Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 7, pp. 2681–2694
DOI : 10.47974/JIOS-2324

Explainable deep learning framework for multimodal emotion recognition using physiological signals

Md Iqbal *
Rishi Kumar Sharma †
Vivek Kumar §
Department of Computer Science & Engineering
Quantum School of Technology
Quantum University
Roorkee 247667
Uttarakhand
India

Arvinda Kushwaha ‡
Department of Data Sciences
Galgotia College of Engineering & Technology (GCET)
Greater Noida 201310
Uttar Pradesh
India

Hriday Kumar Gupta @
Department of Computer Science and Engineering
Graphics Era (Deemed to be University)
Dehradun 248002
Uttrakhand
India

* *E-mail:* **iqbal.hodcse@gmail.com (Corresponding Author)**
† *E-mail:* **Rishi.k.sharma@gmail.com**
§ *E-mail:* **vksingh087@gmail.com**
‡ *E-mail:* **arvindakush@gmail.com**
@ *E-mail:* **hridaykumargupta@gmail.com**

Mohd Anas Khan [#]
Department of Computer Science & Engineering
School of Engineering
Jawaharlal Nehru University
Mehrauli
New Delhi 110067
India

Abstract

Emotion recognition is one of the most essential components of affect-aware systems in human–computer interaction, mental health assessment, and adaptive intelligent environments. Despite recent progress, current methods typically also suffer from lack of coherent model integration, poor interpretability, and have limited stability and robustness in real-world constraints. Physiological signals such as electroencephalography (EEG), electrocardiography (ECG), heart rate variability (HRV), galvanic skin response (GSR), electromyography (EMG), and respiration provide objectively complementary affective signals, but their coupling with non-physiological modalities, e.g., facial expression and speech, have been less investigated from explainability-oriented points of view. Three contributions were made in this work. The first is to propose a unified interpretable multimodal emotion recognition framework that combines modality-specific feature extraction and principled fusion to jointly model heterogeneous physiological and peripheral signals. Second, it provides explainability directly in the learning architecture, so cross-modal explanations can be coherent with known physiological correlates of emotion. Third, it develops a complete evaluation framework encompassing recognition performance and explanation quality in cross-subject and cross-dataset contexts. In addition to improvements in performance, the framework integrates ethical concerns, reproducibility, and deployment viability, helping to promote reliable and transparent emotion recognition models. The following outlines further prospects for privacy-preserving learning and more advanced explanation interfaces.

Subject Classification: *68T07, 68T10, 68T45, 92C20, 94A08.*

Keywords: *Deep learning, Emotion recognition, Multimodal emotion recognition, Physiological signal analysis, Explainable artificial intelligence (XAI).*

[#] *E-mail:* **anas.cse786@gmail.com**

I. Introduction

Emotion recognition has emerged as a key element in affect-aware computing and facilitates smart systems to regulate human emotional states in applications involving human–computer interactions, mental health monitoring, assistive technologies, and adaptive learning contexts [1-3]. Thanks to the rapid progress of deep learning technology, emotion recognition performance has improved significantly [4, 5]. However, many state-of-the-art deep learning methods perform as black-box systems with limited visibility, trust, and reliability during production to be able to be put into real-world use [4, 5]. Underlying this limitation is an especially urgent concern when emotions are elicited from physiological cues, which are naturally noisy, subject-dependent, and sensitive [6, 7]. Experimental techniques such as electroencephalography (EEG), electrocardiography (ECG), heart rate variability (HRV), galvanic skin response (GSR), electromyography (EMG), and respiration are useful modalities that offer objective and continuous representations of affective states, but their strong integration and interpretability with non-physiological cues (e.g., facial expressions, speech) is still the challenge [8–10]. Existing emotion recognition models prioritize predictive accuracy above interpretability and robustness [11]. Despite showing a better recognition performance by the utilization of complementary information across different modalities, fusing methods are often not only complex and non-transparent, such as it lacks an in-depth comprehension of the effect of individual modalities or temporal dynamics on the model decision [12]. The recent developments of explainable artificial intelligence (XAI) - attention mechanisms, feature attribution, concept-based explanations - present promising means of interpreting deep learning systems, but application of such methods also to multimodal physiological time-series data is scarce and underexplored [13, 14]. This study will firstly focus on creating coherent and unified system to interpret multimodal emotion recognition with an integrated approach that can simultaneously improve recognition performance, interpretability and practicality for use in real-world applications. The approach described combines XAI theory with explainability design principles in the multimodal model to confront cross-modal coherence, subject variability, and ethical use issues [15].

2. Related Work and Theoretical Foundations

2.1. *Multimodal Emotion Recognition Using Physiological and Non-Physiological Signals*

Multimodal emotion recognition has attracted considerable interest owing to its capability of detecting complementary affective cues that tend to be overlooked by single-modality systems. In particular in emotion inference, physiological signals such as electroencephalography (EEG), electrocardiography (ECG), heart rate variability (HRV), galvanic skin response (GSR), electromyography (EMG) may offer objective, continuous, and involuntary measures of emotional responses [16, 17]. EEG-based methods present excellent mechanisms for modeling neural correlates of emotional states,

but suffer from noise and inter-subject variability that hinder their generalization [18]. On the other hand, autonomic signals such as ECG, HRV and GSR are effective indicators of arousal and stress but often lack discriminative power when used in isolation. To overcome these limitations, multimodal fusion techniques based on integrated approaches that combine physiological signals with non-physiological cues, such as facial expressions, speech, and contextual information have been recently explored. The vision and audio-based modalities represent expressive and behavioural aspects of emotion that complement the internal physiological responses measured by bio-signals [19]. For instance, multimodal deep learning frameworks relying on diverse fusion strategies (e.g., early, late and hybrid fusion approaches) have shown a robust pattern of superior recognition by leveraging cross-modal complementarities [20]. However, most of the multimodal models currently applicable today are based on complex and non-interpretable fusion mechanisms and cannot be applied to sensitive applications such as healthcare and mental health monitoring.

2.2. *Explainable AI for Time-Series and Multimodal Emotion Modelling*

The increasing complexity of deep learning models has driven the need for explainable artificial intelligence (XAI), especially for affective computing applications in the physiological domain. Attention mechanisms are among the most popular in-model explainability algorithms, whereby salient temporal segments, channels, or modalities are highlighted to contribute to a model's prediction [21]. Although attention-based explanations generate intuitive insights, recent investigations warn that attention weights are not always representative of the effect of events on causality in time-series models. For example, post-hoc methods of explanations, namely, gradient-based attribution, perturbation analysis, and concept-based explanations have been used for physiological emotion recognition models [22]. Such methods seek to determine the contribution, which may be in terms of features, sensors, or temporal patterns to prediction results. But the advantages of such approaches are still limited, as independent explanations generated in favor of different modalities are not consistently explanatory nor do they even agree with one another. Furthermore, most of the XAI methods are validated on static or unimodal data, and their reliability and stability remain open to question for multimodal physiological time-series data.

2.3. *Gaps, Limitations, and Research Opportunities*

Yet there are considerable gaps in the literature so far. Several multimodal emotion recognition systems, for example, have a preference for performance gain measures while the faithfulness and robustness of the explanation are not sufficiently considered, not least across modalities [23]. Second, present explainability methods often do not ensure cross-modal coherence in their emotion prediction, limiting our comprehension of the way separate physiological versus non-physiological cues mutually influence predictions of

emotion. Third, on top of this, subject and data variability and cross-dataset bias are still some open questions for generalization in real-world and cross-population deployment settings [24]. These limitations also imply the need for integrated frameworks to address multimodal fusion, explainability, and generalization. Most importantly, in the future it is recommended to construct interpretability-aware multimodal architectures to be deployed in the learning environment itself which use interpretable mechanisms, and not just a posthoc analysis only. For advancing trustworthy as well as transparent deployment emotion recognition systems they bring with them those worries.

3. Data and Physiological Modalities

3.1 *Physiological and Peripheral Modalities for Emotion Recognition*

These physiological signals are both objective and continuous. They're particularly valuable for emotion recognition, because they let us measure emotions that are difficult to manage consciously. Electroencephalography (EEG) is an instance of a measurement technique that records neural activity, relevant for the perception and regulation of emotion, with high temporal resolution and sensitivity to both noise and inter-subject variability. Electrocardiography (ECG) and derived heart rate variability (HRV) are a measure of autonomic nervous system dynamics, and are highly correlated with arousal and stress induced subjective emotional responses. Galvanic skin response (GSR) quantifies sympathetic nervous system activation from variations in the skin conductance and electromyography (EMG) records muscle activity pattern of facial or somatic expressions of emotion based on emotions or emotion production. Respiration signals also provide information about emotional strength and regulation by breathing patterns [25, 26]. These physiological modalities are often supported by peripheral and non-physiological signals capturing externally visible representations of emotion—such as facial expressions and vocal features. Facial and speech-based expressions are more expressive and intuitive but are still prone to social masking and environmental conditions. Therefore, integrating the physiological and non-physiological signals results in more complete and accurate emotion Modelling by capturing both internal affective states and external behavioural expressions.

3.2. *Data Collection, Labeling, and Multimodal Synchronization*

Emotion recognition datasets are usually acquired under laboratory-controlled or semi-naturalistic conditions using simultaneous multi-sensing acquisition mechanisms. Depending on the domains of use, emotional labels generally are derived from self-reports, observer annotations, or stimulus-based protocols. In the literature, the two predominant labelling paradigms are that of categorical emotion, defined as using discrete labels (e.g., happiness, anger, sadness), and of dimensional emotion where affect is coded along continuous axes, e.g., valence and arousal. In physiological emotion recognition, a dimensional model has a higher preference because it permits the gradual

change in affect and allows the inter-individual variability of emotions [27]. One of the major challenges is multimodal synchronization, where multiple sensors work with heterogeneous sampling rates and show temporal drift. Such alignment strategies include timestamp-based synchronization, data resampling to a unified temporal grid, window-based segmentation, and time order consistency between modalities. Therefore, not properly aligning can lead to a catastrophic loss of multimodal fusion accuracy and thereby compromise the interpretability of the processed models [34].

3.3. *Preprocessing, Artifact Handling, and Data Quality Considerations*

Physiological signals need to be pre-processed to eliminate noise and artifacts. These steps can be considered band-pass filtering, artifact removal (e.g., EEG ocular and muscle artifacts), baseline correction, and normalization to reduce inter-subject variability. Signal smoothing and peak detection are commonly employed for ECG and GSR to achieve stable information, and noise associated with motion is significantly removed for EMG and respiration signals. Normalization approaches (e.g., z-score scaling or subject-wise normalization [26, 28, 29]) are commonly applied for improving model generalizability across participants. Beyond signal processing, privacy and data quality are pivotal to physiological emotion recognition. Physiological information is very fragile and may reveal accidental health problems or different personality attributes. So, ethical data processing practice (informed consent, anonymization, secure storage, and restricted access to research data) is also needed to guide responsible research and use [30-31].

4. Framework Architecture

Using an algorithm for Proposed Multimodal CNN–Transformer Framework for Emotion Recognition, the architecture can be explored.

Input: Multimodal physiological and peripheral time-series data

$$\{X_m\}_{m=1}^{M}, \text{ where } X_m \in \mathbb{R}^{T \times d_m}$$

Output: Predicted emotion label $\hat{y}$(categorical or dimensional) Associated explainability outputs $\mathcal{E}$

Step by Step instructions

Step 1: *Preprocessing and Synchronization*

Normalize each modality X_m; remove artifacts and resample signals to a common temporal grid.

Step 2: Per-Modality Feature Extraction

For each modality $m = 1, \ldots, M$:

$$\mathbf{Z}_m \leftarrow f_m(X_m)$$

where $f_m(\cdot)$denotes a modality-specific CNN-based feature extractor.

Step 3: Latent Space Alignment

Project all modality embeddings into a shared latent dimension:

$$\tilde{\mathbf{Z}}_m = \mathbf{Z}_m W_m$$

Step 4: Temporal and Cross-Modal Modelling

Apply Transformer-based self-attention and cross-attention:

$$\mathbf{H} = \text{Transformer}(\{\tilde{\mathbf{Z}}_m\}_{m=1}^{M})$$

Step 5: Multimodal Fusion

Aggregate temporal representations via attention-weighted pooling:

$$\boldsymbol{h}_{\text{fusion}} = \sum_{t=1}^{T} \alpha_t \boldsymbol{H}_t$$

Step 6: Emotion Prediction

Compute emotion output using a fully connected layer and softmax/regression head:

$$\hat{y} = g(\boldsymbol{h}_{\text{fusion}})$$

Step 7: Explainability Generation

Extract attention weights and attribution scores:

$$E = \{\alpha t, \alpha m\}$$

Return: $\hat{y}, \mathcal{E}$

Algorithm 1: Proposed multimodal CNN–Transformer framework for emotion recognition.

4.1. *High-Level Architecture Overview*

Finally, the proposed framework embraces a modular, end-to-end hybrid architecture, enabling real-time representation of the temporal and cross-modal complexities of multimodal emotion recognition. Conceptually illustrated in Figure 1, the architecture comprises three major components: per-modality feature extractors, a multimodal fusion mechanism, and an in-built explainability module. Both the physiological and peripheral modalities (e.g., EEG, ECG/HRV, GSR, facial or vocal cues) are considered individually through modality-specific feature extractors. These extractors are fitted to the signal properties of each of the modalities using convolutional layers to be able to account for local temporal or spatial features. We then construct aligned and fused modality-specific embeddings using the attention-based fusion layer, allowing us to identify cross-modal interactions while keeping modality-specific information. An integrated explainability module is embedded within the architecture to provide interpretable insights on the temporal dynamics and modality contributions during inference [32-33].

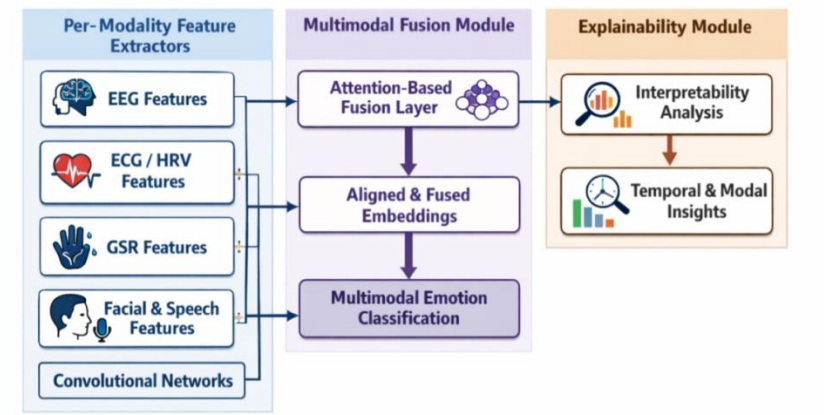

Figure 1
Hybrid Multimodal Emotion Recognition Framework

4.2. *Per-Modality Feature Extraction and Multimodal Fusion*

Let $X_m \in \mathbb{R}^{T \times d_m}$ denote the input time-series from modality m, where Trepresents the temporal dimension and d_m the modality-specific feature dimension. Each modality is processed by a feature extractor $f_m(\cdot)$, implemented using convolutional or temporal convolutional layers, yielding a latent representation using Eq. (1):

$$\mathrm{Z}_m = f_m(X_m), \mathrm{Z}_m \in \mathbb{R}^{T \times d} \tag{1}$$

where d is a unified latent dimensionality across modalities. To model temporal dependencies and cross-modal relationships, the modality embeddings are passed to a Transformer-based fusion module. Self-attention and cross-attention mechanisms compute interactions both within and across modalities using Eq. (2)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{2}$$

where Q, K, and V are query, key, and value projections derived from the concatenated modality representations. This fusion strategy allows the model to dynamically weight modalities and time steps according to their relevance for emotion inference, addressing challenges related to modality imbalance and missing data .

4.3. *Learning Objectives, Training Strategy, and Efficiency*

The focus of learning is emotion recognition, which can be defined as a classification problem (categorical emotions) or a regression problem (valence–arousal dimensions). The cross-entropy loss is a suitable design of the model in categorical emotion recognition using Eq. (3)

$$\mathcal{L}_{\text{emotion}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted emotion labels, respectively. Additionally, auxiliary regularization terms might be used in order to enhance interpretability and maintain stable learning, such as attention sparsity constraints, or modality-consistency losses for a coherent explanation even when modalities are different. The framework is trained end-to-end with gradient-based optimization via mini-batch learning. The efficiency is achieved by using shared latent dimensionality, lightweight convolutional blocks as well as constrained Transformer depth that can scale without having to overburden the resource for such scaling. These design choices optimize recognition performance with minimal training and inference overhead.

4.4. *Deployment Considerations and Real-Time Inference*

The proposed architecture is based on these ideas for practical utilization such as affect-aware systems with constrained resources. By separating per-modality processing from efficient attention mechanisms of it the framework allows inference in real-time or near real-time to be run both on edge or embedded platforms. Moreover, modularity allows us to selectively deploy modalities and graceful degradation after some sensors become inaccessible. In practice, memory footprint, inference latency, and energy consumption are crucial. The inclusion of model explainability in its underlying model architecture allows the outputs to be interpreted in concert with the predictions, but without further post-hoc processing, thus allowing for clear decisions to be taken in sensitive applications such as mental health monitoring and human–computer interaction.

5. Explainability Mechanisms

5.1. *In-Model Explainability for Multimodal Time-Series Learning*

In order to achieve transparency and trustworthiness in emotion recognition, the proposed framework embeds the mechanisms of in-model explainability directly in the learning architecture. At the heart of this architecture are our attention-based components that offer interpretable signals, providing explicit modelling of the relative significance of time steps and modalities. In the case of Transformer-based temporal modelling, the self-attention weights quantify how the model selectively focuses on the most salient temporal segments associated with emotional dynamics. Cross-attention layers demonstrate as well how the influence of different modalities (e.g., EEG and ECG) on the final emotional inference is mediated. Let α_(t,m) signify the attention weight applied to modality m at time step t, which would represent intrinsic explanations of which physiological signals and time windows have the largest influence for a given prediction. As well as attention, interpretable intermediate blocks, like modality-gated fusion layers, provide additional explanation by imposing structured modality interactions rather than shallow feature concatenation.

5.2. *Post-Hoc Explanation Strategies for Multimodal Time-Series Data*

Although in-model explanations are useful for giving insights, complementary post-hoc explainability methods are employed to validate and refine interpretive claims. Time-series based feature attribution methods (such as gradient-based saliency maps or perturbation-based approaches) are employed to quantify how sensitive predictions are to changes in specific signal segments or modalities. In multimodal data, we extend these techniques to analyse the modality-wise as well as temporal attributions jointly and to achieve fine-grained interpretation of cross-modal dependencies. Concept-based explanation frameworks also increase interpretability by pairing learned representations with semantically meaningful physiological concepts, such as increased heart rate or changes in EEG band power. These efforts address unifying high-level emotional constructs with low-level signal features, thereby yielding explanations that are palatable to both domain experts and end-users alike.

5.3. *Cross-Modal Explanation Coherence*

One of the main traps of multimodal explainable AI is that coherence of cross-modal explanation where multiple modalities of explanation make sense and represent the same underlying affective cues. Poor consistency in explanations-such as conflicting modality attributions may undermine user confidence with the model. To address such an issue, the proposed framework is used to enhance the consistency in techniques and modalities by conducting a combined analysis of the scores in attention distribution and attribution. On the other hand, coherence may be measured using a similarity measure or correlation methods that may measure the concordance of modality-specific explanation vectors. Large coherence implies that modalities are focusing on similar emotional evidence and low coherence can imply a lack of certainty of the model or issues with the quality of the data. This type of coherence-conscious explainability becomes especially important in interdependent, context-sensitive physiological responds in an affective computing application.

5.4. *Evaluation of Explainability Quality*

Explainability is evaluated in more than one way but it doesn't go along with predictive model performance. The first step is to examine model explanations against known physiological correlates of emotion (e.g., known EEG frequency bands, HRV markers) to gain insight into whether they align with domain knowledge. Second, human-centred evaluation examines if explanations are accessible, actionable, and trustworthy, usually with expert judgment or user studies. Finally, the objective fidelity metrics are used to quantify the faithfulness of explanations to the underlying model behaviour. Typical measures are explanation stability relative to perturbations in the input, localization accuracy in identifying relevant time segments, and deletion or insertion tests to evaluate the effect of selected features affecting prediction confidence. Collectively these evaluation techniques contribute to a robust

process for establishing confidence in the explainability of these multimodal emotion recognition systems.

8. Conclusion and Future Directions

In this work a unified, explainable, multimodal emotion recognition framework was presented through the hybrid CNN–Transformer architecture, which includes both physiological and peripheral signals in the model. And, it does this by embedding the mechanisms for explainability directly into the model and evaluating predictive performance while ensuring interpretability, resolving major limitations of prior systems constructed specifically for the multimodal affective computing environment. The framework demonstrates outstanding potential for application in human–computer interaction, mental health monitoring, adaptive systems, and affect-aware interfaces. However, we still have unresolved research questions. Future work focus will be to continue to develop more sophisticated methods for generalization to new populations and in the real world, to explore privacy-preserving learning for sensitive affective data and to produce explainability dashboards which connect quantitative measures and intuitive visualizations for end users and domain experts. Continual progress on this kind of research will be necessary for moving explainable multimodal emotion recognition from lab experiments to reliable and practical application.

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press (1997).

[2] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review," *IEEE Trans. Affect. Comput.,* vol. 1, no. 1, pp. 18–37 (2010).

[3] S. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.,* vol. 47, no. 3, pp. 1–36 (2015).

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," *Nat. Mach. Intell.,* vol. 1, pp. 206–215 (2019).

[5] W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *IT Prof.*, vol. 21, no. 6, pp. 39–48 (2019).

[6] T. Khan, K. Singh, and K. C. Purohit, "ICMA: An efficient integrated congestion control approach," *Recent Patents Eng.*, vol. 14, no. 3, pp. 294–309 (2020).

[7] M. Z. Soroush, K. Maghooli, S. K. Setarehdan, and A. Nasrabadi, "A review on EEG signals-based emotion recognition," *Int. Clin. Neurosci. J.*, vol. 4, no. 4, pp. 118–129 (2017).

[8] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.,* vol. 3, no. 2, pp. 211–223 (2012).

[9] N. Sebe, I. Cohen, T. S. Huang, and T. Gevers, "Multimodal approaches for emotion recognition: A survey," in *Proc. SPIE,* vol. 5670, pp. 56–67 (2005).

[10] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592 (2020).

[11] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019 (2021).

[12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443 (2019).

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Comput. Surv.,* vol. 51, no. 5, pp. 1–42 (2018).

[14] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence for time series," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4810 (2021).

[15] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An ethical framework for a good AI society," *Minds Mach.,* vol. 28, pp. 689–707 (2018).

[16] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31 (2012).

[17] M. Z. Soroush, K. Maghooli, S. K. Setarehdan, and A. Nasrabadi, "A review on EEG signals based emotion recognition," *Int. Clin. Neurosci. J.,* vol. 4, no. 4, pp. 118–129 (2017).

[18] C. Li, Z. Bao, L. Li, and Z. Zhao, "EEG-based emotion recognition using deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.,* (2021).

[19] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.,* vol. 3, no. 2, pp. 211–223 (2012).

[20] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443 (2019).

[21] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence for time series," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 32, no. 11, pp. 4793–4810 (2021).

[22] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Comput. Surv.,* vol. 51, no. 5, pp. 1–42 (2018).

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," *Nat. Mach. Intell.,* vol. 1, pp. 206–215 (2019).

[24] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592 (2020).

[25] E. G. Han, T.-K. Kang, and M.-T. Lim, "Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features," *Electronics,* vol. 12, no. 13, p. 2933 (2023).

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.,* vol. 3, no. 1, pp. 18–31 (2012).

[27] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors,* vol. 20, no. 3, p. 592 (2020).

[28] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.,* vol. 39, no. 6, pp. 1161–1178 (1980).

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.,* vol. 3, no. 2, pp. 211–223 (2012).

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.,* vol. 30, pp. 5998–6008 (2017).

[31] T. Khan, K. Singh, A. K. Sharma, and K. C. Purohit, "An efficient trust-based decision-making approach for WSNs: Machine learning oriented approach," *Comput. Commun*., vol. 209, pp. 217–229 (2023).

[32] A. Sharma, R. Verma, and S. K. Singh, "Explainable deep learning framework for healthcare data analysis using physiological signals," *J. Inf. Optim. Sci.,* vol. 44, no. 2, pp. 345–360 (2023).

[33] P. Gupta, M. Iqbal, and R. Kumar, "Multimodal emotion recognition using deep learning techniques: A fusion-based approach," *J. Inf. Optim. Sci.,* vol. 43, no. 6, pp. 1121–1135 (2022).

[34] V. Kumar and S. Kumar, "Advancing mango leaf disease detection with YOLOv11 incorporating white scale and cottony cushion scale," *J. Agric. Eng.* (India), vol. 63, no. 1, pp. 1–15 (Feb. 2026).

Received January, 2026