International Journal of Science, Strategic Management and Technology
Volume **02** Issue **11** November-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

# A Review on AI-Powered SmartInterview System Using BERT-CNN for Semantic Answer

**Chaitali Namdev Nathe[1], Gaurav Vijay Thore[2], Manoj Pravin Sonavane[3], Suraj Sandip Karade[4], Prof. Walke P.R[5]**

[1,2,3,4] Department of Computer Engineering Vidya Niketan College of Engineering, Bota Pune-Nashik Highway Maharashtra, India.

[5]Assistant Professor, Department of Computer Engineering Vidya Niketan College of Engineering, Bota Pune-Nashik Highway Maharashtra, India.

E-Mail IDs:chaitalinathe22@gmail.com, gauravthore2312@gmail.com, sonavanemanoj86@gmail.com, surajkarade30@gmail.com, pratikwalke27@gmail.com

**https://doi.org/**10.55041/ijsmt.v2i11.003

**Cite this Article**: Nathe, C. N., Thore, G. V., Sonavane, M. P. & Karade, S. S. (2026). A Review on AI-Powered SmartInterview System Using BERT-CNN for Semantic Answer. International Journal of Science, Strategic Management and Technology, *02*(11), 1-9. https://doi.org/10.55041/ijsmt.v2i11.003

**License:** This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

**Abstract— The proposed AI Interview Assistant with Voice, Emotion and Skill Analysis is an intelligent Android application developed to help candidates prepare for job interviews through automated assessment and personalized feedback. Traditional interview preparation methods generally depend on self-practice, coaching, or guidance from experienced individuals, which may not provide continuous performance evaluation. To overcome these limitations, the proposed system combines Artificial Intelligence, Machine Learning, Computer Vision, Speech Processing and Natural Language Processing to analyse different aspects of interview performance. A CNN-based model is used to identify facial expressions, while MFCC-based feature extraction and machine-learning techniques are applied to analyse voice characteristics and speech emotions. Speech-to-text technology converts spoken responses into text, which is further processed using NLP techniques to examine answer relevance, important keywords, communication quality and technical skill alignment. The results obtained from facial, voice and textual analysis are combined to generate an overall performance report. The system also provides personalized recommendations, interview improvement tips, learning resources and skill development suggestions. Firebase is used to maintain interview history and monitor performance over multiple practice sessions. The proposed application aims to provide an accessible and interactive interview preparation platform that supports candidates in improving their communication, technical knowledge and behavioural presentation skills.**

**Keywords: Artificial Intelligence, Machine Learning, Android Application, Interview Assessment, Facial Emotion Recognition, Convolutional Neural Network, Speech Emotion Recognition, MFCC, Natural Language Processing, Speech-to-Text, Multimodal Analysis, Skill Assessment, Recommendation System.**

## I. INTRODUCTION

Job interviews play an important role in determining whether a candidate is suitable for a particular position. During an interview, recruiters generally evaluate technical knowledge, communication skills, confidence, problem-solving ability and professional behaviour. However, many candidates find it difficult to present their knowledge effectively because of nervousness, poor communication, lack of preparation or limited interview experience. Traditional preparation methods often depend on

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

mentors, coaching institutes and individual practice. These methods may not always provide detailed performance analysis or continuous feedback.

Recent developments in Artificial Intelligence, Machine Learning, Computer Vision and Natural Language Processing have introduced new possibilities for automated interview preparation. Facial emotion recognition can examine visible expressions, speech processing can analyse vocal characteristics, and NLP can evaluate the information communicated through spoken answers. Combining these techniques allows an intelligent application to examine different aspects of a candidate's interview performance rather than depending on a single evaluation method.

The proposed AI Interview Assistant integrates these technologies into an Android application. During a mock interview, the system captures facial expressions through the camera and records spoken responses through the microphone. A CNN model processes facial information, while MFCC-based voice feature extraction supports speech and emotion analysis. The recorded responses are converted into text using speech recognition technology and evaluated through NLP techniques.z

The system further analyses the relevance of answers, important keywords, communication patterns and alignment with the selected technical domain. Based on these results, an overall performance report is generated. Personalized recommendations are provided to help candidates identify their improvement areas and practise more effectively. Firebase maintains previous interview records so that users can compare their performance over multiple sessions.

The main purpose of this research is to develop an integrated interview preparation platform that combines visual, audio and textual analysis with personalized improvement recommendations in a single mobile application.

## II. Literature Review

[1] Gang Liu, Shifang Cai and Ce Wang, "Speech Emotion Recognition Based on Emotion Perception," EURASIP Journal on Audio, Speech, and Music Processing, Springer, 2023.

Liu, Cai and Wang proposed a speech emotion recognition method based on the way humans perceive emotional information. Their approach uses emotional attributes and multi-task learning to improve the identification of emotions from speech signals. The proposed model was evaluated using the IEMOCAP dataset and achieved improvements of 2.44% in unweighted accuracy and 3.18% in weighted accuracy compared with the reference approach used in their experiments. This research demonstrates that speech emotion recognition can be improved by considering additional emotional information rather than depending only on conventional speech features. It is useful for the proposed interview assistant because the voice analysis module needs to identify changes in speech patterns during candidate responses. However, this study mainly concentrates on speech emotion recognition. The proposed application extends this functionality by combining voice analysis with facial expressions, answer evaluation and personalized interview recommendations.

[2] Tong Liu and Xiaochen Yuan, "Paralinguistic and Spectral Feature Extraction for Speech Emotion Classification Using Machine Learning Techniques," EURASIP Journal on Audio, Speech, and Music Processing, Springer, 2023.

Liu and Yuan developed a speech emotion classification approach using spectral and paralinguistic features. Mel-Frequency Cepstral Coefficients (MFCC) were extracted to represent the spectral characteristics of speech, while openSMILE was used to obtain additional paralinguistic information. The extracted features were evaluated using machine-learning classifiers such as Multi-Layer Perceptron and Support Vector Machine. Experiments were performed using the Berlin emotional speech database under clean and noisy recording conditions. This research is directly relevant to the proposed system because MFCC-based feature extraction is planned for analysing the candidate's voice. Similar processing can be used to examine speech characteristics and emotion-related patterns during mock interviews. The proposed application further combines this voice information with facial expression recognition and NLP-based answer evaluation.

[3] Fabian Stoehr, Benedikt Kämpgen, Lukas Müller, Laura Oleaga Zufiría, Vanesa Junquero, Cristina Merino, Peter Mildenberger and Roman Kloeckner, "Natural Language Processing for Automatic Evaluation of Free-Text Answers – A Feasibility Study Based on the European Diploma in

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

Radiology Examination," Insights into Imaging, Springer, 2023.

Stoehr and the research team investigated the use of Natural Language Processing for evaluating descriptive examination answers. The study focused on reducing the manual effort required to check free-text responses. The developed system used predefined word lists, synonyms and decision-tree learning to compare candidate responses with expected answer information. The evaluation showed F1 scores of 0.26, 0.33 and 0.50 for three selected questions, with better results for the more structured question format. This research indicates that NLP can support automated response assessment, although answer structure and wording differences remain important challenges. It is relevant to the proposed interview assistant because spoken responses can be converted into text and examined for important technical keywords and expected concepts. The proposed system extends this approach by combining textual answer evaluation with facial and voice analysis to produce a broader interview performance report.

[4] Emiliano del Gobbo, Alfonso Guarino, Barbara Cafarelli and Luca Grilli, "GradeAid: A Framework for Automatic Short Answers Grading in Educational Contexts – Design, Implementation and Evaluation," Knowledge and Information Systems, Springer, 2023.

Del Gobbo and the research team introduced GradeAid, an automatic short-answer grading framework designed to evaluate students' descriptive responses. The system analyses both lexical and semantic characteristics of answers and uses regression-based models to estimate answer scores. An important contribution of this research is its support for non-English datasets, along with evaluation using publicly available datasets and an additional research dataset. The approach demonstrates how automated systems can examine the meaning of a response instead of depending entirely on exact keyword matching. This is useful for the proposed interview assistant because candidates may explain the same technical concept using different words. Semantic evaluation can help improve the quality of answer assessment. The proposed system extends this concept by evaluating spoken interview responses and combining the textual assessment with voice and facial information.

[5] Yanling Gan, Luhui Xu, Haiying Xia and colleagues, "Harmonious Mutual Learning for Facial Emotion Recognition," Neural Processing Letters, Springer, 2024.

Gan and the research team proposed a facial emotion recognition framework based on mutual learning between two neural networks. The approach uses attention mechanisms to identify useful facial features and reduce the influence of irrelevant information. It also considers uncertainty between similar facial expression categories through probability-based learning. The study addresses common challenges in facial emotion recognition, such as background variations, partial face obstruction and changes in illumination. This research is useful for the proposed Android application because facial images captured during mock interviews may contain variations in lighting, camera position and facial movement. A suitable CNN-based model can use these research findings to improve visual analysis. However, the proposed application does not limit its assessment to facial expressions and also examines speech and textual answers.

[6] Ruhina Karani, Jay Jani and Sharmishta Desai, "FER-BHARAT: A Lightweight Deep Learning Network for Efficient Unimodal Facial Emotion Recognition in Indian Context," Discover Artificial Intelligence, Springer, 2024.

Karani, Jani and Desai developed FER-BHARAT, a lightweight CNN-based approach for facial emotion recognition designed around the Indian population. The research considered the limitations of existing emotion datasets, which may not represent cultural differences in facial expressions. The proposed models were trained and evaluated using the Indian Spontaneous Expression Dataset and the Indian Semi Acted Facial Expression Database. The study reported accuracy improvements over its baseline models, including improvements of 11.14% for ISED and 4.72% for iSAFE using its proposed CNN approach with manual feature extraction. This research is particularly relevant when developing facial analysis features for Indian users. Its lightweight approach may also provide useful guidance for mobile application development. The proposed system further extends facial analysis by integrating it with voice and technical answer evaluation.

[7] Jong-Yih Kuo, Ti-Feng Hsieh and Ta-Yu Lin, "Constructing Multi-modal Emotion Recognition Model Based on Convolutional Neural Network," Multimedia Tools and Applications, Springer, 2025.
Kuo, Hsieh and Lin introduced a multimodal emotion recognition model that combines audio and visual information using convolutional neural networks. Their research used image information and Mel-spectrogram representations to extract facial

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

and speech-related features. Different fusion techniques were examined to understand how audio and visual information could be combined for emotion classification. The researchers also investigated the influence of image preprocessing on overall model performance using the eNTERFACE05 dataset. This research provides a useful technical foundation for the proposed interview assistant because it demonstrates how audio and visual information can contribute to combined emotion recognition. However, the proposed system adds another important dimension by evaluating the actual content of interview answers through speech-to-text and NLP techniques.

[8] Chhaya Gupta, Nasib Singh Gill, Preeti Gulia, Abhinav Kumar, Hanen Karamti, Demmelash Mollalign Moges and colleagues, "A Multimodal Fusion Model for Real-Time Environment Emotion Recognition Using Audio-Visual-Textual Features," Journal of Big Data, Springer, 2025.

Gupta and the research team proposed a multimodal emotion recognition framework named M-fusHER. The model combines information from audio, video and text to identify human emotional patterns. CNN and multiplicative LSTM techniques are used for feature extraction, while a fine-tuned YOLOv6 model supports facial feature and expression detection from video. The research evaluates the system using established multimodal emotion recognition datasets. This study is closely related to the proposed project because both approaches use multiple information sources rather than depending on a single input type. However, the proposed interview assistant focuses on candidate assessment rather than general environmental emotion recognition. It also includes technical answer evaluation, skill identification, interview scoring and personalized recommendations.

[9] Yan Li, Yapeng Wang, Xu Yang and colleagues, "Speech Emotion Recognition Based on Graph-LSTM Neural Network," EURASIP Journal on Audio, Speech, and Music Processing, Springer, 2023.

Li and the research team introduced a speech emotion recognition method based on Graph Neural Networks and Long Short-Term Memory networks. Their approach represents speech features as a graph so that relationships between different speech segments can be examined. An LSTM-based aggregation mechanism and weighted pooling are used to learn important emotional information from speech sequences. The proposed method was evaluated using the IEMOCAP dataset and achieved 65.39% unweighted accuracy and 71.83% weighted accuracy. The study provides useful information about analysing the sequential nature of spoken responses. During an interview, a candidate's speech may contain variations in tone and delivery across different parts of an answer. The proposed application can use speech feature analysis to provide communication-related feedback, while also combining the results with facial and textual assessment.

[10] Johannes M. Basch and Klaus G. Melchers, "Here's Looking at You: Does Eye Contact in Video Interviews Affect How Applicants are Perceived and Evaluated?" Journal of Business and Psychology, Springer, 2025.

Basch and Melchers investigated the role of eye contact in video-based job interviews. Their research examined how differences in eye-contact direction may influence the way applicants are perceived and evaluated. The authors conducted two experimental studies involving asynchronous video interviews. Their findings indicated that horizontal deviations in eye contact negatively affected perceptions of applicant suitability, while vertical deviations did not produce the same effect on interview performance ratings. This research highlights the importance of visual presentation during online interviews. It is relevant to the proposed application because facial information and visual behaviour are considered during mock interview sessions. The proposed system can provide feedback about visible presentation patterns alongside speech and answer quality. However, eye contact and facial expressions should be treated as practice indicators rather than definitive measures of a candidate's personality or suitability.

## Research Gap Identified

The existing research studies have introduced different techniques for facial emotion recognition, speech emotion analysis, multimodal learning and NLP-based answer evaluation. CNN-based models have improved facial expression classification, while MFCC and machine-learning techniques have supported the identification of emotion-related patterns in speech. Similarly, NLP approaches have made it possible to evaluate written answers based on keywords, semantic meaning and answer relevance. However, most of these studies focus on individual assessment tasks or specific emotion recognition applications rather than providing a complete interview preparation solution. There is a need for an integrated system that combines facial expression analysis, voice characteristics, speech-to-text

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

conversion and technical answer evaluation within a single platform. Existing approaches also provide limited support for personalized interview improvement, skill-based recommendations and continuous performance tracking. To address these limitations, the proposed AI Interview Assistant integrates CNN-based facial emotion recognition, MFCC-based voice analysis and NLP-based answer assessment in an Android application. It combines the outputs of these modules to generate interview performance scores, identify strengths and weaknesses, and provide personalized learning resources and improvement suggestions. Firebase is used to maintain interview history and monitor candidate progress over time. Thus, the proposed system aims to provide a unified and accessible platform for continuous interview practice and performance improvement.

## III. ARCHITECTURE

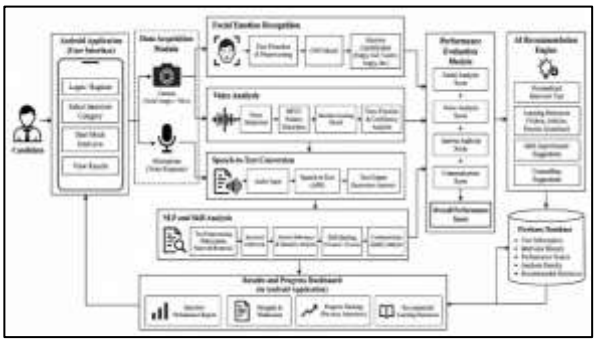


Fig 1: Architecture Diagram

The system architecture of the proposed AI Interview Assistant with Voice, Emotion and Skill Analysis illustrates the complete working process of the Android-based application. It consists of different modules that work together to analyse a candidate's interview performance using Artificial Intelligence, Machine Learning, Computer Vision and Natural Language Processing.

Initially, the candidate accesses the Android application, where registration, login, interview category selection and mock interview options are provided. During the interview session, the data acquisition module collects facial images through the mobile camera and voice responses through the microphone.

The captured facial images are forwarded to the Facial Emotion Recognition module, where face detection and preprocessing are performed before passing the images to a CNN model. The model identifies visible facial expression categories such as happy, sad, neutral and angry. Simultaneously, the recorded voice is processed through noise reduction and MFCC feature extraction. Machine-learning techniques analyse voice characteristics and emotion-related patterns to provide voice assessment results.

The recorded speech is also passed to the Speech-to-Text Conversion module, which converts spoken responses into textual information using Automatic Speech Recognition (ASR). The generated text is processed by the NLP and Skill Analysis module, where text preprocessing, keyword extraction, answer relevance, semantic analysis, technical skill matching and communication quality are evaluated.

The outputs from facial analysis, voice analysis and answer evaluation are transferred to the Performance Evaluation module. These results are combined to generate individual assessment scores and an overall interview performance score. Based on the evaluation, the AI Recommendation Engine provides personalized interview tips, learning resources, skill improvement suggestions and counselling recommendations.

## IV. COMPARATIVE ANALYSIS

| Study | Main approach | Finding relevant to this project |
|---|---|---|
| Liu (2023) [1] | Uses emotional perception and multi-task learning for speech emotion recognition. | Additional emotional features can improve the analysis of voice-based emotional patterns. |
| Liu and Yuan (2023) [2] | Extracts MFCC and paralinguistic features using openSMILE, MLP and SVM classifiers. | MFCC and voice characteristics can support speech emotion analysis during interview responses. |
| Stoehr (2023) [3] | Uses NLP, predefined keywords, synonyms and decision trees for evaluating free-text answers. | Answer evaluation should consider relevant concepts and alternative wordings instead of exact |

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

| | | |
|---|---|---|
| | | keyword matching alone. |
| del Gobbo (2023) [4] | Uses lexical and semantic features with regression models for automatic short-answer grading. | Semantic analysis can help evaluate technical answers that express similar meanings using different words. |
| Gan (2024) [5] | Uses two interacting neural networks, attention mechanisms and probability-based learning for facial emotion recognition. | Facial analysis should consider important expression features and variations in visual information. |
| Karani (2024) [6] | Develops lightweight CNN models for facial emotion recognition using Indian facial expression datasets. | Lightweight facial models can provide useful guidance for mobile-based emotion analysis, especially for Indian users. |
| Kuo (2024/2025) [7] | Combines CNN-based facial and audio features using multimodal fusion techniques. | Audio and visual information can complement each other during emotion recognition. |
| Gupta (2025) [8] | Combines audio, video and textual features using CNN, multiplicative LSTM and YOLOv6. | Multiple input sources can provide broader emotion information than a single modality. |
| Li (2023) [9] | Uses Graph Neural Networks, LSTM aggregation and weighted pooling for speech emotion classification. | Sequential speech information can help examine variations throughout a candidate's spoken response. |
| Basch and Melchers (2024/2025) [10] | Studies the effect of eye-contact direction on applicant perception during video interviews. | Visual presentation is an important aspect of interview practice and can be included in candidate feedback. |

## V. CHALLENGES / LIMITATION

1. Facial Recognition Accuracy: Facial expressions can vary depending on lighting, camera quality, face position and individual differences. These factors may affect the accuracy of CNN-based emotion recognition.

2. Voice Recording Quality: Background noise, pronunciation differences, microphone quality and speaking speed can influence the extraction of MFCC features and voice classification results.

3. Speech-to-Text Accuracy: Technical terminology, regional accents and unclear pronunciation may result in incorrect transcription, which can further affect answer evaluation.

4. NLP-Based Answer Evaluation: Candidates may explain the same technical concept using different words. A keyword-based approach may not always recognize correct answers. Semantic analysis is therefore important.

5. Multimodal Synchronization: Facial frames, voice recordings and converted text must be aligned properly. Incorrect synchronization can affect combined assessment results.

6. Mobile Processing Limitations: Running multiple AI models simultaneously may increase processing time, memory usage and battery consumption on Android devices.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

7. Data Privacy: Facial images, voice recordings and interview results contain personal information. Secure authentication, appropriate encryption and controlled data access are necessary.

8. Dataset Limitations: Models trained using limited datasets may not perform equally well for different languages, accents, age groups and cultural backgrounds.

9. Recommendation Accuracy: The quality of personalized suggestions depends on the correctness of the assessment results and the relevance of the available learning materials.

10. Interpretation of Emotional Information: Facial expressions and vocal characteristics cannot provide a scientifically certain measurement of confidence or psychological condition. Therefore, results should be presented as approximate practice feedback rather than definitive judgments about candidates.

## VI. Future Scope

In the future, the proposed system can be enhanced through the integration of advanced deep-learning models and Large Language Models (LLMs) to improve interview question generation and answer evaluation. The application can generate questions according to a candidate's resume, technical domain, experience level and selected job role. More advanced NLP techniques can be introduced to understand the meaning of descriptive answers, identify missing technical concepts and generate detailed explanations. Multilingual speech recognition can also be integrated to support candidates who prefer practising interviews in different languages. Improved facial and voice analysis models can help provide more detailed feedback about speaking patterns and visual presentation.

The application can also be extended with company-specific interview preparation, job-role recommendations, resume analysis and adaptive mock interview sessions. TensorFlow Lite optimization can help deploy lightweight AI models directly on Android devices, reducing dependency on continuous cloud processing. Integration with Firebase can further support interview history, progress comparison and personalized learning plans. Additional features such as mentor feedback, interview preparation videos, coding assessments and technical quizzes can make the platform more comprehensive. In the long term, the proposed system can be developed into a complete AI-powered career preparation platform that supports candidates in improving their interview performance and professional skills.

## VII. Conclusion

The literature review examines recent developments in facial emotion recognition, speech emotion analysis, multimodal learning and NLP-based answer evaluation. The selected research studies demonstrate that CNN, MFCC, machine-learning classifiers, deep-learning architectures and semantic analysis can support different aspects of intelligent assessment. However, many existing studies concentrate on individual recognition tasks or specific evaluation applications.

The proposed AI Interview Assistant with Voice, Emotion and Skill Analysis aims to integrate these techniques into a single Android-based interview preparation platform. By combining facial expression analysis, voice processing and NLP-based answer evaluation, the system can provide a broader understanding of candidate performance. Personalized recommendations and Firebase-based progress tracking further support continuous interview practice. The proposed approach aims to make interview preparation more accessible, interactive and systematic while helping candidates identify their strengths and areas requiring improvement

## .References

[1] G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, Art. no. 22, 2023.

[2] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, Art. no. 23, 2023.

[3] F. Stoehr, B. Kämpgen, L. Müller, L. Oleaga Zufiría, V. Junquero, C. Merino, P. Mildenberger, and R. Kloeckner, "Natural language processing for automatic evaluation of free-text answers – a feasibility study based on the European Diploma in

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

Radiology examination," Insights into Imaging, vol. 14, Art. no. 150, 2023.

[4] E. del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation," Knowledge and Information Systems, vol. 65, pp. 4295–4334, 2023.

[5] Y. Gan, L. Xu, H. Xia, and colleagues, "Harmonious mutual learning for facial emotion recognition," Neural Processing Letters, vol. 56, Art. no. 96, 2024.

[6] R. Karani, J. Jani, and S. Desai, "FER-BHARAT: A lightweight deep learning network for efficient unimodal facial emotion recognition in Indian context," Discover Artificial Intelligence, vol. 4, Art. no. 35, 2024.

[7] J.-Y. Kuo, T.-F. Hsieh, and T.-Y. Lin, "Constructing multi-modal emotion recognition model based on convolutional neural network," Multimedia Tools and Applications, vol. 84, pp. 31093–31118, 2025.

[8] C. Gupta, N. S. Gill, P. Gulia, A. Kumar, H. Karamti, D. M. Moges, and colleagues, "A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features," Journal of Big Data, vol. 12, Art. no. 256, 2025.

[9] Y. Li, Y. Wang, X. Yang, and colleagues, "Speech emotion recognition based on Graph-LSTM neural network," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, Art. no. 40, 2023.

[10] J. M. Basch and K. G. Melchers, "Here's looking at you: Does eye contact in video interviews affect how applicants are perceived and evaluated?" Journal of Business and Psychology, vol. 40, pp. 669–685, 2025.

[11] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy, vol. 25, no. 10, Art. no. 1440, 2023.

[12] U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and colleagues, "Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects," Multimedia Systems, vol. 30, Art. no. 115, 2024.

[13] J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang, and S. Wang, "Multimodal emotion recognition based on facial expressions, speech, and EEG," IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 396–403, 2024.

[14] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," Artificial Intelligence Review, vol. 58, Art. no. 49, 2025.

[15] G. Kortemeyer, "Performance of the pre-trained large language model GPT-4 on automated short answer grading," Discover Artificial Intelligence, vol. 4, Art. no. 47, 2024.

B. Foundational research papers

[16] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLOS ONE, vol. 13, no. 5, Art. no. e0196391, 2018.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, pp. 335–359, 2008.

[18] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

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

[24] I. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, and colleagues, "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.