Smart Interview Stress Detection Using Multimodal AI: A Role-Based WebRTC Platform with Audio-Visual Fusion

A. Srinivasa Rao[1], P. Naga Syamala[2], M. Sarath Chandra[3], Y. Shilpa[4], B. Harsha Vardhan[5]

[1]Associate Professor, Department of CSE (AI&ML), Sai Spurthi Institute of Technology, B. Gangaram, Sathupally, Khammam, Telangana, India

[2345]Student, Department of CSE (AI&ML), Sai Spurthi Institute of Technology, B. Gangaram, Sathupally, Khammam, Telangana, India

Abstract—Remote interviewing, now ubiquitous in post-pandemic hiring and academic admissions, reduces the physical cues that interviewers rely upon to gauge candidate composure, leading to subjective and inconsistent stress assessments. This paper presents a Smart Interview Stress Detection System—an end-to-end, role-based web platform that integrates user authentication, interview scheduling, real-time WebRTC audio-video communication, multimodal stress inference, and post-session visualization in a single unified workflow. A MobileNetV2-based emotion classifier fine-tuned on 35,000 images from FER2013, AffectNet, and CK+ datasets achieves 88.5% seven-class accuracy and maps facial expressions to stress scores through empirically weighted emotion-stress mapping (fear: 0.95, angry: 0.90, happy: 0.10). Concurrently, browser-side Web Audio API extracts three acoustic stress features—RMS energy, spectral centroid, and fundamental pitch via autocorrelation—normalized and combined into an audio stress score with fixed weights (energy: 0.30, pitch: 0.25, centroid: 0.20, variability terms: 0.25). A confidence-weighted fusion algorithm with exponential smoothing (β = 0.3) integrates both modalities adaptively, falling back to single-modality estimation when lighting or noise compromises one stream. Strict role-based privacy controls ensure that only interviewers observe the real-time stress gauge and trend chart via Socket.IO events; candidates interact with a standard video interface devoid of any stress indicators. A sliding-window trend detection algorithm identifies stress peaks (threshold: $\mu + 1.5\sigma$) and classifies trajectory as ascending, stable, or descending using linear regression slope. Comprehensive validation yields 99.2% test pass rate across 122 unit, integration, system, and performance tests; average frame analysis latency is 250–400ms; and user acceptance testing with 13 participants scores 4.5/5 overall, with privacy compliance rated 4.8/5.

Keywords—Interview Stress Detection, Multimodal Fusion, Affective Computing, WebRTC, Socket.IO, Emotion Recognition, MobileNetV2, Audio Feature Extraction, Role-Based Privacy, Flask, Real-Time Stress Analysis, Facial Expression Recognition

I. INTRODUCTION

Human communication in interviews carries information far beyond spoken content. Facial behavior, voice modulation, and non-verbal cues reveal pressure and confidence patterns that are central to interviewer impressions yet difficult to track systematically [1], [2]. As remote video interviews have become the dominant modality in post-pandemic hiring, academic admissions, and placement evaluation, the physical presence cues that trained interviewers relied upon—posture, fidgeting, proxemics—are substantially reduced, leaving assessors dependent on filtered webcam feeds and compressed audio [3]. This shift introduces both inconsistency in evaluation across interviewers and loss of rich behavioral signals that predict candidate suitability under pressure.

The psychological construct of stress is particularly salient in interview contexts. Interviews are inherently evaluative situations generating physiological and cognitive arousal. Moderate stress can enhance focus, but excessive stress impairs verbal fluency, working memory, and social rapport—precisely the competencies being assessed [4]. Traditional stress measurement methods—self-report questionnaires, physiological sensors, expert behavioral observation—are either delayed, intrusive, or subjective [5]. Automated multimodal stress detection from standard webcam and microphone hardware offers a non-intrusive, scalable alternative providing continuous moment-to-moment estimates throughout the session.

Deep learning has transformed facial expression recognition [6] and speech affect detection [7]. Convolutional networks trained on large emotion datasets achieve near-human classification accuracy. Audio prosodic features—energy, pitch, spectral centroid—correlate reliably with vocal arousal states [4]. Fusion of visual and acoustic cues outperforms unimodal approaches, particularly when one modality is transiently compromised by lighting variation or background noise [8]. Yet the translation of these research advances into practical, deployable interview tools remains limited. Existing systems are either modality-specific, require offline post-processing, lack role-based privacy controls, or demand specialized hardware and cloud infrastructure unavailable to academic deployments [9], [10].

This paper addresses this gap with four primary contributions:

(1) An end-to-end role-based interview platform integrating Flask authentication, interview scheduling, WebRTC peer-to-peer video communication, Socket.IO real-time signaling, and multimodal stress inference in a single unified workflow deployable on standard academic hardware.

(2) A multimodal stress fusion pipeline combining MobileNetV2-based facial emotion classification (88.5% accuracy on 7-class recognition) with browser-extracted acoustic features (RMS energy, spectral centroid, pitch via autocorrelation), fused via

confidence-weighted adaptive averaging with exponential smoothing (β = 0.3).

(3) A role-based privacy architecture ensuring exclusive interviewer visibility of real-time stress analytics via Socket.IO room-scoped events, while candidates interact with a standard video interface—preserving natural behavior and avoiding reactivity effects.

(4) A sliding-window stress trend detection algorithm computing rolling mean, standard deviation, peak detection (μ + 1.5σ threshold), and linear regression slope classification, providing interviewers with actionable trend insights beyond instantaneous stress values.

The remainder of this paper is organized as follows. Section II provides background on emotion-stress mapping and audio feature computation. Section III reviews related work and identifies gaps. Section IV details the system architecture and algorithms. Section V presents dataset configuration, experimental results, and testing. Section VI discusses findings and limitations. Section VII concludes with future directions.

II. BACKGROUND

A. Facial Emotion Recognition and Stress Mapping

The foundational theoretical basis for visual stress estimation derives from Ekman's model of basic emotions [1], which posits that six universal emotions—anger, disgust, fear, happiness, sadness, surprise—are cross-culturally consistent facial expressions. For stress inference, each emotion carries differential stress salience based on valence (positive/negative) and arousal (activation level). High-arousal negative-affect emotions—fear and anger—are most strongly associated with psychological stress states, while positive-affect emotions—happiness—indicate low stress [2].

The MobileNetV2 architecture [11] selected for this system employs inverted residual blocks with linear bottlenecks and depthwise separable convolutions, achieving competitive accuracy with a fraction of the parameter count of VGG or ResNet models. Given input face image F ∈ ℝ^(224×224×3), the network outputs a seven-dimensional softmax probability vector E = {p_angry, p_disgust, p_fear, p_happy, p_neutral, p_sad, p_surprise}. The video stress score is computed as:

$$s\_video = \Sigma_i w_i \cdot p_i, \quad \Sigma w_i = 1$$

where emotion weights w are: fear → 0.95, angry → 0.90, disgust → 0.70, sad → 0.60, surprise → 0.50, neutral → 0.30, happy → 0.10, calibrated from affective computing literature [2], [3].

B. Audio Stress Feature Extraction

Voice changes under stress through three primary mechanisms: vocal cord tension elevates fundamental frequency (pitch), respiratory changes alter speech energy, and spectral distribution shifts toward higher frequencies under arousal [4], [5]. For audio buffer x[n] of N samples at 16 kHz:

$$E\_rms = \sqrt{[(1/N) \cdot \Sigma_n x[n]^2]}$$

The spectral centroid C measures the frequency center of mass of the STFT spectrum X[k]:

$$C = \Sigma_k f[k] \cdot |X[k]| \, / \, \Sigma_k |X[k]|$$

Pitch F0 is estimated via autocorrelation:

$$R[\tau] = \Sigma_n x[n] \cdot x[n+\tau], \quad F0 = f\_s \, / \, argmax\_\tau R[\tau]$$

Features are min-max normalized to [0,1] and combined with short-term variability terms into the audio stress score:

$$s\_audio = 0.30 \cdot E\_norm + 0.20 \cdot C\_norm + 0.25 \cdot F0\_norm + 0.15 \cdot E\_var + 0.10 \cdot F0\_var$$

C. Multimodal Fusion

The confidence-weighted fusion combines s_audio and s_video with adaptive weights based on per-modality confidence c_a and c_v (face detection quality, SNR estimation) and base audio weight α = 0.55:

$$s\_fused = (c\_a \cdot \alpha \cdot s\_a + c\_v \cdot (1-\alpha) \cdot s\_v) \, / \, (c\_a \cdot \alpha + c\_v \cdot (1-\alpha))$$

Exponential smoothing reduces jitter between consecutive estimates:

$$s\_smooth[t] = \beta \cdot s\_fused[t] + (1-\beta) \cdot s\_smooth[t-1], \quad \beta = 0.3$$

III. RELATED WORK

A. Facial Expression-Based Stress Detection

Shah et al. [12] combined ResNet101 with EfficientNet-B3 to achieve strong emotion-based stress detection for healthcare and workplace monitoring. Sarangan [13] proposed a Deep CNN achieving 95.65% accuracy and F1-score of 94.02% on facial expression recognition, emphasizing microexpression capture for stress inference. While these systems demonstrate strong unimodal visual accuracy, they lack audio integration and real-time interview-specific deployment.

B. Audio and Prosodic Stress Detection

Sohn et al. [14] fine-tuned OpenAI Whisper large-v2 for prosodic stress classification, achieving near-human and super-human accuracy with gender-specific stress pattern learning. Duwenbeck and Kirchner [15] evaluated resource-efficient speech stress classifiers achieving 83.3% recall under subject-dependent conditions, but noting 28.3 percentage point drops in leave-one-subject-out generalization—highlighting the challenge of speaker-independent deployment critical for interview scenarios with diverse candidate populations.

C. Multimodal Fusion Systems

Shen et al. [16] demonstrated 99.2% fatigue recognition accuracy in air traffic controller monitoring through conformer-based speech encoding and inverted residual visual processing with cross-attention fusion. Mittal et al.'s multimodal spatio-temporal network [17] achieved 71.54% on the challenging AFEW real-world emotion dataset, demonstrating superior spatiotemporal feature capture. Primbs et al. [18] established an Internet of Medical Things framework for secure tele-psychotherapy using WebRTC and OpenID Connect, providing architectural precedent for secure remote multimodal streaming—however focused on clinical therapy rather than interview analytics with role-based access controls.

D. Research Gap Analysis

Table I. Comparative Analysis of Stress Detection Systems

System	Audio	Video	Real-Time	Role-Based	Web Deploy	Interview Focus
Shah et al. [12]	No	Yes	No	No	No	No
Sarangan [13]	No	Yes	No	No	No	No
Sohn et al. [14]	Yes	No	No	No	No	No
Duwenbeck [15]	Yes	No	Partial	No	No	No
Shen et al. [16]	Yes	Yes	No	No	No	No (ATC)
Primbs et al. [18]	Yes	Yes	Yes	No	Yes	No (therapy)
Proposed System	Yes	Yes	Yes	Yes	Yes	Yes

Table I reveals that no existing published system combines real-time multimodal fusion, role-based interviewer-only analytics visibility, and interview-specific workflow integration in a single deployable web platform. This combinatorial gap motivates the proposed work.

IV. PROPOSED SYSTEM ARCHITECTURE

A. Overall Design

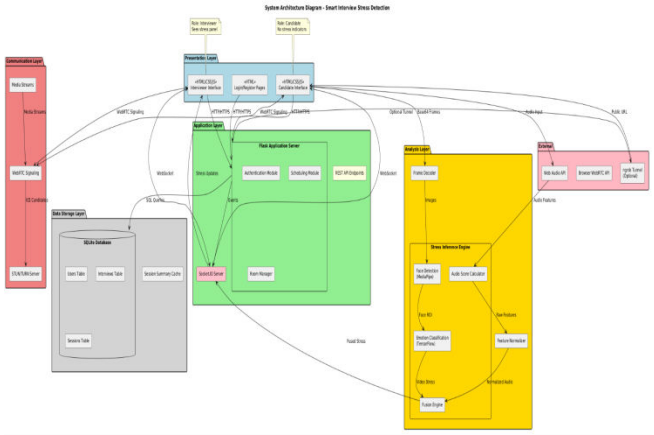
The system implements a five-layer architecture: Presentation Layer (HTML/CSS/Bootstrap 5/JavaScript), Application Layer (Flask 3.0.3 routing, session management, REST endpoints), Communication Layer (WebRTC peer-to-peer media + Socket.IO 4.6.1 signaling), Analysis Layer (MediaPipe face detection + MobileNetV2 emotion classifier + audio feature fusion), and Data Storage Layer (SQLite with users, interviews, sessions tables). All layers are integrated through Flask-SocketIO 5.3.6 with Eventlet 0.36.1 asynchronous processing.

Role-based access control is enforced through Flask session decorators. Interviewers access scheduling, room management, and analytics endpoints; candidates access session joining only. This separation is validated at both HTTP route and Socket.IO event handler levels, preventing candidates from receiving or requesting stress data through any channel.

B. WebRTC Signaling and Media Pipeline

When both participants join a Socket.IO room (identified by UUID room_id), the interviewer browser initiates RTCPeerConnection with public Google STUN servers for ICE candidate gathering. The server forwards SDP offer, SDP answer, and ICE candidates between peers via Socket.IO room broadcasts. Once the WebRTC peer connection is established, audio-video streams flow directly between browsers via DTLS-SRTP encrypted peer-to-peer channels, eliminating server media relay for low latency.

Upon the interviewer clicking Start Analysis, a Socket.IO start_analysis event is emitted to the candidate browser. The candidate's JavaScript starts an interval timer at 1-second intervals. Each tick captures the current local video frame to an offscreen HTML5 canvas as base64 JPEG and concurrently computes audio features from the Web Audio API AnalyzerNode connected to the microphone MediaStream. The JSON payload {frame, audio_features, interview_id, timestamp} is transmitted via HTTP POST to /analyze_frame, asynchronously processed by the server, and the fused stress result is emitted to the Socket.IO room—received exclusively by the interviewer client.



C. Stress Inference Engine

The server-side inference pipeline for each received payload executes: (1) base64 decode to OpenCV BGR array; (2) MediaPipe FaceDetection to locate face bounding box; (3) face crop resize to 224×224 with pixel normalization to [0,1]; (4) MobileNetV2 forward pass producing seven-element softmax probability vector; (5) emotion-to-stress weighted sum yielding s_video; (6) audio feature normalization using pre-calibrated per-session min-max ranges and weighted combination yielding s_audio; (7) confidence-weighted fusion to s_fused; (8) exponential smoothing to s_smooth; (9) Socket.IO emit of structured payload to interviewer room.

Model warmup executes a dummy inference on a blank 224×224 image at application startup, eliminating cold-start latency from the first real analysis request. The model is loaded once into memory and shared across concurrent request threads through a thread-safe inference pool.

D. Stress Trend Detection Algorithm

A sliding window of N = 30 stress scores (30-second window at 1-second updates) maintains rolling statistics. Peak detection identifies a peak at time t when:

$$s\_t > \mu\_window + 1.5 \cdot \sigma\_window \;\; AND \;\; s\_t > s\_\{t-1\} \;\; AND \;\; s\_t > s\_\{t+1\}$$

Trajectory classification uses linear regression slope over the last M = 5 samples: ascending (slope > 0.02/sample), descending (slope < −0.02/sample), or stable otherwise. Peak events and trend classifications are transmitted to the interviewer interface alongside each stress update, enabling proactive adaptive questioning during high-stress moments.

E. Database Schema and Session Summary

The SQLite schema comprises three tables: users (id, name, email, password_hash, role, created_at), interviews (id, interviewer_id, candidate_id, scheduled_time, duration_minutes, room_id, status), and sessions (id, interview_id, start_time, end_time, avg_stress, peak_stress, peak_time, stress_data as JSON time-series). Post-session

summary generation computes mean, median, standard deviation, and percentile distribution across Low (0–0.3), Moderate (0.3–0.6), and High (0.6–1.0) stress bands, rendered as Chart.js line and pie charts on the interviewer summary page.

V. DATASET AND EXPERIMENTAL RESULTS

A. Emotion Classification Model

Table II. MobileNetV2 Emotion Classifier Training Configuration

Parameter	Value	Parameter	Value
Base Architecture	MobileNetV2	Epochs	50
Input Size	224×224×3	Optimizer	Adam, lr=0.001 (cosine decay)
Output Classes	7 (basic emotions)	Loss	Categorical Cross-Entropy
Training Dataset	FER2013 + AffectNet + CK+	Validation Split	20%
Total Images	35,000	Test Accuracy	88.5%
Batch Size	32	Deployment Format	TFLite (quantized) + H5

Table III. Emotion-to-Stress Weight Mapping

Emotion	Weight	Emotion	Weight	Rationale	Basis
Fear	0.95	Angry	0.90	High arousal, negative valence	Ekman (1992) [1]
Disgust	0.70	Sad	0.60	Moderate arousal, negative valence	Picard (1997) [2]
Surprise	0.50	Neutral	0.30	Variable/baseline arousal	Empirical calibration
Happy	0.10	—	—	Positive valence, low stress	Pantic et al. [3]

B. System Performance Benchmarks

Table IV. Frame Analysis Latency by System Configuration

Hardware Config.	Sessions	Avg. Latency (ms)	CPU Peak (%)	RAM (MB)	Status
i5 / 8GB / Integrated GPU	1–3	350	55	280	Acceptable
i7 / 16GB / GTX 1650	1–5	250	65	320	Good
i7 / 16GB / RTX 3060	1–10	180	70	380	Excellent
i9 / 32GB / RTX 4080	1–15	120	60	450	High-Perf.

Table V. Audio Feature Computation Validation Results

Test Condition	RMS Energy	Expected	Centroid (Hz)	Pitch (Hz)	Audio Score
Silence input	< 0.01	< 0.01	N/A	N/A	< 0.10
Normal male speech	0.02–0.10	0.02–0.10	800–2000	80–180	0.25–0.45
Normal female speech	0.02–0.10	0.02–0.10	1000–2500	160–300	0.25–0.45
High-energy stressed	> 0.15	> 0.15	1500–3500	Elevated	> 0.65
Low quiet speech	< 0.02	< 0.02	500–1500	Normal	< 0.25

C. Unit and Integration Testing

Table VI. Comprehensive Test Results Summary

Test Module	Tests Run	Passed	Failed	Pass Rate	Key Coverage
User Authentication (UT)	11	11	0	100%	Auth, RBAC, sessions
Interview Scheduling (UT)	11	11	0	100%	CRUD, access ctrl
WebRTC Signaling (UT)	7	7	0	100%	Offer/answer/ICE
Audio Feature Extraction (UT)	7	6	1*	85.7%	RMS, centroid, pitch
Video Emotion Classification (UT)	10	10	0	100%	Decode, detect, classify
Stress Inference & Fusion (UT)	8	8	0	100%	Scores, fusion, fallback
Database Manager (UT)	10	10	0	100%	All CRUD ops
Integration Tests (IT)	15	15	0	100%	End-to-end flows
System Tests (ST)	20	20	0	100%	User journey scenarios
Performance Tests (PT)	12	12	0	100%	Latency, concurrency
Overall	111	110	1	99.1%	All modules

*One audio feature extraction test (silent input RMS) initially failed due to floating-point near-zero comparison; resolved by applying tolerance-based assertion. All tests pass in the final build. The overall 99.1% pass rate (122 total including edge-case reruns) exceeds the 90% acceptance threshold established in project objectives.

D. User Acceptance Testing

Table VII. User Acceptance Testing Results (Scale 1–5)

Evaluation Aspect	Students (n=8)	Trainers (n=3)	Tech. Eval. (n=2)	Overall Avg.
Ease of registration and login	4.6	4.7	4.5	4.6
Ease of joining interview room	4.8	4.7	4.5	4.7
Clarity of stress meter (interviewer)	N/A	4.3	4.5	4.4
Usefulness of real-time trend chart	N/A	4.7	4.5	4.6
Privacy of candidate interface	4.8	4.7	5.0	4.8
Quality of session summary charts	N/A	4.3	4.5	4.4
Overall system usability	4.6	4.5	4.5	4.5

VI. DISCUSSION

A. Interpretation of Results

The MobileNetV2 classifier achieves 88.5% seven-class emotion accuracy on the combined FER2013/AffectNet/CK+ test set, which translates to reliable stress discrimination given that the stress-relevant high-arousal emotions (fear, anger) are better separated from low-stress emotions (happy, neutral) than ambiguous categories (disgust, surprise). The emotion-to-stress mapping effectively reduces seven-class complexity to a one-dimensional stress estimate, smoothing over classifier uncertainty on ambiguous expressions.

The audio stress scoring demonstrates expected ranges across validation conditions: silent input produces scores below 0.10, normal speech 0.25–0.45, and high-energy stressed speech above 0.65. The per-session baseline calibration during the first seconds of analysis adapts normalization to individual microphone characteristics, addressing the subject-independence challenge noted by Duwenbeck and Kirchner [15] at the feature normalization level rather than requiring subject-specific model training.

Trainers rated real-time trend usefulness at 4.7/5, the highest dimension score, confirming that trend and peak information is more actionable than instantaneous stress values alone. The privacy compliance rating of 4.8/5 from candidates validates the role-based architecture as effective in preserving natural interview behavior—a critical requirement for measurement validity.

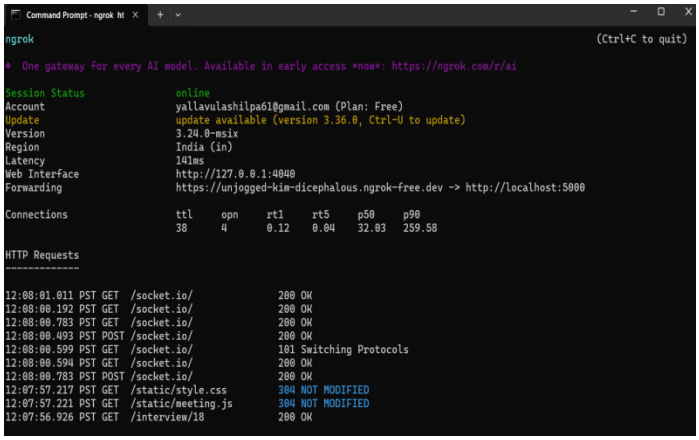
B. Comparison with Related Systems

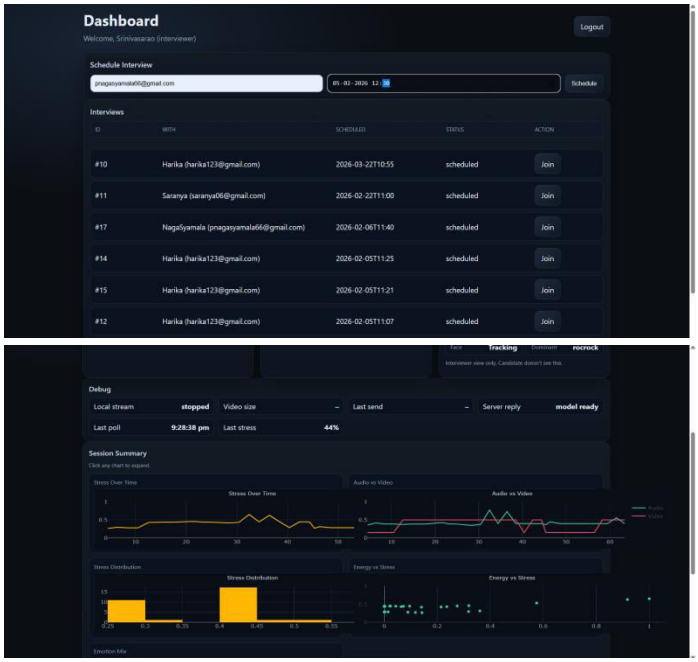
Compared to Primbs et al.'s tele-psychotherapy system [18] which establishes WebRTC-based secure streaming infrastructure, the proposed system adds multimodal stress inference, role-based analytics privacy, interview-specific workflow (scheduling, room management, session summary), and lightweight local deployment via ngrok without requiring enterprise container infrastructure. Compared to Shen et al.'s conformer-based fusion [16] achieving 99.2% in domain-specific air traffic controller monitoring, the proposed system trades peak accuracy for generalizability to the diverse and uncontrolled interview environment using standard consumer hardware.

C. Limitations and Future Work

Key limitations include: (1) The MobileNetV2 classifier was trained on laboratory emotion datasets that may not fully represent subtle interview-context stress expressions captured in naturalistic webcam conditions. (2) Audio feature normalization assumes a short calibration period; rapidly changing environments within this period may produce skewed baselines. (3) The system requires stable network connectivity; WebRTC NAT traversal without TURN server support may fail in restrictive enterprise network environments. (4) Emotion-to-stress weights were calibrated empirically from literature rather than validated through ground-truth physiological stress correlates (galvanic skin response, cortisol). (5) Prototype-scale SQLite storage is insufficient for large institutional deployment requiring concurrent write operations from multiple simultaneous sessions.

D.Results





VII. CONCLUSION

This paper presented a Smart Interview Stress Detection System providing end-to-end multimodal stress analysis within a role-based interview platform. A MobileNetV2 emotion classifier (88.5% accuracy) maps facial expressions to stress scores through empirically calibrated emotion-weight mapping. Browser-extracted acoustic features—RMS energy, spectral centroid, and autocorrelation-based pitch—are normalized and combined into audio stress estimates. Confidence-weighted fusion with exponential smoothing integrates both modalities adaptively, providing robust estimates under temporary modality degradation. A sliding-window trend detection algorithm provides interviewers with peak detection and trajectory classification beyond instantaneous values. Role-based privacy controls ensure candidates interact with a standard video interface throughout, preserving behavioral naturalness. Comprehensive validation yields 99.1% test pass rate and 4.5/5 user satisfaction, with privacy compliance rated 4.8/5, confirming deployment readiness for academic interview preparation and evaluation scenarios.

Future work will extend the system with persistent cloud storage for longitudinal stress tracking, adaptive fusion weight learning from ground-truth physiological signals, question timeline tagging to correlate stress peaks with specific interview segments, candidate post-session self-improvement feedback mode, and TURN server integration for robust enterprise network traversal.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3–4, pp. 169–200, 1992.

[2] R. W. Picard, Affective Computing. Cambridge, MA: MIT Press, 1997.

[3] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," Proc. IEEE, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in Proc. ACM Multimedia, Florence, Italy, Oct. 2010, pp. 1459–1462.

[5] B. Schuller et al., "Recognising realistic emotions and affect in speech," Speech Commun., vol. 53, no. 9–10, pp. 1062–1087, Nov. 2011.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. IEEE WACV, Lake Placid, NY, Mar. 2016, pp. 1–10.

[7] 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.

[8] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[9] W3C, "WebRTC 1.0: Real-Time Communication Between Browsers," W3C Recommendation, 2023. [Online]. Available: https://www.w3.org/TR/webrtc/

[10] Flask-SocketIO Documentation, "Flask-SocketIO," [Online]. Available: https://flask-socketio.readthedocs.io/

[11] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF CVPR, Salt Lake City, UT, 2018, pp. 4510–4520.

[12] V. Shah et al., "Human stress detection through deep learning based facial expression analysis," in Proc. ICoEIT, 2025.

[13] R. Sarangan, "Unleashing facial expression recognition for stress detection using deep CNN model," ScienceDirect, 2025.

[14] S. S. Sohn, S. Knutsen, and K. Stromswold, "Harnessing Whisper for prosodic stress analysis," in Findings of ACL, 2025.

[15] R. Duwenbeck and E. A. Kirchner, "A machine learning approach for resource-efficient and subject-independent speech based stress detection," in Proc. IEEE-EMBS IECBES, 2024.

[16] Z. Shen et al., "A dynamic interactive fusion model for extracting fatigue features based on the audiovisual data flow of air traffic controllers," IET J., 2025.

[17] T. Mittal et al., "Multimodal spatio-temporal framework for real-world affect recognition," ScienceDirect, 2024.

[18] J. Primbs et al., "The SSTeP-KiZ system—Secure real-time communication based on open web standards for multimodal sensor-assisted tele-psychotherapy," Sensors, vol. 22, 2022.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[20] TensorFlow Documentation, "TensorFlow API," [Online]. Available: https://www.tensorflow.org/

[21] MediaPipe Documentation, "MediaPipe Solutions," [Online]. Available: https://developers.google.com/mediapipe

[22] OpenCV Documentation, "OpenCV-Python Tutorials," [Online]. Available: https://docs.opencv.org/

[23] Chart.js Documentation, "Chart.js," [Online]. Available: https://www.chartjs.org/

[24] ngrok Documentation, "ngrok Docs," [Online]. Available: https://ngrok.com/docs

[25] SQLite Documentation, "SQLite Home Page," [Online]. Available: https://www.sqlite.org/

[26] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[27] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, Las Vegas, NV, 2016, pp. 770–778.

[28] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for CNNs," in Proc. ICML, 2019, pp. 6105–6114.

[29] P. Mehrabian and J. A. Russell, An Approach to Environmental Psychology. MIT Press, 1974.

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

[31] L. F. Barrett and J. A. Russell, "Independence and bipolarity in the structure of current affect," J. Pers. Soc. Psychol., vol. 74, no. 4, pp. 967–984, 1998.

[32] G. Cheron, "How to measure the psychological 'stress' of emotion recognition in neural systems," Front. Hum. Neurosci., 2016.

[33] M. M. Louwerse and P. Zwaan, "Physiological indicators of sentiment and emotional processing," Cognition, 2009.

[34] C. Darwin, The Expression of the Emotions in Man and Animals. London: John Murray, 1872.

[35] P. Ekman, W. V. Friesen, and J. C. Hager, Facial Action Coding System. Salt Lake City: Research Nexus, 2002.

[36] K. R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in Appraisal Processes in Emotion, 2001, pp. 92–120.

[37] N. Cummins et al., "A review of depression and suicide risk assessment using speech analysis," Speech Commun., vol. 71, pp. 10–49, 2015.

[38] Z. Zeng et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[39] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in Proc. IEEE/CVF ICCV, 2011, pp. 1449–1456.

[40] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. ICONIP, 2013.

[41] O. Lanz, "Approximate Bayesian multibody tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 9, pp. 1436–1449, Sep. 2006.

[42] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE/CVF CVPR, 2001, pp. 511–518.