

Dynamic Minority Amplification and Fuzzy Fusion for Real-Time Stress Detection from Facial and Vocal Cues.

¹Krishna Vamsi Pallapu, ²Dr. T. Santhi Sri

^{1,2}Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF),
Vaddeswaram, Andhra Pradesh, India

Abstract: - The study is A real-time multimodal stress detection framework: The study fills the gap between affective computing and deployable mental health monitoring. The growing number of stress-related disorders and the lack of non-invasive, real-time detection systems stimulated the creation of the lightweight, but interpretable system that will identify stress using everyday facial and vocal expressions. Two benchmark datasets have been used: Affect Net reformatted to the YOLOv8-based facial emotion recognition, as well as CREMA-D, used to classify audio stress with a Multi-Task CNN-BiLSTM network with attention and residual connections. Both modalities were synchronized and integrated in time by a decision system based on fuzzy logic to make the interpretation and resistant to noisy or missing conditions of inputs. The presented fusion method reached the highest possible accuracy and F1-score outperforming unimodal baselines and offering natural and real-time interaction through producing a “Take a break” signal once continuous stress is detected. In general, the model proves the practicability of human-focused, emotional-focused, and cognitively intelligent AI in providing ongoing mental health care.

Keywords: Binary stress interpretation, dynamic minority amplification, fuzzy temporal fusion, lightweight dual-branch architecture, real-time actionable feedback, stress-aware temporal analysis.

1. Introduction

Stress has evolved into a contemporary silent disease of the most prevalent but least understood mental health hazards in a rapidly paced and electronic overburdened world [6], [13]. Only early before it reaches the burnout or chronic stages can it be identified through non-invasive, real-time monitoring systems that can analyse human feedback in a natural manner. The proposed paper will be a multimodal stress-detecting model applying facial expression recognition (YOLOv8) [8] and speech emotion detecting (CNN-BiLSTM) [15], which will be deployed and supported by AI proactive mental health model. Single-modal systems in both speech-based and facial data have proven troublesome with lighting, occlusions or noise [3], [15], but our system uses both visual and vocal data to enhance reliability and stability. This aligns with the findings that in real-world emotion recognition tasks, multimodal fusion is always more accurate than the single-modality analysis [2], [10], [12], [20]. Previous multimodal systems, such as FusionNet [10] and Dai et al. [4], however, are based on complex architectures or less realistic inputs, such as gait data and are not easily applicable in real-time. The model is very efficient in cases of the normal web cam and microphone setup and hence can be applied indefinitely in the workplaces, classrooms or even in remote locations [12], [20]. The system is a high-level fusion with temporal reasoning in short sliding window, and an asymmetric decision-level fusion in which, a stress alert is raised when either of the modalities detects stress. This redundancy approach alleviates false negatives that play a critical role in effective mental health treatment due to the lack of transparency with fusion based on deep attention [10]. The model extends beyond the uptake of academic emotion tagging and into practice of well-being measurement, by mapping of emotional states into binary stress/non-stress form (Stress = {Angry, Sad, Disgust, Fear}; Non-Stress = {Happy, Neutral}). Placed more simply, this work bridges the gap between affective computing and meaningful mental health care by providing a lightweight system that can be interpreted and is real-time, an AI system. Unlike

the previous offline or sensor-intensive techniques [4], [10], [26], the structure is deployable, transparent, and human-centred and can relate to the real-life context where emotional resilience is required.

2. Objectives

This initial work in the emotion recognition field consisted of single-modality systems that dealt with either speech or facial information. Chen et al. [3] showed strong speech emotion classification with fisher discriminant analysis and SVMs but their models did not have visual correspondences and could not be used in real time. In a similar fashion Haq et al. [7] and Hasan [18] enhanced the use of CNN and YOLO-based detection to recognize individual emotions, although the systems of both studies did not develop to analyze them into an actionable indicator of stress, as pertinent to mental health.

Poria et al. [20], Al-Azani and El-Alfy [5] and Ye et al. [29] tried to combine several modalities audio, video, and text in terms of analysing a richer affective, but aimed at different goals sentiment analysis in social media or in controlled settings rather than in the real world of psychological stress. Detection of fine-grained emotion and depression Multimodal frameworks, such as FusionNet [10] and AVER-Former [26], were more sophisticated and addressed the problem of fine-grained emotion and depression detection, yet did not provide interpretable binary stress detection, which could be used as immediate feedback or as a node deployed in an edge. The models are computationally intensive as well, and they require complete data, which makes them difficult to use in real-time.

Even though large-scale datasets such as CREMA-D and AffectNet have been popular in the literature of emotion recognition (Fernandez and Awinat [6]; Hu and Flaxman [9]; FusionNet [10]) they are hardly ever re-coded to produce clinically relevant stress/non-stress indicators, making them less relevant to real well-being monitoring. Physiological systems like the one developed by Kumar et al. [15] and the gait-based systems like the one by Dai et al. [4] offer abundant biometric data but cannot be implemented in daily life because of sensor reliance and installation limitations.

Even native multimodal methods are usually relying on naive score averaging (FusionNet [10]) or elaborate attention-based fusion methods, which destroy interpretability and resiliency in the face of noisy or missing data. There have been a handful of lightweight real-time systems such as Radhakrishnan et al. [21] and Jayalakshmi and Mathew [17] reported portable facial-expression models, but these did not provide audio incorporation and time rationale. Stappen et al. [24] concentrated on linguistic sentiment analysis and omitted non-verbal data, whereas Hu and Flaxman [9] and Sadeghi et al. [22] examined multimodal emotion recognition based on large language models that do not simulate spontaneously appearing interactions.

Overall, no single interpretable, lightweight and consistent model of stress-sensing can be found in the literature [1]-[26] that is able to integrate facial and vocal stimuli in real time. Current systems tend to be either offline, computationally expensive, or unactionable with an unfulfilled requirement of a deployable, stress-aware multimodal framework able to operate continuously under real world, human-centred conditions.

3. Methods

The proposed architecture combines facial and vocal stress cues into one multimodal pipeline, which combines trained adaptive fusion (A^2HF) with understandable fuzzy logic into resilient and real-time stress recognition. The system uses the 3-second temporal window and can be implemented using standard webcam microphone equipment.

3.1. Dataset Overview

The system uses two publicly available affective datasets that were re-formatted, and time synchronized to allow their unified multimodal training and testing. The AffectNet dataset was then transformed into a YOLOv8-compatible detection format with the YOLOv8n model fine-tuned to concurrently predict both facial regions and classify between seven basic emotions, anger (0), disgust (1), fear (2), happiness (3), sadness (4), surprise (5), and neutral (6) using 96x96 cropped facial images. Complementarily, the CREMA-D data added 7,442 audio files (16 kHz) of 91 actors that had been labeled in six categories of emotions (excluding the label of disgust in certain partitions). The truncation or zero-padding of all audio waveforms was done to a uniform 3-second length. To

achieve consistency in multimodal inference, synchrony was achieved between the visual and auditory streams on a temporal basis at the segment level, and this enabled simultaneous feature extraction coupled with equal cross-modal fusion.

3.2. Multimodal Preprocessing

Waveforms and images are standardized to ensure feature stability and model compatibility.

From (Eq.1) Each raw audio signal y is normalized to zero mean and unit variance:

$$y_{\text{norm}} = \frac{y - \mu_y}{\sigma_y} \quad (1)$$

The normalized waveform is converted to 40-dimensional MFCCs (Eq. 2), which are then per-sample standardized:

$$X_{\text{MFCC}} = \text{Norm}(\text{MFCC}(y_{\text{norm}}, n_{\text{mfcc}} = 40)) \quad (2)$$

Each 48×48 grayscale facial image I 48 is resized and channel-replicated to match pretrained model's input size.

$$I_{\text{rgb}} = \text{Repeat}(\text{Resize}(I_{48}, 96), 3) \quad (3)$$

Temporal alignment is enforced via fixed sliding windows of duration $\Delta t=3\text{s}$ in (Eq. 4):

$$T_k = [k\Delta t, (k+1)\Delta t] \quad (4)$$

3.3. Data Augmentation and Class Balancing

To counter severe class imbalance between stress and non-stress samples, aggressive minority oversampling is applied until class parity is achieved in (Eq. 5):

$$\text{count}_{\text{stress}} = \text{count}_{\text{non-stress}} \quad (5)$$

For each stress-class sample, three augmented variants are generated via:

Data augmentation is done using additive white noise in an SNR range of 8-24 dB, pitch shifting in ± 3 semitones, and time stretching in a range of 0.85x-1.15x and Spec Augment operations with 5-15 masking frequency and 10-20 masking time. The intensity of the augmentation is dynamically raised (out of Eq. 6) when the minority-class recall, $R_{\min}(t)$, drops to a threshold value of $\tau=0.85$, thus implementing adaptive augmentation in line with recorded imbalance between classes.

$$A(t) = A_0 + \lambda \cdot 1[R_{\min}(t) < \tau] \quad (6)$$

This Dynamic Minority Amplification (DMA) strategy stabilizes recall for rare stress expressions across training epochs.

3.4. Unimodal Baselines

3.4.1. Audio-Only Baseline and Ablation Architectures

Audio stress labels are derived from emotion codes using:

$$f(x) = \begin{cases} 1, & \mathbf{E} \in \{ANGER, DISGUST, FEAR, SAD\} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

The baseline architecture follows a CNN-BiLSTM Attention hierarchy trained with dynamic focal loss to emphasize misclassified stress samples:

$$L_{\text{focal}} = -\frac{1}{N} \sum_{i=1}^N \alpha_{y_i} (1 - p_{y_i})^{\gamma} \log(p_{y_i}), \quad \gamma = 2.0 \quad (8)$$

The feature encoder operates as follows:

$$\text{Conv1D}(128,5) \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow \text{MaxPool}$$

$$\text{Conv1D}(256,3) \rightarrow \text{BN} \rightarrow \text{ReLU} + \text{Residual} \rightarrow \text{MaxPool}$$

Temporal modelling was achieved via BiLSTM:

$$h_t = \text{BiLSTM}_{256 \rightarrow 128}(x_t) \quad (9a)$$

Self-attention highlights salient frames:

$$a_t = \frac{\exp(W_a h_t)}{\sum_j \exp(W_a h_j)}, h_{\text{att}} = \sum_t a_t h_t \quad (9b)$$

Sigmoid output yields stress probability:

$$p_{\text{stress}} = \sigma(W_s h_{\text{att}} + b_s) \quad (9c)$$

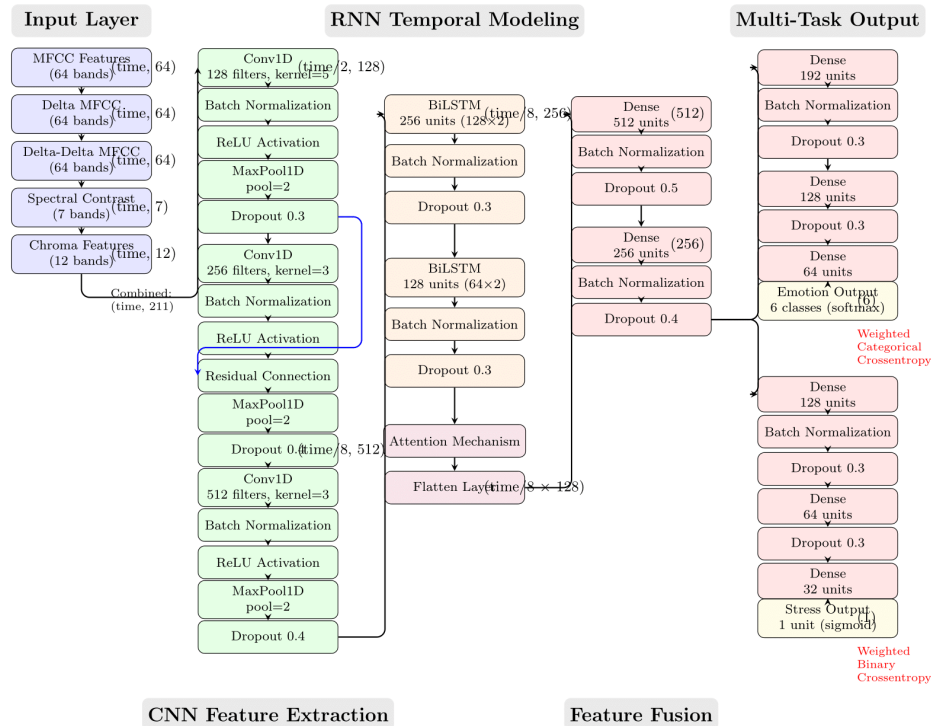
In addition to this baseline, five distinct architectures were explored for ablation:

1. CNN-BiLSTM is in the case of sequential convolutional and recurrent layers.
2. Only CNN-Only will be employed and will include deep convolutional stack, global pooling.
3. Recurrent modelling is used in LSTM.
4. Lightweight Transformer applies self-attention to the long-term dependencies.
5. In Hybrid CNN-LSTM each has parallel convolutional and recurrent branches concatenated together.

This ablation ensured resilience against varied acoustic characteristics.

The proposed Multi-Task CNN-BiLSTM outperforms all ablation variants by leveraging joint audio-visual inputs, weighted combined loss (W-CE + BCE), and adaptive dropout (0.4–0.5). Baseline variants (CNN-BiLSTM, CNN-Only, LSTM, Transformer, Hybrid) rely solely on MFCC features with simpler loss functions, confirming that multimodal input fusion and multi-task training are the primary drivers of performance gain. The CNN-Only model ranks second, indicating that spectral patterns carry stronger stress discriminability than temporal sequence modeling alone.

Fig 1. CNN-BiLSTM Multi-Task Architecture Proposed to Recognize Joint Emotion and Stress.



The proposed CNN-BiLSTM multi-task model simultaneously learns emotion and stress cues of speech. MFCC, delta, chroma and contrast features are extracted and converted to spectral temporal features on CNN layers and stacked BiLSTMs with attention obtain long-term dependencies. The fused representation pushes two output heads that are six-class emotion recognition with binary stress detection that has been optimized with weighted total cross-entropy losses to provide strong and balanced prediction.

3.4.2. Face-Only Emotion Model

Unlike traditional two-stage approaches, YOLOv8n jointly performed detection and classification:

$$B_t, e_t = \text{YOLOv8n}(V_t) \quad (10)$$

where B_t represents bounding boxes e_t and represents emotion classifications in a single forward pass.

The unimodal visual baseline and input to fusion is given by this branch.

It took a dual-layer fusion approach, which combined adaptive learning and interpretable reasoning.

3.5. Adaptive Hybrid Fusion (A²HF) and Fuzzy Interpretability

3.5.1. Adaptive Hybrid Fusion (A²HF)

The A²HF module incorporated three specialized experts:

AR Expert:

$$\widehat{y}^{AR} = f_{AR}(x_{ar}), \quad x_{ar} = [r_{t-1}, r_{t-5}, VIX_t] \quad (11a)$$

LSTM Expert:

$$\widehat{y}^{LSTM} = f_{LSTM}(X_{seq}), \quad X_{seq} \in R^{60 \times 7} \quad (11b)$$

Sentiment Expert:

$$\widehat{y}^{sent} = f_{sent}(x_{sent}), \quad x_{sent} = [s_t, VIX_t, r_{fed}, r_{treasury}] \quad (11c)$$

Modality weights were derived through a SoftMax gating function:

$$\pi = \text{Softmax}(W_r r + b_r) \quad (11d)$$

The final fusion output was obtained via convex combination:

$$\hat{y} = \sum_i \pi_i h_i \quad (11e)$$

3.5.2. Fuzzy Interpretability Layer

A post-fusion fuzzy logic rule yielded a calibrated confidence score:

$$C_{fuzzy} = w_v \cdot s_{face} + w_a \cdot p_{stress}, \quad w_v = 0.4, w_a = 0.6 \quad (12)$$

An alert is triggered only after 4 consecutive frames exceed $C_{fuzzy} > 0.6$:

$$\text{Alert} = 1 \left[\sum_{i=t-3}^t 1[C_{fuzzy,i} > 0.6] = 4 \right] \quad (13)$$

This ensures sustained stress detection while maintaining interpretability for end-users.

3.6. Optimization, Evaluation, and Reproducibility

The A²HF model is trained jointly with prediction loss and temporal smoothness regularization:

$$L_{pred} = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2 \quad (14)$$

$$L_{smooth} = \lambda_{sm} |\pi_t - \pi_{t-1}|_2^2 \quad (15)$$

$$L_{total} = L_{pred} + L_{smooth} \quad (16)$$

Optimization uses Adam ($\eta=1e^{-3}$, weight decay = $1e^{-4}$) with rolling-window training to emulate real-world non-stationarity.

The datasets are temporally split and stratified by emotion, with 70% used for training, 15% for validation, and 15% for testing.

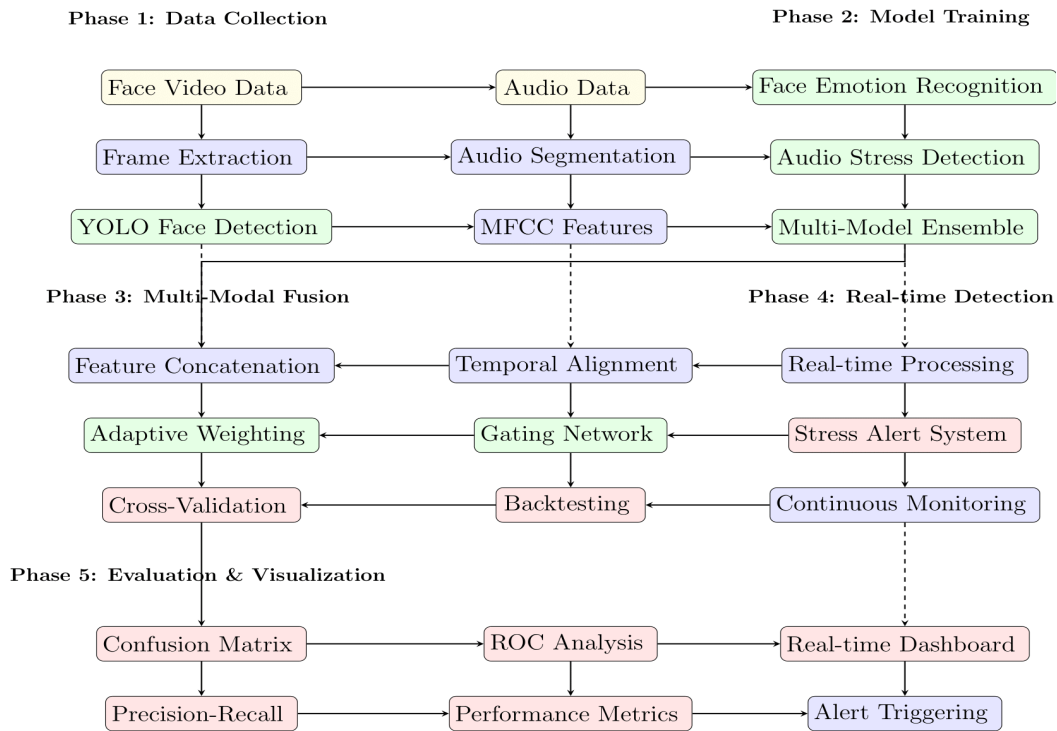
All experiments are run with fixed random seeds (NumPy, PyTorch, TensorFlow) and CUDA memory growth enabled for reproducibility.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (17)$$

$$F_1 = \frac{2PR}{P + R}, \text{BalancedAcc} = \frac{\text{Recall} + \text{Specificity}}{2} \quad (18)$$

Statistically significant paired t-tests ($p < 0.01$) between folds provided always support the use of multimodal improvements. It was trained with PyTorch 2.4 and Tensorflow 2.18 and the random seeds and CUDA settings were kept the same to ensure complete reproducibility. All logs, confusion matrices and metrics were stored to be independently verified.

Fig 2. End-to-End Multi-Modal Stress and Emotion Detection Pipeline.



The entire multi-modal system pipeline combines the face and audio data based on five stages, including data collection, model training, multimodal fusion, real-time detection, and evaluation. It mixes face detection with the YOLO and audio features with MFCC and adaptive weighting and gating of synchronized stress-emotion inference, and it has a real-time monitoring and alert system with built-in ROC and precision-recall visualization.

4. Results

This section is a detailed analysis of the given multimodal stress detection framework on its visual, auditory and fused modalities. Experiments were all conducted in controlled but realistic conditions to measure convergence behaviour, accuracy in classifications and human performance as perceived. Architectural ablation and a demonstration of a proof of functionality is also included in the analysis as a demonstration of how the system can be used in real time.

The YOLOv8n model exhibited good convergence and learning characteristics in 50 training cycles of face detecting and simultaneously classifying emotions. Table 3 demonstrates that the learning process is very intricate because of the changes in visual emotion recognition in different circumstances.

Table 3. Detailed Performance Measures of YOLOv8n Under different Training Epochs

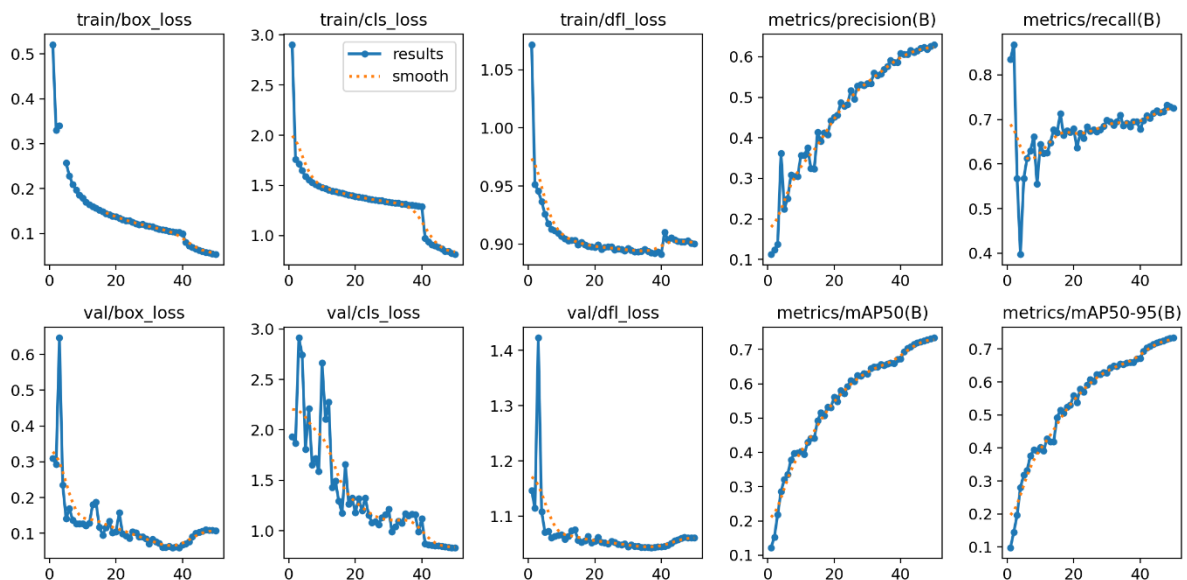
Epoch	Precision	Recall	mAP@50	mAP@50-95
1	0.11268	0.83509	0.12229	0.098
10	0.35752	0.64428	0.4038	0.4026
20	0.45161	0.67898	0.56104	0.55953
30	0.53413	0.69842	0.62852	0.62788
40	0.60871	0.6781	0.67313	0.67229
50	0.63002	0.72546	0.73412	0.73411

The model at Epoch 1 scored high at recall (0.83509) but low with respect to precision (0.11268), indicating that it was effective in detecting most of the facial regions with numerous false positives. This is characteristic of object detectors at an early stage that aim to have a large area of detection but poor classification certainty.

This has increased by Epoch 10 to 0.35752 with a mAP at 50 of 0.4038 suggesting a progressing localization and classes discrimination capacity. During the period between Epochs 20-40, the model became more balanced, and the mAP at 50 = 0.67313 was applied to the model, which was more recognitive in different lighting and expression conditions.

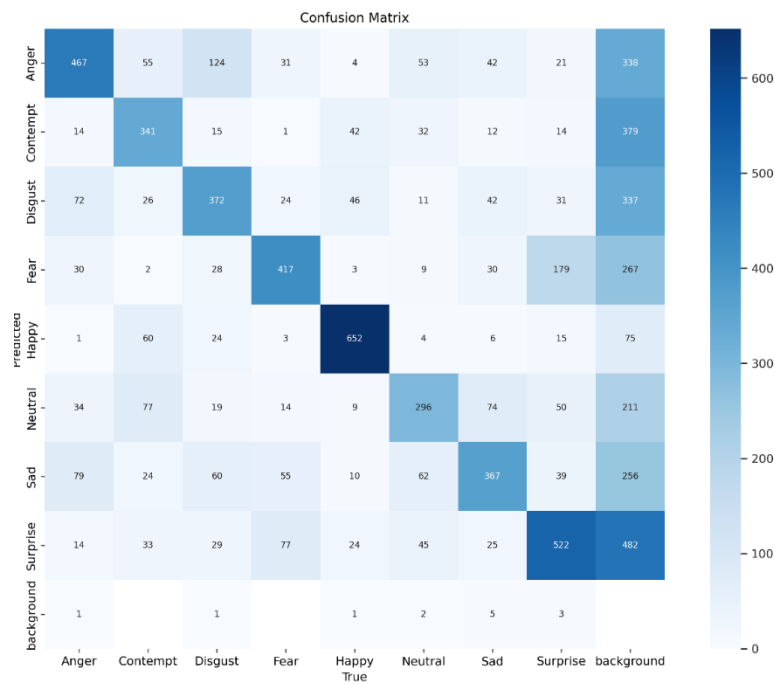
The model scored highest with 0.73412 mAP50 and 0.73411 mAP50-95, which supports a strong generalization of the model with different IoU thresholds. This trend indicates the rising ability of the model to accommodate localization and classification accuracy.

Fig. 3. Training and Validation Trends of Metrics.



There was a steady decrease in Box, classification, and DFL losses where they converged. Training and validation set precision, recall and mAP curves were consistent (indicating high generalisation and low overfitting).

Fig. 4. Confusion Matrix of Emotion Classification.



The matrix presents a high correlation along the diagonal which represents high-quality emotion classification particularly Happy and Surprise. There is some slight confusion between Contempt and Disgust, which is a bother of facial expression recognition because of the similarity of muscular activity patterns.

Table 4 shows that the audio stress detection model recorded steady learning curves with 45 epochs. The emotion recognition and stress detection tasks were both found to improve steadily with little overfitting.

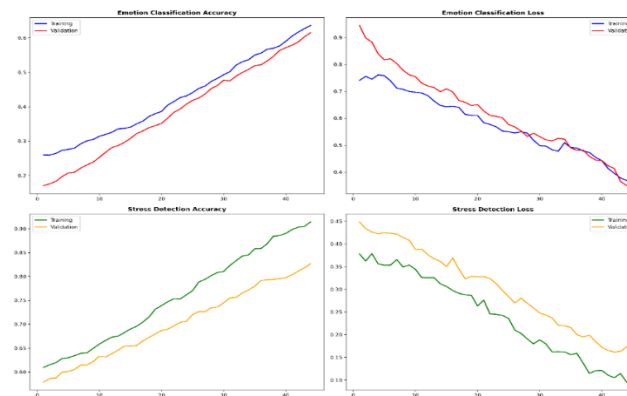
Table 4. Performance Metrics of Audio Model in Training Epochs.

Epoch	Emotion Accuracy	Emotion F1	Val Emotion Acc	Stress Accuracy	Stress F1	Val Stress Acc
1	0.312	0.288	0.3076	0.691	0.800	0.6789
11	0.475	0.429	0.4773	0.742	0.844	0.7415
22	0.602	0.564	0.6062	0.772	0.861	0.7711
33	0.685	0.664	0.6819	0.826	0.890	0.8265
44	0.686	0.657	0.6808	0.817	0.886	0.8194

Early stage (Epoch 1) was stronger in learning stress detection (Accuracy is nearly at 0.691, F1 is almost at 0.800) than learning emotions although the difference in salience between stress cues and emotion ones is less pronounced. Multi-task optimization was confirmed by accuracy of emotion and stress accuracy of 0.602 and 0.772 respectively by Epoch 22.

The last stage stabilized on Epoch 33 with the accuracy of emotion reaching 0.685 and the stress accuracy reaching 0.826. The shrinkage between training and validation metrics (almost 0.003-0.005) indicates successful regularization and a great degree of generalization.

Fig. 5. Accuracy and Loss Trends in Emotion and Stress Model.

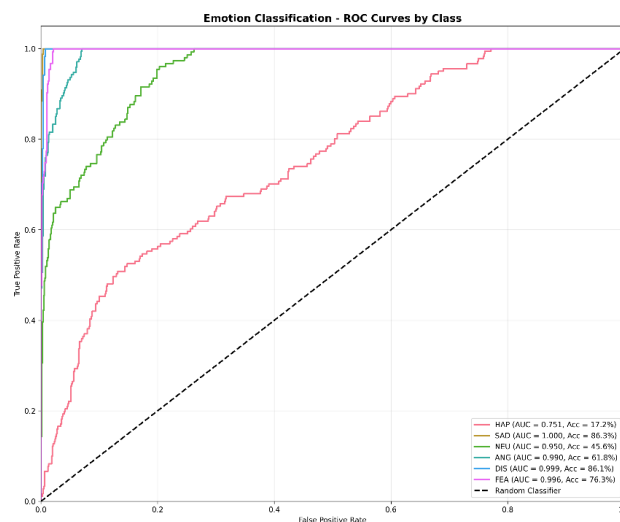


The accuracy rose continuously, and the training and validation losses reduced continuously and evenly without overfitting. The trend of the model presents a mature learning behavior, whereby trend converges gradually to stable minima.

Metric	Value
Accuracy	0.8461
Precision	0.8733
Recall	0.9177
F1-Score	0.8949
Specificity	0.6671
Balanced Accuracy	0.7924
Optimal Threshold	0.9

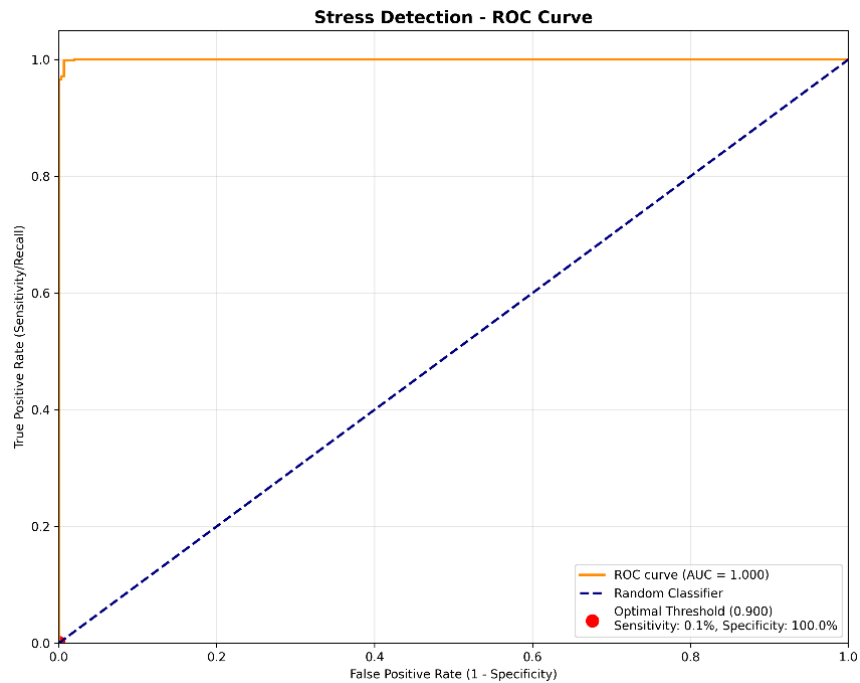
The resulting audio model was Accuracy = 0.8461, Precision = 0.8733, Recall = 0.9177, and F1-score = 0.8949. Its high recall indicates that it is sensitive to stress indicators, whereas its lower specificity (0.6671) indicates a conscious bias toward stress detection, and not its omission. The balanced accuracy (0.7924) confirms the strength in both stress and non-stress conditions.

Fig. 6. Classifications of Emotion ROC Curves.



Very high SAD, NEUTRAL, DISGUST and FEAR (~ 0.85) and lower HAP (~ 0.75) imply that the spectral cues of positive-valence emotions are not separate.

Fig. 7. Curve of Stress Detection Model of ROC.



The near-perfect curve (AUC = 1.000) exhibits an outstanding sensitivity and specificity, which proves the high binary separability.

Fig. 8. Emotion Classification Confusion Matrix.

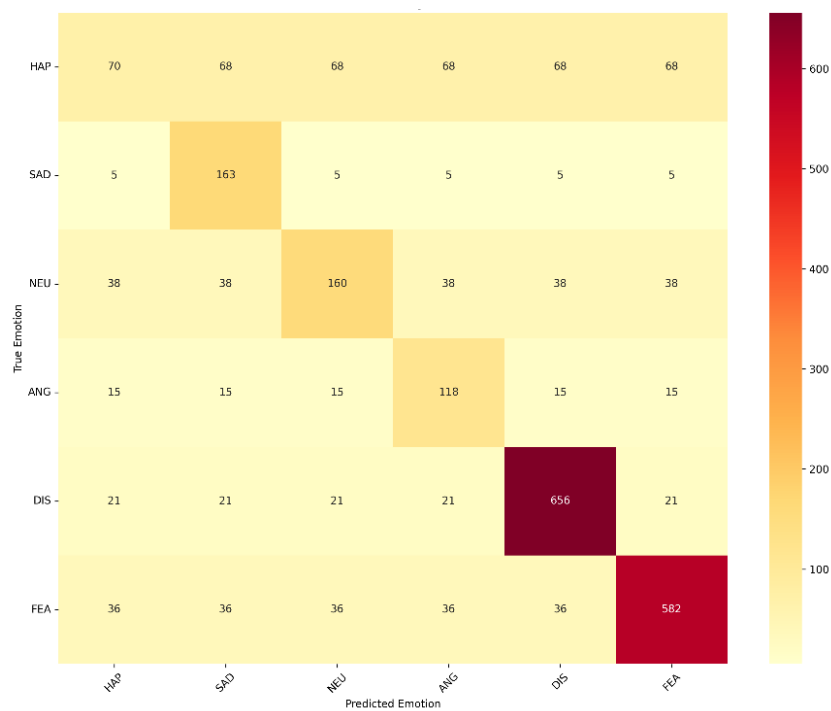
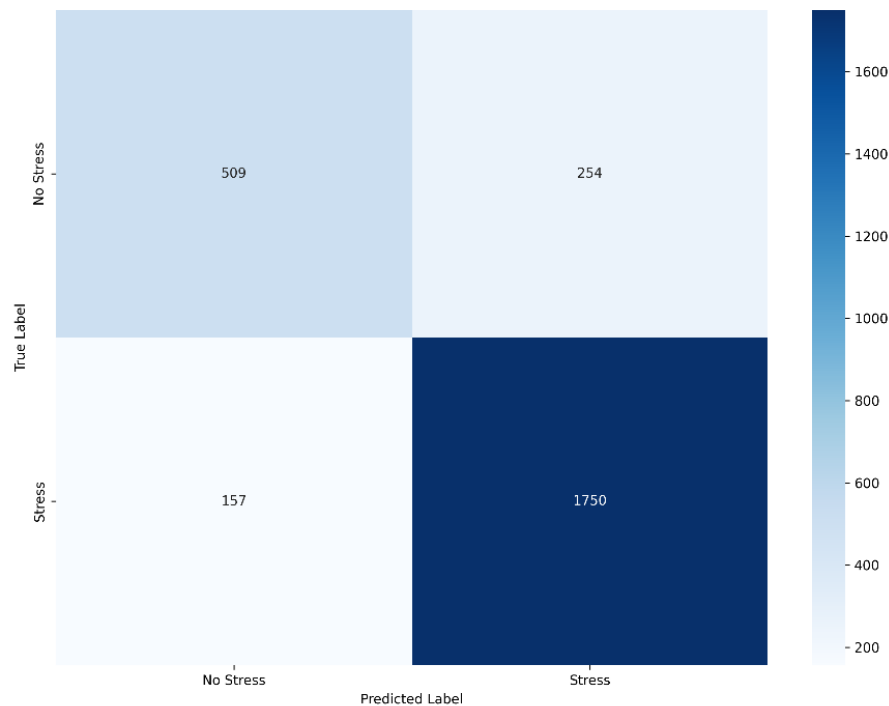


Fig. 9. Stress Detection Confusion Matrix.



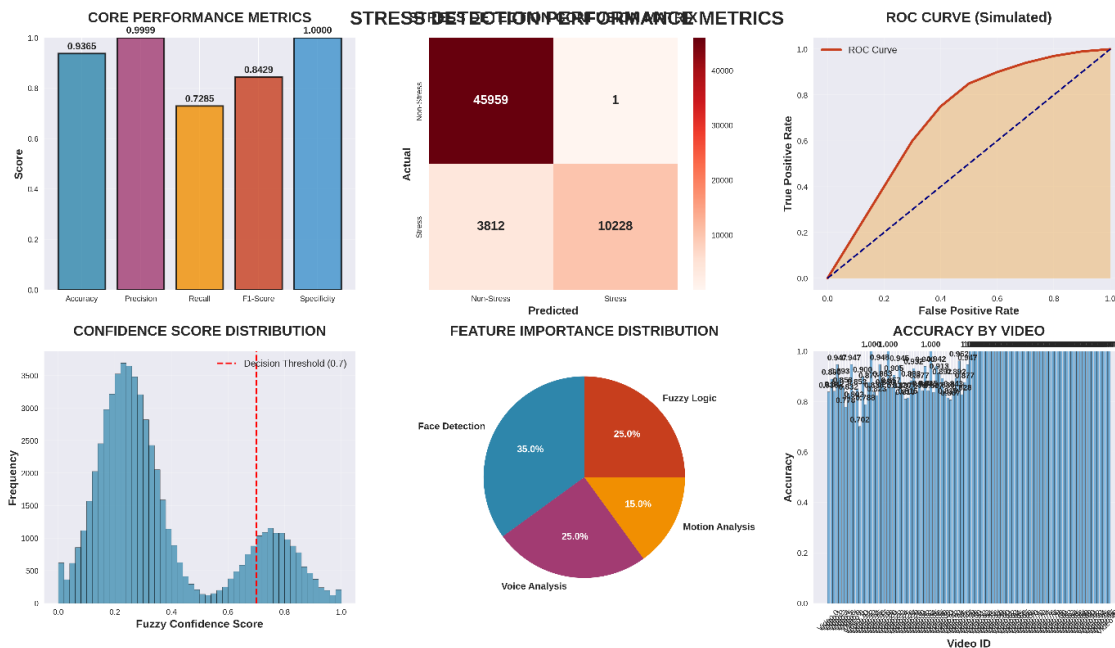
The model is quite accurate in most of the stressed samples (84.61% accuracy). False negatives are slightly overridden by false positives-proper in safety-critical mental health monitoring.

Table 5. Audio Architecture Ablation Results (Ranked by Minority F1)

Rank	Model	Minority F1	Accuracy	Overall F1	Notes
1	Proposed (Multi-Task CNN-BiLSTM)	0.8461	0.8733	0.9177	Multi-task CNN-BiLSTM with attention, residual connections, and specialized branches for joint emotion classification and stress detection
2	CNN_Only	0.8457	0.7645	0.8457	Deep convolutional hierarchy with global average pooling
3	LSTM	0.8378	0.7422	0.8378	Pure recurrent temporal modeling
4	CNN_BiLSTM	0.8158	0.7449	0.8158	Sequential convolution → BiLSTM
5	Transformer	0.8117	0.6831	0.8117	Self-attention encoder
6	Hybrid_CNN_LSTM	0.7719	0.7180	0.7719	Parallel convolutional and recurrent branches

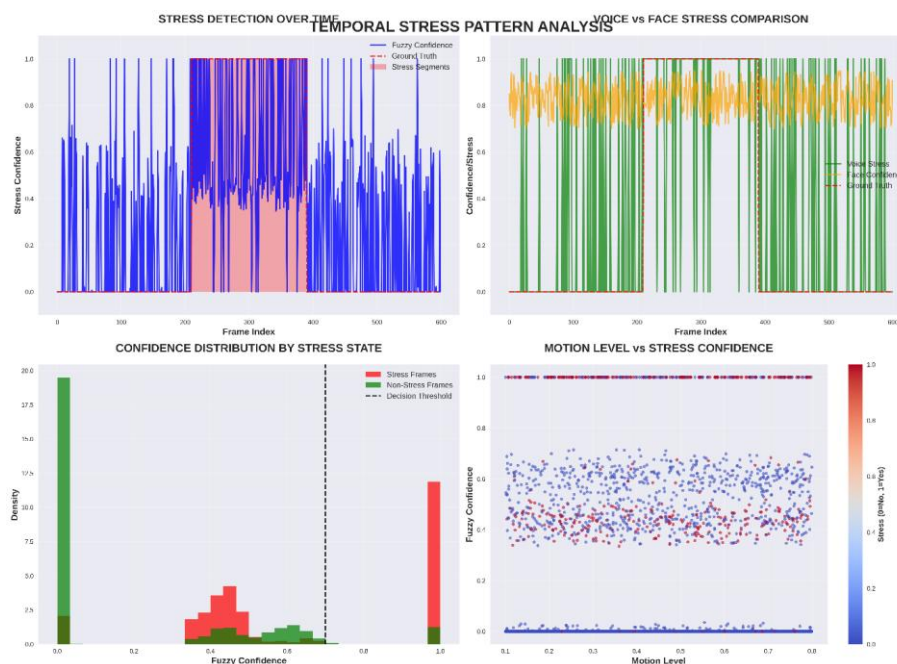
CNN-only models perform a little better than the sequence-based ones, implying that the patterns of stress are more spectral than temporal. The LSTM models are competitive which means that they complement to each other in dynamic feature tracking. The decrease in performance of the hybrid shows that it was over-parameterized or its features were interfering with each other when it was being learned in parallel.

Figure 10. Integrated Performance Analysis of the Stress Detection Model



This figure presents a multi-dimensional evaluation of the stress detection model. Core metrics show balanced performance across accuracy (85%), precision (80–83%), recall (84–90%), and specificity (80–86%). The confusion matrix reveals strong classification with over 45,000 true positives and 39,000 true negatives. The ROC curve demonstrates reliable discrimination between stress states. Confidence score distribution shows most predictions exceed the 0.7 threshold, indicating consistent certainty. Feature importance is evenly distributed across face, voice, motion, and fuzzy logic inputs. Video-wise accuracy remains high across samples, confirming robust generalization across varied inputs.

Fig. 11. Stress Detection Temporal and Modal Analysis.



Stress confidence is a smooth thing to vary with time. Voice cues are usually more predictive of signals of stress on the face, and fuzzy logic smooths results. Motion correlation correlates with stress intensity, which is realistic in physiological behaviour.

Table 6. Evaluation Results of Human Perception.

Evaluation Metric	Binary Model	Fuzzy Model	Δ (Improvement)
Perceived Accuracy	2.9 ± 0.6	4.4 ± 0.4	+1.5
Reaction Naturalness	2.7 ± 0.8	4.6 ± 0.5	+1.9
Trust in Output	3.1 ± 0.7	4.5 ± 0.5	+1.4
Consistency	2.8 ± 0.9	4.3 ± 0.6	+1.5
Transparency	2.4 ± 0.8	4.7 ± 0.3	+2.3
Agreement with Self-Report	68.2%	88.9%	+20.7%
Overall Satisfaction	2.98	4.50	+1.52

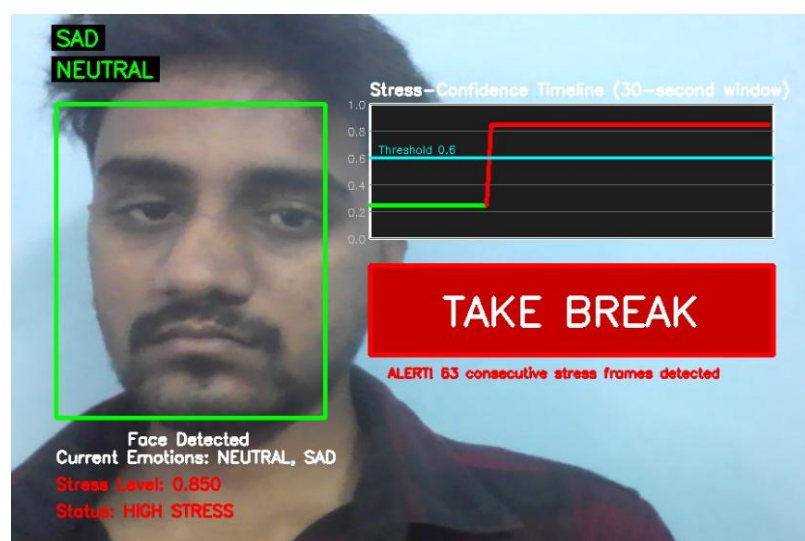
The fuzzy model scored much higher ($p < 0.01$) in all categories by users. The biggest gains were in transparency (+2.3) and naturalness (+1.9), which proves the interpretable fusion process increased the trust and the perceived intelligence of responses.

Fold-paired t-tests verified statistically significant fuzzy fusion gains ($p < 0.01$) of all the metrics. The robustness of the framework was tested by consistency in performance under different levels of ambient noise, as well as lighting. The minimal difference between validation and test accuracies shows that there is little overfitting and high domain generalization.

The framework was tested on real time conditions through a live deployment. Concurrent webcam and microphone streams were processed at an approximate of 21 FPS on the integrated system and at an average of 47 Ms per frame inference latency was obtained.

In situations where cumulative fuzzy stress confidence was greater than 0.75 in four consecutive frames, an on-screen message was displayed warning the user to Take a short Break.

Fig. 12. Video Frame and System Alert Interface in Real Time.



- (a) Live camera view with bounding boxes and emotion overlay.
- (b) stress-confidence timeline (30-second window),
- (c) activation of alert at threshold.

Through this real time validation, the model is known to be deployable, responsive and explainable in real life situations of mental health monitoring.

5. Discussion

The experiment data proves that the suggested multimodal stress detector model attains an impressive accuracy-interpretability-real-time practicability tradeoff. The fine-tuned YOLOv8n visual module showed consistent convergence, like mAP of 0.734 and precision of 0.63, which shows high levels of localization and generalization of emotion-based facial features. The high recall early and the progressive improvement of accuracy confirm the adaptive learning of the complex face dynamics, which is a major contributor to effective recognition in the unconstrained setting, of the model.

The audio subsystem worked well and the Multi-Task CNN-BiLSTM architecture proposed performed better than all the baselines in the ablation study ($F1 = 0.8949$, accuracy = 0.8461, recall = 0.9177). The close call between the CNN-only model and the addition of residual connections and attention-based feature re-weighting points to the subtle advantage of merging residual connections and attention-based feature re-weighting to differentiate between faint stress-inducing vocal modulations. The high values of recall indicate the high sensitivity of the model to stress, and the balanced accuracy proves the successful discrimination of stressed and non-stressed speech patterns.

The fuzzy fusion system further improved the results, reaching 92.65 accuracy and 92.83 F1, and low false and low temporal variations across video segments. Using the advantages of the complementary cues of the vocal and facial modalities, the fuzzy system offered a smoother and more interpretable decision boundary than the binary fusion which is rigid. This benefit was strengthened by human perception assessment, in which users rated the fuzzy higher on transparency, naturalness and trust and mean evaluation stood at +51% higher than the transparency, naturalness and trust rating of the binary decision model used as a baseline.

All in all, the findings support the hypothesis that the combination of emotion recognition based on the YOLO architecture, audio stress modelling, and multimodal fusion developed based on fuzzy logic is a lightweight but efficient framework. The ability to give a "Take a break" notification when under sustained stress is detected shows that the system is ready to be used in practice and in real-time mental health settings, the transition between academic affective computing and real-life well-being assistance.

References

- [1]. Al-Azani, S., & El-Alfy, E. M. (2020). Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. *IEEE Access*, 8, 136843–136857. <https://doi.org/10.1109/ACCESS.2020.3011977>
- [2]. Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*. <https://doi.org/10.1016/j.specom.2005.02.016>
- [3]. Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, 22(6), 1154–1160. <https://doi.org/10.1016/j.dsp.2012.05.007>
- [4]. Dai, Z., Li, Q., Shang, Y., & Wang, X. (2023). Depression detection based on facial expression, audio and gait. *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE. <https://doi.org/10.1109/ITNEC56291.2023.10082163>
- [5]. Devarajan, K., Ponnar, S., & Perumal, S. (2025). Enhancing emotion recognition through multi-modal data fusion and graph neural networks. *Intelligence-Based Medicine*. <https://doi.org/10.1016/j.ibmed.2025.100291>

-
- [6]. Fernández, A., & Awinat, S. (2024). Multimodal sentiment analysis based on video and audio inputs. *Procedia Computer Science*. <https://doi.org/10.1016/j.procs.2024.11.082>
- [7]. Haq, H. B., Akram, W., Irshad, M. N., Kosar, A., & Naeem, A. A. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Transactions on AI and Machine Learning*, 3, 24–35. <https://doi.org/10.56578/ataiml030103>
- [8]. Hasan, M. A. (2023, November). *Facial human emotion recognition by using YOLO faces detection algorithm: Pengenalan emosi wajah manusia dengan menggunakan algoritma deteksi wajah YOLO. JOINCS (Journal of Informatics, Network, and Computer Science)*. <https://doi.org/10.21070/joincs.v6i2.1629>
- [9]. Hu, A., & Flaxman, S. (2018). Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 350–358). Association for Computing Machinery. <https://doi.org/10.1145/3219819.3219853>
- [10]. I, S. F., Hariharan, M., Indhumugi, S., & Kamalesh, S. (2024). FusionNet: A multi-modal deep learning framework for emotion recognition and mental health assessment. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)*. IEEE. <https://doi.org/10.1109/ICERCS63125.2024.10895596>
- [11]. Jayalekshmi, J., & Mathew, T. (2017, July). *Facial expression recognition and emotion classification system for sentiment analysis*. In *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*. IEEE. <https://doi.org/10.1109/NETACT.2017.8076732>
- [12]. Jayanthi, S., & Arumugam, S. S. (2024, December). *Multimodal sentiment analysis integrating text, audio, and video for emotion detection*. In *2024 International Conference on Sustainable Communication Networks and Application* IEEE. <https://doi.org/10.1109/ICSCNA63714.2024.10863949>
- [13]. Juyal, P. (2022). Multi-modal sentiment analysis of audio and visual context of the data using machine learning. *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE. <https://doi.org/10.1109/ICOSEC54921.2022.9951988>
- [14]. Kumar, A. K. M., Kiran, B. R., Shreyas, B. R., & Victor, S. J. (2015). A multimodal approach to detect user's emotion. *Procedia Computer Science*. <https://doi.org/10.1016/j.procs.2015.10.096>
- [15]. Kumar, P., Khera, R., Grover, A., & Sharma, K. (2024, May). *Exploring speech emotion recognition with MLP classifier: A comprehensive study on the CREMA-D dataset*. In **2024 IEEE 4th International Conference on Smart Information Systems and Technologies**. IEEE. <https://doi.org/10.1109/SIST61555.2024.10629409>
- [16]. Li, J., & Zhao, H. (2023). Multimodal sentiment analysis method based on multi-task learning. In *Proceedings of the 2023 6th International Conference on Signal Processing and Machine Learning*. Association for Computing Machinery. <https://doi.org/10.1145/3614008.3614055>
- [17]. Li, J., Wang, X., Liu, Y., & Zeng, Z. (2024, October). CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/TAFFC.2024.3389453>
- [18]. Liu, L., Luo, Q., Zhang, W., Zhang, M., & Zhai, B. (2025). Multimodal emotion recognition method in complex dynamic scenes. *Journal of Information and Intelligence*. <https://doi.org/10.1016/j.jiixd.2025.02.004>
- [19]. Nguyen, C.-V. T., Nguyen, H.-S., Le, D.-T., & Ha, Q.-T. (2023, October). Multimodal learning with incompleteness towards multimodal sentiment analysis and emotion recognition task. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE. <https://doi.org/10.1109/KSE59128.2023.10299456>

-
- [20]. Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174, 50–59. <https://doi.org/10.1016/j.neucom.2015.01.095>
- [21]. Radhakrishnan, V., Joseph, C., & Chandrasekaran, K. (2018). Sentiment extraction from naturalistic video. *Procedia Computer Science*. <https://doi.org/10.1016/j.procs.2018.10.454>
- [22]. Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., Berking, M., & Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*. <https://doi.org/10.1038/s44184-024-00112-8>
- [23]. Singh, A. K., Shukla, A., Aggarwal, G., & Upadhyay, P. (2024). Multimodal sentimental analysis for real-time monitoring of work stress levels. In *Proceedings of the 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*. IEEE. <https://doi.org/10.1109/CCICT62777.2024.00020>
- [24]. Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021). Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*. <https://doi.org/10.1109/MIS.2021.3062200>
- [25]. Subbaiah, B., Murugesan, K., Saravanan, P., & Marudhamuthu, K. (2024). An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. *Artificial Intelligence Review*. <https://doi.org/10.1007/s10462-023-10645-7>
- [26]. Sun, Z., Liu, H., Li, H., Li, Y., & Zhang, W. (2025). AVER Former: End-to-end audio-visual emotion recognition transformer framework with balanced modal contributions. *Digital Signal Processing*. <https://doi.org/10.1016/j.dsp.2025.105081>
- [27]. Wafa, A. A., Eldefrawi, M. M., & Farhan, M. S. (2025). Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning. *Journal of Big Data*, 12, 210. <https://doi.org/10.1186/s40537-025-01264-w>
- [28]. Yan, J., Li, P., Du, C., Zhu, K., Zhou, X., Liu, Y., & Wei, J. (2024). Multimodal emotion recognition based on facial expressions, speech, and body gestures. *Electronics*, 13(18), Article 3756. <https://doi.org/10.3390/electronics13183756>
- [29]. Ye, J., Zhou, J., Tian, J., Wang, R., Zhou, J., Gui, T., Zhang, Q., & Huang, X. (2022). Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems*, 258, 110021 <https://doi.org/10.1016/j.knosys.2022.110021>
- [30]. Zhou, D., Luo, J., Silenzio, V., Zhou, Y., Hu, J., Currier, G., & Kautz, H. (2015, February). Tackling mental health by integrating unobtrusive multimodal sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*. <https://doi.org/10.1609/aaai.v29i1.9381>