Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPECTFT-DSDNET: SPECTRAL MODULATED TEMPORAL FUSION TRANSFORMER ENHANCED FOR EMOTION RECOGNITION

PAGADALA VEERANJANEYULU[1,*], PADMAJA USHA RANI DUDA[2], SINDHU BOIANAPALLI[3], SRI LAKSHMI PARVATHI SITANABOINA[4], RAMYA KRISHNA DHULIPALLA[5]

[1]Department of Information Technology, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology, Hyderabad, India
[2]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Keshav Memorial Engineering College, Hyderabad, India
[3]Department of Computer Science and Engineering, CVR College of Engineering, Mangalpalli, India
[4]Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, India
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India

Email: [1]veeranjaneyulu_p@vnrvjiet.in, [2]padmaja1911@gmail.com, [3]sindhu1209@gmail.com, [4]s.srilakshmiparvathi@gmail.com, [5]d.ramyakrishna@klh.edu.in,

ABSTRACT

Physiological monitoring for emotion recognition provides objective affective state assessments resistant to social masking. However, decoding non-stationary Electroencephalography (EEG) and Electrocardiogram (ECG) signals across subjects using existing deep neural networks remains challenging due to high spectral blurring, amplitude scale mismatches between microvolt and millivolt signals, and temporal information loss. To address this, the Spectral Modulated Temporal Fusion Transformer Enhanced Dynamic Spectrum Driven Network (SpecTFT-DSDNet) is introduced. The proposed architecture utilizes Synchrosqueezed Dynamic Short-Time Fractional Fourier Transform (SSQ-DSTFFT) spectral mapping, enhanced by a novel, learnable spectral modulation block and a Temporal Fusion Transformer (TFT) to address inter-subject variability. These representations are encoded through an Efficient Convolutional Neural Network (CNN) and processed through a Temporal Fusion Transformer (TFT), which is equipped with a Gated Residual Network (GRN) and Variable Selection Network (VSN) to suppress irrelevant noise. A unique learnable spectral modulation block is integrated between the Long Short Term Memory (LSTM) layers and masked attention to ensure that model remains frequency aware. SpecTFT-DSDNet is evaluated under a subject-independent 5-fold cross-validation protocol, by achieving classification accuracy of 98.89% for valence and 98.88% for arousal on DEAP dataset, 99.28% for valence and 99.41% for arousal on DREAMER dataset, and 98.92% for valence 99.58% for arousal on AMIGOS dataset. SpecTFT-DSDNet enhances spectral energy concentration by 68.6% over existing time-frequency transforms while maintaining an inference latency of 12.4 ms per segment. These ensure that using integrated frequency-aware spectral modulation with gated temporal transformer context effectively, which resolves inter-subject signal variance, introducing scalable and computationally viable solution for cross-subject physiological affective computing.

Keywords: *Efficient Convolutional Neural Network, Emotion Recognition, Gated Residual Network, Synchrosqueezed Dynamic Short-Time Fractional Fourier Transform, Temporal Fusion Transformer, Variable Selection Network*

1. INTRODUCTION

Emotion functions as an integrated neurophysiological and evaluative social phenomenon that bridges the gap between internal biological states and external social interactions [1]. Brain-based monitoring provides better ecological validity because it is resistant to social desirability bias. Users easily mask their physiological reactions at the cerebral level [2]. Electroencephalography

(EEG) is a highly advantageous and adaptable method for measuring neural activity using a noninvasive interface to measure cerebral potential [3]. Combining various physiological data points in emotion recognition enhances the classification performance and strengthens the reliability of the results [4]. Electrocardiogram (ECG) and EEG data require the merging of different signal profiles, particularly mismatched sampling frequencies, unique wave patterns, and various noise profiles [5]. EEG captures electrical signals from the outer layers of the brain and provides better temporal resolution for tracking emotional responses. It reflects cortical activity rather than direct signals from deep structures, such as the amygdala, where EEG provides a direct link to the central nervous system [6]. Affective states significantly control both physiological health and the cognitive decision-making process, and the identification of these states is required for advancing clinical diagnosis as well as Human Computer Interaction (HCI) [7]. Research indicates that human affect is categorized into six primary states happiness, sadness, anger, fear, surprise, and hatred, which provide discrete emotions for identifying specific psychological responses [8]. To facilitate classification, emotions are mapped according to valence (positivity/negativity) and arousal (intensity of the stimulus), which allows a continuous coordinate-based representation of affective states [9]. Physiological emotion decoding merges central (EEG) and autonomic (ECG) nervous system signals. However, inter-subject biological variability and non-stationarity causes spectral blurring under the Heisenberg-Gabor limit, causing Deep Learning (DL) models to overfit subject baselines, hindering zero-calibration deployment in real-time affective computing [10]. DL methods frequently fuse EEG data with facial analysis and use Convolutional Neural Networks (CNN) and attention mechanisms to preserve temporal features from facial frames while extracting spatial features from EEG signals [11]. The integration of EEG and ECG data is enhanced by 1D segmentation models, such as MultiResUNet3p, which effectively suppress ocular artifacts [12]. Heart Rate Variability is extracted from ECG recordings, which provides a robust physiological substitute for Autonomic Nervous System (ANS) regulation. Advanced feature extraction methods, such as Amplitude Level Quantization (ALQ), enable the identification of complex affective states by capturing local fluctuations across different frequency bands [13]. The integration of a 1D CNN with Vision Transformer (ViT) allows robust feature extraction from multimodal physiological signals [14]. Transformer-based architectures support self-supervised pre-training by transforming 1D physiological data into 2D scalograms [15]. Empirical evaluation confirms that LSTM is uniquely suitable for modeling the sequential dynamics of physiological time-series data [16]. The inherent difficulty in discriminating between adjacent emotional states with feature distribution is addressed using mutual attention and dense max margin loss functions [17]. The development of a low-complexity sequential architecture that functions with preprocessing pipelines significantly improves the translation utility of affective computing [18]. Hierarchical fusion frameworks use a gated attention mechanism to assign weights to different data sources adaptively [19]. Despite its potential, physiological signal acquisition is hindered by hardware heterogeneity and the requirement for specialized preprocessing pipelines [20].

Emotion recognition from physiological time-series remains difficult due to these three factors. First, a multi-order amplitude scale mismatch exists between scalp EEG signals operating at the microvolt (μV) level and ECG signals operating at the millivolt (mV) level. Second, Short-Time Fourier Transform (STFT) fixes time-windowing which causes severe spectral blurring and suppressing transient neural oscillations. Third, existing temporal attention mechanisms lack explicit frequency awareness, allowing low-frequency muscle artifacts to distort temporal attention weights across subjects. To address these challenges, SpecTFT-DSDNet architecture is introduced by providing three solutions. First, synchrosqueezed dynamic spectral rotation is applied to the architecture as Synchrosqueezed Dynamic Short-Time Fractional Fourier Transform (SSQ-DSTFFT) to rotate time-frequency representations at segment-optimal fractional orders α^* by improving spectral energy sharpness by 68.6% over existing STFT and by resolving fixed-window blurring. Second, frequency-aware spectral modulation is used with a learnable spectral modulation block (α_{mod}) which is integrated between stacked LSTM context layers and masked multi-head attention by applying 1D spectral transformations, which force temporal attention mechanisms to focus on emotionally salient Alpha (8-13 Hz) and Beta (13-30 Hz) rhythms. Third, gated multi-scale temporal fusion is implemented with an overlapping sliding-window segmentation strategy (2.0-second window, 50% overlap), which is coupled with Temporal Fusion Transformer

(TFT), Variable Selection Networks (VSN) and Gated Residual Networks (GRN) to filter out subject-specific physiological noise and model long-range affective transitions.

Existing multimodal Dynamic Spectrum Driven networks (DSDNet) struggle to simultaneously capture non-stationary, frequency-specific, and long-range temporal dynamics of physiological signals. To address these issues, this work develops an efficient, frequency-aware, and temporally robust model that integrates dynamic spectral transformation with gated temporal dependency by achieving state-of-the-art multimodal emotion recognition performance while maintaining computational feasibility.

Changyi Yu et al. [21] employed an Attention-based Inception Capsule Network (AI-CapsNet) designed for EEG motion recognition by integrating spatial, frequency, and temporal features. Raw EEG signals were segmented and filtered into four frequency bands to extract Differential Entropy and Power Spectral Density features. These features were converted to a 3D matrix which is relative to electrode to preserve spatial relationships, which were processed through a coordinate attention mechanism followed by an improved inception module. Finally, a two-layer capsule network with dynamic routing was used to classify the signals into emotional dimensions, such as valence and arousal. However, AI-CapsNet lacks evaluation in real-world scenarios and unverified generalization across different subjects. Kun Chen et al. [22] introduced Hierarchical Attention with 3DCNN CapsNet (HA-CapsNet) for EEG emotion recognition. After extracting features from preprocessing, it undergoes local attention to weight individual channels and intra channel sub bands before being mapped into 2D matrices and stacked into a 3D cube. This cube was refined through Global Attention and fed into a 3D CNN layer to deep spatial structures. Finally, a capsule network is used for the dynamic routing performance in emotional state classification. However, the HA-CapsNet model struggled with cross-subject recognition, which led to significant individual physiological variances. Tianqi Fan et al. [23] employed a Deep CNN, which was incorporated with an improved Convolutional Block Attention Module for emotion recognition. The preprocessed ECG signals were fed into a six-layer CNN backbone using large convolutional kernels and a residual structure to extract deep features while avoiding network degradation. The CBAM adaptively weights these features through channel attention and spatial attention within channels. Finally, the fully connected layer classifies the weighted features into emotional states. However, excessive kernels caused the model to lose detailed information, leading to a decrease in performance.

Zhuozheng Wang and Yihan Wang [24] developed an attention 1D CNN with Gated Recurrent Unit (Att-1DCNN-GRU), which is a fusion of physiological signals for emotion recognition. Feature extraction was performed in the time and frequency domains, and nonlinear functions were used by Random Forest to select the most discriminative features. These refined features were entered into the 1D CNN to capture local features, which were then processed by the GRU to learn temporal dependencies. An attention mechanism weight was an important time stamp before the Softmax layer, which classified the output into Valence and Arousal. However, the Att-1DCNN-GRU model was limited by its sensitivity to individual differences and lacked validation across more diverse environmental contexts. Priyadarshini Samal and Mohammad Farukh Hashmi [25] demonstrated a method for multiple emotion recognition for physiological signals named Dynamic Spectrum Driven network (DSDNet). Dynamic time frequency analysis was completed, which was also referred to as SSQ-DSTFFT, which integrates the synchrosqueezed transform with short time fast fractional fourier transform. This method adaptively extracts features using a dynamic fractional window to better accommodate the non-stationary nature of physiological signals. The resulting spectrograms from both modalities were combined through feature-level fusion and then classified into valence and arousal using an Efficient Convolutional Neural Network (ECNN). However, DSDNet struggled with extensive inter-subject inconsistency in emotion data. Yi Wei et al. [26] demonstrated a multimodal framework designed for emotion recognition called Transformer Capsule Network (TC-Net). EEG signals undergo a process called baseline removal to produce time frequency feature maps. These maps then enter the EEG transformer module, which was used for the EEG Patch Merging technique for local feature extraction, while transformer blocks captured global contextual dependencies across the temporal and frequency domains. Finally, the emotion capsule module used vector neurons and dynamic routing to characterize the spatial features for classification. However, the TC-Net method failed to detect the emotions of multiple features.

Divyashikha Sethia and Neha Gahlan [27] employed a security based method for affective computing. Features were extracted and fused into a single vector, which was then processed by a transformer. This transformer dynamically weights important features to reduce the variation heterogeneity. Locally trained weights were sent to a central server, where scaled weighted federated averaging was performed to collect updates while addressing the imbalanced data distribution between participants. However, the employed model was not evaluated using binary classification and did not capture the full complexity of human emotional signs compared to continuous models. Lia Farokhah et al. [28] developed a simplified 2D CNN with Channel Selection using EEG signals for emotion recognition. The computational complexity is decreased through channel selection method which was evaluated individually through channel performance using the Densenet-201 model by selecting the top 10 most informative channels. These features were then fed into a simplified classification layers to classify emotions into valence and arousal categories. However, the developed model faced issues with generalizability and real-world recording conditions. Xinyi Zhang et al. [30] implemented TPRO-NET model to capture subtle emotional changes for emotion recognition using EEG signals. The implemented model captured nuanced and minuscule emotional changes unlike the traditional model that simplify emotional dimensions. In addition, TPRO-NET model utilized both standard differential entropy (DE) features and enhanced DE features, which are complementary feature sets and boosted the classification performance. The algorithm predicted multiple emotional dimensions in a single run but TPRO-NET's performance was average when tested in subject-independent experiments. This means the model struggles to infer the emotions of unknown individuals using the data of known subjects. The comparative evaluation table for the existing models is mentioned in Table 1.

Table 1: Comparative Evaluation Table for the Existing Models

Method	Input Modalities	Preprocessing	Spectral Mapping	Temporal Model	Complexity	Evaluation Protocol
AI-CapsNet [21]	EEG	Bands Filtering	3D Matrix	Capsule Routing	High	Subject-Dependent
HA-CapsNet [22]	EEG	Channel Selection	3D Cube Mapping	Global Attention	Max	Subject-Independent
Deep CNN [23]	ECG	Resubmission FIR	Large Kernels	6-Layer CNN Back	Med	Intra-Subject Split
DSDNet [25]	EEG + ECG	Notch + Filtering	SSQ-DSTFFT	Efficient CNN	Low	Cross-Subject Pool
Proposed	EEG + ECG	Fast ICA + Z-Score	SSQ-DSTFFT + Learnable Spectral Modulation	TFT with Masked MHA	Moderate	Group 5-Fold Isolation

Despite the high performance, automated emotion recognition models struggle with inter-subject variability, low-frequency noise, spectral blurring from fixed time-window limits, and amplitude scale mismatches between microvolt EEG and millivolt ECG signals. To resolve this problem, this study hypothesizes that combining dynamic fractional spectral rotation with learnable spectral modulation and gated temporal fusion suppresses physiological variance to enable calibration-free cross-subject decoding. Specifically, this research addresses three core questions examining how dynamic fractional-order spectral rotation eliminates fixed-window blurring, how learnable spectral modulation and gated temporal fusion filter physiological noise and scale mismatches, and whether a frequency-aware temporal transformer achieves cross-subject accuracy exceeding 98% across international benchmarks while maintaining inference latency below 15ms. The main contributions of the proposed Spectral Modulated Temporal Fusion Transformer Enhanced Dynamic Spectrum Driven Network (SpecTFT-DSDNet) are as follows:

- Enhancing the temporal Fusion Transformer by adding a learnable spectral modulation block between the LSTM encoder-decoder and masked attention layer. This makes the transformer explicitly frequency-aware, allowing it to focus on emotionally salient rhythms and long-range dependencies while suppressing irrelevant noise through the GRN and VSN.
- The proposed SpecTFT-DSDNet model introduces an overlapping sliding window segmentation process with 50% overlap, which addresses the non-stationary nature of

physiological signals by preventing the loss of transient emotional signals at the window boundaries and allowing the model to capture gradual emotional transitions more effectively.

- The proposed model introduces an overlapping sliding-window segmentation strategy coupled with a Temporal Fusion Transformer (TFT). This integration builds upon a baseline SSQ-DSTFFT spectral mapping using Variable Selection Networks (VSN) and a learnable cross-modal spectral-modulation layer to dynamically weight and optimize feature extraction across temporal scales.

In the following manuscript, the proposed model is explained in Section 2, the experimental results are presented in Section 3, and Section 4 concludes the paper.

2. RESEARCH METHOD

This section details the experimental methodology and signal processing approach developed for evaluating the proposed SpecTFT-DSDNet model across the three benchmark datasets. The model processes a raw multimodal physiological inputs using structured spectral and temporal feature transformations before executing subject isolated affective classification trials.

2.1 Research Design

This study adopts an experimental research design focusing on cross-validation to evaluate cross-subject model generalizability without data leakage. Cross-validation approach follows methods which are mostly used in areas such as signal processing, emotion recognition and neuroscience. The design aligns with cross-subject evaluation frameworks utilized across diverse geographical regions (including European DEAP/AMIGOS and international DREAMER benchmark cohorts), applied industries (ranging from clinical neurological diagnostics and autonomic healthcare monitoring to adaptive human-computer interaction and neuro-ergonomics), and multidisciplinary domains (combining digital signal processing, neurophysiology, and deep neural modeling). By using 5-fold cross-validation method with three different international datasets, experimental design evaluates carefully and mitigates inter-subject biological variance.

2.2 Overview of Proposed Method

The proposed SpecTFT-DSDNet model integrates dynamic spectrum transformation with a Temporal Fusion Transformer to effectively capture the spatial, spectral, and temporal dynamics in physiological signals. These signals exhibit complex non stationary behavior, where emotional responses appear as localized spectral variations and temporal rhythm changes. Figure 1 shows the complete proposed SpecTFT-DSDNet architecture. Therefore, the SpecTFT-DSDNet architecture is designed to progressively transform raw physiological signals into informative spectral representations and subsequently learn temporal dependencies through transformer-based modeling. The overall architecture consists of six major stages: physiological signal acquisition and overlapping sliding-window segmentation, dynamic spectrum transformation, Efficient CNN-based feature encoding, temporal fusion transformer-based dependency learning, spectral modulation, and interpretable attention-based emotion classification.

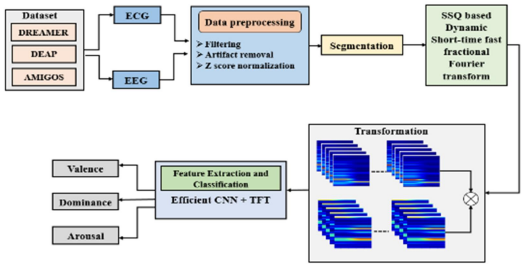


Figure 1: Overall Architecture of Proposed SpecTFT-DSDNet Model for Emotion Recognition.

2.3 Dataset Description

The DEAP [21] dataset is a widely used multimodal benchmark for human affective state analysis. It includes 32 healthy participants who viewed 40 one-minute music video extracts, which are selected to produce a broad range of emotional responses. Physiological signals are recorded using the Biosemic Active Two system, capturing 32-channel EEG at a sampling rate of 512 Hz and eight peripheral physiological signals.

The DREAMER [25] dataset is specifically designed to evaluate the efficiency of low-cost, wireless, off-the-shelf sensors for emotion recognition. It contains recordings from 23 participants exposed to 18 audiovisual movie clips of varying lengths (65–393 sec). Unlike high-density medical systems, DREAMER uses a 14 channel Emotive EPOC EEG headset (128 Hz) and a 2 channel Shimmer2 ECG sensor (256 Hz). Following each stimulus, participants reported their emotional states using the Self-Assessment Manikin for valence, arousal, and dominance on a 5 point likert scale.

The AMIGOS [25] dataset extends beyond individual emotion recognition by incorporating the social context and long-term stimuli. Data were collected from 40 participants across two different experiments. The first involved 16 short video clips (<250 s), and the second involved four long-duration video clips (>14 min). Physiological data consists of 14 channel EEG, ECG, and GSR (Shimmer sensors). A distinguishing feature of AMIGOS is the inclusion of personality traits, where initial mood states such as self-reported and external observer annotations for valence and arousal on a 9-point scale. In addition to dataset, the distribution of samples across emotional classes is analyzed to ensure balanced representation. The inter-subject variability and signal heterogeneity are considered during evaluation to provide realistic evaluation of model performance across different individuals.

The continuous user ratings are mapped into distinct classification targets to evaluate performance. For 1-9 scales in DEAP and AMIGOS dataset, a median split threshold locked at 5.0 categorizes scores into binary low (under 5.0) and high (5.0 and above) classes. For the 1-5 scale in DREAMER, the integers define a 5-class multi-classification baseline. The segment-level sample distributions are systematically extracted across all three datasets following the 2.0-second overlapping sliding-window segmentation to verify the model's high classification accuracy. For DEAP and AMIGOS datasets, the class splits are high balanced. For the 5-class multi-classification task in DREAMER, the extracted windows are distributed near-equally across the integer class.

2.4 Physiological Signal acquisition and Data Preprocessing

The process starts with collecting raw multimodal physiological signals from the DEAP, DREAMER and AMIGO datasets as presented in Equation (1), where T is the total discrete samples from each number of samples t. Physiological signals are naturally non stationary by varying over time due to fluctuating emotional states. Therefore, analyzing the entire a signal as single sequence leads to loss of transient emotional patterns. Before segmentation the raw physiological signals underwent multi stage preprocessing to ensure signal integrity and model robustness. The first filtering process is applied to eliminate power line interference and high frequency noise using a 50 Hz notch filter with a bandpass filter (0.5-45 Hz for EEG and 0.5-40 Hz).

The signals are aligned across a uniform absolute temporal window duration of $\Delta\tau = 2.0$ seconds. All physiological channels are synchronized to a target frequency of 512 Hz to harmonize mismatched native sampling rates across modalities and datasets. In particular, the 128 Hz EEG signals from the DREAMER dataset are upsampled to 512 Hz and the 256 Hz ECG signals from DREAMER dataset are upsampled to 512 Hz, utilizing a polyphaser filter structure with anti-aliasing configurations. The native 512 Hz rate using matching anti-aliasing FIR parameters without rate modification, which establishes a uniform synchronized target rate of 512 Hz across all validation modalities. To preserve emotionally relevant rhythms such as Alpha, Beta, and Theta while removing low frequency drifts. Second, Independent Component Analysis (ICA) is used for artifact removal by decomposing the observed signals $x(t)$ into statistically independent components $S(t)$ as provided in Equation (2), where A is the mixing matrix. Components corresponding to non-cerebral sources such as ocular and muscle noise are identified and set to zero to reconstruct for better signal. The independent components showing a 0.5 – 2.0 Hz peak with an absolute kurtosis $K > 5.0$ are flagged as ocular noise and set to zero during reconstruction. The FastICA algorithm is configured with a parallel solver, a convergence tolerance of $1e-4$, and a fixed random seed of 42 for reproducibility. Finally, to solve inter subject variability and differences in signal amplitude Z score normalization is applied to each channel according to Equation (3), where x_i is the raw sample, μ_i is the mean and σ_i is the standard deviation of the i^{th} channel.

$$S(t) \in \{EEG, ECG\} \quad (1)$$

$$x(t) = As(t) \quad (2)$$

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (3)$$

2.5 Overlapping sliding-window segmentation

To address this by capturing localized variations in emotional states, the continuous signal is partitioned into N overlapping segments. This segmentation process is executed by using a sliding window technique with a fixed window duration of 2.0 seconds ($L = 1024$ samples at 512 Hz) and 50% overlap ratio (1.0-second step size). This window duration strikes the optimal balance between spectral resolution and computational throughput from a signal processing perspective. Window durations shorter than 2.0 seconds induce spectral blurring that degrades low-frequency Theta

band resolution. In contrast, the fast emotional shifts get erased if the window is longer than 2.0 seconds. The tests from 1.0 to 4.0 seconds ensured that a 2.0 second window gives competitive performance and dropping the overlap below 50% breaks up the signal continuity. Each sliding window is defined by a specific, uniform length denoted as L as shown in Equation (4).

$$X = \{x_1, x_2, \dots \dots x_N\} \quad (4)$$

Here, $x_i \in R^L$ denotes the i^{th} signal segment, and N is the total segment count. A 50% overlap is maintained to ensure temporal continuity and prevent the loss of transient emotional triggers at window boundaries. Overlapping segmentation also improves temporal continuity and allows the model to capture gradual changes in emotion more effectively. This segmented representation provides a structured input format for subsequent spectral analyses and deep feature extraction. To synchronize multimodal signals with mismatched sampling rates, the model uses a two-step alignment process. Signals are divided for temporal windows for constant duration rather than a fixed number of samples, ensuring that each segment represents the same physical time interval.

2.6 Dynamic Spectrum Transformation

Each segment x_i is transformed into a spectral map F_i using the SSQ-DSTFFT technique. Unlike standard time-frequency transforms, which suffer from spectral blurring, this method sharpens the signal representation by rotating it in the time-frequency plane at the optimal order. This property enables more accurate localization of oscillation components associated with emotional responses. The fundamental factor of this transformation is rotating the signals in the time-frequency plane at an optimal fractional order α^* as provided in Equation (5). Here, α^* is optimized for each individual signal segment by minimizing fourth-order Renyi Entropy instead of being fixed and subject averaged. This calculates the transformation rotation angle ($\phi = \frac{\alpha^* \pi}{2}$). The method scores 0.86 on sharpness and makes it 68.6% clearer than standard STFT (0.51) and yields sharp map $S_i(t, \omega) = \mathcal{S}\{X_\alpha(u)\}$. Here $\mathcal{S}$ is the synchrosqueezing operator applied to the fractional distribution to eliminate spectral blurring and achieve a sharp energy concentration.

$$X_\alpha(u) = \int_{-\infty}^{\infty} x_i(t)\, K_\alpha(t, u) dt \quad (5)$$

$$K_\alpha(t, u) = \sqrt{\frac{1 - i \cot \phi}{2\pi}} \exp\left(i \frac{t^2 + u^2}{2} \cot\phi - iut \csc \phi\right) \quad (6)$$

The resulting maps are fused into a unified representation as shown in Equation (6), where ω_i is the learnable fusion weight as shown in Equation (7). The multimodal fusion step combines complementary physiological information and improves the representation of emotion.

$$F_{spec} = \sum \omega_i S_i \quad (7)$$

The modality-specific spectral maps S_i are dynamically weighted and fused at spectral level as defined in Equation (7) to form the unified spectral representation F_{spec} to capture cross-modal dynamics prior to spatial filtering. This fused representation is passed to the efficient 2D CNN block in Section 2.7 for localized spatial-spectral feature encoding. Mismatched sampling rates are synchronized using polyphaser interpolation filters to align temporal boundaries across modalities prior to spectral transformation. The learnable fusion weights (ω_i) in Equation (7) are shared across all subjects within each respective dataset during training to capture dataset-wide modal importance but they are learned separately for each dataset because models for DEAP, DREAMER and AMIGOS are trained independently.

2.7 Efficient CNN Feature Encoding

The fused feature is processed by an Efficient CNN to extract spatial spectral features through a convolutional operation using Equation (8), where $W_{m,n}$ is the $k \times k$ kernel and $\sigma(\cdot)$ is the ReLU activation. By loading multiple convolutional layers, the network moves from detecting simple edges in the spectrogram to identifying complex high-level patterns. This step treats the spectral map as a structured image. where the parallel axis shows time and the perpendicular axis is frequency. To prepare data for temporal analysis, the resulting high-dimensional feature maps are flattened into a sequence of d dimensional temporal embeddings $H = \{h_1, h_2, \dots \dots, h_N\}$, where each vector $h_t \epsilon R^d$ provides the compressed latent encoding of the signal spectral state at a specific time step. These embedding's bridge spatial feature extraction and sequential modeling, providing the temporal Fusion Transformer (TFT) with a structured, feature-rich input sequence that preserves both the spectral integrity and chronological order of the original signal.

$$F(i, j) = \sigma\left(\sum_{m,n} W_{m,n} M_{spec}(i + m, j + n) + b\right) (8)$$

The convolution operation performs localized filtering across the spectral map, allowing the network to identify outlines, such as spectral boundaries, ridges, and energy peaks. Convolutional layers are deep features that evolve into higher-level representations that capture complex spectral structures and cross-channel interactions. After convolutional processing, the resulting feature maps are flattened and converted into a sequence of temporal embedding's. Each embedding vector represents the latent representation of the spectral characteristics in a particular time segment. This embedding's maintain both the spectral structure and chronological ordering of the signal, enabling effective sequential modeling in the next step.

2.8 Temporal Fusion Transformer (TFT) for Temporal Dependency Learning

Emotion Recognition from physiological signals requires modeling from short-term variations and long-term temporal dependencies. Conventional sequential models struggle to capture long range dependences due to gradient degradation and limited memory capacity. The proposed SpecTFT-DSDNet uses the TFT [29] architecture to make the model long-range evolution for emotional states which, includes three main steps such as GRN, VSN, and LSTM Encoder Decoder. Unlike standard transformers, the TFT includes a gating mechanism and variable selection to handle inherent noise in physiological data. In TFT, GRN is the fundamental building block that provides adaptive nonlinear processing and allows the model to skip unnecessary transformation based on the input signal complexity. For an input α and static context c, the GRN is defined in Equation (9), where $\eta_1 = W_1\eta_2 + b_1$ and $\eta_2 = ELU(W_2a + W_3c + b_2)$. Here, W_1, W_2, W_3 are learnable weight matrices and b_1, b_2 are biases. The Exponential Linear Unit (ELU) ensures stable gradient flow. The Gated Linear Unit (GLU) acts as a final switch as provided in Equation (10), where $\sigma(\cdot)$ is the sigmoid activation, $\odot$ is the Hadamard product, and γ is the transformed input. This connectivity allows the model to surpass noise at specific time steps.

$$GRN(a,c) = LayerNorm\big(\alpha + GLU(\eta_1)\big) \quad (9)$$

$$GLU(\gamma) = \sigma(W_4\gamma + b_3) \odot (W_5\gamma + b_4) \quad (10)$$

This gating structure acts as an adaptive filter that suppresses irrelevant noise while preserving the emotionally relevant signal patterns. To manage the high dimensionality of multimodal features, the VSN identifies the most salient latent features at each time step t. It produces the importance weights v_t as shown in Equation (11). The filtered features $\widetilde{h_t}$ are then obtained by weighting the individual feature projections, as depicted in Equation (12), where $v_t^{(j)}$ is the contribution of the j^{th} feature. This ensures that the model prioritizes relevant channels during emotional spikes. The individual features within the embedding are denoted as $h_t^{(j)}$ where j ranges from the starting index 1 to the end index d.

$$v_t = Softmax(GRN_v(h_t, c) \quad (11)$$

$$\widetilde{h_t} = \sum_{j=1}^{d} v_t^{(j)} \cdot GRN_\phi\left(h_t^{(j)}\right) \quad (12)$$

The sequence $\tilde{h}_t$ is passed through an LSTM Encoder-Decoder to capture the recurrent sequential context. This layer preserves the local temporal order that the attention mechanism overlooks, as expressed in Equations. (13) and (14), where h_t^{enc} and h_t^{dec} are the hidden states that encode the historical evolution of the physiological response up to time t. To provide temporal context, the decoder receives future known inputs $\phi_{t+1:t+\tau} = \{\phi_{t+1}, \ldots \ldots \phi_{t+\tau}\}$, where ϕ_{t+1} is the immediate next step, and $\phi_{t+\tau}$ is the end of the prediction horizon. This recurrent structure preserves local temporal relationships and models the evolution of the emotional response over time.

$$h_t^{enc} = LSTM\big(\tilde{h}_t - h_{t-1}^{enc}\big) \quad (13)$$

$$h_t^{dec} = LSTM(h_t^{enc} - h_{t-1}^{dec}) \quad (14)$$

The role of LSTM encoder-decoder within TFT layer is to preserve step-by-step chronological order. The multi-head attention layers track long-range patterns and they can overlook localized and chronological context. The stacked LSTM network serves as a sequential anchor that tracks emotional responses smoothly over time. Input arrays initialized at [B,32,1024] are transformed using SSQ-DSTFFT into 2D maps of shape [B, 32,128,64]. Following spatial-spectral flattening and linear projection, the resulting latent embedding's of shape [B, 64, 128] sequentially processed by the LSTM to produce hidden tensor H_{dec}. Next, a1D FFT maps the sequence into a complex frequency representation $F_{freq} \in \mathbb{C}^{[B,64,128]}$. A 1D IFFT returns the modulated features into a time-domain tensor $M_{specmod} \in \mathbb{R}^{[B,64,128]}$ for final masked attention filtering and softmax classification after element-wise multiplication with learnable weight vector α_{mod}. The structure of temporal fusion transformer is represented in Figure 2.

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

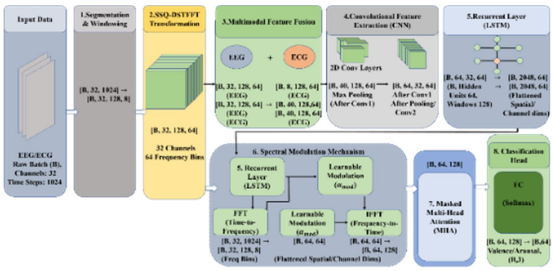


Figure 2: Multi-Dimensional Tensor Data-Flow and Architectural Schema of the Proposed SpecTFT-DSDNet

2.9 Spectral Modulation

Although TFT architecture effectively captures temporal dependencies, it does not explicitly incorporate frequency-specific information, which is important for emotion recognition. To address this, Spectral Modulation is introduced. A connectivity bridge between the sequential decoders, attention layer, and Spectral Modulation ensures that the transformer is frequency-aware. The decoded sequence H_{dec} is projected into the frequency domain using the FFT, as shown in Equation (15). The learnable spectral weight α is applied to amplify emotion-specific rhythms, as expressed in Equation (16). The sequence is then returned to the time domain through Inverse FFT (IFFT), as shown in Equation (17), which forces the attention mechanism to focus on rhythmically significant temporal events. Figure 3 illustrates the structure of the Temporal Fusion Transformer, which is enhanced with Spectral Modulation.

$$F_{freq} = \mathcal{F}(H_{dec}) \in \mathbb{C}^{N \times d_{model}} \quad (15)$$

$$F_{mod} = \alpha_{mod} \odot F_{freq} \quad (16)$$

$$M_{specmod} = \mathcal{F}^{-1}(F_{mod}) \in \mathbb{R}^{N \times d_{model}} \quad (17)$$

Analysis of the learnable spectral weights shows that the model assigned higher importance to the alpha (8-13Hz) and Beta (14-30Hz) frequency bands. This demonstrates that the spectral modulation block prioritizes the neurophysical rhythms most associated with emotional arousal. This mechanism provides a transparent link between transformers attention and biological signal characteristics.

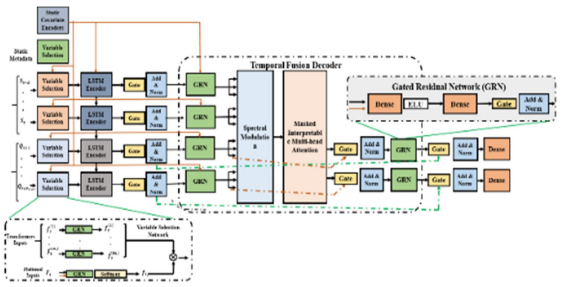


Figure 3: Structure of Temporal Fusion Transformer enhanced with Spectral Modulation.

2.10 Masked Interpretable Attention

The modulated features $F_{specmod}$ are processed by the Masked Interpretable Multi-Head Attention layer, as shown in Equation (18), where $Q\ (Query), K\ (Key) and\ Value\ (Value)$ are linear projections $Q = HW^Q, K = HW^K, V = HW^V$ where $W^Q, W^K, W^V \epsilon R^{d_{model} \times d_k}$ are weight matrices of the input and d_k is the scaling factor. Dividing $\sqrt{d_k}$ scales the variance of the dot product back to 1 when the projection dimension d_k is large. This prevents from pushing into regions of the softmax function with extremely small gradients, so avoiding the vanishing gradient. A casual mask is applied to prevent future segments from influencing earlier predictions, ensuring that the model maintains temporal consistency during emotion recognition.

$$Attention\ (Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (18)$$

2.11 Classification

The final emotion prediction is generated by softmax activation using Equation (19), where z_i is the logit for class i and C is the number of emotional categories. Algorithm is described for SpecTFT-DSDNet method enables the simultaneous prediction of Valence, Arousal, and Dominance. Table 2 describes the notation list used in the equations for proposed method.

$$P(y|x)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (19)$$

Algorithm: SpecTFT-DSDNet

Step 1: Data Preparation and Pre-processing

- Identify the input modalities, specifically EEG and ECG signals using (1).
- Apply ICA to decompose signals into independent components using (2).
- Apply Z score normalization to each channel to handle inter subject variability using (3).
- Segment the continuous signal into N overlapping windows of length L using (4).

Step 2: Feature Extraction and Encoding

- Convert segments into spectral maps using fractional kernel using (5).
- Integrate complementary signals into fused representation using (6).
- Extract spatial-spectral patterns using convolutional kernel and ReLU activation using (7).

Step 3: TFT Modeling

- Use the GRN to skip unnecessary transformation using (8).

- Act as switch to suppress noise using Hadamard product using (9).
- Determine importance weights for latent features at time using (10).
- Apply weights to feature projections using (11).
- Capture local temporal order through LSTM layers using (12) and (13).

Step 4: Spectral Modulation and Attention

- Transform decoded sequences through FFT using (14)
- Amplify emotion-specific rhythm with weight using (15).
- Return to time domain through Inverse FFT using (16).
- Apply MHA to focus on significant temporal events using (17).

Step 5: Output

- Classify by generating probability for emotional categories through softmax using (18).

Table 2: Notation used for SpecTFT-DSDNet Model

Symbol	Description
$S(t), x(t)$	Physiological input signals (EEG and ECG) and observed signal vector
$s(t)$	Statistically independent source components
$S(i)$	Sharpened time-frequency spectral map to the i^{th}modality
$\mathcal{S}$, A	Synchrosqeezing transform operator and Mixing matrix
Z_i	Z-score normalized sample for the i^{th}channel
L, N	Length of the sliding window for segmentation and Total number of signals segments
X	Set of signals segments $\{x_1, x_2, \ldots . x_N\}$
α	Optimal fractional order for signal rotation
$K_\alpha(u, v)$	Fractional kernel for the SSQ-DSTFFT transformation
M_{spec}	Unified fused spectral representation
$W_{m,n}$	Convolutional kernel weights $(k \times k)$
$\sigma(\cdot)$	ReLU activation function
H	Sequence of temporal embedding's $\{h_1, h_2, \ldots h_N\}$
a, c	Input signal and static context for the GRN
γ	Transformed input for GLU
v_t	Importance weights for latent features at time t
h_t^{enc}, h_t^{dec}	Hidden states for the LSTM Encoder and Decoder
$\mathcal{F}, \mathcal{F}^{-1}$	FFT and Inverse FFT
Q, K, V	Query, Key and Value projections for MHA
d_k	Scaling factor for the attention mechanism

3. RESULTS AND DISCUSSION

The SpecTFT-DSDNet method for multimodal emotion recognition is implemented using a high performance hardware and software stack. The setup features such as NVIDIA GPU with 16 GB VRAM, 32 GB RAM, and multi-core CPU. This configuration serves to accelerate intensive operations such as SSQ-DSTFFT, spectral modulation, and transformer-based attention.

It is developed in Python 3.8 and the architecture is leveraged PyTorch 2.0 for its flexible handling of GRN, VSN, and LSTM components. The essential signal process uses mixed precision and the Adam optimizer employs to efficiently process the DEAP, DREAMER, and AMIGOS datasets. This optimized configuration ensures reproducibility and enables the model to achieve performance through efficient training and inference. The proposed model is trained using Adam optimizer with starting learning rate of $1e-4$. To refine optimization, a StepLR scheduler halves the learning rate every 20 epochs and set batch size to 32. The training process spans up to 200 epochs by incorporating early stopping to mitigate overfitting. To further increase generalization, 0.3 drop rate is integrated into the LSTM and GRN layers. The Subject-Independent cross-validation protocol is implemented to prevent data leakage, all signal segments belonging to a specific participant which are kept together. To ensure reproducibility, the core architecture parameters of the hybrid network are standardized. The input sequence length is set to steps to capture long-range dependencies. The embedding's dimension is configured to channels across all projection layers. For the temporal modeling stage, a layer deep LSTM network with a hidden state size of units is deployed. The subsequent multi-head attention mechanism uses parallel attention heads, where the projection dimensionality is for each head. A strict subject-independent 5-fold cross-validation is used, where data is split by person so that a participant's signals are kept strictly together in a single group. An inner subject-independent validation split is implemented to prevent overfitting and leakage during hyper parameter tuning. Here, the active training subjects in each outer fold are partitioned into an inner training set (80% of training subjects) for weight fitting and inner validation set (20% of training subjects) to guide early stopping and hyperparameter selection. Folds are partitioned using dataset-specific subject counts rather than a uniform 80%/20% outer split because the total participant pools differ across datasets. The DEAP data with 32 subjects utilizes 4 folds with a 26/6

train/test subject split and one fold with a 24/8 split. DREAMER dataset with 23 subjects utilizes three folds with an 18/5 split and two folds with a 19/4 split. Preprocessing parameters for Z-score normalization are calculated based on active training subjects and applied to the isolated test fold to guarantee zero information leakage, with core network configurations detailed in Table 3.

Table 3: Experimental Structural Configuration Parameters

Pipeline Component Block	Target Mathematical Symbol	Exact Implemented Configuration Setting
Segmentation Layer	Window Length $\boldsymbol{L}$	2.0 seconds ($\boldsymbol{L = 1024}$) samples / 50% overlap
Efficient 2D CNN	Kernel Space / Stride / Layers	$\mathbf{3 \times 3}$ spatial filters / Stride 2/3 Hidden layers
Latent Vector Space	Embedding Dimension $\boldsymbol{d_{model}}$	128 dimensions
Recurrent Stack	Depth / Hidden state dimension	2 stacked layers / 128 hidden memory units
Spectral Modulation	Vector size $\boldsymbol{\alpha_{mod}}$ / Domain	Array length = 64
Attention Matrix	Heads $\boldsymbol{H}$ / Dimension $\boldsymbol{d_k}$	$\boldsymbol{H = 4}$ parallel heads

3.1 Evaluation Metrics

Accuracy represents the percentage of total emotional states that are accurately predicted by the model, and it is formulated in Equation (20). Here TP is true positive, TN is true negative, FP is false positive and FN is false negative. Precision measures the ability of the model to not label the negative emotional state as positive. It is important for ensuring when the model detects an emotion and it is computed as in Equation (21). Recall is the ability of the model to find all relevant emotional states in the data and ensure that the model does not miss true emotional spikes, and it is computed using Equation (22). F1-Score is the average of Equations. (21) and (22), which is helpful in emotion recognition, where datasets have imbalanced classes, and it is formulated as shown in Equation (23).

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \quad (20)$$

$$Precision = \frac{TP}{TP+FP} \quad (21)$$

$$Recall = \frac{TP}{TP+FN} \quad (22)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recal} \quad (23)$$

Kappa measures the agreement among the model predictions and actual labels for the possibility of occurrence by chance. A higher kappa indicates that the model is learning correct patterns rather than just guessing. The metric is formulated in Equation (24), where p_0 is the observed proportionate accuracy and p_e is the probability of random agreement.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (24)$$

3.2 Quantitative Evaluation

The performance evaluation of the SpecTFT-DSDNet method is presented through a bar graph against state-of-the-art (baseline) models. It includes MultiResUNet, Spatio Temporal Transformer (STT) with CNN, Multi-Head Attention Fusion Neural Network (MMHA-FNN), Parallel fusion of Graph Convolution, and LSTM (PGCL). To ensure better comparison, all baseline models are re-implemented and re-trained within the same environment using the identical SSQ-DSTFFT preprocessing and adaptive segmentation. Figure 4 illustrates the DEAP dataset by considering all five key metrics, which highlights the proposed SpecTFT-DSDNet model for improved results. Figure 5 provides a similar comparison analysis for the DREAMER dataset, whereas Figure 6 focuses on the AMIGOS dataset. Across all three figures, the proposed SpecTFT-DSDNet achieves consistently high values, which shows that the model works more efficiently than the other models.

To present an explanation for classification accuracy, confusion matrices include three primary emotional dimensions: Arousal, Dominance, and Valence. Figure 7 (a) shows the confusion matrices for the AMIGOS, which show high diagonal values that indicate class-wise prediction. Figure 7 (b) illustrates matrices for DEAP, and (c) displays the results for the DREAMER dataset, which includes a larger number of predicted classes than the binary or ternary classifications of the other datasets. These figures visually confirm the model's ability to minimize false positives and negatives across different emotional states. All the baseline architectures are re-implemented under an identical training protocol, hyperparameter budget and subject-independent group 5-fold cross-validation protocol. The baselines are integrated into our proposed preprocessing pipeline by granting them direct access to SSQ-DSTFFT and overlapping sliding-window segmentation layers to eliminate data-cleaning bias. This integration is elevated the baseline accuracy by an average of 2.14% because

these advanced components are absent from original designs.

Table 4 represents the complete numerical classification results under the subject-independent protocol. The model maintains narrow standard deviations of ≤ 0.15%, high F1-scores ≥ 98.07% and Kappa values ≥ 0.915, which validates the cross-subject performance. For the 5, class DREAMER dataset, macro-averaged metrics are reported for precision, recall and F1-score.

Table 4: Comprehensive Performance Metrics under Subject Independent Cross-Validation

Dataset / Dimension	Accuracy (μ±σ)%	Precision (μ±σ)%	Recall (μ±σ)%	F1-Score (μ±σ)%	Kappa (μ±σ)
DEAP Dataset					
Valence	98.89 ± 0.12	98.19 ± 0.18	98.09 ± 0.21	98.14 ± 0.15	0.918 ± 0.012
Arousal	98.88 ± 0.14	98.12 ± 0.22	98.02 ± 0.24	98.07 ± 0.19	0.915 ± 0.014
Dominance	98.91 ± 0.11	98.24 ± 0.15	98.15 ± 0.19	98.19 ± 0.14	0.921 ± 0.011
DREAMER Dataset					
Valence	99.28 ± 0.08	98.58 ± 0.11	98.48 ± 0.13	98.53 ± 0.10	0.923 ± 0.008
Arousal	99.41 ± 0.06	98.72 ± 0.09	98.61 ± 0.11	98.66 ± 0.08	0.934 ± 0.006
Dominance	99.52 ± 0.05	98.89 ± 0.08	98.78 ± 0.10	98.83 ± 0.07	0.941 ± 0.005
AMIGOS Dataset					
Valence	98.92 ± 0.15	98.22 ± 0.21	98.12 ± 0.23	98.17 ± 0.18	0.934 ± 0.015
Arousal	99.58 ± 0.07	98.91 ± 0.10	98.82 ± 0.12	98.86 ± 0.09	0.946 ± 0.007
Dominance	99.59 ± 0.06	98.94 ± 0.09	98.85 ± 0.11	98.89 ± 0.08	0.948 ± 0.006

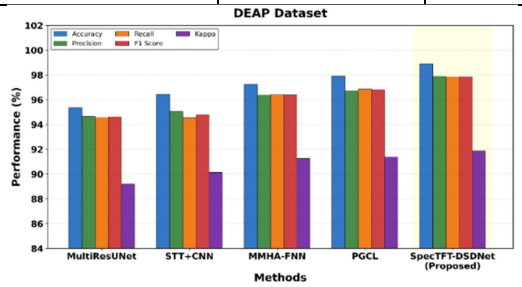


Figure 4: Performance Analysis of SpecTFT-DSDNet Method with Baseline Models for DEAP Dataset.

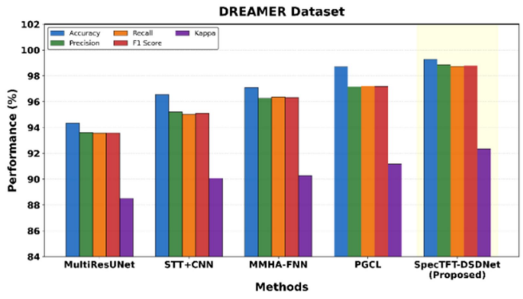


Figure 5: Performance Analysis of SpecTFT-DSDNet Method with Baseline Models for Dreamer Dataset

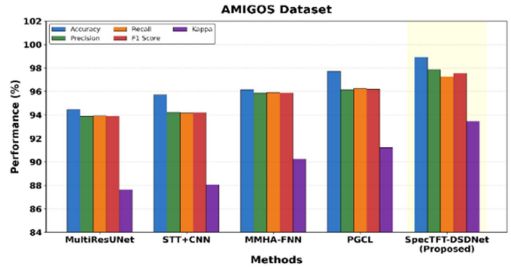


Figure 6: Performance Analysis of SpecTFT-DSDNet Method with Baseline Models for Amigos Dataset

The generalization of SpecTFT-DSDNet is measured using multi stage k- fold cross validation is across all three datasets. Table 5 analyzed for $k\epsilon\{2,3,5,7,9\}$ to evalaute for inter subject variability and potential data imbalances. The results indicate that the model produces optimal results at K=5 for all three datasets by evaluating on valence dimension. This stability across varying folds confirms the effectiveness of integrated spectral modulation and temporal fusion layers in capturing emotional signatures regardless of participant-specific signals. The subject-independent 5-fold cross-validation was strictly enforced across participant boundaries. For 32-subject DEAP dataset, 4 folds contains a 26/6 train/test subject distribution and 1 fold contains 24/8 distribution. For the 23-subject DREAMER dataset, 3 folds contains 18/5 distribution and 2 folds contains 19/4 distribution. The hyperparameter tuning was isolated using an inner validation split within each active training group to prevent data leakage

3.3 Computational Analysis

The proposed SpecTFT-DSDNet model is computationally compared with benchmark architectures such as Shallow CNN, ViT, TFT and Mamba, as shown in Table 6. It is based on four metrics: number of trainable parameters (M), Floating Point Operations (FLOPS) per forward peak, VRAM usage during training for various batch sizes, and training time required per epoch. The proposed SpecTFT-DSDNet model contains 18,426,112 trainable parameters. The system peak

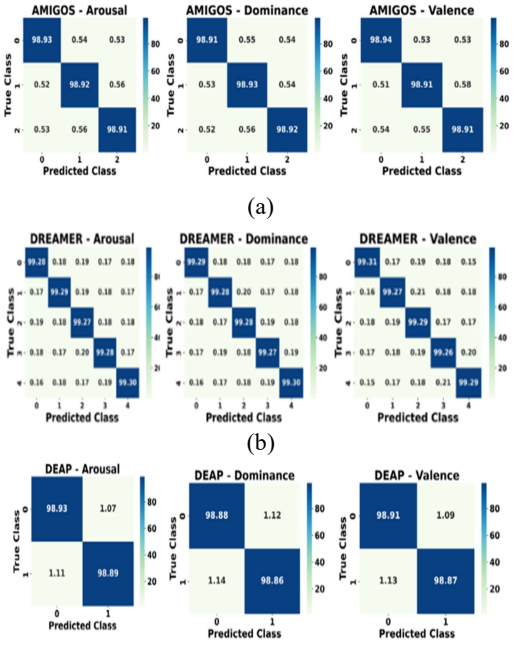
memory tracks at 11.15 GB VRAM despite the intensive operations of SSQ-DSTFFT and spectral modulation block fitting within the physical 16 GB memory ceiling and achieves an inference speed of 12.4 ms per segment as shown in Table 6. To ensure reproducibility, testing was executed on a single NVIDIA GPU with 16 GB VRAM and 32 GB RAM, and a multi-core CPU and PyTorch 2.0. All models are evaluated for inference latency where the proposed method achieved an average inference of 12.4 ms per segment. Table 7 breaks down the isolated module-level latencies across different input segment lengths at a batch size of 32 to evaluate model on low-power hardware. This analysis represents that the primary innovation of spectral modulation block adds a minimal latency overhead of 0.8 ms, while total segment pipeline processing latency of 12.4 ms is obtained.

3.4 Statistical Analysis

To confirm the scientific validity of the results, Table 8 presents a statistical comparison between the SpecTFT-DSDNet model and baseline models. The statistical significance is computed using Nadeau-Bengio Corrected Resampled t-test to account for data overlap across folds. The statistical unit is the mean accuracy per cross-validation split with $K = 5$ observations with 4 degrees of freedom. Rather than assuming a structurally incorrect universal split, Equation (25) employs a dataset specific correction factor (n_{test}/n_{train}=0.25) that represents the mathematical average of actual fold-wise participant distributions: $\rho_{DEAP} = 0.2515$, $\rho_{DREAMER} = 0.2509$ and $\rho_{AMIGOS} = 0.2500$.

$$t = \frac{\frac{1}{K}\sum_{k=1}^{K} d_k}{\sqrt{(\frac{1}{K}+\frac{n_{test}}{n_{train}})\hat{\sigma}^2}} \quad (25)$$

Here, d_k is the pairwise accuracy difference for fold k, Bonferroni multiple-comparison correction tightens the alpha threshold to $\alpha_{adj} = 0.0125$ (four pairwise comparisons, $df = 4$). The statistical testing demonstrates that the proposed model achieved significant improvements over complex architectures such as MultiResUNet and MMHA-FNN across all tested datasets. It also demonstrates competitive and statistically isolated performance against STT+CNN and PGCL in specific validation subsets.



(a) (b) (c)

Figure 7: Confusion Matrix of (a) AMIGOS, (b) DEAP and (c) DREAMER for Arousal, Dominance and Valance

Table 5: K-Fold Cross Validation of the Proposed SpecTFT-DSDNet for DEAP, DREAMER and AMIGOS Dataset for Valence Dimension

	DEAP					DREAMER					AMIGOS				
K-fold	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Kappa (%)	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Kappa (%)	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Kappa (%)
2	95.35	94.65	94.62	94.62	89.20	94.34	93.64	93.54	93.59	88.50	94.46	93.76	93.66	93.71	87.64
3	96.44	95.74	95.64	95.69	90.14	96.55	95.85	95.75	95.80	90.07	95.71	95.01	94.91	94.96	88.07
5	98.89	98.19	98.09	98.14	91.87	99.28	98.58	98.48	98.53	92.35	98.92	98.22	98.12	98.17	93.47
7	97.88	97.18	97.08	97.13	91.21	97.11	96.41	96.31	96.36	90.27	96.16	95.46	95.36	95.41	90.25

9	97.89	97.19	97.09	97.14	91.38	98.71	98.01	97.91	97.96	91.17	97.71	97.01	96.91	96.96	91.22

Table 6: Computational Evaluation of the Proposed SpecTFT-DSDNet against Baseline Models

Model	Parameters (M)	FLOPS (G) per forward pass	PEAK VRAM (GB)-training (batch 32)	Training time per epoch (sec)
Shallow CNN	1.24	0.42	2.15	18.4
DeepConvNet	8.56	3.10	6.80	42.1
ViT	43.12	4.85	14.20	145.2
TFT	24.58	2.10	9.40	88.6
Efficient Multimodal (Mamba)	15.30	1.15	7.60	52.3
SpecTFT-DSDNet (proposed)	18.42	1.84	11.15	10.4

Table 7: Module-Level Execution Latency Breakdown

Algorithmic Block Module	Latency (1.0s)	Latency (2.0s)
SSQ-DSTFFT core transform	2.1ms	3.4 ms
Efficient 2D CNN layer	1.8 ms	2.9 ms
TFT Selection layers (GRN/VSN)	1.1 ms	1.9 ms
Stacked LSTM context layers	1.2 ms	1.7 ms
Learnable spectral modulation	0.5 ms	0.8 ms
Masked multi-head attention	0.9 ms	1.7 ms
Total SpecTFT-DSDNet pipeline	7.6 ms	12.4 ms

Table 8: Pairwise Nadeau-Bengio Corrected T-Test Metrics ($\alpha_{adj} = 0.0125$)

Dataset	Method	Reported t-value	Corrected p (df=4)	Bonferroni Decision $\alpha_{adj} = 0.0125$
DEAP	MultiResUNet	4.812	0.0086	Significant
	STT+CNN	3.945	0.0169	Not Significant
	MMHA-FNN	5.116	0.0069	Significant
	PGCL	3.841	0.0184	Not Significant
DREAMER	MultiResUNet	8.596	0.0010	Significant
	STT+CNN	4.102	0.0148	Not Significant
	MMHA-FNN	6.156	0.0035	Significant
	PGCL	3.960	0.0167	Not Significant
AMIGOS	MultiResUNet	5.314	0.0060	Significant
	STT+CNN	4.672	0.0095	Significant
	MMHA-FNN	5.813	0.0044	Significant
	PGCL	4.146	0.0143	Not Significant

3.5 Ablation Study

Table 9 represents an isolated ablation study with alternative spectral modulation strategies across target modalities and affective coordinates. The results depict that disabling single mechanism, particularly SSQ-DSTFFT, overlapping window segmentation, VSN, GRN and LSTM context causes a direct performance drop. This proves that every module contributes uniquely to the network. In addition, replacing learnable α_{mod} block with constant scalar multiplier and static Gaussian envelopes reduces accuracy across the valence, arousal and dominance dimensions. The dynamic and signal-specific frequency tracking is essential for resolving spectral blurring limitations of baseline DSDNet. The first two rows show that using EEG and ECG separately yields lower performance because of single-sensor limits. By combing EEG and ECG in full SpecTFT-DSDNet model resolves the dimensional tracking confusions. The SpecTFT-DSDNet model address the core limitations of baseline DSDNet such as vulnerability to inter-subject variability and lack of long-range temporal tracking suing three architectural additions as shown:

- An overlapping sliding-window segmentation strategy (50% overlap) helps to capture transient emotional transitions at temporal boundaries.
- A TFT context layer incorporates Gated Residual Networks (GRN) and Variable Selection Networks (VSN) to filter out subject-specific physiological noise.
- A learnable spectral modulation block dynamically weights frequency-specific

rhythms that forces the attention layers to remain frequency-aware of physiological patterns.

Table 9: Ablation Evaluation of Proposed SpecTFT-DSDNet Model based on its Components

Evaluated System Variant Configuration	DEAP Val (%)	DEAP Aro (%)	DEAP Dom (%)	DREAMER Val (%)	DREAMER Aro (%)	DREAMER Dom (%)	AMIGOS Val (%)	AMIGOS Aro (%)	AMIGOS Dom (%)
Modality: EEG Channel Array Only	89.45	88.91	89.04	89.15	89.32	89.90	89.76	89.12	89.24
Modality: ECG Channel Array Only	94.87	94.10	94.22	95.01	94.88	95.30	94.91	94.38	94.52
W/O SSQ-DSTFFT (Standard STFT instead)	97.63	97.11	97.18	97.45	97.52	97.82	97.81	97.15	97.23
W/O overlapping segmentation (0% overlap)	96.18	95.02	95.15	96.22	96.35	96.79	96.42	96.10	96.16
W/O Variable Selection Network (VSN)	98.24	98.10	98.14	98.35	98.41	98.67	98.41	98.29	98.31
W/O Gated Residual Network (GRN)	98.15	97.94	98.02	98.11	98.18	98.40	98.30	98.05	98.12
W/O LSTM Encoder-Decoder Context	98.32	98.11	98.16	98.29	98.36	98.62	98.45	98.33	98.39
W/O Masked Multi-Head Attention	98.40	98.22	98.28	98.39	98.44	98.71	98.50	98.42	98.47
W/O Spectral Modulation Block Entirely	98.72	98.54	98.59	97.92	98.02	98.04	98.58	98.11	98.19
Alternative: Constant Scalar Multiplier	98.75	98.59	98.62	98.10	98.15	98.22	98.64	98.20	98.26
Alternative: Fixed Gaussian Envelope	98.81	98.67	98.70	98.34	98.42	98.54	98.70	98.41	98.49
SpecTFT-DSDNet (Proposed Full Model)	98.89	98.88	98.91	99.28	99.41	99.52	98.92	99.58	99.59

The parameters of the learnable frequency weight vector α_{mod} are extracted over 50 independent runs. This distribution was generated by repeating the subject-independent 5-fold cross-validation over 10 sequential repetitions using random seed 42 as shown in Table 10.

Table 10: Learned Parameter Weight Distributions of the α_{mod} Vector

Neural Oscillatory Rhythm	Target Frequency Band (Hz)	Extracted Weight Mean (μ_w)	Standard Deviation (σ_w)	Structural Target Priority
Delta Waves Array	0.5 – 4.0 Hz	0.124	0.021	Heavily Attenuated
Theta Waves Array	4.0 – 8.0 Hz	0.189	0.034	Baseline Retention
Alpha Cortical Rhythm	8.0 – 13.0 Hz	0.584	0.041	Dominant (Valence Tracking)
Beta Somatosensory	13.0 – 30.0 Hz	0.762	0.053	Peak Focus (Arousal Spikes)
Gamma Waves Array	30.0 – 45.0 Hz	0.141	0.029	Suppressed (Noise Filtered)

3.6 Cross Dataset

To check the applicability of the SpecTFT-DSDNet and ensures generalization standards required for affective computing, a cross dataset validation is performed. This analysis involves training the model on source dataset and evaluating it on target dataset with different configuration and without any domain specific fine tuning. Table 11 presents the significant resilience to domain shift by maintaining average accuracies of 70.49% and 75.59 across all three dimensions. A formal data harmonization protocol was applied to ensure

compatibility during cross-dataset evaluation, where channel matching isolates a common 14-channel subset (F3, F4, AF3, AF4, T7, T8, P7, P8, O1, O2, F7, F8, FC5, FC6) across all sensor caps. In sampling-rate conversion, the signals are resampled to a uniform rate of 512 Hz. Particularly, the 128 Hz DREAMER EEG signals and 256 Hz DREAMER ECG signals are upsampled to 512. The 512 Hz streams from DEAP and AMIGOS are maintained at their native 512 Hz rate, which ensures zero temporal distortion across datasets. The label harmonization process normalizes varying data ranges into a shared continuous space bounded by [-1, +1] through min-max scaling as shown in Equation (26):

$$Y_{norm} = 2 \times \left(\frac{Y_{raw}-Y_{min}}{Y_{max}-Y_{min}}\right) - 1 \quad (26)$$

The binary classification thresholds are kept at zero-median mark ($Y_{norm} = 0.0$), which proves that the reported 70-76% accuracy reflects domain generalization.

Table 11: Cross Dataset Performance across Valence, Arousal and Dominance Across All Three Datasets using 14 Overlapping EEG Channels

Training	Testing	Valence (%)	Arousal (%)	Dominance (%)	Avg accuracy(%)
DEAP	DREAMER	74.35	73.88	72.12	73.45
DREAMER	DEAP	71.18	70.45	69.85	70.49
DREAMER	AMIGOS	76.54	75.92	74.31	75.59
AMIGOS	DREAMER	75.12	74.45	73.08	74.22

3.7 Comparison Study

Table 12 is presented to compare the SpecTFT-DSDNet model with existing literature based models such as AICapsNet, HACapsNet, and TC-Net Transformer. The comparison is categorized into three dimensions: Valence, Arousal, and Dominance. Across all three datasets, the proposed SpecTFT-DSDNet model consistently outperformed existing methods in terms of accuracy. These results highlight the capability of model to overcome the common limitations of inter-subject variability and the loss of detailed information in deep networks. By achieving the highest reported precision in these categories, the results conclude that the proposed SpecTFT-DSDNet model represents spectral and temporal features for multimodal physiological emotion recognition.

Table 12: Comparison Analysis of the Proposed SpecTFT-DSDNet with Existing Models for all Three Datasets (Source-Reference)

Methods	Valence Accuracy (%)	Arousal Accuracy (%)	Dominance Accuracy (%)
DEAP dataset			
AICapsNet [21]	97.83	98.13	98.38
HACapsNet [22]	97.40	97.20	97.60
TC-Net Transformer [26]	98.76	98.81	98.82
SpecTFT-DSDNet (proposed)	98.89	98.88	98.91
TPRO-NET [30]	97.63	97.74	97.88
DREAMER dataset			
AICapsNet [21]	99.16	99.38	99.47
HACapsNet [22]	95.80	96.10	96.30
Deep CNN+CBAM [23]	87.7	87.4	NA
Att-1DCNN-GRU [24]	95.95	94.93	94.91
DSDNet [25]	98.63	98.92	99.22
TC-Net Transformer [26]	98.76	98.81	98.82
TPRO-NET [30]	98.18	98.37	98.93
SpecTFT-DSDNet (proposed)	99.28	99.41	99.52
AMIGOS dataset			
DSD-Net [25]	98.88	99.55	99.44
SpecTFT-DSDNet (proposed)	98.92	99.58	99.59

3.8 Research Implication

The SpecTFT-DSDNet model is developed on the ability to bridge gap between spectral analysis and long-range temporal modeling for emotion recognition. By effectively resolving spectral blurring through SSQ-DSTFFT and introducing frequency-aware spectral modulation block. This research proves that physiological noise is suppressed without losing transient emotional signals.

3.9 Discussion

The final results against study's initial goals ensures that proposed SpecTFT-DSDNet model has complete alignment across all targets. To address the first objective of solving fixed-window spectral blurring, combining SSQ-DSTFFT spectral mapping improved time-frequency energy concentration by 68.6% over computational transformers. To solve second objective of suppressing low-frequency noise and voltage scale mismatches between microvolt EEG and millivolt ECG signals, learnable spectral modulation was integrated with gated temporal fusion, prioritizing emotionally salient Alpha (8-13Hz) and Beta (13-30Hz) rhythms. To achieve the third objective of cross-subject generalizability, the SpecTFT-DSDNet model achieved subject-independent accuracy of 98.89% on DEAP, 99.28% on DREAMER and 98.92% on AMIGOS datasets, while maintaining a real-time segment latency of 12.4 ms. When compared against current state-of-the-art architectures in literature, SpecTFT-DSDNet consistently performs high against AICapsNet (97.83% DEAP valence), HACapsNet (95.80% DREAMER valence), TC-Net (98.76% DEAP valence), and TPRO-NET (97.63% DEAP valence), as well as the baseline DSDNet (98.63% DREAMER valence). Difference between these model and proposed SpecTFT-DSDNet model lies in replacing static frequency feature extractors and unmodulated attention layers with dynamic fractional-order spectral rotation coupled with learnable one-dimensional spectral modulation between recurrent decoders and masked attention layers. While intra-dataset cross-subject generalization is exceptional, zero-shot cross-dataset evaluation exhibits performance degradation (70.49%-75.59% average accuracy) due to unmitigated sensor hardware domain shifts across independent benchmark datasets. Multimodal scope was delimited to the scalp EEG and cardiac ECG dynamics by leaving additional modalities such as electrodermal activity and respiration unintegrated. The segment inference latency of 12.4 milliseconds supports real-time execution where parameter footprint requires quantization and knowledge distillation for deployment on ultra-low-power wearable microcontrollers.

4. CONCLUSION

The research proposes SpecTFT-DSDNet to address cross-subject signal variance and spectral blurring in physiological emotion recognition. The model is evaluated under a strict subject-independent protocol and obtains a cross-subject valence and arousal classification accuracy of 98.89% and 98.88% on DEAP, 99.28% and 99.41% on DREAMER, and 98.92% and 99.58% on AMIGOS, respectively. In context of literature state-of-the-art models, the primary contribution lies in uniting synchrosqueezed dynamic spectral rotation with a learnable 1D spectral modulation block between recurrent decoders and masked attention. Unlike existing capsule and transformer architectures depends on static frequency inputs, this coupling forces temporal attention mechanisms to explicitly track emotionally salient Alpha (8-13 Hz) and Beta (13-30 Hz) rhythms, suppressing physiological noise while improving spectral sharpness by 68.6% over conventional STFT. The proposed model maintains an inference latency of 12.4 ms per segment and SpecTFT-DSDNet shows strong computational feasibility. Key remaining limitations include performance degradation (70%-76%) during zero-shot cross-dataset evaluation due to sensor hardware domain shifts. Ablation study and statistical analysis confirm that overlapping segmentation and gated temporal dependency learning drive these gains. Future work will integrate self-interpretable attention layers to map frequency transitions directly to cortical spatial regions. Future work may extend the architecture by integrating self-interpretable layers which helps to identify which specific spectral bands and brain regions contribute the significant emotional transitions.

REFERENCES

[1] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao, and Y. Zhang, "Self-supervised EEG emotion recognition models based on CNN", *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* Vol. 31, 2023, pp. 1952-1962.

[2] X. Zhang, D. Huang, H. Li, Y. Zhang, Y. Xia, and J. Liu, "Self-training maximum classifier discrepancy for EEG emotion recognition", *CAAI Transactions on Intelligence Technology,* Vol. 8, No. 4, 2023, pp. 1480-1491.

[3] M. Asif, S. Mishra, M.T. Vinodbhai, and U.S. Tiwary, "Emotion recognition using temporally localized emotional events in EEG with naturalistic context: DENS# dataset", *IEEE Access,* Vol. 11, 2023, pp. 39913-39925.

[4] H. D. Le, G. S. Lee, S. H. Kim, S. Kim, and H. J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning", *IEEE Access,* Vol. 11, 2023, pp. 14742-14751.

[5] S. Woo, M. Zubair, S. Lim, and D. Kim, "Deep multimodal emotion recognition using modality-aware attention and proxy-based multimodal loss", *Internet of Things,* Vol. 31, 2025, p. 101562.

[6] B. Zali-Vargahan, A. Charmin, H. Kalbkhani, and S. Barghandan, "Deep time-frequency features and semi-supervised dimension reduction for subject-independent emotion recognition from multi-channel EEG signals", *Biomedical Signal Processing and Control*, vol. 85, 2023, p. 104806.

[7] S. Liu, Z. Wang, Y. An, J. Zhao, Y. Zhao, and Y. D. Zhang, "EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network", *Knowledge-Based Systems,* Vol. 265, 2023, p. 110372.

[8] A. Fang, F. Pan, W. Yu, L. Yang, and P. He, "ECG-based emotion recognition using random convolutional kernel method", *Biomedical Signal Processing and Control,* Vol. 91, 2024, p. 105907.

[9] T. Dhara, P.K. Singh, and M. Mahmud, "A fuzzy ensemble-based deep learning model for EEG-based emotion recognition", *Cognitive Computation,* Vol. 16, No. 3, 2024, pp. 1364-1378.

[10] K. Chen, H. Jing, Q. Liu, Q. Ai, and L. Ma, "A novel Caps-EEGNet combined with channel selection for EEG-based emotion recognition", *Biomedical Signal Processing and Control,* Vol. 86, 2023, p. 105312.

[11] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, "Multimodal emotion recognition from EEG signals and facial expressions", *IEEE Access,* Vol. 11, 2023, pp. 33061-33068.

[12] P. Saha, A.K.A. Kunju, M.E. Majid, S.B.A. Kashem, M. Nashbat, A. Ashraf, M. Hasan, A. Khandakar, M.S. Hossain, A. Alqahtani, and M.E.H. Chowdhury, "Novel multimodal emotion detection method using electroencephalogram and electrocardiogram signals", *Biomedical Signal Processing and Control,* Vol. 92, 2024, p. 106002.

[13] L. Wang, J. Hao, and T.H. Zhou, "ECG multi-emotion recognition based on heart rate variability signal features mining", *Sensors*, Vol. 23, 2023, no. 20, p. 8636.

[14] A.W. Awan, I. Taj, S. Khalid, S.M. Usman, A.S. Imran, and M.U. Akram, "Advancing emotional health assessments: A hybrid deep learning approach using physiological signals for robust emotion recognition", *IEEE Access,* Vol. 12, 2024, pp. 141890-141904.

[15] K. Ali and C.E. Hughes, "A unified biosensor–vision multi-modal transformer network for emotion recognition", *Biomedical Signal Processing and Control,* Vol. 102, 2025, p. 107232.

[16] S.K.B. Sangeetha, R.R. Immanuel, S.K. Mathivanan, J. Cho, and S.V. Easwaramoorthy, "An empirical analysis of multimodal affective computing approaches for advancing emotional intelligence in artificial intelligence for healthcare", *IEEE Access,* Vol. 12, 2024, pp. 114416-114434.

[17] M. Hamdi, N. Alduaiji, M. Elloumi, S. Asklany, F. Almutairy, N.M. Alotaibi, A. Yousef, and G. Abbas, "A novel complemented input matching model for accurate and timely emotion recognition", *Ain Shams Engineering Journal,* Vol. 17, No. 3, 2026, p. 103987.

[18] M. Zubair, S. Woo, S. Lim, and C. Yoon, "Deep representation learning for multimodal emotion recognition using physiological signals", *IEEE Access,* Vol. 12, 2024, pp. 106605-106617.

[19] A. Rodriguez Aguiñaga, M. Ramirez Ramirez, M.D.C. Salgado Soto, and M.D.L.A. Quezada Cisnero, "A multimodal low complexity neural network approach for emotion recognition", *Human Behavior and Emerging Technologies,* Vol. 2024, no. 1, p. 5581443.

[20] J. Yu, Y. Ru, B. Lei, and H. Chen, "GBV-Net: Hierarchical fusion of facial expressions and physiological signals for multimodal emotion recognition", *Sensors,* Vol. 25, No. 20, 2025, p. 6397.

[21] C. Yu, X. Zhang, X. Zhang, and Y. Yang, "EEG emotion recognition using multi-dimensional features with attention-based inception capsule network", *Biomedical Signal Processing and Control,* Vol. 110, 2025, p. 108248.

[22] K. Chen, W. Ruan, Q. Liu, Q. Ai, and L. Ma, "A novel deep learning model combining 3DCNN-CapsNet and hierarchical attention mechanism for EEG emotion recognition", *Neural Networks,* Vol. 186, 2025, p. 107267.

[23] T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, J. Xu, T. Sun, and N. Jiang, "A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition", *Computers in Biology and Medicine,* Vol. 159, 2023, p. 106938.

[24] Z. Wang and Y. Wang, "Emotion recognition based on multimodal physiological electrical

signals", *Frontiers in Neuroscience,* Vol. 19, 2025, p. 1512799.

[25] P. Samal and M.F. Hashmi, "A dynamic spectrum driven network for enhanced multimodal emotion recognition with EEG and ECG signals", *Biocybernetics and Biomedical Engineering,* Vol. 46, No. 1, 2026, pp. 139-161.

[26] Y. Wei, Y. Liu, C. Li, J. Cheng, R. Song, and X. Chen, "TC-Net: A transformer capsule network for EEG-based emotion recognition", Computers in Biology and Medicine, Vol. 152, 2023, p. 106463.

[27] N. Gahlan and D. Sethia, "AFLEMP: Attention-based federated learning for emotion recognition using multi-modal physiological data", *Biomedical Signal Processing and Control,* Vol. 94, 2024, p. 106353, August 2024.

[28] L. Farokhah, R. Sarno, and C. Fatichah, "Simplified 2D CNN architecture with channel selection for emotion recognition using EEG spectrogram", *IEEE Access,* Vol. 11, 2023, pp. 46330–46343.

[29] P.C. Huy, N.Q. Minh, N.D. Tien, and T.T.Q. Anh, "Short-term electricity load forecasting based on temporal fusion transformer model", *IEEE Access,* Vol. 10, 2022, pp. 106296-106304.

[30] X. Zhang, X. Cheng, and H. Liu, "TPRO-NET: An EEG-based emotion recognition method reflecting subtle changes in emotion", *Scientific Reports,* Vol. 14, No. 1, 2024, p. 13491.