International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Entropy-Guided, Optimization-Tuned Fusion for Multimodal Emotion Recognition in Stress and Depression Detection: A Gap Analysis and Proposed Framework

Swati D Bhutekar[1], Dr. Yogini Borole[2], Swati S Chandurkar[3]

[1]PhD Research Scholar, School of Science & Technology, GLocal University, UP, India
[2]Professor, School of Science & Technology, GLocal University, UP, India
[3]Pimpri Chinchwad College of Engineering, Computer Engineering, Pune, India

Abstract
Multimodal emotion recognition (MER) has become a more powerful solution than unimodal approaches for stress and depression screening or other tasks, but the literature presents several common and recurring shortcomings that hinder accuracy and application. This paper provides a systematic gap analysis of the MER literature in the unimodal (facial, speech, text, EEG and physiological) and multimodal fusion research and a detailed description of a proposed AI based deep-learning framework to tackle the identified gaps directly. The results of the gap analysis, which is carried out on sixty-six publications reviewed, result in the following six identified deficiencies: (1) limited multimodal integration by using video, EEG or physiological signals in the same framework; (2) the use of fixed or heuristically chosen fusion weights, without optimization; (3) no multimodal fusion rules mentioned in the reviewed literature based on entropy; (4) limited validation on Indian-context data; (5) scarcity of the use of metaheuristic optimization to optimize multimodal fusion; and (6) limited explicit framing of multimodal emotion recognition in the context of stress and depression detection. In the proposed method, the gaps are addressed by five components: unimodal models for modality-specific speech, facial images and EEG; a novel Deep Belief Network and Deep Recurrent Neural Network for the video and EEG/physiological branches respectively; a cross-modal heterogeneity resolution mechanism using a shared latent space across the two modalities; a novel hybridization of Bat-Rider Optimization Algorithm for the fusion rule; and a validation plan covering the DEAP benchmark and an original Indian-context multimodal dataset. Finally, the paper explicitly traces each part of the proposed approach to the gap it is filling, giving a clear foundation for the design of the framework.

Keywords: Multimodal Emotion Recognition, Gap Analysis, Deep Learning, Fusion Strategies, Renyi entropy, Metaheuristic Optimization, EEG, Stress Detection, Depression Detection.

1. Introduction

Despite the improvement of emotional recognition using a single sensing modality, such modality-specific limitations still exist: Facial expression can be consciously masked, speech can be fooled by acoustic noise, and text can be fooled by voluntary self-report. Multimodal emotion recognition (MER), on the

other hand, utilizing two or more modalities, has gained increasing research interest and is particularly important for stress and depression screening, where the accuracy and the robustness of the result have direct implications for the tested person. Although there is a large and increasing literature base on the MER, this literature base shows uneven progress, with some modality combinations and fusion strategies having been explored extensively, and others, such as combinations and strategies that are directly relevant to stress and depression screening being relatively under-explored.

This paper has two contributions. First, it provides a narrative, evidence-based gap analysis of the MER literature that is more structured than a summary of individual studies to better identify specific, repeated gaps that hinder the current progress of the field. Second, on top of this gap, it proposes an AI-based emotion-recognition framework based on deep-learning technologies, providing a detailed description of the respective designs and their justification on the basis of the identified gaps. This paper is organized as follows: The related literature is reviewed by Modalities and Fusion strategy in Section 2. The structured gap analysis is included in Section 3. In Section 4, the proposed method is presented in its entirety. The proposed approach is explicitly linked to the identified gaps in Section 5. The proposed approach's importance and restrictions are discussed in Section 6 and the paper is concluded in Section 7.

2. Related Work

2.1 Unimodal Emotion Recognition

There has been a vast amount of research based on convolutional architectures for facial-expression recognition [1, 2] and FER2013 [3] is now a well-known benchmark in the field of naturalistic face image collections. Speech emotion recognition has benefited from hybrid convolutional-recurrent architectures [4, 5]. EEG-based emotion recognition has been shown to benefit from Deep Belief Networks [6], with differential entropy [7] and frequency-band/channel-importance analysis [8] providing the dominant feature-extraction approach. Physiological-signal-based recognition has been reviewed comprehensively [9], with weighted multichannel fusion strategies proposed for combining multiple physiological signals [10].

2.2 Multimodal Fusion Strategies

Among multimodal studies, Hassan et al. [11] combined a Deep Belief Network with statistical physiological features via feature-level fusion, achieving the highest reported accuracy (89.53%) among the multimodal studies reviewed for this analysis, evaluated on the DEAP dataset, but did not incorporate EEG or video. Njoku et al. [12] compared early, hybrid, and multi-task fusion of visual, audio, and EEG modalities, representing the only reviewed study to jointly incorporate EEG with audio-visual modalities. Chao et al. [13] used an LSTM-RNN with temporal pooling to fuse audio, visual, and physiological modalities. Mittal et al. [14] proposed a multiplicative, per-sample attention-based fusion mechanism (M3ER), among the highest-accuracy fusion strategies reviewed (82.7%). Hazarika et al. [15] proposed a Conversational Memory Network using attention-based hops to merge modality-specific memory representations (77.6% accuracy, the second-highest reviewed). Baltrusaitis et al. [16] provided an influential taxonomy of multimodal machine learning — representation, translation, alignment, fusion, and co-learning — used to frame the gap analysis presented in Section 3.

2.3 Comparative Summary

Table 1 consolidates the ten multimodal studies most closely related to the present work, summarizing modality, dataset, fusion mechanism, and reported accuracy.

Study	Modality	Dataset	Fusion Mechanism	Accuracy
Chao et al. [13]	Audio, Visual, ECG, EDA	RECOLA	Feature-level, temporal pooling	66.7%
Kahou et al. [17]	Audio, Video	FER+, TFD, custom	Decision-level (DBN+CNN)	47.67%
Tzirakis et al. [18]	Audio, Visual	RECOLA	Decision-level (CNN-ResNet+LSTM)	76%
Sahay et al. [19]	Text, Audio, Visual	CMU-MOSEI	Decision-level (LSTM per modality)	49.1%
Hazarika et al. [15]	Visual, Speech, Text	IEMOCAP	Attention-based memory fusion	77.6%
Hassan et al. [11]	EDA, PPG, EMG	DEAP	Feature-level (DBN+SVM)	89.53%
Siriwardhana et al. [20]	Visual, Audio, Text	IEMOCAP	Hybrid (BERT-like SSL)	73.98%
Priyasad et al. [21]	Text, Audio	IEMOCAP	Feature-level (DCNN, RNN)	80.51%
Mittal et al. [14]	Visual, Speech, Text	IEMOCAP, CMU-MOSEI	Multiplicative attention fusion	82.7%
Njoku et al. [12]	Visual, Audio, EEG	RAVDESS, manual EEG	Early/hybrid/multi-task fusion	78.75%

Table 1: Comparative summary of the ten multimodal emotion-recognition studies most closely related to the present work

3. Gap Analysis

Building on the literature reviewed in Section 2, this section presents a structured gap analysis organized around six specific, evidence-grounded deficiencies in the existing MER literature. Each gap is stated together with the specific evidence from Table 1 and Section 2 that supports it.

Gap 1: Limited Joint Use of Video, EEG, and Physiological Signals

Of the ten multimodal studies summarized in Table 1, only Njoku et al. [12] jointly incorporates EEG alongside video/audio modalities, and no reviewed study combines video, EEG, and physiological signals together within a single framework. Most reviewed studies restrict themselves to two modalities at a time (commonly audio-visual, or text-visual), leaving the specific three-modality combination of video, EEG, and physiological signals — arguably the combination best suited to stress and depression screening, given EEG and physiological signals' resistance to conscious masking — comparatively unexplored.

Gap 2: Reliance on Fixed or Heuristic Fusion Weighting

Across the fusion mechanisms summarized in Table 1, feature-level and decision-level fusion, as used by Chao et al. [13], Kahou et al. [17], Tzirakis et al. [18], Sahay et al. [19], and Hassan et al. [11], rely on fixed combination rules (concatenation, averaging, or a fixed-weight classifier) rather than a fusion weighting scheme that is itself optimized as part of training. Even the more sophisticated attention-based mechanisms of Hazarika et al. [15] and Mittal et al. [14] learn fusion weights implicitly within the neural architecture, rather than through an explicit, interpretable, separately optimized weighting rule.

Gap 3: Absence of Entropy-Based Fusion Rules

None of the ten studies summarized in Table 1 employs an explicit information-theoretic (entropy-based) fusion rule to quantify and exploit the relative confidence of each modality's prediction on a per-sample basis. Wei et al.'s [10] weighted physiological-signal fusion strategy is the closest conceptual precedent identified in the broader literature, but it does not employ an information-theoretic weighting rule of the kind that Renyi entropy provides.

Gap 4: Scarcity of Metaheuristic Optimization for Fusion-Weight Tuning

No study identified in the reviewed literature applies a metaheuristic optimization algorithm (such as a Bat Algorithm, Rider Optimization Algorithm, or a hybrid thereof) specifically to the problem of tuning multimodal fusion weights. Where optimization is discussed in the broader deep-learning literature, it is typically limited to gradient-based hyperparameter tuning applied to the classification network itself, not to an independent, explicit fusion-weighting stage.

Gap 5: Limited Validation on Indian-Context Data

Every dataset referenced in Table 1 (RECOLA, FER+, TFD, CMU-MOSEI, IEMOCAP, DEAP, RAVDESS) was collected from predominantly Western, English-speaking populations. No reviewed multimodal study reports validation on a dataset collected from an Indian subject population, despite well-documented differences in facial-expressiveness norms, speech prosody, and cultural display rules for emotion that may limit the direct transferability of models trained and validated exclusively on Western data.

Gap 6: Limited Explicit Framing Around Stress and Depression Detection

Among the studies reviewed in Section 2, only Tadesse et al. [22] (text-based depression-post detection) and Kumar and Subha [23] (EEG-based depression prediction) explicitly frame their work around depression detection, and neither does so within a multimodal fusion framework. The great majority of the multimodal literature summarized in Table 1 is framed around general-purpose, discrete-emotion classification, leaving the specific translation from recognized emotional states to stress- and depression-relevant risk indicators largely unaddressed within a multimodal fusion context.

Summary of Identified Gaps

#	Gap	Key Supporting Evidence
1	Limited joint use of video, EEG, and physiological signals	Only Njoku et al. [12] combines EEG with audio-visual data among 10 reviewed studies
2	Reliance on fixed/heuristic fusion weighting	8 of 10 reviewed studies use fixed feature- or decision-level fusion

3	Absence of entropy-based fusion rules	0 of 10 reviewed studies use an explicit information-theoretic fusion rule
4	Scarcity of metaheuristic fusion-weight optimization	0 of 10 reviewed studies apply a metaheuristic optimizer to fusion weights
5	Limited Indian-context validation	0 of 10 reviewed studies validate on Indian-context data
6	Limited stress/depression-specific framing	Only 2 of 66 reviewed references address depression detection directly, neither multimodal

Table 2: Summary of the six identified research gaps and their supporting evidence

4. Proposed Method

This section presents the proposed AI-driven, deep-learning-based multimodal emotion-recognition framework, designed as a direct response to the six gaps identified in Section 3. Figure 1 illustrates the complete proposed framework.

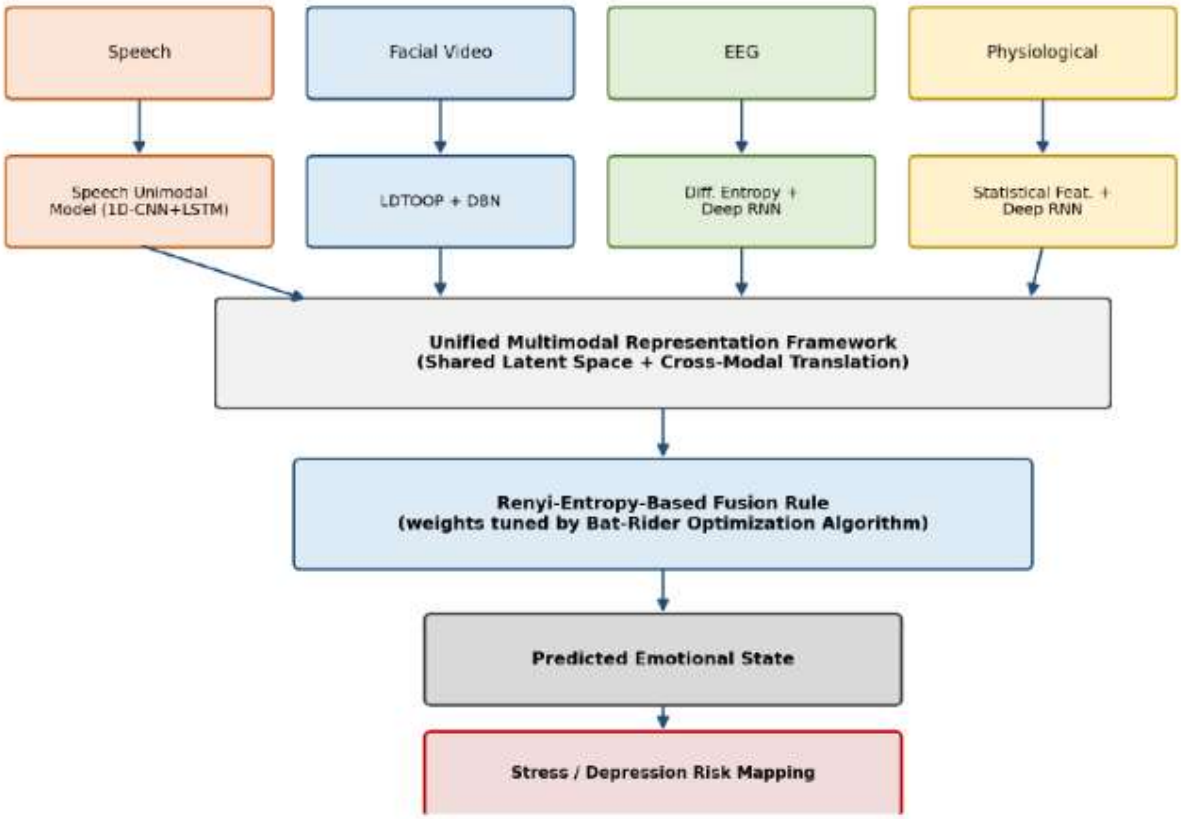

Figure 1: Proposed AI-driven multimodal emotion-recognition framework — end-to-end architecture

4.1 Unimodal Models

Three modality-specific unimodal models are proposed: a speech model combining MFCC and prosodic feature extraction with a hybrid 1D-CNN/LSTM classifier; a facial model combining face detection/alignment with a convolutional neural network; and an EEG model combining band-pass filtering and ICA-based artefact removal with differential-entropy feature extraction and a DBN-informed deep classifier. These models establish validated, modality-specific baselines and supply the feature-extraction pipelines used by the multimodal branches described below.

4.2 Deep Architectures for the Video and EEG/Physiological Branches

Facial video is processed using the Local Directional Ternary Optimal Oriented Pattern (LDTOOP) descriptor, followed by a Deep Belief Network (DBN) using greedy layer-wise unsupervised pre-training (Contrastive Divergence) and supervised fine-tuning, addressing Gap 1 by providing an architecture specifically suited to video-derived features. EEG and physiological signals are processed via per-band differential-entropy and statistical feature extraction, followed by a Deep Recurrent Neural Network (stacked LSTM layers), directly incorporating EEG — the modality identified in Gap 1 as comparatively under-used in the joint multimodal literature.

4.3 Unified Multimodal Representation Framework

To reconcile heterogeneity in sampling rate, dimensionality, and representational format across the video, EEG, and physiological branches, a unified representation framework projects each modality's features into a shared latent space of common dimensionality via modality-specific encoders, with modality-specific decoders enabling both reconstruction and cross-modal translation, directly supporting robustness to missing modalities.

4.4 Renyi-Entropy-Based Fusion with Bat-Rider Optimization

The modality-specific emotion-probability distributions P_m produced by the shared classification interface are combined using a Renyi-entropy-based fusion rule, directly addressing Gap 3:

$$H_\alpha(P_m) = (1/(1-\alpha)) \cdot \log(\Sigma_k p_{m,k}^{\alpha}) \qquad w_m = (1 - H_\alpha(P_m)) / \Sigma_{m'} (1 - H_\alpha(P_{m'}))$$

The order parameter α is tuned using a novel hybrid Bat-Rider Optimization Algorithm (BROA), combining bat-echolocation-inspired global search with rider-optimization-inspired local refinement, directly addressing Gap 4:

$$\alpha^*, \theta^* = \operatorname{argmax}_{\alpha,\theta} Accuracy_{val}(P_{fused}(\alpha, \theta))$$

4.5 Validation Plan

The complete framework is proposed to be validated on the DEAP benchmark dataset and, directly addressing Gap 5, on an original multimodal dataset collected from an Indian subject population under institutional ethical approval. Directly addressing Gap 6, a final evaluation stage maps recognized emotional states to stress- and depression-relevant risk indicators, validated against standardized psychological instruments (Perceived Stress Scale, PHQ-9).

5. Mapping the Proposed Method to the Identified Gaps

Table 3 provides an explicit connection of each gap identified in Section 3 with the specific element of the proposed method described in Section 4 that addresses the gap.

Gap	Addressed By	Section
1. Limited joint video+EEG+physiological fusion	Three-modality framework combining DBN (video) and	4.2

	Deep RNN (EEG/physiological) branches	
2. Fixed/heuristic fusion weighting	Explicit, separately optimized entropy-based fusion weighting	4.4
3. Absence of entropy-based fusion rules	Renyi-entropy-based fusion rule	4.4
4. Scarcity of metaheuristic fusion optimization	Novel hybrid Bat-Rider Optimization Algorithm (BROA)	4.4
5. Limited Indian-context validation	Original Indian-context multimodal dataset collection and validation	4.5
6. Limited stress/depression-specific framing	Explicit stress/depression risk-mapping evaluation stage	4.5

Table 3: Mapping of identified research gaps to corresponding elements of the proposed method

6. Discussion

Overall, the gap analysis given in Section 3 and its direct mapping to the proposed method in Section 5 jointly offer a structured basis for the novelty of the proposed framework: every component is added for the purpose of filling an identified gap, which has been founded on a piece of literature. This gap-driven design approach also elucidates the specific and falsifiable claims the proposed framework makes and which need to be addressed through empirical evaluation: that the combination of all three modalities (video, EEG, and physiological) gives a performance benefit over narrower combinations in the first place; that the fusion performed using entropy-based, BROA-tuned fusion offers a performance benefit compared to fixed-weight fusion; and that the framework generalizes to data from Indian-context and correlates meaningfully with well-known stress/depression instruments.

There are two constraints that should be highlighted for this gap analysis. It is not a systematic review, but a targeted literature search to generate a corpus of sixty-six references; there may be a limited number of relevant studies that are not included. Secondly, gaps in the literature reviewed at the time of writing will be identified in the gap analysis, and as the field of deep learning moves forward some of these gaps may close before the proposed framework of the thesis is fully validated and will be revisited in the discussion chapter of the thesis itself.

7. Conclusion

This paper has documented a detailed gap analysis of the existing multimodal emotion recognition literature and discussed the six identified specific and evidence-based gaps, with a proposed AI-based deep-learning framework specifically addressing each gap. The proposed method directly addresses the existing gap in the field by combining three different types of data i.e., video, EEG and physiological data using purpose-matched deep architectures, resolving the structural heterogeneity between these different data types using a unified representation framework, fusing these different data types using a novel entropy-based, metaheuristically optimized rule, and validating the proposed method on benchmark and original Indian-context data with an explicit stress/depression-relevance evaluation. The results of the

implementation and evaluation of this framework, based on the methodology and validation plan outlined in Section 4.5 and the broader research programme associated with this, will be reported as future work.

References

1. K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," IEEE Trans. Image Processing, vol. 26, no. 9, pp. 4193–4203, 2017.
2. A. Talipu, A. Generosi, M. Mengoni, and L. Giraldi, "Evaluation of deep convolutional neural network architectures for emotion recognition in the wild," 2019 IEEE 23rd ISCT, 2019.
3. I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.
4. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312–323, 2019.
5. M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018.
6. W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using Deep Belief Networks," 2014 IEEE ICME, 2014.
7. R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," 2013 6th Int. IEEE/EMBS Conf. on Neural Engineering (NER), 2013.
8. Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Trans. Autonomous Mental Development, vol. 7, no. 3, pp. 162–175, 2015.
9. M. Egger, M. Ley, and S. Hanke, "Emotion recognition from Physiological Signal Analysis: A Review," Electronic Notes in Theoretical Computer Science, vol. 343, pp. 35–55, 2019.
10. W. Wei, Q. Jia, Y. Feng, and G. Chen, "Emotion recognition based on weighted fusion strategy of Multichannel Physiological Signals," Computational Intelligence and Neuroscience, vol. 2018, pp. 1–9, 2018.
11. M. M. Hassan, Md. G. R. Alam, Md. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," Information Fusion, vol. 51, pp. 10–18, 2019.
12. J. Njoku, Angela C. Caliwag, W. Lim, S. Kim, H.-J. Hwang, and J.-W. Jeong, "Deep Learning Based Data Fusion Methods for Multimodal Emotion Recognition," J. Korean Inst. Comm. Info. Sciences, vol. 47, no. 1, pp. 79–87, 2022.
13. L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," Proc. 5th Int. Workshop on Audio/Visual Emotion Challenge, 2015.
14. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," Proc. AAAI, vol. 34, no. 02, pp. 1359–1367, 2020.
15. D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," Proc. NAACL-HLT, 2018.
16. T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and taxonomy," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

17. S. E. Kahou et al., “Emonets: Multimodal deep learning approaches for emotion recognition in video,” Journal on Multimodal User Interfaces, vol. 10, no. 2, pp. 99–111, 2015.
18. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-End Multimodal Emotion Recognition Using Deep Neural Networks,” IEEE JSTSP, vol. 11, no. 8, pp. 1301–1309, 2017.
19. S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, “Multimodal relational tensor network for sentiment and emotion classification,” Proc. Challenge-HML, 2018.
20. S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, “Jointly fine-tuning ‘bert-like’ self supervised models to improve multimodal speech emotion recognition,” Interspeech 2020, 2020.
21. D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Attention driven fusion for multi-modal emotion recognition,” ICASSP 2020, 2020.
22. M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of Depression-Related Posts in Reddit Social Media Forum,” IEEE Access, vol. 7, pp. 44883–44893, 2019.
23. S. D. Kumar and D. P. Subha, “Prediction of depression from EEG signal using Long short term memory (lstm),” 2019 ICOEI, 2019
24. S. Koelstra et al., “DEAP: A database for emotion analysis using physiological signals,” IEEE Trans. Affective Computing, vol. 3, no. 1, pp. 18–31, 2012.