering a novelty effect in VR, in which initial exposure to

immersive environments temporarily elevates attention independently of deeper cognitive processes. The current experimental design- employing randomized scenario order and a one-day interval- serves to mitigate but does not entirely eliminate this effect. Consequently, longitudinal studies involving repeated VR sessions are required to determine whether the observed attention benefits persist after habituation.

Limitations and future work

Several limitations must be acknowledged. Firstly, the sample size and population diversity are restricted, which may constrain the generalizability to other age groups and learner populations, and may result in overfitting of the model. EEG-guided attention feedback should be evaluated within personalised intervention pipelines, such as children’s attention training or cognitive rehabilitation, to assess real-world feasibility and impact [29, 30]. Although segmentation produced 873 EEG samples, the relatively small number of participants remains a significant constraint for the generalisation of deep learning models. Secondly, the risk of overfitting remains noteworthy considering the dataset size and the utilisation of an 85:15 train–test split for DGCSA evaluation. Future studies should prioritize subject-independent protocols (e.g., leave-one-subject-out or nested cross-validation) and external validation on independent cohorts. Thirdly, EEG acquisition in VR introduces additional confounding factors, including motion-related artefacts and an increased sensory load. While preprocessing involved visual inspection, removal of gross-motion artefacts, and independent component analysis (ICA) with semi-automatic labeling combined with manual confirmation, residual VR-specific artefacts might still influence high-frequency features and the learned connectivity patterns. Fourthly, the current scoring pipeline is based on windowed analysis, and the VR/2D attention scoring depends on non-overlapping short windows aggregated into 90-second summaries. This methodology may overlook finer temporal dynamics and could be sensitive to brief fluctuations. Lastly, the interpretability of the learned dynamic topology remains partial, while ablation studies indicate the benefits of dynamic graphs over static adjacency. Further explainability analyses are required to correlate the learned connectivity changes with physiologically meaningful patterns.

Future studies should validate DGCSA across larger, more diverse cohorts and employ more robust, subject-independent evaluation methods. It is also recommended to incorporate longitudinal virtual reality exposure to measure habituation and isolate effects of novelty. Furthermore, integrating motion and physiological data channels, coupled with stricter artefact quantification, will enhance robustness in virtual reality applications. Expanding interpretability through GNN-specific explainers and connectivity visualizations will further elucidate how dynamic topology learning correlates with attention mechanisms. These measures will serve to strengthen confidence in the deployment of this technology for clinical and educational purposes.

6 Conclusion

This study proposes and validates an EEG-based attention recognition framework that combines a Bayesian-optimized, frequency-band–weighted fusion of features with DGCSA, a

model integrating dynamic graph convolution and self-attention. In a three-class task, DGCSA achieved 91.67% accuracy on a self-collected dataset, outperforming traditional machine-learning and deep-learning baselines. Ablation analysis confirms the complementary contributions of the dynamic graph and attention modules to capturing complex spatiotemporal dependencies. Cross-scenario experiments further show that immersive virtual environments yield significantly higher mean attention scores with lower variability ($p < 0.001$) frequency-band analyses corroborate these behavioral effects, indicating enhanced attentional stability under VR. Collectively, the results establish DGCSA as a reliable tool for assessing cognitive states across learning contexts and provide neurophysiological support for deploying virtual reality to sustain attention in educational settings.

Declarations

Ethics approval and consent to participate

The study received ethical approval from Xi'an Shiyou University (Approval No. 2023-KD-098) and was conducted in accordance with institutional guidelines. Informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Shaanxi Provincial Social Science Funding Project, “Research on Innovative Mechanisms and Paths for Classroom Teaching Evaluation Empowered by Education Digitalization” (Grant No. 2024P019); the Comprehensive Reform Project for Graduate Education at Xi'an Shiyou University (Grant No. 2025XYJG018); and the China Scholarship Council project (Project No. 202308610207).

Authors' contributions

J.Y. and K.X. contributed to writing the original draft, review and editing, visualization, validation, methodology, and data curation. W.Y. and S.K.K.C. contributed to conceptualization, methodology, validation, resources, supervision, project administration, funding acquisition, writing the original draft, and review and editing. All authors reviewed, edited, and approved the final version of the manuscript.

Data availability

The data are available from the corresponding author upon reasonable request, subject to privacy and ethical restrictions.

Use of AI-assisted language tools

This manuscript was polished using Grammarly AI for minor improvements in language and readability. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Clinical trial number

Not applicable.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and contributions to the peer review of this work.

References

1. Mo J, Liang H, Yuan H, Shou Z, Zhang H: **Learning attention characterization based on head pose sight estimation**. *Multimedia Tools and Applications* 2024, **83**(38):85917–85937.
2. Nirmala Devi A, Rathna R: **EEG data based human attention recognition using various machine learning techniques: a review**. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 2024, **11**(7):2299096.

3. Zhang R, Sui L, Shen C, Xu L, Cao J: **EEG-based real-time BCI system using drones for attention visualization.** *Computer Methods in Biomechanics and Biomedical Engineering*:1–9.
4. Peng B, Zhang Y, Wang M, Chen J, Gao D: **T-A-MFFNet: Multi-feature fusion network for EEG analysis and driving fatigue detection based on time domain network and attention network.** *Computational Biology and Chemistry* 2023, **104**:107863.
5. Matto HC, Ihara ES, Cieslowski B, Hines DA, Booth J: **Virtual reality case simulation to build skill competencies in working with substance-engaged clients.** *Social Work Education* 2024, **43**(9):2774–2786.
6. AkgÜN M, Atici B: **The Effects of Immersive Virtual Reality Environments on Students' Academic Achievement: A Meta-analytical and Meta-thematic Study.** *Participatory Educational Research* 2022, **9**:111–131.
7. Ocasio W: **Attention to Attention.** *Organization Science* 2011, **22**(5):1286–1296.
8. Posner MI, Rothbart MK: **Research on attention networks as a model for the integration of psychological science.** *Annu Rev Psychol* 2007, **58**:1–23.
9. Zhang W, Xie Y, Yang T: **Reward salience but not spatial attention dominates the value representation in the orbitofrontal cortex.** *Nature Communications* 2022, **13**(1):6306.
10. miao R, Kato H, Hatori Y, Sato Y, Shioiri S: **Use of facial expressions to estimate level of attention while watching video lectures.** *Journal of Vision* 2023, **23**(9):4726–4726.
11. Tao H, Duan Q: **Hierarchical attention network with progressive feature fusion for facial expression recognition.** *Neural Networks* 2024, **170**:337–348.
12. Casagrande WD, Nakamura-Palacios EM, Frizera-Neto A: **Electroencephalography Neurofeedback Training with Focus on the State of Attention: An Investigation Using Source Localization and Effective Connectivity.** *Sensors* 2024, **24**(18):6056.
13. Pakadang SS, Wibawa AD, Wulandari DP: **EEG Patterns on Learning Concentration Using Short Learning Videos Based on Time-Frequency Domain.** In: *2024 International Seminar on Intelligent Technology and Its Applications (ISITIA): 10–12 July 2024* 2024; 2024: 460–465.
14. Gao Q, Tan Y: **Impact of Different Styles of Online Course Videos on Students' Attention During the COVID-19 Pandemic.** *Frontiers in Public Health* 2022, **Volume 10 - 2022**.
15. Zuo M, Chen X, Sui L: **Evaluation of Machine Learning Algorithms for Classification of Visual Stimulation-Induced EEG Signals in 2D and 3D VR Videos.** *Brain Sciences* 2025, **15**(1):75.
16. Pathak D, Kashyap R: **Neural Correlate-Based E-Learning Validation and Classification Using Convolutional and Long Short-Term Memory Networks.** *Traitement du Signal* 2023, **40**:1457–1467.

17. Rivas F, Sierra-Garcia JE, Camara JM: **Comparison of LSTM- and GRU-Type RNN Networks for Attention and Meditation Prediction on Raw EEG Data from Low-Cost Headsets.** *Electronics* 2025, **14**(4):707.
18. Xu B, Zhang X, Zhang X, Sun B, Wang Y: **An improved graph convolutional neural network for EEG emotion recognition.** *Neural Computing and Applications* 2024, **36**(36):23049–23060.
19. Raikwar S, Mayuri AVR: **Self-attention-based 1DCNN model for multiclass EEG emotion classification.** *The Journal of Supercomputing* 2025, **81**(4):520.
20. Li J, Pan W, Huang H, Pan J, Wang F: **STGATE: Spatial-temporal graph attention network with a transformer encoder for EEG-based emotion recognition.** *Frontiers in Human Neuroscience* 2023, **Volume 17 - 2023**.
21. Li Y, Yang Y, Zheng Q, Liu Y, Wang H, Song S, Zhao P: **Dynamical graph neural network with attention mechanism for epilepsy detection using single channel EEG.** *Medical & Biological Engineering & Computing* 2024, **62**(1):307–326.
22. Gu R, Zhang J, Zhou J, Tong M: **The Baroque music's influence on learning efficiency based on the research of eye movement.** In: *2014 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE): 8–10 Dec. 2014* 2014; 2014: 511–515.
23. Pai W, Fan W, Mei W, Xuebin Q: **Attention grading modeling based on Schulte grid and LSTM.** *Computer Engineering and Applications* 2022, **58**(15):133–140.
24. Jensen O, Tesche CD: **Frontal theta activity in humans increases with memory load in a working memory task.** *Eur J Neurosci* 2002, **15**(8):1395–1399.
25. Clarke AR, Barry RJ, Johnstone S: **Resting state EEG power research in Attention-Deficit/Hyperactivity Disorder: A review update.** *Clinical Neurophysiology* 2020, **131**(7):1463–1479.
26. Foxe JJ, Snyder AC: **The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention.** *Frontiers in Psychology* 2011, **volume 2 - 2011**.
27. Woodman GF, Wang S, Sutterer DW, Reinhart RMG, Fukuda K: **Alpha suppression indexes a spotlight of visual-spatial attention that can shine on both perceptual and memory representations.** *Psychonomic Bulletin & Review* 2022, **29**(3):681–698.
28. Spitzer B, Haegens S: **Beyond the Status Quo: A Role for Beta Oscillations in Endogenous Content (Re)Activation.** *eneuro* 2017, **4**(4):ENEURO.0170–0117.2017.
29. Guedj C, Tyrand R, Badier E, Planchamp L, Stringer M, Zimmermann MO, Férat V, Ha-Vinh Leuchter R, Grouiller F: **Self-Regulation of Attention in Children in a Virtual Classroom Environment: A Feasibility Study.** *Bioengineering (Basel)* 2023, **10**(12).

30. Corrigan N, Păsărelu C-R, Voinescu A: **Immersive virtual reality for improving cognitive deficits in children with ADHD: a systematic review and meta-analysis.** *Virtual Reality* 2023, **27**(4):3545–3564.