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

InterviewShield: A Multimodal AI Framework for Integrity Monitoring in Remote Technical Interviews

Sripradha HN, Yashraj KS, Varshitha H, Vaishnavi D C

Department of Computer Science and Engineering, JNN College of Engineering, Shivamogga, Karnataka, India

Under the guidance of **Mrs. Ayesha Siddiqa** | ayeshasiddiqa@jnnce.ac.in

---

Abstract - Remote technical interviews improve accessibility and reduce recruitment effort, but they also create integrity risks that are difficult to observe from the interview conversation alone. Candidates may be impersonated, receive assistance, leave the interview context, or use external resources not visible to the interviewer. InterviewShield proposes a modular multimodal framework combining computer-vision, audio, behavioral, and browser/system signals into an interpretable risk representation. Frame-level perception is separated from temporal behavior analysis, while a fusion layer aggregates heterogeneous evidence using confidence and time context. The candidate interface includes explicit consent and a pre-interview device check; the recruiter interface presents risk summaries and timestamped supporting evidence. Engineering targets include alert latency below two seconds, at least 80% module-level coverage, and at least 85% fusion-engine coverage. These remain targets rather than verified outcomes, and no numerical accuracy is claimed without a labeled validation set. Prototype testing establishes the modular workflow and identifies audio-analysis reliability as an open validation issue. InterviewShield is intended to support human review, not to make automated hiring or misconduct decisions.

Key Words: Online Interview Integrity, Multimodal AI, Computer Vision, Audio Analysis, Behavioral Monitoring, Risk Scoring.

1. INTRODUCTION

Remote technical interviews have become an important part of modern recruitment because candidates and organizations can participate without sharing a physical location. The same flexibility changes the integrity problem. An interviewer may not be able to determine from the visible stream whether the displayed candidate is the authorized person, whether another person is assisting, whether the candidate repeatedly leaves the interview context, or whether suspicious activity occurs outside the camera view.

Research on AI-assisted interview evaluation demonstrates that facial and speech signals can estimate emotion, confidence, and other performance-related attributes. Gupta et al. proposed a mock interview evaluator using facial-expression analysis and speech features, while Rai et al. described an evaluator combining facial and speech-based emotion analysis [1], [2]. These systems demonstrate multimodal sensing, but focus on candidate development rather than interview integrity.

Remote-proctoring research provides a closer foundation. Atoum et al. combined visual and acoustic observations with user verification, active-window, gaze, speech, and phone-related cues [3]. Özgen et al. proposed a cheating-analysis pipeline using face detection, recognition, object detection, and tracking [4]. Xue et al. combined target detection, tracking, and audio analysis for intelligent invigilation [5]. AutoOEP extends multimodal monitoring with identity verification, gaze and head-pose analysis, prohibited-object detection, and temporal modeling [6].

InterviewShield addresses this gap by correlating browser events and temporal behavioral patterns with visual and acoustic evidence. A single tab switch, gaze deviation, or speech anomaly is not treated as proof of misconduct. Instead, the system collects heterogeneous evidence, assigns confidence, aggregates information over time, and presents a risk-oriented summary to a human reviewer. The central research question is whether such a modular architecture can produce a real-time, interpretable risk representation for remote technical interviews.

2. PROBLEM STATEMENT AND OBJECTIVES

Conventional remote interview platforms may provide video communication, question delivery, recording, or answer evaluation without a unified integrity layer correlating contextual signals. An isolated event is ambiguous: a tab change may be accidental, looking away may be normal thinking behavior, and speech variation

may result from microphone quality or background noise. The challenge is to organize heterogeneous observations into interpretable evidence while preserving uncertainty.

The objectives are to: (1) verify candidate presence and identity-related visual consistency; (2) detect multiple-person presence and suspicious visual patterns; (3) analyze speech and speaker consistency; (4) interpret observations as temporal behavioral patterns; (5) capture browser-level focus and tab events within browser capabilities; (6) combine available signals into Low, Medium, and High risk bands; (7) provide live alerts and historical evidence to recruiters; (8) incorporate explicit consent and privacy information; and (9) evaluate the implementation through unit, integration, end-to-end, performance, security, and edge-case testing.

3. LITERATURE SURVEY

The literature falls into three related directions. First, AI-based interview evaluation combines facial-expression analysis and speech processing to estimate emotion and confidence [1], [2]. Such outputs describe candidate performance rather than the trustworthiness of the interview context.

Second, automated proctoring and cheating detection combine multiple visual and acoustic cues with temporal information [3]–[6]. These studies show how combining modalities can reduce the limitations of a single detector. However, a technical-interview setting also benefits from correlating browser events and temporal behavior with visual and acoustic evidence.

Third, privacy, fairness, and security research highlights the trade-off between detection capability and users' willingness to share visual information. Mukherjee et al. studied privacy-utility and fairness perceptions of visual data in remote proctoring [7]. Burgess et al. analyzed integrity, procedural fairness, security, and privacy risks in remote-proctoring systems [8]. These findings support a human-review model, explicit consent, confidence-aware evidence, and clear limits on browser and sensor capabilities.

4. PROPOSED SYSTEM

InterviewShield consists of four monitoring engines, a fusion and risk engine, a backend event layer, and candidate/recruiter interfaces. Candidate-side media and browser events are processed by the relevant engines; outputs are normalized and forwarded to the fusion layer. The resulting risk representation and evidence are persisted through backend services, and the recruiter dashboard receives live updates. Each engine can be developed, tested, or replaced independently.

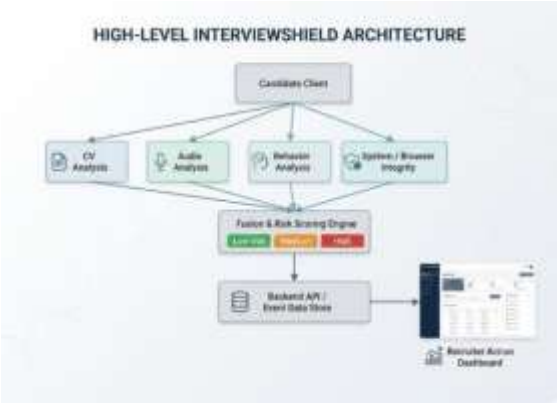


Fig. 1: High-level InterviewShield architecture.

5. METHODOLOGY

A. Computer Vision Engine — The computer-vision (CV) engine is designed for face detection, person counting, identity verification, gaze estimation, attention scoring, and movement-anomaly signals. Face detection produces candidate-presence observations and confidence values; person counting identifies multiple-person conditions. Identity verification compares live observations with a reference captured at session start. Gaze and attention components estimate sustained off-screen behavior, while movement analysis identifies abnormal motion patterns. Frame-level perception is separated from temporal behavior analysis.

B. Audio Analysis Engine — The audio engine is structured around speech-to-text, voice matching, multi-speaker detection, and anomaly scoring. Speech is represented as timestamped segments; voice matching estimates consistency with the session-start speaker; multi-speaker analysis looks for evidence of an additional speaker. Microphone dropouts, background noise, accents, and low-volume speech should reduce confidence rather than automatically increase fraud risk.

C. Behavioral Monitoring Engine — Behavioral monitoring operates over time windows. It maintains a documented pattern library, derives focus trends, and combines repeated patterns with attention information. Repeated gaze deviation together with a hand-to-ear movement may be treated as a possible coaching-related event, but remains an evidence signal rather than a definitive accusation. Temporal analysis is important because individual actions may be normal in isolation.

D. System/Browser Monitoring Engine — The system layer records events observable through browser APIs, including visibility or focus changes, tab switching, and

clipboard interactions. To protect privacy, clipboard monitoring transmits metadata such as content length rather than raw clipboard contents. A normal web application cannot reliably inspect arbitrary operating-system processes or background applications; stronger native monitoring would require a separately permissioned native agent.

E. Candidate and Recruiter Interfaces — The candidate flow begins with an explicit consent and privacy screen followed by a readiness check for camera, microphone, and connectivity. During the interview, events are streamed to the appropriate engines. The recruiter view presents live risk summaries, alert timelines, and timestamped evidence so reviewers can inspect why a score changed.

6. MULTIMODAL FUSION AND RISK SCORING

An individual event should not automatically become a fraud verdict. Each engine produces normalized evidence with confidence and time information. The fusion layer aligns asynchronous event streams into common time windows and combines available signals. The system can continue operating when a signal is unavailable, while making the missing signal explicit.

$$\mathbf{R(t) = D(\Sigma_i\, w_i \cdot s_i(t) \cdot c_i(t))}$$

Here, $s_i(t)$ is the normalized signal from engine i, $c_i(t)$ is its confidence, w_i is the configured weight, and D represents temporal aggregation and decay. This is a design model, not a claim of an empirically optimal classifier. Weights and thresholds should be calibrated using labeled validation data before consequential deployment. Scores are mapped to Low, Medium, and High bands and accompanied by timestamped explanations, such as a second-face event, repeated focus loss, or a low-confidence audio condition.

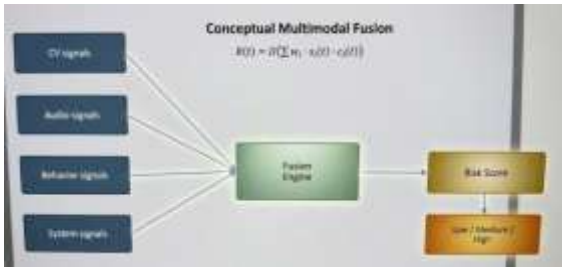


Fig. 2: Conceptual multimodal fusion.

7. RESULTS AND EVALUATION

Evaluation distinguishes measured results from engineering targets. Because a representative labeled interview dataset is not yet available, this paper does not report fabricated accuracy, precision, recall, F1-score, or false-positive rates. The prototype supports independent CV, audio, behavioral, and browser/system pipelines and provides a defined path toward multimodal fusion. Browser-event monitoring is suitable for deterministic functional tests because focus loss and tab switching have observable triggers. CV and behavioral modules can be validated with controlled clean and suspicious clips. Audio analysis remains validation-sensitive due to microphone conditions, background noise, speaker variation, and feature calibration.

Evaluation item	Target / criterion	Status
Module test coverage	≥80%	Engineering target; not claimed achieved
Fusion-engine coverage	≥85%	Engineering target; not claimed achieved
Alert latency	<2 seconds	Requires benchmark evidence
Missing-signal handling	No crash; low-confidence state	Design requirement
Browser event detection	Correct event generation	Controlled functional test
Risk explainability	Timestamped evidence for elevated score	Acceptance criterion
Aggregate accuracy	Precision/recall/F1 on labeled data	Deferred pending dataset

The targets are drawn from the engineering roadmap and are not presented as achieved results. The next experimental stage should construct labeled clean, single-anomaly, and multi-anomaly interview sessions and report precision, recall, false-positive rate, false-negative rate, alert latency, throughput, and resource use for each module and for the fused system.



Fig. 3: End-to-end evaluation workflow.

8. SECURITY, PRIVACY AND HUMAN REVIEW

InterviewShield processes face, voice, behavioral, and browser information, so security and privacy are core requirements. The design includes authenticated role separation for candidates, recruiters, and administrators; protected access to stored evidence; audit logging; and explicit data-retention handling. Browser monitoring is restricted to signals observable within the web security model. Clipboard monitoring avoids storing clipboard contents.

The system is human-in-the-loop. A high risk score is a prioritization signal for review, not an automatic rejection decision. Normal interview behavior, accessibility needs, camera quality, network instability, accent variation, and background noise can resemble anomalies. Transparency, user consent, and careful validation are essential before use in high-stakes hiring contexts [7], [8].

9. LIMITATIONS AND FUTURE SCOPE

Several limitations remain. Browser APIs cannot provide unrestricted visibility into operating-system processes, so claims about background applications or second monitors must remain bounded by browser capabilities. Audio analysis is sensitive to microphones, room acoustics, noise, accents, and speaker overlap. Visual models may degrade under low light, occlusion, camera movement, or poor resolution. Multimodal fusion can reduce reliance on a single noisy signal but cannot eliminate false positives. Robust validation requires representative labeled interview sessions and calibration across devices, environments, and candidate behaviors.

Future work includes stronger temporal models for behavioral analysis, improved speaker diarization and voice embeddings, liveness and anti-spoofing checks, richer browser-integrity signals, optional native monitoring with explicit permission, adaptive confidence thresholds based on device quality, privacy-preserving evidence storage, explainable alert timelines, and large-scale validation across diverse interview conditions. Fairness-aware calibration is particularly important to avoid systematically penalizing benign behavioral or acoustic variation.

10. CONCLUSIONS

InterviewShield presents a multimodal framework for supporting integrity in remote technical interviews. Its contribution is a structured architecture that combines visual, audio, temporal behavioral, and browser-level evidence into an interpretable risk representation rather than a claim of perfect automated fraud detection. The modular design supports independent development and testing, while the fusion layer provides confidence-aware aggregation and temporal context. The system is intended to help recruiters identify sessions or time intervals that deserve closer review; it is not intended to make irreversible hiring or misconduct decisions from a single model output. Consent, privacy safeguards, browser limitations, missing-signal handling, and an honest evaluation strategy provide a foundation for further empirical research and controlled validation.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to the Department of Computer Science and Engineering, JNN College of Engineering, Shivamogga, for providing the academic environment, resources, and encouragement required to undertake this project. The authors also thank their faculty members, peers, and families for their guidance and support throughout the development and evaluation of InterviewShield.

REFERENCES

[1] N. Gupta, A. Sharma, S. Shukla, and P. Kumar, “AI-based mock interview evaluator: An emotion and confidence classifier model,” International Journal of Research in Engineering and Innovation, vol. 9, no. 3, pp. 136–142, 2025, doi: 10.36037/IJREI.2025.9308.

[2] N. S. Rai, A. K. R, A. P, and H. N. R, “AI Based Interview Evaluator: An Emotion and Confidence Classifier,” International Advanced Research Journal in Science, Engineering and Technology, vol. 11, no. 4, 2024, doi: 10.17148/IARJSET.2024.11442.

[3] Y. Atoum, L. Chen, A. X. Liu, S. D. H. Hsu, and X. Liu, “Automated Online Exam Proctoring,” IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1609–1624, 2017, doi: 10.1109/TMM.2017.2656064.

[4] A. C. Özgen, M. U. Öztürk, and U. Bayraktar, “Cheating Detection Pipeline for Online Interviews and Exams,” arXiv:2106.14483, 2021.

[5] J. Xue, W. Wu, and Q. Cheng, “Intelligent invigilator system based on target detection,” Multimedia Tools and Applications, vol. 82, pp. 44673–44695, 2023, doi: 10.1007/s11042-023-15474-y.

[6] A. K. Naveen, B. Singla, R. Wankhade, S. M., R. S., and R. M. R. Guddeti, “AutoOEP—A Multi-modal Framework for Online Exam Proctoring,” arXiv:2509.10887, 2025.

[7] S. Mukherjee, V. Distler, G. Lenzini, and P. Cardoso-Leite, “Balancing the Perception of Cheating Detection, Privacy and Fairness: A Mixed-Methods Study of Visual Data Obfuscation in Remote Proctoring,” in Proceedings of the 2024 European Symposium on Usable Security (EuroUSEC 2024), pp. 337–353, 2024, doi: 10.1145/3688459.3688474.

[8] B. Burgess, A. Ginsberg, E. W. Felten, and S. Cohney, “Watching the watchers: bias and vulnerability in remote proctoring software,” in 31st USENIX Security Symposium (USENIX Security 22), pp. 571–588, 2022.