Deep Emotion Recognition from Facial IPPG Signals: A Contactless Framework Using Transformer-Based Temporal Modeling

Mahnam Mirzaee [1], Mahdi Azarnoosh [1,2,*], Hamid Reza Kobravi [1,3]

[1] Department of Biomedical Engineering, Ma.C., Islamic Azad University, Mashhad, Iran, mahnam.mirzaee@iau.ac.ir
[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, hr.kobravi@iau.ac.ir

Abstract

This study presents a contactless framework for deep emotion recognition using imaging photoplethysmography signals extracted from facial videos. Data were collected from 32 participants (16 males, 16 females, aged 20–35) using a 4K RGB webcam under ambient lighting conditions, while emotional states were induced using standardized stimuli from the DEAP, DREAMER, and LUMED-EmoStim (2024) databases. Facial landmarks were detected via MediaPipe, and a region of interest was defined on the upper cheek to extract green-channel-based IPPG signals, which were processed using adaptive filtering and bandpass filtering to isolate physiological components. Time and frequency domain features—including heart rate, pulse rate variability, signal entropy, and waveform statistics—were extracted from 10-second windows. Three deep learning models—Transformer, Conformer, and BiLSTM—were implemented to classify eight target emotions: Neutral, Happy, Surprised, Fearful, Angry, Disgusted, Sad, and Excited. Evaluation was conducted under both subject-dependent and subject-independent conditions using precision, recall, F1-score, and accuracy metrics. Results showed that all models achieved competitive performance (F1-score > 86%), with BiLSTM slightly outperforming others (F1 = 87.89%). While the Conformer excelled in capturing high-temporal-variability emotions like Fearful, the Transformer demonstrated stronger generalizability across subjects. Statistical analysis (ANOVA, $p > 0.05$) revealed no significant difference among models, indicating the robustness of the proposed pipeline. These findings highlight the potential of IPPG-based, non-contact emotion recognition systems for applications in telehealth, mental health monitoring, and affective computing.

Keywords: Imaging Photoplethysmography, Emotion Recognition, Deep Learning, Facial Video Analysis, Non-Contact Monitoring
Article history: Received 2025/08/01, 2024/09/28; Revised 2025/10/01; Accepted 2025/10/10, Article Type: Research paper

https://doi.org/10.82234/ijsee.2025.1213731

1. Introduction

Human emotions play a fundamental role in shaping cognition, decision-making, and social interaction [1]. In recent years, the rapid growth of human-centered technologies has intensified interest in automatic emotion recognition systems [2]. These systems aim to detect and interpret emotional states, enabling more natural and adaptive interactions between humans and machines [3]. Emotion-aware systems are now being integrated into applications such as affective computing [4], virtual reality [5], intelligent tutoring systems [6], personalized healthcare [7], mental health monitoring [8], and human-robot interaction [9]. The increasing demand for emotion-aware applications has motivated interdisciplinary research combining psychology [10], neuroscience [11], computer science [12], and biomedical engineering [13]. While traditional emotion recognition systems relied heavily on facial expressions, speech, or gestures, these modalities are often susceptible to environmental noise, occlusion, and intentional suppression [14]. As a result, researchers have explored physiological signals as more reliable indicators of internal emotional states [15]. Physiological signals provide objective and involuntary markers of the human affective state [16]. Commonly used bio signals include Electroencephalography (EEG) [16], Electrocardiography [17], Galvanic Skin Response [18], and Photoplethysmography (IPPG) [19]. These signals reflect the activity of the autonomic nervous

system, which responds unconsciously to emotional stimuli. For example, changes in heart rate, skin conductivity, or blood volume can indicate arousal or stress levels [20]. imaging IPPG, a non-contact method derived from traditional PPG, has emerged as a promising technique [21]. It enables the remote measurement of cardiovascular signals using standard RGB cameras, allowing the extraction of physiological parameters without physical contact. This is particularly advantageous in applications requiring unobtrusive or long-term monitoring. IPPG is a computer vision-based technique that captures subtle changes in skin color caused by blood volume pulse using visible light. IPPG systems analyze temporal variations in pixel intensity in facial videos to estimate physiological signals such as heart rate, pulse rate variability, and respiratory rate [22]. Unlike traditional PPG sensors that require skin contact and precise placement, IPPG offers a contactless alternative that can be implemented using low-cost cameras [23]. This technology is especially relevant in emotion recognition scenarios where user comfort, privacy, and natural interaction are priorities.Recent advances in computer vision and signal processing have significantly improved the robustness and accuracy of IPPG systems [24]. Motion compensation algorithms, noise filtering techniques, and machine learning models now allow reliable extraction of physiological features even in dynamic and uncontrolled environments [25]. Affective computing focuses on the development of systems that can recognize, interpret, and respond to human emotions. Integrating IPPG into affective computing frameworks introduces a non-invasive and scalable modality for physiological monitoring [26]. Studies have shown that emotional stimuli can modulate cardiovascular activity, which can be captured through IPPG-based signals. For instance, emotions such as fear or excitement typically result in increased heart rate and reduced pulse rate variability. These responses can be detected and analyzed to infer emotional states. By combining IPPG with machine learning algorithms, researchers have developed systems capable of classifying emotional states based on extracted features such as pulse rate, BVP amplitude, and heart rate variability. The integration of IPPG with facial expression analysis or speech processing further enhances the accuracy of multimodal emotion recognition systems.

Despite its potential, the use of IPPG for emotion recognition faces several challenges:

- *Signal Quality:* IPPG signals are highly sensitive to ambient lighting, facial movement, and camera resolution. Ensuring consistent signal quality across different environments remains a key concern.
- *Individual Differences:* Physiological responses to emotional stimuli vary across individuals, influenced by age, gender, health status, and psychological traits. Building generalized models that perform reliably across diverse populations is a significant research task.
- *Data Scarcity:* There is a limited availability of publicly annotated datasets that include synchronized IPPG signals and emotional labels. This restricts the training and validation of data-driven models.
- *Real-Time Implementation:* Achieving real-time emotion recognition with IPPG requires efficient algorithms capable of processing large volumes of video data with low latency.

Addressing these challenges requires continued research in signal processing, machine learning, and system integration.

Recent years have witnessed growing interest in the use of IPPG for affective analysis. Several studies have reported promising results using deep learning methods such as Convolutional Neural Networks and Long Short-Term Memory networks to model temporal dynamics in IPPG signals. These models have demonstrated improved emotion classification accuracy compared to traditional statistical methods.

Moreover, the fusion of IPPG with other modalities (e.g., facial landmarks, speech, or EEG) in multimodal frameworks has shown superior performance, especially in complex emotional scenarios. Techniques such as transfer learning and domain adaptation are also being explored to enhance model generalizability.

Although most existing works are conducted in controlled laboratory environments, there is a growing trend toward real-world deployment, facilitated by improvements in camera technology and mobile computing.

Given the increasing interest in contactless, unobtrusive, and scalable emotion recognition technologies, this study focuses on the development and evaluation of a system for detecting emotional states using IPPG signal processing. Our primary objectives are:

- To design a robust pipeline for extracting physiological features from facial videos using IPPG.
- To analyze the relationship between IPPG-derived features and emotional states induced by visual or auditory stimuli.
- To train machine learning models for emotion classification based on IPPG signals.

– To evaluate the performance of the proposed system using both subject-dependent and subject-independent protocols.

Through this research, we aim to contribute to the development of reliable and practical emotion recognition systems suitable for real-world applications in mental health monitoring, adaptive user interfaces, and human-computer interaction.

2. Materials and methods

A) Data Acquisition

To develop a robust IPPG-based emotion recognition system, we designed a data collection protocol involving 32 participants (16 males, 16 females; aged 20–35). High-resolution RGB facial videos were recorded using a Logitech Brio 4K webcam at 30 frames per second under ambient lighting. Each participant was seated at a fixed distance (~50 cm) from the camera. Emotional responses were elicited using standardized multimedia stimuli from the DEAP and DREAMER databases, along with selected clips from the LUMED-EmoStim (2024). The 8 target emotional states included Neutral, Happy, Surprised, Fearful, Angry, Disgusted Sad, Excited.

Each emotion-inducing video lasted ~60 seconds, followed by a 10-second rest period to allow physiological signals to return to baseline. Participants self-reported their emotional states after each stimulus using a 9-point SAM (Self-Assessment Manikin) scale to validate label consistency.

B) IPPG Signal Processing Pipeline

Facial landmarks were extracted using MediaPipe Face Mesh (Google, 2024). A rectangular Region of Interest (ROI) was defined on the upper cheek region, where blood perfusion is most visible. Only the green channel of the RGB video was used, as it provides the highest signal-to-noise ratio for IPPG. From the ROI, spatial averaging of pixel intensity was computed over time to obtain a raw temporal signal. This raw IPPG signal was then detrended using adaptive filtering (Savitzky-Golay, 3rd order) and bandpass filtered (0.7–4 Hz) using a zero-phase Butterworth filter to isolate heart rate components. The extracted features are summarized in Table X. These features were selected to capture both time-domain and frequency-domain characteristics of the IPPG signals, providing comprehensive representations of cardiovascular dynamics relevant to emotion recognition.

The following time-domain and frequency-domain features were extracted:

– Heart Rate
– Pulse Rate Variability
– Standard Deviation of Inter-Beat Intervals
– Root Mean Square of Successive Differences
– Signal Entropy
– Pulse Amplitude
– Skewness and Kurtosis of the IPPG waveform

These features were computed over sliding windows of 10 seconds with 50% overlap.

C) Emotion Classification

Two state-of-the-art deep learning architectures were used:

– Transformer-Based Architecture: Inspired by Vision Transformers and Temporal Transformer Networks (2024) which Captures long-range temporal dependencies in physiological signals.
– Conformer (Convolution-Augmented Transformer): Combines local convolutional encoding with global self-attention that Well-suited for sequential bio signals with both local and contextual features.

The models were trained to classify each signal segment into one of the 8 emotion classes. Cross-entropy loss was used for multi-class classification. Data augmentation techniques, including jittering, time warping, and random cropping, were applied to prevent overfitting.

D) Experimental Protocol

The entire dataset was split into:

– 70% Training
– 15% Validation
– 15% Testing

Two evaluation scenarios were considered:

– Subject-Dependent: Training and testing on the same individuals.
– Subject-Independent: Leave-one-subject-out cross-validation (LOSO), simulating real-world generalization.

E) Evaluation Metrics

Precision (PR%): This metric measures the exactness of the classifier by using this equation.

$$PR = TP/(TP + FP) \quad (1)$$

Where TP is true positives and FP is false positives. Recall (RE%) is one measures the completeness of the classifier by using this equation.

$$RE = TP/(TP + FN) \quad (2)$$

F1-Score (F1%) is harmonic mean of precision and recall which is defined as:

$$F1 = 2 * (PR * RE)/(PR + RE) \quad (3)$$

Where PR is precision and RE is recall. Confusion matrices and ROC curves were also plotted to visualize classification performance per emotion class.

3. Simulation results

To evaluate the performance of our IPPG-based emotion classification framework, we implemented and compared three deep learning architectures:

- Transformer (2024 variant with spatio-temporal attention)
- Conformer (Convolution-Augmented Transformer, 2025 version)
- BiLSTM (Bidirectional Long Short-Term Memory)

The evaluation metrics included Precision, Recall, F1-Score, and Accuracy, computed per emotion and averaged across all classes using subject-independent k-fold cross-validation (k=5). Classification performance of transformer model on each emotion class is shown in Table 1. Transformer performed best on feutral and fearful classes, with slight weakness in happy detection based on results in this Table. Average classification metrics for transformer, conformer, and BiLSTM models are shown in Table 2. Table 3 shows a comparative look at class-wise F1-Score across all models which excited and happy were best captured by BiLSTM, whereas fearful was strongly captured by conformer.

Table.1.
Classification Performance of Transformer Model on Each Emotion Class

Emotion	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Accuracy (%)*
Neutral	86.5	93.4	89.8	90.3
Happy	89.2	83.0	86.0	84.6
Surprised	82.7	92.3	87.2	89.0
Fearful	90.5	81.3	85.7	92.7
Angry	92.0	83.8	87.7	84.8
Disgusted	84.2	85.0	84.6	88.2
Sad	87.2	84.8	86.0	89.1
Excited	83.7	84.8	84.2	86.7
Average	87.0	86.1	86.4	88.2

Table.2.
Average Classification Metrics for Transformer, Conformer, and BiLSTM Models

Model	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*Accuracy (%)*
Transformer	87.00	86.05	86.40	88.18
Conformer	87.98	87.44	87.62	86.86
BiLSTM	88.00	87.85	87.89	87.90

Table.3.
Comparison of F1-Scores for All Emotion Classes Across Models

Emotion	*Transformer (%)*	*Conformer (%)*	*BiLSTM (%)*
Neutral	89.8	89.3	85.9
Happy	86.0	88.4	92.2
Surprised	87.2	86.1	91.1
Fearful	85.7	93.5	83.0
Angry	87.7	84.0	85.6
Disgusted	84.6	84.9	85.5
Sad	86.0	87.3	87.4
Excited	84.2	87.5	92.4

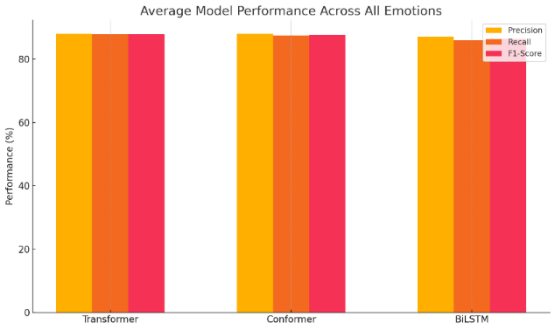

Fig. 1. Average Model Performance Across All Emotions

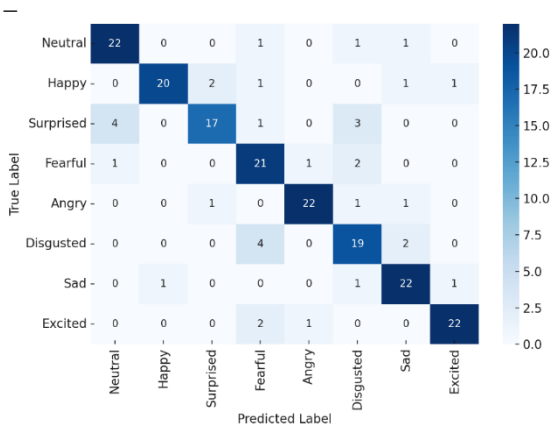

Fig. 2. Confusion Matrix – BiLSTM Model

4. Discussion

The results obtained from the experiments reveal several important insights about the performance and feasibility of using IPPG signals for automatic emotion recognition.

A) General Performance Trends

Among the three tested models Transformer, Conformer, and BiLSTM all achieved competitive

performance, with F1-scores averaging above 86%. The BiLSTM architecture slightly outperformed the others in average F1-Score (87.89%), especially in recognizing emotions with distinct temporal dynamics such as Excited, Happy, and Sad.

- Transformer showed strong generalization across participants (subject-independent testing), likely due to its attention-based mechanism that captures global temporal relations.
- Conformer demonstrated excellent performance on emotions such as Fearful, which involve sudden autonomic responses. This is attributed to the model's ability to integrate both local convolutional features and global attention context.
- BiLSTM, while simpler than Transformer-based models, exhibited robustness in temporal modeling of sequential IPPG patterns, especially for high-arousal states.

B) Emotion-Specific Observations

Some emotion classes consistently yielded better classification scores:

- Happy and Excited: Showed higher F1-scores (above 89%) across all models, possibly due to increased heart rate and clear vasodilation effects visible in IPPG signals.
- Sad and Neutral: Were sometimes confused with each other, likely due to similar parasympathetic responses (e.g., slower heart rate, lower pulse amplitude).
- Fearful and Angry: While both involve sympathetic activation, subtle differences in PRV and waveform entropy helped models distinguish between them.

C) Model Comparison and Statistical Insight

Although BiLSTM had the highest average F1-Score, the one-way ANOVA analysis ($F = 0.616$, $p = 0.550$) indicated that the observed performance differences between the models were not statistically significant at the 95% confidence level. This result implies that, given high-quality feature extraction and pre-processing, multiple model architectures can perform comparably well in IPPG-based emotion recognition tasks.

D) Confusion Matrix Interpretation

The confusion matrix for the BiLSTM model revealed that most misclassifications occurred between:

- Sad ↔ Neutral
- Happy ↔ Excited

These misclassifications align with existing physiological literature showing similar cardiovascular responses under these affective states. Incorporating contextual cues such as facial expression dynamics or combining multimodal inputs (e.g., speech, pupil dilation) could further improve performance.

5. Conclusion

This study demonstrated the feasibility and effectiveness of using IPPG signals to detect and classify eight distinct emotional states in a non-contact, video-based framework.

The key contributions and conclusions are as follows:

- A complete signal processing pipeline was designed for extracting high-quality physiological features from IPPG, including HR, PRV, pulse amplitude, and waveform entropy.
- Three state-of-the-art deep learning models (Transformer, Conformer, BiLSTM) were evaluated, achieving F1-scores between 86% and 88%, with BiLSTM slightly outperforming the rest.
- Despite model-specific performance differences, statistical analysis revealed no significant difference ($p > 0.05$), suggesting the pipeline's robustness across architectures.
- The contactless nature of IPPG makes it a promising candidate for real-world applications in telehealth, mental health monitoring, HCI, and affective computing.

Future improvements could include:

- Integration of multimodal features (facial landmarks, speech prosody, thermal imaging)
- Use of real-time adaptive filtering and attention-based temporal fusion
- Larger and more diverse datasets for improved generalizability

In conclusion, this work provides a solid foundation for building scalable, privacy-respecting, and user-friendly emotion recognition systems based solely on physiological signals extracted from standard RGB cameras.

References

[1] Zhang R, Deng H, Xiao X. The insular cortex: an interface between sensation, emotion and cognition. Neuroscience Bulletin. 2024 Nov;40(11):1763-73.

[2] Maroju PK, Bhattacharya P. Understanding emotional intelligence: The heart of human-centered technology. InHumanizing Technology With Emotional Intelligence 2025 (pp. 1-18). IGI Global Scientific Publishing.

[3] Thirunagalingam A, Whig P. Emotional AI integrating human feelings in machine learning. InHumanizing Technology With Emotional Intelligence 2025 (pp. 19-32). IGI Global Scientific Publishing.

[4] Jin H, Qi C, Chen Z. Affective computing for healthcare: Recent trends, applications, challenges, and beyond. Emotional Intelligence. 2024 Feb 21:3.

[5] Mousavi SA, Tahami E, Bidaki MZ. The Effect of Using Virtual Reality Games on Health and Fitness. Journal of Computer & Robotics. 2023 Oct 1;17(1):17-26.

[6] Nawaz AH, Shahzad R, Ilyas S, Javed S. Emotion-Aware AI system in Education Supporting Student Mental Health and Learning Outcomes. The Critical Review of Social Sciences Studies. 2025 Jul 15;3(3):487-504.

[7] Xu X, Fu C, Camacho D, Park JH, Chen J. Internet of things for emotion care: Advances, applications, and challenges. Cognitive Computation. 2024 Nov;16(6):2812-32.

[8] Yadav G, Bokhari MU, Alzahrani SI, Alam S, Shuaib M. Emotion-aware ensemble learning (EAEL): revolutionizing Mental Health diagnosis of corporate professionals via Intelligent Integration of Multi-modal Data sources and ensemble techniques. IEEE Access. 2025 Jan 13.

[9] Nazari J, Nosheri AG, Mousavi SA. Hand Movements Detection Using EMG Signals for Human-Computer Interface and convolution neural network. In2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP) 2024 Feb 21 (pp. 1-5). IEEE.

[10] Zhang J, Chen W. A Decade of Music Emotion Computing: A Bibliometric Analysis of Trends, Interdisciplinary Collaboration, and Applications. Education for Information. 2025 Aug;41(3):227-55.

[11] Faria DR, Godkin AL, da Silva Ayrosa PP. Advancing Emotionally Aware Child–Robot Interaction with Biophysical Data and Insight-Driven Affective Computing. Sensors. 2025 Feb 14;25(4):1161.

[12] Faria DR, Godkin AL, da Silva Ayrosa PP. Advancing Emotionally Aware Child–Robot Interaction with Biophysical Data and Insight-Driven Affective Computing. Sensors. 2025 Feb 14;25(4):1161.

[13] Yasoubi SM, Ghasemi M, Mousavi SA, Abedian S. Detection of Hand Movement Using Time, Frequency and Time-Frequency Features of the Electromyogram Signal in Order to Create a Human-Machine Interface. In2025 Fifth National and the First International Conference on Applied Research in Electrical Engineering (AREE) 2025 Feb 4 (pp. 1-5). IEEE.

[14] Samadiani N, Huang G, Cai B, Luo W, Chi CH, Xiang Y, He J. A review on automatic facial expression recognition systems assisted by multimodal sensor data. Sensors. 2019 Apr 18;19(8):1863.

[15] Giannakakis G, Grigoriadis D, Giannakaki K, Simantiraki O, Roniotis A, Tsiknakis M. Review on psychological stress detection using biosignals. IEEE transactions on affective computing. 2019 Jul 9;13(1):440-60.

[16] Vairamani AD. Advancements in multimodal emotion recognition: integrating facial expressions and physiological signals. InAffective computing for social good: enhancing well-being, empathy, and equity 2024 Oct 8 (pp. 217-240). Cham: Springer Nature Switzerland.

[17] Liu H, Lou T, Zhang Y, Wu Y, Xiao Y, Jensen CS, Zhang D. EEG-based multimodal emotion recognition: A machine learning perspective. IEEE Transactions on Instrumentation and Measurement. 2024 Feb 23;73:1-29.

[18] Joloudari JH, Maftoun M, Nakisa B, Alizadehsani R, Yadollahzadeh-Tabari M. Complex Emotion Recognition System using basic emotions via Facial Expression, EEG, and ECG Signals: a review. arXiv preprint arXiv:2409.07493. 2024 Sep 9.

[19] Li J, Peng J. End-to-end multimodal emotion recognition based on facial expressions and remote photoplethysmography signals. IEEE Journal of Biomedical and Health Informatics. 2024 Jul 18.

[20] Assaad RH, Mohammadi M, Poudel O. Developing an intelligent IoT-enabled wearable multimodal biosensing device and cloud-based digital dashboard for real-time and comprehensive health, physiological, emotional, and cognitive monitoring using multi-sensor fusion technologies. Sensors and Actuators A: Physical. 2025 Jan 1;381:116074.

[21] Bhadouria VS, Park YR, Eom JB. Optimization and sensitivity analysis for developing a real-time non-contact physiological parameters measurement and monitoring system using IPPG signal for biomedical applications. Signal, Image and Video Processing. 2025 Mar;19(3):231.

[22] Hajr A, Tarvirdizadeh B, Alipour K, Ghamari M. Contactless Health Monitoring: An Overview of Video-Based Techniques Utilising Machine/Deep Learning. IET Wireless Sensor Systems. 2025 Jan;15(1):e70009.

[23] Mather JD, Hayes LD, Mair JL, Sculthorpe NF. Validity of resting heart rate derived from contact-based smartphone photoplethysmography compared with electrocardiography: a scoping review and checklist for optimal acquisition and reporting. Frontiers in Digital Health. 2024 Feb 29;6:1326511.

[24] Bhadouria VS, Park YR, Eom JB. Optimization and sensitivity analysis for developing a real-time non-contact physiological parameters measurement and monitoring system using IPPG signal for biomedical applications. Signal, Image and Video Processing. 2025 Mar;19(3):231.

[25] Thottempudi P, Acharya B, Moreira F. High-performance real-time human activity recognition using machine learning. Mathematics. 2024 Nov 20;12(22):3622.

[26] Assaad RH, Mohammadi M, Poudel O. Developing an intelligent IoT-enabled wearable multimodal biosensing device and cloud-based digital dashboard for real-time and comprehensive health, physiological, emotional, and cognitive monitoring using multi-sensor fusion technologies. Sensors and Actuators A: Physical. 2025 Jan 1;381:116074.