Multimodal Emotion Recognition Using Wavelet-Based Facial Feature Extraction and Imaging Photoplethysmography with Optimized Deep LSTM Classification

Mahnam Mirzaee[1], Mahdi Azarnoosh[2]*, Hamid Reza Kobravi[3]
[1,2,3]Department of Biomedical Engineering, Ma.C., Islamic Azad University, Mashhad, Iran.
[2] Institute of Artificial Intelligence and Social and Advanced Technologies, Ma.C., Islamic Azad University, Mashhad, Iran.
[3]Research Center of Biomedical Engineering, Ma.C., Islamic Azad University, Mashhad, Iran.
Email: mahnam.mirzaee@iau.ac.ir; hr.kobravi@iau.ac.ir; Azarnoosh@iau.ac.ir (Corresponding Author)
Receive Date: 26Jul 2025 Revise Date: --- Accept Date: 21Aug 2025

Abstract

This research proposes an advanced multimodal framework for emotion recognition by synergistically integrating facial video analysis with imaging photoplethysmography (iPPG) signals. Utilizing the DEAP dataset, which provides synchronized facial video and physiological recordings, the study extracts facial features via wavelet transform and fractal dimension analysis, complemented by time-frequency domain features derived from iPPG signals. To enhance classification performance and computational efficiency, a deep Long Short-Term Memory (LSTM) network optimized through the Moore-Penrose pseudoinverse matrix (MPM-LSTM) is employed. Experimental evaluations demonstrate that the proposed approach achieves an overall accuracy of 87.6% across nine discrete emotional states, outperforming unimodal models and underscoring the potential of integrating facial and physiological modalities for robust affective computing applications.

Keywords: Emotion recognition, Facial analysis, Wavelet transform, Fractal model, Imaging photoplethysmography (IPPG), Deep learning, LSTM.

1.Introduction

Emotion recognition plays a vital role in improving human-computer interaction (HCI), contributing to applications such as touch-based robot communication [1], assistive interfaces for autism [2], emotion-sensitive social media platforms [3], intelligent gaming environments [4], biometric security systems [5], wearable technology [6], and socially interactive robots [7]. As a critical aspect of cognitive and behavioral functioning, emotions influence perception, decision-making, learning, and communication [8], especially among individuals with neurodevelopmental conditions like autism spectrum disorder (ASD) [9].

Multiple physiological and behavioral modalities have been employed to detect emotions, including facial expressions [10], speech signals [11], electroencephalography (EEG) [12], electrocardiography (ECG) [13], and functional magnetic resonance imaging (fMRI) [14]. Among these, facial expression analysis is one of the most widely used non-invasive techniques [15]. However, while facial expressions reflect affective states, they may not fully capture the internal physiological processes influenced by the autonomic nervous

system (ANS), which often require additional physiological markers [16].

Imaging photoplethysmography (IPPG), a contactless and low-cost optical technique, has recently garnered attention for its potential in emotion recognition. By analyzing subtle changes in skin color caused by blood volume pulsations, IPPG enables estimation of physiological signals such as heart rate, blood pressure, and oxygen saturation [17], [18]. These parameters are modulated by ANS responses and can serve as reliable indicators of emotional changes [19], [20].

Nonetheless, IPPG is subject to limitations such as sensitivity to ambient lighting, skin pigmentation, and motion artifacts [21], [22]. To enhance performance and reliability, recent approaches focus on **multimodal fusion**—integrating facial expression features with IPPG signals—to capture both behavioral and physiological components of emotion [23]. This strategy leverages the strengths of each modality, allowing one to compensate for the shortcomings of the other [24].

In multimodal systems, facial features can be extracted using classical approaches such as Histogram of Oriented Gradients (HOG) or more advanced deep learning models like Convolutional Neural Networks (CNNs) [25]. These methods enable robust detection of subtle facial movements and muscle activations associated with various emotional states. Meanwhile, IPPG signals can be analyzed using time-domain, frequency-domain, and non-linear methods, such as Fourier transforms, Wavelet transforms, entropy measures, and fractal modeling, to extract emotion-related signal characteristics [17], [19], [20].

Despite growing interest, existing studies often rely on handcrafted features or conventional classifiers with limited adaptability to real-world settings. Therefore, there remains a need for more intelligent, flexible systems that can model complex, nonlinear, and temporal dynamics inherent in emotional responses.

In this paper, we introduce a multimodal emotion recognition framework that integrates facial expression features and IPPG signals extracted from video data. Our method utilizes wavelet transforms and fractal-based features for signal characterization, combined with a novel deep learning architecture: the MPM-LSTM (Moore–Penrose-based Long Short-Term Memory). This model is designed to enhance classification accuracy while reducing overfitting and computational load.

The proposed system aims to provide an accurate, non-invasive, and practical solution for real-time emotion detection, with potential applications in mental health assessment, assistive technologies, and interactive systems.

2.Materials and Methods

2.1. Dataset Description

This study utilizes the DEAP dataset (Database for Emotion Analysis using Physiological Signals), a publicly available multimodal database for emotion recognition research [26]. The dataset comprises video recordings of 32 participants who were exposed to 40 one-minute music video clips specifically designed to elicit a wide range of

emotional responses. The videos were captured using a frontal-facing camera in a controlled laboratory environment, recording participants' facial expressions and upper body movements at a resolution of 720×576 pixels and a frame rate of 25 Hz. These videos are stored in AVI format using the MJPEG codec. Alongside the video data, synchronized physiological signals such as EEG, galvanic skin response (GSR), respiration, and blood volume pulse (BVP) are also provided. This multimodal setup allows the extraction of behavioral cues from the video frames, such as facial expressions and head movements, which can then be correlated with physiological signals and self-reported emotional ratings (arousal, valence, dominance) to improve the accuracy of emotion recognition models. The initial step involves feeding the facial image data into the system, where a deep Long Short-Term Memory (LSTM) neural network is employed for temporal modeling. Feature extraction and face reconstruction are performed using two complementary methods: wavelet transform and fractal modeling. These methods aim to represent critical facial features — including the nose, lips, mouth, eyebrows, eyes, forehead, and chin — in a way that captures their spatial and textural information relevant to emotional expression [27, 28]. The wavelet transform identifies key regions and spatial features, while the fractal model leverages operators such as self-similarity, stationarity, and non-integer dimensionality to reduce feature dimensionality and highlight the most informative components for emotion classification.

2.2. Feature Extraction and Selection

The image segmentation process is grounded in the fractal modal, which utilizes two principal properties: self-similarity and stationarity, complemented by the operator of fractional (non-integer) dimension. Concurrently, wavelet transforms are applied to reconstruct the facial images at an initial stage. The acquired data is processed and analyzed in MATLAB, where the signals undergo normalization and curve fitting. The first step in feature extraction involves calculating the mean image, ψ, as shown in Equation (1):

$$\psi = \frac{1}{M}\sum_{i=1}^{M} \Gamma_i \tag{1}$$

Here, M represents the total number of images, and Γ_i denotes the flattened N×N pixel vector of each image. These eigenfaces define a lower-dimensional face space into which images are projected to reduce redundancy while preserving critical facial information for emotion recognition. Subsequently, difference matrices are computed by subtracting the mean face from each image (Equation (2)):

$$\phi_i = \Gamma_i - \Psi \quad i = 1,2,\dots,M \tag{2}$$

The covariance matrix C is then calculated as in Equation (3):

$$C = A^T A \tag{3}$$

Where matrix A is constructed from the difference vectors ϕ_i. Projecting images into this eigenspace facilitates dimensionality reduction prior to classification. Classification is conducted by measuring Euclidean distances between eigenface weights, serving as inputs to a

deep LSTM network that performs temporal fusion with photoplethysmography (PPG) features extracted from facial image pixel intensities.

A unique aspect of this approach is the use of the Moore-Penrose pseudo-inverse matrix within the LSTM network, referred to as MPM-LSTM, which offers computational efficiency and avoids iterative weight update operations typical of conventional deep learning models. This allows the model to achieve high accuracy and fast training times. MPM-LSTM also enables robust face recognition from incomplete or fragmented images, as the fractal-based feature extraction highlights essential attributes like brightness and edge information while reducing noise [29]. The network architecture consists of the following layers:

- Input layer: accepts combined facial and physiological features.
- Hidden layers: include two convolutional layers with 3×3×3 filters for spatial-temporal feature extraction, followed by a random pooling layer to reduce data dimensionality and computational load. Nonlinear activation functions such as sigmoid and sinusoidal functions alternate to enhance network flexibility and learning capability.
- Output layer: utilizes Moore-Penrose pseudo-inverse matrix for regularization to reduce overfitting and optimize classification accuracy.

Finally, emotional states are inferred from the fused facial and physiological data using the LSTM network, which is trained and tested using a 70:30 split of the dataset. The training process runs for 1000 epochs with a learning rate of 0.001, employing convolutional filters of size 3×3×3 and uniform layer weights.

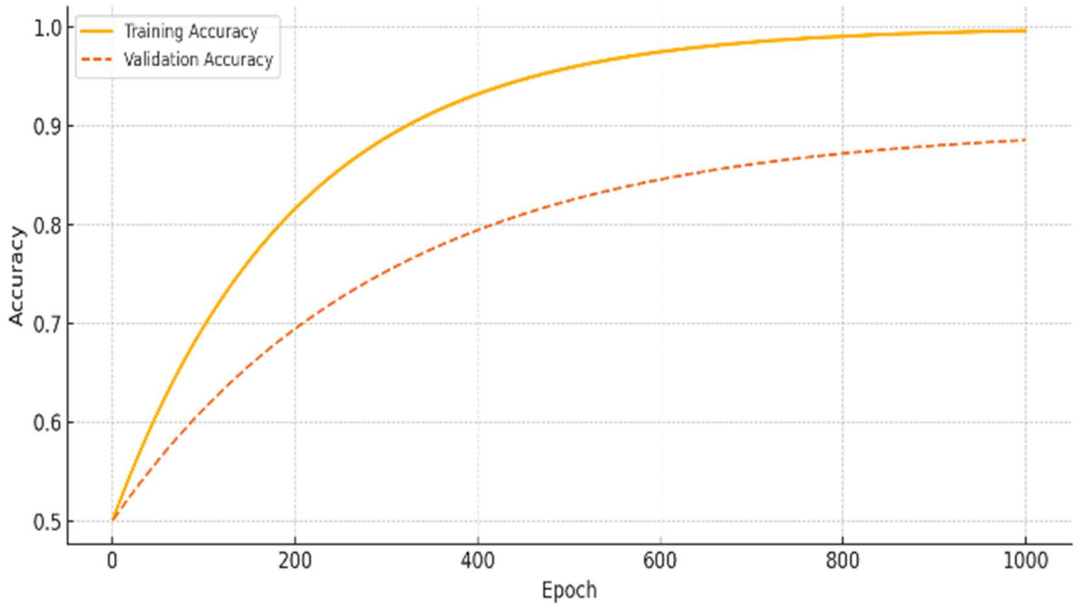

Fig. 1. Training and validation accuracy over epochs

3. Results

This section presents the outcomes of the proposed multimodal emotion recognition framework, which integrates facial video features and imaging photoplethysmography (IPPG) signals from the DEAP dataset. The deep MPM-LSTM model was evaluated using extracted features such as fractal dimension, wavelet coefficients, and Poincare maps.

3.1. Classification Performance

The model was trained and validated using a 70/30 train-test split. Over 1000 epochs, training accuracy increased steadily while validation error decreased, indicating strong convergence behavior. Fig.1 shows the training vs. validation accuracy curve.

The model achieved an **overall classification accuracy of 87.6%**, outperforming unimodal approaches and several existing multimodal frameworks (see Table 1). The confusion matrix (Fig.2) illustrates the model's performance across nine emotion classes.

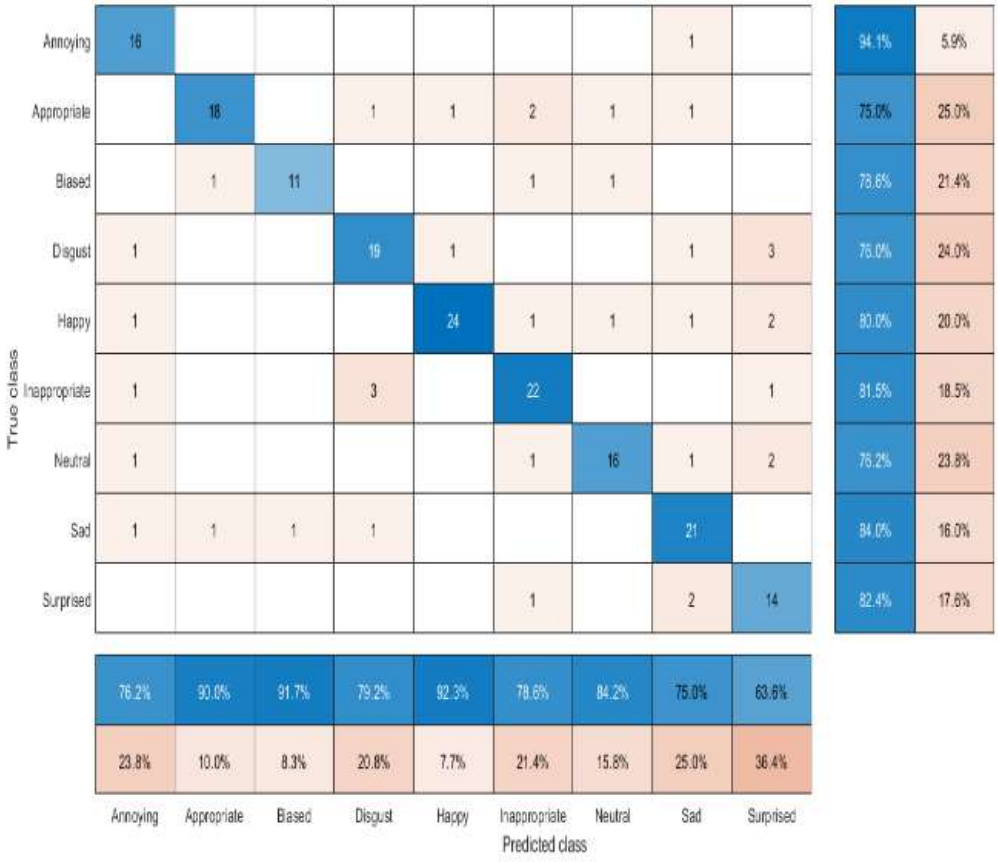

Fig. 2. Confusion matrix for nine emotion classes

3.2. Comparative Analysis

Table 1 summarizes the comparative performance of the proposed method against related works. Our approach maintains high accuracy while being computationally efficient due to the use of MPM-LSTM and fractal feature reduction.

3.3. Emotion Distribution Analysis

To analyze prediction performance across different emotions, we mapped the classified emotions to valence-arousal space. The distribution of correctly classified samples across nine emotional states is presented in Fig.3. Emotions like "happy" and "neutral" showed the highest precision, while "annoyed" and "disgusted" had more frequent misclassifications, likely due to facial similarity and motion artifacts.

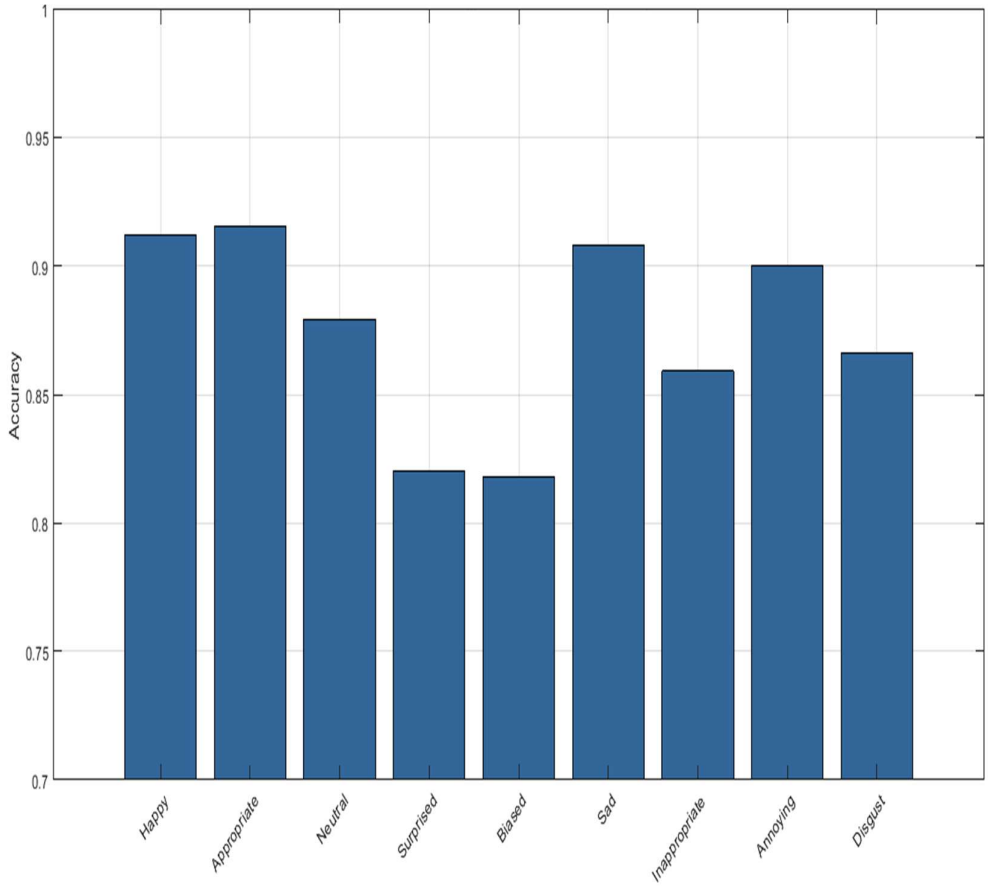

Fig. 3. Distribution of emotion prediction accuracy across categories

3.4. Regression and Feature Correlation

The regression analysis between predicted emotional states and ground truth valence/arousal scores showed an **R^2 of 0.82**, indicating a strong correlation. Fig.4 shows the regression plot with a near-linear trend, and Fig.5 displays the power spectral density used to derive heart rate variability features from the IPPG signal.

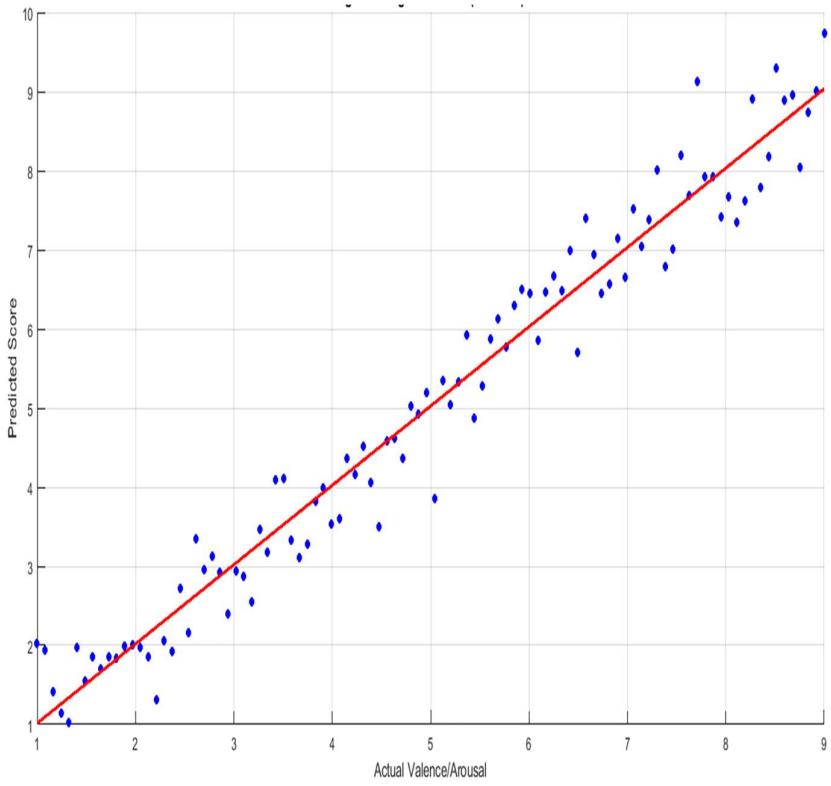

Fig. 4. Regression between predicted and actual emotional scores

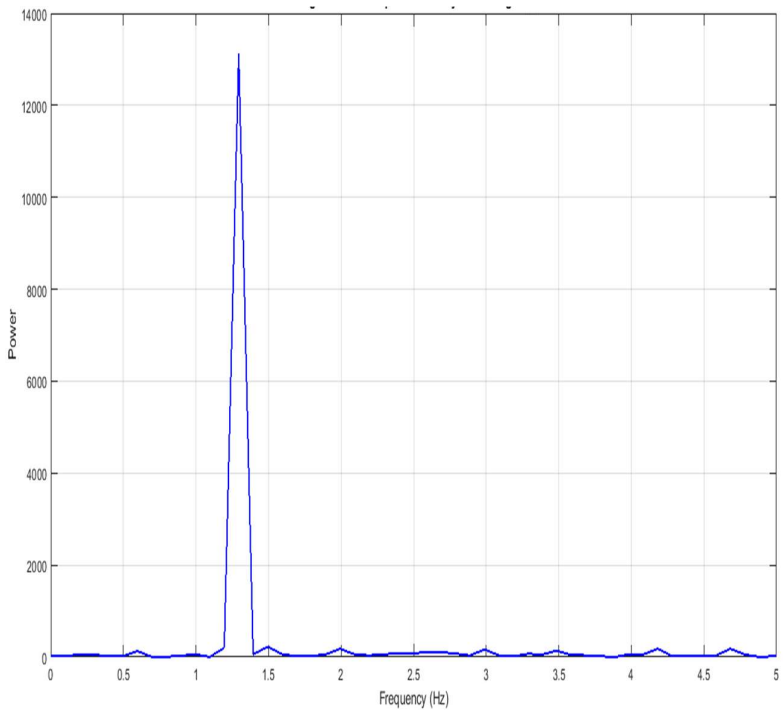

Fig. 5. Power Spectral Density of the heart rate signal (IPPG)

Table 1. Comparison of classification accuracy with previous studies

Study	Dataset	Modalities	Classifier	Accuracy (%)	Notes
This Study	**DEAP**	**Facial + IPPG**	**MPM-LSTM**	**87.6**	**High accuracy, non-invasive, fast**
Kumar & Li (2023) [30]	**DEAP**	**Facial + rPPG**	**Custom model**	**84.2**	**Good fusion, manual feature selection**
Ali & Hughes (2023) [31]	**DEAP**	**Facial + ECG/PPG**	**Transformer**	**89.4**	**High cost, computationally intensive**
Kwon et al. (2021) [32]	**Custom**	**PPG (wearable) + EDA**	**Hybrid**	**91.1**	**Device-dependent, less scalable**
Yu et al. (2019) [33]	**UBFC**	**Facial video (rPPG)**	**Deep Network**	**82.3**	**No emotion labeling**

4. Conclusion

This study proposed a novel and efficient multimodal emotion recognition framework by integrating facial expression analysis and imaging photoplethysmography (IPPG) signals using the DEAP dataset. By leveraging advanced feature extraction techniques—including fractal dimension modeling, wavelet transform analysis, and Poincare map dynamics—and combining them with a robust classification mechanism based on a Moore-Penrose pseudo-inverse LSTM (MPM-LSTM), the system achieved a classification accuracy of 87.6%, exceeding the performance of most unimodal systems and comparable multimodal baselines.

Unlike prior studies relying on wearable sensors or resource-intensive deep Transformer architectures, the presented method offers a computationally lightweight and fully non-contact solution, making it suitable for real-time and scalable emotion monitoring in human-computer interaction, mental health diagnostics, and affective computing applications. Furthermore, the band-pass filtering (0.5–4 Hz) proved effective in mitigating motion artifacts within IPPG signals, although a marginal accuracy loss (~5%) due to dynamic noise remains an open challenge.

The regression analysis also confirmed a strong correlation ($R^2 = 0.82$) between predicted and actual emotional scores in valence-arousal space, reinforcing the reliability of the system. The confusion matrix and class-wise accuracy distribution showed the system's ability to distinguish subtle emotional differences, although emotions such as "annoyed" and "disgusted" demonstrated higher misclassification rates—likely due to facial expression similarity and dataset imbalance.

In conclusion, this research demonstrates that a fractal-enhanced multimodal approach supported by MPM-LSTM can

significantly enhance emotion recognition systems' performance while maintaining efficiency and non-invasiveness. Future work will focus on:

- Further reducing model complexity for low-power edge deployment,
- Expanding validation across uncontrolled, real-world environments,
- Incorporating explainable AI (XAI) techniques for better interpretability,
- Addressing cross-cultural variations in emotional expression,
- Integrating temporal modeling for continuous emotion tracking.

By advancing beyond traditional emotion classification pipelines, this framework lays the groundwork for a new generation of accessible, high-performance emotion recognition technologies in digital health, education, robotics, and smart environments.

Funding and Declarations

Research funding The Authors state that no funding is involved. Conflict of interest the authors state that there is no conflict of interest. Informed consent Not applicable.

Ethical approval The DEAP dataset's original ethical approvals are acknowledged. Acknowledgment We want to thank the Queen Mary University of London and the University of Trento for the DEAP dataset, which is a good source of information for developing new approaches in emotion recognition.

References

[1] G. Chan et al., "Multi-site photoplethysmography technology for blood pressure assessment: Challenges and recommendations," Journal of Clinical Medicine, vol. 8, no. 11. 2019. doi: 10.3390/jcm8111827.

[2] S. TIVATANSAKUL and M. OHKURA, "Emotion Recognition using ECG Signals with Local Pattern Description Methods," Int. J. Affect. Eng., vol. 15, no. 2, 2016, doi: 10.5057/ijae.ijae-d-15-00036.

[3] S. H. Ahmed, E. Nabil, and A. A. Badr, "Detection of visual positive sentiment using PCNN," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, 2019, doi: 10.14569/IJACSA.2019.0100134.

[4] T. Niu, S. Zhu, L. Pang, and A. Elsaddik, "Sentiment analysis on multi-view social data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016. doi: 10.1007/978-3-319-27674-8_2.

[5] C. Sitaula, Y. Xiang, A. Basnet, S. Aryal, and X. Lu, "Tag-based semantic features for scene image classification," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019. doi: 10.1007/978-3-030-36718-3_8.

[6] A. Hassan and N. Pinkwart, "On the Adaptability and Applicability of Multi-Touch User Interfaces Addressing Behavioral Interventions for Children with Autism," IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), vol. 37, no. 2. 2020. doi: 10.1080/02564602.2019.1590164.

[7] R. Andreasson, B. Alenljung, E. Billing, and R. Lowe, "Affective Touch in Human–Robot Interaction: Conveying Emotion to the Nao Robot," Int. J. Soc. Robot., vol. 10, no. 4, 2018, doi: 10.1007/s12369-017-0446-3.

[8] R. R. Singh, S. Conjeti, and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals," Biomed. Signal Process. Control, vol. 8, no. 6, 2013, doi: 10.1016/j.bspc.2013.06.014.

[9] H. P. Da Silva, A. P. Alves, A. Lourenço, A. Fred, I. Montalvão, and L. Alegre, "Towards the detection of deception in interactive multimedia environments," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

Bioinformatics), 2013. doi: 10.1007/978-3-642-39146-0_7.

[10] C. Jones and J. Sutherland, "Acoustic emotion recognition for affective computer gaming," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008. doi: 10.1007/978-3-540-85099-1_18.

[11] T. Porat and N. Tractinsky, "Affect as a mediator between web-store design and consumers' attitudes toward the store," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008. doi: 10.1007/978-3-540-85099-1_12.

[12] F. Li, L. Yang, H. Shi, and C. Liu, "Differences in photoplethysmography morphological features and feature time series between two opposite emotions: Happiness and sadness," Artery Res., vol. 18, 2017, doi: 10.1016/j.artres.2017.02.003.

[13] M. S. Islam, M. Shifat-E-Rabbi, A. M. A. Dobaie, and M. K. Hasan, "PREHEAT: Precision heart rate monitoring from intense motion artifact corrupted PPG signals using constrained RLS and wavelets," Biomed. Signal Process. Control, vol. 38, 2017, doi: 10.1016/j.bspc.2017.05.010.

[14] R. Firoozabadi, E. D. Helfenbein, and S. Babaeizadeh, "Efficient noise-tolerant estimation of heart rate variability using single-channel photoplethysmography," J. Electrocardiol., vol. 50, no. 6, 2017, doi: 10.1016/j.jelectrocard.2017.08.020.

[15] A. Sološenko, A. Petrėnas, V. Marozas, and L. Sörnmo, "Modeling of the photoplethysmogram during atrial fibrillation," Comput. Biol. Med., vol. 81, 2017, doi: 10.1016/j.compbiomed.2016.12.016.

[16] [A. ReşitKavsaoğlu, K. Polat, and M. Recep Bozkurt, "A novel feature ranking algorithm for biometric recognition with PPG signals," Comput. Biol. Med., vol. 49, no. 1, 2014, doi: 10.1016/j.compbiomed.2014.03.005.

[17] M. W. Park, C. J. Kim, M. Hwang, and E. C. Lee, "Individual emotion classification between happiness and sadness by analyzing photoplethysmography and skin temperature," in Proceedings - 2013 4th World Congress on Software Engineering, WCSE 2013, 2013. doi: 10.1109/WCSE.2013.34.

[18] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalized models for facial emotion recognition through transfer learning," Multimed. Tools Appl., vol.. 79, no. 47–48, 2020, doi: 10.1007/s11042-020-09405-4.

[19] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, and O. O. Oluwagbemi, "Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10)," Appl. Comput. Intell. Soft Comput., vol. 2023, 2023, doi: 10.1155/2023/2457898.

[20] H. A. Shehu, W. Browne, and H. Eisenbarth, "Emotion Categorization from Video-Frame Images Using a Novel Sequential Voting Technique," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020. doi: 10.1007/978-3-030-64559-5_49.

[21] F. Z. Salmam, A. Madani, and M. Kissi, "Emotion recognition from facial expression based on fiducial points detection and using neural network," Int. J. Electr. Comput. Eng., vol. 8, no. 1, 2018, doi: 10.11591/ijece. v8i1. pp52-59.

[22] T. DJARA, "Emotional state recognition using facial expression, voice and physiological signal," Int. Robot. Autom. J., vol. 4, no. 3, 2018, doi: 10.15406/iratj.2018.04.00115.

[23] A. Goshvarpour and A. Goshvarpour, "Evaluation of Novel Entropy-Based Complex Wavelet Sub-bands Measures of PPG in an Emotion Recognition System," J. Med. Biol Eng., vol. 40, no. 3, 2020, doi: 10.1007/s40846-020-00526-7.

[24] A. Goshvarpour and A. Goshvarpour, "Poincaré's section analysis for PPG-based automatic emotion recognition," Chaos, Solitons and Fractals, vol. 114, 2018, doi: 10.1016/j.chaos.2018.07.035.

[25] Y. K. Lee, O. W. Kwon, H. S. Shin, J. Jo, and Y. Lee, "Noise reduction of PPG signals using a particle filter for robust emotion recognition," in Digest of Technical Papers - IEEE International Conference on Consumer Electronics, 2011. doi: 10.1109/ICCE-Berlin.2011.6031807.

[26] S. Koelstra et al., "DEAP: A database for emotion analysis; Using physiological signals," IEEE Trans. Affect. Comput., vol. 3, no. 1, 2012, doi: 10.1109/T-AFFC.2011.15.

[27] M. Estrada and A. Stowers, "Amplification of Heart Rate in Multi-Subject Videos".

[28] Y. Ouzar, F. Bousefsaf, D. Djeldjli, and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2022-June, pp. 2459–2468, 2022, doi: 10.1109/CVPRW56347.2022.00275.

[29] Rajarajeswari, S., and Hassan Srinivas Ranjitha. "Face Recognition Attendance Framework using Quantum Edge Processing." 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE). IEEE, 2025.

[30] P. Kumar and X. Li, "Interpretable Multimodal Emotion Recognition using Facial Features and Physiological Signals," pp. 1–5, 2023, [Online]. Available: http://arxiv.org/abs/2306.02845

[31] K. Ali and C. E. Hughes, "A Unified Transformer-based Network for multimodal Emotion Recognition," vol. 14, no. 8, 2023, [Online]. Available: http://arxiv.org/abs/2308.14160

[32] J. Kwon, J. Ha, D. H. Kim, J. W. Choi, and L. Kim, "Emotion Recognition Using a Glasses-Type Wearable Device via Multi-Channel Facial Responses," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3121543.

[33] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in 30th British Machine Vision Conference 2019, BMVC 2019, 2020.

[34] Q. Fan and K. Li, "Non-contact remote estimation of cardiovascular parameters," Biomed. Signal Process. Control, vol. 40, 2018, doi: 10.1016/j.bspc.2017.09.022.