IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3934~3943
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3934-3943

Emotion recognition of electroencephalogram using hybrid convolutional neural networks and vision transformer

Esmeralda Contessa Djamal[1], Revan Vio Endriansyah[1], Daswara Djajasasmita[2]
[1]Department of Informatics, Faculty of Science and Informatics, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
[2]Medical Education Program, Faculty of Medicine, Universitas Jenderal Achmad Yani, Cimahi, Indonesia

Article Info

Article history:

Received Dec 16, 2025
Revised Jun 17, 2026
Accepted Jul 9, 2026

Keywords:

Convolutional neural network
Electroencephalogram
Emotion
Vision transformer
Wavelet

ABSTRACT

Electroencephalogram (EEG) signal-based emotion classification faces challenges in spatio-temporal complexity and inter-channel redundancy. This paper proposes a 2D convolutional neural network (CNN) and vision transformer (ViT) approach. Relevant emotions require signal extraction in the 4-40 Hz frequency band using the discrete wavelet transform (DWT). This study uses the SEED dataset from 15 subjects, with 62 channels reduced to 12. Every 5-second segment is decomposed using DWT into four frequency bands, which are then mapped to a 2D spatial representation for CNNs and ViTs. Experiment results show that the DWT-2D CNN-ViT achieves the best accuracy of 87% with a final loss of 0.087. In the meantime, the CNN-only model (85%), the ViT-only model (77%), and the DWT-only model (47%). Testing with several optimizers also shows that AdamW provides the highest performance with the fastest training time of 12.34 minutes. These results demonstrate that this integration can produce more efficient and stable learning. These findings indicate that the combination of DWT and the CNN-ViT hybrid architecture is effective for accurate, stable EEG-based emotion recognition and is potentially applicable to real-time emotion-monitoring systems.

Corresponding Author:

Esmeralda Contessa Djamal
Department of Informatics, Faculty of Science and Informatics, Universitas Jenderal Achmad Yani
Cimahi, Indonesia
Email: esmeralda.contessa@lecture.unjani.ac.id

1. INTRODUCTION

Automatic emotion recognition is becoming necessary for early detection of emotional disorders and therapy monitoring [1], and as a driver of brain-computer interfaces (BCIs) [2]. One source of emotional information reflecting a person's true and objective condition is electrical signals in the brain [3], recorded via an electroencephalogram (EEG). Accordingly, current research on emotion recognition is of interest [4].

The challenge in emotion recognition is that EEG signals do not have a standard shape and are non-stationary [5], and across multiple channels, data redundancy increases the complexity of emotion identification from EEG signals [6]. Feature extraction is a crucial step in emotion recognition. Usually, it uses the frequency or time-frequency domain, considering the frequency bands relevant to emotion class [7].

This research used data from the SJTU emotion EEG dataset (SEED), provided by the Brain-like Computing and Machine Intelligence (BCMI) laboratory, with three classes: positive, neutral, and negative [8], [9]. These three classes exhibit distinct frequency bands: theta, alpha, beta, and gamma, 4-48 Hz [8], [10]. Positive emotions are generally associated with increases in beta and gamma components, resulting in cognitive enhancement. Positive emotions are generally associated with increased beta and gamma waves, indicating cognitive enhancement. Neutral emotions are balanced across all waves. Negative emotions, on the other hand, are characterized by increased theta and alpha waves. Therefore, frequency band extraction provides distinguishing features for emotion classes [11]. Wavelet transforms are effective for analyzing the EEG signals, enabling time-frequency decomposition. This is achieved by decomposing specific frequency bands to preserve the temporal dynamics essential to emotion. It provides robustness against noise [10].

Previous studies on emotion classification have used convolutional neural networks (CNN), recurrent neural networks [12], and graph neural networks (GNN). CNNs exploit local spatial and spectral correlations among EEG channels and frequency bands. CNNs are effective in capturing discriminative emotional patterns in EEG signals [8], followed by 2D CNNs for multi-channel analysis [13], [14], asymmetric variables across channel pairs [15]. EEG signals from multi-channel arrays often require spatial feature extraction, which can be accomplished with a CNN. Therefore, hybrid CNN-recurrent neural network (RNN) is often used as sequential [16] and parallel [11] approaches. However, these approaches are susceptible to fluctuations and have high computational time.

CNNs are limited to local recognition, while RNNs process sequential data and lose parallel information. Transformers allow each element to be simultaneously and globally spatially connected, similar to the connections of neurons in the brain [17]. Transformers more effectively capture global relations through self-attention mechanisms, making them well-suited for emotion classification from EEG signals. Furthermore, transformer computation is faster for training and capturing spatio-temporal patterns. Research shows that transformer-based models achieve superior performance and robustness in EEG signal classification, allowing for dynamic generalization during learning by reducing irrelevant components or noise [10], [17].

The vision transformer (ViT) has attracted attention in EEG-based emotion recognition due to its robust self-attention mechanism, which enables effective modeling of global dependencies across brain regions and time sequences. When the time-frequency feature extraction results are partitioned into patches, ViT can capture broad contextual relationships that are difficult to model using CNNs or RNNs [17], [18]. ViT relies solely on data-driven learning, making it less effective when trained on medium-sized EEG datasets such as SEED and prone to overfitting. This is especially true when inter-subject variability is high, a challenge that becomes critical as ViT's performance declines significantly [17]. The simplicity of ViT computation allows for faster learning. This is due to its flexibility in handling various EEG signal representations by converting them into image formats, making it compatible with ViT patch tokenization. The patch size is adjustable, allowing ViT to effectively learn hierarchical representations while preserving global structural relationships in multi-channel EEG signals and frequency-band variability. ViT computation can be dynamically scaled and allows for parallel computation by processing all input tokens simultaneously. This method facilitates long-term temporal modeling, addressing RNN issues. ViT allows mapping across EEG signal channels and frequency bands. It contributes to classification in brain mapping.

Previous research used a single-channel ViT [19]. However, EEG signals are recorded from multiple channels and require exploration of the information contained in each frequency band [18]. Therefore, wavelet and CNN need to be combined with ViT. The hybrid model helps improve patch-based ViT learning. Hence, the interpretation and physiological relevance of the learned attention maps are more meaningful. CNN-ViT was used for emotion classification of a single channel [20], [21].

CNN is suitable for feature extraction, but learning gets stuck on neuron localization. Therefore, classification is performed using the ViT to capture the global connectivity of multi-channel EEG signals and various frequency bands with brain connections during specific emotions. The integration of DWT, 2D CNN, and ViT in this paper represents a new step towards improving accuracy, training stability, and efficiency, as well as understanding the relevance of each frequency band to emotions.

The novelty of this research is the integration of a 2D-CNN based on wavelets with ViT for EEG-based emotion classification. Wavelets enable the extraction of frequency bands that correspond to distinct emotion classes. The 2D CNN-ViT architecture allows for more stable and robust generalization of the EEG signal with certain emotions. A hybrid CNN-transformer architecture has been applied to the DEAP dataset from previous research [21].

2. METHOD

2.1. Dataset description

This study employs SEED, a widely used benchmark for EEG-based emotion recognition [8], [9]. The SEED dataset contains EEG recordings from 15 subjects, each participating in 3 experimental sessions. Subjects were exposed to stimuli eliciting three emotional states (positive, neutral, and negative) through emotional film videos. EEG signals were recorded using 62 channels following the international 10–20 system, with a sampling rate of 200 Hz. The reduction from 62 channels to 12 channels was performed using a relevant region-based channel selection approach [15]. Emotions are closely related to the frontal, temporal, and parietal areas. They are FT7, FT8, T7, T8, C5, C6, TP7, TP8, CP5, CP6, P7, and P8 [22]. This approach maintains neurophysiological relevance while reducing redundancy and computational complexity, a particularly beneficial property for transformer-based architectures.

Let $x_c(t) \in R$, denote the raw EEG signal amplitude recorded from the c^{th} EEG channel at a discrete time index t, where c ∈ {1, 2, …, 12}. An EEG trial consisting of 12 channels and T = 1000 points or 5 seconds is represented in (1).

$$\begin{bmatrix} x_1(1) & \cdots & x_1(1000) \\ \vdots & \ddots & \vdots \\ 12(1) & \cdots & x_{12}(1000) \end{bmatrix} \tag{1}$$

2.2. Model design

EEG-based emotion recognition integrates time-frequency analysis, local spatial-spectral feature learning, and global contextual modeling. Raw signals are extracted in the theta, alpha, beta, and gamma frequency bands using discrete wavelet transform (DWT). The presence of each of these bands is strongly correlated with emotional states, facilitating the identification stage [23]. The proposed model, based on a 5-second dataset or 1,000 points, is decomposed using DWT. Then, the 12 channels are extracted using a 2D CNN and mapped using ViT, as shown in Figure 1.

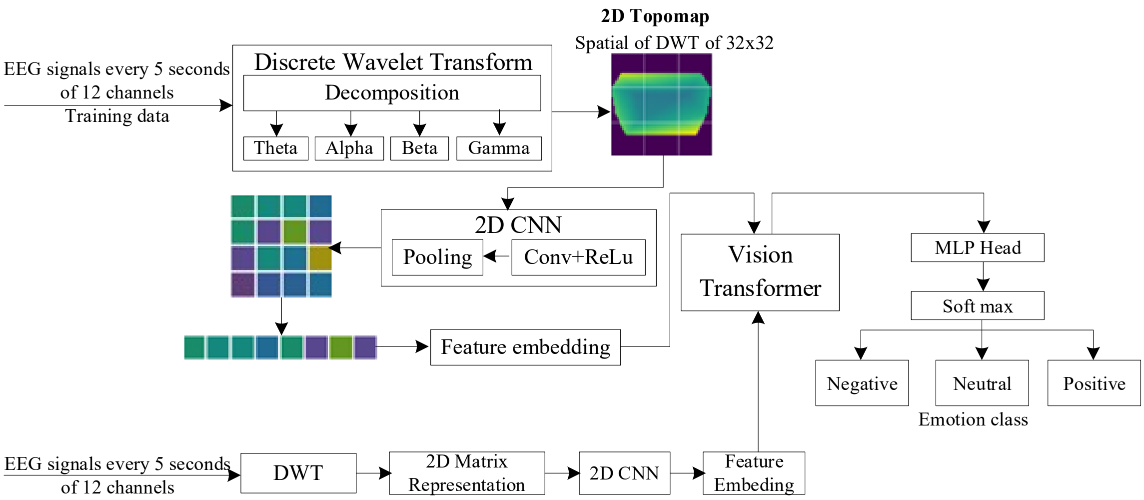


Figure 1. Hybrid 2D CNN-ViT model

Figure 1 describes the complete architecture of the proposed model, including the training and identification stages. During the training stage, EEG data from 15 subjects were collected while they viewed stimulus videos from three emotion classes (positive, neutral, and negative) across 12 main channels. The EEG signals were segmented into five-second segments, a time window duration proven to provide an optimal balance between temporal resolution and stability of emotional features [24]. Then, they were processed using the DWT to extract five frequency bands, including theta, alpha, beta, and gamma waves, which represent cognitive and emotional activity [25]. The DWT results are expressed as a 2D topomap measuring 32 × 32 pixels that displays the distribution of frequency energy against EEG channel positions, an effective technique for preserving spatial information [26]. This topomap is used as input to a 2D CNN to extract local spatial patterns across channels and frequency bands. The result is an 8 × 8 × 256 feature map, which is then flattened into a feature embedding (64 patches × 256 dimensions) and sent to ViT. ViT learns global relationships between patches using a multi-head self-attention mechanism that is superior in capturing long-term dependencies compared to recurrent methods [27]. The output of ViT is passed to the multi-layer perceptron (MLP) head and softmax classifier to determine three emotion classes. The proposed architecture supports real-time emotion prediction, as shown by the bottom path in Figure 1. New EEG data recorded in real time is then processed with DWT → 2D CNN → ViT, and every 5 seconds, it generates a label for one of the three emotion classes.

2.3. Discrete wavelet transforms

After band-pass filtering in the range of 4–48 Hz, the EEG signal is expressed as (2).

$$\tilde{x}_c(t) = \mathcal{F}_{4-48}\{x_c(t)\} \tag{2}$$

The preprocessing stage uses the DWT. This method was chosen because it can analyze non-stationary signals such as EEG and provides detailed information on frequency changes over time. Wavelet-based approaches are superior at capturing abrupt transitions in biomedical signals compared to standard Fourier-based methods [28]. Each EEG signal segment from the previous stage is decomposed to produce five main frequency bands: theta (4–8 Hz), alpha (8–13 Hz), beta (14–30 Hz), and gamma (31–48 Hz) [29], which are related to emotional states. For example, the alpha band often appears in relaxed states, while the beta and gamma bands are more dominant during times of high mental activity or strong emotions [30]. Recent computational neurophysiology studies confirm that power fluctuations in the high-frequency bands (beta and gamma) are the most consistent biomarkers for distinguishing positive and negative emotional valence [25]. Applying the DWT to the filtered EEG signal yields wavelet coefficients as shown in (3).

$$\tilde{x}_c(t) = \sum_k a_{J,k}^{(c)}\, \phi_{J,k}(t) + \sum_{j=1}^{J} \sum_k d_{j,k}^{(c)}\, \psi_{j,k}(t) \tag{3}$$

Where $a_{J,k}^{(c)}$ is channel approximation coefficient c, $d_{j,k}^{(c)}$ is detail coefficient on scale j, and $\phi(t), \psi(t)$ are scale function and wavelet. Wavelet coefficients are grouped into emotion-related EEG bands as (4).

$$\mathrm{W}_c = [\mathrm{W}_c^{\theta}, \mathrm{W}_c^{\alpha}, \mathrm{W}_c^{\beta}, \mathrm{W}_c^{\gamma}] \tag{4}$$

Where θ is 4–8 Hz, α is 8–13 Hz, β is 13–30 Hz, and γ is 30–48 Hz.

The DWT decomposition results for each EEG channel are then arranged into a 2D matrix that represents the energy distribution of each frequency band across the channels. This 2D representation is used as input for a 2D-CNN to recognize spatial patterns between channels and frequency characteristics related to emotions. This transformation to an image matrix format allows the CNN model to exploit spatial correlations between electrodes that are often lost when using only one-dimensional vectors [26]. For each trial, the wavelet features are arranged into a 2D representation as (5).

$$M \in \mathbb{R}^{C \times B} \tag{5}$$

Where C is 12, B is 4 (number of frequency bands), and M is channel–frequency EEG feature map.

2.4. Hybrid 2D convolutional neural network–vision transformer

A hybrid model combining CNN and ViT. At this stage, a convolution operation is applied to extract local patterns from the 2D DWT representation. The convolution operation is formulated in (6).

$$S(i,j) = (X \times K)(i,j) = \sum_m \sum_n X\,(i-m, j-n).\, K(m,n) \tag{6}$$

Where X is the input matrix resulting from the DWT transformation, K is the convolution kernel, and $S(i,j)$ is the convolutional feature map.

CNNs extract important emotional features in specific frequency bands from multi-channel EEG signals. Multiple convolutional layers, activation functions, and max pooling operations reduce dimensionality without losing information, which are key advantages of CNNs. Meanwhile, ViT models global relationships among features via a self-attention mechanism [31]. The output of the CNN is converted into a fixed-size patch embedding. ViT utilizes two attention heads and one encoder layer, each equipped with multi-head self-attention and a feed-forward network. The CNN feature map is divided into N patches as (7).

$$Z_0 = [z_0^1, z_0^2, \ldots, z_0^N] \tag{7}$$

$$z_0^i = p_i E + e_i \tag{8}$$

Where p_i is patch EEG i-th, E is matrix embedding, and e_i is positional encoding.

The output of ViT is passed to a fully connected layer and finally determines the emotion category: positive, neutral, or negative using softmax [32]. The function use (9).

$$\hat{y} = softmax(WczL + bc) \quad (9)$$

Where $\hat{y} \in \mathbb{R}^3$ is the probability of class: positive, neutral, and negative. The dataset is split 80% for training and 20% for testing, a standard split ratio used in the SEED dataset to avoid overfitting and validate model generalization [33]. The pre-experimental results determined the parameters of batch size 4, AdamW optimizer, and learning rate 0.0001, as in Table 1.

Table 1. Architecture of hybrid DWT-2D CNN-ViT

Layer	Configuration	Output shape	Number of parameters
Input layer	EEG (32 × 32 × 3)	32 × 32 × 3	0
Conv2D_1	32 filter, kernel 3 × 3, ReLU	30 × 30 × 32	896
MaxPooling2D_1	Pool 2 × 2	15 × 15 × 32	0
Conv2D_2	64 filter, kernel 3 × 3, ReLU	13 × 13 × 64	18,496
Flatten	-	10816	0
Patch embedding	16 × 16 patch	16 × 256	-
ViT encoder	1 layer, 2 head, MLP dim 256	16 × 256	-
Dense (softmax)	Output = 3 classes	3	771
	Number of parameters		20,000

Table 1 shows that the CNN component extracts spatial features from the DWT results, while ViT learns global relationships across channels and time. This model configuration improves the model's ability to recognize emotional patterns, as measured by accuracy, stability, and time. Throughout the training process, accuracy values increased steadily, indicating that the model effectively and efficiently learned spatial patterns and relationships among EEG channels. The AdamW optimizer was selected based on its ability to separate weight decay from the gradient adaptation step, a common weakness of the standard Adam. Comparative studies have shown that AdamW converges faster and generalizes better than classical optimization methods such as RMSProp or regular Adam [34].

The model training process uses a cross-entropy loss function to minimize prediction errors in multi-class classification. This function works by calculating the logarithmic difference between the softmax-predicted probabilities and the actual class labels, thereby forcing the model to increase the confidence in the correct class [35]. All experiments were implemented using the PyTorch framework on a graphics processing unit (GPU)-powered computing environment for accelerated CNN and self-attention matrix operations.

3. RESULTS AND DISCUSSION

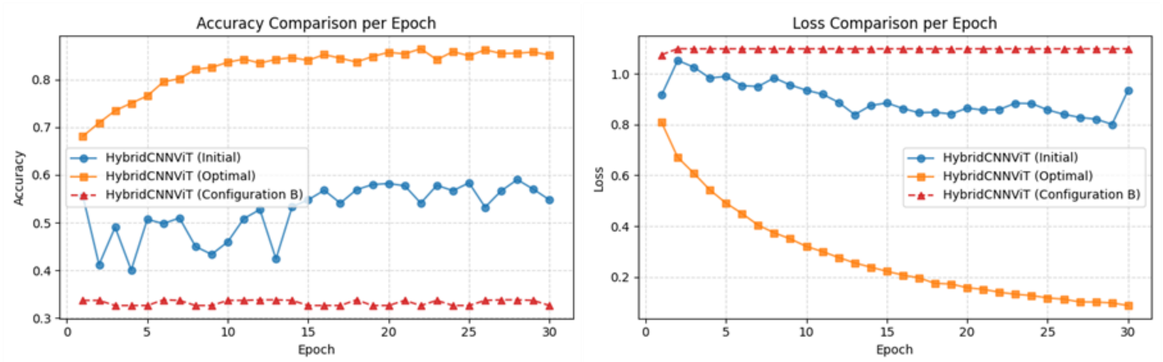
3.1. Configuration design experiment

To obtain the optimal model configuration, experiments were conducted on several key parameters of the hybrid CNN–ViT architecture, including positional embeddings, ViT dimensions, the number of attention heads, the number of encoder layers, and the learning rate, which are presented in Table 2. Table 2 shows that increasing model capacity does not always result in better performance. In configuration B, the ViT dimension was increased to 256 with eight attention heads, but accuracy actually decreased by 33%. Configurations A and B are too complex, leading to unstable learning and overfitting. Conversely, configuration C is more lightweight and achieves 87% accuracy. This simpler architecture can adjust the learning rate and improve training efficiency while maintaining gradient stability. The optimal configuration (C) achieved the highest performance with 87% accuracy, significantly outperforming the initial setup (A) by 28%. Furthermore, the minimal loss of 0.10 in configuration C indicates a significantly more precise model compared to the others. These results confirm that the applied optimization successfully stabilized the training process and maximized learning ability.

Table 2. Performance comparison between configurations

Configuration	Positional embedding	ViT (dim-head-layer)	Learning rate	Accuracy (%)	Loss	Convergence remarks
A (initial)	Random	128 – 4 – 2	10-3	59	0.80	Highly fluctuating
B	Random	256 – 8 – 2	10-3	33	1.10	Model not learning
C (optimal)	Null	128 – 2 – 1	10-4	87	0.10	Stability and fast

Figure 2 shows the transient-state performance from Table 3 of three model designs. Figure 2(a) shows the accuracy performance. Initially, they have an accuracy of 59% with fluctuations. Configuration A appears unstable due to the random embedding. However, the optimal configuration B continues to improve and reaches 87%. Configuration B remains low, indicating underfitting due to excessive complexity and a poor learning rate. For the loss value shown in Figure 2(b), the orange curve decreases rapidly from 0.809 to 0.087, indicating solid convergence is occurring. The initialization results remain high and unstable, while configuration B fails to significantly reduce the loss, indicating underfitting.



(a) (b)

Figure 2. Performance of hybrid CNN-ViT of (a) accuracy and (b) loss

Table 3 presents a detailed comparison of precision, recall, and F1-scores for each class. As observed, the proposed improvements resulted in significant performance gains across all categories. In the negative class, precision increased sharply from 36% to 84%, while recall rose from 12% to 80%, indicating that the optimized model successfully identifies previously missed negative emotions. Similar enhancements are evident in the neutral class, where precision improved from 47% to 82%, with recall remaining stable at a high level (85% to 86%). Meanwhile, the positive class achieved the most consistent results, with precision rising from 80% to 94% and recall from 66% to 93%, culminating in a high F1-score of 93%. These metrics confirm that the optimal configuration offers superior sensitivity and precision compared to the initial model.

Table 3. Performance comparison between the initial and the improvisation

Class	Precision (%)		Recall (%)		F1-score (%)	
	Initial	Optimal	Initial	Optimal	Initial	Optimal
Negative	36	84	12	80	18	82
Neutral	47	82	85	86	60	84
Positive	80	94	66	93	72	93

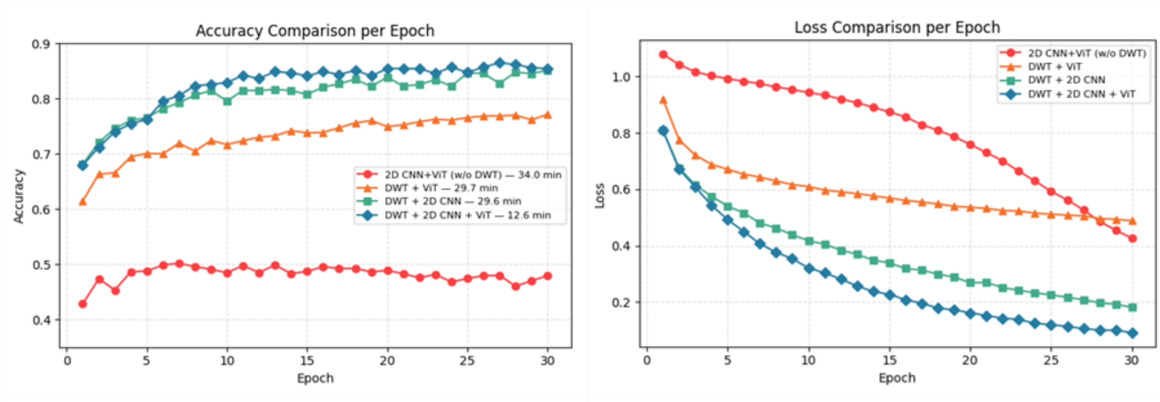
3.2. Effect of hybrid component: ablation testing

The model configuration was ablated by removing one of the components of DWT, 2D CNN, and ViT, on its performance, as shown in Table 4. It is shown that DWT improves accuracy by 38.76% and reduces time by 21.68 minutes. Meanwhile, 2D CNN spatial extraction improves accuracy by 9.53% and reduces time by 17.3 minutes. ViT achieves an accuracy increase of 1.58% but reduces computation time significantly, namely 40%, characteristic of training.

Table 4. Ablation testing of DWT, 2D CNN, ViT

Model	Accuracy (%)	Loss	Learning time (minutes)
2D CNN+ViT (without DWT)	47.93 (↓)	0.42	34.02 (↓)
DWT+ViT	77.16 (↑)	0.48	29.70 (↓)
DWT+2D CNN	85.11 (↑)	0.18	29.65 (↓)
DWT+2D CNN+ViT	86.69 (↑)	0.08	12.34 (↓)

Figure 3 shows that the model without DWT has an accuracy below 50% and a loss above 1.0 during training. In contrast, the DWT + 2D CNN + ViT model shows fast and stable convergence starting from the 10th epoch. The combination of DWT, CNN, and ViT is efficient and robust. The performance in each case in Table 4 can be illustrated in Figure 3(a) for accuracy, which presents a marked improvement over configurations without DWT. Then, the feature reduces dimensionality while still preserving local spatial features for use in CNNs. Simultaneously, ViT is also added, achieving the highest accuracy, with improved training stability and reduced computation time. Figure 3(b) presents the corresponding loss comparison. The DWT–CNN–ViT model provides the lowest loss value, converges, and is stable. These results confirm that DWT plays a crucial role in extracting features in relevant frequency bands. CNN effectively models local spatial patterns, and ViT captures simultaneous spatio-temporal relationships through self-attention.



(a) (b)

Figure 3. Performance of each component of (a) accuracy and (b) loss

3.3. Effect of learning weight correction techniques

After determining the configuration and main components, several optimization algorithms were assessed for their impact on training. The optimizers compared were Adam, AdamW, AdaMax, NAdam, and RMSProp, all with learning rate of 1×10^{-4}. The performance comparison results are presented in Table 5.

Table 5. Comparison of some optimizers

Optimizer	Accuracy (%)	Loss	Learning time (minutes)
Adam	86	0.091	12.35
AdamW	87	0.087	12.34
AdaMax	83	0.128	13.50
NAdam	86	0.093	18.04
RMSProp	86	0.105	15.38

Based on Table 6, all optimizers demonstrated relatively stable performance with accuracy above 82%. The differences in results across optimizers were not significant, indicating that the model had achieved architectural stability; thus, changes to the optimization algorithm did not significantly affect the final performance. The more influential factors affecting accuracy were model configuration, architectural components (DWT, 2D CNN, and ViT), and learning rate adjustments. Nevertheless, AdamW was chosen as the primary optimizer because it provided the best balance between gradient stability, 87% accuracy, and an efficient learning time of 12 minutes.

3.4. Comparison with previous research

To assess the effectiveness of the proposed model, a comparison was conducted with several previous studies on the SEED dataset. The goal was to evaluate the performance improvement of the DWT + hybrid 2D CNN-ViT method compared to CNN- and RNN-based approaches, as presented in Table 6. The performance of the proposed method was tested on the same dataset, SEED. Table 6 shows that the DWT + hybrid 2D CNN-ViT approach successfully surpassed previous research. The proposed method achieved an accuracy of 86.69%, higher than the CNN-only and CNN-RNN approaches. The integration of DWT, CNN, and ViT improves the model's ability to recognize emotional patterns more accurately, efficiently, and stably and has the potential to be applied to emotion monitoring.

Table 6. Comparison with previous research

Method	Accuracy (%)
Differential entropy (DE) + SVM and graph regularized extreme learning machine (GELM) [36]	79
Wavelet + 2D CNN [13]	83
Wavelet + multiple CNN [14]	81
Hybrid parallel 2D CNN-GRU [11]	84
DWT + hybrid 2D CNN-ViT (proposed methods)	87

4. CONCLUSION

The research shows a way to emotion recognition of EEG signals using a hybrid wavelet transform, the 2D CNN, and the ViT. This seems to improve the recognition of emotional patterns and the robustness overall. Wavelet transform with DWT gets the relevant frequency bands and handles the non-stationary aspects of the EEG signals. The 2D CNN captures local spatial features, while the ViT learns global relationships between features through self-attention, enabling it to capture more complex patterns than a single CNN method. The combination of these three methods provided good classification performance in distinguishing emotional states. The evaluation metrics include accuracy, precision, recall, and F1-score, which gave good performance. This approach demonstrates high potential for application in BCI systems and real-time emotional monitoring. Architectural configurations such as the number of attention heads, encoder layers, parameter selection, and weight correction techniques or optimizers significantly influence the accuracy, stability, and robustness of emotion recognition. The AdamW optimizer provides accuracy improvements, although not significantly more than the other architectural configurations. Overall, the recognition performance of the three emotion classes is relatively similar. However, the positive emotion class provides superior accuracy compared to the other classes. This is not only due to the characteristics of training data, but also because the positive emotion class is relatively easy to recognize from its more distinctive frequency band, which is enhanced by the use of wavelet transforms. Future work will focus on extending the proposed models for cross-dataset evaluation, exploring adaptive channel selection, and investigating lightweight transformer variants to improve generalization and real-time performance.

FUNDING INFORMATION

This research is funded by the *penelitian fundamental regular* 2025 from the Indonesian Directorate General of Research and Community Service, Ministry of Higher Education, Science, and Technology under contract 125/C3/DT.05.00/PL/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Esmeralda Contessa Djamal	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓
Revan Vio Endriansyah		✓	✓			✓	✓	✓	✓	✓	✓			
Daswara Djajasasmita		✓		✓	✓	✓			✓	✓		✓		

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the SEED dataset consists of EEG datasets provided by the BCMI laboratory, available at https://bcmi.sjtu.edu.cn/home/seed/, with 5-second segmentations. Coding of the experiment available in GitHub at https://github.com/RevanVE/eeg-emotion-classification-with-VIT.git.

REFERENCES

[1] P. Leszczełowska and N. Dawidowska, "Systematic literature review for emotion recognition from EEG signals," *New Trends in Database and Information Systems (ADBIS 2022)*. Cham, Switzerland: Springer, pp. 467–475, 2022, doi: 10.1007/978-3-031-15743-1_43.
[2] Y. Wang *et al.*, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 83–84, pp. 19–52, Jul. 2022, doi: 10.1016/j.inffus.2022.03.009.
[3] J. N. Schneider, M. Matyjek, A. Weigand, I. Dziobek, and T. R. Brick, "Subjective and objective difficulty of emotional facial expression perception from dynamic stimuli," *PLOS ONE*, vol. 17, no. 6, Jun. 2022, doi: 10.1371/journal.pone.0269156.
[4] W. Ma, Y. Zheng, T. Li, Z. Li, Y. Li, and L. Wang, "A comprehensive review of deep learning in EEG-based emotion recognition: classifications, trends, and practical implications," *PeerJ Computer Science*, vol. 10, May 2024, doi: 10.7717/peerj-cs.2065.
[5] A. Chaddad, Y. Wu, R. Kateb, and A. Bouridane, "Electroencephalography signal processing: a comprehensive review and analysis of methods and techniques," *Sensors*, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146434.
[6] Z. Zhang, J. M. Fort, and L. G. Mateu, "Mini review: challenges in EEG emotion recognition," *Frontiers in Psychology*, vol. 14, Jan. 2024, doi: 10.3389/fpsyg.2023.1289816.
[7] J. Wang and M. Wang, "Review of the emotional feature extraction and classification using EEG signals," *Cognitive Robotics*, vol. 1, pp. 29–40, 2021, doi: 10.1016/j.cogr.2021.04.001.
[8] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, Sep. 2015, doi: 10.1109/TAMD.2015.2431497.
[9] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering*, Nov. 2013, pp. 81–84, doi: 10.1109/NER.2013.6695876.
[10] K. Kamble and J. Sengupta, "A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27269–27304, Jul. 2023, doi: 10.1007/s11042-023-14489-9.
[11] N. A. Putri, E. C. Djamal, F. Nugraha, and F. Kasyidi, "EEG emotion recognition using parallel hybrid convolutional-recurrent neural networks," in *2022 International Conference on Data Science and Its Applications*, Jul. 2022, pp. 24–29, doi: 10.1109/ICoDSA55874.2022.9862853.
[12] H. M. M. Hasan and M. A. Islam, "Emotion recognition from Bengali speech using RNN modulation-based categorization," in *2020 Third International Conference on Smart Systems and Inventive Technology*, Aug. 2020, pp. 1131–1136, doi: 10.1109/ICSSIT48917.2020.9214196.
[13] R. H. D. Haqque, E. C. Djamal, and A. Wulandari, "Emotion recognition of EEG signals using wavelet filter and convolutional neural networks," in *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications*, Sep. 2021, pp. 1–6, doi: 10.1109/ICAICTA53211.2021.9640279.
[14] H. D. Khairunissa, E. C. Djamal, and A. Wulandari, "Multiple convolutional neural networks in EEG emotion recognition," in *2021 4th International Conference of Computer and Informatics Engineering*, Sep. 2021, pp. 30–35, doi: 10.1109/IC2IE53219.2021.9649215.
[15] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul. 2018, pp. 1561–1567, doi: 10.24963/ijcai.2018/216.
[16] Y. Ma *et al.*, "Emotion recognition model of EEG signals based on double attention mechanism," *Brain Sciences*, vol. 14, no. 12, Dec. 2024, doi: 10.3390/brainsci14121289.
[17] W. Tao *et al.*, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 382–393, Jan. 2023, doi: 10.1109/TAFFC.2020.3025777.
[18] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, pp. 1–21.
[19] A. Arjun, A. S. Rajpoot, and M. R. Panicker, "Introducing attention mechanism for EEG signals: emotion recognition with vision transformers," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, Nov. 2021, pp. 5723–5726, doi: 10.1109/EMBC46164.2021.9629837.
[20] M. Mansour, L. N. Mahmoud, I. Sadek, and N. M. Salem, "Emotion recognition from EEG signals using vision transformer and CNN," in *2025 International Telecommunications Conference (ITC-Egypt)*, Jul. 2025, pp. 259–264, doi: 10.1109/ITC-Egypt66095.2025.11186562.
[21] K. Devarajan, S. Ponnan, and S. Perumal, "Hybrid CNN-transformer architecture for enhanced EEG-based emotion recognition: capturing local and global dependencies with self-attention mechanisms," *Discover Computing*, vol. 28, no. 1, May 2025, doi: 10.1007/s10791-025-09596-0.
[22] Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for EEG emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143303–143311, 2019, doi: 10.1109/ACCESS.2019.2944273.
[23] A. Subasi, T. Tuncer, S. Dogan, D. Tanko, and U. Sakoglu, "EEG-based emotion recognition using tunable Q wavelet transform and rotation forest ensemble classifier," *Biomedical Signal Processing and Control*, vol. 68, Jul. 2021, doi: 10.1016/j.bspc.2021.102648.
[24] P. Keelawat, N. Thammasan, M. Numao, and B. Kijsirikul, "A comparative study of window size and channel arrangement on EEG-emotion recognition using deep CNN," *Sensors*, vol. 21, no. 5, Mar. 2021, doi: 10.3390/s21051678.
[25] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021, doi: 10.1109/TNNLS.2020.3008938.
[26] S. K. Gatfan, "A review on deep learning for electroencephalogram signal classification," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 16, no. 1, pp. 137–151, Mar. 2024, doi: 10.29304/jqcsm.2024.16.11453.
[27] E. Vafaei and M. Hosseini, "Transformers in EEG analysis: a review of architectures and applications in motor imagery, seizure, and emotion classification," *Sensors*, vol. 25, no. 5, Feb. 2025, doi: 10.3390/s25051293.
[28] S. Li, X. Lyu, L. Zhao, Z. Chen, A. Gong, and Y. Fu, "Identification of emotion using electroencephalogram by tunable Q-factor wavelet transform and binary gray wolf optimization," *Frontiers in Computational Neuroscience*, vol. 15, Sep. 2021, doi: 10.3389/fncom.2021.732763.
[29] O. Saha, M. S. Mahmud, S. A. Fattah, and M. Saquib, "Automatic emotion recognition from multi-band EEG data based on a deep learning scheme with effective channel attention," *IEEE Access*, vol. 11, pp. 2342–2350, 2023, doi: 10.1109/ACCESS.2022.3224725.

[30] M. A. Maria, M. A. H. Akhand, A. B. M. A. Hossain, M. A. S. Kamal, and K. Yamada, "A comparative study on prominent connectivity features for emotion recognition from EEG," *IEEE Access*, vol. 11, pp. 37809–37831, 2023, doi: 10.1109/ACCESS.2023.3264845.
[31] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "EEG-based emotion recognition with deep convolutional neural networks," *Biomedical Engineering / Biomedizinische Technik*, vol. 66, no. 1, pp. 43–57, Feb. 2021, doi: 10.1515/bmt-2019-0306.
[32] T. Dhara and P. K. Singh, "Emotion recognition from EEG data using hybrid deep learning approach," in *Frontiers of ICT in Healthcare*, Singapore: Springer, 2023, pp. 179–189, doi: 10.1007/978-981-19-5191-6_15.
[33] A. Prakash and A. Poulose, "Electroencephalogram-based emotion recognition: a comparative analysis of supervised machine learning algorithms," *Data Science and Management*, vol. 8, no. 3, pp. 342–360, Sep. 2025, doi: 10.1016/j.dsm.2024.12.004.
[34] C. Desai, "Comparative analysis of optimizers in deep neural networks," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 10, pp. 959–962, 2020.
[35] Z. Fan, F. Chen, X. Xia, and Y. Liu, "EEG emotion classification based on graph convolutional network," *Applied Sciences*, vol. 14, no. 2, Jan. 2024, doi: 10.3390/app14020726.
[36] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, Jul. 2019, doi: 10.1109/TAFFC.2017.2712143.

BIOGRAPHIES OF AUTHORS

Esmeralda Contessa Djamal received a bachelor's degree in Engineering Physics from Institut Teknologi Bandung in 1994, a master's degree in Instrument and Control from the same university in 1998. From her dissertation until now, her research in EEG signal processing has finished her doctoral program at Institut Teknologi Bandung in 2005. She is an associate professor in the Department of Informatics, Universitas Jenderal Achmad Yani. For more than 20 years, she has researched EEG signal processing and machine learning. She can be contacted at email: esmeralda.contessa@lecture.unjani.ac.id.

Revan Vio Endriansyah is an undergraduate student in Informatics from Universitas Jenderal Achmad Yani. His thesis focused on emotion identification from EEG signals using a transformer. He is now a machine learning researcher and system developer. He can be contacted at email: revanvio22@if.unjani.ac.id.

Daswara Djajasasmita is a neurologist and lecturer at the Medical Education Program, Faculty of Medicine at Universitas Jenderal Achmad Yani. His research interests are stroke and post-stroke management, including signal processing. He can be contacted at email: daswaradj@gmail.com.