EEG-Based Emotion Recognition Using a Hybrid CNN–Transformer Model

Duy Nguyen
Graduate School of Science and Technology
Gunma University
Gunma, Japan
t232b604@gunma-u.ac.jp

Minh Tuan Nguyen
Faculty of Telecommunications
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
nmtuan@ptit.edu.vn

Kou Yamada
Graduate School of Science and Technology
Gunma University
Gunma, Japan
yamada@gunma-u.ac.jp

***Abstract*—Accurate recognition of human emotions is essential in many fields such as education, healthcare, and entertainment, and is particularly valuable for improving the adaptability of brain–computer interface (BCI) systems in human–computer interactions. Therefore, in this study, we propose a hybrid deep learning approach that combines a one-dimensional convolutional neural network (1D-CNN) with a Transformer encoder, referred to as the 1D-CNN–Transformer, for emotion classification based on electroencephalogram (EEG) signals using two subsets of selected channels. RF-RFE is employed to select the most informative EEG channels, followed by the extraction of time-domain and frequency-domain features across four frequency bands decomposed by the discrete wavelet transform (DWT). These features are evaluated using three models including 1D-CNN, Multilayer perceptron, and the proposed 1D-CNN–Transformer, with subject-wise 5-fold cross-validation on the validation dataset. Among the models, the hybrid 1D-CNN-transformer model achieves the best results, with 70.0% accuracy, 72.9% precision, 82.7% recall, and a 76.7% F1-score for valence, and 67.5% accuracy, 68.89% precision, 82.71% recall, and a 73.9% F1-score for arousal.**

***Index Terms*—Emotion Recognition, Discrete Wavelet Transform, Deep Learning, Transformer Model, Channel Selection.**

I. Introduction

Emotions play an important role in the human experience, influencing behavior, mental health, relationships, and interactions with technology. Therefore, the application of emotion recognition technologies offers significant advantages in many fields, including brain-computer interfaces (BCI), healthcare, security, e-commerce, education, and entertainment [1]. An intelligent emotion detection system can be used with either physiological or non-physiological emotion detection methods. While non-physiological signals, such as facial expressions, speech, and behavioral patterns, can be consciously controlled or concealed by the user, making them less reliable for emotion recognition [2]. On the other hand, physiological electroencephalogram (EEG) signals have gained significant prominence in the field of affective computing due to their safety, cost-effectiveness, non-invasiveness, user-friendliness, portability, and ability to maintain high temporal resolution. The reliability and objectivity of EEG signals in reflecting emotions, which are independent of human intention [3], have led to EEG being widely used in emotion recognition [4]. EEG-based emotion recognition provides an objective and physiologically based approach to understanding, identifying, and applying emotional information, with profound implications for improving mental health [5], enhancing human-computer interactions, and improving the quality of human-computer interactions.

Deep learning techniques have been extensively applied to capture complex spatial–temporal patterns in EEG data, thereby improving the performance of emotion recognition systems. Long short-term memory networks (LSTM), and convolutional neural networks (CNNs) are widely used [6], [7]. Recently, the Transformer architecture has demonstrated powerful capabilities for modeling long-range dependencies and has achieved state-of-the-art performance in various domains [8].

Motivated by the effectiveness of CNN models and transformer model, this study proposes an efficient hybrid 1D-CNN-transformer model for EEG-based emotion recognition. The main contributions of this work are as follows:

1) A feature extraction framework is designed by combining discrete wavelet transform (DWT) with time and frequency domain features, ensuring informative representations of EEG signals.
2) Channel selection is performed using the Random forest combined with recursive feature elimination (RF-RFE) method to identify the most discriminative channels for valence and arousal classification.
3) Two hybrid 1D-CNN–Transformer architectures are proposed, tailored separately for valence and arousal models to enhance performance of classification.

The remaining part of the paper is organized as follows: Section II presents the proposed method. The simulation results and comparison are presented in Section III. Finally, Section IV summarizes the research.

II. Method

The proposed method is illustrated in Fig. 1, consisting of four main steps. First, the data are preprocessed, after which DWT is applied to decompose the EEG signals into four frequency bands, including theta (4-8Hz), alpha (8-16Hz), beta (16-32Hz), and gamma (32-45Hz). In the second step,

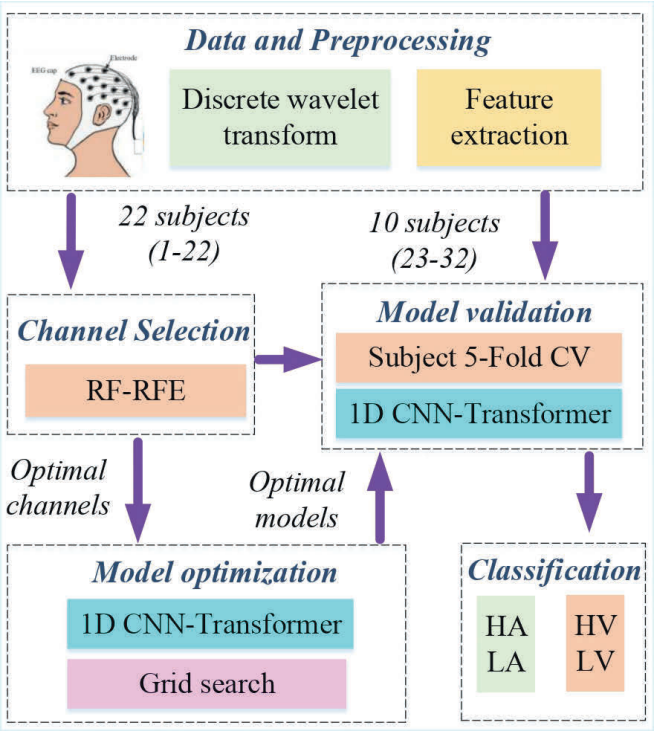

Fig. 1. Flowchart of the method

the combination of the Random forest model and the Recursive feature elimination (RF-RFE) algorithm is employed for channel selection. Next, the hybrid CNN-Transformer model is optimized using grid search. Finally, model validation is performed with subject-wise 5-fold cross-validation (CV), where 80% of the subjects are used for training and 20% for testing in each of the five iterations.

A. Data and preprocessing

1) Data: The database for emotion analysis using physiological signals (DEAP) [9] is a widely applied benchmark for EEG-based emotion recognition research. The dataset consists of recordings from 32 participants including 16 males and 16 females, aged between 19 and 37. Each participant watched 40 one-minute music videos and then rated them on a scale from 1 to 9 for valence, arousal, dominance, and liking. EEG signals were recorded for each trial, starting 3 seconds before the start of the video and continuing until the end of the video, resulting in a total duration of 63 seconds per trial. Data collection used 40 channels, including 32 EEG electrodes and 8 peripheral physiological sensors. The original recordings were sampled at 512 Hz and then downsampled to 128 Hz during preprocessing. Electrooculogram (EOG) artifacts were reduced using a band pass filter in the range of 4 to 45 Hz.

After preprocessing and removing the 3-second baseline, the DEAP dataset is structured in a 32×40×32×8064 format (subject×trials×channels× samples), while the corresponding sentiment labels were arranged in a 40×4 matrix (trials× labels), including Valence, arousal, liking and dominance.

The DEAP dataset is divided into 22 subjects for training (participants from 1 to 22) and 10 subjects for validation (participant from 23 to 32).

2) Labeling: In this paper, two emotional dimensions are chosen to label emotions, namely valence and arousal. Each dimension is scored on a scale from 1 to 9, and a threshold of 4.5 is employed. Binary classification is performed for two scenarios using this threshold: high/low valence (HV/LV) and high/low arousal (HA/LA).

3) Signal decompotion: DWT is used to decompose the signal into four bands:

- Theta (4–8 Hz): associated with memory retrieval and emotional regulation.
- Alpha (8–16 Hz): reflects a state of relaxed alertness and is typically suppressed during periods of high arousal.
- Beta (16–32 Hz): linked to active concentration and anxiety.
- Gamma (32–45 Hz): related to cross-modal sensory integration and peak arousal.

4) Feature extraction: For each of the four frequency bands and each channel, a set of frequency- and time-domain features is extracted. The frequency-domain features include band power, spectral entropy, power spectral density, relative intensity ratio, maximum of power spectral density, maximum frequency, spectral peak, and median frequency. In addition, time-domain features are computed, such as Hjorth parameters (activity, mobility, and complexity), zero-crossing rate, mean, root mean square, variance, skewness, and kurtosis. Therefore, for each channel and band, 16 features are extracted, resulting in a total of 2048 features.

B. Channel selection

To reduce dimensionality and eliminate redundant information, channel selection is performed using the Random Forest–Recursive Feature Elimination (RF-RFE) method. Random Forest provides an importance ranking of EEG channels based on their contribution to classification performance. RFE then iteratively removes the least important channels, retraining the model at each step, until the optimal subset of channels is obtained. This approach ensures that the most informative EEG channels are preserved, improving both classification accuracy and computational efficiency.

C. Model optimization

The proposed model is a hybrid architecture that integrates a one-dimensional convolutional neural network (1D-CNN) with a Transformer encoder. The 1D-CNN acts as an automatic feature extractor, capturing local temporal patterns and short-range dependencies from EEG signals, while the Transformer enhances the modeling of long-range dependencies through self-attention mechanisms. To optimize the proposed model, a grid search strategy is applied for hyperparameter tuning, ensuring robust performance.

The 1D-CNN part is organized into n convolutional blocks. Each block comprises two 1D convolutional layers, each followed by a ReLU activation and batch normalization,

and ends with a max-pooling layer. The Transformer part consists of m encoder layers. Each layer includes a multi-head self-attention mechanism, layer normalization, and a position-wise feed-forward network with residual connections. The self-attention captures long-range dependencies across the extracted features, while the feed-forward layers enhance non-linear representations. Finally, the outputs of the 1D-CNN-Transformer model are passed through fully connected layers and a sigmoid activation for emotion classification.

In order to validate the effectiveness of the proposed CNN-Transformer model, its performance is compared with two baseline models: 1D-CNN and a multilayer perceptron (MLP). This comparison highlights the improvements achieved by incorporating the Transformer module into the CNN framework.

D. Model validation

The optimal model configuration and the selected EEG channels are evaluated on the validation set using subject-wise 5-fold cross-validation. In this scheme, 10 subjects are divided into five folds, with 80% of the subjects used for training and 20% for testing in each iteration. This process is repeated five times so that each subject is included in the test set once. Subject-wise cross-validation ensures that the evaluation reflects generalization across different participants, thereby providing a reliable assessment of the proposed model and the selected channel subset.

III. Simulation Results and Comparison

A. Channel selection

By applying the RF-RFE method, the optimal EEG channels were identified for each emotional dimension. In the arousal model, 10 channels are identified as optimal, namely FC1, T7, C3, Pz, F8, P3, F7, CP6, T8, and Fz. For the valence model, the selected set consists of 10 channels, including Cz, T7, O2, Oz, CP6, FC6, O1, Fz, and AF4. Therefore, a total of 640 features for each model.

B. Optimal architecture of the model

1) Valance model classification: The MLP model is composed of three fully connected layers. The 1D-CNN model consists of four convolutional blocks. The hybrid 1D-CNN–Transformer model integrates three convolutional blocks, with two Transformer encoder layers.

2) Arousal model classification: The MLP model includes four fully connected layers. The 1D-CNN architecture contains four convolutional blocks. The hybrid 1D-CNN–Transformer model consists of four convolutional blocks combined with two Transformer encoder layers.

C. Model validation

We use performance metrics as accuracy (Acc), precision (Pre), Recall (Re), and F1-score (F1) to evaluate the performance of the DL models. Acc measures the proportion of correctly predicted samples over the total number of samples. Pre quantifies the proportion of correctly predicted positive observations among all predicted positives. Re, also known as sensitivity, measures the proportion of actual positives that are correctly identified. The F1 represents the harmonic mean of precision and recall, reflecting the ability of models to identify high and low valence and arousal.

TABLE I
PERFORMANCE OF MODELS ON THE VALIDATION SET

Model	Acc(%)	Pre(%)	Re(%)	F1(%)
		Valence		
1D-CNN	62.0±2.45	63.21±3.7	96.3±7.4	76.5±2.0
MLP	58.75±9.4	66.1±7.3	80.74±18.3	70.2±10.7
Proposed	**70.0±5.5**	**72.9±6.5**	**82.7±7.4**	**76.7±8.9**
		Arousal		
1D-CNN	57.50±2.62	59.84±2.36	85.22±10.3	68.16±9.7
MLP	54.0±10.79	61.2±8.7	64.09±21.6	57.8±15.8
Proposed	**67.5±6.7**	**68.89±9.0**	**82.71±9.8**	**73.9±7.5**

The performance of the proposed hybrid 1D-CNN–Transformer model outperforms the baseline models (1D-CNN and MLP) on both valence and arousal classification tasks, as shown in Table I. For valence, the proposed model achieves the highest Acc of 70.0%, Pre of 72.9%, Re of 82.7%, and F1 of 76.7%, surpassing the 1D-CNN and MLP models across all evaluation metrics. Similarly, for arousal, the proposed model attains 67.5% Acc, 68.9% Pre, 82.7% Re, and 73.9% F1, which represent clear improvements compared to the baseline methods. These results demonstrate that the integration of the Transformer with 1D-CNN effectively enhances feature representation by capturing both local and long-range dependencies, thereby improving the robustness and generalization of emotion recognition.

D. Comparison with Existing Works

A comparison between the our proposed hybrid 1D-CNN–Transformer model and existing approaches is presented in Table II. In [10], EEGNet was applied directly to raw EEG signals using eight channels, achieving Acc of 61.4% accuracy and F1 of 60.2% for valence classification. In [11], 3D spatial–spectral features extracted using the Welch method across different sub-bands, combined with a Graph Convolutional Network (GCN) model, yielded 57.7% accuracy for valence and 58.3% accuracy for arousal. By contrast, our proposed method, which incorporates RF-RFE channel selection, handcrafted feature extraction, and the 1D-CNN–Transformer architecture, achieves significantly higher performance with 70.0% accuracy and 76.7% F1-score for valence, as well as 67.5% accuracy and 73.9% F1-score for arousal. These results clearly demonstrate the effectiveness of integrating channel selection and hybrid deep learning for EEG-based emotion recognition.

IV. Conclusion

Emotion recognition plays an important role in BCI systems, with applications across domains such as education,

TABLE II
COMPARISON OF THE PROPOSED MODEL WITH EXISTING WORKS

		Valence		Arousal	
Ref	Method	Acc	F1	Acc	F1
[10] (2025)	+ 8 channels + Raw signals + EEGNet model	61.4	60.2	-	-
[11] (2023)	+ 3D spatial-spectral + GCN model	57.7	-	58.3	-
Our	+ RF-RFE for channel selection + 16 extracted features + 1D-CNN-transformer	70.0	76.7	67.5	73.9

healthcare, and entertainment. In this study, we proposed a hybrid deep learning model that combines 1D-CNN with Transformer encoder for EEG-based emotion classification. EEG signals were decomposed into four frequency bands using the DWT, and 16 features were extracted from each band. To reduce dimensionality and retain the most informative information, channel selection was performed using the RF-RFE method, resulting in two subsets of 10 channels for valence and arousal classification. The proposed 1D-CNN–Transformer model was compared with baseline models, including 1D-CNN and MLP, using subject-wise 5-fold CV. The experimental results show that the hybrid model outperforms the baselines. These results demonstrate the effectiveness of integrating 1D-CNN with Transformer encoders to capture both local and long-range dependencies, highlighting the potential of the proposed model for robust emotion recognition applications.

REFERENCES

[1] Imtiaz, Md Niaz, and Naimul Khan. "Enhanced cross-dataset electroencephalogram-based emotion recognition using unsupervised domain adaptation." Computers in Biology and Medicine 184 (2025): 109394.

[2] Nandini, Durgesh, Jyoti Yadav, Asha Rani, and Vijander Singh. "Enhancing emotion detection with non-invasive multi-channel EEG and hybrid deep learning architecture." Iranian Journal of Science and Technology, Transactions of Electrical Engineering 48, no. 3 (2024): 1229-1248.

[3] Yang, Yimin, et al. "EEG-based emotion recognition using hierarchical network with subnetwork nodes." IEEE Transactions on Cognitive and Developmental Systems 10.2 (2017): 408-419.

[4] García-Martínez, Beatriz, et al. "A review on nonlinear methods using electroencephalographic recordings for emotion recognition." IEEE Transactions on Affective Computing 12.3 (2019): 801-820.

[5] Li, Qiaomei, et al. "DEMA: Deep EEG-first multi-physiological affect model for emotion recognition." Biomedical Signal Processing and Control 99 (2025): 106812.

[6] Sharif, SK Mastan, et al. "Improved LSTM-Squeeze net architecture for brain activity detection using EEG with improved feature set." Biomedical Signal Processing and Control 101 (2025): 107222.

[7] Nguyen, Duy, Minh Tuan Nguyen, and Kou Yamada. "Electroencephalogram Based Emotion Recognition Using Hybrid Intelligent Method and Discrete Wavelet Transform." Applied Sciences 15.5 (2025): 2328.

[8] Devarajan, Kasthuri, Suresh Ponnan, and Sundresan Perumal. "Hybrid CNN-transformer architecture for enhanced EEG-based emotion recognition: capturing local and global dependencies with self-attention mechanisms." Discover Computing 28.1 (2025): 87.

[9] Koelstra, Sander, et al. "Deap: A database for emotion analysis; using physiological signals." IEEE transactions on affective computing 3.1 (2011): 18-31.

[10] Thiruselvam, S. V., and M. Ramasubba Reddy. "Frontal EEG correlation based human emotion identification and classification." Physical and Engineering Sciences in Medicine 48.1 (2025): 121-132.

[11] Cui, Gaochao, Xueyuan Li, and Hideaki Touyama. "Emotion recognition based on group phase locking value using convolutional neural network." Scientific Reports 13.1 (2023): 3769.