VNU Journal of Science: Com. Science & Com. Eng., Vol. 41, No. 2 (2025) 16–28

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Combination of Facial Expressions and EEG for Multimodal Emotion Recognition

Le Ngoc Thanh[1], Nguyen Hong Thinh[1]*

[1] *VNU University of Engineering and Technology, Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Received 25th February 2025;
Revised 16th June 2025; Accepted 10th December 2025

Abstract: Emotion recognition is crucial in various fields, particularly in human-machine interaction. Although previous research has focused predominantly on using facial expressions or electroencephalogram (EEG) signals as the sole input for emotion recognition, integrating multiple data types into a single system remains an underexplored area. This paper aims to contribute to this evolving field by exploring a multimodal emotion recognition approach that combines facial expressions and EEG signals. For facial expression recognition, two deep learning models are developed: one for detecting facial emotions and another for classifying them. The focus is on improving the efficiency of the recognition model by leveraging modern machine learning techniques and minimizing the number of parameters. In parallel, a small dataset is collected using the Emotiv FLEX 2 Saline 32-channel EEG headset, capturing four distinct emotional states. The processing of EEG signals involves a complete workflow, from pre-processing to feature extraction, followed by mapping the features into a format suitable for emotion recognition. The evaluation results demonstrate the effectiveness of the proposed method. The facial expression recognition model achieves an accuracy of 92.5%, while the EEG-based recognition model achieves an impressive 98.7%. Furthermore, combining the output of both models improves performance, as the integration of facial expressions and EEG signals compensates for the limitations of using either data type individually.

Keywords: Facial Expression, Emotional Recognition, EEG, Deep learning.

1. Introduction

Emotions are a core element of the human experience, shaping psychological processes and behaviors such as information processing, decision-making, and social interactions. The emotions play a critical role in both personal life and professional settings, significantly affecting communication and emotional intelligence, the ability to recognize, understand of human being.

*Corresponding author.
E-mail address: hongthinh.nguyen@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.4566

Emotional intelligence is the key to successful interpersonal interactions and effective collaboration. The ability to accurately recognize emotions is growing in importance, not just in human-machine interaction, but across various fields such as healthcare, education, virtual reality, and entertainment.

In recent years, emotion recognition have generally been classified into two categories: those based on physiological signals and those based on non-physiological signals. Non-physiological signals include facial expressions, gestures, and voice. However, in some cases, the reliability of these methods can be limited, as individuals can control these physical signs, such as facial expressions and speech, to mask their true emotions. On the other hand, physiological signals, such as electroencephalograms (EEG), body temperature (T), electrocardiograms (ECG), electromyograms (EMG), galvanic skin response (GSR), and respiration (RSP), are less influenced by subjective factors and can more accurately reflect emotional states. Among these, EEG signal analysis is particularly popular in emotion research due to its non-invasive nature and the availability of affordable EEG devices, which provide accessible solutions for emotion recognition applications.

Although emotion recognition using physiological signals offers greater reliability, it has certain limitations. These signals are often weak and susceptible to interference from external noise, such as electromagnetic or mechanical noise, as well as physiological noise from the body. As a result, the practical applicability of physiological signal-based methods is more constrained compared to non-physiological signals.

Although numerous studies have explored using either type of signal independently for emotion recognition, few have investigated combining both types of signals. In this paper, we propose a multi-modal approach that integrates both non-physiological (facial expressions) and physiological (EEG) signals. The framework focuses on a discrete emotion classification model, which, due to data limitations, will classify four common emotional states: Happiness, Sadness, Fear, and Neutral. By combining both signal types, this approach addresses the weaknesses inherent in each, improving the overall accuracy of emotion recognition.

2. Related Work

2.1. Facial Expression Recognition

Facial expression recognition has become a prominent research area in the field of artificial intelligence and computer vision. In recent years, two main approaches have emerged: (i) classical methods, which involve image processing, feature extraction, and the application of traditional machine learning algorithms, and (ii) deep learning methods, where convolutional neural networks (CNNs) are applied, leveraging advances in computational power.

A typical traditional facial expression recognition system follows a multi-step process: image collection, preprocessing, feature extraction, and then classification using models such as K-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest, or linear regression. However, traditional methods have notable limitations. One challenge is to find features that generalize well across a wide range of emotions and diverse topics, considering factors such as age, gender, and culture. Additionally, the extraction and classification stages are typically handled separately, making it difficult to improve the overall performance of the system.

In contrast, deep learning methods address these issues by using end-to-end learning with automatic feature extraction, thus eliminating the need for manual feature engineering and overcoming the optimization challenges faced by classical approaches. Today, deep

learning methods dominate the field of emotion recognition and classification, with convolutional neural networks (CNNs) and recurring neural networks (RNNs) being among the most popular architectures. Additionally, newer approaches such as multitask networks and generative adversarial networks (GANs) have also gained attention because of their potential to enhance recognition accuracy and generate synthetic data for training.

Data are critical in deep learning; the larger and more diverse the dataset, the better the model performance. Several facial emotion recognition datasets have been developed to support deep learning, including the Facial Emotion Recognition 2013 (FER 2013) dataset and the Extended Cohn-Kanade Dataset (CK+). Data augmentation techniques are also commonly used to expand the dataset and improve model performance.

2.2. EEG-based Emotion Recognition

EEG signals are considered reliable physiological indicators, as they reflect the electrical activity of neurons in the human cerebral cortex. Compared to non-physiological data (such as facial expressions or gestures), EEG signals provide a more objective and consistent source of information for emotion assessment.

Emotion recognition methods based on EEG signals can be broadly categorized into two primary approaches: feature extraction combined with traditional machine learning and deep learning.

In traditional machine learning-based methods, features are manually extracted from EEG signals and subsequently fed into classifiers such as Naive Bayes, Support Vector Machines (SVM), and others for emotion recognition. Lin et al. (2018) [1] outlined the typical workflow for emotion recognition using traditional machine learning, which includes steps such as emotional stimulus presentation, signal acquisition, feature extraction, and classification. However, a significant limitation of these methods lies in their reliance on linear models, which struggle to effectively handle the non-linear, high-dimensional nature of EEG data, making it difficult to accurately distinguish between emotional states.

In contrast, deep learning-based approaches automate the feature extraction process and use models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNN) [2] to perform end-to-end emotion recognition. Deep learning methods excel at addressing the non-linearity of EEG signals by learning complex, hierarchical feature representations, enabling more accurate and efficient emotion classification. Furthermore, deep learning models can perform full mappings from input to output, which significantly simplifies the recognition process compared to traditional methods. [3] demonstrated the effectiveness of Deep Belief Networks (DBN) in emotion recognition from EEG signals, highlighting the potential of deep neural architectures for this task.

Recently, advanced deep learning techniques, including CNNs, LSTMs, Generative Adversarial Networks (GANs), and other neural network models, have gained widespread adoption in emotion recognition research using EEG signals, showing considerable improvements in accuracy and robustness compared to traditional approaches [4].

In emotion classification research, two primary approaches are typically employed: intra-subject and inter-subject classification. Intra-subject classification involves training and testing the model using data from the same individual, meaning the model is optimized to recognize emotions based on the specific neural patterns of that person. In contrast, inter-subject classification involves training the model on data from one group of individuals and testing it on data from different individuals. This approach presents a greater challenge, as the model must

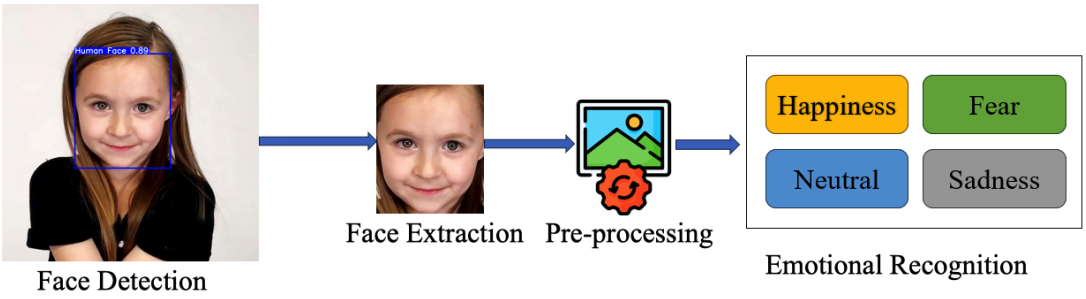

Figure 1. Flowchart Illustrating the Facial Expression Recognition Methodology.

generalize across individuals, making it possible to perform emotion classification for a subject without prior training on that subject's data.

Inter-subject classification is inherently more difficult due to the significant variability in EEG signals across individuals. EEG data does not exhibit a fixed correlation with emotional responses between different individuals, as emotional expression is influenced by numerous personal factors such as personality, culture, gender, education, prior experiences, and environmental context. As a result, individuals may exhibit distinct neurophysiological or behavioral responses, even when exposed to the same emotional stimulus. Consequently, the EEG signature of a given emotional state is unlikely to be consistent across individuals, which complicates the development of a universal classifier that can accurately predict emotions for any person. This issue—designing a robust, generalized emotion recognition model that performs well across diverse individuals—remains a significant challenge in the field.

Moreover, the use of EEG for emotion classification is subject to certain technical limitations. Different EEG systems may employ varying electrode types and configurations, which can influence both the quality and duration of the signal collection, thus introducing variability in the data. Additionally, EEG measurement is highly sensitive to motion artifacts; even slight body movements during data collection can introduce significant interference. Environmental factors, such as background noise or electromagnetic interference from nearby electronic equipment, can also distort the signals, leading to false or misleading results. These challenges underscore the complexity of developing reliable emotion recognition systems based on EEG data.

2.3. Multimodal Data Fusion for Emotion Recognition.

Recently, there has been a growing trend of emotion recognition models that integrate EEG data with recorded facial feature videos, as combining multimodal features can enhance performance. Tan et al. (2021) [5] introduced a multimodal emotion recognition approach that leverages both facial expression images and EEG data to develop a human-robot interaction system. In a similar vein, Aguinaga et al. (2021) [6] proposed a two-stage deep learning model for emotional state recognition by linking facial expressions with brain signals. They used facial expressions as indicators of emotional responses, which were then used to extract corresponding EEG segments for analysis. Saffaryazdi et al. (2022) [7] found that facial micro-expressions are more reliable than macro-expressions in conveying emotions and suggested combining these micro-expressions with brain and physiological signals for more

accurate emotion detection. Sun et al. (2020) [8] investigated the relationship between spontaneous human facial expressions and multimodal brain activity, using wearable sensors to capture data from functional near-infrared spectroscopy (fNIRS) and electroencephalogram (EEG) signals. Lastly, Wang et al. (2023) [9] proposed a deep learning model for multimodal emotion recognition that extracts both facial features and spatial information from EEG signals, merging these two types of data before passing them through a classifier to identify emotions.

3. Image-based Emotion Recognition

Figure 1 illustrates the workflow for facial expression recognition in our study. The initial step involves detecting faces within a video and subsequently extracting face images from each frame. These images are then preprocessed and fed into our proposed model for emotion recognition. For face detection, the YOLOv8 algorithm [10] was employed due to its real-time processing capabilities and high accuracy. The extracted faces were resized to 112 pixels in width and height and normalized.

The deep learning model for facial expression recognition was developed based on the architecture presented in [11]. Our proposed model builds upon the MobileNetV2 architecture by integrating a patch extraction block and a self-attention layer. From the original MobileNetV2 architecture, we removed the final 29 layers, which were deemed unsuitable for our task, and froze the remaining layers. Freezing the original network's layers accelerates the training process, as a higher learning rate can be used compared to fine-tuning the weights.

The initial layers derived from MobileNetV2 are responsible for extracting the basic features of the input, while the patch extraction block is subsequently concatenated to learn detailed features. This block comprises three distinct layers: two depth-wise separable convolutions and one point-wise convolution. The two consecutive depth-wise convolution layers divide the preceding layer's feature map into four smaller patches and learn higher-level features. The use of depth-wise separable convolutions helps improve the model's performance while reducing the number of weights. The point-wise convolution at the end of the block is responsible for aggregating features from different channels and creating new features. The output feature map of the truncated MobileNetV2 is appended with the layers of the patch extraction block, as shown in Figure 2.

The classifier is designed with three main layers, including two fully connected layers and one self-attention layer. Specifically, the self-attention layer is inserted between the two fully connected layers to enhance the model's ability to learn from features. This self-attention layer operates based on the dot product of vectors derived from the features to calculate attention weights. The softmax activation function is used to normalize these attention weights, ensuring that they have positive values and sum to one. These attention weights represent the importance of each feature, enabling the model to focus on the most relevant features.

4. EEG-based Emotion Recognition

Effective analysis of raw EEG signals, which are 1D time series, demands a specialized approach. Unlike conventional signals, EEG is characterized by its multi-channel nature, with electrode locations corresponding to crucial cortical areas. Furthermore, different frequency bands (such as alpha, beta, theta, and delta waves) within EEG carry distinct information. Therefore, to fully harness EEG's potential, it is imperative to capture its temporal, frequency, and spatial dimensions. Our method for extracting temporal and frequency information involves transforming EEG into a time-frequency

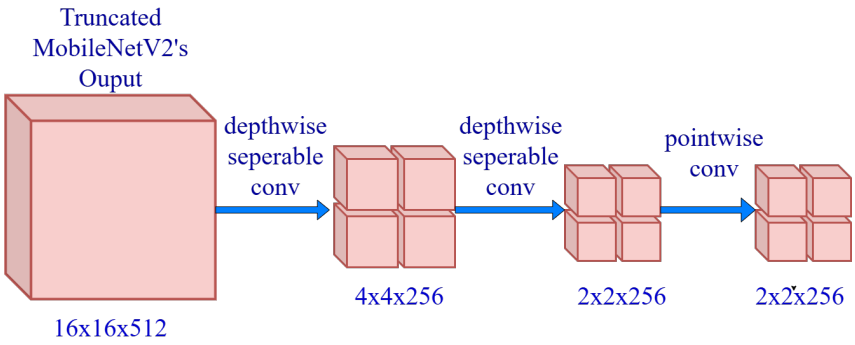

Figure 2. Integration of the Patch Extraction Module with the Truncated MobileNetV2 Backbone.

representation (spectrogram). This spectrogram is then fed into the Inception V3 network. InceptionV3, originally designed for image recognition, proves highly effective here because its inception modules excel at identifying the intricate, multi-scale patterns present in the time-frequency characteristics of EEG. Our detailed signal processing framework is outlined below.

4.1. Pre Processing

Signal preprocessing and feature extraction are crucial for raw EEG signals obtained from the headcap system. Typically, optimal frequency bands for EEG-based emotion recognition range from 1 to 51 Hz, with a particular focus on the beta and gamma bands [12].

Initially, raw signals are sampled at 200 Hz, and a Butterworth band-pass filter is applied to separate the signals into three distinct bands: 1-14 Hz (delta-theta-alpha), 14-31 Hz (beta), and 31-51 Hz (gamma). These correspond to the respective spectra, and the preprocessing steps are depicted in Figure 3. Subsequently, the filtered signals are normalized to the range [0, 1]. The three spectral signals are then split into smaller segments and transformed into three (229×229) 2D spectrograms. These spectrograms are used as input to the InceptionV3 model for feature extraction. Finally, the extracted features are aggregated and arranged in a spatial map. The details of these transformations are presented in the following sections.

4.2. Spectrogram Generation

The transformation of signals from the time domain to 2D spectrograms visualizes the signal amplitude across time and frequency, while also enabling the efficient integration of these signals into a deep learning model for feature extraction. After normalization, the signals are segmented into n non-overlapping samples. The value of the variable n and the frame size (window size) used in the Short-Time Fourier Transform (STFT) are determined by the desired dimensions of the spectrogram and the hop size. The hop size represents the number of samples by which the window shifts during each STFT computation. Selecting an appropriate hop size is critical; a value that is too small results in consecutive windows containing highly redundant signal segments, whereas a large hop size creates a spectrogram that is sparse and deficient in temporal resolution. The transformation includes the following steps: performing STFT to convert the signal from the time domain to the frequency domain; computing the spectral power; and converting the spectral power to a logarithmic scale (dB).

4.3. Feature Map Generation

To extract the features from data in the form of images (2D spectrograms), convolutional neural networks such as ResNet and Inception are commonly used. According to the experiment

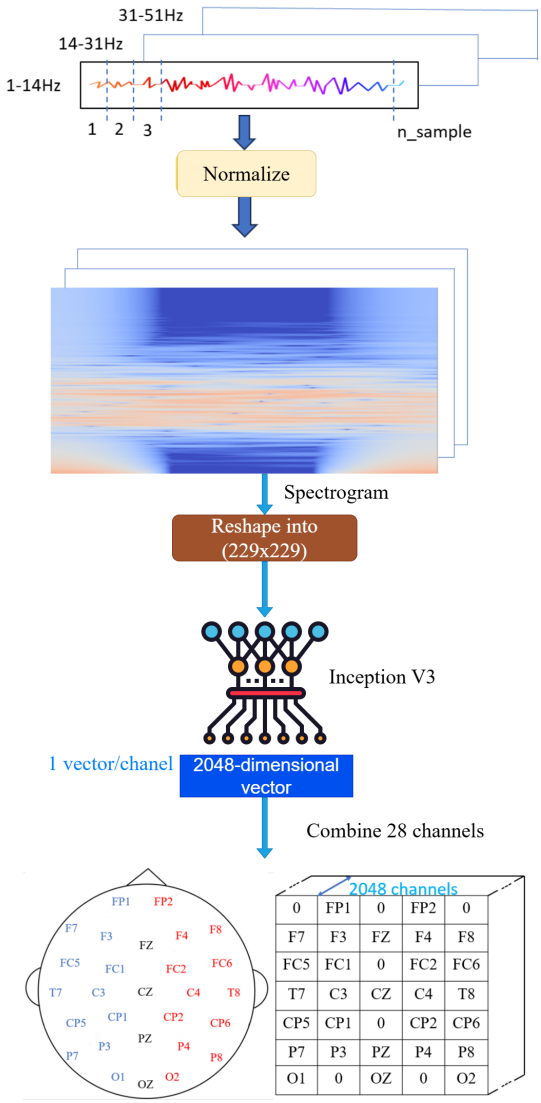

Figure 3. The Preprocessing Pipeline for EEG Signals.

from [13], Inception Net extracts all the features captured by Resnet backbone, along with extra features thanks to its multi-scale convolution architecture; therefore, this study adopted a pretrained InceptionV3 model as a feature extractor. The EEG signals employed in this study comprised 28 channels. After processing these channels sequentially through the InceptionV3 model, we obtained 28 corresponding feature vectors. The spatial distribution of these channels on the scalp, as defined by the 10-20 electrode placement system, allowed us to arrange the feature vectors into a spatial map. This map, representing a top-down view of the scalp, effectively embeds the spatial information of the channels, thereby enhancing the deep learning model's ability to learn features from adjacent channels. We found that a (7×5) map provided an optimal representation for the 28 channels. Considering that the InceptionV3 model produces output vectors of dimension 2048, the spatially arranged feature vectors formed a map that can be conceptualized as a (7×5) image with 2048 channels.

4.4. The Recognition Model

The signals are transformed into maps with dimensions $(7 \times 5 \times 2048)$, necessitating a deep learning classification model specifically designed for this input format. The network architecture is straightforward, commencing with three sequential convolutional layers for spatial feature extraction, utilizing kernel sizes of (3×3), (2×2), and (1×1), respectively. These layers are followed by a max-pooling layer to reduce computational complexity and a flattening layer to convert the data from a three-dimensional to a one-dimensional format. The classifier comprises two fully connected layers (FCLs), with the second layer serving as the final classification layer to compute probabilities for the four emotion labels.

5. Experiments and Results

5.1. Dataset

Facial Expression Dataset: For facial expression-based emotion recognition, we utilized images from the Real-world Affective Database (RAF-DB). This large-scale, diverse database, collected from internet sources, features varying lighting conditions and demographics. The RAF-DB dataset includes 52% female, 43%

male, and 5% uncertain gender. Racially, it comprises 77% White, 8% African-American, and 15% Asian individuals. While RAF-DB contains images depicting seven common emotional states (happiness, sadness, surprise, fear, anger, disgust, and neutral), this study focused on four: happiness, sadness, fear, and neutral.

EEG Signals Dataset: To enhance data diversity, we independently collected an EEG signal dataset in the Signal and Systems Laboratory at the Faculty of Electronics and Telecommunications, University of Engineering and Technology. We conducted EEG data collection from 15 students, but due to the connection between the EEG measurement equipment and the scalp, it is sometimes not good; resulting in a very noisy signal. Using these data can lead to errors in the deep learning machine learning model. Therefore, after reviewing all the recordings; we selected data from 7 people, whose data collection process achieved a signal quality reported by the device of 80% or higher. In addition, the data participants also watched many different video clips; and the data collection process was very long (15-20 minutes). In summary, our EEG data were measured from 7 subjects, all of whom were male aging from 20 to 22 years old. For comparison, the SEED dataset [14] has a total of 15 subjects, and the DEAP dataset[15] contains signals from 32 participants. This dataset focused on the recognition of four emotional states: happiness, sadness, fear, and a neutral state. This selection was made because certain emotions are challenging to reliably induce through video or short clip stimuli, leading us to collect data for three strong emotions (happy, sad, scared) and a neutral resting state.

The experimental setup, illustrated in Figure 4, involved recording data from individual participants. Each participant wore an EEG headset and sat before a computer screen, with facial expressions captured by a synchronized camera. Participants watched emotion-eliciting videos corresponding to the target emotion. Although the FLEX 2 Saline headset has 32 channels, only 28 were available for analysis after accounting for synchronization channels. Due to challenges with signal quality related to thick hair, the final EEG dataset included only male participants, aged 19 to 22, who met the data quality standards.

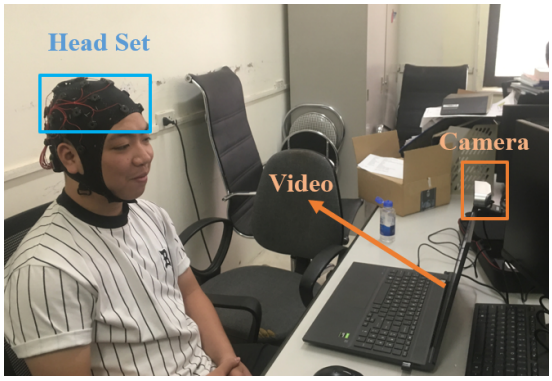

Figure 4. The Experiment Setup.

5.2. *Emotion Recognition Using Facial Expression*

These EEG data were then mixed for processing. 80% of the data was used for training and 20% for testing. The training was conducted for a maximum of 30 epochs, with early stopping implemented to terminate training if no improvement in validation accuracy was observed over a consecutive number of epochs. The initial learning rate was set to (1×10^{-3}). To mitigate overfitting, the learning rate was configured to decrease when the model exhibited no improvement, with a minimum allowable value of (1×10^{-6}).

The trained model achieved strong performance, with training and validation accuracies of 95% and 90%, respectively. The evaluation metrics for the test dataset are presented in Table 1. When evaluated on the test dataset, the model achieved an accuracy

exceeding 92%. The minimal difference in accuracy across the training, validation, and test datasets indicates that the model effectively learned the underlying data features, demonstrating good generalization, and avoiding over-fitting. With this high accuracy, the model fulfills the requirements for practical emotion recognition applications.

Table 1. Quantitative Results on Facial Expression Dataset (%)

Accuracy	Precision	Recall	F1 Score
92	92	92	91

As shown in Figure 5, the model's performance on the test set reveals a tendency to confuse sadness with the neutral state and sadness with fear. This can be attributed to the fact that subtle expressions of sadness are often difficult to discern from a neutral expression, leading to misclassification. Similarly, while fear is often visibly expressed, it shares some visual similarities with sadness, particularly when tears are present. In contrast, happiness exhibits distinct facial features, resulting in more accurate classification. The model demonstrates high accuracy in recognizing happiness and neutral emotions, but lower accuracy for sadness and fear. Specifically, the model frequently misclassifies faces expressing sadness and fear. Conversely, it rarely misclassifies neutral emotions, but often incorrectly classifies other emotional states as neutral. This evaluation highlights the importance of combining facial expression data with EEG signals in this research. By integrating both sources of information, the system aims to improve emotion recognition accuracy, especially in cases where emotions are not clearly visible in facial expressions.

5.3. Emotion Recognition Using EEG Signal

The EEG dataset was initially split into 15% for testing. The remaining 85% was further divided into 85% for training and 15%

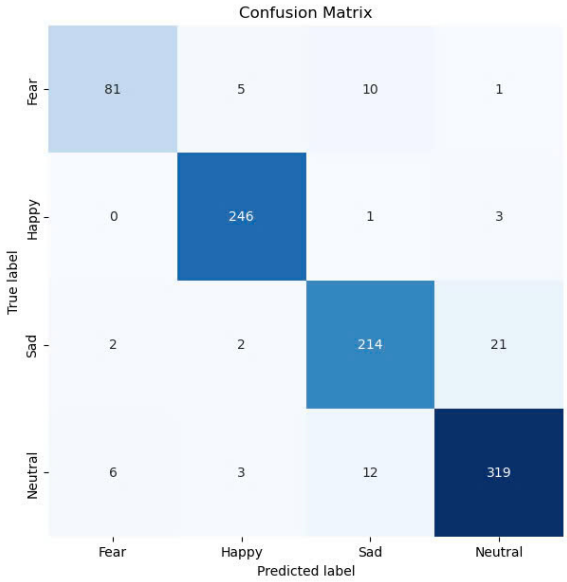

Figure 5. Confusion Matrix of the Facial Expression Recognition Model.

for validation, resulting in final proportions of 72.25% training, 12.75% validation, and 15% testing. The model was trained for up to 100 epochs with an initial learning rate of (1×10^{-4}). The learning rate was gradually decreased during training if no improvement in validation performance was observed, with a minimum allowable value of (1×10^{-6}). Due to the application of early stopping, the training process terminated before reaching the 100th epoch, and the weights of the best-performing model were restored instead of the final state's weights.

The optimal model achieved accuracies of 99.91% and 98.98% on the training and validation datasets, respectively. Evaluation metrics for the test dataset are presented in Table 2 and Figure 6. The results indicate that the model exhibited lower accuracy for samples associated with happiness compared to other emotional states. Conversely, the neutral state displayed the lowest recall value, indicating that the model frequently failed to detect signal segments representing this state.

Despite achieving favorable evaluation

Table 2. Quantitative Results on EEG Dataset (%)

Accuracy	Precision	Recall	F1 Score
98.71	98.71	98.77	98.72

metrics, the trained model exhibits several limitations. Due to inter-individual variability in EEG signals within the same emotional state, a model trained solely on data from a single individual demonstrates poor generalization when tested on signals from other individuals. Furthermore, the model's performance is not consistent throughout the entire EEG signal recording. This inconsistency stems from the participants' difficulty in maintaining a sustained emotional state during the measurement process, leading to intermittent emotional responses. Lastly, the time-intensive preprocessing of EEG signals renders the current model unsuitable for real-time applications.

Several feasible approaches are proposed for future work to mitigate the issue of inter-subject variability in EEG signals. First, we can augment and diversify our dataset by collecting EEG signals from individuals of different age groups, genders, and cultural backgrounds. In parallel, we can leverage advanced machine learning techniques to improve model generalization. Specifically, domain adaptation and transfer learning are used to better align data distributions and optimize deep learning model performance. Additionally, we are exploring the use of generative models, such as GANs, for data augmentation. Another strategy involves incorporating time-domain features into the model architecture. These features are processed through fully connected and attention layers, and then concatenated with the output of convolutional layers before the final classification stage.

To benchmark our proposed method, we trained two established models, EEG-ITNet [16] and SST-EmotionNet [17], using data collected in our laboratory. These models are specifically

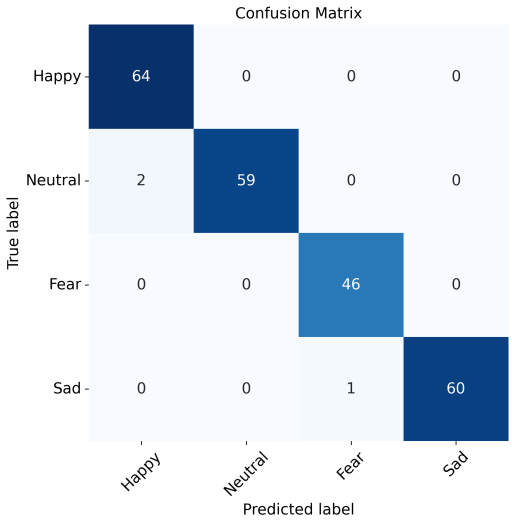

Figure 6. Confusion Matrix for EEG-based Emotion Recognition Model.

designed for EEG-based emotion recognition and have demonstrated strong performance on widely used datasets such as SEED, SEED-IV, and OpenBCI. A comparative analysis of the results is presented in Table 3. Our proposed method outperformed both EEG-ITNet and SST-EmotionNet across all four evaluation metrics, validating its effectiveness for emotion recognition from small EEG datasets. EEG-ITNet also yielded positive results, albeit with lower performance than our proposed method. SST-EmotionNet, however, exhibited significantly lower performance, indicating its limited suitability for our experimental data. While SST-EmotionNet may be optimized for larger datasets, such as SEED or SEED-IV, it appears less effective for smaller datasets.

5.4. Results of Multimodal Fusion

As previously mentioned, both facial expression-based and EEG-based emotion recognition models possess distinct strengths and limitations. Integrating these two data modalities leverages the advantages of each, resulting in more accurate and reliable outcomes. During experiments, it was observed that there was a

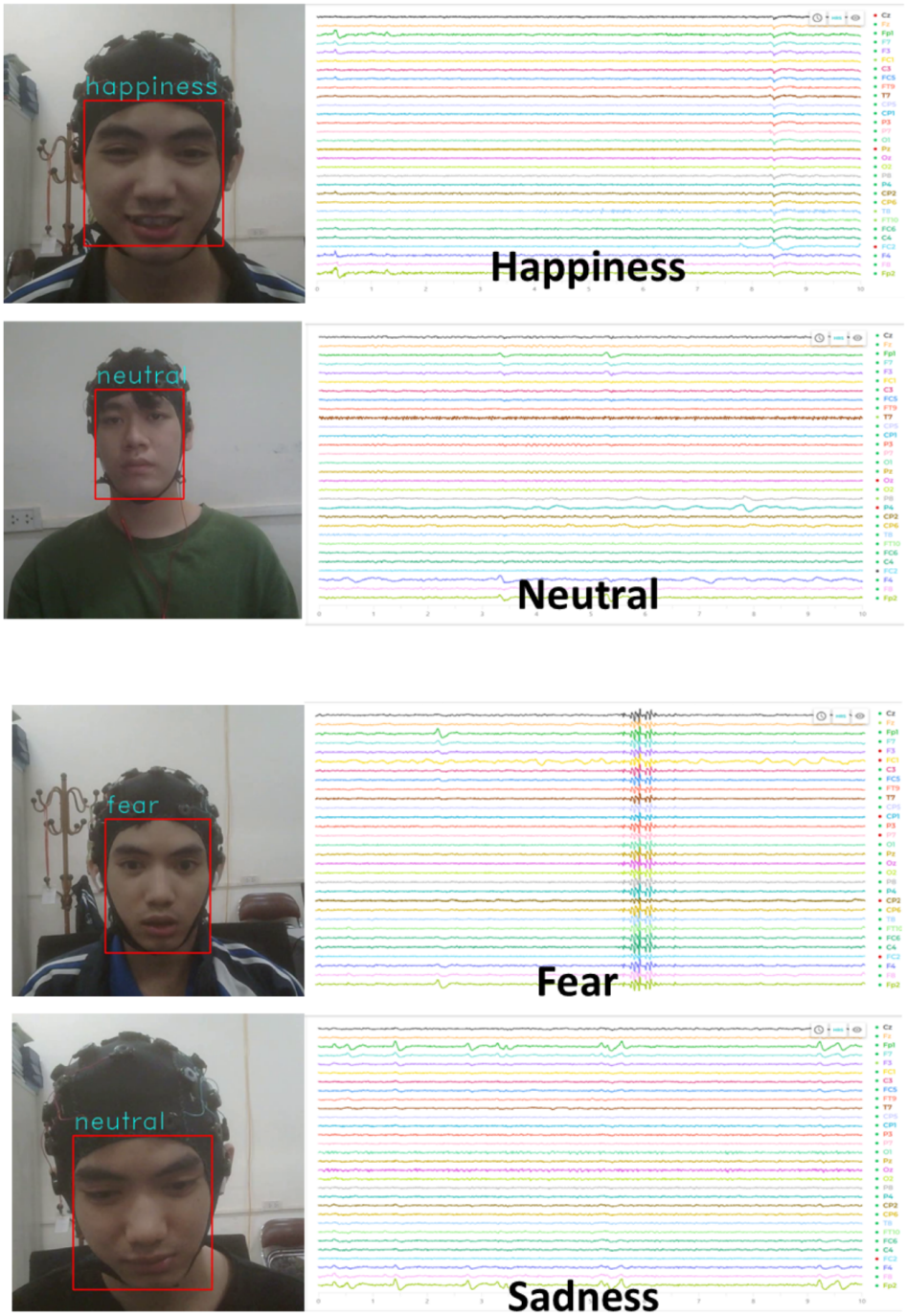

Figure 7. Multimodal Fusion Results: Combining EEG and Facial Expressions.

Table 3. Performance Comparison (%)

Method	Accuracy	Precision	Recall	F1 Score
SST-EmotionNet	74.15	77.41	73.03	74.60
EEG-ITNet	94.61	94.81	93.89	93.62
Proposed method	**98.71**	**98.71**	**98.77**	**98.72**

2-second delay between the facial expressions captured by the camera and the corresponding EEG signals displayed in the Emotiv software. This was due to the delay in the wireless connection and data transmission from the EEG helmet (Emotiv Flex 2) and the computer to receive the data. To ensure synchronization, the first 2 seconds of EEG data were discarded to compensate for the observed 2-second delay in the EEG signal relative to the video recording.

For each evoked emotion, a 3-second EEG segment was selected to ensure sufficient information for preprocessing and input modeling (3 s is the duration of each EEG epoch used in the framework for preprocessing and extraction features [18]). While real-time emotion recognition from videos is feasible, the average time required for raw EEG signal preprocessing and inference was 1.5 seconds.

In Figure 7, happiness, neutral, and fear have clear facial expressions, thus, the image-based model can accurately recognize these emotions. Simultaneously, predictions based on the corresponding EEG signal segments also show similar emotional states, demonstrating the compatibility between the two data types. However, for sadness, the participant did not have significant expressions. This led the image-based model to misclassify it as neutral. However, the predictions based on the EEG signal segment correctly identified the emotional state as sadness. These findings highlight the crucial role of EEG signals in overcoming the limitations of image-based methods, particularly for emotions with subtle or unclear expressions.

6. Conclusion

In this paper, we present a novel framework for emotion recognition that combines facial expressions with EEG signals. For facial expression analysis, we introduce a new network architecture based on MobileNet, enhanced with an Attention mechanism. This design not only improves recognition performance, but also ensures that the model remains compact and efficient, making it suitable for real-world deployment with limited computational resources.

For EEG signal processing, we propose a method that generates spectral maps, preserving critical information within each frequency band while maintaining the spatial distribution of signal power across electrode locations. This approach enables more effective encoding of EEG data compared to raw signal inputs, leading to enhanced performance when processed by deep learning models.

By fusing the emotion recognition outputs from both facial expression and EEG data, our system achieves superior performance, particularly in cases where emotional expressions are subtle or ambiguous. The flexibility of our framework allows for future extensions to recognize a wider range of emotions and accommodate diverse subject characteristics. Additionally, due to its compact architecture, the system is ideally suited for embedded applications, such as robotic systems, where resource constraints are a key consideration.

References

[1] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, X. Yang, A Review of Emotion Recognition Using Physiological Signals, Sensors 18 (7) (2018) 2074–2115, https://doi.org/10.3390/s18072074.

[2] A. Craik, Y. He, J. L. Contreras-Vidal, Deep Learning for Electroencephalogram (EEG) Classification Tasks: A Review, Journal of neural engineering 16 (3) (2019) 031001, https://doi.org/10.1088/1741-2552/ab0ab5.

[3] F. Movahedi, J. L. Coyle, E. Sejdić, Deep Belief Networks for Electroencephalography: A Review of Recent Contributions and Future Outlooks, IEEE journal of biomedical and health informatics 22 (3) (2017) 642–652, https://doi.org/10.1109/JBHI.2017.2727218.

[4] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. L. García, J. M. Gorriz, U. R. Acharya, Emotion Recognition in EEG signals Using Deep Larning Methods: A review, Computers in Biology and Medicine 165 (2023) 107450, https://doi.org/10.1016/j.compbiomed.2023.107450.

[5] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, C. F. Caiafa, A Multimodal Emotion Recognition Method Based on Facial Expressions and Electroencephalography, Biomedical Signal Processing and Control 70 (2021) 103029, https://doi.org/10.1016/j.bspc.2021.103029.

[6] A. R. Aguiñaga, D. E. Hernandez, A. Quezada, A. Calvillo Téllez, Emotion Recognition by Correlating Facial Expressions and EEG Analysis, Applied Sciences 11 (15) (2021) 6987, https://doi.org/10.3390/app11156987.

[7] N. Saffaryazdi, S. T. Wasim, K. Dileep, A. F. Nia, S. Nanayakkara, E. Broadbent, M. Billinghurst, Using Facial Micro-Expressions in Combination with EEG and Physiological Signals for Emotion Recognition, Frontiers in Psychology 13 (2022) 864047, https://doi.org/10.3389/fpsyg.2022.864047.

[8] Y. Sun, H. Ayaz, A. N. Akansu, Multimodal Affective State Assessment Using Fnirs+ EEG and Spontaneous Facial Expression, Brain sciences 10 (2) (2020) 85, https://doi.org/10.3390/brainsci10020085.

[9] S. Wang, J. Qu, Y. Zhang, Y. Zhang, Multimodal Emotion Recognition from EEG Signals and Facial Expressions, IEEE Access 11 (2023) 33061–33068, https://doi.org/10.1109/ACCESS.2023.3263670.

[10] R. Varghese, M. Sambath, Yolov8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness, in: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS), IEEE, 2024, pp. 1–6, https://doi.org/10.1109/ADICS58448.2024.10533619.

[11] J. L. Ngwe, K. M. Lim, C. P. Lee, T. S. Ong, A. Alqahtani, PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition, IEEE Access 12 (2024) 79327–79341, https://doi.org/10.1109/ACCESS.2024.3407108.

[12] V. Jadhav, N. Tiwari, M. Chawla, EEG-Based Emotion Recognition Using Transfer Learning Based Feature Extraction and Convolutional Neural Network, in: ITM Web of Conferences, Vol. 53, EDP Sciences, 2023, p. 02011, https://doi.org/10.1051/itmconf/20235302011.

[13] D. McNeely-White, J. R. Beveridge, B. A. Draper, Inception And Resnet Features Are (Almost) Equivalent, Cognitive Systems Research 59 (2020) 312–318, https://doi.org/10.1016/j.cogsys.2019.10.004.

[14] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, A. Cichocki, Emotionmeter: A Multimodal Framework for Recognizing Human Emotions, IEEE transactions on cybernetics 49 (3) (2018) 1110–1122, https://doi.org/10.1109/TCYB.2018.2797176.

[15] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A Database for Emotion Analysis Using Physiological Signals, IEEE transactions on affective computing 3 (1) (2011) 18–31, https://doi.org/10.1109/T-AFFC.2011.15.

[16] A. Salami, J. Andreu-Perez, H. Gillmeister, EEG-ITNet: An Explainable Inception Temporal Convolutional Network for Motor Imagery Classification, IEEE Access 10 (2022) 36672–36685, https://doi.org/10.1109/ACCESS.2022.3161489.

[17] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, J. Wang, SST-EmotionNet: Spatial-Spectral-Temporal Based Attention 3D Dense Network for EEG Emotion Recognition, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2909–2917, https://doi.org/10.1145/3394171.3413724.

[18] D. Ouyang, Y. Yuan, G. Li, Z. Guo, The Effect of Time Window Length on EEG-Based Emotion Recognition, Sensors 22 (13) (2022) 4939, https://doi.org/10.3390/s22134939.