Depression Detection Using Temporal-frequency Spatial Transformer Enhanced with Cross-frequency Interaction Layer

Monika Puttaramaiah [1], Sneha S Bagalkot [2], and Amogh Pramod Kulkarni [3,*]

[1] Department of Artificial Intelligence and Machine Learning, B.M.S. College of Engineering, Bangalore, India
[2] Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore, India
[3] Department of Information Science and Engineering, Sai Vidya Institute of Technology, Bangalore, India
Email: monika.mel@bmsce.ac.in (M.P.); snehasb.cse@bmsce.ac.in (S.S.B.); amogh.pk@saividya.ac.in (A.P.K.)
*Corresponding author

***Abstract*—Depression is a major global mental health disorder that severely affects individual well-being and societal productivity. Electroencephalography (EEG) is a widely used non-invasive biomarker for capturing the direct electrical activity of the brain. Existing methods often rely on extensive manual feature engineering and fail to capture important spatio-temporal interactions and electrode topologies. To overcome this, the Cross-Frequency Interaction Temporal-Frequency Spatial Transformer (CFI-TFST) is introduced for depression detection. The proposed CFI-TFST method begins with Frequency-Aware Feature Representation that uses Discrete Cosine Transform (DCT) combined with a Frequency Attention (FA) mechanism to extract discriminative spectral information directly from EEG feature maps. By integrating the Cross-Frequency Interaction (CFI) layer, the proposed model captures nonlinear cross-frequency dynamics through cross-spectral feature representations. The Sub-Domain Knowledge Alignment (SDKA) module enhances these features to improve generalization across different data domains. Furthermore, the proposed CFI-TFST model uses a lightweight EEG-Conformer baseline to reduce the transformer block to a single-head attention layer to prevent overfitting while maintaining relatively low computational complexity. The experimental results indicate that the CFI-TFST method achieves epoch-level classification accuracy under subject-disjoint grouped cross-validation of 95.24% and 95.87% for the MODMA and PRED + CT datasets, respectively.**

***Keywords*—depression, discrete cosine transform, electroencephalography, frequency attention, cross-frequency interaction, temporal-frequency spatial transformer**

I. Introduction

Depression is recognized as a significant mental health disorder in modern society by continuously affecting the thought process and reduces involvement in daily activities [1]. According to the World Health Organization, an estimated 280 million individuals worldwide are affected by depression, highlighting its designation as an important global public health challenge [2]. Depression affects individual well-being and societal functioning through psychological symptoms, increased suicide risk and disease progression, leading to reduced productivity and increased disability [3]. Therefore, monitoring mental health and the early identification of depressive symptoms have become critical priorities in the present healthcare systems, and psychiatric evaluation using standardized screening instruments, such as the Patient Health Questionnaire, plays an important role in detection and intervention [4]. Electroencephalography (EEG) is a non-invasive neuroimaging modality recognized as a valuable digital biomarker of cognition. By capturing fine-grained fluctuations in brain activity, EEG provides insights into neural dynamics and has been widely applied in medical settings to monitor electrical activity under different conditions [5]. EEG enables objective detection of depression through physiological signal analysis, and its benefits include portability, low cost, ease of use and the absence of radiation, which makes it suitable for dynamic monitoring of depression [6]. The pathogenesis of depression is multifactorial, involving genetic susceptibility and psychosocial stressors. Beyond clinical conditions, depression reduces occupational performance, destabilizes family and social life and represents a risk factor for suicide [7]. Studies of cerebral functional and structural abnormalities associated with depression have extended current investigation by using EEG with different medical techniques [8]. Compared to traditional Machine Learning (ML) models, Deep Learning (DL) provides notable advantages for EEG data processing. By automatically extracting hierarchical features from raw input, the model captures nonlinear relationships and complex spatial-temporal structures that manual features overlook [9]. Accurate identification of depression requires monitoring symptoms over a defined period to avoid misclassification with related psychiatric conditions such as anxiety disorder [10]. In EEG analysis, coherence

Manuscript received May 11, 2026; revised June 30, 2026; accepted September 1, 2026; published November 25, 2026.

is widely recognized as a strong indicator of integrated neuronal activity by measuring the linear correlation between frequency components of signals recorded from different brain areas or electrodes [11]. Over the years, neuroscience has increasingly investigated depression through studies by focusing on brain mechanism, which increase the understanding of biological and neural pathways [12]. Various studies have explored EEG-based techniques for depression, in which DL methods applied in this domain are generally classified into Convolutional Neural Networks and Graph Neural Networks (GNN) [13]. These methods are specifically developed to model spatial structural relationships among EEG channels, enabling the extraction of cross-channel features that reflect integrated neural activity [14]. EEG-based depression detection is constrained by two primary issues, such as frequently bypassing extensive feature engineering, which depends on raw inputs. Another issue is that the spatial topology of electrodes involved in spatio-temporal interactions among brain regions is often neglected [15]. Early treatment during the initial stages of depression is essential to treat disease progression and reduce the risk of developing treatment-resistant forms [16]. Stress-related anxious states show beta activity and reduced alpha coherence, while depression is associated with abnormal theta and alpha rhythms [17]. Individuals experiencing depression frequently use social media platforms, where they express their thoughts, emotions and feelings through posts [18]. Depression is a multifaceted condition, but EEG-based computational techniques have been combined with behavioral insights from social media, enabling more accurate detection [19]. Together, these advances show the increasing applicability of integrating physiological signals to develop a depression detection model [20]. The existing Temporal-Frequency Spatial Transformer with Sub-Domain Knowledge Alignment (TFST-SDKA) model relies heavily on extensive feature representations derived from raw inputs. The architecture treats frequency bands as independent vectors, failing to model nonlinear interdependencies across frequency scales, and ignores the spatial topology. The main objective is to develop a depression detection model that uses a Cross-Frequency Interaction (CFI) layer to computationally model cross-frequency interactions and Sub-Domain Knowledge Alignment Module (SDKA) module to optimize feature alignment across heterogeneous EEG datasets.

The contributions of the proposed Cross-Frequency Interaction Temporal-Frequency Spatial Transformer (CFI-TFST) model are as follows:

- The proposed CFI-TFST model uses Discrete Cosine Transform (DCT) combined with a channel attention mechanism to directly extract spectral information from EEG feature maps instead of depending on manual bandpass filtering.
- It introduces a Cross-Frequency Interaction (CFI) layer to model nonlinear interdependencies among frequency bands, capturing complex cross-spectral feature representations.
- The Frequency Attention (FA) mechanism extends conventional channel attention, which divides the EEG input into multiple channel segments and applies DCT to each segment.

This manuscript is organized as follows: Related works are presented in Section II, the proposed methodology is explained in Section III, the experimental results and discussion are described in Section IV, and Section V concludes the paper.

II. Related Works

Zhang *et al.* [21] introduced a DL model for speech-based depression detection using 30-second audio samples. The proposed model extracted three feature sets: Mel spectrogram, Mel Frequency Cepstral Coefficients (MFCC), and short-term energy. The data were then processed using a hybrid Convolutional Neural Network (CNN) and Bi-directional Gated Recurrent Unit (Bi-GRU). The CNN-Bi-GRU model enhanced emotional support through privacy-focused screening and includes short-term energy features that capture signs of depression. However, the developed model struggled with long-form, semi-structured audio samples, where patterns varied unpredictably.

De *et al.* [22] introduced a hybrid model, SLiTRANet integrated with the Stockwell transform which was used to exploit spectro-temporal information and Linear Graph Convolutional Network to capture spatial dependencies within brain networks. The method employed the Stockwell transform for time-frequency analysis, which generated a 2D representation fed into the LGCN, followed by a transformer for classification. However, the introduced model did not include longitudinal data, which were needed to identify trends and trajectories of depressive symptoms over time.

Das and Naskar [23] demonstrated the Spectro_CNN model for depression detection, which used a multimodal method combining MFCC with features extracted from audio spectrograms through CNN. The combined features were then fed into a binary classifier for depression or non-depression detection. The Spectro_CNN effectively handles the issue of class imbalance by combining multiple short audio samples from an individual into a single file and using noise reduction techniques. However, MFCC features alone did not capture complex spatial data patterns, leading to reduced performance, particularly in noisy environments.

Xu *et al.* [24] employed the TFST-SDKA model for depression detection, which used FA with DCT to convert EEG feature maps into the frequency domain, capturing multi-frequency information associated with depression. This process enhanced the model's representation capability and expands the feature space. The SDKA module classifies subjects into different subdomains based on the labels. However, TFST-SDKA used FA, which relied on predefined frequency band features and struggled to adapt to different variations.

Sam *et al.* [25] utilized a Spiking Neural Network (SNN) and Long Short Term Memory (LSTM) structure to classify specific depression levels using raw EEG signals. The process involved two main steps: unsupervised learning within a 3D brain template-structured SNN

reservoir using the spike time-dependent plasticity learning rule to model brain activities. The utilized model was biologically plausible, as the SNN model mimics internal brain activities and provides a visual interpretation of structural alterations linked to depression. However, the interpretation capability of the SNN-LSTM models lacked in-depth analysis of specific brain mechanism and require enhancement of SNN hyperparameters.

Esmi *et al.* [26] developed a synchronized multimodal transformer for depression detection that combined EEG signals and clinical interview data. The method starts with signal preparation using Flexible Temporal Sequence Matching (FTSM) to prioritize the signal channels and align asynchronous audio-EEG data. This was followed by Signal Conversion, which transformed the EEG into a 2D representation and audio into spectrograms and text tokens. A multi-head cross-attention mechanism was fused at the token level, and a self-attention block captured long-term dependencies before classification. However, the employed model faced difficulties in generalizability across different populations, susceptibility to noise, and ethical risks regarding data confidentiality.

Xi *et al.* [27] introduced a framework that combined temporal, spatial, and frequency domain features of EEG to improve depression recognition. The introduced method starts with an initial EEG channel selection phase using frequency-domain weighting to automatically adjust the channel weights and select the most relevant channels across five frequency bands. A Multiscale Spatiotemporal Convolutional Attention Network was used to extract features, where the Multiscale Convolutional Attention module captured spatial features. Temporal Trend Aware Self-Attention extracted long-term temporal features by examining these correlations. However, the proposed model lacks generalizability across different populations.

Wang *et al.* [28] developed a model for depression using sleep data acquired by wearable devices through ensemble models. The process begins with the extraction of nine sleep features from activity data. This feature set undergoes preprocessing, including handling missing values using an improved Generative Adversarial Interpolation Network. Finally, the bagging ensemble learning method integrates multiple weak classifiers to produce the final depression classification. However, the developed ensemble model struggled with generalizability and was not extensively tested in non-clinical settings.

III. Proposed Method

The proposed CFI-TFST model is designed for objective depression detection using EEG signals. It addresses the limitation of manual feature engineering to capture intricate and nonlinear relationships within brain activity. The CFI-TFST model uses a Frequency-Aware Feature Representation based on the DCT and FA mechanisms, which extract discriminative spectral information directly from EEG feature maps. To capture the neural dynamics across different rhythms, the CFI layer models the computational interdependencies between frequency bands. The Sub-Domain Knowledge Alignment (SDKA) module uses Sub-Domain Maximum Mean Discrepancy (SMMD) to improve the model's generalization capability. Fig. 1 shows the overall architecture of the CFI-TFST model.

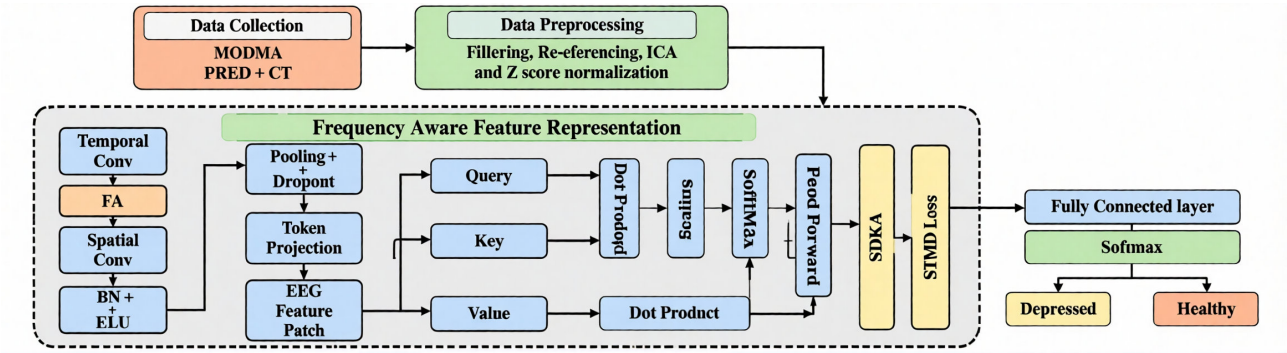

Fig. 1. Overall architecture of CFI-TFST for EEG-based depression detection.

A. Dataset Description

The MODMA dataset [29] contains resting-state EEG recordings from 53 participants, including 24 individuals diagnosed with depression and 29 healthy controls. It uses a 128-channel HydroCel Geodesic Sensor Net with a sampling rate of 250 Hz and Cz electrodes as the reference point. For analysis, the continuous EEG recordings are divided into non-overlapping dataset-specific epochs specifically 2-second windows for the MODMA dataset and 5-second windows for the PRED+CT dataset. This increases the diversity of the training set without introducing redundant temporal patterns while ensuring reliable evaluation and prevent subject-level data leakage, the classification is evaluated using a subject-level split integrated within a 5-fold Grouped K-Fold cross-validation protocol, ensuring that all epochs belonging to a single participant remain in the same fold across both training and testing phases. To maintain structural equality across evaluation folds, participants are assigned using subject IDs into five discrete blocks. Folds 1 through 4 consists of 11 unique subjects with 5 depressed individuals and 6 healthy controls, whereas Fold 5 consists of remaining 9 unique subjects with 4 depressed individuals and 5 healthy controls.

The PRED + CT dataset represents a larger group of 121 participants consisting of 46 depression subjects and 75 healthy controls [30]. The recordings are captured using a

67-channel device at a higher sampling rate of 500 Hz. This dataset used a 5s window size to leverage the higher resolution for identifying lower frequency rhythms. A subject-level partition is maintained within the Grouped 5-fold cross-validation loops for binary classification, ensuring that no overlapping participant data occurs between the training and testing sets within the cross-validation loops. Subjects are dynamically allocated into the 5 folds by their distinct subject IDs according to fold-level distributions. Fold 1 contains 25 unique subjects (10 depressed, 15 healthy controls), whereas Folds 2 through 5 each consists of 24 unique subjects (9 depressed, 15 healthy controls) to accommodate the total of 121 participants.

B. Data Preprocessing

To preprocess EEG signals from the MODMA and PRED + CT datasets, eliminating non-neural noise using a bandpass filter (0.5–50 Hz) is required for N^{th} order Butterworth filter, as shown in Eq. (1), to isolate frequencies between 0.5 Hz and 50 Hz, where $\omega_O = \sqrt{\omega_L - \omega_H}$, $BW = \omega_H - \omega_L$ and N is the filter order. This range is important as it preserves all key clinical digital measures, such as Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–50 Hz). It reflects the neural signatures of depression, such as Frontal Alpha Asymmetry. By setting these bounds, high-frequency (>50 Hz) noise is removed, and power line interference and low-frequency artifacts (<0.5 Hz) caused by sweat, skin resistance, and physical movement can bias the model. The raw EEG signals are segmented into uniform epoch dimensions before entering the feature extraction. Specifically, for the MODMA dataset, the input tensor shape entering the model is defined as [B, 128,500], representing [Batch Size, Channels, Time Steps]. For the PRED+CT dataset, the input tensor shape is precisely [B, 62, 2500]. The 62 channels are actually fed into neural network because the bad electrodes are removed during preprocessing from the recorded 67 channels. These fixed tensor topologies ensure consistent batch matrix operations across all downsample and filtering layers. Filtering is followed by referencing to the Common Average Reference (CAR) where the mean activity of all channels is subtracted from each individual electrode using Eq. (2).

$$|H(j\omega)| = \frac{1}{\sqrt{1+\left(\frac{\omega^2-\omega_O^2}{\omega \cdot BW}\right)^{2N}}} \tag{1}$$

$$x_i^{avg}(t) = x_i(t) - \frac{1}{n}\sum_{j=1}^{n} x_j(t) \tag{2}$$

Here, $x_i^{avg}(t)$ is the cleaned potential at the electrode i at time t after re-referencing, $x_i(t)$ is the raw potential originally recorded at electrode i, n is the total number of channels, and $\frac{1}{n}\sum_{j=1}^{n} x_j(t)$ is the instantaneous average of all scalp electrodes representing common-mode noise. To handle biological artifacts such as eye blinks, Independent Component Analysis (ICA) is applied to decompose the observed signals X into independent sources S through linear transformation, as shown in Eq. (3). The artifact components are then identified and removed before reconstructing the cleaned signals. The independent components are identified as artifacts based on three specific criteria: topographical scalp maps, power spectral density, and temporal patterns. The components exhibiting these non-neural signatures are removed before signal reconstruction. Finally, the segmented dataset-specific epochs undergo amplitude normalization through a Z-score transformation as shown in Eq. (4). Here, z is the standard amplitude, x is the raw EEG sample value, μ is the mean amplitude, σ is the standard deviation. This process ensures that variations in scalp impedance between the MODMA and PRED + CT groups did not affect feature extraction.

$$S = WX \tag{3}$$

$$z = \frac{(x-\mu)}{\sigma} \tag{4}$$

C. Frequency-Aware Feature Representation

The end-to-end method for frequency-based feature extraction is described in this section. Unlike traditional methods that depend on manual bandpass filtering ($\alpha, \beta, \theta, \delta, \gamma$), the proposed CFI-TFST model directly extracts discriminative spectral information from EEG feature maps. The proposed CFI-TFST model is built on the Discrete Cosine Transform (DCT) combined with a standard channel attention mechanism, as expressed in Eq. (5). Here, $P_{m,n}^{p,q}$ is the basis for the transformation, m and n are the frequency components, p and q identify the specific spatial locations within the feature map, and M and N are the total dimensions of the map.

$$P_{m,n}^{p,q} = \cos\left(\frac{\pi m}{M}\left(p+\frac{1}{2}\right)\right) cos\left(\frac{\pi n}{N}\left(q+\frac{1}{2}\right)\right) \tag{5}$$

To transform the input EEG feature map into the frequency domain, the DCT is calculated using Eq. (6), as follows: This process decomposes the spatial EEG feature map into a set of orthogonal frequency components that represent the signal energy distribution across different spectral bases.

$$DCT_{m,n} = \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} DCT_{m,n} P_{m,n}^{p,q} \tag{6}$$

To identify depression features within a CNN, a channel attention mechanism is used, which compresses the information in each channel into a single scalar value to determine its importance. If $T \in R^{D\times M\times N}$ represents the EEG feature maps across D channels, the weighting is calculated as in Eq. (7). Here, $W_{gt} \in R^D$ is the resulting weight vector, $FF(\cdot)$ is the fully connected layer, where D represents the total number of EEG channels, and $Reduce(\cdot)$ is a function that collapses the spatial dimension $M \times N$ into a single value per channel. Once the weights are determined, the original maps are applied, as shown in Eq. (8), where $\widehat{T_k}$, ... is the final scaled output,

and each specific channel k is multiplied by its corresponding weight W_{gt_k}.

$$W_{gt} = Sigmoid\ \Big(FF\big(Reduce(T)\big)\Big) \tag{7}$$

D. Frequency Attention (FA)

The FA mechanism is designed to expand the scope of standard channel attention. The vector A learned selection mechanism was employed to compress computational overhead and filter out non-spectral noise. A 2D Discrete Cosine Transform (2D-DCT) is executed across the spatial-temporal dimensions of the intermediate feature maps to convert channel-time tensors into spatial-frequency representation. This results in 2D spectral output and it is concatenated to form a unified spectral descriptor, which is then passed through a fully connected mapping layer and sigmoid activation function. Subsequently, the spatial-frequency components are automatically sorted in descending order according to their calculated weight values. The top-K ($K = 5$) components with the highest attention coefficients are retained for downstream processing. By incorporating a broader frequency range, the limitations of Global Average Pooling (GAP) are mitigated. The lowest frequency component in Eq. (8) is significantly simplified, where $m = 0$ and $n = 0$.

$$DCT_{m,n} = \sum_{p=0}^{M-1}\sum_{n=0}^{N-1} t_{p,q} cos\left(\frac{0}{M}(p+\frac{1}{2})\right) \times \cos\left(\frac{0}{N}(q+\frac{1}{2})\right) = \sum_{p=0}^{M-1}\sum_{n=0}^{N-1} t_{p,q} = GAP(t)MN \tag{8}$$

In this specific state, $DCT_{0,0}$ is mathematically proportional to the GAP, which proves that the standard attention mechanism extracts the lowest-frequency data. For complex neurological conditions that depend on this minimal frequency range, accurate identification is insufficient. To capture additional spectral profiles, the input T is divided into r different channel segments, such as $[t^0, t^1, \dots, t^{r-1}]$. Here, each segment t^l exists in $R^{D^l \times M \times N}$ where D^l is calculated as D/r. Each segment is then transformed into its own DCT spectrum according to Eq. (9), where l ranges from 0 to $r-1$. The variables $[a_l, b_l]$ are the specific vectors $Spec^l \in r^{D^l}$.

$$Spec^l = DCT_{a_l,b_l}(t^l) = \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} t_{,l,p,q}^{\ l}\ P_{m,n}^{p,q} \tag{9}$$

The individual spectral components are then combined to create combined representation shown in Eq. (10). The resulting vector $Spec \in R^D$ serves as a multi-spectral weighting tool that represents the different frequency information across all channels.

$$Spec = Concat([Spec^0, Spec^1, \dots, Spec^{r-1}]) \tag{10}$$

To finalize the attention weights, the combined vector undergoes a nonlinear transformation in Eq. (11), which produces the final FA weights that determine which features are highlighted.

$$FA = Sigmoid\big(FF(Spec)\big) \tag{11}$$

To maximize efficiency, a two-stage selection process is used, which evaluates the contribution of different frequencies and selects the top-K informative frequency components. This ensures that selected frequency components are fused into the proposed model.

E. Cross-Frequency Interaction (CFI) Layer

Standard FA treats different frequency bands as independent and fails to account for the relationships between them; thus, the CFI layer is introduced. The selected bands are passed to the CFI module to capture the non-linear cross-frequency relationships. In depression, neural processes involve complex connections between bands. The CFI layer models these nonlinear interdependencies, allowing the system to capture combined spectral features. By calculating the pairwise interactions between frequencies, the CFI enhances the FA output with cross-spectral information. Let $B \in R^{D \times U \times V}$ be the output of the FA module. Here, the dimension V denotes the number of selected target frequency components, D denotes the channel feature embedding dimension and U is the localized segment length. This models cross-frequency interaction features, justifying the CFI layer setup. It is important to note that this interaction tensor functions as a mathematical feature-map approximation optimized for classification, rather than signal-processing metric. The CFI layer starts by constructing the band interaction tensor G using the outer products across different frequencies, which is defined in Eq. (12). Fig. 2 shows the structure of the FA enhanced with the CFI layer.

$$G_{i,j} = \sum_{k=1}^{V} B_{:,:,i} \odot B_{:,:,j} \cdot \sigma(W_g(B_{:,:,i} - B_{:,:,j})) \tag{12}$$

Here, $\odot$ denotes element-wise multiplication, σ is a non-linear activation function, such as ReLU. The term $W_g \in R^{U \times U}$ is a learnable weight matrix designed to weight the differences between frequency components. Indices i and j range across the frequency dimensions from 1 to V allowing the model to compare every pair of frequencies. Once the interaction tensor G is established, it is improved using a two-dimensional convolutional operation. This helps the model capture localized patterns of interplay, as expressed in Eq. (13), where K_c is the kernel used for the 2D convolution, and the layer normalization ensures stability throughout the training process.

$$H = Conv2D(G, K_c) + LayerNorm\ (G) \tag{13}$$

The final step involved combining the CFI with original data. The enhanced output $\hat{B}$ is calculated as shown in Eq. (14), where V is the frequency dimension and H is the refined interaction map.

$$\hat{B} = B + Softmax\left(\frac{H}{\sqrt{V}}\right) \cdot B \tag{14}$$

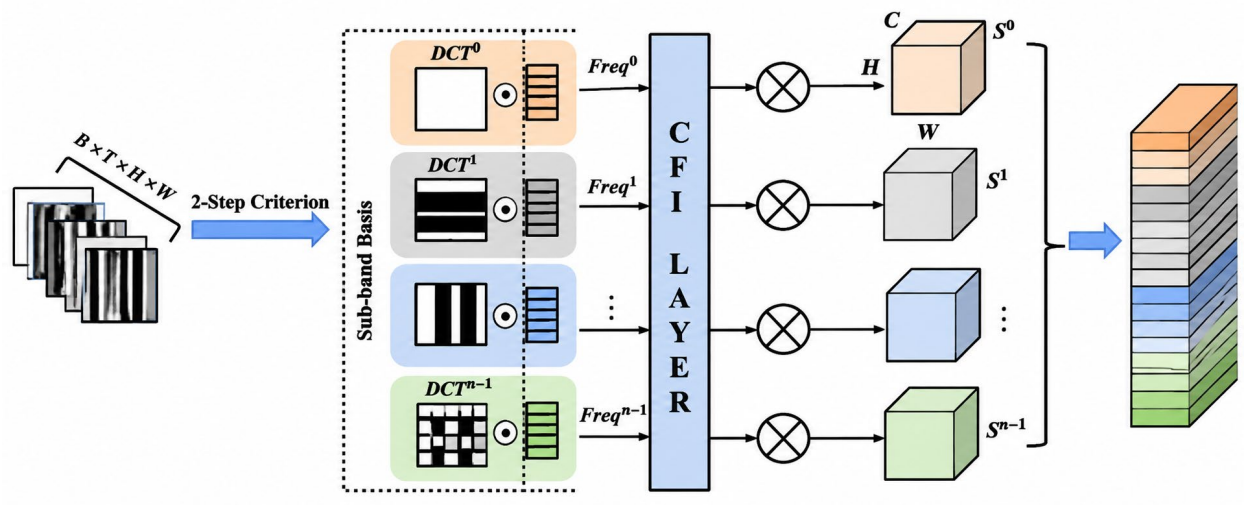

Fig. 2. Structure of frequency attention enhanced with CFI layer.

This structure uses residual linkage, which ensures that the original EEG features are preserved while being enhanced by a CFI layer. While the traditional method measures temporal filtering, the CFI layer approximates these dynamics by constructing a Band Interaction Tensor. This tensor uses an outer product to capture nonlinear co-modulation between the spectral components. The 2D convolutional operation then extracts localized coupling signatures from the interaction map, allowing the model to learn the complex dependencies between low-frequency rhythms and high-frequency oscillations associated with depression symptoms.

F. Sub-Domain Knowledge Alignment Module (SDKA)

The SDKA module focuses on aligning specific classes across different datasets. It uses the Sub-Domain Maximum Mean Discrepancy (SMMD) to provide a more granular measurement of differences than traditional methods. The SMMD is an advanced version of the standard Maximum Mean Discrepancy, which is defined in Eq. (15), where r^p and r^q are samples from the source and target domains, respectively.

$$SMMD(u,v) = \triangleq E_l[\left\| E_{u^{(l)}(r)}[\Psi(r^p)] - E_{v^{(l)}(r)}[\Psi(r^q)] \right\|)^2 \quad (15)$$

The variables $u^{(l)}$, $v^{(l)}$ represent the distributions for a specific class l while $\Psi(\cdot)$ is mapped into a high-dimensional Hilbert space, where E_l signifies the average expectation across both classes. The weights for each sample are calculated based on their labels, as shown in Eq. (16), where z_{il} denotes the label component.

$$c_i^L = \frac{z_{il}}{\sum(r_i,z_i)\in E^{z_{il}}} \quad (16)$$

When applied to the actual activation features of the model (w_i^p, w_j^q), the SMMD loss is expressed using the kernel function, as shown in Eq. (17). This kernel-based method measures the similarity between pairs of samples within and across domains.

$$Loss_{SMMD} = \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{i=1}^{o_p}\sum_{i^l=1}^{o_p} c^{p,l}, c^{p,l} k\left(w_i^p, w_{i^l}^p\right) + \sum_{j=1}^{o_p}\sum_{j^l=1}^{o_p} c^{q,l}, c^{q,l} k\left(w_i^q, w_{i^l}^q\right) - 2\sum_{i=1}^{o_p}\sum_{j^l=1}^{o_q} c^{p,l}, c^{q,l} k\left(w_i^p, w_{i^l}^q\right)\right] \quad (17)$$

The EEG-Conformer is used as a baseline network, where the EEG data is sparse in terms of semantics, and the network is kept lightweight to prevent overfitting. It is simplified to use a single-head attention and a single feed-forward layer. The proposed CFI-TFST model is trained using a combined loss function, as shown in Eq. (18), where β is the weighting factor that determines the influence of the SMMD alignment on the total training process. Algorithm 1 provides an overview of the proposed CFI-TFST.

$$Loss_{Total} = Loss_{CrossEntropy} + \beta Loss_{SMMD} \quad (18)$$

Algorithm 1. Process of the proposed CFI-TFST framework for framework for frequency-aware EEG feature extraction and cross-domain depression classification

Input: Raw EEG signals X, Labels Y, Weighting factor β
Output: Predicted Class, Total Loss
Step 1: Data preprocessing
To isolate the frequencies between 0.5 and 50 Hz, the Nth-order Butterworth filter is applied using Eq. (1).
In Re-referencing, the CAR is calculated to eliminate common-mode noise using Eq. (2).
Artifact removal decomposes the signals into independent sources through linear transformation using Eq. (3).
During normalization, the continuous recordings are segmented into dataset-specific epochs (2sec for MODMA, 5sec for PRED+CT) and applied Z-score normalization using Eq. (4).

Step 2: Frequency Aware Feature Representation
The DCT bias defines the transformation bias for spatio to frequency mapping.
Spectral extraction uses DCT to transform EEG feature maps into the frequency domain using the basis function in Eq. (5) and reconstruction in Eq. (6).
Channel Attention computes important weights for individual channels by collapsing spatial dimensions into a scalar weight vector using Eq. (7).
Multi-Spectral segmentation divides the input T into r channel segments and transforms each into its own DCT spectrum using Eq. (9).
The individual spectral components are concatenated into a vector by weight integration using Eq. (10)
Apply nonlinear transformation is applied to produce final FA weights using Eq. (11).
Step 3: CFI layer
For nonlinear modelling, construct a band interaction tensor to capture the interdependencies between frequency pairs using Eq. (12).
2D convolution and layer normalization are applied to the interaction tensor to stabilize the training process using Eq. (13)
For residual fusion, the cross-frequency interactions are combined back into the original data while preserving the raw features using Eq. (14).
Step 4: SDKA
Distribution alignment measures the differences between the source and target domains using the SMMD, as defined in Eq. (15).
Label weighting calculates the weight of each sample based on the class labels using Eq. (16).
The final kernel-based SMMD loss for alignment is computed using Eq. (17), and the model is trained using the total loss function using Eq. (18).

IV. Experimental Results

The proposed CFI-TFST model is implemented on Windows 11 Pro on an Intel Core i7-12700K processor with a base clock speed of 3.60 GHz and implemented using the PyTorch 2.1.0 DL framework in a Python 3.10 environment. It is equipped with 64 GB of RAM and a 2 TB SSD for extensive multitasking and storage capacity. The graphical setup features a dual-GPU configuration utilizing integrated Intel UHD Graphics 770 and NVIDIA GeForce RTX 3050. Although the standard RTX 3050 has an 8 GB VRAM configuration, it reports a substantial 40 GB of total available graphics memory. Table I provides details of the specific hyperparameter settings and architectural configuration used to train and evaluate the proposed CFI-TFST model. The parameters are designed to accommodate the differing input dimensions of the MODMA and PRED + CT datasets such as variations in electrode counts and sampling rates, while maintaining consistency in the transformer block, optimization technique and SDKA module setting to ensure a consistent comparative analysis. During training, high-strength regularization is applied, including a 0.5 drop rate and 0.2 L2 penalty. The proposed model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 64. To mitigate the risk of overfitting given the deep parameterized architecture, an early stopping mechanism is implemented. This mechanism monitors the validation loss with a patience threshold of 15 epochs such as the training terminates automatically if no lower validation loss is recorded for 15 consecutive iterations, restoring the optimal weights. The baseline models are reconstructed at their input interfaces to accept single-modality neural tensors to ensure fair reproducible comparison under a unified experimental architecture as shown:

- CNN-GRU: The temporal input was reconstructed by mapping the multi-channel spatial dimension of EEG epoch tensor into multivariate feature dimension of initial 1D convolutional layers. This enables the GRU to track cross-channel temporal sequences.
- Spectro-CNN: The raw EEG channels are processed using Short-Time Fourier Transform (STFT) to compute 2D time-frequency power spectral density spectrograms per channel. The spatial-spectral maps are then fed to into the 2D CNN layers.
- Synchronized Multimodal Transformer: The linguistic and paralinguistic feature branches were entirely decoupled from the architecture. The cross attention fusion block is reconfigured to function as an intra-modal temporal attention. This evaluates the cross-correlations between localized temporal partitions of multi-channel EED data.

TABLE I. Hyperparameter Configuration for the Proposed CFI-TFST Model

Category	Hyperparameter	MODMA dataset	PRED + CT
Convolutional features	Temporal Conv Kernel	(1,25)	(1,25)
	Spatial Conv Kernel	(128,1)	(62,1)
Transformer Block	Self-Attention Heads	1	1
	Encoder layers	1	1
	Dropout rate	0.5	0.5
	Top-K Frequency Components ($\boldsymbol{K}$)	5	5
Classification	FC layers	3	3
Optimization	Optimizer type	Adam	Adam
	L2 regularization	0.2	0.2
	L2 scheduler	Exponential Decay	Exponential Decay
	Initial learning rate	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-4}}$
	Learning rate decay	0.1 (per seconds)	0.1 (per seconds)
	SMMD weight	0.5	0.7
Execution	Total epochs	50	50
	SDKA batch size	64	64

A. Evaluation Metrics

Accuracy measures the proportions of correctly classified samples (both depressed and non-depressed), and the formula is defined in Eq. (19), where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

Precision represents the probability that a person flagged as depressed by the model is actually depressed and is calculated using Eq. (20).

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

Recall quantifies the probability that the model detects depression when it is present, and is measured using Eq. (21).

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

The F1 Score is the harmonic mean of precision and recall, and is computed using Eq. (22)

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{22}$$

B. Quantitative Evaluation

Figs. 3 and 4 present a comparative analysis of the baseline models by comparing them with the proposed model. Baseline models such as the Multi-Stage Hemispheric Asymmetry Fusion Network (MSHAF-Net), Local-Global Feature Fusion (LGFF-Net), and Depression Hybrid Neural Network (DeepHNN) are analyzed for both utilized datasets. The performance is evaluated across four key metrics, where the proposed CFI-TFST model consistently outperformed the baseline models, and the combination of the CFI layer with FA improves performance.

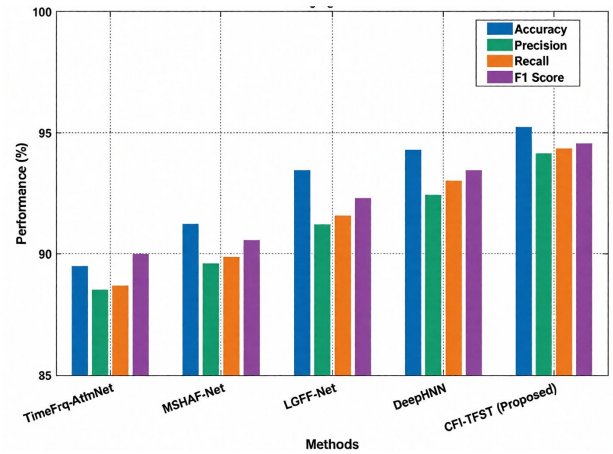

Fig. 3. Quantitative analysis of proposed CFI-TFST with baseline models for MODMA dataset.

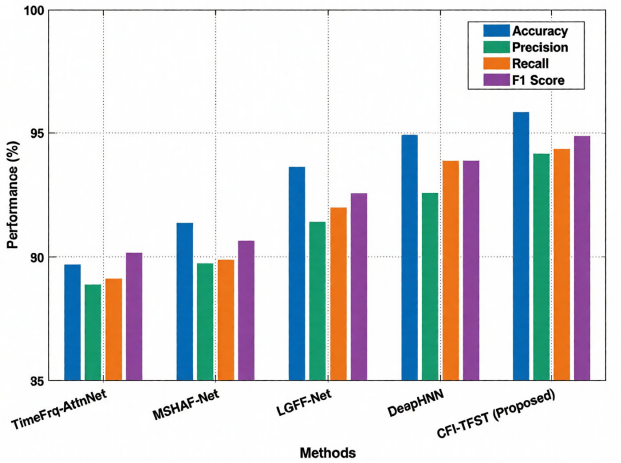

Fig. 4. Quantitative analysis of proposed CFI-TFST with baseline models for PRED + CT dataset.

Table II presents the experimental results for the proposed CFI-TFST method across different K-Fold cross-validation configurations using both datasets. To guarantee that no subject leakage occurs during cross-validation, a Grouped K-Fold strategy is implemented. In this setup, the folds are split based on subject IDs rather than individual epochs where all epochs belong to a single participant, which is kept entirely within the same fold during both training and testing phases. The results indicate that the proposed CFI-TFST achieves better performance at K = 5, consistently outperforming the baseline models and demonstrating its effectiveness in generalizing across different datasets. A lower k has a high bias, whereas a high k has a high variance in error estimation. K = 5 has been empirically shown to provide an appropriate balance that does not suffer from excessively high bias. For PRED + CT dataset with total 121 subjects, Fold 1 consists of 25 unique subjects with 10 depressed individuals and 15 healthy controls, whereas Folds 2 through 5 each consists of 24 unique subjects with 9 depressed individuals and 15 healthy controls. In MODMA dataset with 53 total subjects, Folds 1 through 4 consists of 11 unique subjects with 5 depressed individuals and 6 healthy controls, whereas Fold 5 consists of remaining 9 unique subjects with 4 depressed individuals and 5 healthy controls. To eliminate any risk of data leakage, all hyperparameter optimization procedures are performed within the training partitions of each cross-validation loop. The respective testing folds remained completely unobserved during model configuration. For the primary configuration (K = 5), the proposed model shows generalization across small-subject EEG distributions, maintaining negligible fold-wise variance. For the MODMA dataset, individual raw fold accuracy are closely clustered between a minimum of 95.12% and maximum of 95.35%, Early stopping is activated uniformly for these individual training runs across an epoch distribution of 39 ± 3.7. Similarly, for the PRED+CT dataset, individual raw fold accuracies range from 95.49% to 96.15%. The network reached convergence rapidly, with the early-stopping mechanism across an epoch distribution of 36.0 ± 3.4.

TABLE II. K-FOLD CROSS-VALIDATION OF PROPOSED CFI-TFST FOR MODMA AND PRED + CT DATASETS

Dataset	K-Fold validation	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
MODMA dataset	$K = 2$	89.52 ± 0.65	88.54 ± 0.72	88.69 ± 0.68	90.01 ± 0.59
	$K = 3$	91.24 ± 0.48	89.61 ± 0.55	89.87 ± 0.51	90.56 ± 0.42
	$K = 4$	93.47 ± 0.39	91.23 ± 0.42	91.56 ± 0.40	92.31 ± 0.35
	$K = 5$	95.24 ± 0.25	94.17 ± 0.31	94.36 ± 0.28	94.57 ± 0.22
	$K = 6$	94.31 ± 0.32	92.45 ± 0.38	93.01 ± 0.34	93.48 ± 0.29
PRED + CT dataset	$K = 2$	89.69 ± 0.61	88.88 ± 0.69	89.12 ± 0.64	90.18 ± 0.55
	$K = 3$	91.37 ± 0.45	89.75 ± 0.52	89.90 ± 0.47	90.64 ± 0.39
	$K = 4$	93.64 ± 0.36	91.42 ± 0.40	91.98 ± 0.38	92.57 ± 0.31
	$K = 5$	95.87 ± 0.22	94.17 ± 0.28	94.36 ± 0.25	94.89 ± 0.19
	$K = 6$	94.93 ± 0.29	92.59 ± 0.35	93.87 ± 0.31	93.88 ± 0.26

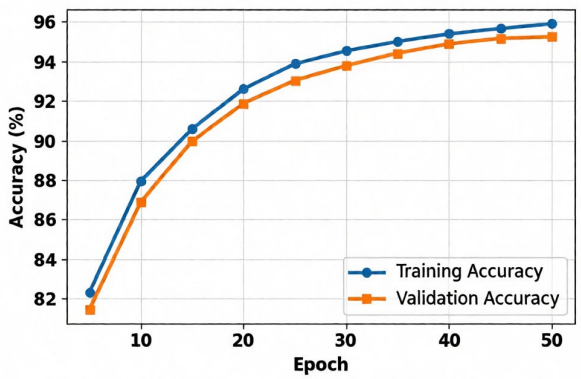

Fig. 5. Epoch-level learning curves for MODMA dataset (Fold 4, Random Seed 42): (a) Training and validation accuracy curve across 50 training epochs (b) Training loss and validation loss curves.

To demonstrate that the highly parameterized components do not induce overfitting on small datasets, the learning progression curves are analyzed in Fig. 5. Specifically, Fig. 5(a) illustrates the step-wise trajectory of training accuracy and validation accuracy across 50 training epochs. Concurrently, Fig. 5(b) traces the continuous minimization behavior of the corresponding training loss and validation loss curves. Throughout the training cycle, the validation trajectories are closely similar to the training curves with minimal divergence. The training accuracy reaches 95.91% while validation accuracy reaches stable convergence at 95.24%. Similarly, the validation loss smoothly decreases to 0.148 following the training loss. No widening performance gap is observed in either evaluation plot. The parallel and stable convergence behavior provides evidence of enhanced generalization. These curves correspond to the execution of Fold 4 during cross-validation on the MODMA dataset, which is initialized with a fixed random seed of 42 for reproducibility. To prevent the model from memorizing training noise, the early stopping mechanism monitored the validation loss curve and training was terminated uniformly.

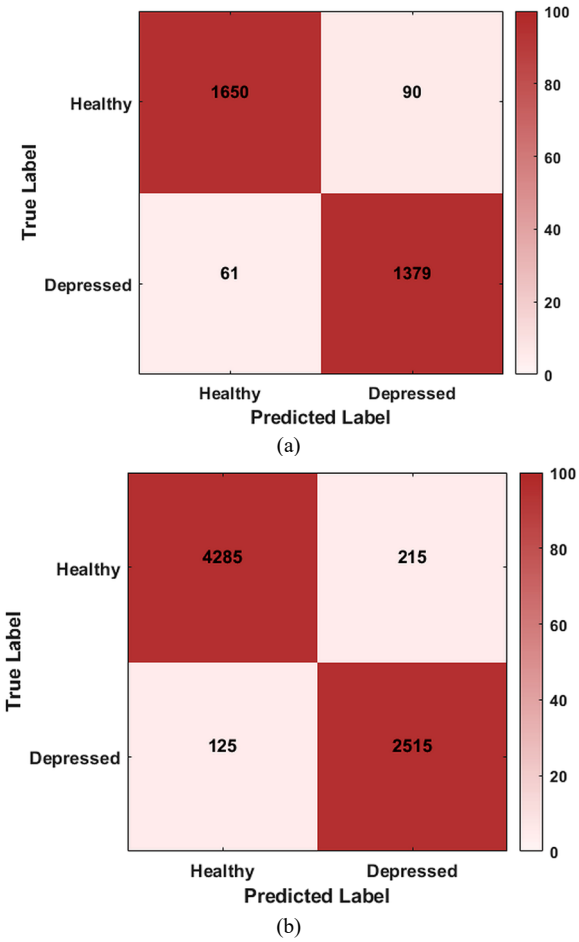

Fig. 6. Epoch-level Confusion Matrices for (a) MODMA (2-sec epochs) and (b) PRED + CT dataset (5-sec epochs).

Fig. 6 shows the confusion matrices for the CFI-TFST method across two different datasets to evaluate the classification. The confusion matrix is presented based on a representative cross-validation scenario instead of an average of multiple fold values to ensure concrete representation of raw prediction counts. It highlights the ability to accurately distinguish between Healthy and Depressed states using test epochs from the selected

evaluation fold. The matrices compare the True Labels against Predicted Labels with a high concentration of samples along the principal diagnosis, showing high sensitivity and specificity of the proposed models. The minimal off-diagonal values across both datasets indicate the credibility of the integrated CFI and spatial transformer modules, which reduces misclassifications. The data partitioning follows a subject-level strategy to isolate participants and eliminate the risk of data leakage but the model classification performance is evaluated at the epoch level. No subject-level aggregation and voting is applied to compile the metrics.

C. Computational and Statistical Analysis

Table III presents the computational analysis of the proposed CFI-TFST model, detailing the efficiency of its individual modules. It outlines the time and memory complexity of key components, including data preprocessing, convolutional layers, transformers, and other architectural configurations used to improve computational performance. To substantiate theoretical bounds with physical metrics, Table IV shows the performance against baseline models under identical conditions using an NVIDIA RTX 3050 GPU with 8 GB of dedicated VRAM. All evaluations are conducted using a batch size of 64 and input matrix tensor shapes for MODMA and for PRED+CT. Latency measurements are averaged over 100 inference runs following 10 initial warmup runs to eliminate caching bias. While integrating tracking components, the proposed model limits its operational memory overhead to 710 MB and trains at 4.5 s per epoch, demonstrating high computational efficiency.

TABLE III. COMPUTATIONAL ANALYSIS OF THE PROPOSED CFI-TFST BASED ON THE MODULES

Module	Time complexity	Memory complexity	Configurations
Data preprocessing segmenting (2-sec for MODMA and 5-sec epochs for PRED+CT)	$\boldsymbol{O(N)}$	$\boldsymbol{O(N)}$	Linear scan for filtering and segmenting (2-sec for MODMA and 5-sec epochs for PRED+CT)
Convolutional Layer	$\boldsymbol{O(N \cdot C \cdot K)}$	$\boldsymbol{O(C \cdot D)}$	Dependent on kernel size (K = 25) and channel count (C = 62)
Transformer Encoder	$\boldsymbol{O(N^2 \cdot D)}$	$\boldsymbol{O(N^2 + N \cdot D)}$	Self-attention matrix scales quadratically with sequence length N
SDKA module	$\boldsymbol{O(B \cdot D^2)}$	$\boldsymbol{O(B \cdot D)}$	Domain alignment calculations per batch B
Classifiers	$\boldsymbol{O(D \cdot H)}$	$\boldsymbol{O(D \cdot H)}$	Matrix multiplications for hidden units H
Total Inference	Low medium	$\sim$ **40 GB**	Optimized by using only 1 layer and 1 head in the transformer

TABLE IV. EMPIRICAL COMPUTATIONAL RESOURCE COMPARISON AGAINST BASELINE METHODS

Models	Trainable parameters (M)	Training Time per epoch (s)	inference Latency per batch (ms)	peak GPU memory usage (MB)
CNN-GRU	0.18	2.1	8.4	420
Spectro CNN	0.35	3.4	10.1	510
Deephnn	1.12	5.8	14.6	890
CFI-TFST (proposed)	0.68	4.5	12.8	710

To evaluate whether the performance improvements produced by the proposed method are statistically valid, pairwise comparisons are performed against baseline models as shown in Table V. The metric units used for the statistical validation are the overall classification accuracies obtained over a 2-times repeated 5-fold cross-validation scheme that yields 10 distinct fold-level observations (K = 10). Because cross-validation splits depend on overlapping training samples across different runs, standard independence assumptions are violated making a traditional Student's t-test statistically invalid. To mathematically adjust for this sample dependence, a Corrected Resampled Paired t-test is applied using the corrected t-value formula $t = \frac{\bar{d}}{\sqrt{s^2(\frac{1}{K}+\frac{n_2}{n_1})}}$. Here, $\bar{d}$ is the mean of the differences, s^2 is the variance of the differences. The parameter K is set to 10 fold-level observations generated across the cross-validation partitions. The sample sizes n_1 and n_2 are defined based on the independent subject-level partitions rather than raw epoch counts. As the epochs from same participant lack statistical indepdence, treating them as independent units falsely reduces the calculated variance. Here n_1 denotes the training set size (encompasses approximately 80% of the data) and n_2 denotes testing set size (holds approximately 20% of the data) in the 5-fold cross-validation distribution, which establish operational ratio of $\frac{n_1}{n_2} \approx 0.25$ across the statistical evaluation loops. Normality is verified through the Shapiro-Wilk test ($p > 0.05$). With 9 degrees of freedom (df = 9), the proposed model performed statistically significant over all baselines across both datasets, with p-values lower than the standard ($\alpha = 0.05$) threshold. The null hypothesis (H_0) suggests that there is no significant difference in depression detection performance between the proposed model and the baseline models. Conversely the alternative hypothesis (H_1) states that integrating the CFI and FA layers significantly enhances classification performance. To control the error rate arising from multiple pairwise comparisons, a sequential step-down Holm-Bonferroni adjustment is applied per dataset. The resulting Holm-adjusted p-values, paired difference mean values with their standard deviations, corrected 95% Confidence Interval (CI), and Cohens d effect indicate statistical differences under the adopted test.

TABLE V. STATISTICAL ANALYSIS OF PROPOSED CFI-TFST FOR BOTH MODMA AND PRED + CT DATASET

Dataset	CFI-TFST vs. Baseline Methods	Baseline accuracy mean ± SD (%)	Paired Diff. (Mean±Standard Deviation)	Corrected t-value	Raw test *p* value	Corrected 95% CI	Cohen's d	Holm Adjusted p value
MODMA	CNN-GRU	94.25±0.32	0.99±0.46	3.65	0.0053	[0.38,1.60]	2.16	0.0159
	Spectro CNN	94.36±0.35	0.88±0.43	3.42	0.0076	[0.30,1.46]	2.02	0.0159
	DeepHNN	94.85±0.28	0.39±0.23	2.89	0.0179	[0.08,0.70]	1.71	0.0179
PRED + CT	CNN-GRU	93. 65±0.41	2.22±0.77	4.88	0.0009	[1.19,3.25]	2.89	0.0027
	Spectro CNN	94.02±0.38	1.85±0.74	4.21	0.0023	[0.86,2.84]	2.49	0.0046
	DeepHNN	94.67±0.31	1.20±0.57	3.54	0.0063	[0.43,1.97]	2.09	0.0063

D. Ablation Study

The proposed CFI-TFST model across the MODMA and PRED + CT datasets is evaluated based on the architectural components, as shown in Table VI. It systematically evaluates the incremental performance gains provided by each integrated component, including the FA, CFI layer, and SDKA, across all four key metrics. This progressive evaluation shows how the integration of frequency-aware features, non-linear inter-band modelling, and domain adaptation collectively drives the model to its highest observed performance.

TABLE VI. ABLATION EVALUATION OF THE PROPOSED CFI-TFST FOR MODMA AND PRED + CT DATASETS

dataset	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score(%)
MODMA dataset	TFST	93.24 ± 0.42	92.87 ± 0.45	93.01 ± 0.41	93.25 ± 0.39
	TFST + FA	94.02 ± 0.38	93.01 ± 0.41	93.45 ± 0.36	93.54 ± 0.34
	TFST+FA+CFI	94.65 ± 0.31	93.97 ± 0.35	94.05 ± 0.32	94.21 ± 0.28
	TFST+FA+CFI + SDKA	95.24 ± 0.25	94.17 ± 0.31	94.36 ± 0.28	94.57 ± 0.22
PRED + CT	TFST	93.02 ± 0.45	92.98 ± 0.48	93.05 ± 0.44	93.05 ± 0.41
	TFST + FA	93.25 ± 0.41	92.98 ± 0.43	93.05 ± 0.39	93.22 ± 0.37
	TFST+FA+CFI	94.35 ± 0.34	93.87 ± 0.38	93.65 ± 0.32	94.29 ± 0.29
	TFST+FA+CFI + SDKA	95.87 ± 0.22	94.17 ± 0.28	94.36 ± 0.25	94.89 ± 0.19

E. Cross-Dataset Transfer Validation

To validate the domain generalization capability of the proposed framework, cross-dataset transfer experiments are performed between the MODMA and PRED + CT datasets. In these experiments, the model is trained on one dataset (source domain) and evaluated on the completely unseen secondary dataset (target domain) without any parameter updates or fine-tuning. Because the two datasets use different sensor layouts, the 128 channels of MODMA and 62 operational channels of PRED + CT are mapped to a shared subset of 60 unique electrodes configured under international 10-10 system. Table VII presents the cross-dataset transfer performance under subject-level isolation. Specifically, the model is trained on MODMA and tested on PRED + CT, while the model trained on PRED + CT is verified against all 53 unique target subjects of the MODMA dataset to avoid cross-study data contamination. Baseline models experience performance degradation when shifting domains, highlighting their vulnerability to device and cohort variations. Conversely, the proposed model maintains improved generalization metrics, achieving an accuracy of 89.34% on the PRED + CT target domain and 88.65% when transferred to the MODMA target domain. This performance suggests improved cross-dataset transfer under the evaluated harmonization protocol. Spatial alignment is performed using spherical spline interpolation to map missing sensor projections, and time-series data are uniformly resampled to 250 Hz.

TABLE VII. CROSS-DATASET EVALUATION RESULTS FOR DOMAIN GENERALIZATION (SOURCE FOR TRAINING →TARGET FOR TESTING)

Model	MODMA→PRED+CT Accuracy (%)	MODMA→PRED+CT F1 Score (%)	PRED+CT→MODMA Accuracy (%)	PRED+CT→MODMA F1 Score (%)
CNN-GRU	82.45 ± 0.68	81.12 ± 0.72	81.3 ± 60.65	80.05 ± 0.70
Spectro CNN	83.12 ± 0.62	82.04 ± 0.65	82.50 ± 0.59	81.42 ± 0.63
DeepHNN	85.67 ± 0.55	84.50 ± 0.58	84.15 ± 0.51	83.10 ± 0.54
CFI-TFST (Proposed)	89.34 ± 0.41	88.21 ± 0.45	88.65 ± 0.38	87.48 ± 0.42

The 60 specific electrodes used for this shared feature space are Fp1, Fp2 (Left and Right Pre-Frontal), AF3, AF4 (Left and Right Anterior-Frontal), F7, F5, F3, F1 (Left Outer, Middle, and Inner Frontal), Fz (Frontal Midline), F2, F4, F6, F8 (Right Inner, Middle, and Outer Frontal FT7, FT8 (Left and Right Frontotemporal) FC5, FC3, FC1 (Left Outer, Middle, and Inner Front central), FCz (Frontocentral Midline), FC2, FC4, FC6 (Right Inner, Middle, and Outer Frontocentral), T7, T8 (Left and Right Temporal) , C5, C3, C1 (Left Outer, Middle, and Inner Central), Cz (Central Midline), C2, C4, C6 (Right Inner, Middle, and Outer Central), TP7 (Temporoparietal), CP5, CP3, CP1 (Left Outer, Middle, and Inner Centroparietal), CPz (Central Midline), CP2, CP4, CP6 (Right Inner, Middle, Centro-Paritel), P7, P5, P3, P1 (Left Outer, Middle, and Inner Parietal), Pz (Parietal Middle), P2, P4, P6, P8 (Right Inner, Middle, and Outer Parietal rows), PO7, PO5, PO3 (Left Outer, Middle, and Outer Parietal), POz (Parieto-Occipital Midline), PO4, PO6, PO8 (Right Inner, Middle, and Outer Parieto-Occipital), O1, O2 (Left

and Right Occipital), Oz (Occipital Midline, and CB1 (Left Cerebellar).

F. Comparison Study

Table VIII presents a detailed comparative study evaluating the performance of the CFI-TFST method against existing methods for depression detection. By benchmarking against the accuracy and F1 Score, the results highlight the improved performance of the CFI-TFST method over existing models such as CNN-GRU [21], Spectro-CNN [23], and Synchronized Multimodal Transformer [26]. To ensure fair and reproducible evaluation, all existing models are re-implemented and executed within the exact same experimental environment for EEG baselines and cross-modality contextual modalities. All models are executed in an identical experimental environment using the same subject-level validation parameters to ensure an objective architectural comparison. Rather than relying on historical source-reported metrics that use varying data pipeline, all models are subjected to the proposed configurations preprocessing and validation parameters. The baseline models such as (CNN-GRU, Spectro-CNN and Synchronized Transformer) are not evaluated in PRED + CT dataset because of architectural limitations when processing long temporal inputs. In preprocessing phase, a 5-second window has been chosen to capture slow and low-frequency brain rhythms. Although, this extended window increases the size of input data and standard baseline architectures lack attention mechanisms needed to track patterns over this length of time that leads to training failures. As these models struggles to handle large volume of data, the PRED + CT analysis focuses solely on TFST-SDKA and CFI-TFST models.

TABLE VIII. COMPARISON EVALUATION OF THE PROPOSED MODEL WITH THE EXISTING MODEL (RE-IMPLEMENTED BASED ON SAME CONFIGURATION)

Dataset	Methods	Accuracy	F1 Score
MODMA	CNN-GRU [21]	94.25±0.35	92.00±0.38
	Spectro-CNN [23]	94.36±0.31	89.00±0.42
	TFST-SDKA [24]	92.45±0.44	91.30±0.47
	Synchronized Multimodal Transformer [26]	91.22±0.25	91.23±0.22
	CFI-TFST (Proposed)	95.24±0.39	94.57±0.41
PRED + CT	TFST-SDKA [24]	93.16±0.52	90.91±0.50
	CFI-TFST (Proposed)	95.87±0.22	94.89±0.19

G. Discussion

The experimental results show that the proposed CFI-TFST model comparatively improves EEG-based depression detection performance compared with existing models. The impact of cross-frequency modelling is provided by an ablation study, where isolating the CFI layer improves performance over the TFST + FA model. Specifically, the model achieved accuracies of 95.24% on the MODMA dataset and 95.87% on the PRED + CT dataset, outperforming existing architectures. This development is primarily attributed to the integration of the CFI layer with the FA mechanism. By explicitly modelling nonlinear interdependencies between different frequency bands, the proposed model captures complex cross-frequency features that standard attention mechanisms overlook. The implementation of K-Fold cross-validation confirmed the stability of the model across varying data segments. The numerical difference between the accuracy shown in the confusion matrix of Fig. 6(b) and the mean value reported in Table II occurs because Table II represents the average performance, whereas Fig. 6(b) displays the raw epoch-level classification counts from a single representative cross-validation scenario. The SDKA module also plays an important role in ensuring generalization by using the SMMD to align specific classes across different datasets. Despite its advanced feature extraction capabilities, the model remains significant by using a lightweight EEG-Conformer baseline with simplified attention layers, making it suitable for clinical applications.

V. CONCLUSION

This manuscript introduces the CFI-TFST model, which improves the objective detection of depression using EEG signals. By integrating the CFI layer with the FA mechanism, the proposed CFI-TFST captures complex nonlinear cross-frequency interactions, that are not captured by traditional feature engineering methods. The experimental results across both datasets demonstrated that the proposed CFI-TFST model outperforms the baseline models by achieving higher accuracy. These results consistently outperform the state-of-the-art models, namely, DeepHNN and TFST-SDKA. The implementation of the SDKA module ensured that the proposed CFI-TFST model maintained improved transferability between MODMA and PRED + CT. This study provides a scalable framework for research-oriented depression screening using physiological biomarkers for monitoring mental health in clinical settings. Future models can be optimized for integration into wearable IoT devices, such as smartwatches and EEG-integrated headbands, which enable continuous, non-invasive mental health monitoring in daily life settings.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Monika Puttaramaiah conceptualized and designed the study, supervised the overall research work, and contributed to manuscript writing and critical revision.

Sneha S Bagalkot conducted the research work, performed data collection and experimentation, and contributed to drafting the manuscript. Amogh Pramod Kulkarni carried out the data analysis, validation of results, and assisted in the interpretation of findings. All authors reviewed, revised, and approved the final version of the manuscript.

REFERENCES

[1] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, and B. Alhaqbani, "High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.,* vol. 20, no. 4, pp. 2587–2597, March 2023.

[2] S. Ren and J. Song, "A graph-based method for automatic graded diagnosis of depression using EEG signals," *Biomed. Signal Process. Control*, vol. 100, 106973, February 2025.

[3] L. Zhou, B. Hu, and Z. H. Guan, "Gaborformer: A method for depression detection through hybrid acoustic feature extraction and fusion," *Journal of Affective Disorders*., vol. 400, 121172, May 2026.

[4] C. Sun, M. Jiang, L. Gao, Y. Xin, and Y. Dong, "A novel study for depression detecting using audio signals based on graph neural network," *Biomed. Signal Process. Control*, vol. 88, 105675, February 2024.

[5] Y. Dou, T. Xing, X. Zhao, X. Chen, J. Zhou, and S. Peng, "Spatio-temporal representation learning with selective state space models for EEG-based depression detection," *Biomed. Signal Process. Control,* vol. 112, 108707, February 2026.

[6] X. Li, G. Huang, Y. Liu, and K. S. Hong, "EEG-based depression detection using a local–global feature fusion deep learning network," *Biomed. Signal Process. Control*, vol. 118, 109681, June 2026.

[7] G. Tasci, H. W. Loh, P. D. Barua, M. Baygin, B. Tasci, S. Dogan, T. Tuncer, E. E. Palmer, R.S. Tan, and U. R. Acharya, "Automated accurate detection of depression using twin Pascal's triangles lattice pattern with EEG signals," *Knowledge-Based Syst*., vol. 260, 110190, January 2023.

[8] Y. Zhou, X. Yu, H. Lin, R. Li, J. Liang, X. Shi, and Y. Luo, "A shallow 2D self-attention-CNN framework for depression severity identification using EEG functional connectivity network," *Biomed. Signal Process. Control*, vol. 115, 109397, April 2026.

[9] Z. Wu, G. Luo, H. Zhang, Z. Tian, X. Hu, and D. Xu, "TwoM: A model for joint time–frequency representation and spatial filtering for EEG-based depression recognition," *Biomed. Signal Process. Control*, vol. 113, 108977, March 2026.

[10] B. Taşcı, "Multilevel hybrid handcrafted feature extraction based depression recognition method using speech," *Journal of Affective Disorders*, vol. 364, pp. 9–19, November 2024.

[11] A. Ksibi, M. Zakariah, L. J. Menzli, O. Saidani, L. Almuqren, and R. A. M. Hanafieh, "Electroencephalography-based depression detection using multiple machine learning techniques," *Diagnostics*, vol. 13, no. 10, 1779, May 2023.

[12] J. Shen, K. Li, H. Liang, Z. Zhao, Y. Ma, J. Wu, J. Zhang, Y. Zhang, and B. Hu, "HEMAsNet: A hemisphere asymmetry network inspired by the brain for depression recognition from electroencephalogram signals," *IEEE J. Biomed. Health. Inf.,* vol. 28, no. 9, pp. 5247–5259, May 2024.

[13] G. Luo, H. Rao, P. An, Y. Li, R. Hong, W. Chen, and S. Chen, "Exploring adaptive graph topologies and temporal graph networks for EEG-based depression detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3947–3957, September 2023.

[14] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Syst. Appl.,* vol. 249, 123834, September 2024.

[15] Z. Wang, C. Hu, W. Liu, X. Zhou, and X. Zhao, "EEG-based high-performance depression state recognition," *Front. Neurosci*., vol. 17, 1301214, January 2024.

[16] B. Zhang, D. Wei, G. Yan, X. Li, Y. Su, and H. Cai, "Spatial–temporal EEG fusion based on neural network for major depressive disorder detection," *Interdiscip. Sci.: Comput. Life Sci*., vol. 15, no. 4, pp. 542–559, May 2023.

[17] M. Xia, Y. Zhang, Y. Wu, and X. Wang, "An end-to-end deep learning model for EEG-based major depressive disorder classification," *IEEE Access*, vol. 11, pp. 41337–41347, April 2023.

[18] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer-based models for identifying stress and depression in social media," *IEEE Trans. Comput. Social Syst*., vol. 11, no. 2, pp. 1979–1990, June 2023.

[19] S. A. M. Mane and A. Shinde, "StressNet: Hybrid model of LSTM and CNN for stress detection from Electroencephalogram signal (EEG)," *Results Control Optim*., vol. 11, 100231, May 2023.

[20] R. F. Dougherty, P. Clarke, M. Atli, J. Kuc, D. Schlosser, B. W. Dunlop, D. J. Hellerstein, S. T. Aaronson, S. Zisook, A. H. Young, and R. Carhart-Harris, G. M. Goodwin, and G. A. Ryslik "Psilocybin therapy for treatment resistant depression: Prediction of clinical outcome by natural language processing," *Psychopharmacology*, vol. 242, no. 7, pp. 1553–1561, August 2025.

[21] S. Zhang, Z. Zheng, D. He, H. Zhu, Y. Gu, Y. Hu, X. Lv, T. Zhu, and W. Zhang, "Speech-Based Depression Detection: Enhancing Emotional Support via Clinical Data," *IEEE Trans. Affective Comput*., December 2025.

[22] S. De, A. Singh, V. Tiwari, H. Patel, G. N. Vivekananda, and D. S. Rajput, "Slitranet: An EEG-based automated diagnosis framework for major depressive disorder monitoring using a novel LGCN and transformer-based hybrid deep learning approach," *IEEE Access*, vol. 12, pp. 173109–173126, November 2024.

[23] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomed. Signal Process. Control*, vol. 90, no. 1–3, 105898, April 2024.

[24] C. Y. Xu, F. Y. Fan, L. X. Zhao, L. C. Jin, Y. H. Zhang, and Q. H. Meng, "Major depressive disorder detection via temporal-frequency-spatial transformer with sub-domain knowledge alignment using EEG," *Neural Networks*, vol. 193, 107965, January 2026.

[25] A. Sam, R. Boostani, S. Hashempour, M. Taghavi, and S. Sanei, "Depression identification using EEG signals via a hybrid of LSTM and spiking neural networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4725–4737, November 2023.

[26] N. Esmi, A. Shahbahrami, G. Gaydadjiev, and P. de Jonge, "Multimodal transformer for depression detection based on EEG and interview data," *Biomed. Signal Process. Control*, vol. 113, 109039, March 2026.

[27] Y. Xi, Y. Chen, T. Meng, Z. Lan, and L. Zhang, "Depression detection based on the temporal-spatial-frequency feature fusion of EEG," *Biomed. Signal Process. Control*, vol. 100, 106930, February 2025.

[28] W. Wang, J. Chen, Y. Hu, H. Liu, J. Chen, T. R. Gadekallu, L. Garg, M. Guizani, and X. Hu, "Integration of artificial intelligence and wearable internet of things for mental health detection," *Int. J. Cognit. Comput. Eng.*, vol. 5, pp. 307–315, July 2024.

[29] MODMA Dataset. [Online]. Available: https://modma.lzu.edu.cn/data/index/

[30] J. F. Cavanagh, A. Napolitano, C. Wu, and A. Mueen, "The Patient Repository for EEG Data + Computational Tools (PRED + CT)," *Front Neuroinform.*, vol. 11, 67, 2017. doi: 10.3389/fninf.2017.00067

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).