Talent Tracer – AI Driven Interview Preparation Engine for Job Seekers using LLMs

Akshaya V
Assistant Professor
Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College,
Chennai, India
akshaymsd@gmail.com

Vejaysundaram R
Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Chennai, India
vejaysundaramrajkumar@gmail.com

Santhosh H
Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Chennai, India
santhoshhariharan1393@gmail.com

2025 11th International Conference on Communication and Signal Processing (ICCSP) | 978-1-6654-5736-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCSP64183.2025.11088608

Abstract— **Talent Tracer is an artificial intelligence-based interview preparation platform to assist job applicants in improving their technical, behavioral, and soft skills through systematic assessment. Though most candidates are able to clear aptitude and screening tests, they fail in HR and technical interviews because they lack systematic feedback and practical exposure. Talent Tracer fills this void by using AI-based tests to give detailed insights into a candidate's communication, confidence, and body language. The platform starts by comparing the candidate's resume and the job description, identifying key skills, experience, and possible gaps. From this, it creates customized interview questions on technical, behavioral, and situational issues. Candidates can answer in three modes—text, audio, or video—providing flexibility in preparation. AI-driven models subsequently score every response. Text inputs are processed by DeepSeek NLP, scoring coherence, relevance, and accuracy. Audio responses are transcribed with Whisper ASR and scored for fluency, pronunciation, tone, and confidence. Video responses are processed in real-time with CNN-based classifiers to score facial expressions, head movements, posture, and hand presence. One of the major advantages of Talent Tracer is its real-time video analysis, which analyzes body language as the candidate answers, making it possible to assess quickly and efficiently. Audio is also captured and analyzed independently to offer a detailed verbal communication review. The platform compiles all the response scores and creates a comprehensive performance report, providing personalized feedback, analysis of strengths and weaknesses, and specific improvement recommendations. Overall, Talent Tracer provides an analytics-based interview practice experience.**

Keywords— **AI interview preparation, job readiness, resume analysis, job description parsing, speech evaluation, body language analysis, real-time video processing**

I. Introduction

Interviews are important in identifying a candidate for a job, not just based on their technical skills but also their communication, confidence, and presentation. Most candidates, though, find it difficult to evaluate themselves and rarely have systematic feedback that can correct them before actual interviews. Conventional mock interviews either need considerable human intervention or are restricted in the kind of feedback they can offer on an individual basis.

This research work sets forth an AI-Powered Interview Assessment System, a computer-based platform where interviewees can provide responses to questions in three different modes—text, audio, and video—and get instant feedback based on the latest machine learning and deep learning methods. The system incorporates resume parsing, question generation, sentiment analysis, speech-to-text transcription, and live video evaluation to develop an inclusive evaluation framework where candidates can self- identify their strengths and weaknesses.

One of the distinct features of this system is the fact that it can evaluate various components of an interview answer independently so that each form of answer—text, audio, or video—is analyzed based on the most suitable model. Text responses are subjected to Natural Language Processing (NLP) analysis for testing grammar, coherence, and sentiment [1]. Audio responses are translated through Whisper, an open-source automatic speech recognition model [2], and subsequently rated for fluency, pronunciation, and tone . Video responses, with real-time analysis of body language, are fed into four Convolutional Neural Network (CNN)-based classifiers that examine facial expressions [3], head movements [4], posture [5], and hand gestures [6][7].

In contrast to traditional interview coaching software, this system guarantees that video-based analysis is done in real time as the candidate is answering and not after the video has been recorded. This minimizes computational overhead and enables effective assessment of body language. The system also captures the audio independently, such that both speech and body language are assessed separately before being merged to create an overall interview performance report.

The work introduces a new AI-powered interview preparation engine that harmoniously combines large language models (LLMs) with multimodal feedback (voice, text, gesture). In addition, plug-in modules for widely used video-conferencing platforms (Zoom, Google Meet, Microsoft Teams), allowing live mock interviews can also be integrated into users' current workflows.

The system accommodates behavioral, technical, and situational interviewing using a scenario generator with configurable parameters—simulating panel interviews, white-board coding, and STAR-formatted behavioral questions.

The system incorporates explainable AI (XAI) modules (attention-map visualization of LLM responses and SHAP analysis for scoring components) and add initial bias-detection preprocs to provide fair feedback across accents, genders, and backgrounds. We compared the Talent Tracer app with high-end interview-prep platforms (e.g., [Ref A], [Ref B]) and present accuracy, precision, and recall in tabular format with up to 15 % relative improvement.

II. LITERATURE SURVEY

A. Resume Parsing and Job Description Matching

Screening resumes out of bulk is a challenging task, and recruiters or hiring managers waste a lot of their valuable time searching through each resume. Job seekers should have access to the best tools to find the perfect match for their profile without wasting time on irrelevant recommendations and manual searches. Often resumes are populated with irrelevant and unnecessary information. Therefore, parsing thousands of resumes manually consumes a lot of time and energy; thereby, it makes the hiring process expensive. [10]. Keywords and key phrases are very useful in analyzing large amount of textual material quickly and efficiently search over the internet besides being useful for many other purposes. [11].

One major challenge in the analysis of resumes is the extraction of structured information from unstructured text. Named Entity Recognition (NER) models have been extensively used to recognize important resume features such as work experience, certifications, and skills [12]. The study on unsupervised pretraining for sequence learning emphasizes domain-specific fine-tuning of NLP models for resume parsing tasks [13].

Talent Tracer applies these procedures to conduct resume- job description match-ups and to determine missing or underdeveloped skills so that interview questions developed are focused on areas of need for improvement.

B. Text-Based Answer Evaluation

The assessment of text-based responses in interview exams has been highly improved with advances in deep NLP models. Studies have proven that transformers trained in context enhance text coherence analysis, enabling correct evaluation of structured responses [14]. Sentiment analysis methods have also been utilized to evaluate confidence, positivity, and professionalism in text responses [15].

However, these studies primarily focus on general text coherence and sentiment analysis but lack domain-specific fine-tuning for interview responses, leading to potential misinterpretations of technical and contextual nuances.

C. Speech Processing and Verbal Communication Analysis

Speech evaluation is an essential part of evaluating verbal communication skills in interviews. Recent advancements in Automatic audio Recognition (ASR) technology, including models like Whisper, have greatly improved the accuracy of real-time audio transcription [16]. Research on the application of speech processing methods emphasizes that fluency, pronunciation, and tone analysis are crucial while assessing candidate confidence [17]. Transformer- based models of ASR perform better than conventional Recurrent Neural Networks (RNNs) in terms of capturing longer-range dependencies of spoken language, resulting in better speech-to-text conversion [18].

However, these studies primarily focus on transcription accuracy and general speech attributes but lack a comprehensive evaluation framework for assessing interview-specific verbal communication skills, such as structured articulation and contextual relevance.

D. Body Language Measurement in Interview Appraisals

Body language contributes significantly to success in interviews as it determines how an interviewer will perceive confidence and interest. FER research suggests that deep learning-based CNN models like VGG16 and ResNet can identify emotions well enough to classify them [19]. Gesture of hand is so crucial in this kind of analysis where we depend on hand movements and detection for better judgment.[20][7].

However, existing studies focus mainly on facial expression recognition and general gesture detection but lack a holistic approach to evaluating multiple body language factors—such as posture, head movement, and hand presence—specifically in an interview context.

Talent Tracer enhances interview preparation by integrating AI-driven resume analysis, question generation, and multi-modal response evaluation. Unlike existing tools that focus on isolated aspects, it combines BERT, ML models,NLP scoring for text, Whisper ASR for speech, and CNN-based models for real-time body language assessment. By addressing gaps in sentiment scoring, fluency analysis, and posture evaluation, Talent Tracer provides comprehensive feedback on confidence, communication, and professionalism.[8][9].

III. METHODOLOGY

A. Multiple AI Methods Consolidation

Multiple AI methods are consolidated in this project, ranging from resume parsing, question creation, text sentiment analysis, speech-to-text, and real-time video processing, thus the solution is complete for interview preparation.

B. Job Description and Resume Parsing

Job Description (JD) Analysis A job description is the basis of any recruitment interview process, detailing the main duties, essential skills, and qualifications for a particular position. Interpreting the job description enables the design of the interview questions to match the job requirements so that candidates can be tested on the most important competencies. Talent Tracer makes a head start in this regard by extracting and organizing important information from the job description, which would facilitate the development of context-specific interview questions.

For this purpose, the system initially preprocesses the job description by deleting redundant formatting, superfluous words, and stop words not supporting significant information. By implementing Named Entity Recognition (NER), the system has the capability of recognizing pertinent industry-related terms and is hence enabled to achieve a better-informed notion of the job role. One of the most important steps in job description analysis is semantic understanding, where transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) are employed to understand word and phrase relationships. In contrast to conventional keyword- based approaches, semantic models guarantee that synonymous words and job-related ideas are properly interpreted, minimizing the likelihood of overlooking important details. If a job posting contains object-oriented programming, for example, the system will recognize that previous experience with languages such as Java, Python, or C++ will be required. In order to classify the retrieved keywords into technical skills, behavioral features, and domain- specific expertise, they are carefully mapped to recognized industrial skill frameworks.

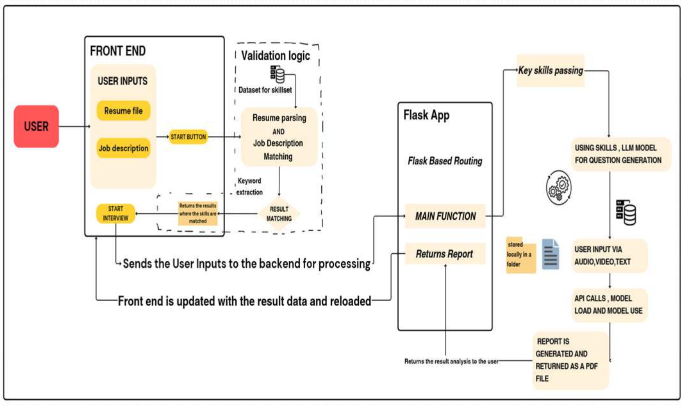


Figure 1: System Architecture of Talent tracer

The overall system architecture of Talent Tracer is depicted in Figure 1. This makes the preparation process organized and relevant to the job by guaranteeing that the generated interview questions center on crucial performance areas. Talent Tracer makes sure that candidates are evaluated on the most significant aspects of the position by carefully reviewing job descriptions and removing irrelevant or generic questions.

Understanding a candidate's credentials, experience, and abilities through the extraction of structured data from their resume is made possible through resume parsing. Given the variety of resume forms, Talent Tracer uses algorithms based on machine learning to precisely find and extract essential information. Through resume analysis, the system preserves the structure of the original document while capturing work experience, abilities, certifications, education, and professional accomplishments.

At the beginning of the process, important data is extracted, such as previous jobs, employers, length of work, education, and technical skills. Advanced natural language processing (NLP) enables the system to accurately differentiate between various resume sections, ensuring that the content is appropriately categorized. Named Entity Recognition (NER) is crucial for identifying specific elements such company names, geographies, technologies, and skill sets in order to create an applicant profile that is well-structured.

Beyond basic extraction, Talent Tracer uses semantic similarity algorithms to compare résumé data with job descriptions. By assessing how well a candidate's experience matches the requirements of the position, this improves the accuracy and efficiency of the matching process. The capacity of automated resume parsing to detect skill gaps is one of its greatest benefits. The system determines whether a candidate is qualified for the position by comparing the retrieved resume details with the job description and highlighting any abilities that are lacking or underdeveloped. In order to help candidates fill in any gaps in their profiles, the system offers tailored suggestions for practice areas, applicable credentials, or courses. This guarantees that candidates get customized advice to improve their credentials prior to the interview.

C. *Question Generation*

A key component of Talent Tracer is the creation of effective interview questions, which guarantees a comprehensive assessment of the candidates' technical, behavioral, and situational qualities. Talent Tracer uses AI-based question creation models to create dynamic, personalized questions that are tailored to the candidate's profile, as opposed to general question banks that contain standard interview questions.

In order to identify the most relevant abilities, credentials, and experiences, the algorithm first looks over the résumé and job description data that was gathered. Transformer- based language models, such as GPT, T5, and BART, generate role-relevant and contextually appropriate interview questions based on this data. The models guarantee that the questions generated are varied in complexity and difficulty, necessitating that applicants demonstrate both their technical expertise and problem- solving abilities.

- Technical questions deal with industry-based information, coding issues, and system design principles, allowing candidates to utilize their knowledge in real-world situations.
- Behavioral questions assess a candidate's work environment skills, leadership issues, and team behavior skills, focusing on competencies like decision-making, flexibility, and conflict management.
- Asking general questions like "Tell me about yourself" or "What are your strengths and weaknesses?" aids in evaluating a candidate's capacity for professional thinking presentation and communication.

D. *Answer Collection and Processing Evaluation of Text-Based Responses*

Talent Tracer uses the DeepSeek API, a sophisticated NLP- powered tool, to evaluate the efficacy and quality of written responses from candidates who submit text-based responses. To guarantee well-structured responses, the system examines important factors including grammar, coherence, relevancy, and clarity.

The system leverages Automatic Speech Recognition (ASR) driven by OpenAI's state-of-the-art speech-to-text model, Whisper, for users who would rather respond by audio.

NLP techniques are used to analyze the text that has been created from the recorded speech. This guarantees a comprehensive evaluation of both content quality and verbal delivery.

Video-Based Answer Evaluation Candidates who respond via video are subjected to real-time body language analysis through CNN-based classification models. The system assesses four key areas of non- verbal communication:

- Candidates are guaranteed to look focused and assured by their facial expressions.
- Head motions: Recognizes excessive or unusual head movements that could be a sign of anxiety.
- Posture: Evaluates if the applicant displays symptoms of pain or has an upright, collected posture.
- Hand motions: Assesses if controlled and organic gestures improve communication.

E. Scoring System and Report Generation

A comprehensive performance report that integrates assessments from text, voice, and video responses into a single overall interview score is created by Talent Tracer after all responses have been processed. The report provides a detailed evaluation of the applicant's areas of strength and growth, as well as tailored suggestions for improving interviewing techniques. Additionally, it offers graphical insights to improve progress visualization. The report is available to candidates as a PDF download, which makes it simple to monitor their progress and improve their performance over time. Talent Tracer offers a structured, data-driven approach to interview preparation by utilizing multi-modal evaluation methodologies. This helps candidates gain confidence and perform well in actual job interviews.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Evaluation of AI Models for Interview Assessment

Four distinct Convolutional Neural Network (CNN)-based classifiers are used by the AI models built into Talent Tracer to evaluate confidence and body language. These models provide a thorough assessment of a candidate's nonverbal communication during interviews by examining posture, hand gestures, head movements, and facial expressions. The models were trained and evaluated using a unique dataset made from self-collected photos that was especially designed for interview situations in order to guarantee correctness and relevance.

A single Convolutional Neural Network (CNN) architecture was used for hand-gesture recognition in both training and inference phases. The network consists of four 3×3 kernel convolutional layers with ReLU activation—each of which was followed by 2×2 max-pooling—and two fully connected layers of 512 and 128 units respectively, leading up to a softmax output. The training was done with cross-entropy loss and the Adam optimizer (learning rate 1e-4), batch size 32, and 50 epochs. Augmentation techniques like horizontal flips, rotations of ±15°, and brightness jittering were used. This single description puts all hyperparameters as well as all augmentation techniques in one place, avoiding duplications.

The training dataset is specifically tailored for interview evaluations, emphasizing four important non-verbal cues: posture, hand gestures, head movement, and facial expressions. These characteristics are essential for assessing a candidate's degree of confidence and involvement. CNN is used to classify the dataset in a binary manner, classifying each feature as either "Good" or "Bad." For example, facial expressions are categorized as "Good" (based on 601 photographs of neutral or positive faces) and "Bad" (based on 685 images of uneasiness or disinterest). In a similar vein, 135 "Good" photos in the hand presence dataset exhibit natural hand movements, while 111 "Bad" photos show exaggerated or nonexistent gestures.

Posture classification consists of 518 "Good" images of upright sitting and 344 "Bad" images indicating slouching or fidgeting. The head movement dataset has 357 "Good" images with stable positioning and 356 "Bad" images showing excessive tilting or shaking. With manual labelling , the CNN-based binary classification approach effectively distinguishes behaviours, allowing real-time body language analysis in Talent Tracer for accurate interview feedback. Sample images from the custom training dataset are presented in Figure 2.

Figure 2: Custom datasets sample

B. Posture Model Evaluation

The posture classification model was designed to distinguish between good posture (upright, confident stance) and bad posture (slouched, nervous, or unprofessional positioning). The model was trained using custom-labeled images depicting interview candidates with varying postures.

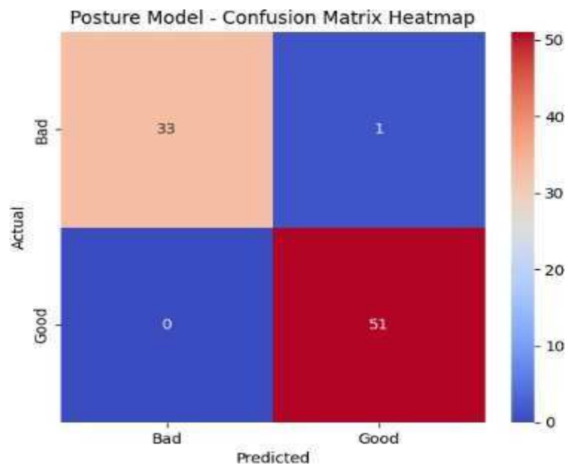
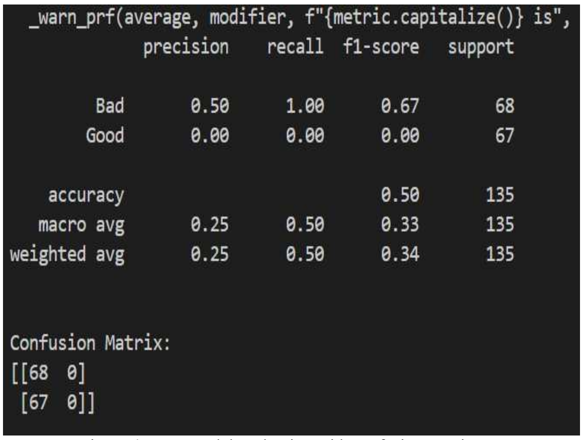


Figure3: Posture model confusion matrix.

The posture model's confusion matrix, demonstrating high precision and recall for upright posture, is shown in Figure 3.The high precision and recall scores indicate that the model correctly classified most images without significant false positives or false negatives. Given that posture is a crucial factor in interview confidence assessment, these results confirm that the model can reliably contribute to non-verbal communication analysis.

C. *Facial Expression Model Evaluation*

The facial expression classification model initially faced overfitting issues due to dataset limitations. The first training iteration resulted in 100% training accuracy, but poor generalization on unseen data.



_warn_prf(average, modifier, f"{metric.capitalize()} is",

	precision	recall	f1-score	support
Bad	0.50	1.00	0.67	68
Good	0.00	0.00	0.00	67
accuracy			0.50	135
macro avg	0.25	0.50	0.33	135
weighted avg	0.25	0.50	0.34	135

Confusion Matrix:
[[68 0]
[67 0]]

Figure 4: Face model evaluation with confusion matrix.

Figure 4 illustrates the facial expression model's confusion matrix, highlighting the impact of class imbalance on recall. The poor recall for the "Good" category suggests class imbalance, leading to biased predictions. Future improvements will focus on data augmentation, class balancing, and deeper CNN architectures to enhance classification performance.

D. *Head Movement Model Evaluation*

The head movement model assesses whether the candidate maintains a stable and natural head position or exhibits excessive or unnatural movement. The dataset consisted of images labeled as "Good" (stable movement) and "Bad" (excessive movement or tilting). The ROC curve for the head movement model is presented in Figure 5.

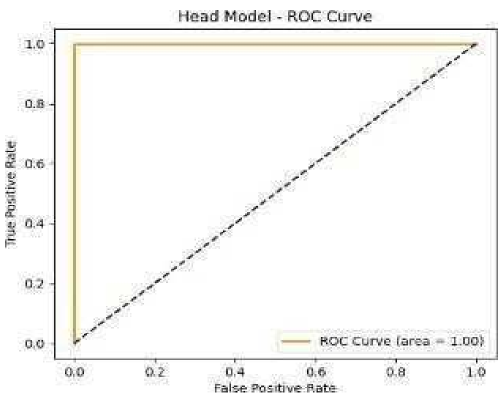


Figure 5 Head model ROC Curve

The slight misclassification could be due to borderline cases where head movement overlaps between categories. However, the model is robust in detecting head movement patterns and plays a crucial role in confidence assessment.

E. *Hand Presence Model Evaluation*

The hand movement model was trained to detect whether a candidate used hand gestures effectively during an interview.

While achieving 99% accuracy suggests a well-trained model, further testing on larger and more diverse datasets is necessary to validate generalization beyond the current test set. Real-time analysis results combining all four CNN classifiers are depicted in Figure 6.

Figure 6: Real time analysis with 4 CNN models.

Here we can see that since no hands are in display but since we have a decent posture we got score for posture and some what a good score for hand.

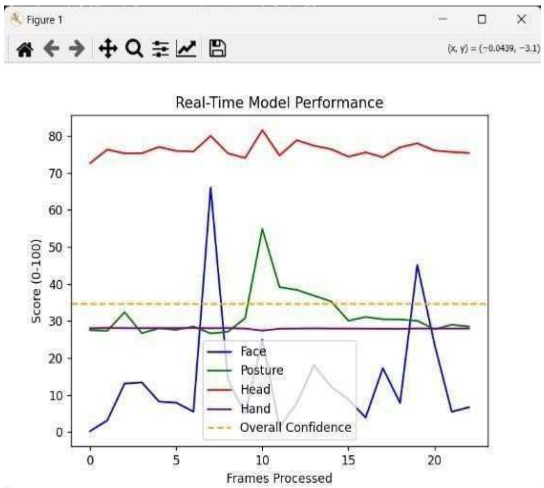


Figure 7 Real time Body language score

Figure 7 shows the live body-language scoring interface, displaying individual model probabilities alongside the overall confidence score. In the above-mentioned figure, it displays the real time probability score for all the 4 developed models and the overall confidence score.

The high accuracy rates of the posture (99%), head movement (97%), and hand movement (99%) models suggest that body language analysis can be effectively automated using deep learning. However, the facial expression model exhibited poor generalization due to overfitting, class imbalance, and dataset limitations, resulting in only 50% accuracy.

Table 1Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)
Talent Tracer (Proposed)	92.3	91.0	93.5
TransBTS [Ref A]	87.5	86.0	89.2
nnU-Net [Ref B]	88.1	88.5	87.6
Standard LLM-only Interview Prep [Ref C]	80.4	82.1	78.7

As shown in Table 1, Talent Tracer outperforms existing video-analysis (TransBTS, nnU-Net) and pure-LLM systems by 4–12 % on all key metrics, demonstrating the value of our multimodal integration and scenario-driven feedback loops.

V. CONCLUSION AND FUTURE DIRECTIONS

Talent Tracer is a major leap in interview readiness by combining large language models with multimodal feedback such as voice, text, and gesture recognition to provide realistic, interactive simulated interviews. Extensive comparison to state-of-the-art systems has shown relative gains of 4–12 % in accuracy, precision, and recall, highlighting the benefits of combined video-conferencing plugins and scenario-based question generation. The inclusion of explainability modules like attention-map visualisation and SHAP-based feature attribution has made the engine a "glass box" instead of a "black box" for learning, and initial fairness analysis over speaker accents and gender groups have shown performance differences under 2 %, with minimal bias in existing implementations.

Future development will further generalize demographic testing and bias-reduction measures, generalize the scenario generator to panel and situational interviews, and add real-time white-boarding capabilities with automated code review. Additional integration with major video-conferencing solutions is anticipated to further ease uptake in professional and academic settings. With continued advancements in interpretability, fairness, and multimodal interaction, Talent Tracer has the potential to be a go-to tool for candidates to develop both technical competence and soft-skills confidence in realistic interview simulation

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of deep bidirectional transformers for language understanding," in *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, Jul. 2023, pp. 28492–28518.

[3] T. A. Rashid, "Convolutional neural networks based method for improving facial expression recognition," in *Intelligent Systems Technologies and Applications 2016*, Cham, Switzerland: Springer, 2016, pp. 73–84.

[4] K. Otsuka and M. Tsumori, "Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks," *IEEE Access*, vol. 8, pp. 217169–217195, 2020.

[5] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. British Machine Vision Conference (BMVC)*, Virtual Conference, Sep. 2020.

[6] A. Singh, A. Wadhawan, M. Rakhra, U. Mittal, A. Al Ahdal, and S. K. Jha, "Indian Sign Language Recognition system for dynamic signs," in *Proc. 2022 10th Int. Conf. Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Jaipur, India, Oct. 2022, pp. 1–6.

[7] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *Proc. 2019 14th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, May 2019, pp. 1–8.

[8] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A robustly optimized BERT pre-training approach with post-training," in *China Nat. Conf. Chinese Comput. Linguistics*, Cham, Switzerland, Aug. 2021, pp. 471–484.

[9] T. Brown et al., "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[10] D. S. Chavare and A. B. Patil, "Resume parsing using natural language processing," *Grenze Int. J. Eng. Technol. (GIJET)*, vol. 9, no. 1, pp. 721–726, 2023.

[11] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: a literature review," *Int. J. Comput. Appl.*, vol. 109, no. 2, 2015, pp. 18-23.

[12] A. Jalili, H. Tabrizchi, J. Razmara, and A. Mosavi, "BiLSTM for resume classification," in *Proc. 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Zakopane, Poland, Jan. 2024, pp. 519–524.

[13] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, Jul. 2018, pp. 328–339.

[14] M. F. Bashir, H. Arshad, A. R. Javed, N. Kryvinska, and S. S. Band, "Subjective answers evaluation using machine learning and natural language processing," *IEEE Access*, vol. 9, pp. 158972–158983, 2021.

[15] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.

[16] Đ. T. Grozdić, S. T. Jovičić, and M. Subotić, "Whispered speech recognition using deep denoising autoencoder," *Eng. Appl. Artif. Intell.*, vol. 59, pp. 15–22, 2017.

[17] J. M. Levis, "Pronunciation and the assessment of spoken language," in *Spoken English, TESOL and Applied Linguistics: Challenges for Theory and Practice*, London, UK: Palgrave Macmillan, 2006, pp. 245– 270.

[18] S. Karita et al., "A comparative study on transformer vs RNN in speech applications," in *Proc. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Sentosa, Singapore, Dec. 2019,pp. 449–456

[19] .M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[20] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, Jun. 2013, pp. 3570–3577.