



## OPEN Emotion recognition with multiple physiological parameters based on ensemble learning

Yilong Liao<sup>1,2,3</sup>, Yuan Gao<sup>1,2,3</sup>, Fang Wang<sup>1,2,3</sup>, Li Zhang<sup>1,2</sup>✉, Zhenrong Xu<sup>1,2</sup> & Yifan Wu<sup>1,2</sup>

Emotion recognition is a key research area in artificial intelligence, playing a critical role in enhancing human-computer interaction and optimizing user experience design. This study explores the application and effectiveness of ensemble learning methods for emotion recognition based on multiple physiological parameters. A dataset was systematically constructed by preprocessing data from electroencephalogram (EEG), galvanic skin response (GSR), skin temperature (ST), and heart rate (HR) collected from 38 subjects while watching short videos. We proposed a hybrid model framework combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, trained and optimized using a random seed initialization strategy and a cosine annealing warm restart strategy. To further enhance performance, various strategies were designed and evaluated. The results showed that applying advanced preprocessing techniques significantly improved data quality, while the hybrid model effectively leveraged the advantages of both CNN and LSTM. Incorporating the cosine annealing warm restart strategy further boosted model performance. Using a soft voting ensemble method, the proposed approach achieved a 96.21% accuracy rate in classifying seven emotions—calm, happy, disgust, surprise, anger, sad, and fear, indicating its ability to accurately capture emotional responses to short videos. This study presents an innovative approach to emotion recognition using multiple physiological parameters, demonstrating the potential of ensemble learning for complex tasks. It offers valuable insights for the development of effective applications.

**Keywords** Ensemble learning, Emotion recognition, Cosine annealing warm restart, Multiple physiological parameters, Data processing

Emotion recognition is a core research topic in artificial intelligence focused on analyzing and understanding human emotional states. This technology plays a pivotal role in enhancing human-computer interaction, advancing mental health management, and optimizing user experience design. In recent years, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been widely applied in emotion recognition and related fields, demonstrating remarkable performance<sup>1,2</sup>. For instance, Febrian<sup>2</sup> proposed a Bidirectional LSTM (Bi-LSTM)-CNN model that achieved outstanding results in facial expression recognition. However, occlusions pose a significant challenge to facial expression analysis<sup>3</sup>. Even with multi-task auxiliary correction (MTAC) methods<sup>4</sup>, maintaining stability in complex scenarios remains a formidable task. Beyond facial expressions, CNNs and LSTMs have also proven highly effective in physiological signal-based emotion recognition. These models have been successfully applied to classify emotions using signals such as electrocardiogram (ECG)<sup>5</sup>, electrodermal activity (EDA)<sup>6</sup>, and electromyography (EMG)<sup>7</sup>, further validating the potential of physiological signals in this domain. Du<sup>8</sup> introduced the ATDD-LSTM model, which effectively captures nonlinear relationships between EEG electrodes, significantly improving EEG-based emotion recognition. Additionally, Chakravarthi<sup>9</sup> developed a CNN-LSTM hybrid framework based on ResNet-152, achieving highly accurate emotion classification. In another study<sup>10</sup>, a CNN-LSTM hybrid architecture integrated with a stochastic fractal search optimization algorithm, further improved recognition performance.

Despite significant progress in deep learning-based unimodal emotion recognition, inherent limitations persist across different approaches. For instance, text-based emotion recognition heavily relies on semantic understanding, making it difficult to detect implicit emotions<sup>11,12</sup>. Similarly, single physiological signal-based emotion recognition faces challenges related to individual differences, signal variability, and environmental noise, which limit its generalization capabilities. To address these issues, integrating multiple physiological

<sup>1</sup>School of Biomedical Engineering, South-Central Minzu University, Wuhan 430074, China. <sup>2</sup>South-Central Minzu University, No.182, Minzu Avenue, Hongshan District, Wuhan City, Hubei Province, China. <sup>3</sup>Yilong Liao, Yuan Gao and Fang Wang are the co-first authors of the paper. ✉email: zhangli1996@163.com

signals has emerged as a promising strategy to enhance both generalization and stability. For example, self-supervised multimodal representation learning<sup>13</sup>, EEG and ECG fusion<sup>14</sup>, and combining EEG with electrooculography (EOG)<sup>15</sup> have significantly improved emotion recognition accuracy. Additionally, studies by Li<sup>16</sup> and Kang<sup>17</sup> emphasized the effectiveness of extracting features from multiple physiological signals for emotion classification, validating the importance of multimodal data utilization. These studies enhanced emotion recognition performance and broadened its applications, such as in the emotional assessment of gamers<sup>18</sup> and real-time conversation emotion recognition<sup>19</sup>. Notably, Wang<sup>20</sup> and Wang<sup>21</sup> developed advanced emotion recognition strategies tailored for noisy environments, combining EEG with other modalities, such as speech and facial expressions.

The high dimensionality and non-linear characteristics of multi-parameter data challenge single algorithms in effectively processing this complexity. Ensemble learning, which combines multiple models for a single prediction task, has proven effective in enhancing model generalization. By integrating different machine learning models, Zhang<sup>22</sup> significantly improved the performance of emotion recognition algorithms using physiological signals. The two-tier ensemble strategy of deep convolutional neural network models shows the powerful potential of combining deep learning with ensemble learning<sup>23</sup>. In speech emotion recognition, Aishwarya<sup>24</sup> significantly improved accuracy by using various feature extraction techniques and ensemble methods, such as CatBoost and Voting classifiers. Furthermore, ensemble learning's applications extend beyond labs, showing potential in real-world scenarios and maintaining high accuracy in noisy environments, as validated by numerous studies<sup>25,26</sup>. Zaman<sup>27</sup> and De<sup>28</sup> explored ensemble learning in specific applications, such as driver emotion recognition and personalized treatment for Alzheimer's patients, highlighting its diverse applications and benefits for specific user groups. Finally, Subasi's<sup>29</sup> EEG processing method using ensemble classifiers achieved extremely high accuracy, demonstrating the value of ensemble learning in specialized fields. These findings provide robust evidence for advancements in emotion recognition technology and offer valuable insights for future research directions.

In summary, deep learning algorithms such as CNNs and LSTMs exhibit great potential in emotion recognition. Meanwhile, ensemble learning improves accuracy and robustness by integrating multiple physiological parameters, offering strong support for the development of more stable and efficient recognition systems. Therefore, this study focuses on leveraging ensemble learning methods to recognize emotions from various physiological signals. Specifically, we collected EEG, GSR, ST, and HR signals from subjects watching short videos. These multimodal data were used to construct a multi-parameter emotion recognition dataset. To expand the dataset size, we applied a sliding window technique for data segmentation. A base model was then developed combining CNN and LSTM architectures, and trained using random initialization (RI) and cosine annealing warm restarts (CAWR), with key model weights saved during training. To further optimize performance, a soft voting method was implemented for decision-level fusion. Experimental results showed that the proposed method achieved 96.21% accuracy in multi-task emotion recognition, significantly outperforming single-model performance. These findings validate the effectiveness and robustness of ensemble learning for emotion recognition based on multiple physiological parameters.

The structure of this article is as follows: Sect. [Dataset and preprocessing](#) outlines the collection and preprocessing of the multi-physiological parameter dataset for emotion recognition. Section [Data collection](#) introduces the proposed emotion recognition framework and model optimization strategies, including the base model architecture integrating CNN and LSTM units, as well as the optimization techniques (RI and CAWR). It also details the soft voting ensemble method, which combines models saved during training to achieve precise identification of seven emotions. Section [Results and discussion](#) delves into the analysis and discussion of results obtained from various optimization strategies and ensemble learning experiments. Section [Conclusion](#) summarizes the key contributions of this work and outlines its broader implications.

## Dataset and preprocessing

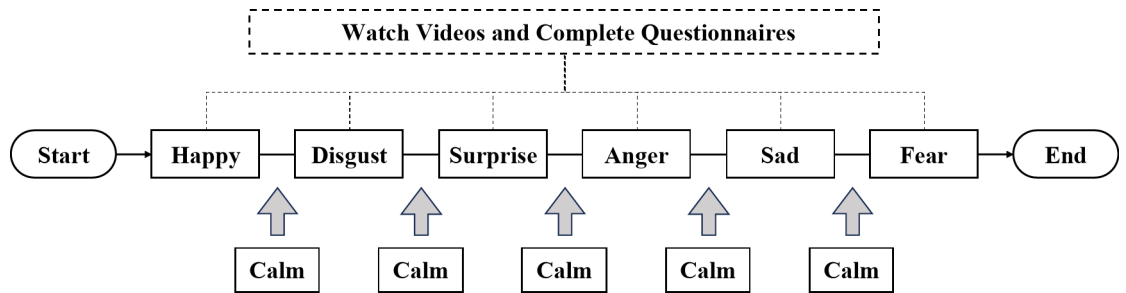
### Data collection

In this study, we developed the Self-built dataset (SELF dataset), a multi-parameter emotion recognition dataset, by carefully selecting videos specifically designed to elicit basic emotions. A total of 45 participants aged 18–28, all with no history of psychological or mental disorders, were recruited. During the viewing of short videos, physiological data were collected from the participants, including single-channel EEG, GSR, ST, and HR. After a thorough process of data cleaning and validating, the final dataset comprised complete and reliable data from 38 participants.

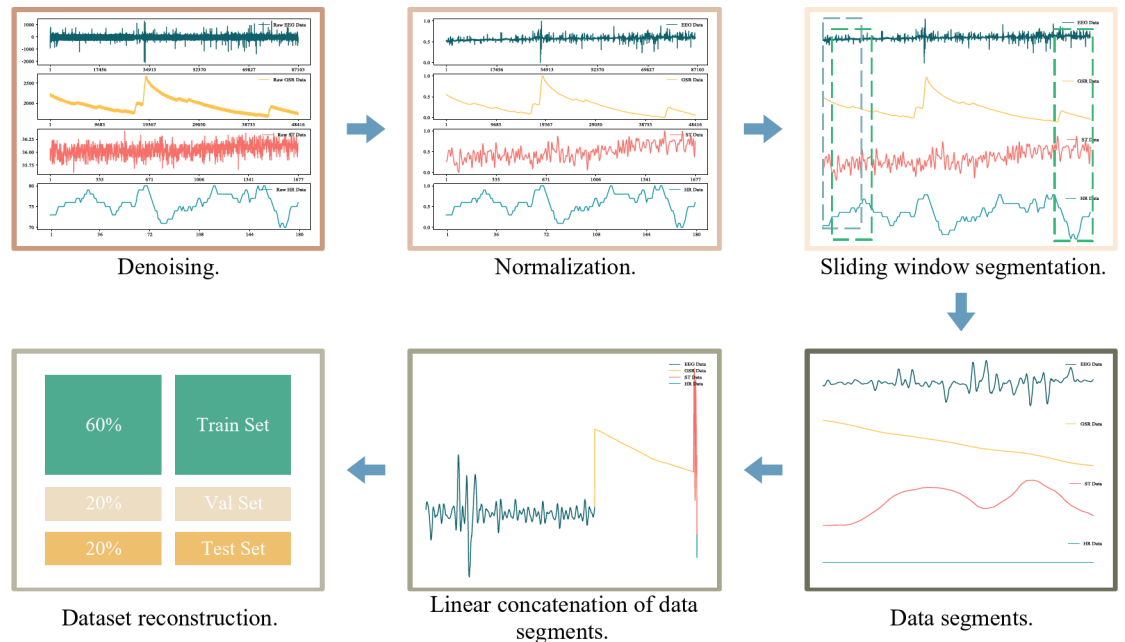
This study was approved by the Scientific Research Ethics and Science and Technology Safety Committee of South-Central Minzu University, Grant Number: 2022-scuec- 106. Informed consent was signed by all study participants. Eligibility and exclusion criteria for subject recruitment: a convenience sample from healthy adults (18 years or older) who responded to the recruitment flyer. Subjects with a history of major disease were excluded.

Figure 1 illustrates the emotion induction process. The emotion induction materials consisted of six sets of short videos and one long video designed to elicit seven emotions: Calm, Happy, Disgust, Surprise, Anger, Sad, and Fear. Participants watched each set of 10 videos sequentially. Video lengths ranged from 76 to 584 s. After each video, participants completed an emotion questionnaire to report their feelings. After each set of videos, participants watched a neutral video to return to their emotional baseline, preparing for the next set of emotion induction.

Data collection equipment included single-channel EEG, GSR, ST, and HR sensors. Sampling frequencies were: EEG at 512 Hz, GSR at 300 Hz, ST at 10 Hz, and HR at 1 Hz. Sensors connected to a computer via serial ports ensured real-time, continuous data reception during video playback. Each data point was timestamped to 1-millisecond precision and saved as a txt file. After data collection, experimenters manually verified timestamps to ensure accurate alignment between physiological data and viewed videos. Emotion questionnaire results labeled physiological data with discrete emotion tags. Finally, data were cleaned, timestamps aligned, and both



**Fig. 1.** Emotion induction process.



**Fig. 2.** SELF Dataset Preprocessing Workflow.

data and labels saved as.csv files for analysis and research. For a detailed description of the dataset construction process and its usability validation, please refer to<sup>30</sup>. Detailed dataset information is available at: <https://github.com/LiaoEpoch/Dataset-for-Emotion-Recognition-from-Physiological-Signals-Induced-by-Short-Videos.git>.

To ensure a fair evaluation of data usability and model performance, this study further validates its findings using the SWELL<sup>31</sup> and WESAD<sup>32</sup> datasets. The SWELL dataset includes physiological signals recording from 25 participants engaged in a three-hour knowledge work task. To control experimental conditions, stressors were introduced by exposing participants to neutral, interruption, and time-pressure scenarios. This study utilizes the raw ECG signals sampled at 2048 Hz and assesses emotional states using the Self-Assessment Manikin (SAM) scale. Based on valence and arousal scores, with a threshold of 4, emotions are classified into High Valence (HV,  $\geq 4$ ) and Low Valence (LV,  $< 4$ ), as well as High Arousal (HA,  $\geq 4$ ) and Low Arousal (LA,  $< 4$ ), resulting in four emotional categories: HVHA, HVLA, LVHA, and LVLA. The WESAD dataset primarily explores wearable-based stress and emotion detection, containing physiological recordings from 15 participants under four emotional states: Neutral, Stress, Amusement, and Meditation. For experimental analysis, this study selects raw ECG, EDA, and temperature (Temp) signals, all sampled at 700 Hz, for experimental analysis. These datasets provide a robust foundation for assessing the generalizability and effectiveness of the proposed emotion recognition framework.

### Data preprocessing

This study used a systematic data preprocessing workflow to ensure data quality and analytical accuracy. Figure 2 illustrates the data preprocessing steps of the SELF dataset.

**Data denoising and normalization** In the SELF dataset, wavelet denoising is applied to EEG, GSR, and ST signals to eliminate high-frequency noise. Specifically, the raw signals are first decomposed into different scales of components using the wavelet transform. This decomposition separates the signal into approximation coefficients (cA) and detail coefficients (cD). A soft thresholding technique is then applied to the cD to suppress noise.

This step involves setting coefficients below a threshold to zero (removing noise) and shrinking larger coefficients (preserving signal characteristics but reducing noise). After thresholding, the signal is reconstructed using the inverse wavelet transform, combining the denoised cD and the original cA components. For EEG signals, a six-level decomposition is performed using the db4 wavelet, with cD1 and cD2 removed and soft thresholding applied to cD3–cD6. GSR signals undergo a nine-level decomposition with the db5 wavelet, where cD1 and cD2 are discarded, and soft thresholding is applied to cD3–cD8. ST signals are processed using a four-level decomposition with the db4 wavelet, removing cD1–cD3 and applying soft thresholding to cD4. The threshold value is empirically set to 0.6745 for all signal types. To ensure uniformity across signals of varying scales, min-max normalization is applied as a preprocessing step. This normalization step scales the data to a consistent range, facilitating subsequent analysis and comparison.

**Sliding window segmentation** To address the limited sample size of the SELF dataset, data augmentation is performed using a partially overlapping sliding window approach. Specifically, the preprocessed EEG, GSR, ST, and HR signals are segmented using sliding windows. This method enhances the dataset's size and diversity by generating multiple overlapping subsequences from the original signals, thereby improving the robustness of subsequent analysis. Each segment is 5 s long, with a window length of  $5 \times F_{Signal}$  and a sliding step of  $1 \times F_{Signal}$ , where  $F_{Signal}$  denotes the sampling frequency of each signal. Emotion labels are preserved across subsequences to maintain label accuracy.

**Linear concatenation of data segments** The EEG, GSR, ST, and HR data from the same participant under the same video condition are concatenated in index order to construct multimodal feature inputs.

In contrast to the preprocessing approach applied to the SELF dataset, the SWELL and WESAD datasets do not undergo denoising or normalization. Instead, data segmentation is performed using a non-overlapping sliding window strategy. For the SWELL dataset, raw ECG signals are divided into 2-second windows, each containing  $2 \times 2048$  data points, with a step size equal to the window length. Similarly, the WESAD dataset adopts a 5-second window, where each segment consists of  $5 \times 700$  data points, also with a step size matching the window length.

To evaluate the model's generalization capability to unseen participants, all data from the last participant in each dataset are reserved as an external validation set and processed according to their respective preprocessing pipelines.

After processing, the SELF dataset comprises 45,586 data segments, each containing 4,115 data points and labeled with seven emotional states: Calm, Happy, Disgust, Surprise, Anger, Sad, and Fear. The SWELL dataset generates 91,112 single-channel ECG segments, each with 4,096 data points, categorized into four emotion classes: HVHA, HVLA, LVHA, and LVLA. The WESAD dataset produces 8,973 three-channel physiological signal segments, each consisting of 3,500 data points, annotated with four states: Baseline, Amusement, Meditation, and Stress. Excluding the external validation set, the remaining data are randomly split into training, validation, and test sets in a 6:2:2 ratio for model training, hyperparameter tuning, and final evaluation.

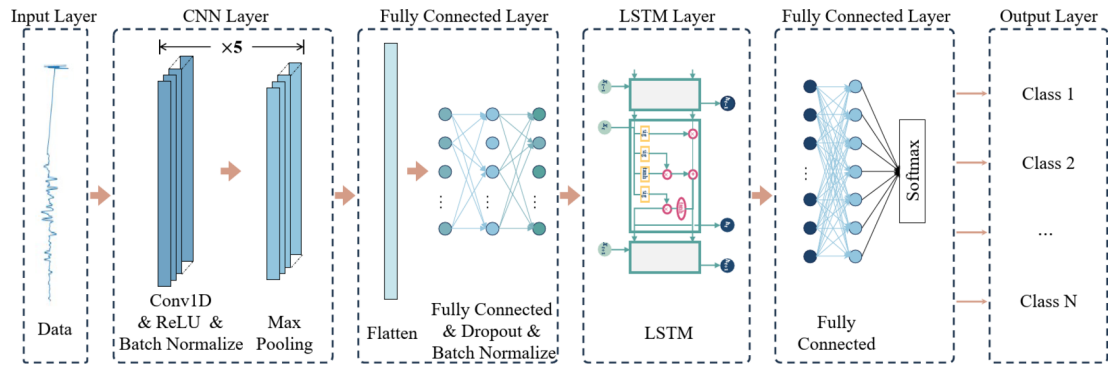
## Emotion recognition

This study aims to construct a hybrid model combining CNN and LSTM to evaluate its performance in emotion recognition using multiple physiological parameters. The model is initialized randomly, and systematic training and testing are conducted using the preprocessed dataset to evaluate its performance under five different initialization configurations. A CAWR strategy is introduced to optimize model convergence and prevent it from trapping in local optima. This strategy helps the model in searching for the optimal solution in the global search space, enhancing overall performance. The model state is saved after each training epoch for further analysis and comparison. Additionally, predictions from different models are fused at the decision layer using the soft voting method. By comparing model performance before and after fusion, the impact of ensemble learning strategies on emotion recognition accuracy can be assessed.

## Base model

Traditional CNNs struggle to capture complex signal patterns. For long time-series data, they typically require deeper architectures to expand the receptive field and extract global information<sup>33</sup>, leading to a significant increase in computational cost. In contrast, LSTMs excel in temporal sequence modeling<sup>34</sup> but face high computational and memory demands when processing ultra-long sequences, along with potential gradient vanishing issues<sup>35</sup>. To overcome these challenges, we propose a hybrid CNN-LSTM framework for multimodal emotion recognition, as illustrated in Fig. 3. The framework first employs a CNN module to extract signal features and reduce dimensionality, obtaining a compact representation of long time-series data. Then, an LSTM module models and integrates contextual dependencies within these low-dimensional features. Finally, a classification layer generates the emotion recognition results, effectively tackling the complexities of multimodal long-sequence emotion classification.

Due to variations in data formats across different datasets, the model inputs also differ accordingly. Specifically, the SELF dataset input consists of a single-channel signal with a length of 4,115, the SWELL dataset input is a single-channel sequence of length 4,096, and the WESAD dataset input comprises a three-channel signal with a length of 3,500. The CNN layer comprises five one-dimensional convolutional layers, each followed by L2 regularization, ReLU activation, batch normalization, and max-pooling layers. The L2 regularization factor is set to 0.01. The outputs of these convolutional layers are flattened into one-dimensional vectors and fed into three fully connected layers with Dropout. The dropout rate is fixed at 0.4. These vectors then pass through the LSTM layer to capture long-term dependencies. The final output layer contains either 7 or 4 nodes, depending on the classification task, and employs the Softmax activation function to produce class probabilities. During training,



**Fig. 3.** Base model structure.

the model uses the Adam optimizer with the cross-entropy loss function, evaluating performance based on accuracy.

### Random initialization

Model parameters are randomly initialized using five different random seeds (3, 6, 9, 12, 15). All network layers, including convolutional and fully connected layers, use the default initialization method provided by the Keras framework. Five models with different initializations are trained, and each model is saved upon completion of training. The validation set is used to compare the impact of different initialization methods on model performance, aiming to identify the optimal configuration for improving emotion recognition accuracy.

### Cosine annealing with warm restarts

In this study, CAWR is employed in addition to RI to optimize the learning rate scheduling process. Within each training cycle, the learning rate decreases gradually following a cosine curve. At the end of each cycle, it resets to a higher initial value before starting the next annealing cycle. This method aims to mitigate the problem of local optima associated with traditional monotonically decreasing learning rate strategies, while ensuring the model retains adequate exploration capability during training<sup>36</sup>.

Let  $Epochs$  denote the current training epoch, starting from 1.  $T$  represents the cycle length in epochs.  $\eta_{max}$  is the initial learning rate at the beginning of each cycle, set as the maximum learning rate threshold.  $\eta_{min}$  is the minimum learning rate threshold. Thus, the relative position  $x$  of the current training epoch within the cycle can be calculated using Eq. 1.

$$x = \left( \frac{Epochs}{T} \right) \% 1 \quad (1)$$

The symbol % denotes the modulo operation, ensuring that the relative position  $x$  remains within the range  $[0, 1)$ . Equation 2 calculates the cosine annealing factor  $\alpha$  based on the relative position  $x$  of the current training epoch within the cycle.

$$\alpha = \frac{1}{2} (1 + \cos(\pi \times x)) \quad (2)$$

The learning rate  $\eta_{new}$  is updated using the cosine annealing factor  $\alpha$  as described in Eq. 3.

$$\eta_{new} = \max[\eta_{min}, \eta_{min} + (\eta_{max} - \eta_{min}) \times \alpha] \quad (3)$$

Equation 4 summarizes the formula for updating the learning rate.

$$\eta_{new} = \max \left\{ \eta_{min}, \eta_{min} + \frac{1}{2} (\eta_{max} - \eta_{min}) \left[ 1 + \cos \left( \pi \times \left( \frac{Epochs}{T} \right) \% 1 \right) \right] \right\} \quad (4)$$

Formula 4 illustrates the dynamic adjustment of the learning rate based on the current training epoch (Epochs), ensuring it oscillates between the maximum learning rate  $\eta_{max}$  and the minimum learning rate  $\eta_{min}$  in a cosine annealing manner throughout training. The modulo operation ensures the learning rate completes a cosine decay cycle within each cycle  $T$ , restarting at the end to begin a new decay process from a higher initial learning rate. In this study, the number of epochs is set to 300, with a cosine annealing schedule of period  $T = 60$ . The maximum learning rate is  $\eta_{max} = 0.001$ , and the minimum learning rate is  $\eta_{min} = 0$ . Figure 4 illustrates the curve of the learning rate changing throughout the training epochs.

Figure 4 illustrates how the learning rate decreases gradually from  $\eta_{max} = 0.001$  to  $\eta_{min} = 0$  within each cycle, following a cosine curve. At the start of each new cycle, the learning rate resets to  $\eta_{max}$ , facilitating periodic adjustment and restarting of the learning rate.

This study saved critical model weights during training to thoroughly evaluate the impact of different initialization conditions on model performance. During 300 epochs, each initialized random model saved its state at the end of every cycle. Therefore, with five initializations and five cycles, 25 states were saved for

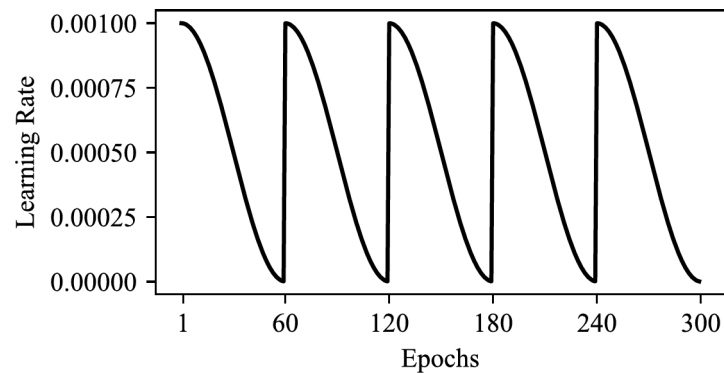


Fig. 4. Learning rate change curve.

| Dataset | Random seed | Train accuracy (%) | Val accuracy (%) | Test accuracy (%) |
|---------|-------------|--------------------|------------------|-------------------|
| SELF    | 3           | 78.30              | 65.90            | 65.30             |
|         | 6           | 72.55              | 57.58            | 60.16             |
|         | 9           | 84.21              | 67.51            | 69.51             |
|         | 12          | 79.81              | 63.58            | 66.01             |
|         | 15          | 55.86              | 43.31            | 48.24             |
| SWELL   | 3           | 91.68              | 91.81            | 92.75             |
|         | 6           | 92.47              | 92.43            | 94.06             |
|         | 9           | 94.49              | 94.15            | 95.53             |
|         | 12          | 85.61              | 86.50            | 88.37             |
|         | 15          | 57.75              | 42.62            | 57.61             |
| WESAD   | 3           | 97.19              | 94.15            | 95.10             |
|         | 6           | 96.84              | 90.81            | 93.85             |
|         | 9           | 97.17              | 96.06            | 96.06             |
|         | 12          | 96.28              | 93.37            | 93.43             |
|         | 15          | 96.02              | 91.76            | 94.92             |

Table 1. Training and testing results of the base model with RI strategy.

the models. Comparing model states under different initialization conditions and at various training points provides a deeper understanding of how initialization conditions and periodic training strategies influence model learning capacity.

Model ensemble

To enhance predictive performance, reduce overfitting, and improve model robustness, we applied a soft voting method to ensemble models from the 25 saved states. Specifically, the model ensemble was approached from three perspectives: first, integrating models with identical random seeds but varying training epochs to assess how initialization conditions affect predictive performance. Second, integrating models with different random seed initializations but the same number of training epochs to evaluate the effect of training progress variations on model performance. Finally, integrating all models saved during training to achieve comprehensive performance optimization.

Results and discussion

Discussion of random initialization results

This study aimed to investigate the impact of random seed initialization on model performance. The same base model was initialized with five different random seeds (3, 6, 9, 12, 15), and trained for 300 epochs. The batch sizes for the SELF, SWELL, and WESAD datasets are set to 512, 128, and 128, respectively. Models were trained on the training dataset and evaluated continuously on the validation dataset. After training, each model was saved, and their performance on the test dataset was evaluated. Table 1 presents the training and test accuracies of five models initialized with different random seeds across the three datasets.

The results indicate that random seed initialization significantly impacts model training and generalization. For the SELF dataset, seed 9 achieves the highest accuracy across the training, validation, and test sets, whereas seed 15 may cause the model to fall into a local optimum, resulting in performance degradation. For the SWELL dataset, where overall classification accuracy is relatively high, seed 9 delivers the best test performance, whereas seed 15 impedes model convergence. For the WESAD dataset, although model performance remains

stable across different initializations, seed 9 consistently outperforms others. Overall, the choice of random seed is critical for model training and generalization. Seed 9 helps guide the model toward a more stable and generalizable solution, providing a foundation for further optimization strategies and the development of more robust emotion recognition methods.

### Discussion of CAWR results

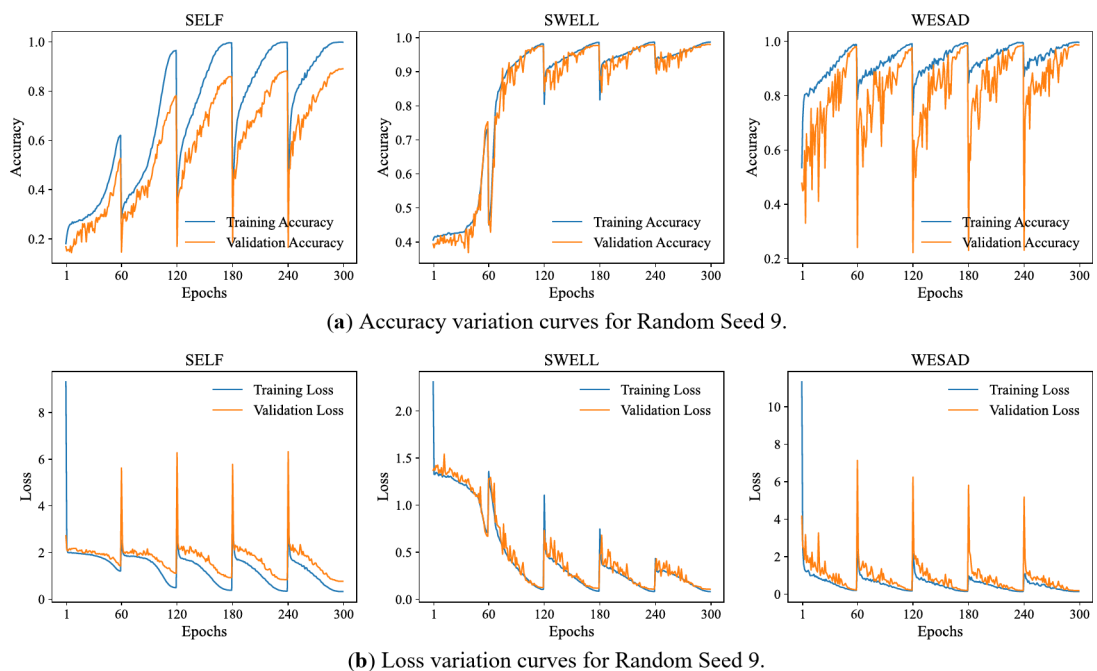
This study introduced the CAWR strategy alongside random seed initialization to adjust the learning rate, aiming to optimize the training effectiveness of the base model. Five different random seeds (3, 6, 9, 12, 15) were used. The dynamic learning rate cycle was  $T = 60$  epochs, with training conducted for 300 epochs. The batch sizes for the SELF, SWELL, and WESAD datasets are set to 512, 128, and 128, respectively. The maximum and minimum learning rates were set at  $\eta_{max} = 0.001$  and  $\eta_{min} = 0$ , respectively. Throughout the training, the model's performance was assessed using the validation set, and the model's state was saved at the end of each learning rate cycle. The accuracy and loss curves during training with random seed 9 for the three datasets are shown in Fig. 5.

Figure 5 clearly illustrates the periodic variation of model accuracy and loss values during the CAWR learning rate process. Throughout the annealing and restart cycle, both training and validation accuracies exhibit steady improvement, while loss values gradually decrease. At the end of each annealing cycle, resetting the learning rate to its initial maximum value enables the model to escape local optima and explore a wider solution space. This strategy significantly enhances model performance, achieving peak validation accuracies of 89.04%, 98.00%, and 98.69% on the SELF, SWELL, and WESAD datasets, respectively. Additionally, the CAWR strategy reduces the number of training epochs required to reach high accuracy. For instance, on the SELF dataset, the model initialized with random seed 9 achieves comparable results to the baseline model trained for 300 epochs in just 100 epochs, when using the CAWR strategy. This improvement not only enhances efficiency but also substantially reduces training time and computational cost.

As the number of training epochs increases, the model shows varying degrees of overfitting. To comprehensively assess the model's performance, the saved models at the end of training are tested using the testing dataset. Table 2 presents the testing results and performance comparisons before and after introducing the dynamic learning rate.

Table 2 shows that applying the CAWR learning rate adjustment strategy significantly improves the test accuracy of all models, with a notable average increase. For the SELF dataset, the model initialized with random seed 15 achieves an accuracy increase from 48.25 to 86.99%, demonstrating that CAWR effectively mitigates underfitting. Similarly, in the SWELL dataset, the accuracy of the seed 15 model rises from 57.61 to 97.35%, further validating CAWR's ability to enhance model robustness. Even in the WESAD dataset, where baseline model accuracy is already relatively high, CAWR still delivers consistent performance gains. Additionally, CAWR significantly reduces performance variability caused by random seed initialization. Under the RI training strategy, the accuracy gap between the best and worst models in the SELF dataset is 21.26%, whereas CAWR narrows this discrepancy to just 2.54%, greatly enhancing training stability and reproducibility.

LSTM effectively retains long-term dependencies through its gating mechanism, alleviating the vanishing gradient problem inherent in traditional RNNs, but it remains susceptible to gradient explosion. To address



**Fig. 5.** Experimental results of the base model with CAWR strategy.

| Dataset | Random seed | Accuracy (%) |           |             |
|---------|-------------|--------------|-----------|-------------|
|         |             | RI           | RI + CAWR | Improvement |
| SELF    | 3           | 65.30        | 88.75     | 23.45       |
|         | 6           | 60.16        | 88.10     | 27.94       |
|         | 9           | 69.51        | 89.53     | 20.02       |
|         | 12          | 66.01        | 87.81     | 21.80       |
|         | 15          | 48.25        | 86.99     | 38.74       |
| SWELL   | 3           | 92.75        | 97.43     | 4.68        |
|         | 6           | 94.06        | 96.03     | 1.97        |
|         | 9           | 95.53        | 97.82     | 2.29        |
|         | 12          | 88.37        | 96.50     | 8.13        |
|         | 15          | 57.61        | 97.35     | 39.74       |
| WESAD   | 3           | 95.10        | 98.21     | 3.11        |
|         | 6           | 93.85        | 98.03     | 4.18        |
|         | 9           | 96.06        | 98.21     | 2.15        |
|         | 12          | 93.43        | 98.15     | 4.72        |
|         | 15          | 94.92        | 97.79     | 2.87        |

**Table 2.** The comparison of testing results with and without CAWR.

| Ensemble perspective               |                   | Soft voting accuracy (%) |              |              |
|------------------------------------|-------------------|--------------------------|--------------|--------------|
|                                    |                   | SELF                     | SWELL        | WESAD        |
| Same Random Seed, Different Epochs | Random seed of 3  | 91.12                    | 97.79        | 98.45        |
|                                    | Random seed of 6  | 90.14                    | 96.46        | 98.51        |
|                                    | Random seed of 9  | 92.34                    | 98.10        | 98.33        |
|                                    | Random seed of 12 | 90.23                    | 96.59        | 98.15        |
|                                    | Random seed of 15 | 87.71                    | 97.88        | 98.27        |
| Same Epoch, Different Random Seeds | Epoch of 60       | 58.66                    | 94.50        | 98.39        |
|                                    | Epoch of 120      | 84.22                    | 98.01        | 98.45        |
|                                    | Epoch of 180      | 93.68                    | 98.17        | 98.27        |
|                                    | Epoch of 240      | 95.27                    | 98.17        | 98.51        |
|                                    | Epoch of 300      | <b>96.21</b>             | <b>98.26</b> | <b>98.81</b> |
| All Models                         |                   | 95.36                    | 98.22        | 98.45        |

**Table 3.** Ensemble learning results.

this, CAWR adopts a monotonically decreasing learning rate strategy within each cycle, gradually reducing the learning rate to zero. This approach suppresses instability caused by large gradients and facilitates model convergence. Furthermore, the compatibility between LSTM’s activation functions (Sigmoid for gating and Tanh for state updates) and the cosine annealing strategy further enhances optimization. By periodically adjusting the learning rate, CAWR optimizes weight updates at different training phases, accelerates convergence toward the global optimum, and enhances multi-class classification accuracy. Notably, under constrained training iterations, this strategy maximizes the model’s potential, significantly enhancing overall performance. These results validate the effectiveness of CAWR in multi-physiological signal-based emotion recognition tasks.

Ensemble learning results discussion

The soft voting method was used to combine models from 25 different states, including models with the same random seed but different training epochs, different random seeds but the same training epochs, and all models saved during training. Table 3 presents the test accuracy of different ensemble strategies.

The results indicate that all ensemble approaches significantly improve the model’s generalization capability. Whether by varying training epochs with a fixed random seed, using different random seeds with fixed training epochs, or employing a full-model ensemble, ensemble learning effectively addresses both underfitting and overfitting, thereby enhancing prediction stability. The SELF dataset exhibits the most pronounced benefits from these strategies. Ensembling across different training epochs enriches feature representation, while ensembling models trained with different random seeds reduces the impact of initialization variability. The full-model ensemble achieves the highest generalization performance. These findings highlight that a well-designed ensemble strategy not only improves classification accuracy but also enhances model robustness, making it a powerful tool for physiological signal analysis.

After 300 training epochs, the soft voting ensemble strategy, which integrates five models with different random initializations, achieves the highest classification accuracy across all three datasets. Table 4 further

evaluates the classification consistency for each emotion category, including precision, recall, and F1-score. These metrics offer a comprehensive evaluation of the model's performance and its balance across emotion classes.

Overall, the model demonstrates high classification performance across all datasets and emotion categories, highlighting its effectiveness in emotion recognition. However, the difficulty of distinguishing between classes varies. In the SELF dataset, all emotion categories achieve F1 scores above 95%. Recognition of Calm, Happy, Anger, Sad, and Fear is well-balanced, whereas Disgust and Surprise exhibit lower recall rates, likely due to high intra-class variability or feature overlap causing misclassification. In the SWELL dataset, all categories achieve F1 scores exceeding 98%, indicating that the model can accurately distinguish between high/low arousal and valence states. However, the recall rate for LVHA is slightly lower, possibly because the boundary between high-arousal/low-valence samples is less distinct. In contrast, HVLA and LVLA show greater stability, likely due to the even distribution of low-arousal state data. For the WESAD dataset, all categories achieve F1 scores above 96%. Recognition of Baseline and Stress is the most accurate, likely because their physiological signal patterns are distinct. However, Amusement shows a lower recall rate, with some samples misclassified as Baseline or Meditation, reflecting lower separability in physiological signals.

To evaluate the model's generalization capability to unseen subjects or videos, we applied a segment-wise prediction and voting strategy for external validation. Predictions were generated at the segment level on external datasets, with final classifications determined through soft voting. The results showed classification accuracies of 42.86%, 50%, and 50% on the SELF, SWELL, and WESAD datasets, respectively. These findings indicate that the model's high accuracy on internal data did not generalize effectively to external data. The primary limitation likely arises from individual differences in physiological signals, which significantly impact cross-subject emotion recognition. Future research should prioritize strategies to reduce inter-subject variability, such as domain adaptation or personalized modeling, to enhance the model's generalizability and enable broader real-world applications.

To validate the effectiveness of our proposed method, we reproduced and compared several representative models from existing studies. Specifically, we strictly followed the original architectures and hyperparameter settings described in the respective papers to replicate DCNN<sup>37</sup>, CNN<sup>38</sup>, and Res2 Net<sup>39</sup>. Additionally, we implemented two widely recognized general-purpose models, DeepConvNet<sup>40</sup> and EEGNet<sup>41</sup>, ensuring that their parameter configurations remained consistent with their original implementations. Given the limitations of RNN, GRU, and LSTM in modeling ultra-long time series, we developed hybrid frameworks, CNN + RNN and CNN + GRU, to assess different temporal modeling strategies. Table 5 summarizes the F1 scores of each model across the SELF, SWELL, and WESAD datasets, providing a comprehensive comparison of their classification performance.

As demonstrated in Table 5, our approach achieves the best performance across all datasets, outperforming existing models in terms of F1 scores. The SELF dataset has a relatively small sample size and multimodal and complex signal characteristics. While Res2 Net, CNN + RNN, and CNN + GRU outperform traditional CNN and lightweight models, our method further enhances feature learning, achieving superior results. The SWELL dataset consists of large-scale, single-channel ECG signals, which are uniform but extensive in quantity. Models with enhanced feature modeling capabilities, such as Res2 Net, CNN + RNN, and CNN + GRU, achieve high F1 scores, and our method further optimizes performance on this basis. The WESAD dataset contains multi-channel physiological signals, offering rich emotional information. Multi-channel temporal modeling significantly enhances the performance of Res2 Net, CNN + RNN, and CNN + GRU. However, our method surpasses all existing models, demonstrating its strength in integrating and analyzing multimodal physiological signals.

| Dataset | Emotions   | Precision (%) | Recall (%) | F1 Score (%) |
|---------|------------|---------------|------------|--------------|
| SELF    | Calm       | 96.63         | 96.80      | 96.71        |
|         | Happy      | 96.89         | 95.76      | 96.32        |
|         | Disgust    | 97.01         | 94.19      | 95.58        |
|         | Surprise   | 94.34         | 96.30      | 95.31        |
|         | Anger      | 96.50         | 96.66      | 96.58        |
|         | Sad        | 96.15         | 96.59      | 96.37        |
|         | Fear       | 96.26         | 97.07      | 96.66        |
| SWELL   | HVHA       | 97.61         | 98.62      | 98.12        |
|         | HVLA       | 98.17         | 98.17      | 98.17        |
|         | LVHA       | 99.11         | 97.91      | 98.50        |
|         | LVLA       | 97.96         | 98.18      | 98.07        |
| WESAD   | Baseline   | 99.69         | 99.54      | 99.62        |
|         | Amusement  | 96.85         | 95.56      | 96.20        |
|         | Meditation | 97.68         | 99.29      | 98.48        |
|         | Stress     | 99.73         | 98.92      | 99.33        |

Table 4. Performance evaluation of the ensemble model.

| Method                    | F1 Score (%) |              |              |
|---------------------------|--------------|--------------|--------------|
|                           | SELF         | SWELL        | WESAD        |
| DCNN <sup>37</sup>        | 32.60        | 49.03        | 76.73        |
| CNN <sup>38</sup>         | 35.92        | 65.77        | 65.39        |
| Res2 Net <sup>39</sup>    | 92.80        | 97.41        | 97.88        |
| DeepConvNet <sup>40</sup> | 16.84        | 24.38        | 43.05        |
| EEGNet <sup>41</sup>      | 25.99        | 41.64        | 71.17        |
| CNN + RNN                 | 90.64        | 94.61        | 97.06        |
| CNN + GRU                 | 89.31        | 95.09        | 96.69        |
| <b>This Study</b>         | <b>96.21</b> | <b>98.15</b> | <b>98.40</b> |

**Table 5.** Comparison with other methods.

In conclusion, our method demonstrates superior performance in complex emotion classification tasks, proving its competitiveness with state-of-the-art techniques. This study highlights the significant potential of ensemble learning in multimodal emotion recognition. By combining random seed initialization with cosine annealing learning rate adjustment, we effectively enhance model performance. The soft voting strategy not only improves classification accuracy but also ensures stability across diverse emotion categories. Furthermore, extending training epochs and optimizing the learning rate maximize the model's potential, significantly boosting overall classification performance. These findings underscore the importance of well-designed ensemble strategies and optimized training protocols in addressing complex classification challenges.

Conclusion

This study investigates the application and effectiveness of ensemble learning methods in multimodal emotion recognition using multiple physiological parameters. A high-quality dataset was constructed through systematic preprocessing of EEG, GSR, ST, and HR data collected from 38 subjects watching short videos, including denoising, normalization, sliding window segmentation, and linear concatenation. A hybrid CNN-LSTM framework was developed, integrating random seed initialization, the CAWR strategy, and ensemble methods to optimize training and enhance performance.

The results demonstrate that systematic preprocessing significantly improves data quality, providing a robust foundation for model training. The hybrid CNN-LSTM framework effectively combines the strengths of CNN and LSTM in processing spatial and temporal information. The CAWR strategy accelerates convergence and enhances performance, outperforming traditional training methods. Soft voting across models trained with different random seeds further improves generalization and robustness.

In conclusion, this study demonstrates the effectiveness of ensemble learning and dynamic learning rate adjustment mechanisms in improving emotion recognition performance. By integrating advanced technologies, the proposed method achieves higher accuracy and stability in multimodal emotion recognition tasks, particularly in capturing emotional responses to short videos. These findings highlight the potential of ensemble learning in complex emotion recognition tasks and provide a solid foundation for future research and applications.

Data availability

Data and code are publicly available at <https://github.com/LiaoEpoch/Dataset-for-Emotion-Recognition-from-Physiological-Signals-Induced-by-Short-Videos.git>. The SWELL dataset supporting this study is available at <https://cs.ru.nl/~skoldijk/SWELL-KW/Dataset.html>, and the WESAD dataset can be accessed at [https://ubi29.informatik.uni-siegen.de/usi/data\\_wesad.html](https://ubi29.informatik.uni-siegen.de/usi/data_wesad.html). Both datasets are publicly available but require permission for access. This study obtained official authorization to use SWELL and WESAD.

Received: 26 June 2024; Accepted: 31 March 2025  
Published online: 06 June 2025

References

1. Mishra, S. R. et al. Real time human action recognition using triggered frame extraction and a typical CNN heuristic[J]. *Pattern Recognit. Lett.* **135**, 329–336 (2020).
2. Febrian, R. et al. Facial expression recognition using bidirectional LSTM-CNN[J]. *Procedia Comput. Sci.* **216**(2022): 39–47. (2023).
3. Rinck, M., Primbs, M. A., Verpaalen, I. A. & Bijlstra, G. Face masks impair facial emotion recognition and induce specific emotion confusions. *Cogn. Research: Principles Implications.* **7** (1), 83 (2022).
4. Liu, Y., Zhang, X., Kauttonen, J. & Zhao, G. Uncertain facial expression recognition via Multi-task assisted correction. *IEEE Trans. Multimedia* **26**, 2531–2543 (2024).
5. Sarkar, P. & Etemad, A. Self-supervised ECG representation learning for emotion recognition[J]. *IEEE Trans. Affect. Comput.* **13** (3), 1541–1554 (2020).
6. Sánchez-Reolid, R. et al. One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity[J]. *Biomed. Signal Process. Control.* **71**, 103203 (2022).
7. Awais, M. et al. LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19[J]. *IEEE Internet Things J.* **8** (23), 16863–16871 (2020).
8. Du, X. et al. An efficient LSTM network for emotion recognition from multichannel EEG signals[J]. *IEEE Trans. Affect. Comput.* **13** (3), 1528–1540 (2020).

9. Chakravarthi, B. et al. EEG-based emotion recognition using hybrid CNN and LSTM classification[J]. *Front. Comput. Neurosci.* **16**, 1019776 (2022).
10. Abdelhamid, A. A. et al. Robust speech emotion recognition using CNN + LSTM based on stochastic fractal search optimization algorithm[J]. *Ieee Access.* **10**, 49265–49284 (2022).
11. Zhu, Z. & Mao, K. Knowledge-based BERT word embedding fine-tuning for emotion recognition. *Neurocomputing* **552**, 126488 (2023).
12. Zhou, Y., Kang, X. & Ren, F. Prompt consistency for Multi-label textual emotion detection. *IEEE Trans. Affect. Comput.* **15**(1), 121–129 (2023).
13. Wu, Y., Daoudi, M. & Amad, A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Trans. Affect. Comput.* **15**(1), 157–172 (2023).
14. Wang, X., Zhang, J., He, C., Wu, H. & Cheng, L. A novel emotion recognition method based on the feature fusion of Single-Lead EEG and ECG signals. *IEEE Internet Things J.* **11**(5), 8746–8756 (2023).
15. Yin, J. et al. Research on multimodal emotion recognition based on fusion of electroencephalogram and electrooculography. *IEEE Trans. Instrum. Meas.* **73**, 1–12 (2024).
16. Li, Q., Liu, Y., Yan, F., Zhang, Q. & Liu, C. Emotion recognition based on multiple physiological signals. *Biomed. Signal Process. Control.* **85**, 104989 (2023).
17. Kang, T. K. Emotion recognition using Short-Term Multi-Physiological signals. *KSII Trans. Internet Inform. Syst. (TIIS).* **16** (3), 1076–1094 (2022).
18. Li, R., Ding, J. & Ning, H. Emotion arousal assessment based on multimodal physiological signals for game users. *IEEE Trans. Affect. Comput.* **14**(4), 2582–2594 (2023).
19. Zitouni, M. S., Park, C. Y., Lee, U., Hadjileontiadis, L. J. & Khandoker, A. LSTM-modeling of emotion recognition using peripheral physiological signals in naturalistic conversations. *IEEE J. Biomedical Health Inf.* **27** (2), 912–923 (2022).
20. Wang, Q., Wang, M., Yang, Y. & Zhang, X. Multi-modal emotion recognition using EEG and speech signals. *Comput. Biol. Med.* **149**, 105907 (2022).
21. Wang, S., Qu, J., Zhang, Y. & Zhang, Y. Multimodal emotion recognition from EEG signals and facial expressions. *IEEE Access.* **11**, 33061–33068 (2023).
22. Zhang, Q., Zhang, H., Zhou, K. & Zhang, L. Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition. *Tsinghua Sci. Technol.* **28** (4), 673–685 (2023).
23. Hussain, M., AboAlSamh, H. A. & Ullah, I. Emotion recognition system based on two-level ensemble of deep-convolutional neural network models. *IEEE Access.* **11**, 16875–16895 (2023).
24. Aishwarya, N., Kaur, K. & Seemakurthy, K. A computationally efficient speech emotion recognition system employing machine learning classifiers and ensemble learning. *Int. J. Speech Technol.* **27**(1), 239–254 (2024).
25. Alturki, N. et al. Convolutional neural network and ensemble machine learning model for optimizing performance of emotion recognition in wild. *Multimed. Tools Appl.* <https://doi.org/10.1007/s11042-024-18178-z> (2024).
26. Iyer, A., Das, S. S., Teotia, R., Maheshwari, S. & Sharma, R. R. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. *Multimedia Tools Appl.* **82** (4), 4883–4896 (2023).
27. Zaman, K. et al. A novel driver emotion recognition system based on deep ensemble classification. *Complex. Intell. Syst.* **9** (6), 6927–6952 (2023).
28. de Santana, M. A., Fonseca, F. S., Torcate, A. S. & dos Santos, W. P. Emotion Recognition from Multimodal Data: a machine learning approach combining classical and hybrid deep architectures. *Res. Biomedical Eng.*, **39**(3), 613–638 (2023).
29. Subasi, A., Mian & Qaisar EEG-based emotion recognition using modified covariance and ensemble classifiers. *J. Ambient Intell. Humaniz. Comput.* **15** (1), 575–591 (2024).
30. Liao, Y. et al. Exploring emotional experiences and dataset construction in the era of short videos based on physiological signals[J]. *Biomed. Signal Process. Control.* **96**, 106648 (2024).
31. Koldijk, S. et al. The swell knowledge work dataset for stress and user modeling research[C]//Proceedings of the 16th international conference on multimodal interaction. : 291–298. (2014).
32. Schmidt, P. et al. Introducing wesad, a multimodal dataset for wearable stress and affect detection[C]//Proceedings of the 20th ACM international conference on multimodal interaction. : 400–408. (2018).
33. Li, H. et al. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network[J]. *Biomed. Signal Process. Control.* **72**, 103342 (2022).
34. Lindemann, B. et al. A survey on long short-term memory networks for time series prediction[J]. *Procedia Cirp.* **99**, 650–655 (2021).
35. Landi, F. et al. Working memory connections for LSTM[J]. *Neural Netw.* **144**, 334–341 (2021).
36. Huang, G. et al. Q. Snapshot ensembles: Train 1, get m for free. *ArXiv Preprint arXiv.* 170400109. <https://doi.org/10.48550/arXiv.1704.00109> (2017).
37. Zhang, Y., Cheng, C. & Zhang, Y. Multimodal emotion recognition based on manifold learning and Convolution neural network. *Multimedia Tools Appl.* **81** (23), 33253–33268 (2022).
38. Oh, S., Lee, J. Y. & Kim, D. K. The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals. *Sensors* **20** (3), 866 (2020).
39. Wu, Y., Liu, W. & Li, Q. Research on Emotion Recognition Based on Parameter Transfer and Res2Net. In 2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC) (pp. 515–518). IEEE. (2023), November.
40. Gemein, L. A. W. et al. Machine-learning-based diagnostics of EEG pathology[J]. *NeuroImage* **220**, 117021 (2020).
41. Lawhern, V. J. et al. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces[J]. *J. Neural Eng.* **15** (5), 056013 (2018).

## Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities of South-Central Minzu University (Grant Number: CZQ23031 and CZQ23029).

## Author contributions

Conceptualization, Y. L., Y. G., F. W. and L. Z.; Methodology, Y. L., Y. G. and F. W.; Investigation, Y. L., Y. W., and Z. X.; Writing – Original Draft, Y. L.; Writing – Review & Editing, Y. L., Y. G., F. W. and L. Z.; Funding Acquisition, Y. G., F. W., and L. Z.; Resources, Y. G., F. Wang.; Supervision, Y. G., F. W., and L. Z. All authors reviewed the manuscript.

## Funding

This work was supported by the Fundamental Research Funds for the Central Universities of South-Central Minzu University (Grant Number: CZQ23031 and CZQ23029).

## Declarations

### Competing interests

The authors declare no competing interests.

### Approval for human experiments

The datasets used in this study were approved by the respective authorities overseeing SWELL and WESAD. All methods followed their ethical guidelines and regulations. The construction of the SELF dataset was approved by the Research Ethics and Technology Security Committee of South-Central Minzu University (Approval No. 2022-scuec- 106.). Informed consent was obtained from all study participants. Additionally, other experiments involving these three datasets were approved by the Research Ethics and Science and Technology Safety Committee of South-Central Minzu University (Approval No. 2022-scuec- 106.).

### Additional information

**Correspondence** and requests for materials should be addressed to L.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025