IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SkillSight: An AI-Powered Platform for Soft Skills Interview & Assessment

[1]Aditya Tighare, [2]Vedant Mete, [3]Proshanjeet Chanda, [4]Sanket Rupnavar, [5]Kavita Patil

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor

[1,2,3,4, 5] Computer Science and Business Systems Rajarshi Shahu College of Engineering Pune, India

Abstract: The integration of Artificial Intelligence (AI) into recruitment processes has changed traditional hiring methodologies, enabling more subtle assessments of candidates' competencies. This paper presents an AI-driven mock interview system that evaluates both technical proficiencies and soft skills through multimodal analysis. The system employs computer vision techniques to analyse facial expressions and gestures, utilizing datasets like FER+ for emotion recognition. Natural Language Processing (NLP) models, such as BERT, are finetuned to assess the semantic relevance and clarity of candidates' responses. Additionally, computational paralinguistics extract vocal features—intonation, pitch, and speaking rate—to infer affective states and communication effectiveness. By integrating these modalities, the system provides comprehensive feedback, translating complex model outputs into actionable insights, thereby facilitating targeted improvements in interviewee performance. This approach not only enhances the objectivity and efficiency of candidate evaluations but also addresses ethical considerations by ensuring interpretability and transparency in AI-driven assessments.

Index Terms **- Artificial Intelligence, Mock Interview System, Multimodal Analysis, Facial Expression Recognition, Natural Language Processing, BERT, Computational Paralinguistics, Soft Skills Assessment.**

I. INTRODUCTION

In today's competitive job market, the ability to excel in interviews hinges not only on technical expertise but also on mastery of soft skills such as communication, emotional intelligence, and adaptability. Despite their critical importance, traditional interview preparation methods often fall short in addressing these aspects, offering limited scope for person- aized feedback and non-verbal communication analysis.

To bridge this gap, SkillSight emerges as an innovative, AI powered platform designed to revolutionize how individuals prepare for interviews. By simulating real-world scenarios, the platform integrates advanced technologies such as natural language processing (NLP), facial expression recognition, and behavioural analysis to assess candidates comprehensively. It provides users with tailored feedback on verbal and non-verbal performance metrics, empowering them to refine their soft skills effectively.

This paper delves into the practical implementation of SkillSight, detailing its system architecture, core methodologies, and technological underpinnings. From real-time video and audio data capture to adaptive learning recommendations, SkillSight leverages state-of-the-art AI solutions to create an immersive and impactful interview practice environment. By addressing challenges such as scalability, data privacy, and real-time processing, the platform demonstrates its potential as a transformative tool for enhancing employability and professional growth.

II. Problem Statement

In the modern job market, excelling in interviews requires a balance of technical knowledge and soft skills such as effective communication, emotional intelligence, and adaptability. However, traditional interview preparation methods fall short in addressing critical aspects of soft skills development. Existing systems often lack personalized feedback, fail to analyse nonverbal communication such as facial expressions and gestures, and do not provide a comprehensive, immersive simulation of real-world interview scenarios.

Additionally, with the growing prevalence of virtual interviews, job seekers face challenges in presenting themselves effectively online, where non-verbal cues play a significant role in creating a positive impression. Fresh graduates and professionals alike struggle to refine these skills due to the absence of platforms that integrate AI- driven insights and dynamic, adaptive learning. Employers also face inefficiencies in traditional hiring processes, where assessing candidates' soft skills often requires multiple interview rounds. This increases hiring costs and fails to identify candidates who might struggle due to underdeveloped interpersonal abilities, leading to higher turnover rates.

SkillSight aims to address these challenges by providing an AI- powered solution that offers real-time, data-driven feedback on verbal and non-verbal communication, enabling job seekers to refine their soft skills effectively while reducing inefficiencies for employers.

III. Literature Survey

With the increasing adoption of AI-driven solutions in employability enhancement, various systems have emerged focusing on mock interviews, candidate evaluation, and skill development. These systems utilize advances in deep learning, natural language processing, and multimodal analysis to assess both the technical and non-technical aspects of candidate performance. This section presents a critical review of related works relevant to the proposed system.

Vaibhav Sharma et al. [1] provide a comprehensive review of AI-based interview systems, highlighting the emergence of intelligent systems capable of automating interview evaluations through NLP, facial analysis, and sentiment detection. The review also underlines the need for integrating soft skill evaluation as a complementary metric to traditional answer correctness.

Sahil Temgire et al. [2] proposed a real-time mock interview system using deep learning techniques. Their system focuses primarily on textual analysis and basic emotion recognition from audio input, laying foundational work in AI-based interactive interviews. However, it lacks an adaptive learning component and deeper multimodal fusion.

Al Asefer and Zainal Abidin [3] emphasize the significance of soft skills such as communication, emotional intelligence, and adaptability in enhancing employability. Their findings justify the inclusion of soft skill analysis in automated interview platforms, aligning with the core objective of our system.

Lamri and Lubart [4] introduce a framework for reconciling hard and soft skills under a unified competency model. Their work informs the design of composite scoring systems that evaluate

candidates not only on the basis of correct answers but also their delivery, behaviour, and emotional state.

Noel Jaymon et al. [5] presented a system for real-time emotion detection using deep learning. Their work on facial expression recognition using CNNs has influenced the facial emotion sub- module in our architecture, particularly in leveraging FER+ and AffectNet datasets for emotion classification

Rodriguez et al. [6] raise ethical concerns about AI-based interview systems, especially in terms of bias in NLP algorithms. Their study reinforces the need for fairness-aware design in candidate evaluation engines, motivating our use of interpretable features and explainable AI modules.

Guodong Guo and Na Zhang [7] explore the challenges in deep learning for face recognition in uncontrolled environments, providing important guidelines on model robustness and generalization—critical in real-time video interviews under diverse user settings.

Tengfe Song et al. [8] contribute a multimodal physiological emotion dataset that supports the development of systems capable of discrete emotion recognition using video and audio. This inspires our integrated emotion and prosody analysis for richer soft skill inference.

Mittal et al. [9] review classical and deep learning methods for speech and speaker recognition. Their study informs the design of the acoustic processing pipeline within our system, particularly the use of spectrogram-based CNNs and transformer based acoustic encoders.

In addition to research literature, commercial platforms like HireVue [10] and Google's Interview Warmup [11] showcase industry interest in automated interviews. While these platforms provide large-scale interview simulation, they often lack open scientific evaluation metrics and adaptive feedback mechanisms, which our system aims to provide.

Jayaram et al. [12] edited Bridging the Skills Gap, focusing on addressing the disconnect between industry needs and the skills of job seekers. The volume explores technical and vocational education frameworks to improve training programs and better align them with the evolving job market.

Moreover, policy and training institutions such as the USAID

[13] and SHRM [14] have stressed the importance of quantifiable soft skill training. Our system aligns with this vision by offering data-driven, personalized improvement suggestions based on soft skill deficiencies.

Mollahosseini et al. [15] introduced AffectNet, a large-scale dataset for facial expression recognition in real-world settings, annotated for both categorical emotions and continuous dimensions like valence and arousal. The system uses pretrained CNN models based on AffectNet to detect emotions from video frames during mock interviews. Unlike AffectNet's focus on emotion detection, the current system integrates emotion, gesture, and prosodic analysis for a comprehensive soft skills evaluation.

Devlin et al. [16] developed BERT, a model that enables deep bidirectional context understanding in language, widely used for tasks like sentiment analysis. The system incorporates finetuned BERT to assess the accuracy, clarity, and relevance of interviewee responses, adding multi- dimensional evaluation by integrating audio-visual cues and feedback generation.

Schuller and Batliner [17] highlighted the role of vocal features such as pitch, prosody, and speaking rate in interpreting emotional states. The system uses these features, extracted through MFCCs and pitch analysis, to infer traits like clarity, nervousness, and confidence, enhancing soft skill assessment.

Lipton [18] discussed the importance of model interpretability and user-centric explanations in deep learning. The system includes interpretable feedback modules, translating complex model outputs into actionable advice, such as "Speak with more energy" or "Avoid filler words," to improve usability and educational value.

IV. System Overview

The architecture of the proposed mock interview system is designed to facilitate end-to-end automation of interview simulation, response evaluation, and personalized feedback delivery, with a specific emphasis on soft skill assessment and answer accuracy. The system comprises three primary functional layers: User Interaction, Response Acquisition and Analysis, and Feedback and Personalization.

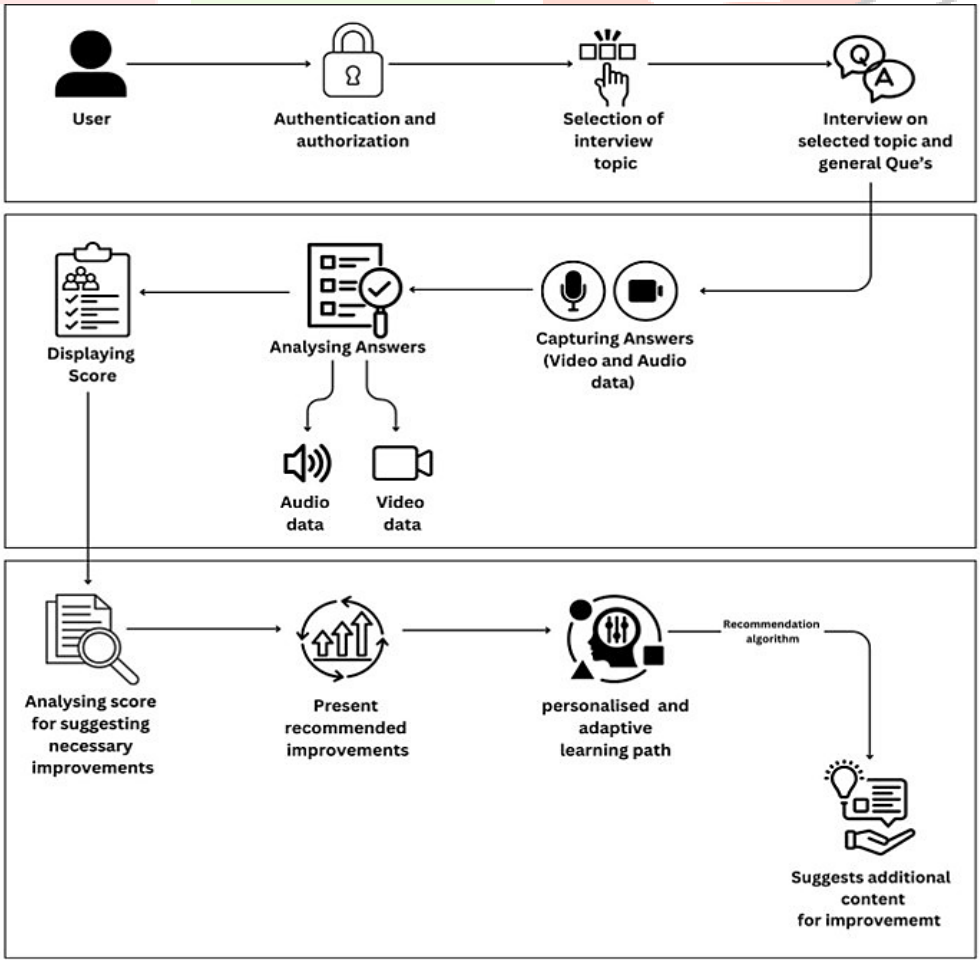

Fig. 1. System Architecture

4.1 User Interaction Layer

This layer handles the initial engagement with the system. The process initiates with secure authentication and authorization of users, ensuring access control and data integrity. Once authenticated, users proceed to select an interview topic from a predefined set of domains. Subsequently, the system conducts an automated interview session, presenting a mix of domain-specific and general behavioural questions.

4.2 Response Acquisition and Analysis Layer

During the interview, the system leverages integrated media modules to capture multimodal data, specifically video and audio responses from the user. This data is passed to the analysis module, which incorporates speech recognition and computer vision algorithms to extract and analyse verbal and non-verbal cues. Key soft skill indicators such as tone, pitch, speech clarity, facial expressions, eye contact, and body language are evaluated in conjunction with semantic accuracy of the responses.

The analysis engine computes a composite score, representing the user's performance across communication, confidence, domain knowledge, and behavioural metrics. This score is visualized and presented to the user in a structured format.

4.3 Feedback and Personalization Layer

Following score generation, the system performs a diagnostic analysis to identify strengths and deficiencies. Based on the evaluation, a recommendation engine suggests targeted improvements and presents a personalized, adaptive learning path. This path is dynamically generated using content-based filtering and rule-based logic that align with the user's specific improvement areas.

To reinforce learning, the system also suggests supplementary resources—such as curated videos, reading materials, and practice exercises—tailored to the user's weak zones. This closed feedback loop ensures continuous self-improvement, mimicking real-world interview preparation.

V. METHODOLOGY

This section outlines the detailed methodological pipeline used in the implementation of the mock interview system. The system is modular, and each module contributes to the goal of simulating realistic interview conditions, analysing both verbal and non-verbal performance, and recommending personalized improvements.

Users interact with the system via a secure web-based interface. Upon successful authentication, they proceed to select a domain from a predefined dataset $D=d1,d2,...,dn$. Based on the selected topic, the system initiates an interview simulation by generating a sequence of questions $Q=q1,q2,...,qk$ comprising both technical and behavioural types:

5.1 Data Acquisition Module:

During the interview, real-time audio-visual data is captured through the user's device camera and microphone.

Audio Stream a(t): Time-series of spoken audio input. Video Stream v(t): Frame-wise visual data including facial expressions and gestures. Both a(t)) and v(t) are timestamp-synchronized and stored as tuples $D=(a1,v1),...,(ak,vk)$ for each question-response pair.

5.2 Answer Analysis Module:

This module performs multi-modal analysis using signal processing and machine learning models. The analysis includes:

A. Video data analysis:

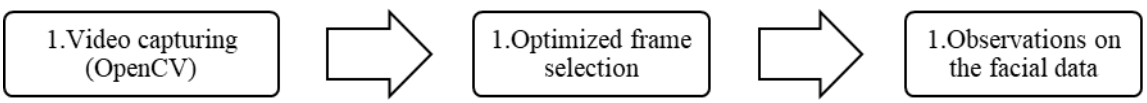

Fig. 2. Video analysis module

To extract non-verbal behavioural indicators, the system employs a computer vision pipeline that processes the user's video stream in real time. The procedure involves the following stages:

i. Video Capturing: The system captures continuous video data using the OpenCV library. The video stream is divided into frames $vi=f1,f2,...,fn$, which are then subject to optimized frame selection strategies to reduce computational redundancy while preserving expressive variance.

ii. Frame Selection: An adaptive frame selection technique is employed to retain frames with significant motion or facial variation. This improves the relevance and accuracy of downstream inference.

iii. Facial Expression and Gesture Recognition: Selected frames are passed through a pre-trained Convolutional Neural Network (CNN) model fcnn(fi) trained on FER+ datasets. The model outputs a vector of probabilities across emotion classes:

$$Ei = f_{CNN}(f_i) = [p_{happy}, p_{sad}, p_{angry}, , \ldots, p_{neutral}] \quad (1)$$

The highest scoring emotion is selected as the primary facial state of the frame. In parallel, gesture attributes such as head nods, hand movements are inferred using pose estimation techniques.

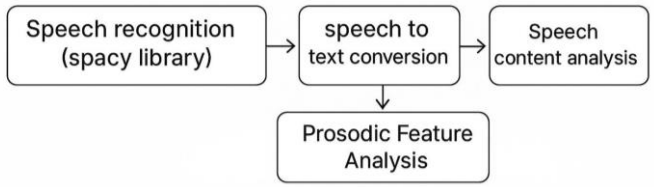

Fig. 3. audio analysis module

B.Audio data analysis:

i. Speech Content Analysis:

We use an Automatic Speech Recognition (ASR) model – $f_{ASR}(\alpha i) \rightarrow$ Ti to convert audio αi into a text transcript Ti. This transcript is semantically evaluated against a ground-truth vector Gi using cosine similarity:

$$\text{Accuracy}_i = \cos(T_i, G_i) = \frac{T_i \cdot G_i}{||Ti||\,||Gi||} \quad (2)$$

ii. Prosodic Feature Analysis:

Features such as pitch (f0), speech rate, energy, and pauses are extracted from the speech signal. A prosodic score is computed using a weighted linear combination:

$$\text{ProsodicScore} = \alpha 1 \cdot \text{PitchVariance} + \alpha 2 \cdot \text{EnergyMean} - \alpha 3 \cdot \text{SilenceDuration} \quad (3)$$

5.3 Aggregated Soft Skill Scoring:

A weighted fusion of audio and video insights gives an overall soft skill score:

$$SoftSkill = \beta 1 \cdot ProsodicScore + \beta 2 \cdot GestureScore + \beta 3 \cdot FacialEmotion \quad (4)$$

5.4 Report Generation

Scoring and Evaluation Engine The final score for each response is computed as:

$$FinalScorei = \gamma 1 \cdot Accuracyi + \gamma 2 \cdot SoftSkilli \quad (5)$$

This score is normalized to a scale of 0–100 and presented to the user through an intuitive dashboard. The dashboard visualizes:

- Individual question scores
- Communication metrics
- Confidence heatmaps over time

5.5 Feedback Recommendation Module

The score vector S=s1,...,sk is passed to a recommendation algorithm $f_R(S)$, which performs:

- Error diagnosis: Identify below-threshold scores.
- Learning path generation: Construct a directed acyclic graph (DAG) of learning objectives where each node represents a micro-skill (e.g., “maintain eye contact”, “answer concisely”).

$$L = f_R(S) = DAG(M, E) \quad (6)$$

Where M are micro-skills and E are learning dependencies. Content resources C=c1,c2,...,cn are mapped to M using a relevance score computed by semantic similarity between skill descriptions and content metadata.

This methodology ensures a seamless flow from user interaction to actionable feedback, enabling users to refine their soft skills effectively.

VI. Conclusion

The proposed system implements a multimodal framework for soft skill evaluation during mock interviews by integrating computer vision, speech processing, and natural language understanding techniques. Facial expressions are analyzed using pre-trained CNN model, while vocal attributes such as pitch, energy, and speaking rate are used to infer prosodic and paralinguistic features. Additionally, semantic content of interview responses is assessed using fine-tuned BERT models. These components collectively enable emotion detection, gesture interpretation, and response quality analysis. Interpretable feedback modules convert these analyses into user-friendly suggestions, offering an effective and scalable approach for soft skill development.

VII. Future Work

Future work will focus on expanding the system's dataset to include more diverse user profiles for improved generalization. Enhancements will also include refining the synchronization between audio-visual modalities and improving the temporal resolution of gesture and emotion tracking. Further development of the feedback module will aim to incorporate more nuanced performance metrics and adaptive feedback strategies based on user history. Additionally, continuous evaluation through user studies will be conducted to assess system effectiveness and usability in real-world training environments.

References

[1] V. Sharma, V. K. Nishad, S. Chaurasia, N. Raj, and K. Devi, "Unveiling the Potential of AI: A Comprehensive Review of AI-Based Interview Systems," *Int. J. Res. Anal. Rev. (IJRAR)*, vol. 10, no. 2, May 2023.

[2] S. Temgire, A. Butte, R. Patil, V. Nanekar, and S. Gavhane, "Real Time Mock Interview using Deep Learning," Int. J. Eng. Res. Technol. (IJERT), vol. 10, no. 05, May 2021.

[3] M. A. Al Asefer and N. S. Zainal Abidin, "Soft Skills and Graduates' Employability in the 21st Century from Employers' Perspectives: A Review of Literature," Int. J. Infrastruct. Res. Manag., vol. 9, no. 2, Dec. 2021.

[4] J. Lamri and T. Lubart, "Reconciling Hard Skills and Soft Skills in a Common Framework: The Generic Skills Component Approach," J. Intell., vol. 11, p. 107, 2023.

[5] N. Jaymon, S. Nagdeote, A. Yadav, and R. Rodrigues, "Real Time Emotion Detection Using Deep Learning," in Proc. Int. Conf. Adv. Elect., Comput., Commun. Sustain. Technol. (ICAECT), 2021, doi: 10.1109/ICAECT49130.2021.9392584.

[6] M. D. Rodriguez, E. Garcia, and S. Srinivasan, "Addressing Bias in AI-Based Interview Systems: A Case Study of Natural Language Processing Algorithms," Proc. ACM Hum.-Comput. Interact., 2022.

[7] G. Guo and N. Zhang, "What is the Challenge for Deep Learning in Unconstrained Face Recognition?" in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), 2018.

[8] T. Song, W. Zheng, C. Lu, Y. Yuanzong, X. Zhang, and Z. Cui, "MPED: A Multi-Modal Physiological Emotion Database for Discrete Emotion Recognition," IEEE Access, vol. 7, pp. 12177–12191, 2019.

[9] A. Mittal, M. Dua, and S. Dua, "Classical and Deep Learning Data Processing Techniques for Speech and Speaker Recognitions," in Signals and Communication Technology, Springer, Cham, 2021.

[10] HireVue, "Video Interviewing and Recruitment Solutions." [Online]. Available: https://www.hirevue.com/

[11] Interview Warmup - Grow with Google. Available: https://grow.google/interview-warmup/

[12] S. Jayaram, W. Munge, B. Adamson, D. Sorrell, and N. Jain (eds.), Bridging the Skills Gap, Tech. and Voc. Educ. and Training, vol. 26, Springer, Cham.

[13] United States Agency for International Development (USAID), "Measuring Soft Skills Life Skills in International Youth Development Programs: A Review and Inventory of Tools."

[14] Society for Human Resource Management. Available: https://www.shrm.org/

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affect. Comput., 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL, 2019.

[17] B. Schuller and A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2014.

[18] Z. C. Lipton, "The Mythos of Model Interpretability," Commun. ACM, 2018.