Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066702

ARTICLE

EEG Scalogram Analysis in Emotion Recognition: A Swin Transformer and TCN-Based Approach

Selime Tuba Pesen and Mehmet Ali Altuncu*

Department of Computer Engineering, Kocaeli University, Kocaeli, 41380, Türkiye

*Corresponding Author: Mehmet Ali Altuncu. Email: mehmetali.altuncu@kocaeli.edu.tr

Received: 15 April 2025; Accepted: 20 June 2025; Published: 30 July 2025

ABSTRACT: EEG signals are widely used in emotion recognition due to their ability to reflect involuntary physiological responses. However, the high dimensionality of EEG signals and their continuous variability in the time-frequency plane make their analysis challenging. Therefore, advanced deep learning methods are needed to extract meaningful features and improve classification performance. This study proposes a hybrid model that integrates the Swin Transformer and Temporal Convolutional Network (TCN) mechanisms for EEG-based emotion recognition. EEG signals are first converted into scalogram images using Continuous Wavelet Transform (CWT), and classification is performed on these images. Swin Transformer is used to extract spatial features in scalogram images, and the TCN method is used to learn long-term dependencies. In addition, attention mechanisms are integrated to highlight the essential features extracted from both models. The effectiveness of the proposed model has been tested on the SEED dataset, widely used in the field of emotion recognition, and it has consistently achieved high performance across all emotional classes, with accuracy, precision, recall, and F1-score values of 97.53%, 97.54%, 97.53%, and 97.54%, respectively. Compared to traditional transfer learning models, the proposed approach achieved an accuracy increase of 1.43% over ResNet-101, 1.81% over DenseNet-201, and 2.44% over VGG-19. In addition, the proposed model outperformed many recent CNN, RNN, and Transformer-based methods reported in the literature.

KEYWORDS: Continuous wavelet transform; EEG; emotion recognition; Swin Transformer; temporal convolutional network

1 Introduction

Emotion is defined as an individual's response to environmental, physiological, and psychological stimuli and affects behavior, decision-making processes, and social interactions in daily life [1]. Emotion recognition is utilized in various fields, including computer-aided learning, e-commerce, banking, military and aviation, healthcare systems, and call center applications. In these systems, the more accurately the user's emotional state is identified, the more efficient communication becomes [2].

The perception and evaluation of emotions are directly related to the mental and neural activities of the individual. Physiological signals are usually used to measure these activities. Non-physiological signals include behavioral indicators such as gestures and body language, facial expressions, voice and speech, and eye movements [3]. Some physiological signals, such as EEG, can capture involuntary neural responses that individuals cannot control. Therefore, they provide more objective and reliable measurements than non-physiological signals [4]. Due to these advantages, many researchers have used EEG signals in emotion recognition studies. The low cost, portability, and ability of EEG devices to provide rich signal information at high temporal resolution are the main reasons for this preference [5].

In EEG-based emotion recognition, a common approach is to generate visual representations of the signals and process them using CNN-based models. Bagherzadeh et al. [6] aimed to compare the performance of the ensemble method against individual CNN models. First, EEG signals were converted into scalogram images in the time-frequency domain using CWT. Then, the obtained images were processed separately with five pre-trained CNN models: AlexNet, VGG-19, Inception-v1, Inception-v3, and ResNet-18. The classification decision was made using majority voting, a type of ensemble method. Analysis shows that the ensemble method yields better results compared to a single CNN model. Cai et al. [3] proposed a model based on Swin Transformer architecture using features obtained by combining spatial and frequency information in EEG signals. The effectiveness of the proposed model was demonstrated with experiments using data augmentation techniques on SEED and SEED-IV datasets. In the study, the use of window-based attention and shifted window partitions improved the overall performance of the model. Ke et al. [7] transformed EEG signals into a vector by combining features of frequencies in four different brain regions (frontal, parietal, temporal, occipital). They eliminated the mismatch between dimensions using a fully connected network. The feature vectors they obtained were fed to the Transformer, and self-attention calculations were performed. They also used Capsule Networks to identify the connections between local and global features. The proposed model indicated that the frontal lobe region performed better in emotion recognition compared to other brain regions.

In recent years, hybrid approaches utilizing attention mechanisms and Transformer architectures [8] have emerged as a prominent method for modeling long-range dependencies in EEG data. Xu et al. [9] aimed to develop an emotion recognition system using data in time, frequency, and spatial dimensions in EEG signals. First, EEG signals were divided into multiple frequency bands, and Power Spectral Density (PSD) and Differential Entropy (DE) features were extracted. Then, Transformer blocks were employed to assign attention weights to the most informative features in each dimension. Thus, the model's workload was reduced by decreasing the number of channels. As a result of the study, it was observed that the spectral Transformer block, which enables the selection of features, particularly in the frequency dimension, has the greatest impact on the model's performance. Li et al. [10] proposed a model that transforms frequency, time, and spatial information in EEG signals into a graph structure. They used graph-based learning and top-k connections for spatial data, the Temporal Convolutional Network (TCN) for temporal information, and the power spectral density and differential entropy methods for frequency information. Additionally, attention mechanisms were employed to highlight key features in each type of information. As a result of the experimental findings, better classification performance was obtained compared to graph-based methods in the literature. The hybrid system proposed by Liu et al. [11] consists of two main stages. In the first stage, noise reduction and feature extraction are performed on EEG signals using temporal and spatial convolutions. In the second stage, the extracted features are processed through a multi-head attention mechanism to model long-range dependencies. It is reported that the proposed model achieves higher classification accuracy compared to CNN and LSTM-based models. Liu et al. [12] introduced a novel self-attention mechanism that can simultaneously model both temporal and channel-wise dependencies. In their model, preprocessed two-dimensional raw EEG signals are directly used as input. The attention mechanism is structured hierarchically, enabling a coarse-to-fine computation strategy that reduces computational costs while preserving rich feature representations. This mechanism demonstrates superior performance compared to models that perform attention in a single dimension. In the model proposed by Gong et al. [13], EEG signals are first segmented into one-second intervals, and features are extracted across five distinct frequency bands. Attention mechanisms are then applied to emphasize information-rich components of the signal. Finally, CNN and Transformer-based architectures are integrated to perform emotion classification.

EEG signals contain both temporal and spectral information. Extracting these components with high time-frequency resolution in a way that preserves the time-varying and complex patterns in the signal provides more accurate results in emotion recognition [14]. For this purpose, analysis methods such as the Fourier transform (FT), the short-time Fourier transform (STFT), and the wavelet transform (WT) are widely used. The FT cannot preserve temporal locality, whereas the STFT, due to its fixed window size, is limited in resolving both low and high-frequency components. In contrast, the WT performs multiresolution time-frequency analysis using bandpass filters with variable bandwidths, enabling it to better adapt to the diverse frequency characteristics of EEG signals [15]. The outputs of such analysis can be transformed into rich visual representations (e.g., scalograms) that retain both detailed temporal and spectral content. These visualizations, in turn, provide more informative inputs to deep learning models, enhancing their ability to learn meaningful features.

Recent studies on emotion recognition using EEG signals have shown that CNN-based methods are generally preferred. However, unlike CNNs, Transformer-based architectures can learn long-range dependencies more effectively by directly modeling the global context. In particular, the Swin Transformer preserves global context and significantly reduces computational cost compared to methods such as the Vision Transformer (ViT) [16] due to its shifted window mechanism [17]. The Vanilla Transformer [18], initially designed for sequential data processing, and the Graph Attention Network (GAT) [19], developed for graph-based analysis, both demonstrate limited effectiveness in modeling visual EEG representations that contain spatial structures, such as scalograms [20]. On the other hand, 2D-CNNs are effective in capturing short-term local dependencies but are insufficient in modeling long-term temporal relationships. As a result, models like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Temporal Convolutional Network (TCN) are commonly used to capture such long-range dependencies. Compared to other time-series models, TCN significantly mitigates the vanishing gradient problem through residual connections and enables faster training due to its parallel computational structure [21]. In this study, the Swin Transformer is employed for spatial feature extraction, while the Temporal Convolutional Network (TCN) is utilized to capture temporal dependencies. Accordingly, a modular hybrid architecture is proposed that independently models spatial and temporal information. Compared to fully integrated spatio-temporal architectures, this design enables more efficient training and lower computational complexity due to reduced parameterization and architectural decoupling.

In this study, we propose an innovative model for emotion recognition from EEG signals using Swin Transformer, TCN, and cross-attention mechanisms. The contributions of the study can be listed as follows:

I. To effectively model non-stationary EEG signals, these signals were transformed into scalogram images using the Continuous Wavelet Transform (CWT). This transform preserves both temporal and frequency components, allowing the model to learn complex patterns and local features in the signals more effectively.

II. The self-attention mechanism in Swin Transformer is used to extract spatial context from scalogram images. In addition, the spatial features extracted by the Swin Transformer were enhanced with CBAM (Convolutional Block Attention Module) to emphasize important regions. The channel attention and spatial attention components of CBAM enable the model to focus on only the most critical information.

III. A TCN block with dilated convolutions and residual connections is used to model long-term temporal dependencies in scalogram images efficiently.

IV. Spatial and temporal features of EEG signals are combined for more comprehensive learning of the model.

V. The proposed model demonstrated higher classification performance compared to some recent CNN, RNN, and Transformer methods on the SEED dataset, achieving accuracy improvements of up to 6.7%. Furthermore, comparable classification accuracy was achieved using a simpler and more modular architecture instead of more complex structures such as LSTM.

The rest of the paper is organized as follows. Section 2 provides detailed information about the dataset, the proposed model's architecture, and the methods employed. Section 3 analyzes the performance of the model, and Section 4 discusses the results obtained. Finally, in Section 5, the conclusion and future work are presented.

2 Materials and Methods

The architecture of the proposed model is shown in Fig. 1. As can be seen in Fig. 1, the proposed model consists of three main blocks. First, in the Swin + CBAM block, 224 × 224 scalogram images are divided into 4 × 4 windows and converted into 96-dimensional embed vectors. The Swin Transformer's shifted window mechanism was used to capture spatial features. Important feature channels were identified using the channel attention mechanism (SE Block) within the CBAM component, while the spatial attention mechanism (implemented via a 7 × 7 convolution) was employed to refine spatial feature maps by emphasizing informative regions. The feature vector of size 768 obtained in this block was transferred to the TCN + Attention block. In the TCN + Attention block, two different dilated 1D convolution filters (3 × 3, dilation = 2, and dilation = 4) were used to capture temporal dependencies. As a result, 1024- and 512-dimensional feature maps were created. In addition, batch normalization, dropout, and GELU activation were applied to the TCN layer. To mitigate information loss in the features generated by the TCN, the SE channel attention mechanism was employed to convert the features into a 512-dimensional vector, which was then passed to the fusion stage. In the fusion layer, 768-dimensional spatial features from the Swin Transformer and 512-dimensional temporal features from the TCN are combined to form a 1280-dimensional feature vector. The 1280-dimensional features extracted by the model are then processed in the 1024-dimensional fully connected (FC) layer. In the classification stage, three classes of emotional states, namely 'Positive', 'Neutral', and 'Negative', were detected.

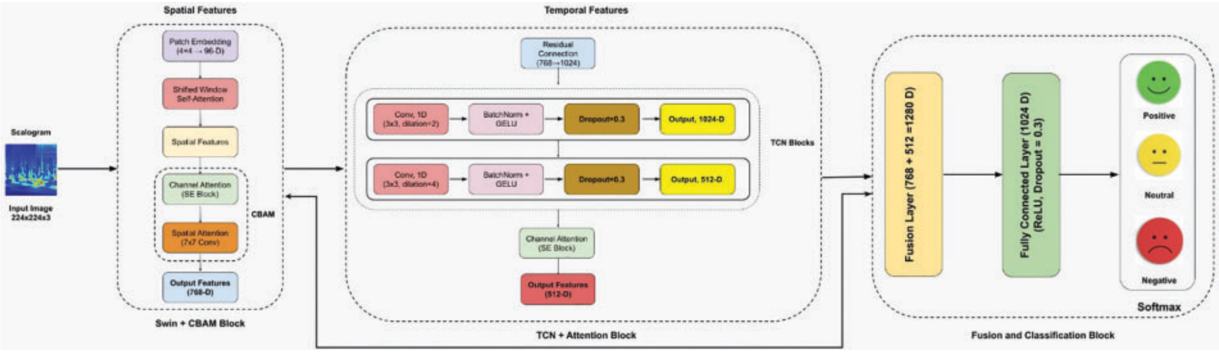
Spatial Features
Temporal Features
Scalogram
Input Image 224x224x3
Patch Embedding (4×4 → 96-D)
Shifted Window Self-Attention
Spatial Features
Channel Attention (SE Block)
Spatial Attention (7x7 Conv)
CBAM
Output Features (768-D)
Swin + CBAM Block
Residual Connection (768→1024)
Conv, 1D (3x3, dilation=2)
BatchNorm + GELU
Dropout=0.3
Output, 1024-D
Conv, 1D (3x3, dilation=4)
BatchNorm + GELU
Dropout=0.3
Output, 512-D
TCN Blocks
Channel Attention (SE Block)
Output Features (512-D)
TCN + Attention Block
Fusion Layer (768 + 512 =1280 D)
Fully Connected Layer (1024 D) (ReLU, Dropout = 0.3)
Positive
Neutral
Negative
Softmax
Fusion and Classification Block

Figure 1: The architecture of the proposed model

2.1 SEED Dataset

The SEED [22] dataset used in this study contains EEG signals from the emotional responses of 15 participants. The EEG recordings were acquired using a 62-channel NeuroScan system and were preprocessed using a bandpass filter to remove noise and artifacts. Each participant attended three sessions. In the experiments, participants were shown 15 emotionally labeled movie clips in each session. Participants rated their emotional reactions after watching each clip, and these ratings were categorized as positive (+1), neutral (0), and negative (−1) [23].

2.2 Scalogram Generation Using Continuous Wavelet Transform (CWT)

In this study, CWT is used to analyze the time-frequency components of EEG signals. CWT allows both high and low-frequency components to be examined at different scales over time. For this reason, CWT is widely preferred for analyzing biomedical signals such as time-varying EEG. CWT decomposes a signal into different frequency components and uses wavelet functions to examine how each component changes over time [24]. The CWT for a signal is defined as in Eq. (1):

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

In Eq. (1), $x\left(t\right)$ is the input signal, $\psi(t)$ is the wavelet function, a is the scale parameter (which determines the frequency information), b is the shift parameter (which determines the time information), and $*$ is the complex conjugate operator.

The Morlet wavelet function, one of the most preferred wavelet functions, is used in this study. The Morlet wavelet function is defined as in Eq. (2).

$$\psi\left(\mathrm{t}\right)=\mathrm{e}^{\mathrm{j}2\pi\mathrm{f}_0\mathrm{t}}\mathrm{e}^{-\mathrm{t}^2/2} \tag{2}$$

In Eq. (2), f_0 denotes the center frequency of the wavelet. In this study, the center frequency is set to 10 Hz, and the scale parameter is set to a value between 1 and 64. The chosen wavelet configuration provides an effective balance between adequate temporal resolution at high frequencies and excellent frequency resolution at low frequencies.

In this study, the wavelet transform was performed separately for each EEG signal in the SEED dataset. Thus, a total of 45 × 15 × 62 = 41,850 scalogram images were generated according to the formula number of samples (N) × number of trials (T) × number of channels (C). Since there are an equal number of EEG recordings for each class in the SEED dataset, the scalogram images were equally distributed, with 13,950 images for each class. These CWT-derived images were used as input data for the proposed model. Representative scalogram examples are displayed in Fig. 2.

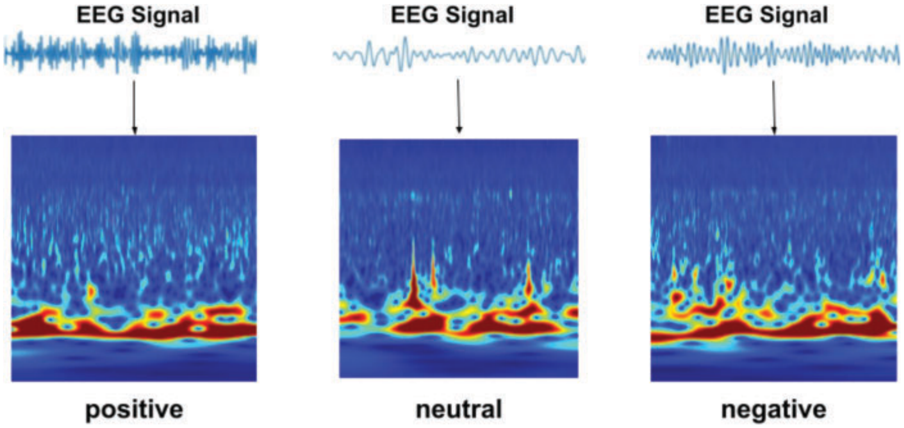

Figure 2: Example scalogram images obtained using CWT

2.3 Swin Transformer + CBAM Block

Swin Transformer and CBAM attention mechanisms are used to extract spatial features from EEG scalogram images. Swin Transformer is a hierarchical Transformer architecture proposed by Liu et al. [25]. In this architecture, the spatial resolution is reduced, and the channel width is increased through patch splitting and merging operations [26]. Furthermore, the sliding window-based attention mechanism enables the model to capture both local and global patterns in images, making Swin Transformer an efficient architecture for spatial feature extraction.

In the study, scalogram images are given as input to the Swin Transformer model. In Swin Transformer, each input image I(x, y) is divided into 4 × 4 patches, and each patch (X_p) is transformed into a 96-dimensional vector (Z) as in Eq. (3).

$$Z = W_p.X_p + b_p \tag{3}$$

In Eq. (3), W_p represents the learnable weight matrix and b_p represents the bias term.

In Swin Transformer, learning the relationships between patches within the window is accomplished through Shifted Window Self-Attention [25]. For each self-attention operation, the formula in Eq. (4) is used:

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

In Eq. (4), Q, K and V denote the query, key, and value matrices, respectively. The term $\sqrt{d_k}$ (the square root of the key dimension) is used as a scaling factor.

CBAM was employed to further enhance the spatial features extracted by the Swin Transformer. It utilizes channel and spatial attention mechanisms to refine the intermediate feature maps, allowing the model to focus on the most informative regions and learn more meaningful representations. These attention mechanisms are applied to the feature maps generated by the Swin Transformer, as illustrated in Eq. (5).

$$F' = M_S\left(M_c\left(F\right)\right) = M_S.M_c.F \tag{5}$$

In Eq. (5):

- F = represents 768-dimensional attribute maps extracted from Swin Transformer;
- M_c = represents importance weights learned by the channel attention mechanism;
- M_S = represents importance weights learned by the spatial attention mechanism;
- F′ = represents the attribute maps processed by CBAM and passed to the classification layer of the model.

The 768-D spatial features obtained after CBAM were transferred to the TCN + Attention block for learning temporal dependencies and to the Fusion and Classification block for final classification.

2.4 TCN + Attention Block

TCN and attentional mechanisms were used to enhance the learning of spatial features extracted from EEG scalogram images in a time series context. TCN, developed to model time series, can effectively capture both short- and long-term dependencies, thanks to dilated convolutions that leave gaps between the core elements. In this study, two different dilation ratios (d = 2 and d = 4) were used to explore these dependencies. The mathematical expression for the 1D convolution used in the model is provided in Eq. (6).

$$y\left(t\right) = \sum_{k=0}^{K-1} w\left(k\right) \cdot x\left(t - k \cdot d\right) \tag{6}$$

In Eq. (6):

- y(t) = represents the value of the output sequence at time step t;
- x (t) = represents the value of the input sequence at time step t;
- w(k) = represents convolution kernel weights;
- k = represents the kernel size;
- *d* = represents the dilation rate;
- k = represents element indices;
- t − k.d = represents the offset introduced by the dilation in the input data.

Batch Normalization was applied to both convolutional layers in the TCN block to stabilize the training process and ensure faster network convergence. GELU was chosen as the activation function, and a dropout rate of 0.3 was set to prevent overfitting. Temporal features extracted from the TCN were reduced to 512 dimensions with dilated convolutions, and the most important features were highlighted by applying a channel-based attention mechanism (SE Block). These features were transferred to the Fusion and Classification Block.

2.5 Fusion and Classification Block

In this block, 768-D spatial and 512-D temporal features from the previous two blocks are directly combined (Eq. (7)) to form a 1280-D feature vector. Thus, both features are evaluated together.

$$F = \left[F_{spatial}; F_{temporal}\right] \tag{7}$$

After the feature fusion stage, the 1280-dimensional feature vectors were fed into a fully connected layer, where they were compressed to 1024 dimensions to reduce parameter complexity and improve learning efficiency. ReLU activation and a dropout rate of 0.3 were applied to prevent overfitting. Finally, a Softmax activation function was used in the output layer to classify the input into three distinct emotional categories: positive, neutral, and negative.

2.6 Training Configuration

Experiments on the proposed model were conducted on a computer equipped with an Intel Core i5 processor and 16 GB of RAM, running the Microsoft Windows operating system. The creation and training of the model were performed using the PyTorch deep learning library in the JupyterLab development environment. The Scikit-learn library was used for performance evaluation metrics. The basic configurations used for training and evaluation of the model are presented in Table 1. The scalogram images were resized to 224 × 224 pixels and converted into a tensor format to meet the input requirements of the Swin Transformer architecture. The AdamW optimization algorithm, which is widely used in Transformer-based models, was employed. To prevent overfitting, the learning rate was set to 1e−4 and the weight decay to 1e−3. Additionally, cross-entropy loss incorporating label smoothing (0.1) was used to enhance the model's generalization ability and reduce overconfident predictions.

Table 1 : Experimental settings

Setting	Value
Optimization	AdamW
Learning rate (lr)	1e−4

(Continued)

Table 1 (continued)

Setting	Value
Weight decay	1e−3
Loss function	Cross Entropy Loss + Label Smoothing (0.1)
Number of epochs	100
Batch size	8
Data split	70% training–30% validation
Data transformations	Normalize Resize (224 × 224) to Tensor

2.7 Evaluation Criteria

The model's performance was evaluated using Accuracy, Precision, Recall, and F1-score. To compute these metrics, the values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were derived from the confusion matrix, as shown in Eqs. (8)–(11).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1} - \text{score:} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

3 Results

The confusion matrix values obtained after the training of the proposed model are given in Fig. 3. According to the confusion matrix results, the proposed model achieved an accuracy of 97.53%. In addition, the precision, recall, and F1-score values calculated separately for each class are presented in Table 2.

The classification metrics presented in Table 2 demonstrate that the proposed model achieves high and balanced performance across all three emotional classes. The close Precision, Recall, and F1-score values among the classes indicate that the model maintains consistent classification performance across all emotional categories. Notably, the high recall rate of 98.13% for the neutral class—which is typically difficult to classify due to its ambiguous nature—indicates that the model can effectively distinguish not only strong emotional states but also subtle emotional transitions. These performance levels provide a solid basis for evaluating the model's applicability in real-world scenarios such as mood monitoring and user response analysis.

Fig. 4 presents the graph showing the training and validation accuracy values obtained at each epoch during the training process of the proposed model. Upon examining the graph, it is observed that the learning process progresses rapidly, with accuracy values reaching around 90% within the first few epochs. This suggests that the model rapidly captures essential patterns from the training data. Furthermore, the small difference between training and validation accuracy throughout the process suggests that the model generalizes well to the validation data without overfitting. These findings demonstrate that the proposed architecture ensures efficient and stable learning.

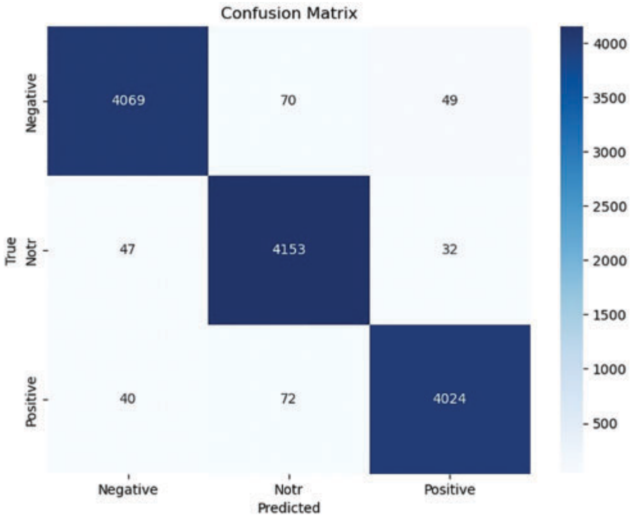

Figure 3: Confusion matrix of the proposed model

Table 2: Classification metrics

Category	Precision (%)	Recall (%)	F1-score (%)
Negative	97.91	97.16	97.53
Neutral	96.69	98.13	97.41
Positive	98.03	97.29	97.66

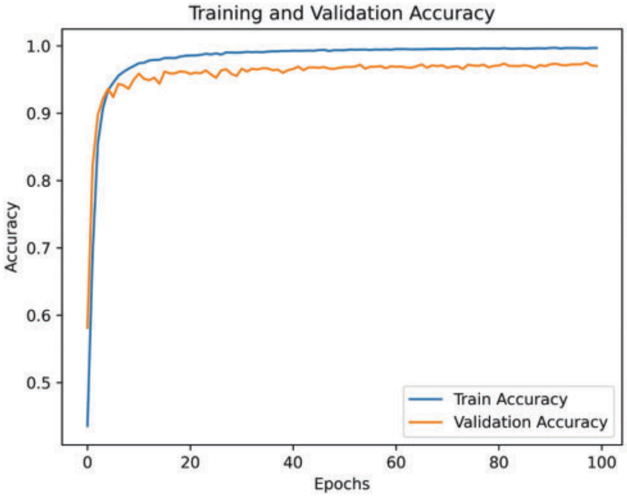

Figure 4: Training and validation accuracy per epoch

The training and validation losses of the proposed model across epochs are presented in Fig. 5. Upon analyzing the graph, it is observed that the training loss stabilizes at a low level after a certain number of epochs. Although there are slight fluctuations in the validation loss, its overall trend closely follows that of the training loss. The consistent and closely aligned behavior of the training and validation losses indicates that the model avoids overfitting and demonstrates strong generalization capability. These findings suggest that the proposed architecture learns without overfitting and is capable of maintaining high performance on unseen data.

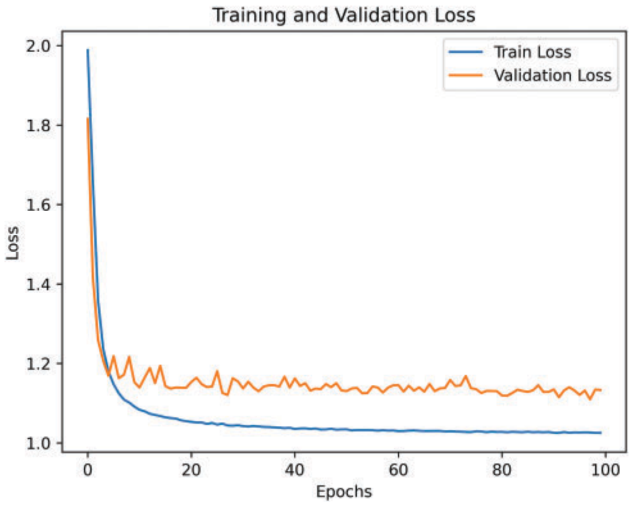

Figure 5: Training and validation loss per epoch

Fig. 6 presents the Receiver Operating Characteristic (ROC) curves of the proposed model for three different emotion classes: positive, neutral, and negative. Upon examining the graph, it is observed that the Area Under the Curve (AUC) values are notably high for all classes: positive = 0.99, neutral = 0.99, and negative = 1.00. These scores indicate that the model exhibits strong discriminative ability among classes and maintains very low false positive rates. In particular, the perfect AUC score of 1.00 for the negative class suggests that the model did not misclassify any negative instances as positive. These findings demonstrate that the model can accurately recognize each emotional class with high confidence and achieves strong overall classification performance.

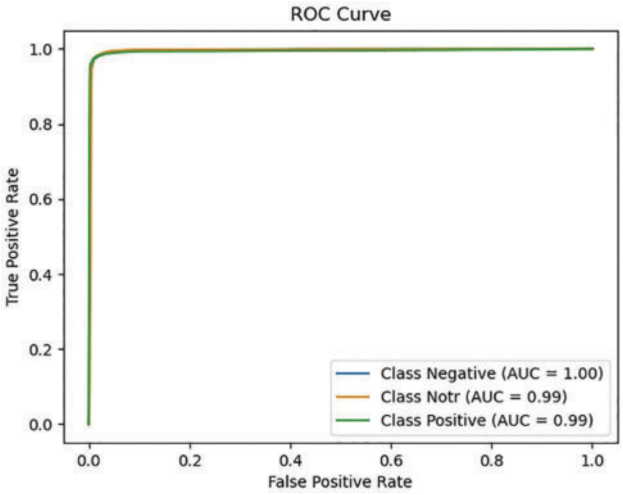

Figure 6: ROC Curve and AUC scores for each class

4 Discussion

To evaluate the performance of the proposed model, the model was compared with the DenseNet-201, VGG-19, and ResNet-101 transfer learning models. To make the comparison fair and consistent, the hyperparameters given in Table 1 were used in the training of all three models. The results obtained are shown in Table 3.

Table 3: Performance comparison of the proposed model with DenseNet-201, VGG-19, and ResNet-101 models

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet-201	95.72	95.77	95.72	95.74
VGG-19	95.09	95.1	95.09	95.09
ResNet-101	96.1	96.15	96.09	96.12
Swin + TCN + Attention	97.53	97.54	97.53	97.54

When the results in Table 3 are examined, the proposed Transformer- and TCN-based model yields the best results in all metrics, achieving 97.53% accuracy, 97.54% precision, 97.53% recall, and 97.54% F1-score. Additionally, the proposed model demonstrated the best performance across all classes (negative, neutral, and positive). The second-best performance was achieved with the ResNet-101 model, yielding 96.10% accuracy, 96.15% precision, 96.09% recall, and 96.12% F1-score. However, ResNet-101 exhibited comparatively lower effectiveness, particularly in the negative class. The DenseNet-201 model ranks third in the performance ranking with 95.72% accuracy, 95.77% precision, 95.72% recall, and 95.74% F1-score values. Although the model gave balanced results in positive and neutral classes, it made more errors in the negative class than in the others. The VGG-19 model exhibited the lowest performance among the four methods with 95.09% accuracy, 95.10% precision, 95.09% recall, and 95.09% F1-score. The VGG-19 model performed worse than the other models, especially in distinguishing the positive class.

The computational efficiency of the proposed model is presented in Table 4, compared with DenseNet-201, VGG-19, and ResNet-101. The comparison includes the total number of parameters, the inference time per sample, and the training time per epoch for each model. The proposed model has a significantly lighter architecture compared to VGG-19 (139.58 million parameters), with 33.27 million parameters, similar to ResNet-101 (42.5 million parameters) and more complex than DenseNet-201 (18 million parameters). In terms of training time, the proposed model takes 1218.96 s per epoch, which is shorter than VGG-19 but longer than ResNet-101 and DenseNet-201. The proposed model demonstrates the highest inference latency per sample (414.99 ms) among the compared architectures, which is likely attributable to the computational complexity introduced by its multi-layered structure. Therefore, additional optimizations may be necessary to enhance computational performance in real-time scenarios.

The comparison of the proposed model with recent studies using the SEED dataset is presented in Table 5. The proposed model achieved a higher classification accuracy of 97.53% compared to CNN- and RNN-based approaches such as those by Yuvaraj et al. [27], Trujillo et al. [28], Dai et al. [29], and Vujji et al. [30]. Although the models proposed by Xu et al. [9] and Asif et al. [31] reported similar accuracy levels (97.17% and 97.68%, respectively), their architectures involve more complex components, including multiple attention blocks or recurrent structures such as LSTM. In contrast, the proposed model consists of two separate modules: Swin Transformer and TCN, which operate independently to extract spatial and temporal features, respectively. This modular and decoupled design enhances the model's flexibility and interpretability. In conclusion, the proposed model offers a competitive and practical alternative for EEG-based emotion classification.

Table 4: Comparison of computational efficiency between the proposed model and the DenseNet-201, VGG-19, and ResNet-101 models

	Number of parameters	Inference time (ms)	Epoch time (s)
DenseNet-201	18,098,691	242.72	761.54
VGG-19	139,582,531	107.30	1325.13
ResNet-101	42,506,307	169.04	728.99
Swin + TCN + Attention	33,273,227	414.99	1218.96

Table 5: Comparison of the proposed model with current studies in the literature using the SEED dataset

Study	Method	Accuracy (%)
Yuvaraj et al. [27]	3D-CNN + ELM + Post-Processing	90.85
Quan et al. [32]	Multi-source Transfer Learning + MR-VAE	92.83
Trujillo et al. [28]	Kernel PCA + Radial Basis Function + Random Forest	93.20
Chen et al. [33]	Variational Autoencoder + Capsule Network with Trial Correction	93.48
Zhu et al. [34]	Multiple Class Domain Adaptation + SLAC—Source Label Adaptive Correction + Target Label Prediction	93.57
Zang et al. [35]	Contrastive Reinforced Transfer Learning	93.57
Zhang et al. [36]	4D Feature Representations + Multiple Attention Mechanisms	93.93
Dai et al. [29]	CNN + RNN + Contrastive Learning	95.16
Vujji et al. [30]	Variational Mode Decomposition (VMD) + Grid Search SVM	95.80
Xu et al. [9]	Attention-Based Multiple Dimensions EEG Transformer (AMDET)	97.17
Asif et al. [31]	CNN-LSTM	97.68
Our study	Swin Transformer + TCN + Attention	97.53

Although the proposed hybrid model achieves high accuracy and balanced classification performance, several limitations should be addressed in future studies. First, the model has an inference time of 414.99 ms, which limits its suitability for real-time applications that require rapid response. Second, the training and evaluation processes were conducted solely on the SEED dataset. Although SEED is widely used, relying on a single dataset may limit the model's generalizability. In future work, evaluating the model on diverse EEG datasets will be essential to assess its robustness to varying signal characteristics. Furthermore, the practical applicability of the model can be expanded by integrating it into real-time systems for domains such as human-computer interaction and psychological analysis.

5 Conclusions

In this study, a hybrid model combining the Swin Transformer and Temporal Convolutional Network (TCN) is proposed for emotion recognition using EEG signals. The proposed model has a structure that evaluates both temporal and spatial features. The performance of the model was assessed on the SEED dataset, yielding accuracy, precision, recall, and F1 score values of 97.53%, 97.54%, 97.53%, and 97.54%, respectively. The findings confirm that the proposed model performs with both high accuracy and class balance in emotion classification.

When the proposed model is compared to transfer learning methods, it is observed that it significantly increases the success of emotion recognition. The proposed model achieved 1.43% higher accuracy than ResNet-101, 1.81% higher accuracy than DenseNet-201, and 2.44% higher accuracy than VGG-19. When the Precision, Recall, and F1-score values are also examined, the proposed model demonstrated superior performance compared to transfer learning-based methods. In addition, the proposed model has achieved higher accuracy rates than CNN, RNN, and Transformer-based approaches for EEG-based emotion recognition. These findings suggest that Transformer-based time series modeling approaches yield improved performance in EEG-based emotion recognition.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Mehmet Ali Altuncu; methodology, Selime Tuba Pesen, and Mehmet Ali Altuncu; software, Selime Tuba Pesen; validation, Selime Tuba Pesen, and Mehmet Ali Altuncu; formal analysis, Mehmet Ali Altuncu; investigation, Selime Tuba Pesen; resources, Selime Tuba Pesen; data curation, Selime Tuba Pesen; writing—original draft preparation, Selime Tuba Pesen; writing—review and editing, Selime Tuba Pesen, and Mehmet Ali Altuncu; visualization, Selime Tuba Pesen, and Mehmet Ali Altuncu; supervision, Mehmet Ali Altuncu; project administration, Selime Tuba Pesen, and Mehmet Ali Altuncu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available at http://bcmi.sjtu.edu.cn/~seed/index.html (accessed on 19 June 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Samal P, Hashmi MF. Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: a review. Artif Intell Rev. 2024;57(3):50. doi:10.1007/s10462-023-10690-2.
2. Sönmez YÜ, Varol A. In-depth investigation of speech emotion recognition studies from past to present: the importance of emotion recognition from speech signal for AI. Intell Syst Appl. 2024;22:200351. doi:10.1016/j.iswa.2024.200351.
3. Cai M, Chen J, Hua C, Wen G, Fu R. EEG emotion recognition using EEG-SWTNS neural network through EEG spectral image. Inf Sci. 2024;680:121198. doi:10.1016/j.ins.2023.121.
4. Li Q, Liu Y, Yan F, Zhang Q, Liu C. Emotion recognition based on multiple physiological signals. Biomed Signal Process Control. 2023;85(4):104989. doi:10.1016/j.bspc.2023.104989.
5. Rahman MM, Sarkar AK, Hossain MA, Hossain MS, Islam MR, Hossain MB, et al. Recognition of human emotions using EEG signals: a review. Comput Biol Med. 2021;136(2):104696. doi:10.1016/j.compbiomed.2021.104696.
6. Bagherzadeh S, Maghooli K, Shalbaf A, Maghsoudi A. Emotion recognition using continuous wavelet transform and ensemble of convolutional neural networks through transfer learning from electroencephalogram signal. Front Biomed Technol. 2023;10(1):47–56. doi:10.18502/fbt.v10i1.12073.
7. Ke S, Ma C, Li W, Lv J, Zou L. Multi-region and multi-band electroencephalogram emotion recognition based on self-attention and capsule network. Appl Sci. 2024;14(2):702. doi:10.3390/app14020702.
8. Anwar A, Khalifa Y, Coyle JL, Sejdic E. Transformers in biosignal analysis: a review. Inf Fusion. 2024;114(2):102697. doi:10.1016/j.inffus.2024.102697.
9. Xu Y, Li J, Zhang X, Zhang X, Zhang Z, Guo Y. AMDET: attention based multiple dimensions EEG Transformer for emotion recognition. IEEE Trans Affect Comput. 2024;15(3):1067–77. doi:10.1109/TAFFC.2023.3318321.

10. Li C, Wang F, Zhao Z, Wang H, Schuller BW. Attention-based temporal graph representation learning for EEG-based emotion recognition. IEEE J Biomed Health Inform. 2024;28(10):5755–67. doi:10.1109/JBHI.2024.1234567.
11. Liu R, Chao Y, Ma X, Sha X, Sun L, Li S, et al. Attention-based interpretable transformer framework for EEG emotion recognition. Front Neurosci. 2024;18:1320645. doi:10.3389/fnins.2024.1320645.
12. Liu Y, Zhou Y, Zhang D. TCT: temporal and channel transformer for EEG-based emotion recognition. Proc IEEE Int Symp Comput Based Med Syst. 2022;2022:366–71. doi:10.1109/CBMS55023.2022.00072.
13. Gong L, Li M, Zhang T, Chen W. EEG emotion recognition using attention-based convolutional transformer neural network. Biomed Signal Process Control. 2023;84(4):104835. doi:10.1016/j.bspc.2023.104835.
14. Garg D, Verma GK, Singh AK. EEG-based emotion recognition using MobileNet Recurrent Neural Network with time-frequency features. Appl Soft Comput. 2024;154:111338. doi:10.1016/j.asoc.2023.111338.
15. Riaz F, Hassan A, Rehman S, Niazi IK, Dremstrup K. EMD-based temporal and spectral features for the classification of EEG signals using supervised learning. IEEE Trans Neural Syst Rehabil Eng. 2015;24(1):28–35. doi:10.1109/TNSRE.2015.2479255.
16. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.00321.
17. Chen Z, Ma M, Li T, Wang H, Li C. Long sequence time-series forecasting with deep learning: a survey. Inf Fusion. 2023;97(5):101819. doi:10.1016/j.inffus.2023.101819.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv: 1706.03762. 2017. doi:10.48550/arXiv.1706.03762.
19. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv:1710.10903. 2017 DOI 10.48550/arXiv.1710.10903.
20. Vafaei E, Hosseini M. Transformers in EEG analysis: a review of architectures and applications in motor imagery, seizure, and emotion classification. Sensors. 2025;25(5):1293. doi:10.3390/s25051293.
21. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271. 2018. doi:10.48550/arxiv.1803.01271.
22. Zheng WL, Lu BL. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans Autom Ment Dev. 2015;7(3):162–75. doi:10.1109/TAMD.2015.2431497.
23. Duan RN, Zhu JY, Lu BL. Differential entropy feature for EEG-based emotion classification. In: Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER); 2012 Nov 6–8; San Diego, CA, USA. doi:10.1109/NER.2013.6695910.
24. Arts LP, Van den Broek EL. The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time-frequency analysis. Nat Comput Sci. 2022;2(1):47–58. doi:10.1038/s43588-021-00179-9.
25. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. doi:10.1109/ICCV48922.2021.00987.
26. Yao D, Shao Y. A data efficient Transformer based on Swin Transformer. Vis Comput. 2024;40(4):2589–98. doi:10.1007/s00371-022-02535-9.
27. Yuvaraj R, Baranwal A, Prince AA, Murugappan M, Mohammed JS. Emotion recognition from spatio-temporal representation of EEG signals via 3D-CNN with ensemble learning techniques. Brain Sci. 2023;13(4):685. doi:10.3390/brainsci13040685.
28. Trujillo L, Hernandez DE, Rodriguez A, Monroy O, Villanueva O. Effects of feature reduction on emotion recognition using EEG signals and machine learning. Expert Syst. 2024;41(8):e13577. doi:10.1111/exsy.13577.
29. Dai S, Li M, Wu X, Ju X, Li X, Yang J, et al. Contrastive learning of EEG representation of brain area for emotion recognition. IEEE Trans Instrum Meas. 2025;74:2506913. doi:10.1109/TIM.2025.1234567.
30. Vujji A, Pusarla N, Singh A, Tripathi S. Emotion recognition using VMD domain bandwidth and spectral features. Int J Inf Technol. 2025;74:2403–11. doi:10.1007/s41870-025-00876-3.
31. Asif M, Mishra S, Vinodbhai MT, Tiwary US. Emotion recognition using temporally localized emotional events in EEG with naturalistic context: DENS# dataset. IEEE Access. 2023;11:39913–25. doi:10.1109/ACCESS.2023.3281234.

32. Quan J, Li Y, Wang L, He R, Yang S, Guo L. EEG-based cross-subject emotion recognition using multi-source domain transfer learning. Biomed Signal Process Control. 2023;84(1):104741. doi:10.1016/j.bspc.2023.104741.
33. Chen H, Li J, He H, Sun S, Zhu J, Li X, et al. VAE-CapsNet: a common emotion information extractor for cross-subject emotion recognition. Knowl Based Syst. 2025;311(2):113018. doi:10.1016/j.knosys.2025.113018.
34. Zhu L, Xu M, Huang A, Zhang J, Tan X. Multiple class transfer learning framework with source label adaptive correction for EEG emotion recognition. Biomed Signal Process Control. 2025;104:107536. doi:10.1016/j.bspc.2024.107536.
35. Zang Z, Yu X, Fu B, Liu Y, Ge SS. Contrastive reinforced transfer learning for EEG-based emotion recognition with consideration of individual differences. Biomed Signal Process Control. 2025;106:107622. doi:10.1016/j.bspc.2025.107622.
36. Zhang Y, Qu J, Zhang Q, Cheng C. EEG-based emotion recognition based on 4D feature representations and multiple attention mechanisms. Biomed Signal Process Control. 2025;103(16):107432. doi:10.1016/j.bspc.2024.107432.