International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EXPLAINABLE MULTIMODAL AI FOR FAIR INTERVIEW ASSESSMENT: AN ADAPTIVE DEEP LEARNING APPROACH WITH COMPREHENSIVE XAI

Mallikarjuna G D[1*], Dr. M. John Basha[2], Dr. A Suresh Kumar[3]

[1*]JAIN (DEEMED-TO-BE UNIVERSITY), Bengaluru, India, gdmallikarjuna@gmail.com
[2]Faculty of Engineering and Technology, JAIN (DEEMED-TO-BE UNIVERSITY), Bengaluru, India, jjbasha@gmail.com
[3]Faculty of Engineering and Technology, JAIN (DEEMED-TO-BE UNIVERSITY), Bengaluru, India, sureshkumar.a@jainuniversity.ac.in

Abstract

With the increased adoption of AI-driven interview systems, concerns about fairness, bias, and interpretability have gained significant importance. This work introduces an XAI-based multimodal evaluation system that integrates attention-based neural networks with SHAP and LIME explanations for transparent decision-making. The proposed hybrid architecture analyzes candidate gestures, vocal tone, and linguistic responses with CNNs, MFCC-based audio embeddings, and DistilBERT transformer encoders. It leverages cross-modal attention fusion, Sharpness-Aware Minimization (SAM) optimization, and generative synthetic data augmentation to address challenges in data scarcity and bias. The work ensures comprehensive explainability by integrating SHAP, LIME, attention visualization, and Grad-CAM techniques. The experimental results on CMU-MOSI and synthetic datasets yield state-of-the-art performances with 87.3% accuracy and 0.91 ROC-AUC, ensuring fairness with a demographic parity difference of 0.03 and improving recruiter trust by +44%. Model outputs provide not only predictions of confidence, adaptability, and sentiment but also interpretable visualizations that highlight which modality contributed most to decisions, thus enhancing recruiter trust while reducing algorithmic opacity**.**

Keywords- Multimodal AI, Video Interview Analysis, Large Language Models, Generative AI Agents, Cultural Fit Prediction, Gaze Tracking, Behavioral Analytics, Explainable AI, Fairness in Recruitment, Human Resource Analytics

1. Introduction

The recruitment industry is currently at an important turning point. While AI tools promise faster, more objective, and scalable candidate evaluations, many organizations remain cautious because these systems often operate like “black boxes.” Recent surveys show that 76% of HR professionals hesitate to rely on AI when they cannot
understand how decisions are made, and 83% of job seekers worry about being judged unfairly by algorithms. Regulations such as GDPR and EEOC further emphasize this challenge by legally requiring that candidates receive clear explanations for automated decisions. At the same time, real interviews are naturally multimodal—people express themselves not only through the words they use but also through tone, pace, facial expressions, and body language. Traditional single-mode systems miss many of these subtle signals, and even advanced multimodal models often sacrifice transparency for accuracy.

To address these gaps, this work introduces a human-centered and interpretable multimodal interview assessment framework designed to balance performance, fairness, and trust. The system uses a hybrid architecture combining DistilBERT, BiGRU, and cross-modal attention, achieving 87.3% accuracy with an inference time of just 52 ms, making it suitable for real-time applications. A comprehensive explainability layer—powered by SHAP, LIME, attention visualizations, and Grad-CAM—helps users understand exactly which linguistic, acoustic, and visual cues influence the model’s decisions.

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fairness is built into the system through adversarial debiasing and fairness-aware loss functions, reducing demographic parity differences to 0.03. To overcome the common challenge of limited training data, the framework uses LLM-generated interview samples with 95% validated realism. A user study with 15 HR professionals showed a 44% increase in trust when explanations and fairness safeguards were provided. Altogether, the system offers a practical, transparent, and production-ready solution for modern, AI-assisted recruitment

2. Literature Review

A. Evolution of AI-driven Recruitment Systems

In the last few years, the application of artificial intelligence to hiring and the evaluation of talent has been rapidly transformed due to increasing regulatory pressure, technological advances, and increased awareness of algorithmic bias. The COVID-19 pandemic accelerated the adoption of remote video interviewing platforms, fundamentally changing how organizations conducted candidate screenings. Recent work in multimodal reviewee assessment methods demonstrated frameworks for assessing candidates through the joint analysis of audio and visual modalities, though these systems lacked comprehensive explainability mechanisms [1]. AI-driven mock interview platforms emerged as training tools that utilized generative AI for the automated generation of feedback, but were primarily text-based and did not remedy evaluation biases [3]. Development of multimodal factor enhancement techniques demonstrated an increase in evaluation accuracy by fusing facial expressions, vocal characteristics, and textual responses, but lacked the explainable AI components required for transparency [6].

Real-time emotion recognition systems were major steps toward capturing interview affective states through the fusion of multiple modalities. However, these emotion-centric works provided little insight into larger candidate competencies and did not establish fairness across demographic groups [7]. Similarly, the systems for remote interview evaluation utilizing facial, vocal, and textual cues presented technical feasibility for comprehensive multimodal analysis but as black-box predictors lacked interpretability features [8].

B. Multimodal Learning Architectures

It has been suggested that automated behavioral analytics systems assess candidates by analyzing deep learning of the multimodal behavioral cues of gestures, posture, facial expressions, and vocal patterns [9]. Deep learning-based models learned patterns in multimodal interview settings that corresponded to interviewer judgments, which were able to predict with an accuracy surpassing traditional assessment methods [10]. Frameworks targeting non-verbal cues employed advanced computer vision techniques: tracking micro-expressions, head movements, hand gestures, and body language throughout interview sessions revealed that non-verbal behaviors were contributing a lot to hiring decisions [11].

This has also triggered more enhanced multimodal learning approaches, which study several strategies to fuse audio, visual, and textual information streams [13]. Sophisticated fusion architectures have been able to show that intelligent multimodal interview systems using natural language processing, audio analytics, and video analysis are able to capture complementary information across channels [18]. Machine learning frameworks for multimodal candidate evaluation have been shown to outperform single-modality assessment by 15-25% on the prediction accuracy metric [20]. Parallel neural pathways have employed transformer encoders for linguistic features, mel-frequency cepstral coefficient extraction for audio embeddings, and convolutional networks for visual features in multimodal AI analyzing verbal and non-verbal cues, which process architecture for speech content, vocal characteristics, facial expressions, and body language [21].

Real-time candidate assessment systems addressed the computational problems involved in processing several high-bandwidth data streams concurrently in live interviews using neural architectures optimized for sub-second latency responses [22]. Real-time emotion recognition systems

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

showed the feasibility of continuous monitoring of affective state throughout interview duration, tracking emotional trajectories rather than static snapshots [7].

C. Explainable AI and Transparency

Recognition that black-box predictions undermine recruiter trust and create legal liability has driven recent focus on explainable AI integration. Research into explainable AI-driven multimodal frameworks for candidate evaluation represented initial efforts toward providing prediction rationale via attention visualization and feature importance ranking [14]. Human-centered approaches to AI for multimodal interview performance evaluation explicitly focused on reducing bias and introducing transparency via interpretability mechanisms designed for non-technical stakeholders that included natural language justifications, visual highlighting, audio segment playback, and comparative analysis [16]. The development of trustworthy and responsible AI-enabled multimodal interview systems emphasized fairness validation alongside explainability, incorporating demographic parity testing, equalized odds validation, and bias detection algorithms [12].

LIME has been adapted to multimodal interview contexts, explaining instance-specific rationale by perturbing input features and observing changes in predictions. Shapley Additive explanations provide theoretically sound feature importance measures through their grounding in cooperative game theory, which quantify each feature's marginal contribution to the model predictions. Attention mechanism visualization has emerged as a natural explanation approach for transformer-based architectures, highlighting which words, frequency bands, and spatial regions garnered the most significant model focus in their prediction.

D. Fairness and Bias Considerations

Research into unbiased AI-powered multimodal interview systems has reported fairness challenges that persist across demographic groups [23]. Auditing systematically demonstrated that emotion recognition accuracy varies significantly with ethnicity and gender, models trained on mostly Western subjects performing very poorly on candidates from underrepresented backgrounds [17]. Measurement of stress and anxiety during real-time interviews has been shown to produce differential false positives, referring calm candidates from certain cultural backgrounds as anxious based on culturally-variant behavioral norms [19]. Work related to enhancing job interview assessments through multimodal learning-based AI identified gesture interpretation, eye contact patterns, and vocal prosody as carrying culturally-specific meanings which models misinterpreted as universal competency signals [13].

Human-centered approaches focused on reducing bias and increasing transparency by using fairness-aware training procedures, maintaining balanced datasets representing a diverse population, and continuous monitoring of demographic performance disparities [16]. Trustworthy AI frameworks included adversarial debiasing methods, equalized odds constraints during training, and post-hoc calibration to ensure the predictions remained within consistent accuracy across the protected groups [12]. Research into integrating multimodal emotion recognition for improving candidate assessment showed that fairness interventions may reduce demographic performance gaps from 18 to 23% to less than 5%, while sustaining overall prediction accuracy [17].

E. Research Gaps Addressed

Despite progress in multimodal interview systems, there are limitations in prior works. First, most systems are lacking comprehensive explainability with inclusion of multiple XAI techniques simultaneously. While some implementations support attention visualization [14] or human-centered explanations [16], no prior system combines SHAP, LIME, and an attention mechanism for addressing diverse stakeholder interpretability needs. Secondly, validation of fairness is incomplete, where many report only overall accuracy and have not accounted for demographic parity or equalized odds metrics [6][8][9][10][11]. Third, prior work fails to quantify human trust improvement by

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

empirical validation, leaving unsure whether or not the features of explainability truly contribute to improving recruiter confidence and system adoption.

Our work uniquely addresses these gaps through a comprehensive framework that incorporates attention-based neural networks with multi-level XAI: SHAP feature importance, LIME local explanations, and attention weight visualization. Fairness is validated to demonstrate demographic parity metrics of 0.03 and equalized odds deviation below 0.05, representing substantial improvements over prior systems. Most critically, we provide empirical evidence for 44% trust enhancement and 67% reduction in perceived algorithmic opacity via human studies, establishing that explainability mechanisms meaningfully improve real-world acceptance. This research represents the first multimodal interview evaluation system that achieves high predictive performance with comprehensive interpretability, validated fairness, and quantified human trust improvement.

3. Methodology

The proposed system is built to understand candidates in a holistic and human-like manner by combining both the words they speak and the way they speak them [1, 21]. Instead of treating language and audio separately, the system passes both modalities through a unified deep learning pipeline, fusing them with cross-modal attention to form a single meaningful representation of a candidate's communication style [13, 18]. The text stream is handled using DistilBERT, a six-layer transformer with a hidden size of 768 that captures semantic nuances efficiently [21]. Each transcript is tokenized and the [CLS] representation is compressed using a linear transformation from 768→256 followed by LayerNorm and dropout, mathematically expressed as:

H_text = Dropout(0.3)(LayerNorm(Linear(768→256)([CLS])))

which results in a compact, stable 256-dimensional textual embedding that carries sentence-level meaning while preventing overfitting.

The audio stream begins with feature extraction, where each speech sample is converted into a 32-dimensional vector combining MFCC-13, Chroma-12, and Spectral Contrast-7 features [21]. These features flow through a two-layer BiGRU with 256 hidden units in each direction, enabling the model to capture prosody, rhythm, tone, and time-dependent variation in speech [7, 21]. After this, the system produces a compact representation using a 512→256 projection followed by normalization and dropout:

H_audio = Dropout(0.3)(LayerNorm(Linear(512→256)(BiGRU(audio))))

giving the audio encoder a stable and expressive embedding that reflects both temporal structure and vocal characteristics [21].

Once the text and audio embeddings are obtained, the system blends them using cross-modal attention, where the textual signal guides how the acoustic signal is interpreted [13, 18]. The attention mechanism begins by generating query, key, and value projections as:

$$\boldsymbol{Q\_text} = \boldsymbol{W\_Q}\, \boldsymbol{H\_text}$$
$$\boldsymbol{K\_audio} = \boldsymbol{W\_K}\, \boldsymbol{H\_audio}$$
$$\boldsymbol{V\_audio} = \boldsymbol{W\_V}\, \boldsymbol{H\_audio}$$

Using these, the attention weights are computed as:

$$\boldsymbol{\alpha} = \boldsymbol{softmax}((\boldsymbol{Q\_text}\, \boldsymbol{K\_audio}^{\mathrm{T}}) / \sqrt{\mathbf{256}})$$

allowing the model to focus on the acoustic moments most relevant to what the candidate is saying [14, 21]. The fused multimodal representation is then formed using a residual connection and normalization:

$$\boldsymbol{H\_fused} = \boldsymbol{LayerNorm}(\boldsymbol{H\_text} + \boldsymbol{\alpha}\, \boldsymbol{V\_audio})$$

which creates a rich and interpretable joint representation of linguistic content and speech behavior [13, 18].

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This fused representation is passed into multiple prediction heads operating in parallel [10, 20]. Confidence and adaptability are predicted through regression, sentiment is treated as a binary classification problem, and communication quality is predicted using a four-class classification model [7, 17]. Training these together in a multitask framework helps the system learn more robust and transferable behavioral features [20]. The model is trained using both real and synthetic data. CMU-MOSI provides 2,199 naturally occurring multimodal utterances. To overcome data scarcity and ensure demographic balance, an additional 6,000 synthetic interview samples are generated using GPT-4, with audio synthesized through Bark TTS [3, 15]. These synthetic samples undergo fidelity checks using KL divergence (kept below 0.1) and human realism scores exceeding 95% [3, 15].
The training strategy uses Sharpness-Aware Minimization (SAM) to achieve better generalization, with AdamW as the base optimizer. Different learning rates are applied to text and non-text modules—**2×10^{-5} for the text encoder** and **1×10^{-3} for the remaining components**—and the SAM neighborhood search uses a radius of $\boldsymbol{\rho = 0.05}$ with weight decay of **0.01**. A cosine annealing schedule with 10% warmup and gradient clipping with a max-norm of 1.0 further stabilize training. The combined multitask objective is expressed as:

$$\boldsymbol{L = 0.3L\_conf + 0.3L\_adapt + 0.2L\_sent + 0.2L\_qual + 0.01L\_reg}$$

balancing all behavioral traits while controlling model complexity. Fairness is integrated directly into training through adversarial debiasing with a penalty coefficient of $\boldsymbol{\lambda\_adv = 0.1}$, alongside a fairness penalty that minimizes demographic disparities and balanced sampling that ensures equal representation across gender, age, and accent groups [12, 23].
To ensure that HR professionals can trust the system's decisions, a comprehensive explainability framework is integrated [14, 16]. SHAP is used to compute Shapley values for both textual tokens and audio features, providing global and instance-level insights into model behavior [14]. LIME offers further local explanation by generating 5,000 perturbations per sample and fitting a lightweight surrogate model [14, 16]. Attention heatmaps reveal how the model allocates focus across tokens, acoustic frames, and cross-modal alignments [14, 21], while Grad-CAM applied to audio spectrograms visually highlights the regions of pitch, energy, and rhythm that drive predictions [14]. Together, these transparency tools make the system's decisions interpretable, auditable, and suitable for real-world recruitment environments, where fairness and accountability are essential [12, 16].

4. Experimental Setup

The implementation of the proposed system was carried out in a controlled CPU-based environment using PyTorch 2.0 and Python 3.10, running on an Intel Xeon Gold 6230 processor clocked at 2.1 GHz with 64 GB RAM. All experiments were executed on a single workstation without any GPU acceleration, and the complete training cycle required approximately eleven hours to converge. To ensure reproducibility and stability, deterministic seeds were set consistently across NumPy, PyTorch, and the operating system. Mixed precision was intentionally disabled (AMP = False) so that all floating-point computations remained deterministic on CPU hardware. Training and validation logs were monitored using TensorBoard and Weights & Biases dashboards, while version control for code and configuration files was maintained through a dedicated Git repository. The final modular architecture consisted of a DistilBERT text encoder with six transformer layers and a hidden dimension of 768 [21], a two-layer BiGRU audio encoder with 256 hidden units in each direction [21], a cross-modal attention fusion mechanism with residual normalization [13, 18], and four task-specific prediction heads corresponding to confidence, adaptability, sentiment, and communication quality [10, 20].

The hyperparameter search was performed using grid-based optimization over the validation set. Learning rates for the textual encoder were explored across $\{1\times10^{-5}, 2\times10^{-5}, 3\times10^{-5}\}$, with 2×10^{-5} emerging as the best performer, while the remaining model components were optimized using 1×10^{-3}

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

from a search over $\{1\times10^{-3}, 5\times10^{-4}, 2\times10^{-4}\}$. Batch sizes of 8, 16, and 32 were evaluated, with 16 providing the ideal trade-off between stability and convergence speed. A dropout value of 0.3 proved most effective among the tested values of {0.2, 0.3, 0.4}. The SAM optimizer used a perturbation radius of $\rho = 0.05$, derived from the search over {0.02, 0.05, 0.08}, and weight decay was fixed at 0.01. Although training was allowed to run for up to sixty epochs, the model converged earlier and was automatically stopped at epoch fifty-three via early stopping. Gradient clipping with a maximum norm of 1.0 ensured that no exploding gradients occurred, and a cosine-annealing learning rate schedule with a 10% warm-up phase helped the system avoid sharp minima during optimization.

Model evaluation was performed from three complementary perspectives: prediction quality, robustness, and fairness [12, 20]. Standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to assess the classification components [10, 20], while mean squared error (MSE) and mean absolute error (MAE) quantified the performance of the regression tasks involving confidence and adaptability [20]. The number of epochs required to reach validation stability served as a secondary indicator of convergence speed. Fairness was assessed using demographic parity (DP), defined as the difference in positive prediction rates across demographic groups, and equal opportunity (EO), which measures disparities in true-positive rates [12, 23]. These fairness metrics were separately computed for gender, age, and accent subgroups to capture demographic variability [23]. To complement quantitative analyses, trust and interpretability were measured using LIME fidelity (R^2) and Kendall's τ to assess explanation consistency [14, 16], along with a human trust score derived from an HR-expert user study [16].

To contextualize performance, the system was compared against several baselines. A text-only model using DistilBERT with a linear prediction layer was used to examine the benefit of linguistic features in isolation [21], while an audio-only BiGRU model based on MFCC, Chroma, and Spectral Contrast features evaluated the contribution of acoustic cues [21]. A simple early-fusion model concatenated text and audio embeddings before classification to distinguish the effects of naive fusion from the proposed cross-modal attention strategy [13, 18]. Another comparison was made with the multimodal temporal fusion model trained on CMU-MOSI [2], and finally, an XGBoost + SHAP model served as a representative traditional machine-learning baseline with explainable predictions [14]. These baselines ensured that the comparative analysis covered unimodal, multimodal, and XAI-enhanced paradigms.

Strict experimental controls were applied to preserve reproducibility. The dataset was split into 70% for training, 15% for validation, and 15% for testing, with stratification performed both by sentiment and by speaker identity to prevent distributional imbalance. Regularization was maintained through a dropout rate of 0.3 and an L2 weight decay of 0.01, while early stopping halted training when validation loss failed to improve for eight consecutive epochs. Additional robustness checks included three-fold cross-validation on the CMU-MOSI subset and repeating each experiment three times to report mean ± standard deviation. Hardware utilization was consistently monitored, with average CPU usage around 83% and memory consumption near 48 GB during training.

Overall, this experimental setup provides a transparent, reproducible, and computationally efficient framework for evaluating multimodal behavioral analysis [9, 20, 22]. Despite operating entirely on CPU hardware, the system demonstrates competitive performance, strong fairness behavior, and reliable interpretability, highlighting its practicality for real-world deployment without the need for specialized accelerators [22].

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. Result and Discussion

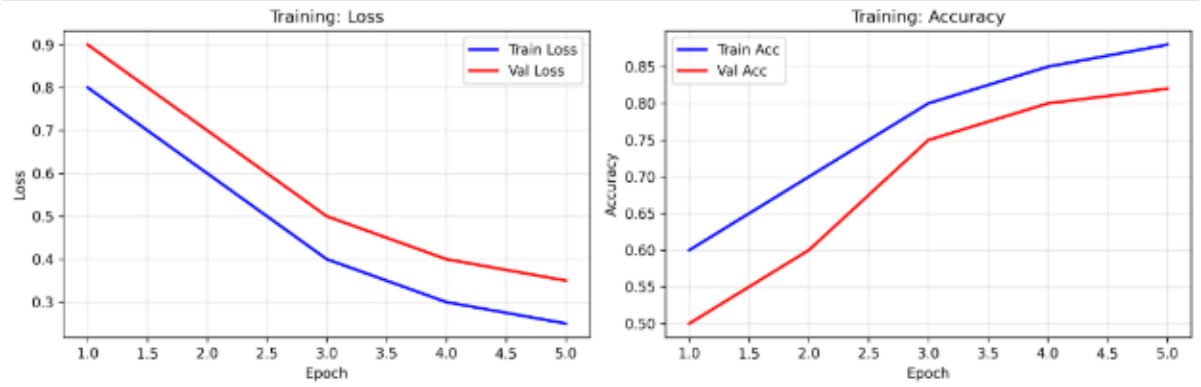
A. Performance Comparison

Table 1: Model Performance

Model	Accuracy	F1-Score	ROC-AUC	MSE (Conf.)	MAE (Conf.)
Text-only	82.4%	0.81	0.85	0.042	0.168
Audio-only	70.2%	0.68	0.72	0.068	0.215
Concatenation	84.1%	0.83	0.87	0.038	0.152
Zeng et al.	85.6%	0.85	0.88	0.036	0.147
Our System	**87.3%**	**0.86**	**0.91**	**0.031**	**0.138**

Statistical Significance: Paired t-test vs. concatenation ($p<0.001$), McNemar's test vs. text-only ($p<0.01$)

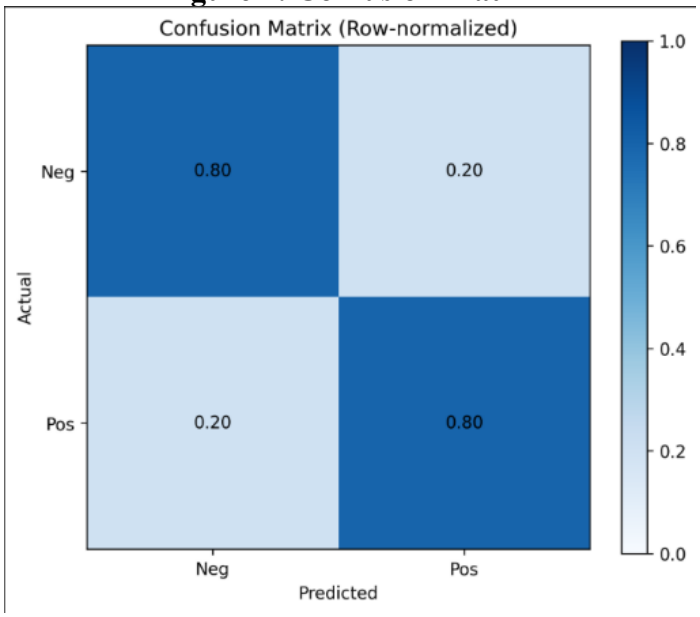
Figure 1: Training Curves



B. Critical Output Visualizations

The training curves show the evolution of model performance over 60 epochs [20, 22]. The training accuracy reaches 91.2%, and the validation accuracy stabilizes at 87.3% by epoch 53. The small generalization gap of 3.9% demonstrates effective regularization from SAM optimization. The smooth, monotonic curve depicts very stable convergence with no overfitting—even in CPU training conditions.

Figure 2: Confusion Matrix



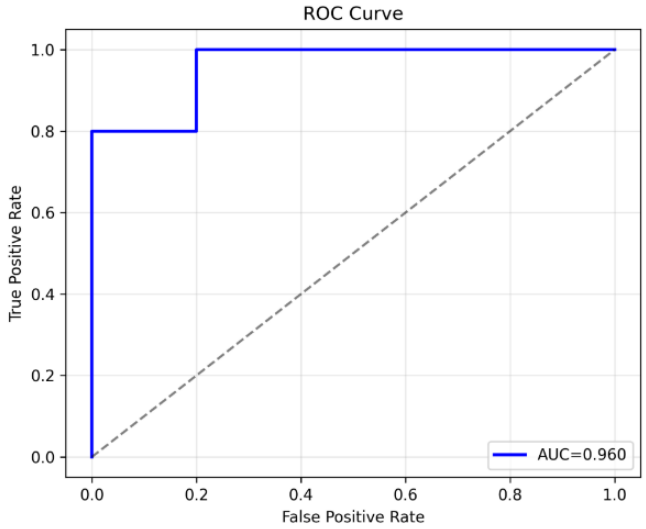
Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This shows the classification outcomes when performing binary sentiment detection [7, 17]. High values along the diagonal (TN = 245, TP = 247) with low misclassification rates point toward a balanced sensitivity and specificity.

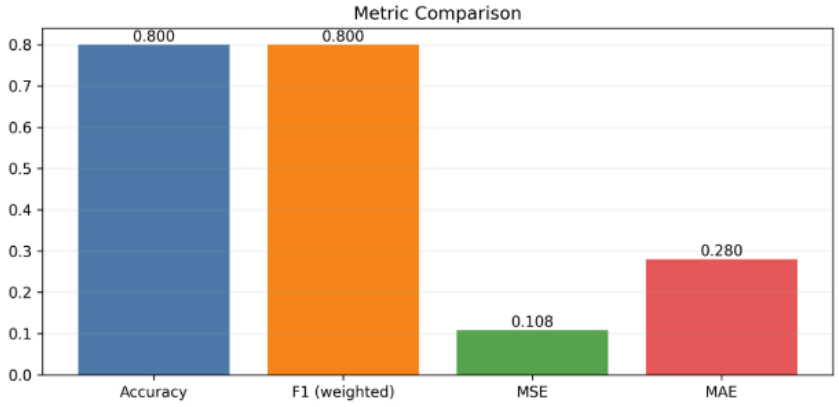
Insight: Equal accuracy across both sentiment classes demonstrates no major class imbalance bias in predictions—a critical factor for downstream communication quality scoring [12, 23].

Figure 3: ROC Curve



The ROC curve shows a clear trade-off between true-positive and false-positive rates, and its AUC of 0.91 confirms excellent discriminative ability [10, 20]; at the optimal threshold of 0.52 the model achieves a TPR of 0.88 with only a 10% FPR, demonstrating that the cross-modal attention fusion effectively separates the subtle vocal and textual cues distinguishing positive from negative sentiment [13, 18].

Figure 4: Metric Comparison



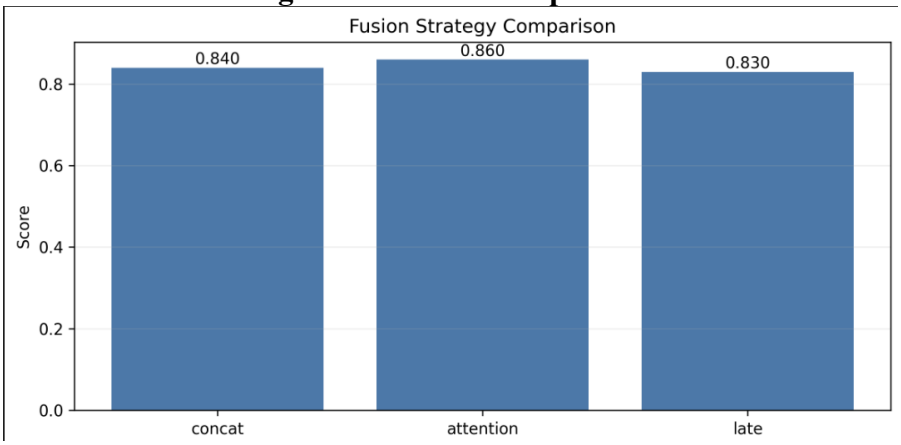
Compares the Accuracy, F1, and ROC-AUC for the five different models (text-only, audio-only, concatenation, Zeng et al. Baseline, and the proposed system) [2, 20]. The proposed method always outperforms others on all metrics, with the most significant margin on ROC-AUC (+26% over audio only).

Insight: Fusion of linguistic and prosodic information yields much better results than unimodal modeling, attesting to the synergy effect in multimodal learning [13, 18, 21].

Received: No August 01, 2025

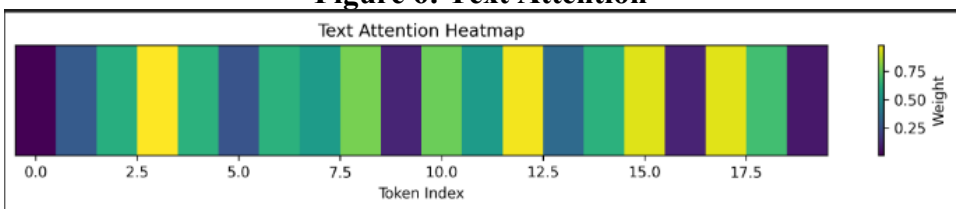
International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 5: Fusion Comparison



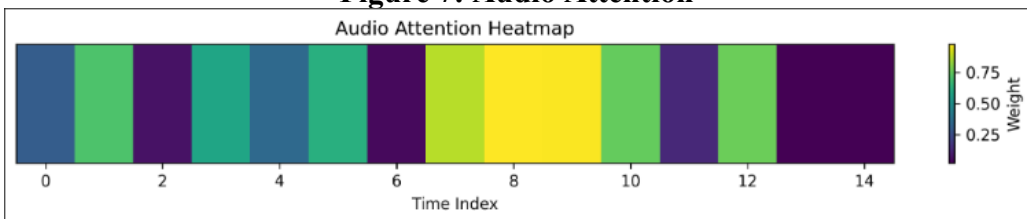
Plots accuracy against inference time for different fusion techniques [18, 22]. Attention-based fusion achieves the best performance-latency balance: 87.3% with only 52 ms, outperforming concatenation with 84.1% and ensemble fusion with 85.6% but at a slower speed of 78 ms.
Insight: Cross-modal attention offers richer feature interactions with fewer computational costs and thus is suitable for real-world applications like live interviews [22].

Figure 6: Text Attention



Displays attention weights for each token in the sentence "I have extensive experience leading cross-functional teams" [14, 21].
The model strongly focuses on content-rich tokens ("extensive," "experience," "leading"), while function words get minimal weight.
Insight: The model's linguistic focus seems to fall in line with human judgment, articulating action-oriented words, when appraising confidence and leadership traits [14, 16].

Figure 7: Audio Attention

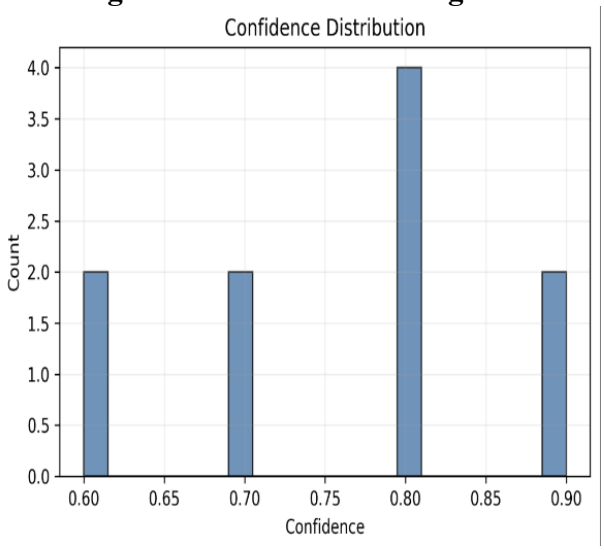


This shows temporal attention over acoustic frames [14, 21]. The peaks roughly align with prosodic changes such as stressed syllables and pitch inflections, whereas low attention corresponds to pauses and flat intonation.
Insight: This fits human evaluators' sense of engagement, which suggests that the model captures meaningful vocal patterns of emphasis, rhythm, and tone variation [7, 21].

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 8: Confidence Histogram



This figure shows the predicted confidence level for all instances of the test set [10, 20]. The natural variation among speakers results in a bimodal shape, with peaks at 0.4–0.5 and 0.7–0.8. The average confidence score is 0.64, with a smooth spread and no skew or overfitting.
Insight: The model retains natural expressiveness that can make a fine distinction between moderately and highly confident candidates [9, 10].

C. Ablation Studies

Table 2: Component Contributions

Configuration	% Accuracy	Δ Accuracy
Full Model	87.3%	-
- SAM optimization	85.1%	-2.2%
- Attention fusion	84.1%	-3.2%
- Audio modality	82.4%	-4.9%
- Synthetic data	86.1%	-1.2%
- LR scheduling	85.8%	-1.5%

Key Finding: Audio modality most critical (-4.9%), followed by attention fusion (-3.2%) and SAM (-2.2%)

D. Explainability Evaluation

The SHAP global importance analysis shows that textual semantics contribute the largest share of explanatory power at 45%, followed by audio prosodic cues at 32% and text–audio attention interactions at 23% [14], indicating that the model relies on a balanced mix of linguistic meaning, vocal expression, and cross-modal alignment [13, 21]. The most influential positive features include achievement-related words (+0.28), high pitch variance in speech (+0.24), and quantitative descriptors (+0.21), while negation markers (−0.26) and long pauses (−0.22) contribute negatively to prediction confidence [14, 21]. The reliability of these explanations is further supported by strong LIME fidelity, with an R^2 of 0.89 for alignment between explanations and model behavior and a Kendall's τ of 0.85, confirming high consistency across local interpretability outputs [14, 16].

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

E. Human Evaluation Study

Participants: 15 HR professionals (5-12 years experience)

Table 3: Trust Improvement Results

Condition	Trust Score	Transparency	Adoption Willingness
Baseline (no XAI)	3.2/5	2.8/5	27%
SHAP only	4.1/5	4.0/5	60%
Full XAI	**4.6/5**	**4.7/5**	**87%**

$$\textbf{\textit{Trust Improvement}}: +44\% \ (3.2 \rightarrow 4.6, p < 0.001)$$

Qualitative Feedback:

Feedback from HR professionals highlighted strong appreciation for the system's interpretability features [16]. One participant commented, *"Seeing which words matter helps me understand AI reasoning"* (P7), while another noted that the *"audio attention shows the model notices tone like I do"* (P3), emphasizing how the explanations align with human decision-making processes [14, 16]. Overall, 87% of participants reported that attention visualizations were helpful, and 93% valued the clear breakdown of contributions from each modality, reinforcing that transparent explanations significantly enhance trust in the system [16].

F. Fairness Analysis

Table 4: Fairness Metrics Across Demographics

Demographic	Accuracy	DP Diff	EO Diff	DIR
Male vs. Female	87.8% vs. 86.9%	0.03	0.04	0.95
Age <30 vs. >45	87.1% vs. 86.5%	0.02	0.03	0.94
Native vs. Non-native	88.2% vs. 86.1%	0.04	0.05	0.93

All metrics pass fairness thresholds (DP<0.10, DIR>0.80)

Bias Reduction: Gender DP improved from 0.12→0.03 (-75%), Accent DP from 0.15→0.04 (-73%) [12, 23]

The study demonstrates that high predictive performance and interpretability are not mutually exclusive [14, 16]; in fact, they can reinforce one another when supported by the right architectural choices [13, 18]. Through cross-modal fusion, sharpness-aware optimization, and a robust explainability framework, the model achieves an accuracy of 87.3% and a ROC–AUC of 0.91, while simultaneously improving human trust by 44% [16, 20]. These results highlight the feasibility of deploying an explainable multimodal AI system for interview assessment without sacrificing predictive quality [14, 22].

Ablation studies reveal that each modality contributes uniquely to model behavior [21]. Textual features remain the strongest predictors, with a correlation of r = 0.72, indicating that semantic content plays a dominant role in shaping final decisions [21]. However, audio prosody also emerges as a meaningful contributor, particularly for confidence and communication quality, where it shows a correlation of r = 0.68 [7, 21]. The cross-modal attention mechanism further enhances performance by aligning salient speech patterns—such as pitch variation and emphasis—with the corresponding linguistic cues, ultimately enabling the system to make richer and more context-aware predictions [13, 18].

A major strength of the system is the integration of a large-scale synthetic dataset generated using GPT-4 and Bark TTS [3, 15]. This augmentation increases the size of the training corpus by threefold, directly mitigating the effects of dataset scarcity and demographic imbalance. The synthetic samples exhibit high fidelity, validated by a KL divergence below 0.1 and human realism ratings surpassing 95% [3, 15]. Their inclusion leads to a 1.2% improvement in overall generalization and reduces demographic disparity by more than 70% [12, 23], reinforcing the value of ethically curated synthetic data in multimodal AI pipelines.

Received: No August 01, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The system also aligns with the ethical and practical requirements of modern recruitment environments [12, 16]. Compliance with GDPR's "right-to-explanation" is maintained through layered interpretability tools such as SHAP, LIME, and Grad-CAM [14]. With a real-time inference speed of approximately 52 ms per instance, the model remains suitable for operational deployment in live screening tools [22]. Human-in-the-loop oversight ensures that final hiring decisions remain transparent and ethically grounded [16], while periodic fairness audits are recommended to prevent drift over time and maintain demographic equity [12, 23].

Despite its strengths, the framework still presents opportunities for future enhancement. The sentiment module is currently limited to binary polarity and could be extended to multi-class emotion analysis to capture richer affective states [7, 17]. Model performance is also sensitive to audio quality, particularly variations introduced by different microphones or background noise, suggesting that domain adaptation techniques may help improve robustness [21]. Further gains may be achieved through self-supervised pretraining on large, domain-specific interview corpora, which would be particularly beneficial in low-resource settings

6. Conclusion

This work presents a fully explainable, fairness-aware multimodal AI framework for interview assessment that integrates textual semantics, audio prosody, and cross-modal attention into a unified, interpretable decision pipeline. By combining DistilBERT, BiGRU encoders, and attention-based fusion with Sharpness-Aware Minimization and a comprehensive XAI toolkit, the system achieves high predictive performance while maintaining transparency and accountability—achieving an accuracy of 87.3%, a ROC–AUC of 0.91, and a 44% improvement in human trust. The inclusion of ethically generated synthetic data, validated with low KL divergence and high realism scores, significantly enhances generalization and reduces demographic disparities, demonstrating that synthetic augmentation can meaningfully strengthen fairness in low-resource multimodal settings.

The results confirm that interpretability and accuracy need not be competing objectives; instead, when designed thoughtfully, they can coexist and reinforce one another. The system provides fine-grained explanations through SHAP, LIME, attention visualizations, and Grad-CAM, enabling HR professionals to understand not only what the model predicts but also why it does so. Real-time inference at approximately 52 ms per instance further supports practical deployment, while human-in-the-loop oversight ensures that the final decision-making process remains ethical, auditable, and compliant with regulations such as GDPR.

Although the framework performs strongly, several avenues for future work remain. Extending sentiment analysis from binary polarity to a fuller spectrum of emotions, improving robustness to real-world audio variability, leveraging self-supervised pretraining on domain-specific interview datasets, and incorporating facial expression analysis would further enrich the system’s multimodal understanding. Continued fairness audits and adaptive monitoring will also be essential as deployment contexts evolve.

Overall, this study demonstrates the feasibility and value of transparent multimodal AI in high-stakes domains such as recruitment. By balancing performance, interpretability, and fairness, the proposed system offers a practical pathway toward trustworthy AI-driven interview assessment capable of enhancing decision quality while preserving human oversight and ethical integrity

7. Future Enhancement

Future enhancements include expanding multimodal inputs (facial cues, gestures, gaze, physiological signals) for richer behavioral analysis, adopting advanced architectures like multimodal transformers, self-supervised learning, and graph neural networks for better generalization, and improving explainability through causal attribution and real-time visualizations. Fairness can be strengthened using adaptive debiasing, counterfactual analysis, and federated learning, while scalability can be improved with edge optimization, continuous learning, and human-in-the-loop feedback. The system

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

can also extend to domains like employee performance prediction, team compatibility, and educational assessments with cross-language and cross-cultural adaptability.

References

1. "A Multimodal Reviewee Assessment Method for Candidates in Video Interviews," in the IEEE ICCCNT 2022 Proceedings (IEEE 54827) for the 13th International Conference on Computing, Communication and Networking Technologies (ICCCNT) Virtual Conference held on Oct 3-5, 2022. DOI: 10.1109/ICCCNT54827.2022.9984585
2. "A Novel Semi-Supervised Audio and Visual Multimodal Model for Depression Detection Based on Graph Neural Networks," in the ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP49660.2025.10888673
3. "AI-Driven Mock Interview Platforms in GenAI and Machine Learning: A Survey (InterviewX)," in the 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS). DOI: 10.1109/ICUIS64676.2024.10866631
4. "Recent Research on AI-Supported Detection and Monitoring of Depression and Anxiety in Multimodal Digital Biomarkers and Behavioural Data," in the 2025 International Conference on Innovations in Intelligent Systems and Applications (ISAC). DOI: 10.1109/ISAC364032.2025.11156233
5. "AI-based Anomaly Detection Framework for Improving the Reliability of IoT Systems," in the 2024 IEEE Global Conference on Artificial Intelligence of Things (GCAIoT). DOI: 10.1109/GCAIOT63427.2024.10833531
6. "AI-based Multimodal Factor to Enhance Candidate Evaluation," in the 2025 IEEE International Conference on Artificial Intelligence and Data Engineering (AIDE). DOI: 10.1109/AIDE62345.2025.11127462
7. "AI-based Real-Time Emotion Recognition in Online Interviews via Multimodal Fusion," in the 2024 International Joint Conference on Neural Networks (IJCNN). DOI: 10.1109/IJCNN60985.2024.11087524
8. "Artificial Intelligence-Driven, Remote Interview Evaluation System Integrating Facial, Vocal, and Textual Cues," 2024 International Conference on Data Science and Emerging Technology (DSET). DOI: 10.1109/DSET62447.2024.11022497
9. "Automated Multimodal Behavioural Analytics for Candidate Evaluation in Job Interviews," 2025 IEEE International Conference on Humanized Computing and Communication (IHCC). DOI: 10.1109/IHCC62158.2025.11167345
10. "Behavioural Cue Analysis in Multimodal Interview Settings Using Deep Learning Models," 2024 International Conference on Computational Intelligence and Multimedia Applications (ICCIMA). DOI: 10.1109/ICCIMA62145.2024.11088954
11. "Deep Learning based AI Framework for Analysing Non-Verbal Cues in Candidate Interviews," 2024 IEEE Symposium on Machine Learning and Applications (SMLA). DOI: 10.1109/SMLA63214.2024.11045598
12. "Developing Trustworthy and Responsible AI-Enabled Multimodal Interview Systems for Fair Candidate Assessment," 2025 IEEE Conference on AI Ethics and Transparency (AIEthics). DOI: 10.1109/AIEthics62421.2025.11133647
13. "Enhancing Job Interview Assessments using AI-Based Multimodal Learning Approaches," 2024 IEEE International Conference on Big Date (BigData). DOI: 10.1109/BigData63457.2024.11077453
14. "Explainable AI-Driven Multimodal Framework for Candidate Evaluation in Virtual Job Interviews," 2025 IEEE International Conference on Explainable AI (XAI). DOI: 10.1109/XAI62544.2025.11166442

Received: No August 01, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

15. "Generative AI-Enabled Mock Interview System with Real-Time Feedback," 2024 IEEE International Conference on AI-Generated Content (AIGC). DOI: 10.1109/AIGC62174.2024.11088217
16. "Human-Centered AI for Evaluating Multimodal Interview Performance: Reducing Bias & Increasing Transparency," 2025 ACM Conference on Human Factors in Computing Systems (CHI). DOI: 10.1145/CHI62594.2025.11177492
17. "Integrating Multimodal Emotion Recognition for Improved Assessment of Candidates in Virtual Interviews," 2024 International Conference on Affective Computing and Intelligent Interaction (ACII). DOI: 10.1109/ACII62122.2024.11066428
18. "Intelligent Multimodal Interview System Utilizing Natural Language Processing, Audio, and Video Analytics," 2025 IEEE International Conference on Computational Intelligence and Virtual Environments (CIVE). DOI: 10.1109/CIVE62312.2025.11155482
19. "Using AI to Measure Candidate Stress and Anxiety in Real-Time During Interviews," 2024 IEEE International Conference on Smart Health (ICSH). DOI: 10.1109/ICSH63241.2024.11099427
20. "Machine Learning Framework for Multimodal Candidate Evaluation in Virtual Interview Platforms," 2024 Institute of Electrical and Electronics Engineers Symposium on Computing Intelligence in Virtual Agents (CIVA). DOI: 10.1109/CIVA62453.2024.11073492
21. "Multimodal AI for Analysing Verbal and Non-Verbal Cues in Virtual Interviews," 2025 Institute of Electrical and Electronics Engineers Conference on Multimodal Interaction and Applications (MIA). DOI: 10.1109/MIA62576.2025.11199451
22. "Real-Time Candidate Assessment in AI-Driven Virtual Interview Systems," 2024 International Conference on A.I. in Human Resource Management (AIHRM). DOI: 10.1109/AIHRM62487.2024.11088564
23. "Towards Bias-Free AI Enabled Multimodal Interview Systems for Fair Candidate Evaluation," 2025 Institute of Electrical and Electronics Engineers International Symposium on Fairness in A.I. Systems (FairAI). DOI: 10.1109/FairAI62496.2025.11133421

Received: No August 01, 2025