A Four-Track Multimodal Pipeline for Scalable Video Interview Assessment: Facial Affect, Gaze, Paralinguistic Features, and Zero-Shot NLP Fusion with Explainable Job Matching

Hafiz M. Abdullah, Asghar Ali, and Muhammad Bilal
Department of Computer Science, FAST School of Computing,
National University of Computer and Emerging Sciences (FAST-NUCES), Karachi Campus, Pakistan
Corresponding author: Hafiz M. Abdullah (k224489@nu.edu.pk)

This work was supervised by Engr. AbdurRahman (Project Supervisor), FAST School of Computing, FAST-NUCES Karachi. Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science, Spring 2026.

ABSTRACT Modern recruitment processes increasingly demand scalable, objective, and data-driven mechanisms for evaluating candidates. This paper presents the Multimodal Interview Analysis (MMIA) system, a fully implemented and production-ready artificial intelligence platform that automatically evaluates recorded video interviews across four complementary dimensions: facial emotion recognition, gaze and head pose estimation, acoustic and paralinguistic feature extraction, and natural language content analysis. Video frames are processed by a ResNet-50 backbone pre-trained on AffectNet (450,000 faces) and temporally smoothed by an LSTM network trained on Aff-Wild2. A total of 468 three-dimensional facial landmarks are extracted using MediaPipe FaceMesh and consumed by the solvePnP routine in OpenCV to produce per-frame Euler head-pose angles along with an eye-contact percentage. Audio is stripped using moviepy, transcribed by OpenAI Whisper Base, and then analysed in parallel by the librosa signal-processing library (pitch via probabilistic YIN, RMS energy, zero-crossing rate, and 13 MFCCs) alongside two Hugging Face transformer models: RoBERTa for three-class sentiment and BART-MNLI for zero-shot communication-style classification across eight labels. A transparent and fully traceable weighted fusion formula combining Emotion (30%), Gaze (20%), Acoustic (25%), and Verbal (25%) produces a single composite score on a scale of 0 to 100, complete with a four-tier grade and natural-language recruiter highlights. The platform also incorporates an Explainable Job Matcher (XAI module) that leverages gradient-boosted regression and semantic skill matching via the all-MiniLM-L6-v2 sentence-embedding model to produce interpretable student-to-job compatibility scores together with human-readable gap analyses and action plans. The system processes a one-minute interview in 25 to 30 seconds on CPU, achieving 3 to 5 times faster performance with CUDA acceleration. All candidate data are discarded after analysis, ensuring complete privacy preservation. Evaluation confirms that the scoring formulas faithfully reflect ground-truth recruiter judgements, and the XAI module achieves an R^2 of 0.94 on held-out data. Together, these components directly address transparency, bias reduction, scalability, and privacy, which represent the four principal obstacles to responsible AI adoption in recruitment.

INDEX TERMS multimodal interview analysis, facial emotion recognition, acoustic feature extraction, natural language processing, explainable AI, job matching, ResNet-50, LSTM, BART-MNLI, RoBERTa, candidate assessment, decision-level fusion, semantic skill matching

I. Introduction

Recruitment has always been one of the most consequential decision-making activities within any organisation. In an era where candidate pools routinely number in the thousands, human interviewers face an impossible trade-off between thoroughness and efficiency. Traditional in-person and live video interviews suffer from well-documented cognitive biases including confirmation bias, affinity bias, and halo effects, all of which reduce both fairness and predictive validity [12]. At the same time, purely algorithmic shortlisting tools that rely on keywords or rule-based criteria fail to capture the rich behavioural signals that expert human interviewers attend to, such as vocal confidence, facial engagement, emotional

congruence, and communicative fluency.

Multimodal analysis has emerged as a principled response to this tension [1]. By integrating verbal and nonverbal communication signals within a single computational framework, a system can replicate and in some respects surpass the depth of a skilled human assessor, while operating at scale and with algorithmic consistency. Prior work has established that acoustic features correlate with perceived confidence [7], that facial expressions carry predictive information about interview outcomes [9], that eye contact influences interviewer ratings [10], and that NLP-based sentiment and filler-word counts serve as reliable proxies for speech fluency [8]. However, most existing implementations address only one or two of these channels simultaneously, rely on closed proprietary systems, or lack transparent scoring formulas that recruiters can audit.

This paper makes the following contributions:

- A complete, open, and production-deployed multimodal interview analysis pipeline (MMIA) covering all four established behavioural dimensions of candidate assessment, namely emotion, gaze, acoustics, and verbal content, implemented as a FastAPI microservice.
- A fully transparent and mathematically grounded weighted fusion scoring model whose every output is traceable to raw candidate behaviour, directly addressing the black-box criticism commonly directed at AI recruitment tools.
- An explainable job-matching module (XAI) that augments candidate evaluation with gradient-boosted regression and semantic skill embeddings, producing structured human-readable explanations alongside match scores.
- Empirical validation of the composite scoring formula against recruiter-aligned benchmarks, and cross-validated performance metrics for the XAI regression model.
- A privacy-by-design architecture that retains no candidate data after processing, satisfying GDPR-aligned requirements.

The remainder of this paper is organised as follows. Section II reviews relevant literature. Section III states system requirements. Section IV describes the overall architecture. Section V details each implemented component. Section VI reports evaluation results. Section VII addresses ethics and bias. Section VIII concludes.

II. Related Work

A. Theoretical Foundations of Multimodal Communication

Communication is a complex and multifaceted process involving spoken language, facial expressions, gestures, posture, tone of voice, and physiological signals [3]. Mehrabian's seminal experiments demonstrated that in cases of incongruent verbal and nonverbal cues, receivers tend to weight nonverbal channels more heavily [2]. The Multimodal Interaction Model [4] further emphasises the dynamic and integrative nature of these channels, while neurocognitive evidence confirms that multimodal stimuli enhance comprehension and affective processing beyond what unimodal inputs can achieve [5].

In the recruitment context, where first impressions can determine a candidate's fate, this interplay is particularly significant. The theoretical foundation motivating the MMIA design is therefore one of complementary channels: each modality yields information the others cannot provide, and only their fusion delivers a holistic candidate profile [1].

B. Verbal Communication and Acoustic Analysis

Verbal communication has traditionally been the cornerstone of candidate evaluation [6]. Beyond semantic content, paralinguistic features such as pitch, loudness, intonation, and speech rate carry emotional weight and personality signals [7]. Hesitation markers like "um", "uh", and "like" correlate with reduced perceived fluency and preparation [8]. Transformer-based models such as BERT and RoBERTa [17] enable context-aware semantic analysis and sentiment detection with state-of-the-art accuracy, while zero-shot classifiers built on natural-language inference models like BART-MNLI [18] allow communication-style classification without domain-specific fine-tuning.

C. Nonverbal Cue Analysis

Facial expressions are among the most thoroughly studied nonverbal cues in interview research. Ekman and Friesen's Facial Action Coding System (FACS) [9] systematically catalogues the muscle movements underlying six universal emotions, namely happiness, fear, anger, surprise, sadness, and disgust, plus a neutral state. Deep convolutional networks, particularly ResNet architectures [13] trained on large labelled datasets such as AffectNet [14], achieve near-human accuracy on per-frame emotion classification. Temporal smoothing via LSTM networks mitigates frame-level jitter caused by micro-expressions and lighting artefacts [15].

Eye gaze is a critical engagement signal: sustained eye contact is associated with higher interviewer ratings [10], while gaze aversion can trigger impressions of evasiveness. Landmark-based head-pose estimation using OpenCV's Perspective-n-Point algorithm provides Euler angles (pitch, yaw, roll) from which camera gaze can be inferred without additional hardware [16].

D. Multimodal Fusion Strategies

Three canonical fusion strategies exist: feature-level, decision-level, and hybrid [21]. Feature-level fusion concatenates raw feature vectors before classification, capturing inter-modal correlations at the cost of dimensionality and synchronisation complexity. Decision-level fusion processes each modality independently and then combines the resulting scores, offering modularity and interpretability [22]. Hybrid approaches combine both layers for optimal performance [23].

The MMIA system deliberately adopts decision-level fusion to maintain per-modality transparency, enabling recruiters to inspect and override individual dimension scores when needed.

Multimodal sentiment analysis frameworks [19] and multimodal transformers that capture cross-modal temporal dynamics [20] represent the current state of the art in research settings but impose heavy computational requirements and reduced transparency. The MMIA design therefore prioritises deployability and auditability over marginal accuracy gains.

E. Explainable AI in High-Stakes Decision-Making

The black-box problem is especially acute wherever automated systems produce decisions that directly affect individuals' lives. In finance, AI systems for credit scoring, insurance underwriting, and fraud detection have been shown to perpetuate historical biases when trained on unrepresentative data [29], [30]. Explainable AI techniques such as LIME [24] and SHAP [25] have emerged as leading approaches for post-hoc explanation of arbitrary model outputs. Recent surveys confirm that XAI adoption significantly improves stakeholder trust and enables legal compliance with right-to-explanation provisions in regulations such as the EU's GDPR [26], [27].

In the recruitment domain, Raji et al. [28] document how opaque AI hiring tools have drawn considerable regulatory scrutiny. The MMIA platform addresses this directly: every component of the composite score is derived from a transparent formula, and the XAI job-matching module surfaces feature importances and natural-language rationales rather than bare numeric scores.

F. Bias and Fairness

Algorithmic bias arising from non-representative training data is a central concern in AI-driven hiring [31]. Buolamwini and Gebru [32] documented significant accuracy disparities in commercial face-analysis systems across demographic groups. Fairness-enhancing techniques such as adversarial debiasing, disparate impact analysis, and fairness constraints have been proposed to mitigate these risks [33]. The MMIA platform mitigates bias through uniform and fixed scoring weights applied identically to every candidate, and by relying on pre-trained models validated on demographically diverse benchmarks, with AffectNet alone covering 450,000 labelled faces.

G. Practical Implementation Gap

Despite the rich theoretical literature, a notable gap exists between research prototypes and production-ready, auditable multimodal interview analysis systems. Most academic systems process data offline in notebook environments, lack structured API interfaces, and do not address privacy-preserving data handling. The MMIA system directly fills this gap by implementing all four modalities end-to-end in a FastAPI microservice with stateless, privacy-preserving data handling and structured JSON outputs suitable for integration into existing Applicant Tracking Systems (ATS).

III. System Requirements

A. Functional Requirements

The following functional requirements were derived from the research literature, project roadmap, and the reference paper by Alexander [1].

FR-1: The system shall accept an MP4 video recording via a REST API endpoint together with candidate name and role metadata.

FR-2: The system shall detect faces in every video frame and extract 468 three-dimensional facial landmarks per face.

FR-3: The system shall classify per-frame facial emotion into seven categories and apply temporal smoothing over a sliding window.

FR-4: The system shall estimate head pose (pitch, yaw, roll) and compute the eye-contact percentage per video.

FR-5: The system shall extract the audio stream, transcribe it to text with timestamps, and detect the spoken language automatically.

FR-6: The system shall extract paralinguistic features including fundamental frequency, RMS energy, zero-crossing rate, and MFCCs directly from the audio waveform.

FR-7: The system shall classify transcript sentiment and communication style, and count filler words.

FR-8: The system shall compute a weighted composite score from 0 to 100, assign a grade, and generate recruiter highlights.

FR-9: The system shall support batch submission of multiple videos for sequential processing.

FR-10: The system shall delete all temporary files immediately after processing and retain no candidate data.

FR-11: The XAI module shall match a student profile to a job posting and return a weighted match score together with a human-readable explanation and action plan.

B. Non-Functional Requirements

Performance: The system shall process a one-minute video in under 60 seconds on a CPU-only host, with a target of 25 to 30 seconds.

Privacy: No candidate audio, video, or derived features shall be written to persistent storage.

Transparency: Every numeric contribution to the composite score shall be fully traceable to a documented formula.

Scalability: The FastAPI backend shall support asynchronous file upload and stateless processing, permitting horizontal scaling.

Portability: The system shall run on any host with Python 3.8 or higher and standard POSIX-compatible libraries; GPU acceleration shall remain optional.

C. Use Cases

UC-1 — Single-Candidate Assessment: A recruiter submits a recorded video interview. The system returns a structured JSON report containing the composite score, a four-dimensional breakdown, and recruiter highlights within 30 seconds.

UC-2 — Batch Screening: An HR team submits twenty videos via the batch-analyze endpoint. The system processes them sequentially and returns an array of reports, allowing side-by-side ranking.

UC-3 — Job-Match Explanation: A career-services counsellor submits a student profile and a target job description to the XAI module. The system returns a compatibility score, a breakdown of matched and missing skills, and a prioritised three-horizon action plan.

IV. System Design

A. High-Level Architecture

The MMIA platform is decomposed into four concurrent analysis tracks and a fusion layer, all orchestrated by a FastAPI application server. Fig. 1 illustrates the full pipeline.

Input: MP4 video file
Track A — Facial Pipeline
MediaPipe FaceMesh → 468 landmarks/frame
ResNet-50 → per-frame 7-class emotion
LSTM (10-frame window) → smoothed emotion timeline
OpenCV solvePnP → pitch/yaw/roll, eye-contact %
Track B — Audio Extraction
moviepy → 16 kHz PCM mono WAV
Track C — Acoustic Pipeline
librosa → F0 (pYIN), RMS, ZCR, 13 MFCCs
Track D — Verbal Pipeline
Whisper Base → transcript + timestamps
RoBERTa → 3-class sentiment
BART-MNLI → 8-label communication style
Regex detector → filler-word count & ratio
Fusion Layer
candidate_scorer.py → composite score (0–100)
Grade + recruiter highlights
Output: Structured JSON response

Fig. 1. MMIA end-to-end processing pipeline.

B. Facial Analysis Sub-system

The facial sub-system operates frame by frame. For each frame, MediaPipe FaceMesh provides 468 landmarks in normalised image coordinates. The `FaceDetector` class converts these to pixel coordinates and extracts a bounding box for the face region. The face crop is resized to 224×224 pixels and fed to the ResNet-50 backbone, which produces a feature embedding. The `EmotionPredictor` class maintains a FIFO deque of length 10; the LSTM network reads the deque at each step and outputs a smoothed probability distribution over the seven emotion classes.

Head pose is estimated using six stable 2D-to-3D correspondences (nose tip, chin, bilateral eye outer corners, and bilateral mouth corners) with a pinhole camera model whose focal length is set to the frame width. The `SOLVEPNP_ITERATIVE` flag in `cv2.solvePnP` recovers the rotation vector, subsequently decomposed into Euler angles via `cv2.RQDecomp3x3`. A geometric frontalness proxy based on the ratio of the nose's lateral offset to the inter-eye distance provides a fallback for frames where the solver encounters degeneracy. Eye contact is declared when $|\text{yaw}| < 22^\circ$ and $|\text{pitch}| < 20^\circ$, or when the geometric proxy satisfies its own threshold.

C. Acoustic Analysis Sub-system

Audio is extracted from video to a 16 kHz mono PCM WAV file. The acoustic analyser uses the probabilistic YIN algorithm in librosa to extract the fundamental frequency (F0), discarding unvoiced frames. RMS energy is computed over the full waveform, the zero-crossing rate serves as a speech-rate proxy, and 13 MFCC coefficients capture voice timbre. All features are reduced to scalar statistics before being passed to the scorer.

D. NLP Analysis Sub-system

The transcript analyser applies three sequential analyses to the Whisper transcript:

1) **Sentiment**: A RoBERTa-based sentiment pipeline (CardiffNLP) returns probabilities for POSITIVE, NEUTRAL, and NEGATIVE. The top label and score are retained.
2) **Communication style**: The BART large MNLI zero-shot classifier is invoked with eight candidate labels: confident, hesitant, enthusiastic, nervous, knowledgeable, vague, assertive, and uncertain. Multi-label inference is enabled so that a candidate may score highly on multiple styles simultaneously.
3) **Filler words**: A lexicon of 15 single-word and six multi-word hesitation markers is matched against the transcript; the filler ratio is the count divided by total word count.

Both transformer models are loaded from a local cache populated by a one-time download script, eliminating inference-time internet dependence.

E. Composite Scoring and Fusion

The scoring function implements decision-level fusion via a weighted average of four normalised dimension scores, each on the interval $[0, 1]$.

Emotion Score (E):

$$E = 0.6\, p^{+} + 0.4\,(1 - p^{-}) \tag{1}$$

where p^+ is the proportion of video time showing positive emotions (Happiness, Surprise) and p^- is the proportion showing negative emotions (Fear, Disgust, Anger, Sadness).

Gaze Score (G):

$$G = \frac{\text{eye-contact \%}}{100} \tag{2}$$

Acoustic Score (A):

$$A = \frac{1}{2}\left[\min\left(\frac{\sigma_{f_0}}{30}, 1\right) + \min\left(\frac{\overline{\text{RMS}}}{0.08}, 1\right)\right] \tag{3}$$

where σ_{f_0} is the standard deviation of voiced F0 in Hz and $\overline{\text{RMS}}$ is mean RMS energy.

Verbal Score (V):

$$V = 0.5\, s + 0.5\, \max(0,\, 1 - 5\, r_f) \tag{4}$$

where $s = 1.0$ (POSITIVE), 0.6 (NEUTRAL), or 0.2 (NEGATIVE), and r_f is the filler ratio.

Composite Score (C):

$$C = 100 \times (0.30\, E + 0.20\, G + 0.25\, A + 0.25\, V) \tag{5}$$

If any modality is unavailable (for example, if no speech is detected), its weight is redistributed proportionally among the remaining modalities, ensuring a valid score even under degraded conditions.

F. *Explainable Job Matching (XAI Module)*

The XAI module evaluates student-to-job compatibility along five dimensions: skills match (40%), education eligibility (20%), relevant projects (15%), relevant coursework (15%), and CGPA performance (10%).

Skill matching uses sentence embeddings from the all-MiniLM-L6-v2 model [34]. For each required skill, the cosine similarity between its embedding and each student skill embedding is computed; a match is accepted if cosine similarity exceeds 0.40. A proficiency multiplier further modulates the match credit based on required versus actual proficiency levels. The overall match score is computed by a `GradientBoostingRegressor` trained on synthetic but statistically representative data, with grid-searched hyperparameters evaluated via 5-fold cross-validation.

Human-readable explanations are generated by the feedback generator, which maps the match score and feature importances to structured natural-language statements covering strengths, gaps prioritised by severity, and a three-horizon action plan (immediate, short-term, and long-term).

G. *API Design*

The FastAPI application exposes six endpoints, as summarised in Table I.

All models are lazily loaded and cached as module-level singletons on first request, eliminating reload overhead for subsequent calls. Temporary video and audio files are explicitly deleted in `finally` blocks regardless of whether processing succeeds or raises an exception.

TABLE I
MMIA API Endpoints

Endpoint	Method	Purpose
`/`	GET	API info and health check
`/health`	GET	Model load status
`/analyze`	POST	Facial emotion only
`/analyze-with-transcription`	POST	Full MMIA (4 modalities)
`/analyze-with-transcription-new`	POST	Full MMIA with gaze
`/batch-analyze`	POST	Sequential batch processing

V. Implementation

A. *Technology Stack*

The backend is implemented in Python 3.8 and above. Table II summarises the principal libraries and models used throughout the system.

B. *Facial Emotion Recognition*

The `EmotionPredictor` class encapsulates two PyTorch models. The ResNet-50 backbone is loaded from pre-trained weights and placed in evaluation mode before inference. A custom preprocessing transform flips RGB to BGR and subtracts the AffectNet channel means before passing the 224×224 crop to the backbone. The backbone's feature extraction method produces a vector that is appended to a deque of length 10. The LSTM reads the full deque at every step and outputs a smoothed probability distribution over the seven emotion classes:

```
model = ResNet50(num_classes=7, channels=3)
state_dict = torch.load(
    MODEL_BACKBONE_PATH, map_location=device)
model.load_state_dict(state_dict)
model.eval()

lstm_input = np.vstack(list(self.lstm_features))
lstm_tensor = torch.from_numpy(
    lstm_input).unsqueeze(0).to(device)
output = self.lstm_model(lstm_tensor)
predicted_idx = np.argmax(
    output.cpu().numpy()[0])
emotion = EMOTION_LABELS[predicted_idx]
```

The LSTM acts as a temporal low-pass filter, suppressing single-frame mispredictions caused by transient occlusions or micro-expressions.

TABLE II
TECHNOLOGY STACK

Component	Library / Model	Purpose
API framework	FastAPI + Uvicorn	REST endpoints
Face detection	MediaPipe 0.10	468-landmark FaceMesh
Emotion (frame)	ResNet-50 (AffectNet)	7-class classification
Emotion (temporal)	LSTM (Aff-Wild2)	10-frame smoothing
Head pose	OpenCV solvePnP	Euler angle estimation
Audio extraction	moviepy	MP4 to WAV
Transcription	Whisper Base	ASR with timestamps
Acoustic features	librosa 0.10	F0, RMS, ZCR, MFCCs
Sentiment	RoBERTa (CardiffNLP)	3-class sentiment
Style classifier	BART-MNLI (Meta AI)	8-label zero-shot
Skill embedding	all-MiniLM-L6-v2	Semantic skill matching
Job-match ML	GradientBoostingRegressor	Match score regression
Deep learning	PyTorch 2.x	ResNet-50 and LSTM

C. *Head Pose and Eye Contact Estimation*

A minimal Perspective-n-Point problem is established with six facial anchors whose canonical 3D coordinates are known in a scaled face model. The six anchor points are the nose tip, chin, left and right eye outer corners, and left and right mouth corners. After solvePnP, the rotation vector is decomposed to yield Euler angles wrapped to $[-90°, 90°]$. Eye contact is computed as:

$$\text{eye-contact} = \mathbf{1}\Big[(|\text{yaw}| < 22° \wedge |\text{pitch}| < 20°) \vee (\delta_{\text{yaw}} < 0.45 \wedge \delta_{\text{pitch}} < 0.28)\Big] \tag{6}$$

where δ_{yaw} and δ_{pitch} are geometric proxy values derived from landmark ratios.

D. *Acoustic Feature Extraction*

The acoustic analyser extracts five feature groups:

1) **Pitch (F0)**: The probabilistic YIN algorithm with fmin $\approx$ 65 Hz and fmax $\approx$ 2093 Hz returns frame-level F0, a voiced flag array, and voiced probabilities. Statistics including mean, standard deviation, minimum, maximum, and range are computed over voiced frames only.
2) **Energy**: RMS energy computed over the full waveform using librosa.
3) **Speech rate proxy**: Mean value of the zero-crossing rate.
4) **Voice texture**: 13 MFCC coefficients averaged across time.
5) **Voiced ratio**: Proportion of frames where the voiced flag is True.

An interpretation layer maps raw statistics to human-readable labels such as high, moderate, or low confidence; high, moderate, or low energy; and expressive, moderate, or monotone pitch variety.

E. *NLP Analysis*

Both transformer pipelines are instantiated as module-level singletons with lazy loading. The sentiment pipeline loads the CardiffNLP RoBERTa model from a local cache and returns all three class scores with truncation to 512 tokens. The filler-word detector uses a two-pass approach: single tokens are matched against a 15-entry lexicon after stripping punctuation, and phrase occurrences are counted for six multi-word fillers such as “you know”, “i mean”, and “kind of”. The filler ratio feeds directly into the verbal score formula.

F. *Composite Scoring and Grade Assignment*

The scoring function applies Eqs. (1) through (5) after validating data availability for each modality. The grade-assignment thresholds are defined in Table III.

TABLE III
GRADE THRESHOLDS

Score Range	Grade	Interpretation
80 to 100	Excellent	Strong candidate
60 to 79	Good	Solid performer
40 to 59	Average	Acceptable with gaps
0 to 39	Needs Review	Human review recommended

Recruiter highlights are generated by a rule-based function that applies four conditions (for example, Happiness exceeding 30% of video, eye-contact exceeding 70%, filler ratio below 3%, and pitch standard deviation exceeding 20 Hz) to produce one to four plain-language observation sentences.

G. *Explainable Job Matching*

The job matcher computes five component scores. The skill score uses semantic matching: for each required skill, the cosine similarity between embeddings is compared against a threshold of 0.40. A proficiency multiplier scales the match credit, and a seniority multiplier based on the student’s batch year ranges from 1.00 times for freshmen up to 1.15 times for students four or more years into their programme.

The gradient-boosted regressor is trained with grid-searched hyperparameters (n_estimators in {100, 200}, learning_rate in {0.05, 0.1}, max_depth in {4, 5, 6}) evaluated via 5-fold cross-validation. Advanced feature engineering creates 15 derived features including technical strength, academic strength, profile balance, critical score, and polynomial interaction terms.

H. Privacy Architecture

All temporary files are managed via named temporary file objects with explicit deletion in `finally` blocks. No candidate audio, video, frame crops, embeddings, or scores are written to a database. The system is entirely stateless: two identical API calls with the same video will return identical results, and no session or candidate state is stored between calls.

VI. Evaluation

A. Evaluation Strategy

Evaluation covers four areas: component-level validation of the scoring formulas against recruiter-aligned benchmarks; performance profiling of the end-to-end pipeline; statistical evaluation of the XAI regression model; and a worked example demonstrating the full scoring trace.

B. Scoring Formula Validation

The composite scoring formula was validated by applying it to five controlled test recordings, each annotated by two independent human recruiters on a 0 to 100 scale. Table IV reports the mean absolute error (MAE) between the system score and the mean human score per dimension.

TABLE IV
Scoring Formula Validation Against Human Recruiters (5 Test Videos)

Dimension	MAE	Pearson r
Emotion	6.2	0.87
Gaze / Eye Contact	4.8	0.91
Acoustic	8.1	0.83
Verbal / NLP	5.4	0.89
Composite	4.9	0.92

The composite correlation of 0.92 indicates strong alignment between algorithmic and human assessments. The acoustic dimension shows the highest MAE (8.1), consistent with the known subjectivity of human vocal quality judgements.

C. Worked Scoring Example

Table V traces the composite score calculation for a representative candidate video, yielding a final score of 81.6 in the Excellent tier.

Emotion: $p^+ = 0.60$, $p^- = 0.05 \Rightarrow E = 0.36 + 0.38 = 0.74$. *Acoustic*: $\sigma_{f_0} = 28\,\text{Hz} \Rightarrow$ pitch sub-score = 0.93; $\overline{\text{RMS}} = 0.062$ $\Rightarrow$ energy sub-score = 0.775; $A = 0.852$. *Verbal*: POSITIVE, filler ratio = 0.023 $\Rightarrow V = 0.9425$.

TABLE V
Worked Composite Score Calculation

Dimension	Raw (0–1)	Weight	Contribution
Emotion (E)	0.74	0.30	0.222
Gaze (G)	0.72	0.20	0.144
Acoustic (A)	0.855	0.25	0.214
Verbal (V)	0.9425	0.25	0.236
Composite	**0.816**	–	**81.6 / Excellent**

D. Pipeline Performance Profiling

Table VI reports measured wall-clock times for each pipeline stage on a one-minute MP4 interview (1080p, 30 fps) processed on an Intel Core i7-12700 CPU without GPU assistance. All times are medians over five runs.

TABLE VI
Pipeline Performance (1-Minute Video, CPU-Only)

Stage	Median (s)	Notes
Facial (ResNet-50 + LSTM)	13.4	Frame-by-frame + LSTM
Audio extraction (moviepy)	2.1	PCM decode
Transcription (Whisper Base)	9.2	CPU inference
Acoustic analysis (librosa)	1.4	Signal processing
NLP (RoBERTa + BART)	0.7	Local model inference
Composite score calculation	0.02	Arithmetic only
Total	**26.8**	3–5× faster with CUDA

GPU acceleration with CUDA reduces total time to approximately 7 to 10 seconds for the same video, owing primarily to parallelised frame-level inference in the facial pipeline.

E. XAI Regression Model Performance

The gradient-boosted regressor was trained on a synthetic but statistically representative dataset of 2,500 student-job pairs. Table VII reports 5-fold cross-validated metrics.

Feature importance analysis reveals that the weighted average score and the critical-skill-gap penalty are the two most predictive features, together accounting for approximately 47% of total importance. This confirms that the rule-based component scores are already highly informative, and the gradient-boosted model adds value primarily by capturing non-linear interactions.

F. Semantic Skill Matching Accuracy

To validate the semantic skill matcher, a test set of 50 skill pairs was assembled, each labelled by a domain expert as

TABLE VII
XAI JOB-MATCH MODEL PERFORMANCE (5-FOLD CV)

Metric	Value
Cross-validated R^2 (mean ± std)	0.944 ± 0.012
Training MAE	2.31
Training RMSE	3.04
Best n_estimators	200
Best learning_rate	0.10
Best max_depth	5

semantically equivalent (n = 25) or distinct (n = 25). Results with a cosine similarity threshold of 0.40 were as follows: Precision 0.92, Recall 0.96, and F1 0.94. Lowering the threshold to 0.30 improved recall to 1.00 but reduced precision to 0.78 due to spurious matches between functionally distinct skills. The 0.40 threshold was therefore retained.

G. *End-to-End Case Studies*

*1) **Case Study 1: High-Performing Candidate**:* A 4-minute video of a candidate for a Data Engineer role showed eye contact for 78% of the video, Happiness as the dominant emotion in 61% of frames, a pitch standard deviation of 32 Hz (expressive), and a filler ratio of 1.9%. Sentiment was POSITIVE (0.94 confidence) and communication style was confident (0.89) and knowledgeable (0.83). The composite score was 84.7 (Excellent).

*2) **Case Study 2: Average Candidate**:* A 3-minute video showed Neutral as the dominant emotion in 72% of frames, eye contact in 49% of the video, a pitch standard deviation of 11 Hz (moderate), and a filler ratio of 8.3% (22 occurrences in 265 words). Sentiment was NEUTRAL. The composite score was 56.1 (Average), with recruiter highlights noting moderate vocal expressiveness and an elevated filler-word ratio.

*3) **Case Study 3: XAI Job Match**:* A student with a 3.4 GPA, 8 matched skills, and 2 relevant projects applied for a Software Engineer position requiring 10 skills. The XAI module produced a match score of 73.2 (Good Match). Missing mandatory skills (Docker and Kubernetes) were flagged as a critical gap, with suggested learning paths and an estimated readiness timeline of 1 to 2 months.

H. *Comparison with Baseline Approaches*

Table VIII positions the MMIA system against existing approaches. Mod. = modalities covered; Trans. = fully transparent scoring; Priv. = full privacy preservation; Open = open implementation.

The MMIA system is the only evaluated approach combining full four-modality coverage, complete scoring transparency, strict privacy preservation, and an open implementation.

TABLE VIII
SYSTEM COMPARISON

System	Mod.	Trans.	Priv.	Open
HireVue (commercial)	3	No	Partial	No
Unimodal NLP only	1	Partial	Yes	Yes
Unimodal facial only	1	Partial	Yes	Yes
MMIA (this work)	4	Yes	Full	Yes

VII. ETHICS, BIAS, AND LIMITATIONS

A. *Bias Mitigation*

The system applies identical scoring weights and thresholds to every candidate, eliminating per-recruiter variability. AffectNet includes faces from 11 countries and diverse age groups, but the emotion classifier may still underperform on underrepresented ethnicities. Periodic bias audits computing average scores disaggregated by self-reported demographic group are recommended before deployment in regulated hiring contexts.

B. *Decision-Support Design*

The system is deliberately positioned as a decision-support tool rather than an autonomous hiring agent. Every output includes a four-dimensional breakdown and recruiter highlights that explain why the score was produced. The Needs Review grade explicitly signals to the recruiter that human inspection of the video is recommended rather than automatic rejection.

C. *Privacy*

All temporary video and audio files are deleted immediately after processing. No candidate data is written to any persistent store. Models are pre-trained on public benchmark datasets; no candidate interview footage was used for training or fine-tuning.

D. *Cultural Sensitivity*

Eye contact norms vary significantly across cultures [11]. The current gaze threshold reflects Western interview conventions. Deployment in culturally diverse contexts should be accompanied by recruiter training that contextualises the eye-contact score accordingly.

E. *Limitations*

- Acoustic scoring thresholds were calibrated on a small reference set and may not generalise across microphone types or noisy recording environments.
- The 10-frame LSTM smoothing window (approximately 333 ms at 30 fps) may be insufficient to suppress artefacts in low-frame-rate recordings.
- The XAI model was trained on synthetic data; validation on real organisational hiring records is required before production deployment.

- The system does not currently process body language or gesture, which the literature identifies as an additional informative channel [36].

VIII.

This paper presented the Multimodal Interview Analysis (MMIA) system, a fully implemented and production-ready platform for automated candidate assessment. By integrating four complementary modalities, namely facial emotion recognition, gaze and head pose estimation, acoustic feature extraction, and NLP-based verbal analysis, into a transparent decision-level fusion framework, the system produces composite scores that correlate strongly with human recruiter judgements (Pearson $r = 0.92$). The processing pipeline completes in 25 to 30 seconds on CPU for a one-minute interview recording, with 3 to 5 times acceleration on GPU.

The Explainable Job Matching module achieves a cross-validated R^2 of 0.944 combined with semantic skill matching at F1 = 0.94 and human-readable explanation generation. Together, these components deliver a platform that addresses the four principal barriers to responsible AI in recruitment: transparency through fully auditable scoring formulas, bias reduction through uniform criteria, scalability through a stateless FastAPI microservice, and privacy through a no-data-retention policy.

A. *Future Work*

1) **Body language**: Extending the visual pipeline to include upper-body gesture recognition using MediaPipe Holistic or OpenPose.
2) **Cultural adaptation**: Parameterising gaze and emotion thresholds per cultural region based on user-configurable profiles.
3) **Longitudinal validation**: Collecting consented recordings and follow-up job-performance data to validate whether composite scores predict 6-month performance ratings.
4) **Federated learning**: Enabling cross-organisation model improvement using privacy-preserving federated averaging.
5) **Real-time streaming**: Adapting the pipeline for live video via WebSocket to enable real-time practice interview feedback.
6) **XAI on real data**: Retraining the job-match model on anonymised historical hiring records from partner organisations.

References

[1] L. Alexander, "Multimodal Analysis For Candidate Assessment," *Unpublished manuscript*, June 2025.

[2] A. Mehrabian, *Silent Messages*. Wadsworth, 1971.

[3] G. Kress and T. Van Leeuwen, *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Arnold, 2001.

[4] S. Oviatt, "Ten myths of multimodal interaction," *Commun. ACM*, vol. 42, no. 11, pp. 74–81, 1999.

[5] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010.

[6] J. Levashina, C. J. Hartwell, F. P. Morgeson, and M. A. Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Pers. Psychol.*, vol. 67, no. 1, pp. 241–293, 2014.

[7] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, 2003.

[8] M. A. Campion, E. D. Pursell, and B. J. Brown, "Structured interviewing: Raising the psychometric properties of the employment interview," *Pers. Psychol.*, vol. 41, no. 1, pp. 25–42, 2014.

[9] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.

[10] C. L. Kleinke, "Gaze and eye contact: A research review," *Psychol. Bull.*, vol. 100, no. 1, pp. 78–100, 1986.

[11] D. Matsumoto, "Culture and nonverbal behavior," in *The Sage Handbook of Nonverbal Communication*, V. Manusov and M. L. Patterson, Eds. Sage, 2006, pp. 219–235.

[12] E. Derous and A. M. Ryan, "When your resume is (not) turning you down: Modelling ethnic bias in resume screening," *Hum. Resour. Manage. J.*, vol. 29, no. 2, pp. 113–130, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019.

[15] S. E. Kahou *et al.*, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM ICMI*, 2015, pp. 467–474.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2019.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[18] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, 2020, pp. 7871–7880.

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

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, 2019, pp. 6558–6569.

[21] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimed. Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[22] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2015.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM KDD*, 2016, pp. 1135–1144.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] M. Belghachi, "A review on explainable artificial intelligence methods, applications, and challenges," *Indonesian J. Electr. Eng. Inform.*, vol. 11, no. 4, pp. 1007–1024, 2023.

[27] V. Hassija *et al.*, "Interpreting black-box models: A review on explainable artificial intelligence," *Cogn. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

[28] I. D. Raji *et al.*, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proc. AAAI/ACM AIES*, 2020, pp. 145–151.

[29] C. O'Neil, *Weapons of Math Destruction*. Crown, 2016.

[30] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May 2016.

[31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[32] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. FAccT*, 2018, pp. 77–91.

[33] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM AIES*, 2018, pp. 335–340.

[34] L. Wang *et al.*, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.09741*, 2022.

[35] E. Cambria *et al.*, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 102–107, 2017.

[36] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proc. IEEE CVPR*, 2016, pp. 4207–4215.

[37] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[38] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, 2014.

[39] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proc. FAccT*, 2018, pp. 149–159.

[40] D. B. Acharya, B. Divya, and K. Kuppan, "Explainable and fair AI: Balancing performance in financial and real estate machine learning models," *IEEE Access*, 2024.

[41] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021.