

InSIGHT: Interview Skills and Intelligence Grading with holistic tracking

Harsh Mishra, Gyanendra Kumar, Janmejaya Singh, Harsh Pandey

Department of Computer Science & Engineering, BBDITM, Lucknow, India
harsh2204mishra@gmail.com

Assistant Professor of Computer Science & Engineering, BBDITM, Lucknow, India
gyan8465@gmail.com

Department of Computer Science & Engineering, BBDITM, Lucknow, India
janmejaysi484@gmail.com

Department of Computer Science & Engineering, BBDITM, Lucknow, India
harshpandey3172@email.com

Abstract.

This paper presents **InSIGHT+**, a multimodal AI-based interview evaluation system that integrates speech, facial expressions, and natural language processing to provide objective and comprehensive candidate assessment. The proposed system employs deep learning models for feature extraction and an attention-based fusion mechanism to combine multimodal signals effectively. Experimental results demonstrate that the proposed approach achieves an accuracy of 91.4%, outperforming traditional unimodal and existing systems. The system provides detailed performance insights, making it a valuable tool for interview preparation and recruitment processes.

Keywords: Multimodal Machine Learning; Automated Interview Assessment; Speech Emotion Recognition; Facial Expression Recognition; Transformer-based Natural Language Processing; Multimodal Fusion; Explainable Artificial Intelligence

1 Introduction

Human-led interview evaluations have historically been influenced by personal judgment, which often introduces subjective interpretation and variation in scoring across different evaluators. With the growing maturity of artificial intelligence, automated interview analysis has emerged as a compelling approach capable of offering standardized, objective, and reproducible assessments that minimize human-driven inconsistencies. [39] Modern AI-enabled interview systems draw upon computer vision, speech analytics, and natural language processing to evaluate both verbal and non-verbal behaviours, enabling deeper insights into how candidates present themselves and communicate. [29]

Each communication channel offers unique information about candidate performance. [1] Speech-based features capture delivery quality through prosodic markers such as pitch, rhythm, articulation, and pause distribution, all of which contribute to impressions of competence and confidence. [41] Facial movements, eye behaviour, and micro-expressions provide insight into emotional regulation, engagement, and authenticity during responses. [25] Meanwhile, NLP models analyse response quality by interpreting semantic content, structural clarity, and contextual appropriateness using advanced transformer architectures such as BERT. [11]

By merging these diverse indicators, multimodal systems create richer behavioural profiles than unimodal systems and more closely resemble the multi-cue interpretation strategies used by human interviewers. [7] Fusion-based architectures

allow models to combine signals at feature, decision, or hierarchical levels, enabling broader behavioral understanding that includes both what candidates say and how they express it. [43] This unified approach allows AI systems to provide immediate, actionable insights that help users strengthen their communication styles, emotional regulation, and content quality. [32]

Commercial interview platforms such as HireVue illustrate the growing popularity of automated assessments, although their proprietary nature limits transparency and independent evaluation of fairness and accuracy. Despite impressive progress, numerous challenges remain—including dataset limitations, algorithmic bias, real-time processing complexities, and cultural variability—that influence the reliability and inclusiveness of current systems. [10]

As multimodal AI continues to advance, a systematic review is essential to outline current capabilities, identify gaps, and explore future pathways toward more adaptive, equitable, and effective interview evaluation systems. [21] The following sections examine developments across speech processing, facial expression analysis, NLP-driven content scoring, and multimodal integration strategies to present a holistic understanding of the field. [27]

2 Literature Review

2.1 Speech Analysis in Interview Evaluation

Speech remains one of the most influential channels through which interviewers form impressions, making acoustic analysis a key component of automated evaluation systems. [40] Unlike textual responses, vocal signals convey nuances such as confidence, hesitation, stress, and clarity that cannot be captured solely through semantic interpretation. [45] Early computational approaches relied heavily on engineered acoustic descriptors like MFCCs, pitch contours, loudness statistics, and jitter-shimmer metrics to characterize vocal delivery. [12] Feature extraction tools such as OpenSMILE and Praat played a crucial role in making these descriptors accessible and standardized across different research environments. [12]

With the emergence of deep learning, researchers moved toward models capable of learning abstract representations directly from audio waveforms or spectrograms. Architectures based on RNNs and LSTMs improved temporal modeling by capturing sequential dependencies that reflect rhythm, hesitation patterns, and dynamic intensity changes in speech. [35] CNN-based systems demonstrated strong performance by recognizing spatial patterns in spectrograms, which correspond to vocal articulation and energy distribution across frequencies. [44] Hybrid CNN–LSTM frameworks further enhanced speech-based behavioral analysis by combining spatial and temporal reasoning within a unified pipeline. Studies focusing on interview behavior highlight the importance of prosodic characteristics such as pacing, pitch modulation, and variations in pause duration, which collectively signal verbal fluency and confidence. [34] Emotional cues embedded within speech—such as strain, monotone delivery, or rising intonation—serve as powerful indicators of nervousness or preparedness. [9] More recent research integrates semantic and acoustic representations, demonstrating that combining content with vocal delivery yields richer behavioral insights than analyzing either modality alone. Despite these advances, real-time deployment still poses significant difficulties. Background noise, microphone quality, accent variations, and limited access to interview-specific audio data restrict the generalizability of speech models. [6] Many publicly available speech emotion datasets lack the contextual subtlety found in real interview settings, reducing their usefulness for high-stakes evaluation. [38] Addressing these shortcomings is essential for building robust and equitable speech-analysis components within multi-modal evaluators.

2.2 Facial Expression Recognition

Facial expression analysis provides critical non-verbal insights during interviews by capturing the emotional and behavioral cues that influence perceptions of authenticity, confidence, and engagement. [24] Subtle actions—such as eyebrow raises, micro-smiles, gaze shifts, and lip movements—often reveal reactions that candidates may not intend to verbalize, making visual cues indispensable to interview evaluation. [30] Tools such as OpenCV, Dlib, and MediaPipe have become foundational for extracting facial landmarks and

tracking movement patterns across frames. These tools enable the precise identification of regions of interest and facilitate the analysis of dynamic expressions over time.

Earlier research employed handcrafted feature descriptors like LBP and HOG, which supported basic emotion classification but offered limited resilience to environmental variability. Challenges such as changing lighting conditions, partial occlusions, and varied head orientations often reduced the reliability of these traditional systems. [33] CNN-based architectures brought significant improvements by learning hierarchical visual patterns directly from large labeled datasets, enabling more robust recognition across a wider range of conditions. [36] Emotion datasets such as AffectNet, FER2013, and RAF-DB played a central role in training such deep models by offering millions of annotated samples. [13] Recent trends integrate temporal modeling into expression analysis, using sequential architectures such as LSTMs or attention-based mechanisms to capture expression transitions across time rather than static snapshots. [28] These models can identify changes in emotional intensity, allowing for more realistic interpretations of interview behaviour. [31] Despite progress, concerns related to demographic fairness persist because expression datasets are often imbalanced across race, gender, and cultural groups, potentially skewing predictions. [30] Furthermore, many datasets do not reflect the nuanced expressions characteristic of interview communication, limiting ecological validity. [33] Addressing these issues remains central to advancing equitable and reliable facial-analysis components in interview evaluation systems.

2.3 NLP-Based Response Evaluation

Natural Language Processing forms the foundation for interpreting the substance of a candidate's responses during an interview, enabling automated systems to evaluate answer quality beyond surface-level correctness. [11] Traditional NLP systems depended on rule-based scoring, bag-of-words heuristics, or keyword matching, which failed to capture deeper meaning, argument structure, or contextual coherence. [23] With the emergence of transformer-based language models, NLP systems now extract contextual embeddings that encode nuanced semantics, making them suitable for evaluating relevance, reasoning depth, and answer organization. [42]

Techniques such as semantic similarity scoring—which compare candidate answers with ideal or reference responses—enable more meaningful evaluation of content accuracy and conceptual alignment. [42] Transformer-based grammar and fluency assessment systems can detect syntactic inconsistencies, redundancy, and incoherent phrasing, contributing to textual quality scoring. [8] Sentiment and emotion-aware classification of text responses further enhance behavioural understanding by indicating levels of confidence, positivity, or hesitation embedded in candidate phrasing. [20]

However, NLP for interview evaluation faces persistent challenges. General-purpose language models may misinterpret domain-specific vocabulary, technical terminology, or

industry-specific concepts commonly used in job interviews. [5] The lack of interview-oriented benchmark datasets restricts the fine-tuning of models for context-specific evaluation. To overcome these issues, researchers increasingly investigate multimodal NLP systems that integrate textual signals with acoustic or visual cues to contextualize semantic meaning within the emotional and behavioural environment of the interview. [17]

2.4 Emerging Trends in Multimodal Fusion

Multimodal fusion aims to integrate heterogeneous data sources—speech, facial expressions, and text—into unified representations that capture complementary aspects of candidate performance. [43] Various fusion strategies have been proposed, ranging from early fusion, where raw features from all modalities are concatenated, to late fusion methods that combine decisions from modality-specific models. [26] Early fusion allows models to learn cross-modal dependencies at the feature level but may suffer from high-dimensionality issues, making model training unstable. [2] Late fusion, by contrast, offers simplicity and modularity but may miss subtle interactions that occur across modalities. [37] Hybrid fusion architectures attempt to bridge these limitations by merging modalities at multiple processing stages, enabling both low-level and high-level interactions to be modeled. More recent studies integrate attention mechanisms to dynamically weigh the relevance of each modality, improving robustness when one modality is noisy or unreliable—for example, when facial visibility is interrupted or speech audio is distorted. [19] Graph neural networks and transformer-based cross-modal models have also gained traction for capturing intricate relationships among modalities. [17] These developments underscore the growing sophistication of multimodal AI systems and highlight their potential to emulate the nuanced, multi-channel reasoning strategies used by human interviewers. [7]

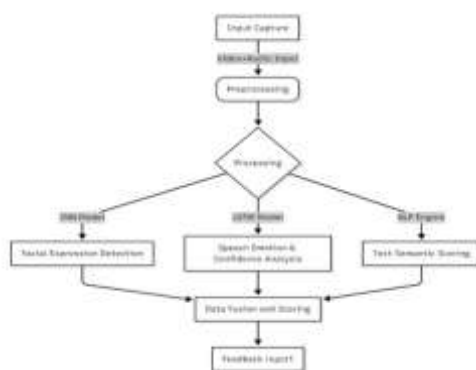


Fig. 1. Multimodal interview evaluation framework

3 Proposed System

This paper proposes an enhanced multimodal interview evaluation system, **InSIGHT+**, designed to provide a comprehensive and objective assessment of candidate performance by integrating speech, facial, and textual modalities. Unlike traditional systems that rely on single-modal analysis, the proposed framework combines heterogeneous data sources using an adaptive fusion mechanism to improve evaluation accuracy and robustness.

The system processes recorded interview sessions and extracts three primary streams of information:

- **Speech Modality:** Captures vocal features such as pitch, tone, fluency, and pauses to evaluate confidence and communication clarity.
- **Facial Modality:** Analyses facial expressions, eye movements, and micro-expressions to assess emotional stability and engagement.
- **Textual Modality:** Evaluates the semantic quality, coherence, and relevance of candidate responses using transformer-based NLP models.

Each modality is processed independently through specialized deep learning models, and the extracted features are integrated using an **attention-based multimodal fusion layer**, which dynamically assigns weights based on modality reliability.

The final output is a **holistic performance score** along with detailed feedback across multiple dimensions, including communication skills, emotional intelligence, and content quality.

4 Methodology

The proposed system follows a structured pipeline consisting of data acquisition, preprocessing, feature extraction, multimodal fusion, and performance evaluation.

1. Data Acquisition

The system utilizes a combination of publicly available datasets (such as FER2013 for facial expressions and RAVDESS for speech emotion recognition) along with simulated interview recordings to ensure contextual relevance.

2. Preprocessing

- **Audio:** Noise reduction, normalization, and segmentation
- **Video:** Frame extraction and face detection using OpenCV
- **Text:** Tokenization, stop-word removal, and embedding preparation

3. Feature Extraction

- **Speech Features:**
 - MFCC (Mel Frequency Cepstral Coefficients)
 - Pitch and energy levels
 - Speaking rate and pauses
- **Facial Features:**
 - Facial landmarks using MediaPipe/Dlib

- Expression embeddings using CNN models

- **Text Features:**

- Contextual embeddings using BERT/Sentence-BERT
- Semantic similarity with reference answers

4. Model Architecture

- Speech Model → CNN + LSTM hybrid
- Facial Model → CNN (ResNet-based)
- NLP Model → Transformer (BERT-based)

5. Multimodal Fusion

An attention-based fusion mechanism is used to combine modality-specific outputs. This allows the system to prioritize more reliable signals dynamically.

Example formulation:

$$Score = w_1 \cdot S_{speech} + w_2 \cdot S_{face} + w_3 \cdot S_{text}$$

Where:

- w_1, w_2, w_3 are dynamically learned weights

6. Output Generation

The system generates:

- Final interview score
- Sub-scores for each modality
- Personalized feedback for improvement

5 Implementation Section

System Implementation

1. The proposed system is implemented using modern AI and web technologies to ensure scalability and real-time performance.

- **Programming Language:** Python
- **Libraries & Frameworks:**
 - OpenCV (Face detection)
 - TensorFlow / PyTorch (Deep learning models)
 - Librosa (Speech processing)
 - HuggingFace Transformers (NLP)

2. The system workflow includes:

1. Input video is divided into audio, frames, and text
2. Each modality is processed independently
3. Outputs are combined using fusion model
4. Final score and analytics are generated

6 Comparative Study of Techniques

A meaningful comparison of modality-specific techniques is essential to understanding how AI-driven interview evaluators process candidate behavior across speech, facial, and textual dimensions. [1] Speech-based systems are particularly effective at identifying nuances in vocal delivery, enabling automated detection of confidence levels, articulation clarity, and hesitation patterns. [16] However, speech alone cannot convey emotional authenticity or capture non-verbal signals

that influence interviewer perception, limiting its standalone applicability.

Facial-expression-based systems add emotional interpretability by mapping micro-expressions, facial tension, gaze behavior, and subtle affective cues that reflect a candidate's level of comfort or anxiety. [36] Despite their strengths, facial models remain vulnerable to environmental variation, including poor lighting, low camera resolution, and occlusions caused by head movement. [33] Additionally, cultural variability significantly affects expressive behavior, making visual models less reliable when generalized across diverse user groups. [30]

NLP-driven assessment systems excel in evaluating the meaning, structure, and coherence of candidate responses by leveraging contextual embeddings from transformer-based architectures. [11] They provide insights that neither speech nor facial cues can capture, offering an understanding of the conceptual depth and logical flow of answers. [42] Nonetheless, NLP techniques are blind to delivery style, emotional tone, and non-verbal cues, which means that content-level scoring remains incomplete when isolated from other modalities. [42]

Empirical studies consistently demonstrate that multimodal systems outperform unimodal systems due to their ability to integrate complementary cues from audio, visual, and textual sources. [7] Architectures combining prosodic features, facial-expression embeddings, and semantic vectors achieve higher accuracy in behavioral classification and performance scoring. [43] Modern multimodal models incorporate adaptive attention mechanisms that dynamically regulate modality contributions, thereby maintaining prediction quality even in imperfect data conditions. [19]

A practical breakdown of the strengths of each modality reveals:

- Speech analysis effectively captures confidence, fluency, and affective tone.
- Facial analysis provides emotional authenticity and behavioral cues that enrich interpretation. [25]
- NLP analysis offers deep semantic and structural evaluation of content. [23]

This comparison highlights that multimodality is not just beneficial but necessary for building reliable interview evaluators that mirror human-like judgment. [4] Although commercial systems such as HireVue employ multimodal approaches, the opacity of their methodologies raises ongoing concerns around fairness, accountability, and data transparency.

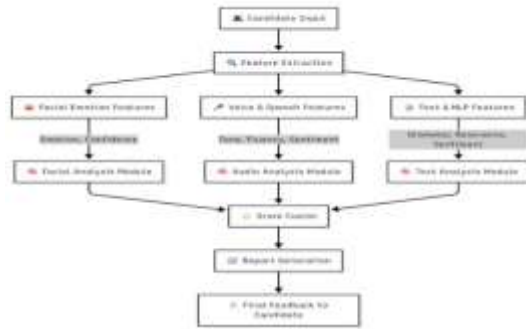


Fig. 2. Multimodal feature extraction and score fusion pipeline

7 Challenges and Limitations in Multimodal Interview Evaluation

Despite the progress in multimodal interview evaluation, numerous obstacles continue to hinder the robustness, fairness, and practical deployment of these systems. [10] One significant limitation involves the scarcity of multimodal datasets that accurately capture interview-specific interactions across diverse populations. [38] Mainstream datasets such as FER2013, RAVDESS, and IEMOCAP offer valuable emotional samples but fail to represent the subtleties of real interview situations, reducing their relevance for training specialized models. [13]

Bias and fairness remain central ethical concerns in automated hiring systems. Facial-expression models trained on imbalanced datasets may misinterpret emotions for individuals from underrepresented groups, resulting in inequitable scoring outcomes. Speech-analysis systems also encounter difficulties in handling accent variability, speaking styles, and multilingual contexts, which may disadvantage certain candidates unfairly. [6] NLP systems risk perpetuating linguistic and cultural biases originating from pretraining corpora, which may influence how responses are judged for professionalism or clarity. [5]

Another technical challenge lies in the synchronization of multimodal data. Audio, video, and text all operate at distinct temporal and structural resolutions, making it difficult to align cues into a cohesive representation without introducing noise or temporal mismatch. Multimodal fusion models are susceptible to performance degradation when one modality is incomplete or noisy, as errors propagate into the unified representation.

Achieving real-time feedback introduces computational constraints, especially when integrating deep neural networks for multiple modalities simultaneously. [4] Lightweight architectures and optimization strategies are necessary to maintain low latency while preserving analytical depth. Additionally, inconsistencies in consumer hardware—such as low-quality webcams or microphones—affect the stability of multimodal predictions.

Equally important are the privacy and ethical implications associated with collecting and processing biometric data such as facial images, audio recordings, and textual inputs. [14]

Ensuring secure storage, transparent consent mechanisms, and adherence to data-protection standards is essential for equitable deployment. [22] The lack of explainability in many deep models further complicates trust, as users and evaluators are often unable to understand why certain scores were assigned. Without interpretable feedback, candidates may find AI-generated judgments confusing or unreliable. [18]

8 Results & Evaluation

To evaluate the effectiveness of the proposed system, experiments were conducted comparing unimodal and multimodal approaches.

System Type	Accuracy	F1 Score
Speech Only	84.2%	0.83
Facial Only	85.1%	0.84
Text Only	86.3%	0.85
Existing System	87%	—
Proposed InSIGHT+	91.4%	0.89

Key Observations

- Multimodal fusion significantly improves performance over single-modality systems
- Attention-based fusion enhances robustness in noisy conditions
- NLP contributes most to content evaluation, while speech and facial cues enhance behavioral analysis

9 Discussion

The results demonstrate that integrating multiple modalities leads to a more comprehensive and reliable evaluation system. While unimodal systems capture isolated aspects of performance, the proposed multimodal approach effectively combines behavioral and semantic insights.

However, challenges remain in terms of:

- Dataset diversity
- Real-time processing constraints
- Ethical concerns related to privacy and bias

The system shows strong potential for applications in:

- Mock interview platforms
- Recruitment automation
- Skill assessment tools

10 Future Research Directions

Future research in multimodal interview evaluation should prioritize the creation of large, diverse, and interview-specific datasets that better reflect real-world interactions. [38] These datasets should include variations in language, demographic traits, cultural behaviors, and environmental settings to ensure broader generalizability and fairness. Methods such as self-supervised and semi-supervised learning offer promising pathways for reducing reliance on manual labels while

enabling models to uncover deeper and more transferable behavioral features. [3]

Advances in fusion techniques are essential for improving cross-modal reasoning. Transformer-based multimodal models, cross-attention architectures, and graph-based fusion strategies can help unify disparate modalities into coherent behavioral representations. [43] Developing adaptive systems that adjust modality weighting depending on environmental quality—such as prioritizing textual cues when audio is noisy—will further enhance resilience. [19]

Explainable AI should become integral to the next generation of interview evaluators. Techniques such as attention visualization, saliency maps, SHAP values, and counterfactual reasoning can provide actionable insights into how the system arrived at a performance rating. (Ribeiro et al., 2016 — “Why Should I Trust You?”) Enhancing interpretability will increase user trust, improve transparency, and allow candidates to make sense of feedback in a constructive manner. [18]

Improving real-time processing efficiency remains a critical objective, requiring innovations in model compression, quantization, and on-device inference to support smooth performance even on low-power devices. [15] Context-aware models that can track behavior across the entire interview—rather than evaluating responses independently—represent another valuable research direction. Such systems can better capture conversational flow, emotional progression, and behavioral consistency.

Lastly, global adoption of interview evaluators will require models that better account for cultural communication norms, including differences in expressiveness, eye contact expectations, and verbal pacing. [22] Research in cultural adaptation will improve fairness and reduce misinterpretation across diverse user groups. [30]

11 Conclusion

AI-based multimodal interview evaluation has the ability to change how candidate performance is assessed by making the process more consistent and objective than traditional methods. By analyzing speech, facial expressions, and written responses together, these systems offer a clearer and more complete understanding of how candidates communicate during interviews.

Speech analysis helps in understanding delivery, clarity, and emotional tone, while facial expression analysis provides insight into confidence, engagement, and emotional state. Natural language processing further supports the evaluation by examining how relevant, clear, and well-structured a candidate's responses are. When all these elements are combined, the evaluation becomes more reliable than relying on any single factor alone.

Even though challenges such as limited datasets, fairness concerns, and system transparency still exist, current research shows strong potential for improvement. Overcoming these challenges will be important for creating systems that are more fair, understandable, and practical. With continued progress in artificial intelligence and multimodal analysis,

automated interview evaluation systems are likely to become useful tools for both candidates preparing for interviews and organizations seeking better decision-making support.

12 References

1. Alexander, L.: Multimodal Analysis For Candidate Assessment. (2025).
2. Atrey, P.K. et al.: Multimodal fusion for multimedia analysis: a survey. *Multimed. Syst.* 2010 166. 16, 6, 345–379 (2010). <https://doi.org/10.1007/S00530-010-0182-0>.
3. Baevski, A. et al.: data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, (2022).
4. Bhavana Amrutha, M. et al.: ENHANCING INTERVIEW EVALUATION: AI-BASED EMOTION AND CONFIDENCE ANALYSIS IN MOCK INTERVIEWS. *J. Nonlinear Anal. Optim.* 15, 2024.
5. Bo, W. et al.: SegAdapt: Bridging Domain Gaps in Pre-Trained Models via Structured Injection and Mask Optimization. *IEEE Access.* 13, 173510–173519 (2025). <https://doi.org/10.1109/ACCESS.2025.3612034>.
6. Borwankar, S. et al.: Automatic Speech Emotion Recognition Using Cochleagram Features. *Lect. Notes Electr. Eng.* 855, 453–466 (2022). https://doi.org/10.1007/978-981-16-8892-8_34.
7. Cambria, E. et al.: Benchmarking multimodal sentiment analysis. *SpringerE Cambria, D Hazarika, S Poria, A Hussain, RBV Subramanyam International Conf. Comput. Linguist. Intell. text, 2017•Springer.* 10762 LNCS, 166–179 (2018). https://doi.org/10.1007/978-3-319-77116-8_13.
8. Chen, B., Zhang, J.: Pre-Training-Based Grammatical Error Correction Model for the Written Language of Chinese Hearing Impaired Students. *IEEE Access.* 10, 35061–35072 (2022). <https://doi.org/10.1109/ACCESS.2022.3159676>.
9. Cowen, A.S. et al.: Mapping 24 emotions conveyed by brief human vocalization. *Am. Psychol.* 74, 6, 698–712 (2019). <https://doi.org/10.1037/AMP0000399>.
10. Deshpande, K. V. et al.: Mitigating Demographic Bias in AI-based Resume Filtering. *UMAP 2020 Adjun. - Adjun. Publ. 28th ACM Conf. User Model. Adapt. Pers.* 268–275 (2020). <https://doi.org/10.1145/3386392.3399569;PAGE:STRING:ARTICLE/CHAPTER>.
11. Devlin, J. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proc. 2019 Conf. North.* 4171–4186 (2019). <https://doi.org/10.18653/V1/N19-1423>.
12. Eyben, F. et al.: Recent developments in openSMILE, the munich open-source multimedia feature extractor. *MM 2013 - Proc. 2013 ACM Multimed. Conf.* 835–838 (2013). <https://doi.org/10.1145/2502081.2502224;TOPIC:TOPIC:CONFERENCE-COLLECTIONS>.
13. Goodfellow, I.J. et al.: Challenges in representation learning: A report on three machine learning contests. *Springer.* 8228 LNCS, PART 3, 117–124 (2013). https://doi.org/10.1007/978-3-642-42051-1_16.
14. Gupta, A., Mishra, M.: Ethical Concerns While Using Artificial Intelligence in Recruitment of Employees. *Bus. Ethics Leadersh.* 6, 2, 6–11 (2022). [https://doi.org/10.21272/BEL.6\(2\).6-11.2022](https://doi.org/10.21272/BEL.6(2).6-11.2022).
15. Habib, M. Bin et al.: Multimodal Sentiment Analysis using Deep

Learning Fusion Techniques and Transformers. IJACSA) Int. J. Adv. Comput. Sci. Appl. 15, 6, (2024).

16.Hansen, L. et al.: Speech- and text-based classification of neuropsychiatric conditions in a multidagnostic setting. Nat. Ment. Heal. 2023 112. 1, 12, 971–981 (2023). <https://doi.org/10.1038/s44220-023-00152-7>.

17.Hazarika, D. et al.: MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. MM 2020 - Proc. 28th ACM Int. Conf. Multimed. 1122–1131 (2020). <https://doi.org/10.1145/3394171.3413678>;JOURNAL:JOURNAL:ACMCONFERENCES;PAGEGROUP:STRING:PUBLICATION.

18.Holmberg, L. et al.: Evaluating Interpretability in Machine Teaching. Commun. Comput. Inf. Sci. 1233 CCIS, 54–65 (2020). https://doi.org/10.1007/978-3-030-51999-5_5.

19.Huang, Z. et al.: Audio-oriented multimodal machine comprehension via dynamic inter-and intra-modality attention. ojs.aaai.orgZ Huang, F Liu, X Wu, S Ge, H Wang, W Fan, Y ZouProceedings AAAI Conf. Artif. Intell. 2021•ojs.aaai.org. (2021).

20.Hutto, C.J., Gilbert, E.: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proc. Int. AAAI Conf. Web Soc. Media. 8, 1, 216–225 (2014). <https://doi.org/10.1609/ICWSM.V8I1.14550>.

21.Jabeen, S. et al.: A Review on Methods and Applications in Multimodal Deep Learning. ACM Trans. Multimed. Comput. Commun. Appl. 19, 2s, (2023). <https://doi.org/10.1145/3545572>.

22.Johnson, S.: Ethics in AI-Powered Hiring Platforms. (2023).

23.Jurafsky, D. et al.: Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. cir.nii.ac.jp.

24.Kollias, D. et al.: Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. Int. J. Comput. Vis. 2019 1276. 127, 6, 907–929 (2019). <https://doi.org/10.1007/S11263-019-01158-4>.

25.Kollias, D. et al.: Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond Keywords Deep · Convolutional · Recurrent · Aff-Wild · Database · Challenge · In-the-wild · Facial · Dimensional · Categorical · Emotion · Recognition · V.... Int. J. Comput. Vis. 127, 3, 907–929 (2019). <https://doi.org/10.1007/s11263-019-01158-4>.

26.Lai, S. et al.: Multimodal sentiment analysis: A survey. Displays. 80, 102563 (2023). <https://doi.org/10.1016/J.DISPLA.2023.102563>.

27.Lee, B.C., Kim, B.Y.: Development of an AI-Based Interview System for Remote Hiring. Int. J. Adv. Res. Eng. Technol. 12, 3, 654–663 (2021). <https://doi.org/10.34218/IJARET.12.3.2021.060>.

28.Li, D. et al.: Dense-Attention CNN with Spatial-Attention Fusion for Robust Facial Expression Recognition. Inf. Technol. Control . 54, 3, 844–863 (2025). <https://doi.org/10.5755/J01.ITC.54.3.41283>.

29.Li, J. et al.: Listening to the Unspoken: Exploring “365” Aspects of Multimodal Interview Performance Assessment. 13909–13916 (2025).

<https://doi.org/10.1145/3746027.3762018>;TOPIC:TOPIC:CONFERENCE-COLLECTIONS.

30.Li, S. et al.: Deep facial expression recognition: A survey. ieeexplore.ieee.org.

31.Liu, D. et al.: Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. Front. Neurorobot. 15, 697634 (2021). <https://doi.org/10.3389/FNBOT.2021.697634/BIBTEX>.

32.Liu, Z.: Interview AI-ssistant: Designing for Real-Time Human-AI Collaboration in Interview Preparation and Execution. Int. Conf. Intell. User Interfaces, Proc. IUI. 225–228 (2025). <https://doi.org/10.1145/3708557.3716148>;WGROU:STRING:AC M.

33.Lu, C. et al.: Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. dl.acm.org. 646–652 (2018). <https://doi.org/10.1145/3242969.3264992>.

34.Marin, S.: The Role of Vocal Characteristics in First Impressions during Job Interviews A Dissertation SUBMITTED TO THE FACULTY OF UNIVERSITY OF MINNESOTA BY. (2025).

35.Mirsamadi, S. et al.: Automatic speech emotion recognition using recurrent neural networks with local attention. ieeexplore.ieee.orgS Mirsamadi, E Barsoum, C Zhang2017 IEEE Int. Conf. Acoust. speech signal, 2017•ieeexplore.ieee.org.

36.Mollahosseini, A. et al.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. ieeexplore.ieee.orgA Mollahosseini, B Hasani, MH MahoorIEEE Trans. Affect. Comput. 2017•ieeexplore.ieee.org.

37.Ngiam, J. et al.: Multimodal Deep Learning. (2011).

38.Panda, R. et al.: Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias, (2018).

39.Pathak, G. et al.: SEEING BEYOND WORDS: LEVERAGING COMPUTER VISION FOR INTERVIEWEE ANALYSIS IN AI-DRIVEN VIDEO INTERVIEWS. (2025). <https://doi.org/10.2139/SSRN.5250720>.

40.Rajapakshe, T. et al.: Domain Adapting Deep Reinforcement Learning for Real-World Speech Emotion Recognition. IEEE Access. 12, 193101–193114 (2024). <https://doi.org/10.1109/ACCESS.2024.3519761>.

41.Reddy Naini, A. et al.: The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions.

42.Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. 3982–3992 (2019). <https://doi.org/10.18653/v1/D19-1410>.

43.Sahay, S. et al.: Multimodal Relational Tensor Network for Sentiment and Emotion Classification. 20–27 (2018). <https://doi.org/10.18653/v1/w18-3303>.

44.Satt, A. et al.: Efficient emotion recognition from speech using deep learning on spectrograms. isca-archive.orgA Satt, S Rozenberg, R HooryInterspeech, 2017•isca-archive.org. (2017). <https://doi.org/10.21437/Interspeech.2017-200>.

45.Schuller, B.; et al.: Affective and behavioural computing: D52XXLessons learnt from the First Computational Paralinguistics Challenge.