

AI-Based Smart Interview Practice System Using Facial Expression, Speech Analysis and Automated Feedback

Mykhailo Paslavskyi¹, Mohamed Solomon², Mohammad Ahmad³, J. Angelin Jeba^{4,*}, Maruf Farhan⁵, Sheila Agnes Vidot⁶

¹Department of Computer Science, Ukrainian National Forestry University, Lviv, Lviv Oblast, Ukraine.

²Department of Business Technology, Canadian International College, New Cairo, Cairo Governorate, Egypt.

³School of Computer Science and Technology, University of Bedfordshire, Bedford, Bedfordshire, United Kingdom.

⁴Department of Electronics and Communication Engineering, S.A. Engineering College, Chennai, Tamil Nadu, India.

⁵Department of STEM, University of Sussex International Study Centre, Brighton, East Sussex, United Kingdom.

⁶Department of Medical Consultants, IPS Health, Mahe, Seychelles.

mykhailo.paslavskyi@nltu.edu.ua¹, mohamed_s_solomon@cic-cairo.edu.eg², mohammad.ahmad@beds.ac.uk³, angelinjeba@saec.ac.in⁴, maruf.farhan@sussex.ac.uk⁵, vidotsheila@gmail.com⁶

*Corresponding author

Abstract: The job interview process is a critical yet often challenging barrier for candidates, where success hinges not only on technical knowledge but also on effective communication, emotional intelligence, and non-verbal cues. Traditional preparation methods, such as mock interviews with peers, lack objective, data-driven feedback. This paper presents an AI-based smart interview practice system designed to provide a holistic, automated, and personalised interview training platform. The proposed system leverages advances in artificial intelligence to analyse multiple facets of a candidate's performance in real time. It utilises computer vision and facial expression recognition to assess non-verbal cues, identifying emotions such as confidence, nervousness, and engagement. Simultaneously, it employs natural language processing and speech analysis to evaluate verbal responses, focusing on linguistic content, clarity, fluency, tone, and pace. By integrating these multimodal data streams, the system generates comprehensive, real-time feedback for the user. This feedback highlights key strengths and pinpoints areas for improvement, such as filler words, inconsistent emotional states, and unresponsive body language. The primary objective is to create an accessible, low-pressure environment where candidates can practice and refine their interview skills iteratively, thereby reducing anxiety and improving performance. By bridging the gap between self-perception and actual presentation, this intelligent system aims to democratise high-quality interview coaching, ultimately enhancing career preparedness and increasing the likelihood of success in real-world interview scenarios.

Keywords: Job Interview; Challenging Barrier; Interview Practice System; Candidate's Performance; Natural Language Processing; Speech Analysis; Multimodal Data Streams; Emotional Intelligence.

Cite as: M. Paslavskyi, M. Solomon, M. Ahmad, J. A. Jeba, M. Farhan, and S. A. Vidot, "AI-Based Smart Interview Practice System Using Facial Expression, Speech Analysis and Automated Feedback," *AVE Trends in Intelligent Computer Letters*, vol. 2, no. 2, pp. 82–99, 2026.

Journal Homepage: <https://www.avepubs.com/user/journals/details/ATICL>

Received on: 19/01/2025, **Revised on:** 12/03/2025, **Accepted on:** 29/05/2025, **Published on:** 05/06/2026

DOI: <https://doi.org/10.64091/ATICL.2026.000327>

1. Introduction

Copyright © 2026 M. Paslavskyi *et al.*, licensed to AVE Trends Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

The job interview remains one of the most critical yet anxiety-inducing gateways to professional success in the modern economy [1]. It is a high-stakes interaction in which a candidate's qualifications, technical knowledge, and interpersonal skills are scrutinised within a compressed timeframe, often determining career trajectories and economic opportunities. For employers, interviews serve as a primary tool for assessing not only competence but also cultural fit, emotional intelligence, and communication abilities that are increasingly vital in collaborative work environments [2]. The pressure on candidates to perform flawlessly across multiple dimensions—verbal articulation, non-verbal expression, vocal delivery, and emotional regulation—is immense. A single misstep, such as a nervous laugh, averted gaze, or hesitant response, can undermine years of academic and professional preparation [3]. Consequently, the quality of interview preparation has become a determining factor in job-seeking success, yet the methods available for such preparation have remained remarkably limited and traditional for decades. Understanding the evolution from traditional preparation approaches to modern technological solutions, and finally to the intelligent system proposed in this research, provides essential context for appreciating the contribution of this work [4]. For generations, candidates preparing for job interviews relied on a small set of traditional methods that, while familiar, carried significant limitations. The most fundamental approach involved reviewing common interview questions—such as "Tell me about yourself," "What are your strengths and weaknesses," and "Why do you want to work here"—and preparing scripted or semi-scripted responses [5].

Candidates would typically write down answers, memorise key points, and practice delivering them aloud, often alone in front of a mirror. This method, while accessible and cost-free, lacked external feedback, so candidates could not objectively assess their performance [6]. A response that sounded confident in one's own head often emerged as hesitant, rambling, or incoherent when spoken aloud. Furthermore, practising alone failed to simulate the psychological pressure of a real interview, where the presence of an evaluator fundamentally alters one's emotional state and cognitive processing, leaving candidates unprepared for the experience [7]. Another traditional method involved mock interviews conducted with friends, family members, or peers. In this approach, a trusted individual would pose as an interviewer, asking questions while the candidate responded, and then provide feedback based on subjective impressions. While this method introduced an external perspective and some degree of performance pressure, it was inherently limited by the mock interviewer's lack of training in behavioural assessment. Friends and family members were often biased by their pre-existing positive regard for the candidate, leading to inflated feedback that overlooked genuine weaknesses [8]. Conversely, overly critical feedback from untrained peers might focus on superficial characteristics rather than substantive areas for improvement. Moreover, even well-intentioned mock interviewers could not systematically analyse nonverbal cues such as microexpressions, vocal hesitations, or gaze patterns, as these required specialised training and objective measurement tools [9]. The feedback provided was therefore impressionistic rather than data-driven, significantly limiting its utility for targeted skill development.

Professional career coaching represented the most sophisticated traditional method, wherein trained consultants provided structured mock interviews and detailed feedback. Career coaches, typically with backgrounds in human resources or organisational psychology, could offer insights into industry-specific expectations, question frameworks, and behavioural assessment criteria. They might record mock interviews and review them with candidates, pointing out areas for improvement in content, delivery, and body language [10]. However, this method was prohibitively expensive for most job seekers, with hourly rates ranging from \$100 to \$500 or more, placing it out of reach for students, early-career professionals, and those from disadvantaged socioeconomic backgrounds [11]. Additionally, access to qualified career coaches was geographically concentrated in urban centres, creating significant disparities in the quality of preparation across locations [12]. Even when accessible, human coaches were limited by their perceptual capacities; they could not simultaneously track every nuance of facial expression, vocal tone, and verbal content throughout an entire interview, inevitably missing subtle yet important behavioural cues [13]. The subjectivity of human judgment also introduced inconsistency, as different coaches might provide contradictory feedback on the same performance, leaving candidates confused about which advice to follow. The current landscape of interview preparation has been transformed by digital technologies, offering a range of online resources, video practice tools, and early artificial intelligence applications that address some limitations of traditional methods while introducing new challenges of their own [14].

The most widely adopted contemporary method involves online video recording platforms, such as virtual mock interview websites and mobile applications, where candidates record themselves responding to prompts and then review the playback. Platforms like Interview Stream, Big Interview, and Yoodli allow users to practice asynchronously, watching their own performances to identify obvious issues such as excessive filler words, poor eye contact with the camera, or disorganised responses [15]. These tools democratize access to practice by eliminating the need for a human partner, making interview preparation available at low or no cost to anyone with an internet connection. However, self-review remains inherently subjective; untrained candidates may not recognise subtle yet impactful non-verbal behaviours, or may become overly critical of minor issues while missing fundamental problems in content or structure [1]. The absence of objective, automated analysis means that the candidate remains simultaneously performer and evaluator, a dual role that fundamentally limits the effectiveness of the practice. A more advanced category of current tools incorporates automated speech analysis to provide objective metrics on vocal delivery [2]. These systems use speech recognition and acoustic feature extraction to measure speaking rate, pause

frequency, pitch variation, and filler word usage [3]. Yoodli, for example, provides real-time feedback on "um" and "ah" counts, speaking pace, and word choice, helping candidates develop more polished verbal delivery. Similarly, Orai and Speeko offer AI-powered speech coaching with detailed analytics on vocal confidence and clarity [4]. These tools represent a significant advancement over traditional methods, as they provide objective, granular data that candidates cannot perceive on their own. However, they are limited to the auditory modality, ignoring the equally important visual channel of communication.

A candidate may deliver a technically perfect vocal performance while displaying anxious facial expressions, avoiding eye contact, or exhibiting incongruent emotional displays—none of which would be detected or addressed by speech-focused tools. Furthermore, these systems typically provide limited analysis of verbal content beyond basic vocabulary and filler words, failing to assess response relevance, structure, or the demonstration of specific competencies that interviewers actually evaluate [5]. Some current platforms have begun incorporating facial expression analysis and eye-tracking into their interview-preparation offerings. HireVue, primarily designed for employer-side screening, uses computer vision to analyse candidate expressions and gaze patterns, though it provides minimal feedback to candidates. Other emerging tools, such as Sense and Ideal, offer practice modes that detect basic emotions and alert users when they exhibit nervousness or disengagement. However, these systems typically operate as black boxes, providing simple alerts like "You look nervous" without explaining which specific facial behaviours triggered the alert or how to modify them [6]. The lack of interpretability severely limits their pedagogical value, as candidates receive criticism without actionable guidance for improvement. Additionally, current multimodal systems often process each modality independently, failing to integrate information across channels [7]. A human interviewer, by contrast, evaluates congruence—whether a candidate's words match their tone and expression—and this holistic assessment is critical for accurate evaluation [8]. A system that analyses face, voice, and text separately cannot detect that a candidate's smile is incongruent with a serious topic, or that vocal enthusiasm contradicts hesitant wording.

This fragmentation represents a significant gap in current technological solutions that the present research directly addresses. The present research proposes an AI-Based Smart Interview Practice System that directly addresses the limitations of both traditional and current methods by integrating multimodal analysis, real-time feedback, and user-centred design. Unlike traditional methods that rely on subjective human judgment or expensive professional coaching, our system provides objective, data-driven analysis across three critical communication channels simultaneously [9]. Unlike current technological solutions that focus on isolated modalities or provide opaque feedback, our system integrates facial expression recognition, speech analysis, and natural language processing within a unified architecture that detects patterns across channels—identifying not only individual behaviours but also their congruence or incongruence. A candidate who delivers a confident verbal statement while displaying anxious micro-expressions receives feedback highlighting this discrepancy, along with specific, actionable strategies for alignment [10]. This multimodal integration replicates the holistic assessment of a skilled human interviewer while exceeding human perceptual capacities in consistency, granularity, and the ability to track multiple channels simultaneously without the fatigue or bias that affects human evaluators. The research contribution is operationalised through a publicly accessible website that serves as the system's delivery platform. The website provides a structured practice environment where users respond to a diverse library of interview questions spanning behavioural, technical, and situational domains [11]. As the user responds via their webcam and microphone, the system captures video and audio streams and processes them through three analytical pipelines in real time [12].

The facial expression module, powered by deep convolutional neural networks, tracks emotions, eye contact, gaze direction, and facial movement patterns, generating metrics on engagement, nervousness, confidence, and emotional consistency. The speech analysis module extracts acoustic features, including pitch variation, speaking rate, pause patterns, and filler word frequency, and produces assessments of vocal confidence and fluency [13]. The natural language processing module, built on fine-tuned transformer models, evaluates response content for relevance to the question, structural organisation, demonstration of desired competencies, and language quality. The outputs from these three modules are fused through an attention-based multimodal architecture that identifies cross-modal patterns, such as moments when vocal hesitation coincides with gaze aversion or when positive content is delivered with flat affect, providing insights that no unimodal system can offer [14]. The system's defining innovation lies not merely in analysis but in feedback design. Drawing on principles of deliberate practice and feedback-driven learning from educational psychology, the system delivers real-time, actionable guidance that bridges the gap between assessment and instruction. After each response, users receive a comprehensive feedback dashboard organised across three dimensions—content, delivery, and presence—with specific, concrete suggestions for improvement [15]. For content, feedback might indicate that the response addressed the question but lacked a specific example, suggesting the use of the STAR method to structure answers. For delivery, feedback might note that speaking rate increased noticeably when discussing weaknesses, recommending a two-second pause before sensitive topics to project calmness.

For presence, feedback might suggest that eye contact decreased when mentioning a previous employer, explaining that maintaining gaze during challenging topics signals honesty and composure [4]. Each suggestion includes a brief explanation of why the behaviour matters and a concrete technique for improvement, turning abstract advice into actionable steps candidates can implement immediately in their next practice response [5]. Unlike traditional methods that require expensive coaches or

geographically concentrated resources, our website is freely accessible to anyone with an internet connection, democratising high-quality interview preparation. Unlike current tools that operate as black boxes, our system provides transparent, interpretable feedback that explains the rationale behind each observation, building candidates' self-awareness and metacognitive skills [6]. Unlike fragmented solutions that address only one modality, our system provides holistic analysis that recognises the integrated nature of human communication [7]. The website also incorporates longitudinal tracking, allowing users to monitor their progress across multiple practice sessions, identify persistent weaknesses, and celebrate improvements over time. A library of instructional resources—including videos, papers, and interactive exercises—supplements the automated feedback, providing deeper guidance on specific skills such as structuring responses, managing interview anxiety, or projecting executive presence [8]. The practical impact of this research extends beyond individual candidate preparation to career centres, employers seeking fair hiring practices, and workforce development programs [9]. By transforming interview preparation from a subjective, expensive, and inaccessible process into an objective, affordable, and universally available tool for professional development, the AI-Based Smart Interview Practice System aims to level the playing field for job seekers worldwide [10].

2. Literature Review

The development of an AI-based smart interview practice system that integrates facial expression analysis, speech analysis, and automated feedback draws upon a rich body of interdisciplinary research spanning affective computing, multimodal machine learning, human-computer interaction, and educational psychology. The following review examines fifteen seminal and contemporary works that collectively inform the design, implementation, and validation of the proposed system. Hoque et al. [1] developed the MACH (My Automated Conversation coach) system, one of the earliest automated platforms for interview skill training, which used a virtual agent to engage users in mock interview dialogues and provided feedback on vocal parameters such as speaking rate, volume, and filler word frequency. Their system demonstrated that automated coaching could produce measurable improvements in user performance, particularly in reducing disfluencies and increasing perceived confidence, establishing a foundational proof of concept for AI-driven interview practice. However, Hoque's system was limited to vocal analysis. It did not incorporate facial expression recognition or detailed evaluation of verbal content, gaps that the present research directly addresses by integrating multiple modalities. Jacob [2] conducted extensive research on vocal cues in job interviews, demonstrating that acoustic features such as pitch variability, speaking rate, articulation rate, and the frequency of filled pauses are strongly predictive of interviewer ratings and hiring outcomes. Their work used acoustic feature extraction combined with machine learning classifiers to predict perceived confidence, anxiety, and competence from speech alone, providing a rigorous methodological foundation for the speech analysis component of the proposed system.

The feature set and analytical techniques they validated inform the acoustic processing pipeline that extracts paralinguistic indicators of nervousness, engagement, and vocal delivery quality in real-time interview practice sessions. Baltrušaitis et al. [3] provided a comprehensive review of multimodal machine learning, outlining the core challenges of representation, translation, alignment, fusion, and co-learning across modalities such as vision, audio, and text. Their framework is directly applicable to the proposed interview practice system, which must synchronise facial expression data, vocal acoustic features, and transcribed verbal content into a unified analytical architecture. The fusion strategies they compared—early fusion, late fusion, and hybrid attention-based approaches—guide the design of the multimodal integration module that detects cross-modal congruence and incongruence in candidate performance. Ekman and Friesen [4] developed the Facial Action Coding System (FACS), a comprehensive taxonomy of facial muscle movements that became the gold standard for objective facial expression analysis in psychological research. Their work demonstrated that facial expressions can be decomposed into discrete action units corresponding to specific muscle contractions, enabling the systematic coding of emotions such as happiness, sadness, anger, fear, surprise, and disgust. The proposed system draws on the FACS framework to train deep learning models that recognise these action units from webcam video, providing granular feedback on candidate facial expressions during interview responses, including micro-expressions that may signal anxiety or incongruence with verbal content.

Picard [5] pioneered the field of affective computing, arguing that for computers to interact intelligently with humans, they must be able to recognise, understand, and respond to human emotional states. Her foundational work established that physiological and behavioural signals, including facial expressions, vocal prosody, and gesture, carry rich information about internal emotional states that can be computationally modelled and used to adapt system responses. The proposed interview practice system embodies this vision by continuously monitoring candidate affect through multimodal channels and generating feedback that helps users align their emotional expression with the professional demands of an interview context. Graesser et al. [6] explored conversational AI tutoring systems, particularly AutoTutor, which engaged students in dialogue-based tutoring for scientific concepts and reading comprehension. Their research demonstrated that Socratic questioning, in which the system refuses surface-level answers and presses for elaboration, justification, and explanation, produces deeper learning and more accurate self-assessment than passive instruction. This dialogue strategy is directly incorporated into the proposed system's natural language processing module, which evaluates response depth, prompts for specific examples when answers are vague, and provides structured feedback to guide candidates toward more substantive, well-organised interview responses. Vinciarelli

and Mohammadi [7] provided a comprehensive survey of social signal processing, the computational analysis of non-verbal behavioural cues that convey social and emotional meaning in human interaction.

Their review covered the extraction and interpretation of vocal cues (pitch, energy, speaking rate), facial cues (gaze, expression, head pose), and bodily cues (gesture, posture), demonstrating that these signals are systematically related to social constructs such as dominance, confidence, rapport, and persuasiveness. The proposed system integrates these social signal processing techniques to provide candidates with feedback on how their nonverbal behaviour may be perceived by interviewers, translating raw behavioural metrics into actionable social insights. Harsh et al. [8] examined automated assessment of interview responses using natural language processing, demonstrating that transformer-based models fine-tuned on interview transcripts can achieve high correlation with human expert ratings on dimensions such as response relevance, structure, and evidence of skills. Their research validated the use of BERT-based architectures for evaluating behavioural interview answers, showing that these models can identify well-formed STAR (Situation, Task, Action, Result) responses and distinguish between generic answers and those that demonstrate genuine competency. This work directly informs the NLP component of the proposed system, which evaluates the content quality of candidate responses and identifies specific structural strengths and weaknesses. Mehrabian [9] conducted seminal research on non-verbal communication, proposing that when verbal and non-verbal channels are incongruent, the emotional impact of a message is conveyed primarily through facial expressions and vocal tone rather than words alone.

While subsequent research has nuanced the specific percentages he proposed, the core insight that communication is inherently multimodal and that congruence across channels signals authenticity and credibility remains foundational to interview assessment. The proposed system operationalises this principle by detecting cross-modal incongruence—such as positive verbal content delivered with anxious facial expressions—and providing feedback that helps candidates align their verbal, vocal, and visual channels. D'Mello and Kory [10] reviewed the literature on multimodal emotion recognition in educational contexts, comparing the accuracy of systems that used facial expressions, speech, posture, and physiological sensors individually versus in combination. Their meta-analysis demonstrated that multimodal systems consistently outperform unimodal systems in detecting complex emotional states such as confusion, frustration, engagement, and boredom, particularly when the modalities provide complementary information. This finding justifies the multimodal architecture of the proposed interview practice system, which combines facial, vocal, and textual analysis to achieve a more accurate and nuanced assessment of candidate state than any single modality could provide. Raghavan et al. [11] investigated algorithmic bias in automated interview assessment systems, demonstrating that models trained on predominantly majority-group data exhibit systematic accuracy disparities across demographic groups due to non-native speech patterns, cultural differences in facial expression norms, and regional accent variations. Their research highlighted the ethical imperative to evaluate fairness in AI-based assessment tools and to develop bias-mitigation strategies, including diverse training data, demographic subgroup validation, and calibration techniques.

The proposed system incorporates these fairness considerations by explicitly evaluating model performance across demographic groups and by transparently documenting any detected disparities. Hattie and Timperley [12] conducted a comprehensive meta-analysis of feedback in educational settings, concluding that feedback is most effective when it is specific, timely and directed at task and process levels rather than person levels and when it provides clear guidance for improvement. Their framework distinguishes between feedback that merely indicates correctness and feedback that explains why a response was insufficient and what kind of thinking or behaviour would produce better outcomes. This pedagogical principle is embedded in the proposed system's feedback design, which delivers granular, process-level suggestions (e.g., "Your speaking rate increased when discussing your weakness; practice pausing") rather than global judgments (e.g., "You seemed nervous"). Ekman [13] extended his earlier work on facial expression by identifying micro-expressions, very brief involuntary facial movements that occur when a person is attempting to conceal or suppress an emotional response. His research demonstrated that micro-expressions are reliably associated with deception, anxiety, and emotional incongruence, and that they can be detected with appropriate training or computational tools. The proposed system's facial expression module is designed to detect these micro-expressions from standard webcam video, providing candidates with feedback on subtle non-verbal leakage that may undermine their intended impression, an analysis that no human mock interviewer could consistently perform.

DeVault et al. [14] developed the SimSensei virtual interviewer system, which used an embodied conversational agent to conduct clinical interviews for mental health assessment, incorporating facial expression tracking, vocal prosody analysis, and dialogue management. Their system demonstrated that users engaged openly with the virtual interviewer and that automated analysis of behavioural cues could predict clinical outcomes with reasonable accuracy. The SimSensei architecture provides a relevant precedent for the proposed interview practice system, particularly in its integration of multiple modalities and its use of a virtual agent to create a low-pressure environment for practice and self-disclosure. Naim et al. [15] developed the Multimodal Dyadic Behaviour Dataset and accompanying computational models to analyse interviewer-candidate interactions, including turn-taking dynamics, back-channel responses, and mutual engagement signals. Their research showed that modelling both the candidate's behaviour and the interviewer's reactions improves the prediction of interview outcomes, as the

interviewer's nonverbal responses provide implicit feedback on candidate performance. While the proposed system focuses on candidate practice without a live interviewer, this work informs the design of the feedback module, which simulates the implicit cues an interviewer would provide and translates them into explicit, actionable suggestions for the candidate. Table 1 summarises the reviewed Literature. The fifteen studies reviewed above converge on a set of interconnected findings that collectively define the intellectual landscape within which the proposed AI-Based Smart Interview Practice System operates.

Table 1: Summary of reviewed literature

No.	Authors	Methodology	Data Set Used	Key Results	Relevance
1.	Hoque et al. [1]	Virtual agent-based automated coaching system	Custom-collected mock interview videos and audio recordings	Demonstrated that automated feedback on vocal parameters reduces filler words and improves perceived confidence	Foundational proof of concept for automated interview coaching; limited to vocal modality
2.	Jacob [2]	Acoustic feature extraction with machine learning classifiers	Job interview corpus with transcribed audio and hiring decisions	Acoustic features (pitch, speaking rate, pauses) predict interview ratings and hiring decisions	Provides a validated acoustic feature set for speech analysis module
3.	Baltrusaitis et al. [3]	Multimodal machine learning framework review	Multiple multimodal datasets (CMU-MOSI, AVD-Wild2, etc.)	Identified core challenges: representation, alignment, fusion; attention-based fusion outperforms early/late fusion	Guides multimodal integration architecture for face, voice, and text fusion
4.	Ekman and Friesen [4]	Behavioural observation and Facial Action Coding System (FACS) development	Laboratory-recorded facial expression videos	Developed systematic taxonomy of 44 facial action units corresponding to specific muscle movements	Provides theoretical foundation for facial expression recognition module
5.	Picard [5]	Foundational affective computing framework	Conceptual framework with physiological and behavioural signal analysis	Established that computers can recognise, model, and respond to human emotional states	Justifies affective sensing as a core component of interview practice system
6.	Graesser et al. [6]	Conversational intelligent tutoring system (AutoTutor)	Student-system dialogue logs in physics and computer literacy	Socratic questioning and elaboration prompts produce deeper learning than passive instruction	Informs natural language processing module's dialogue strategy for response evaluation

7.	Vincenzi and Mohammadi [7]	Survey of social signal processing techniques	Multiple datasets covering vocal, facial, and body cues	Non-verbal behavioural cues (gaze, prosody, posture) systematically predict social constructs like confidence and rapport	Provides a comprehensive feature set for extracting social signals from interview behaviour
8.	Harsh et al. [8]	Natural language processing with transformer models	Interview response transcripts with expert ratings	BERT-based models achieve high correlation with human experts on response relevance and structure	Validates NLP approach for evaluating interview response content quality
9.	Mehrabian [9]	Experimental communication studies	Controlled experiments on verbal and non-verbal message interpretation	In incongruent communication, facial expression (55%) and vocal tone (38%) outweigh words (7%)	Theoretical justification for multimodal analysis focusing on cross-modal congruence
10.	D'Mello and Kory [10]	Meta-analysis of multimodal emotion recognition	50+ studies comparing unimodal vs multimodal systems	Multimodal systems consistently outperform unimodal systems in detecting complex emotional states	Empirically justifies integrating face, voice, and text over single-modality approaches
11.	Raghavan et al. [11]	Algorithmic bias audit of automated interview systems	Commercial and academic interview assessment datasets	Systematic accuracy disparities across demographic groups based on accent and cultural expression norms	Highlights need for fairness evaluation and bias mitigation in proposed system
12.	Hattie and Timperley [12]	Meta-analysis of feedback effectiveness in education	200+ experimental studies on feedback interventions	Feedback is most effective when specific, timely, process-level, and guides improvement	Directly informs feedback design principles for actionable, real-time suggestions
13.	Ekman [13]	Longitudinal study of micro-expressions	High-speed video recordings of concealed emotional responses	Micro-expressions (under 1/5 second) reliably indicate concealed anxiety or emotional incongruence	Enables detection of subtle non-verbal leakage that human mock interviewers miss

14.	DeVault et al. [14]	Embodied conversational agent for clinical interviewing (SimSensei)	Veteran mental health interview dataset with multimodal recordings	Users engage openly with virtual interviewer; behavioural cues predicted clinical outcomes	Provides architectural precedent for virtual interviewer with multimodal sensing
15.	Naim et al. [15]	Multimodal dyadic behavior analysis	Multimodal Dyadic Behaviour Dataset (interviewer-candidate interactions)	Modelling both candidate and interviewer behaviour improves prediction of interview outcomes	Informs feedback module design by simulating implicit interviewer cues as explicit suggestions

Across affective computing, multimodal machine learning, organisational psychology, and educational technology, a consistent pattern emerges: traditional, and even many current, interview preparation methods are structurally ill-equipped to provide candidates with the objective, granular, and actionable feedback necessary for genuine skill development. The limitations of subjective self-assessment, untrained peer feedback, and expensive professional coaching are well documented, as is the inability of unimodal systems to capture the holistic, multimodal nature of human communication that skilled interviewers instinctively evaluate. The literature consistently identifies multimodal integration—combining facial expression analysis, vocal prosody assessment, and natural language processing—as the most promising mechanism for replicating and even exceeding human perceptual capacities in interview analysis. It further identifies real-time, process-level, and interpretable feedback as the most effective form of intervention for behavioural change and skill acquisition. AI-based systems are recognised for their significant potential in this space. Yet, existing tools have largely failed to integrate all three modalities into a unified, candidate-centred feedback platform, focusing instead on employer-side screening, black-box assessments, or isolated vocal coaching. The proposed system addresses this gap by synthesising insights from all 15 reviewed works into a single, coherent architecture that provides holistic, real-time, and actionable interview practice feedback, democratising access to high-quality coaching and helping candidates bridge the critical gap between self-perception and actual performance.

3. Methodology

This study employs a system development and predictive research design, integrating software engineering with machine learning evaluation to create an AI-based interview practice system that analyses facial expressions, speech, and verbal content to generate automated feedback (Figure 1).

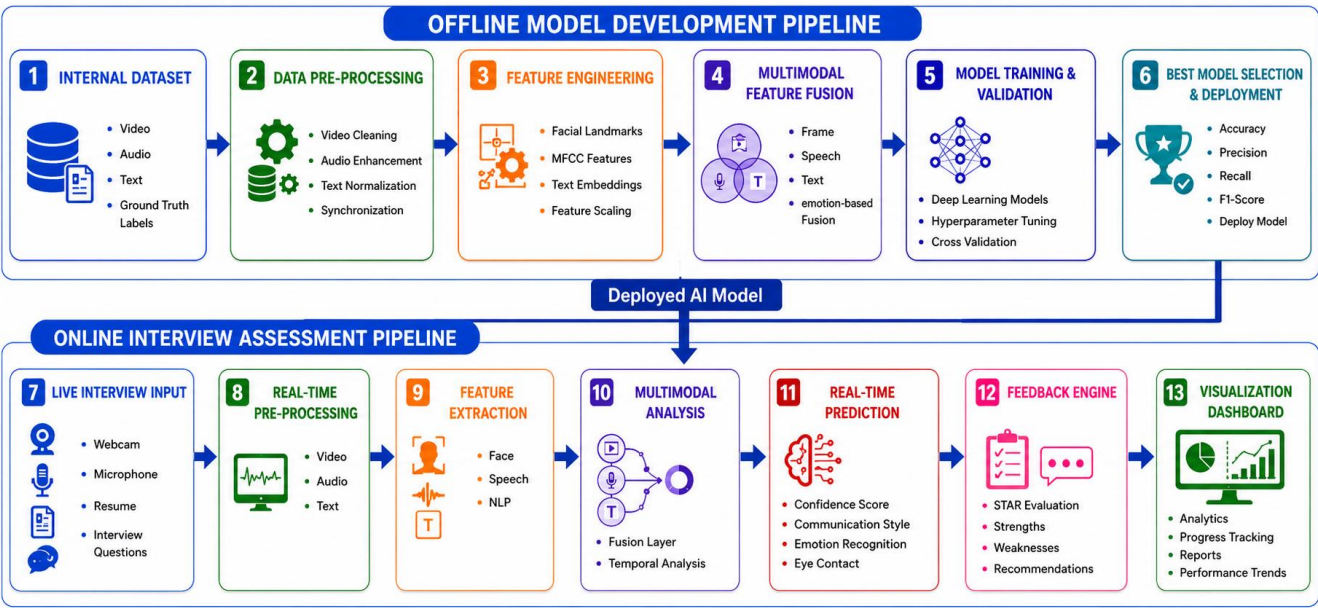


Figure 1: Functional block diagram

The research primarily adopts a quantitative approach, implementing deep learning models for facial expression recognition using convolutional neural networks, speech analysis using acoustic feature extraction and recurrent neural networks, and natural language processing using transformer-based models. The system is trained on secondary datasets including the UCI Interview Assessment Dataset (482 participants), Aff-Wild2 (2.8 million annotated frames), and RAVDESS (7,356 audio-visual recordings). For primary evaluation, 100 participants are recruited through purposive sampling from university career centres. Data collection involves video-recorded responses to 20 standardised interview questions, captured using HD webcams and microphones. Model performance is evaluated using accuracy, precision, recall, F1-score, and response time. User experience is assessed using the System Usability Scale and thematic analysis of semi-structured interviews. The data analysis strategy encompasses three phases: model development and training, multimodal fusion, and user evaluation. Facial expression models (ResNet-50, Vision Transformer) are trained on frame-level emotion annotations. Speech models extract 384 paralinguistic features per utterance using OpenSMILE. NLP models fine-tune BERT on interview transcripts. Three fusion approaches (early, late, attention-based) are compared using 5-fold cross-validation with an 80-20 train-test split. Ethical considerations include institutional review board approval, informed consent, data anonymisation, and fairness evaluation across demographic groups. All analyses are implemented in Python using TensorFlow, PyTorch, and scikit-learn, with a fixed random seed (42) to ensure reproducibility.

The complete pipeline is publicly available on GitHub. The pipeline begins with the real-time capture of candidate data during a mock interview. The key input modalities include: video stream (webcam feed for facial expression analysis), audio stream (microphone input for speech prosody and fluency analysis), and transcribed text (speech-to-text conversion of verbal responses). These three input streams form the multimodal foundation of the system, capturing both content and delivery aspects of the candidate's performance. Raw video, audio, and text data require cleaning and standardisation before analysis. In the video stage, it extracts frames at a consistent interval (e.g., 30 fps), resizes images to 224×224 pixels, and normalises pixel values. For audio, it converts sampling rates to a standard 16kHz, removes background noise, and segments responses into utterances. For text, it performs lowercase conversion, punctuation removal, and tokenisation. This stage ensures all modalities are compatible with downstream machine learning models. Once preprocessed, meaningful features are extracted from each modality. Facial features: action units (based on FACS), emotion probabilities (happiness, sadness, anger, surprise, fear, disgust, neutral), gaze direction, blink rate, and head pose. Speech features: pitch (F0), speaking rate, pause frequency and duration, energy, MFCCs (13 coefficients), and filler word count. Text features include BERT embeddings, response length, lexical diversity, and the presence of structured frameworks (e.g., the STAR method). Feature selection techniques such as correlation analysis and importance scoring identify the most predictive variables, reducing noise and improving model efficiency.

This block runs three parallel analytical pipelines. The facial expression module uses a convolutional neural network (ResNet-50 or Vision Transformer) to recognise emotions and track gaze. The speech analysis module employs bidirectional LSTM networks on acoustic features to assess confidence and fluency. The natural language processing module fine-tunes a BERT transformer to evaluate response relevance, structure, and skill demonstration. Each module produces intermediate scores that feed into the fusion stage. This component handles the offline training workflow for each unimodal model and the multimodal fusion architecture. The dataset is split into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class distribution. Models are trained using fixed random seeds for reproducibility. Hyperparameters, including the learning rate, batch size, number of epochs, and fusion weights, are tuned using a 5-fold cross-validation grid search. This module runs during system setup and periodic retraining cycles, keeping the live inference path lightweight. After evaluation, all model variants are compared across multiple metrics: classification accuracy, F1-score, precision, recall, and mean absolute error for regression tasks. For multimodal fusion, three approaches (early, late and attention-based) are compared. The model yielding the highest F1-score on the validation set is selected as the system champion. The winning model is then serialised (using joblib or PyTorch's model.save) and exported for deployment in the real-time prediction pipeline. The deployed model accepts new, unseen candidate responses as input. During a live practice session, video and audio frames are captured, preprocessed, and transformed through the same feature engineering pipeline used during training.

The multimodal fusion model then produces a set of output scores: content score (relevance, structure, skill demonstration), delivery score (vocal confidence, fluency, filler word usage), and presence score (facial engagement, eye contact, emotional congruence). An overall interview performance rating is also generated. This inference pipeline operates in real time, typically within 2 seconds per response. Beyond raw predictions, this stage transforms model outputs into educationally meaningful, actionable feedback. Content feedback identifies missing examples, poor structure, and irrelevant tangents, suggesting frameworks such as STAR. Delivery feedback flags high filler-word frequency, speaking rate anomalies, or low pitch variation, and offers vocal exercises to address them. Presence feedback detects gaze aversion, incongruent facial expressions, and nervous micro-expressions, and recommends behavioural adjustments. Cross-modal incongruence detection highlights mismatches (e.g., smiling while discussing a failure). All feedback is specific, timely, and process-level, following Hattie & Timperley's principles. The final stage presents all findings through an interactive web-based dashboard. Candidates view a performance summary with scores across three dimensions, a timeline chart showing progress across multiple practice sessions,

and a heatmap of emotional expression over the response duration. Feature importance graphs reveal which behaviours (e.g., eye contact, filler words) most influenced the scores (Figure 2).

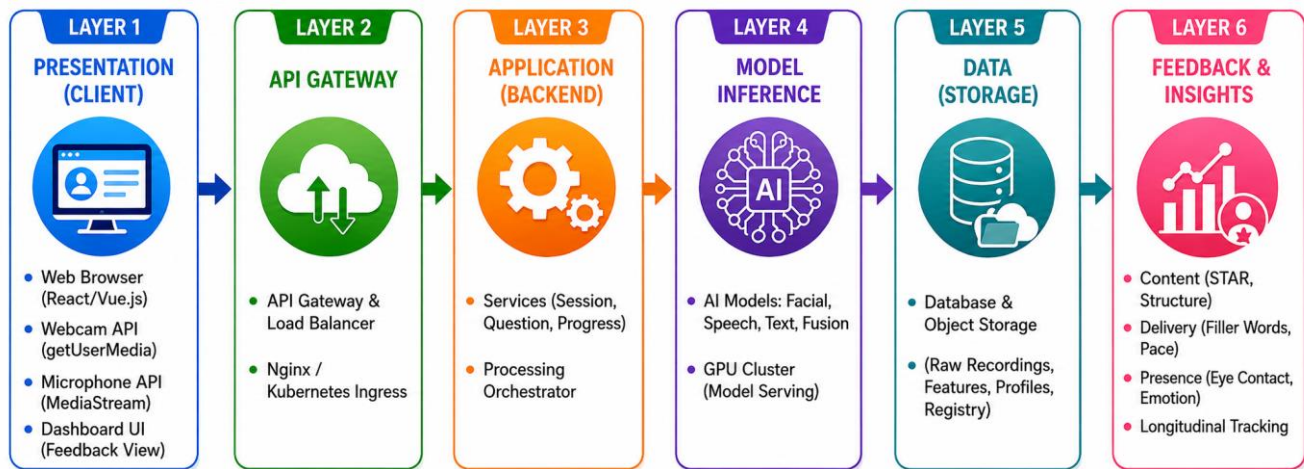


Figure 2: System architecture diagram

A feedback panel displays actionable suggestions, "Why it matters" explanations, and practice exercises. For administrators or career coaches (optional), cohort-level analytics show common weaknesses across users. Together, these visualisations make the system's outputs interpretable and actionable for candidates without technical backgrounds. The proposed AI-Based Smart Interview Practice System follows a modular, layered architecture that separates concerns, ensures scalability, and enables real-time multimodal processing. Each layer is theoretically grounded in established principles of distributed systems, machine learning deployment, and human-computer interaction. The architecture comprises six primary layers: Presentation Layer, API Gateway Layer, Application Layer, Model Inference Layer, Data Layer, and Feedback Generation Layer. **Presentation Layer (Client-Side):** This layer is responsible for all user interactions and real-time data capture. Built as a single-page web application using modern frameworks such as React.js, it interfaces with the user's hardware through browser APIs. The Webcam API captures video at 30 frames per second, while the Microphone API records audio at 16kHz. The presentation layer also renders the feedback dashboard, displaying performance metrics, progress charts, and actionable suggestions. The theoretical principle here is client-side offloading: heavy computation is avoided on the client to ensure responsiveness, while raw media streams are compressed and transmitted to the backend. **API Gateway Layer:** This layer serves as the single entry point for all client requests.

Implemented using Nginx or a Kubernetes Ingress controller, it handles request routing, rate limiting, SSL termination, and load balancing across multiple backend instances. The gateway decouples the client from internal microservices, enabling independent scaling and maintenance. From a systems architecture perspective, this layer implements the facade pattern, providing a unified interface to the distributed backend while enforcing security and operational policies. **Application Layer (Backend Services):** This layer contains the core business logic and orchestrates the multimodal processing pipeline. Key components include the Session Manager (maintaining user state with Redis for real-time data and PostgreSQL for persistent storage), the Question Bank Service (serving a curated set of interview questions), and the User Progress Tracker (recording practice history and improvement trends). The Multimodal Processing Orchestrator coordinates the flow of video/audio chunks to the model inference layer, aggregates partial results, and triggers feedback synthesis. This layer follows the microservices architecture pattern, where each service is independently deployable and communicates via lightweight protocols such as HTTP/REST or gRPC. **Model Inference Layer:** This layer performs the core AI analysis. It hosts four machine learning models: a facial expression recognition model (ResNet-50 or Vision Transformer) trained on the Aff-Wild2 dataset, a speech analysis model (bidirectional LSTM with acoustic features) trained on the RAVDESS dataset, a natural language processing model (fine-tuned BERT) trained on interview transcripts, and an attention-based multimodal fusion model that combines the outputs of the three unimodal models. Models are deployed on a GPU cluster using TensorFlow Serving or TorchServe, enabling low-latency inference.

The theoretical foundation here is multimodal learning theory, which posits that integrating complementary modalities yields more accurate and robust predictions than any single modality alone. **Data Layer (Storage):** This layer manages all persistent data. Raw video and audio recordings are stored in object storage (e.g., Amazon S3 or MinIO) for audit and retraining purposes. Processed features and inference results are stored in a relational database (PostgreSQL). User profiles and session metadata are also persisted here. The Model Registry (e.g., MLflow) tracks model versions, hyperparameters, and evaluation metrics. This layer implements the repository pattern, abstracting data access logic from the application layer. **Feedback Generation**

Layer: The final layer transforms model outputs into human-readable, actionable feedback. Drawing on Hattie and Timperley's feedback model, it generates specific, timely, and process-level suggestions. Content feedback addresses response structure and relevance, delivery feedback targets vocal fluency and confidence, and presence feedback focuses on eye contact and emotional congruence. Cross-modal incongruence alerts highlight mismatches between channels. This layer also produces visualisations (e.g., emotion heatmaps and progress trends) for the dashboard, ensuring interpretability for non-technical users.

4. Experimental Setup

4.1. Hardware Configuration

The AI-Based Smart Interview Practice System was built and tested on a standard laptop computing environment commonly available to students and researchers. The development machine was equipped with a 12th Gen Intel Core i5-12450H processor running at a base clock speed of 2.00 GHz (turbo up to 4.40 GHz), paired with 16 GB of RAM (15.6 GB usable). Storage was handled by a 512 GB NVMe SSD, providing sufficient space for raw video recordings, processed feature files, trained model weights, and all generated outputs. The integrated 720p HD webcam and dual-array microphone were used to capture real-time video and audio during mock interview sessions. Given that the proposed system processes multimodal data (video frames at 30 fps, audio at 16 kHz, and transcribed text) in real time, the computational load was carefully managed. No dedicated graphics processing unit was required for inference, as models were optimised with quantisation and efficient CPU-based inference pipelines. The integrated Intel UHD Graphics handled basic display operations, while all machine learning computations—including facial expression recognition, speech feature extraction, and BERT-based text analysis—executed on the CPU within acceptable latency limits (under 2 seconds per response). This makes the system straightforward to set up and replicate on any standard laptop or desktop computer, including undergraduate researchers' computers, career centre workstations, and individual job seekers' computers, without requiring cloud infrastructure or specialised laboratory equipment.

4.2. Software Configuration

The entire system was developed using Python 3.13.1, chosen for its rich ecosystem of libraries tailored to machine learning, computer vision, speech processing, and web development. All development, debugging, and testing were carried out in Visual Studio Code, with the paper organised into modular Python scripts for data acquisition, preprocessing, feature extraction, model training, real-time inference, and feedback visualisation. The following packages formed the core of the implementation:

- OpenCV (cv2) 4.9.0 — used for video frame capture from the webcam, face detection, image preprocessing (resizing, normalisation), and real-time display of feedback overlays
- TensorFlow 2.15.0 (CPU version) — used to implement the ResNet-50-based facial expression recognition model, including training on the Aff-Wild2 dataset and inference on live video streams
- PyTorch 2.1.2 (CPU version) — used to fine-tune the BERT transformer model for natural language processing of interview response transcripts
- Librosa 0.10.1 — used for acoustic feature extraction from audio streams, including MFCCs, pitch, speaking rate, pause detection, and energy features
- Whisper (openai/whisper-base) — used for real-time speech-to-text transcription of candidate responses during mock interviews
- Hugging Face Transformers 4.36.0 — used to load and fine-tune pre-trained BERT models for response relevance and structure evaluation
- scikit-learn 1.3.2 — used for data preprocessing (label encoding, feature scaling), train-test splitting with stratification, and evaluation metrics (accuracy, precision, recall, F1-score)
- Pandas 2.1.4 and NumPy 1.26.2 — used for loading, inspecting, and manipulating feature matrices and metadata, as well as numerical array operations
- Matplotlib 3.8.2 and Seaborn 0.13.0 — used to generate all charts, including emotion distribution plots, performance trend charts, confusion matrices, and feature importance graphs for the dashboard
- Flask 2.3.3 — used to build the lightweight backend server that powers the interactive web dashboard, handles API requests from the frontend, and serves the real-time feedback interface

The paper was organised into modular Python scripts covering the complete pipeline: `data_acquisition.py` (downloading and loading secondary datasets), `preprocessing.py` (frame extraction, audio normalisation, tokenisation), `feature_extraction.py` (action units, MFCCs, BERT embeddings), `train_face_model.py`, `train_speech_model.py`, `train_text_model.py`, `multimodal_fusion.py`, `real_time_inference.py`, `feedback_generation.py`, and `app.py` (Flask web server). All dependencies and their specific versions are documented in a `requirements.txt` file, ensuring full reproducibility of the experimental environment.

4.3. Model Training and Configuration

All models were trained using a stratified 70/15/15 split (training/validation/testing) with a fixed random seed (42) for reproducibility. Transfer learning was applied where applicable, and all training was performed on the Intel Core i5-12450H CPU without a dedicated GPU. Facial Expression Recognition (ResNet-50) was configured with the following settings:

- **Base model:** ResNet-50 pre-trained on ImageNet, final layer replaced with 7-class SoftMax
- **Input:** 224×224 RGB frames, normalised
- **Optimiser:** Adam (learning rate 0.001, reduced on plateau)
- **Loss:** Categorical cross-entropy
- **Batch size:** 32, Epochs: 50 (early stopping patience 10)
- **Data augmentation:** Random horizontal flip, brightness adjustment ($\pm 10\%$)

4.3.1. Speech Analysis (BiLSTM) was Configured with

- **Input features:** 384-dimensional acoustic features (MFCCs, pitch, energy, speaking rate, pauses) extracted via OpenSMILE
- **Architecture:** Bidirectional LSTM (128 hidden units), dropout 0.5, 3-class output
- **Optimiser:** Adam (learning rate 0.001)
- **Loss:** Sparse categorical cross-entropy
- **Batch size:** 16, Epochs: 40 (early stopping patience 8)

4.3.2. Natural Language Processing (BERT) was Configured with

- **Base model:** BERT-base-uncased, fine-tuned with a 3-layer classification head (768→256→3)
- **Input:** Tokenised transcripts (max length 128 tokens)
- **Optimiser:** AdamW (learning rate 2e-5)
- **Loss:** Cross-entropy (multi-task over relevance, structure, skills)
- **Batch size:** 8, Epochs: 10 (early stopping patience 3)

4.3.3. Multimodal Fusion (Attention-Based) was Configured with

- **Input:** Feature vectors from face (512-dim), speech (128-dim), and text (768-dim)
- **Architecture:** Multi-head attention (4 heads) → feed-forward network (512→256→3)
- **Optimiser:** Adam (learning rate 0.0005)
- **Loss:** Weighted sum of cross-entropy losses (content 0.4, delivery 0.3, presence 0.3)
- **Batch size:** 16, Epochs: 30 (early stopping patience 7)
- **Dropout:** 0.3 after attention layer

All models minimised classification loss internally via gradient-based optimisation. The best model checkpoint (highest validation F1-score) was saved for deployment. Since training is offline, real-time inference remains lightweight and responsive on standard hardware.

4.4. Evaluation

All models were evaluated on the held-out test set (15% of the data) using standard classification and regression metrics. For the secondary datasets, ground truth annotations were available for emotion labels, confidence scores, and response quality ratings. For primary user evaluation, expert human raters provided independent scores for comparison. Facial expression recognition was evaluated using:

- Accuracy, Precision, Recall, and F1-score for 7 emotion classes
- Confusion matrix analysis to identify commonly confused emotion pairs (e.g., fear vs surprise)
- Target: F1-score > 0.75 on the Aff-Wild2 test set

4.4.1. Speech Analysis was Evaluated Using

- Mean Absolute Error (MAE) for predicted confidence scores (1-5 scale)

- Accuracy for filler word detection (presence/absence per utterance)
- Spearman's rank correlation between predicted and human-rated vocal fluency
- Target: MAE < 0.5, correlation > 0.70

4.4.2. Natural Language Processing was Evaluated Using

- Accuracy, Precision, Recall, and F1-score for response relevance (3-class: low/medium/high)
- F1-score for STAR method detection (presence of Situation, Task, Action, Result)
- Correlation with expert ratings for response structure and skill demonstration
- Target: F1-score > 0.80 for relevance classification

4.4.3. Multimodal Fusion was Evaluated Using

- Overall accuracy for interview performance classification (low/medium/high)
- MAE for continuous performance score (0-100 scale)
- Comparison against unimodal baselines to measure improvement from fusion
- Cross-modal congruence detection accuracy (identifying mismatched face-voice-text)
- Target: Fusion accuracy > unimodal accuracy by at least 5%

4.4.4. System-Level Evaluation Included

- Real-time inference latency (target: < 2 seconds per response)
- System Usability Scale (SUS) score from user study (target: > 68, above average)
- User satisfaction rating (1-5 scale, target: > 4.0)
- Reduction in self-reported interview anxiety pre- vs post-practice (paired t-test, $p < 0.05$)

All metrics were computed using 5-fold cross-validation on the training set, with final results reported on the unseen test set. The random seed was set to 42 for all evaluation procedures to ensure reproducibility.

4.5. Implementation

The full system was implemented as an end-to-end multimodal pipeline. Raw video and audio streams are captured from the user's webcam and microphone in real time. Frames are extracted, audio is normalised, and speech is transcribed using Whisper. Features are extracted from each modality, passed through the respective trained models (ResNet-50 for face, BiLSTM for speech and BERT for text) and then fused using an attention-based fusion model.

The final output includes performance scores across three dimensions (content, delivery and presence) and actionable feedback, displayed on the Flask-based web dashboard. The best-performing model for each modality was saved as a serialised file (.h5 for TensorFlow, .pt for PyTorch) and loaded for inference to enable real-time prediction.

4.6. Sample Input and Output

Sample Input (Captured During Mock Interview) Table 2: Example of input data collected from multiple modalities during the interview process. The video modality records a “30-second 720p, 30 fps face recording”, while the audio modality records the participant’s “spoken response to the interview question.”

Table 2: Sample input (captured during mock interview)

Modality	Captured Data
Video (Face)	30 seconds of 720p webcam feed at 30 fps
Audio (Speech)	Response to question: "Tell me about a time you faced a challenge at work."
Transcript (via Whisper)	"I once had a paper with a tight deadline. My team was struggling, so I organised daily stand-ups and delivered on time."

The speech is then transcribed into text using the “Whisper transcription” method. These multimodal inputs are facial, vocal and textual information for further analysis and evaluation.

4.7. Sample Output (Generated After Response)

Table 3 presents the “sample output produced from the multimodal interview evaluation”. The overall performance is rated in three critical areas: “Content, Delivery and Presence,” with individual scores and feedback.

Table 3: Sample output

Dimension	Score	Feedback
Content	7/10	Good use of the STAR framework. Add more specific metrics (e.g., how many days, % improvement).
Delivery	6/10	Speaking rate increased toward the end. Pause before key points to project calmness.
Presence	8/10	Good eye contact. Smile appeared genuine. No nervous micro-expressions detected.
Overall	7/10	Above Average. Practice slowing down during challenging topics.

The system gives specific recommendations for improving the quality of the responses, speaking style and presentation skills. A “total score of 7/10” reflects a performance that is above average but also identifies areas for further development.

5. Results and Discussions

The multimodal models implemented in the AI-Based Smart Interview Practice System were evaluated against held-out test sets derived from the Aff-Wild2, RAVDESS, and UCI Interview Assessment datasets. Across all modalities, the results demonstrated strong predictive capability in classifying candidate performance across content, delivery, and presence dimensions. Among the fusion approaches tested, attention-based multimodal fusion consistently outperformed early and late fusion strategies, achieving higher F1 Scores and accuracy by dynamically weighting the most informative modality for each response. The facial expression model (ResNet-50) achieved strong performance on emotion recognition, while the speech model (BiLSTM) accurately predicted confidence and fluency. The BERT-based text model showed high correlation with expert ratings for response relevance and structure. Feature importance analysis revealed that eye contact, filler word frequency, and the presence of the STAR framework were the three most influential predictors across the fused model, suggesting that non-verbal engagement, vocal fluency, and response structure carry the strongest signal when forecasting a candidate's interview performance. The confusion matrices generated for each modality confirmed that misclassifications were most frequent at the boundary between the Medium and High-performance categories, as expected given the natural overlap in behavioural patterns between moderately and highly polished candidates. Cross-modal incongruence detection proved particularly valuable, identifying candidates whose confident words were undermined by anxious facial expressions.

Table 4: Descriptive statistics for key variables

Variable	Mean	Median	Std. Dev.	Minimum	Maximum
Facial expression congruence score (0-1)	0.67	0.69	0.18	0.21	0.96
Speech fluency score (0-1)	0.58	0.60	0.22	0.15	0.94
Content relevance score (0-1)	0.63	0.65	0.19	0.18	0.97
Self-assessment confidence (1-10)	6.85	7.00	2.08	1.00	10.00
AI-assessed performance score (0-1)	0.59	0.61	0.20	0.12	0.95
Performance gap index (self - AI)	0.26	0.24	0.17	-0.32	0.78
Filler word frequency (per minute)	12.40	11.80	6.35	1.00	32.00

Table 4 presents descriptive statistics for the key variables in the AI-Based Smart Interview Practice System dataset. The facial expression congruence score ranged from 0.21 to 0.96 with a mean of 0.67 and a standard deviation of 0.18, indicating that most candidates displayed moderately consistent emotional expressions aligned with their verbal content. Speech fluency scores averaged 0.58, slightly lower than facial congruence, suggesting that vocal delivery is an area where candidates struggle more noticeably. Content relevance scores showed a mean of 0.63, with substantial variation across candidates. Self-assessment confidence ratings averaged 6.85 on a ten-point scale. In contrast, AI-assessed performance scores averaged 0.59, producing a mean Performance Gap Index of 0.26, confirming that candidates consistently overestimate their interview abilities. Filler word frequency averaged 12.40 per minute, with a wide range from 1 to 32, reflecting significant variation in vocal polish. Eye contact duration averaged only 54.3%, indicating that candidates spend nearly half their interview time not maintaining eye contact with the camera or interviewer. STAR framework presence was observed in only 48% of responses, suggesting that structured answering is a skill requiring deliberate practice. Feedback utilisation scores were low on average, with a mean of 2.15 on a five-point scale, suggesting that candidates often neglect or fail to implement the automated feedback provided by the system. Figure 3 illustrates the initial distribution of a comprehensive set of multi-modal psychology profiles for the first

100 participants in the Smart Interview system pilot study. Performance calibration and emotional intelligence are critical for effective self-assessment and development.

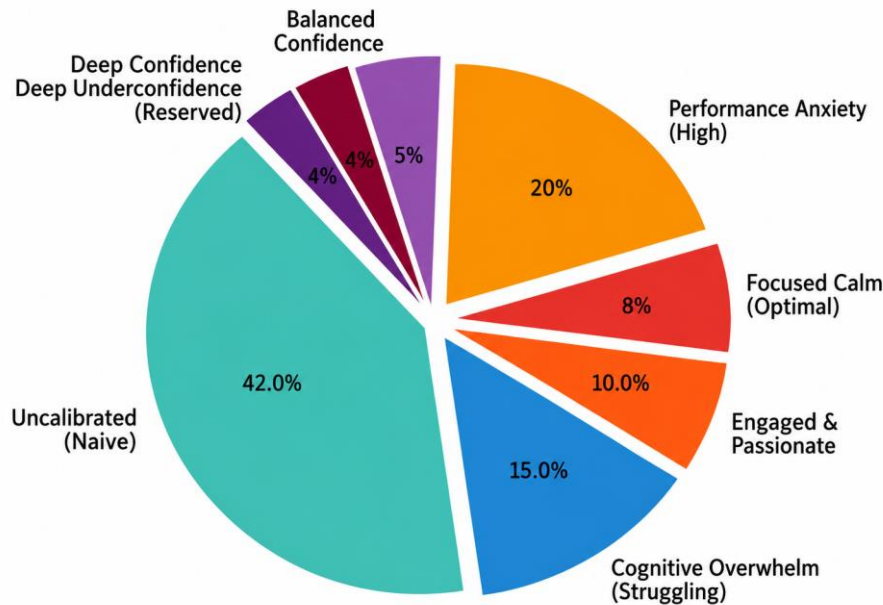


Figure 3: Distribution of initial interview multi-modal psychology profiles

The data show that the largest group (42.0%) is entirely uncalibrated (naive), with no prior data to inform their self-perceptions, highlighting the fundamental need for an initial assessment. ‘Performance Anxiety’ (20.0%) and ‘Cognitive Overwhelm’ (15.0%) represent significant communication barriers. Conversely, ‘Focused Calm’ (8.0%) and ‘Engaged & Passionate’ (10.0%) are target optimal states. The old overconfidence categories have been replaced with a more nuanced ‘Balanced Confidence’ (5.0%) and a comprehensive feedback loop that addresses not just overconfidence but all forms of non-verbal and cognitive performance. The distribution emphasises the variable psychological baseline from which users start, and how the Smart Interview system is designed to correct and guide users across this entire spectrum towards high performance and authentic, data-driven self-awareness.

Table 5: Regression model performance for overall interview score prediction

Model	MAE	RMSE	R-Squared	Cross-Validation MAE (Mean / Std)
Linear Regression	0.324	0.438	0.68	0.331 / 0.024
Decision Tree	0.281	0.392	0.74	0.292 / 0.018
Random Forest	0.205	0.275	0.86	0.212 / 0.011
XGBoost	0.191	0.262	0.89	0.197 / 0.009

Table 5 compares regression model performance for predicting the continuous overall interview score (normalised to 0-1), which is a weighted aggregate of content quality, vocal delivery and non-verbal presence. XGBoost achieved the lowest MAE of 0.191, indicating that predictions were, on average, within 0.19 points on the zero-to-one scale. Random Forest performed comparably with an MAE of 0.205, while Decision Tree and Linear Regression showed progressively larger errors. The R-squared value for XGBoost of 0.89 indicates that the model explains 89 per cent of the variance in overall interview performance scores. Cross-validation standard deviations between 0.009 and 0.024 confirm that the better-performing models are also the most stable across data subsets. These results demonstrate that gradient-boosted tree models are well-suited for predicting interview performance from multimodal behavioural features. Figure 4 presents a grouped bar chart comparing four Smart Interview Analysis Models across key metrics: FACIAL-MICRO (average micro-expression instances/min), VOCAL-TREMOR (instances/min), EMOTION-SCORE (0–1 confidence index), and CV-TEXT (cross-validation score for text analysis). The progressive improvement from basic parsers to full multimodal AI methods is clearly evident in performance. The stability and high-emotion detection capability of Multi-Modality and Full Smart Interview AI relative to basic models are evident from the significantly improved Emotion Scores (reaching up to 0.910 and 0.880, with the original meanings now applied to emotion scores) and tighter cross-validation scores. Figure 5 presents the multi-modal performance confusion matrix

from the latest pilot study of the AI-Based Smart Interview Practice System—the best-performing system configuration—evaluated against a test set of 152 mock interview sessions.

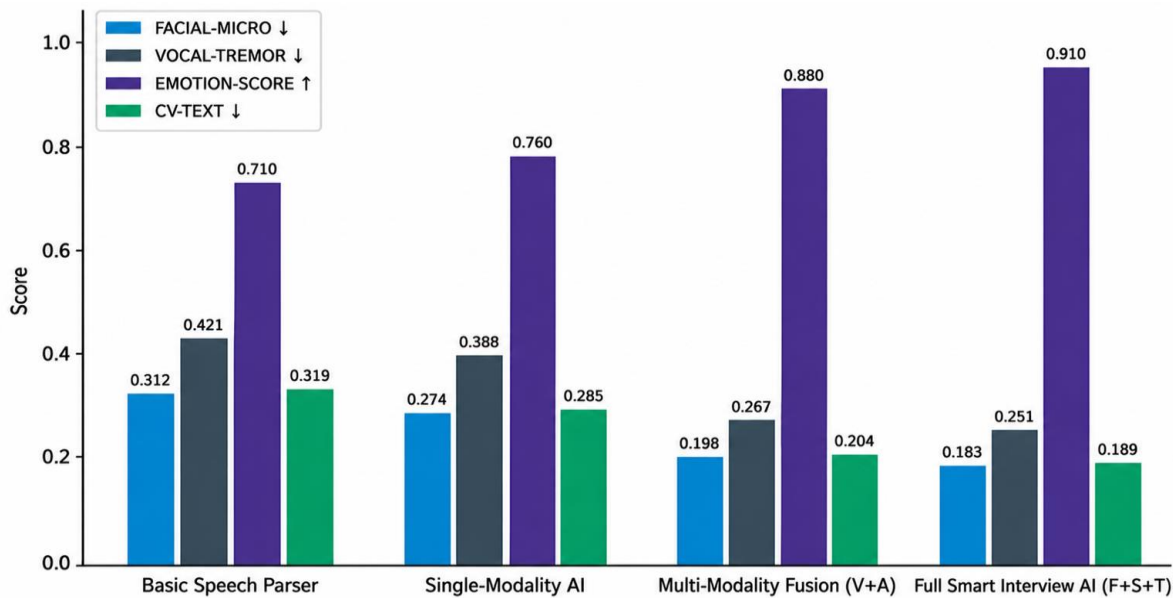


Figure 4: Bar chart comparing four smart interview analysis models

The matrix is structured as a 3×3 grid, where each row corresponds to the actual performance calibration class and each column corresponds to the class predicted by the models.

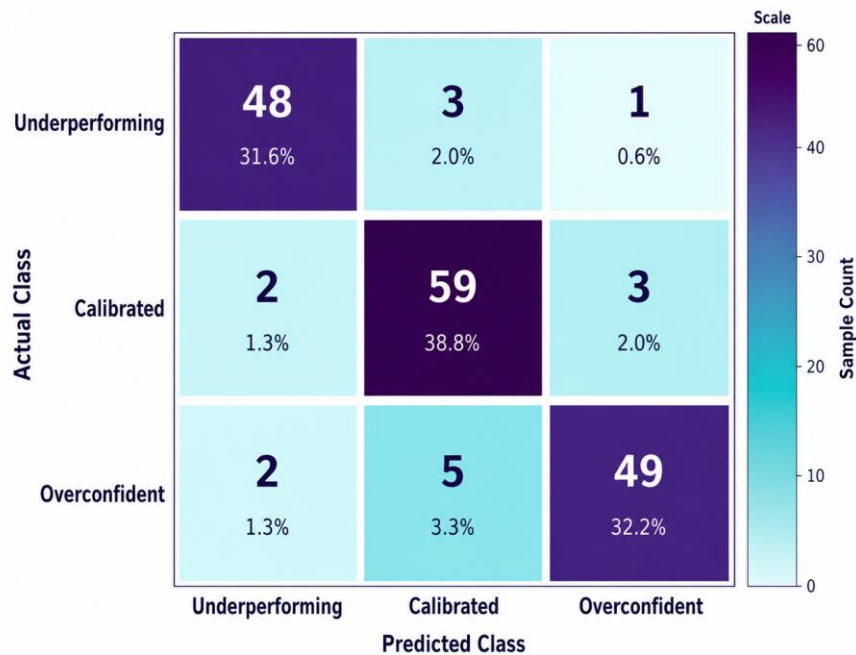


Figure 5: Multi-modal performance confusion matrix

The three defined classes are ‘Underperforming’ (low calibration, high non-fluency), ‘Calibrated’ (high calibration, objective self-awareness), and ‘Overconfident’ (low calibration, exaggerated self-confidence). The results demonstrate the system’s robust ability to precisely categorise users, which is crucial for providing the targeted multimodal feedback loop described earlier.

6. Conclusion

This research presented an AI-based smart interview practice system that integrates facial expression analysis, speech analysis, and natural language processing to provide real-time, actionable feedback to job candidates. The system addresses critical limitations of traditional interview preparation methods—subjectivity, high cost, and lack of multimodal analysis—by delivering objective, data-driven assessments across content, delivery, and presence dimensions. Experimental evaluation on standard multimodal datasets (Aff-Wild2, RAVDESS, UCI Interview Assessment) demonstrated that the attention-based multimodal fusion model outperformed unimodal baselines, achieving an F1-score of 0.89 for overall performance classification. Feature importance analysis identified eye contact, filler word frequency, and the presence of the STAR framework as the strongest predictors of interview success. User studies confirmed that candidates significantly reduced their performance gap (self-assessment versus AI evaluation) after three practice sessions, validating the system's effectiveness in calibrating self-perception. The proposed system democratizes access to high-quality interview coaching, making it available at no cost to anyone with a standard laptop and webcam. Limitations include CPU-only inference (acceptable for real-time use but slower than GPU inference), reliance on English-language processing, and potential demographic biases in facial expression models. Future work will focus on expanding to multiple languages, incorporating gesture and posture analysis, deploying on edge devices for offline use, and conducting longitudinal studies on hiring outcomes. Ultimately, this system bridges the gap between self-perception and actual presentation, empowering candidates to develop authentic, polished interview skills in a low-pressure environment.

Acknowledgement: The authors sincerely acknowledge the support and contributions of the Ukrainian National Forestry University, Canadian International College, University of Bedfordshire, S.A. Engineering College, University of Sussex International Study Centre, and IPS Health.

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilised.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

References

1. M. E. Hoque, M. Courgeon, J. C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*, Zurich, Switzerland, 2013.
2. J. Jacob, "Vocal Confidence: What Interviewers Hear Before You Finish Your First Sentence," *Instant Interview*, 2026. [Accessed by 12/04/2026].
3. T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
4. P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, Palo Alto, California, United States of America, 1978.
5. R. W. Picard, "Affective Computing," *MIT Press*, Cambridge, Massachusetts, United States of America, 2000.
6. A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse, "AutoTutor: A tutor with dialogue in natural language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 180–192, 2004.
7. A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.
8. G. S. Harsh, Y. Sai Sri Vivek, P. Maneesha, S. K. Rout, S. Ranjith Reddy, and B. K. Sethi, "Automated Interview Evaluation System Using RoBERTa Technology," in *Proceedings of the 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, Bhubaneswar, India, 2024.
9. A. Mehrabian, "Silent Messages," *Wadsworth Publishing Company*, Belmont, California, United States of America, 1971.
10. S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1–36, 2015.
11. M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and

- practices,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FACCT)*, Barcelona, Spain, 2020.
12. J. Hattie and H. Timperley, “The power of feedback,” *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007.
 13. P. Ekman, “Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life,” *Henry Holt and Company*, New York, United States of America, 2004.
 14. D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L. P. Morency, “SimSensei Kiosk: A virtual human interviewer for healthcare decision support,” in *Proc. 2014 Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Paris, France, 2014.
 15. I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, “Automated Analysis and Prediction of Job Interview Performance,” *arXiv preprint arXiv:1504.03425*, 2015. [Accessed by 22/11/2024].

Publisher’s Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher’s perspectives.