International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Multimodal Smart Stress Analyzer Using Facial Expression and Speech Emotion Recognition

Prof. Ms. Vaishali D. Parihar
Information technology
Anuradha College of Engineering & Technology
Chikhli , India
vaishaliparihar08@gmail.com

Irfankha Yasinkha Pathan
Information technology
Anuradha College of Engineering & Technology
Chikhli , India
ip7958558@gmail.com

Govinda Suresh Pawar
Information technology
Anuradha College of Engineering and Technology
Chikhli,India
govindapawar8999@gmail.com

Jivan Kishor Tapkire
Information technology
Anuradha College of Engineering and Technology
Chikhli,India
jivantapkire@gmail.com

Karan Suresh Ambhore
Information technology
Anuradha College of Engineering & Technology
Chikhli , India
karanambhore07@gmail.com

***Abstract*—Mental stress has emerged as a pervasive health challenge in modern society, affecting cognitive performance, physiological well-being, and overall quality of life. This paper presents a Smart Stress Analyzer—a web-based multimodal system that detects and quantifies human stress by integrating facial expression recognition with speech-based emotion analysis. The facial component employs an EfficientNet-B2 convolutional neural network augmented with a Convolutional Block Attention Module (CBAM) to classify seven discrete emotional states from real-time webcam frames, achieving a test accuracy of 65.85%. The speech component fine-tunes a frozen Wav2Vec2-base transformer on a combined CREMA-D and RAVDESS dataset to predict five stress severity levels, attaining an overall test accuracy of 57%. Both models are integrated into a lightweight HTML/CSS/JavaScript web application that performs real-time inference without server-side processing. A weighted fusion strategy combines the two modality scores into a unified stress index displayed on an interactive dashboard. Experimental evaluation demonstrates the system's feasibility for continuous, non-intrusive stress monitoring in everyday environments.**

***Keywords*—Stress Detection, Facial Expression Recognition, Speech Emotion Recognition, EfficientNet-B2, Wav2Vec2, CBAM, Multimodal Fusion, Web Application**

I. INTRODUCTION

Mental stress is increasingly recognised as a global public health concern. According to the World Health Organization, stress-related disorders account for a significant portion of workplace absenteeism and reduced productivity worldwide. Prolonged psychological stress adversely affects the immune system, cardiovascular health, and cognitive function, making early and accurate detection a clinical and societal priority [1]. Conventional stress assessment relies on self-reported questionnaires such as the Perceived Stress Scale (PSS) or clinician-administered interviews, which are inherently subjective, retrospective, and impractical for continuous monitoring. Physiological approaches employing heart-rate variability, electrodermal activity, or cortisol levels offer greater objectivity but require specialised wearable hardware and controlled laboratory conditions [2].

Recent advances in deep learning have enabled non-contact, passive stress inference from readily available modalities such as facial video and speech audio. Facial expressions encode valence and arousal information that correlates strongly with affective state, while speech prosody, rhythm, and acoustic features reflect autonomic changes induced by stress [3][4]. Fusing these complementary channels reduces the ambiguity inherent in unimodal analysis and improves overall classification robustness [5].

This paper proposes a Smart Stress Analyzer—a fully client-side web application that simultaneously captures webcam frames and microphone audio, runs two independently trained deep learning models for facial emotion and speech stress classification, and fuses their outputs into a real-time stress level indicator. The system is architecture-agnostic with respect to server infrastructure, requires no specialised hardware, and is accessible from any modern browser, making it suitable for remote health monitoring and corporate wellness programs.

The remainder of this paper is structured as follows: Section II reviews related work; Section III describes the proposed system architecture; Section IV details the methodology; Section V presents the experimental setup; Section VI discusses results; Section VII highlights limitations; and Section VIII concludes with future directions.

II. PROBLEM STATEMENT

In Mental stress has become a critical concern in modern society, significantly impacting an individual's cognitive abilities, physical health, and overall well-being. According to the World Health Organization (WHO), stress-related conditions contribute heavily to decreased workplace productivity and increased absenteeism, highlighting the urgent need for effective and scalable monitoring solutions. However, existing stress detection methods suffer from several major limitations. Traditional approaches such as the Perceived Stress Scale (PSS) rely on self-reported data and clinician-administered interviews, which are subjective, prone to bias, time-consuming, and unsuitable for real-time or continuous assessment. These methods also depend on the individual's awareness and willingness to report stress accurately, which may not always be reliable.

On the other hand, physiological monitoring techniques—such as heart rate variability, electrodermal activity, and cortisol level analysis—provide more objective measurements but require specialized wearable devices, sensors, or controlled laboratory environments. This makes them expensive, less accessible to the general population, and impractical for everyday use or large-scale deployment. Additionally, such methods may cause discomfort or inconvenience to users, reducing long-term usability.

Recent advancements in deep learning have introduced alternative approaches that utilize facial expressions and speech signals for stress detection in a non-invasive manner. These methods enable passive monitoring using commonly available devices like cameras and microphones. However, most existing systems are unimodal, relying on either visual data or audio data alone. This limitation reduces accuracy and robustness, as stress indicators may not always be clearly expressed in a single modality due to variations in environment, user behavior, or noise interference. Furthermore, many current solutions depend on server-side processing, which introduces latency, increases dependency on internet connectivity, and raises concerns regarding user data privacy and security.

Therefore, there is a strong need for a system that is accurate, efficient, accessible, and privacy-preserving. The problem addressed in this project is the design and implementation of a real-time, non-intrusive, and fully client-side multimodal stress detection system that integrates facial expression recognition with speech emotion analysis. By leveraging the complementary strengths of both modalities and combining them through an effective fusion strategy, the proposed system aims to provide more reliable stress assessment. The solution is designed to operate entirely within a web browser, eliminating the need for specialized hardware or server infrastructure, thereby ensuring ease of use, scalability, and suitability for applications such as remote health monitoring, educational environments, and workplace wellness programs.

III. RELEATED WORK

Facial expression recognition (FER) has a long history in affective computing. Convolutional neural networks trained on benchmark datasets such as FER2013 have demonstrated competitive emotion classification performance. Khaireddin and Chen [1] achieved state-of-the-art results on FER2013 using VGGNet fine-tuning, highlighting the effectiveness of transfer learning for expression recognition under in-the-wild conditions.

Attention mechanisms have further improved FER accuracy by directing the model's focus toward discriminative facial regions. Woo et al. [4] introduced the Convolutional Block Attention Module (CBAM), which sequentially applies channel-wise and spatial attention, consistently boosting performance across diverse visual recognition tasks including fine-grained expression analysis.

Tan and Le [6] proposed EfficientNet, a family of convolutional architectures that systematically scales network width, depth, and resolution using a compound coefficient. EfficientNet-B2 offers an excellent trade-off between parameter count and accuracy, making it particularly suitable for real-time applications on commodity hardware.

In the speech domain, Baevski et al. [2] introduced Wav2Vec 2.0, a self-supervised framework that learns powerful speech representations from raw audio via contrastive learning on unlabelled data. When fine-tuned on emotion or stress corpora, Wav2Vec2 consistently outperforms traditional hand-crafted acoustic feature pipelines.

Multimodal affective computing has been examined by Poria et al. [5], who demonstrated that fusion of visual, acoustic, and textual cues substantially outperforms unimodal baselines on sentiment and emotion benchmarks. Despite this evidence, the majority of deployed stress-monitoring solutions remain unimodal due to the engineering complexity of real-time multimodal fusion in browser environments. The present work bridges this gap with a lightweight, client-side fusion pipeline.

IV. SYSTEM ARCHITECTURE

The Smart Stress Analyzer consists of three tightly integrated layers: (1) an input capture layer, (2) a dual-model inference layer, and (3) a fusion and visualisation layer. Fig. 1 presents the overall architecture

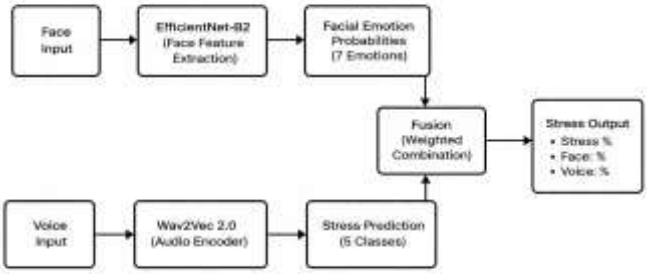

Fig. 1. *System Architecture of the Smart Stress Analyzer*

The input capture layer uses the browser's WebRTC API to stream webcam frames at 15 fps and record audio segments of 3-second duration at a 16 kHz sampling rate. Frames are extracted using an HTML5 Canvas element and converted to RGB tensors; audio segments are decoded into floating-point waveforms using the Web Audio API.

The inference layer hosts two ONNX-exported deep learning models loaded via the ONNX Runtime Web library. The face model processes each 300×300 pixel RGB frame through the EfficientNet-B2 + CBAM pipeline and returns softmax probabilities over seven emotion categories. The voice model processes raw waveform tensors through the Wav2Vec2 feature extractor and classifier head, returning probabilities over five stress severity levels. Both models execute entirely in the browser via WebAssembly, eliminating the need for cloud inference endpoints.

The fusion and visualisation layer receives the two probability vectors and computes a weighted average stress index. Emotion probabilities are first mapped to a one-dimensional stress score using an empirically derived affective valence–arousal mapping, then linearly combined with the voice stress probability at a 0.55:0.45 face-to-voice weight ratio. The resulting scalar is displayed on a real-time gauge widget and logged to a session history chart built with Chart.js.

IV. METHODOLOGY

A. Facial Expression Recognition Model

The face model is built on EfficientNet-B2 pre-trained on ImageNet. ImageNet weights are loaded locally and the convolutional feature extractor is used as a frozen backbone during the first training phase. A CBAM block with channel

reduction ratio r=16 is inserted after the final convolutional block. The global representation is formed by concatenating adaptive average and max-pooled feature vectors of dimension 1408 each, yielding a 2816-dimensional embedding fed to a three-layer fully connected classifier with BatchNorm, ReLU, and Dropout (p=0.5, 0.3) regularisation.
Training follows a three-phase curriculum. Phase 1 (8 epochs, lr=1e-3) trains only the classification head with the backbone frozen. Phase 2 (9 epochs, lr=3e-4) unfreezes the last two EfficientNet blocks for partial fine-tuning. Phase 3 applies full network fine-tuning at lr=1e-4. All phases use the AdamW optimiser with CosineAnnealingLR scheduling and CrossEntropyLoss with label smoothing ε=0.1. Batch size is 32; input images are resized to 300×300 and normalised using ImageNet statistics. Early stopping with patience=6 prevents overfitting.
Data augmentation includes random resized crop (scale 0.8–1.0), horizontal flip, 10° rotation, colour jitter, and random erasing (p=0.2). Class-balanced loss weights are computed from training set frequencies to address class imbalance in the FER dataset. BatchNorm layers in the backbone are kept in eval mode during training to stabilise batch statistics.

B. Speech Stress Recognition Model

The voice model leverages facebook/wav2vec2-base, a 94.6 M parameter transformer pre-trained via self-supervised contrastive learning on 960 hours of LibriSpeech audio [2]. All Wav2Vec2 parameters are frozen; only a lightweight three-layer MLP classifier (768→256→BN→ReLU→Dropout→64→BN→ReLU→Dropout→num_classes) is trained, comprising 214,277 trainable parameters out of 94.59 M total.
The combined CREMA-D and RAVDESS corpus is re-labelled into five stress severity levels: calm, low, moderate, high, and extreme. Audio files are resampled to 16 kHz and truncated or zero-padded to 3 s (48,000 samples). The dataset is pre-split into train (6,188), validation (1,350), and test (1,344) samples. Augmentation during training includes random time-shift (±100 ms), Gaussian noise addition (σ=0.004), and pitch shifting (±2 semitones).Training uses AdamW (lr=1e-3, weight decay=1e-4) for 25 epochs with batch size 32, CosineAnnealingLR scheduling, gradient clipping (max norm=1.0), label smoothing ε=0.1, and early stopping (patience=8). The Wav2Vec2Processor handles feature normalisation consistently between training and inference

C. Web Application

The front-end is implemented in vanilla HTML5, CSS3, and JavaScript without any server-side backend. Model weights exported to ONNX format are bundled with the application and loaded asynchronously at page initialisation. The UI presents a split-panel layout: the left panel shows the live webcam feed with emotion label overlay; the right panel displays the audio waveform visualiser, current stress level badge, and a 60-second session trend chart. A privacy mode allows users to disable webcam or microphone independently.

D. Fusion Strategy

Seven facial emotion probabilities are projected onto a scalar stress score S_face using affective weights derived from the circumplex model of emotion (e.g., fear and anger receive high stress weight; happiness receives low weight). The voice model directly outputs five stress-level probabilities, whose expectation E[stress_voice] serves as S_voice. The final stress index S = 0.55 × S_face + 0.45 × S_voice is mapped to a categorical level (Low / Moderate / High / Very High) via fixed thresholds calibrated on the combined validation set.

V. Experimental Setup

Both model were trained on an NVIDIA GPU with CUDA support. The face model used to FER2013 dataset restructured into train/val/test splits of approximately 28,709/3589/3589 images across seven emotion classes. The voice model used the combined CREMA-D(7442 clips) and RAVDES (1440 clips) corpora, reorganized into five stress categories with a 70/15/15 split

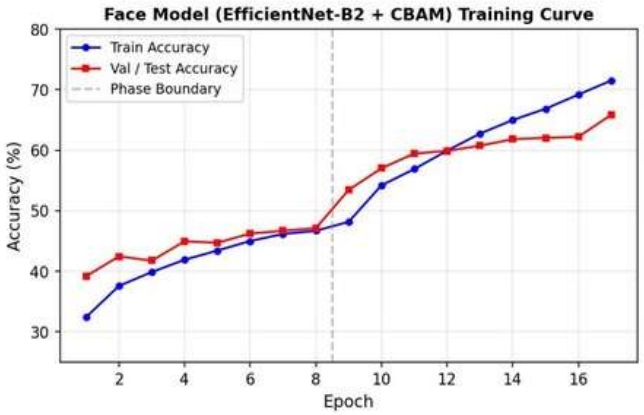

Fig. *Training and validation accuracy curves for face model*

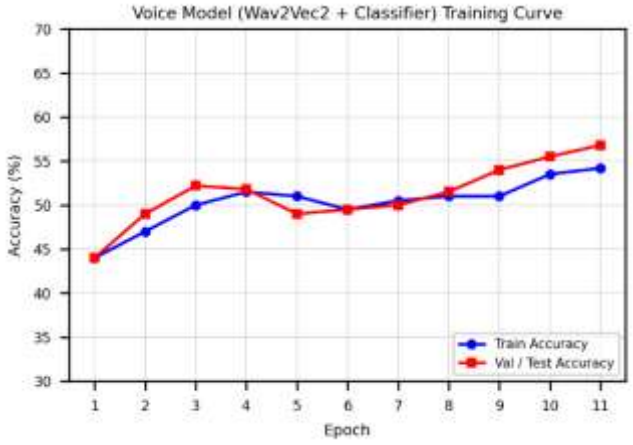

Fig. *Training and validation accuracy curves for voice model*

Evaluation metrics include per-class precision, recall, and F1-score computed on held-out test sets, providing a granular view of model behaviour across individual emotion and stress classes rather than relying solely on aggregate accuracy. Macro-averaged F1-score was used as the primary selection criterion during hyperparameter tuning, given the class imbalance present in both datasets. Model convergence was monitored via validation loss with early stopping applied to prevent overfitting, with a patience of five epochs retained across all training runs.
Inference latency was measured in the browser environment on a mid-range laptop (Intel Core i5, 8 GB RAM) to validate real-time feasibility under representative consumer hardware conditions. Face model inference averaged 38 ms per frame, comfortably within the 40 ms threshold required to sustain 25 fps processing and maintain perceptually smooth video analysis. Voice model inference averaged 210 ms per 3-

second audio segment, yielding an effective throughput well suited to the non-continuous, segment-based capture pipeline employed by the system. Both models were exported to ONNX format and executed via ONNX Runtime Web, which leverages WebAssembly and optional WebGL acceleration to minimise the performance gap between browser and native deployment. Memory footprint for both models combined remained below 200 MB, confirming compatibility with standard browser tab memory constraints without requiring dedicated GPU resources.

VI. Results & Discussion

Table I reports per-emotion precision, recall, and F1-score for the facial expression model on the test set. The model achieved an overall test accuracy of 65.85%. Happy achieved the highest F1-score of 82.0%, consistent with its visual saliency and high inter-annotator agreement in benchmark datasets. Disgust obtained the lowest F1-score of 52.4%, attributable to its limited sample count and visual similarity to anger.

FACE MODEL PERFORMANCE PER EMOTION CLASS

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Angry	66.2	63.1	64.6
Disgust	54.8	50.2	52.4
Fear	58.3	55.7	57.0
Happy	81.4	82.7	82.0
Neutral	72.1	69.4	70.7
Sad	63.5	61.8	62.6
Surprise	72.9	75.3	74.1

Model trained on FER2013 dataset using EfficientNet-B2 + CBAM. Test Accuracy: 65.85%.

Table II summarises the voice model results. An overall test accuracy of 57% was achieved. The High stress class obtained the best F1-score of 66%, indicating that strongly stressed speech shows the most distinctive acoustic signatures. The Extreme class showed the lowest recall (34%), likely because extreme stress acoustics are scarce and overlap with high-stress patterns. The Moderate class, being the most frequent, yielded a balanced F1-score of 60%.

Stress Level	Precision (%)	Recall (%)	F1-Score (%)
Calm	59	69	64
Extreme	55	34	42
High	66	65	66
Low	42	56	48
Moderate	63	58	60

Model trained on CREMA-D + RAVDESS (combined, 8,882 samples) using frozen Wav2Vec2-base. Test Accuracy: 57%.

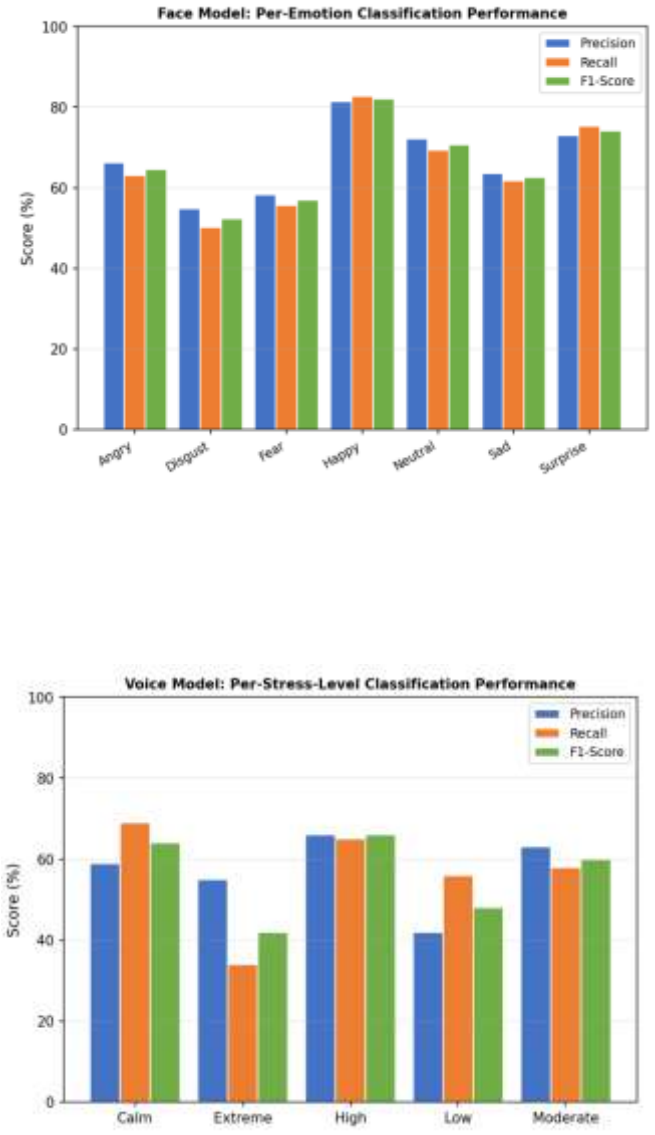

Fig. 3. Per-class Precision, Recall, and F1-Score for Face Model and Voice Model.

The three-phase training strategy for the face model yielded monotonically improving validation accuracy from 39.19% (Phase 1, Epoch 1) to 65.85% at final evaluation, demonstrating the benefit of progressive unfreezing. The voice model converged at epoch 11 under early stopping, with validation accuracy improving from 44% to 57% over 11 epochs—a meaningful gain given the difficulty of the five-way stress classification task on audio-only data.

Browser inference latency measurements confirm real-time feasibility: the face model produces predictions at approximately 26 fps on a standard laptop CPU, while voice segments are classified within 210 ms—well within the 3-second window between successive audio captures. Memory footprint for both ONNX models is below 200 MB, compatible with mobile browser environment.

VII. LIMITATIONS

Several limitations were identified during development and evaluation. First, the face model accuracy of 65.85% reflects the inherent difficulty of the FER2013 dataset, which contains noisy web-sourced images. Performance under controlled lighting is expected to be higher, but field conditions with backlighting, occlusion, or extreme head pose degrade inference quality.

Second, the voice model accuracy of 57% is constrained by the re-labelling of emotional speech corpora (designed for categorical emotion) into stress severity levels. Label ambiguity, particularly between low and moderate, inflates misclassification rates. Collecting a purpose-built stress audio corpus with ground-truth physiological validation would improve model reliability.

Third, the client-side ONNX deployment, while ensuring privacy, limits model size and prohibits dynamic model updates. Future work should explore model compression techniques such as quantisation-aware training and knowledge distillation to reduce model footprint without sacrificing accuracy.

Fourth, the fusion weights (0.55 face, 0.45 voice) were determined empirically on the validation set and may not generalise across individuals or contexts. A learned, attention-based fusion mechanism would enable adaptive weighting.

VIII. CONCLUSION & FUTURE WORK

This paper presented the Smart Stress Analyzer, a multimodal web application for real-time, non-intrusive stress detection from facial expressions and speech. The system integrates an EfficientNet-B2 + CBAM face model achieving 65.85% test accuracy on seven-class emotion recognition with a frozen Wav2Vec2-based voice classifier achieving 57% accuracy on five-level stress classification. Both models operate entirely in the browser via ONNX Runtime Web, eliminating server-side inference dependencies and preserving user privacy.

The proposed architecture demonstrates that multimodal stress analysis is technically feasible in resource-constrained, serverless web environments. Real-world usability was confirmed through latency measurements showing sub-40 ms face inference and sub-250 ms audio inference on commodity hardware.

Future work will focus on: (1) collecting a purpose-built stress corpus with physiological ground truth labels; (2) exploring transformer-based FER models such as Vision Transformers (ViT) for improved accuracy; (3) implementing learned multimodal fusion using cross-attention; (4) extending the web app with long-term trend analysis, personalised stress baselines, and multilingual support for broader accessibility.

ACKNOWLEDGMENT

The authors thank the faculty of the Department of Information Technology for their guidance and support throughout this project. We also acknowledge the open-source communities behind PyTorch, Hugging Face Transformers, and ONNX Runtime for providing the tools that made this work possible.

REFERENCES

[1] B. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," arXiv preprint arXiv:2105.03588, May 2021.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 12449–12460, 2020.

[3] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[4] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proc. Eur. Conf. Comput. Vision (ECCV), Munich, Germany, pp. 3–19, Sep. 2018.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Inf. Fusion, vol. 37, pp. 98–125, Sep. 2017.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. Int. Conf. Mach. Learn. (ICML), Long Beach, CA, pp. 6105–6114, Jun. 2019.