**Comparative Study**

1. **DEVELOPMENT OF AN AI-BASED INTERVIEW SYSTEM FOR REMOTE HIRING**

International Journal of Advanced Research in Engineering and Technology (IJARET)

B C Lee

Ph.D. Candidate of Seoul Business School, Seoul School of Integrated Sciences and

Technologies (aSSIST), Seoul, South Korea

B Y Kim

Ph.D. Candidate of Seoul Business School, Seoul School of Integrated Sciences and

Technologies (aSSIST), Seoul, South Korea

Methodology : V4 is the AI tech that extracts features from visual, vocal, verbal, and vital information, based on video and voice information. Visual technology analyzes 68 facial points to assess emotions like joy, sadness, and anger by examining delicate movements. It provides insights into eye and head movements, enabling a comprehensive understanding of an applicant’s emotional responses and reactions. Auditory technology analyzes vocal waves to assess utterance time, speed, and volume, while verbal technology examines linguistic habits and word usage. Vital data, including blood flow and pulse, are collected to determine emotional state and dishonesty. Voice data in videos are converted into text using the Google STT API and analyzed through a natural language processing engine. The development environment includes a WAS area based on Linux Ubuntu and a Deep Learning area on Linux CentOS, using MongoDB and MySQL, respectively. Data is exchanged via JSON protocol and results provided through RESTful API. The AI interviewer evaluates applicants' biological signals and responses, assessing emotions and genuineness from facial expressions, voice, vocabulary, and pulse. Interviews are recorded, and AI evaluation sheets are generated for personnel managers. The AI system, trained on over 400,000 interview videos, achieves high evaluation reliability with a Pearson score of 0.88.

Validity : As each evaluation item's classification accuracy was analyzed for the AI-based job

interview evaluation model, the accuracy was as high as 70% in every evaluation item.

The accuracy rate of performance competency was 75%.

relationship competency 72%

organizational fitness 70.2%

fitness to the official position 82%

the level of satisfaction was 85.5% on average

highest score (92) was given to the item of efficiency in the

evaluation process.

1. **Doing more with less: Using AI-based Big Interview to combine exam preparation and interview practice**

H. Kevin Fulk

Heidi L. Dent

William A. Kapakos

Barbara Jo White

Methodology : Big Interview, an interview preparation tool, features training videos, job-related content, and a diverse question database. Users can create and record interview questions and set evaluation criteria. The AI tool reviews responses based on various criteria but not user-created ones. Institutional licenses cost $3,000-$13,000. In Spring 2022, a survey in 14 business course sections (11 in-person, 3 online) assessed student responses. Activities included recording video responses to questions asked by male and female voices and reviewing AI-generated scores and feedback. Students reflected on their use of Big Interview compared to traditional interview and exam preparation methods.

Validity : A total of 172 students completed the survey, but not all completed both the video assignment and the survey. Surveys were excluded if videos were too short (less than 30 seconds), if students took the survey more than once in the same class, completed assignments in multiple classes, or selected the wrong class section. After excluding 29 surveys, 143 remained in the sample. A one-sample T-test revealed students liked using Big Interview for exam preparation (M = 4.36, SD = 1.85, p = .048) but did not find it significantly easier or more time-efficient than written review questions. Students reported learning more with Big Interview (M = 4.49, SD = 1.83, p = .011) and found its AI feedback accurate (M = 5.61, SD = 1.38, p < .001). For job interview preparation, students preferred Big Interview over written reports (M = 6.25, SD = 1.30, p < .001), found it more educational, spent more time using it, and found it easier and accurate. No significant differences were found for gender, age, or major. Face-to-face students preferred Big Interview more than online students.

1. **Artificial Intelligence Based System for Preliminary Rounds of Recruitment Process**

Sayli Uttarwar, Simran Gambani, Tej Thakkar, and Nikahat Mulla

Department of Information Technology, Sardar Patel Institute of Technology, Mumbai 400058, India

Methodology : The five modules that the system can be broken down into are Video/Audio to Text, Question Answering System, Candidate Response Evaluation, Adaptive

Question Generation and Report Generation, explained in detail as follows.

The system uses multithreading for interviews, capturing video (OpenCV) and audio (Pyaudio), merging them with FFMPEG, and creating four files: video, audio, merged, and text from audio via Google Speech-to-Text API. Video is analyzed for facial expressions, and audio for speech-to-text and vocabulary. The QA system, designed for Java interviews, uses a dataset inspired by SQuAD, scraped from various sources, and processed with word embeddings and token features using NLP techniques. The system evaluates candidate answers by comparing them with QA system-generated answers using similarity measures like Cosine Similarity and sentiment analysis. An adaptive question generation adjusts difficulty based on candidate performance. The system provides detailed candidate performance reports, supporting interviewer decisions. Experimental analysis involved comparing different embedding and similarity techniques, implementing the most efficient ones. The embedding layer used a modified skip-gram model and GloVe vectors, with spaCy for feature extraction. The system's QA component, tested with Java method passages, answers related questions, showing the QA process's effectiveness and response time.

Validity : For the system to accurately evaluate candidates, the QA component must perform optimally. We modified Facebook's original open-domain QA model to cater to a closed-domain Java dataset. Answers are judged using EM (exact match) and F1 scores, achieving 69.76% and 78.6% respectively over 40 epochs. Our model's EM score surpassed other QA models, while the F1 score was comparable. Though not better overall, our modified neural network-based model performed well in terms of relevance and correctness. Further improvements to our LSTM-based architecture could enhance performance.

1. **Dear Computer on My Desk, Which Candidate Fits Best? An Assessment of Candidates’ Perception of Assessment Quality When Using AI in Personnel Selection**

Jessica Schick, Sebastian Fischer

Methodology : The study targeted students aged 18-28, inviting around 200 via online platforms, resulting in 96 valid responses (53% female, 44% male, 3% diverse). Conducted in German from January to March 2020, the survey took about six minutes. Using a cross-sectional vignette design, participants assessed AI in job application scenarios. Each received five randomized vignettes out of 27 possible combinations of AI complexity, intangibility, and reliability. After each vignette, participants rated perceived assessment quality and ranked AI complexities. Data were analyzed using MANOVA in SPSS 25.0.

Validity : The study utilized MANOVA and post hoc analyses to examine AI complexity, intangibility, and reliability on assessment quality perception. Significant effects were found for AI complexity (ηp2=0.045) and AI intangibility (ηp2=0.02), but the three-way interaction showed marginal significance (ηp2=0.025).

Hypothesis 1: AI complexity negatively affects candidates' perception. Significant relationships were found with knowledge (ηp2=0.044) and strengths and weaknesses (ηp2=0.026), but not with motivation. The complex AI method "robotic interview" was perceived negatively compared to algorithms and speech analysis.

Hypothesis 2: AI intangibility negatively affects perception. Significant relationships were noted with knowledge (ηp2=0.025) and strengths and weaknesses (ηp2=0.017), but not with motivation. Personality evaluation was seen as lower quality compared to skills and job performance forecasts, leading to the rejection of Hypothesis 2.

Hypothesis 3: AI reliability positively affects perception. No significant relationships were found with motivation, knowledge, or strengths and weaknesses, leading to the rejection of Hypothesis 3.

Hypothesis 4a: High AI complexity and intangibility positively impact perception if AI reliability is high. Supported partly by significant differences in performance forecasts.

Hypothesis 4b: High AI complexity and low intangibility positively impact perception if AI reliability is low. Supported partly by significant differences in skill evaluations.

Hypothesis 4c: AI complexity and reliability are additive, with low complexity and high reliability leading to better perceptions. Partly supported by significant differences in skill evaluations.

1. **Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes?**

**Hung-Yue Suena, Mavis Yi-Ching Chena, Shih-Hao Lub**

Method :

Design: This experimental study compared applicant and rater responses across SVI, AVI, and AVI-AI conditions. Initially, three senior HR professionals piloted the study to validate questionnaires and hypotheses. The sponsor posted a job description for HR managers, attracting 180 qualified applicants who submitted resumes and signed consent forms.

Sample Composition: Each interview type had 60 applicants, with varied gender, age, and education distribution.

Interviews and Media Tools: AVI-AI and AVI used the HRDA platform; SVIs were conducted via Skype. All interview types followed a structured format with behaviorally oriented questions assessing communication skills. In AVI-AI, AI algorithms analyzed interview responses; in AVI, interviews were recorded for later viewing. SVI interviews were live.

Ratings and Measures: Raters, trained in a 2-hour session, scored interviews using a 5-point scale. Initial impressions and physical appearance ratings were also collected. Applicants completed questionnaires on favorability and perceived fairness.

Data Analysis: Analysis was performed using IBM SPSS v23, employing Chi-square, ANOVA, ANCOVA, MANCOVA, and multiple linear regression methods to test hypotheses. Control variables included gender, age, education, interview motivation, and video interview experience.

Validity : Applicant initial impression (Mean = 3.57) and appearance (Mean = 4.02) were highly correlated with interview score (Mean = 3.23). Favourability towards the interview process (Mean = 3.22) and perceived fairness (Mean = 3.50) were slightly correlated with interview score and initial impression. Higher favourability was linked to higher appearance ratings and motivation (Mean = 3.75). Applicants reported more video interview experiences in SVI (Mean = 1.40) and preferred SVI over AVI or AVI-AI.

Chi-square tests and ANOVAs found no significant differences in control variables among groups except for education, which was included as a covariate. ANCOVA showed initial impression and appearance significantly influenced interview scores, supporting H1 and H2. H3 was fully supported and H4 partially supported, indicating stronger effects of initial impression in SVI.

MANCOVA results indicated that synchrony significantly affected favourability towards the interview process but not perceived fairness, supporting H5 but not H6. ANCOVA comparing AVI and AVI-AI found no significant effect of AI decision agents on perceived fairness, so H7 was not supported. Education level did not significantly impact favourability or perceived fairness.

1. **Personality Recognition & Video Interview Analysis**

Supriya Anand

SIT (Department of CSE)

Symbiosis International University,

Pune- 411045

Nihar Gupta

SIT (Department of CSE)

Symbiosis International University,

Pune- 411045

Mayesh Mulay

SIT (Department of CSE)

Symbiosis International University,

Pune- 411045

Abhimanyu Sherawat

SIT (Department of CSE)

Symbiosis International University, Pune- 411045

Asst. Prof. Rupali Gangarde

SIT (Department of CSE)

Symbiosis International University, Pune- 411045

Methodology :

Interpreting CNN and DNN Models

Convolutional Neural Networks (CNN) are deep neural networks that excel in image processing with minimal preprocessing. CNNs mimic the visual cortex's connectivity patterns, utilizing convolution to extract analytical features and fully connected layers for final predictions. The CNN architecture is inspired by the primate cortical area, with layers recognizing progressively complex information. Neurons in a CNN form a 3D structure, each focusing on a part of the image, ultimately providing a probability vector for classification.

Deep Neural Networks (DNN) transform abstract definitions into human-understandable domains. A DNN consists of neurons in layers, each receiving inputs from the previous layer and performing computations. The neurons collaboratively form a dynamic nonlinear mapping from input to output.

Resume Parsing

Resume parsing technology extracts data from resumes, simplifying the screening process by identifying contact information, skills, job history, and education. Using OCR and deep learning algorithms, resume parsers achieve high accuracy, effectively modeling the context of words to handle variations in company names, institutions, and degrees.

Speech Emotion Recognition

Emotion plays a significant role in human interaction and decision-making. Automatic emotion recognition, using hand-crafted features and models like CNN and CNN Alex Net, identifies emotions from speech data. Advanced models like DBN and SVM classifiers enhance accuracy. Sequential models like LSTM further refine emotion detection.

Facial Emotion Recognition

Facial recognition maps facial features and compares them to a database for identification. Algorithms measure facial geometry to create a facial signature, aiding in emotion detection. Machine learning algorithms, such as Support Vector Machines, extract features for accurate classification. The Facial Action Coding System (FACS) quantifies facial movements, essential for recognizing expressions.

Company Verification

Using resume parsing, DNN models analyze profiles from Naukri.com to recommend suitable job roles based on skills. Keras library aids in developing these deep learning models, applied to 25 job roles and 10,000 profiles for job recommendations.

Data Processing

Data collection involved recording interviewees' audio and visual responses. Data labeling mapped facial features for recognition. Pretrained Inception-v3 dataset standardized images for feature extraction. The DNN model for resume parsing used data from Naukri.com, while voice and facial emotion recognition models used datasets from Kaggle. Training involved using TensorFlow and Keras to predict emotions, with implementation strategies ensuring a smooth transition from manual to automated systems.

Validity : The project successfully parsed resumes and provided job recommendations from 25 roles using a CNN model. In speech recognition, the model accurately identified gender and classified emotions, improving from an initial accuracy of 0.5500 to 0.9150, and reducing data loss from 1.2252 to 0.1619. In video emotion recognition, the system identified emotions like happiness and anger, with accuracy increasing from 0.2560 to 0.9476 and data loss decreasing from 1.8090 to 0.1471. The model effectively assists in job recommendations and recruitment processes.

1. **TensorFlow-based Automatic Personality Recognition Used in Asynchronous Video Interviews**

Methodology : Personality refers to individual differences in thinking, feeling, and behaving. It predicts job performance and cultural fit, with the "big five" traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—commonly used for assessment. Self-rating and observer-rating approaches measure these traits, with self-ratings often foundational in predicting workplace behaviour.

Personality computing leverages AI to assess personality through automatic personality recognition (APR), perception (APP), and synthesis (APS). APR uses audio-visual data to recognize self-assessed traits, while APP predicts observer-rated traits using distal cues, as proximal cues are challenging to measure. APS involves artificial agents exhibiting human-like behaviours perceived as specific traits.

Data for this study was collected using AVI software in a real job interview context. 120 applicants for HR positions were recorded answering structured interview questions, and their self-rated big five traits were collected using an online survey. Facial expressions were captured using a pretrained Inception-v3 dataset and analysed using OpenCV and Dlib.

The APR model was trained with a CNN using Python and TensorFlow. Features were normalized, and the model included convolutional, pooling, and fully connected layers. Inputs were grayscale images of facial expressions, and outputs were self-rated big five traits. The dataset was split 50-50 for testing and validation, with 4,000 training iterations.

The model's performance was assessed using concurrent validity, Pearson correlation coefficient, coefficient of determination (R²), and mean square error (MSE). These metrics evaluated the correlation between the APR and self-reported inventory, with higher R² and lower MSE indicating better model performance.

Validity : Prior to assessing the performance of our automatic personality recognition (APR), we validated the self-reported personality traits using IBM’s SPSS v23. Construct validity was confirmed with factor loadings above 0.6 and a Kaiser-Meyer-Olkin (KMO) value over 0.8. Internal consistency reliability was strong, with Cronbach's alpha values exceeding 0.7 for all traits: openness to experience (0.75), conscientiousness (0.83), extraversion (0.88), agreeableness (0.80), and neuroticism (0.84).

The AI model, built using TensorFlow, successfully learned and predicted the big five personality traits. The Pearson correlation coefficients for each trait ranged from 0.966 to 0.976, with R² values between 0.933 and 0.952, all significant at p < 0.01. The mean square error (MSE) for each trait was between 0.053 and 0.120. Higher R² values indicate a better model fit, and lower MSE values indicate smaller estimation errors. The classification accuracy averaged 95.36%, demonstrating the model's effectiveness in predicting self-assessed big five personality scores.

1. **Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It**

Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea1, and Mohammed (Ehsan) Hoque

ROC HCI, Department of Computer Science, University of Rochester

ROC HCI, Department of Electrical and Computer Engineering, University of Rochester

Methodology :

Data Collection:

Study Setup: Interviews were conducted in a room with a desk, two chairs, and two wall-mounted cameras with microphones to capture facial expressions and audio.

Participants: Initially, 90 juniors, all native English speakers, participated. Professional career counsellors conducted the interviews before and after an intervention, with participants receiving $50. The resumes of the top 5% were forwarded to sponsoring organizations. Finally, 69 participants (26 male, 43 female) allowed their videos to be used for research.

Procedure: Each session involved five questions focusing on behavioural and social skills. The total video duration was 10.5 hours, averaging 4.7 minutes per interview. This dataset is one of the largest collections of job interview videos under realistic settings.

Data Labelling:

Due to the subjective nature of human judgment, Amazon Mechanical Turk workers rated interview performance using 16 assessment questions on a seven-point Likert scale. Questions focused on overall performance and specific behavioural dimensions like warmth, presence, competence, and content. Ten Turkers were selected based on agreement with counsellors on five control videos. After optimization, nine ratings per video were used to establish ground truth ratings.

Prediction Framework:

We extracted 82 features from the videos and trained two regression models: Support Vector Regression (SVR) and Lasso. The goals were to predict Turkers’ ratings on overall performance and behavioural traits and to understand the importance of individual features.

Feature Extraction:

Prosodic Features: Using PRAAT, we analyzed pitch, vocal intensity, formants, spectral energy, pause duration, and other speech characteristics.

Lexical Features: Transcriptions by Amazon Mechanical Turk workers included filler words. We used psycholinguistic word categories from LIWC and topics from Latent Dirichlet Allocation (LDA). Additional features included words per second, unique words per second, word count, and filler words per second.

Facial Features: Faces were detected using Shore framework, and an AdaBoost classifier distinguished between neutral and smiling faces. We also extracted head gestures like nods and shakes.

Score Prediction:

The combined features were normalized, and SVR and Lasso models, which performed best, were used to predict interview scores and specific traits.

Validity : We analyzed the quality and reliability of Turkers' ratings using Krippendorff's Alpha to measure inter-rater agreement, finding good agreement for traits like engagement, excitement, and friendliness, but low agreement for traits like stress and speaking rate. We trained 16 regression models (SVR and Lasso) on automatically extracted features from interview videos to predict ratings for 16 traits. Using 1000 independent trials with 80% training and 20% testing data splits, we measured prediction accuracy via correlation coefficients. High prediction accuracy (r > 0.70) was achieved for overall performance, hiring recommendations, engagement, excitement, and friendliness. Traits with low prediction accuracy either had low inter-rater agreement or lacked key features, such as eye contact.

Feature analysis revealed that prosodic features were most significant for predicting engagement and excitement, while lexical and prosodic features were important for overall performance and hiring recommendations. Facial features, particularly smiles, were significant for predicting friendliness. Recommendations indicated that higher speaking rates, fluency, positive emotion words, and smiles correlated with better interview performance. Conversely, filler words and negative emotions were detrimental. The models highlighted the importance of prosody, speaking style, and facial expressions in interview success.

1. **Trusting Virtual Agents: The Effect of Personality**

MICHELLE X. ZHOU, Juji, Inc.

GLORIA MARK, University of California, Irvine

JINGYI LI, University of California, Berkeley

HUAHAI YANG, Juji, Inc.

Methodology : To build a personality inference engine, data from Twitter, specifically tweets from 15 million users, was collected. A subset of one million users' tweets was used as training data due to Twitter's public availability, diverse topics, and user base. This diversity helps mitigate potential biases and previous research supports the effectiveness of Twitter data in measuring personality traits. Tweets were examined to identify trait-related evidence based on existing literature, such as references to music and poetry indicating Artistic Interests. Some evidence, like vocabulary richness for Intellectual Curiosity, was computed automatically.

Identifying individual tweets as trait evidence is impractical; instead, Twitter users were associated with traits based on known characteristics. For example, Lady Gaga was linked with Artistic Interests due to her profession. This method allowed for the collection of linguistic cues unique to groups with shared traits, which were used as potential evidence for personality traits.

The relationships between these linguistic cues and traits were quantified using Item Response Theory (IRT). This approach models traits as latent factors and cues as observed items. The occurrence rate and discriminative power of each cue were estimated using formulas and algorithms like Expectation Maximization. This model was then applied to user-generated text to compute trait scores for the Big 5 personality factors and their facets.

To validate the inferred traits, the model's reliability and validity were assessed. Analysing tweets from 100,000 users, the model achieved acceptable reliability (Cronbach’s α >= 0.8) for 19 out of 35 traits with at least 1,000 words of data. Validity was tested by comparing inferred traits with those from a traditional personality test (IPIP-300) taken by 475 student volunteers. Pearson correlation coefficients between the two datasets ranged from -0.088 to 0.294, with 13 traits showing significant correlations (≥0.1). The model’s coefficients were stronger than those in previous studies, attributed to the richer linguistic evidence used.

The inferred traits' validity was further supported by empirical evidence. Hiring managers found suitable candidates based on inferred traits, which matched in-person assessments. Additionally, user perceptions of an AI interviewer, Kaya, were influenced by personality similarity and extraversion, consistent with inferred traits. However, rigorously validating the model is a complex, long-term task. Future work involves collecting longitudinal data to compare inferred traits with real-world behaviours, such as job or academic performance, against self-reported scores.

Validity : In our study, we explored user perceptions and interactions with two AI agents, Kaya and Albert, through multiple deployments. Users rated Kaya as calm, cheerful, and warm, and Albert as calm, rational, and assertive. In high-stakes job interviews, users' perceptions aligned more with the intended personalities of the agents. For instance, Kaya was rated significantly more like a friend than Albert, and Albert was slightly more trusted.

We found that context significantly affected user engagement. Job interviewees rated the agents higher than university students, likely due to the high-stakes nature of job applications. Internship applicants, who were mainly Computer Science majors, were more positive about Kaya, possibly due to their familiarity with AI.

In terms of communication preferences, 83% of users preferred text-based interactions, while 17% preferred voice-based conversations.

We examined the impact of agent personality on users’ willingness to confide in and listen to the agents. Users trusted Albert more and were more willing to confide in him, possibly due to his rational and assertive nature. Trust was a key factor in users' willingness to impress and confide in an agent. Contrary to our hypothesis, users were more honest and more willing to confide in Albert than in Kaya.

When analysing users' willingness to listen, Albert was preferred over Kaya, suggesting that users are more inclined to listen to an assertive and reserved agent rather than one with a similar personality.

We also studied the influence of interview context and individual differences on user behaviour. In high-stakes contexts, users were more likely to confide in Kaya if they were open-minded and cooperative but less likely if they were extroverted or excitement-seeking. The IM Scale indicated higher image inflation in high-stakes contexts, particularly among conscientious males.

Perceived trust in Kaya was influenced by personality similarities and traits such as extroversion and cheerfulness. For Albert, trust was associated with conservativeness. Users rated agent quality and enjoyment higher when they perceived the agent’s personality as similar to their own. Extroverted and cheerful users were more willing to listen to Kaya.

Overall, our findings highlight that trust and personality alignment significantly influence user interactions with AI agents, and context plays a crucial role in shaping these perceptions and behaviours.

1. **Hire Me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behaviour**

Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez, Member, IEEE

Methodology : In this study, a corpus of 62 employment interviews was collected to evaluate job candidates for a marketing position involving recruiting participants for psychology studies. The job, advertised through various channels across three Swiss universities, attracted mainly students, with 90% being Bachelor and Master students, 4.8% PhD students, and 3.2% employed individuals. The average age was 24 years, with a predominance of female applicants (45 females, 17 males). Applicants completed consent forms and psychology questionnaires before participating in a structured behavioral interview, designed to elicit a range of behavioral responses through a consistent sequence of questions.

The interviews, totaling 670 minutes (average duration: 11 minutes), were recorded using two HD cameras and a Microcone microphone array. These recordings captured both audio and visual data, with manual synchronization of audio-video streams. Five hirability scores were assessed: communication, persuasion, conscientiousness, stress resistance, and overall hiring decision. These scores were annotated by a Master student in organizational psychology and validated by a second coder, demonstrating good inter-rater agreement.

The dataset's privacy-sensitive nature prevents public availability, but extracted features and hirability scores can be obtained by contacting the authors. The study extracted nonverbal features from audio and visual modalities, including speaking activity, prosody, head nods, overall visual motion, and head region optical flow. Applicant behavior, such as speaking time, turn duration, and fluency, was correlated with hirability ratings, indicating that fluent and expressive candidates were perceived as more hirable. Interviewer behaviour, including speaking activity and visual back-channelling, was also correlated with hirability, suggesting the interviewer's behaviour adapted to the candidate's perceived quality.

Mutual short utterances between the interviewer and applicant were negatively correlated with hirability, possibly indicating that fluent candidates who required fewer clarifications were rated higher. The study proposed a regression framework to infer hirability scores, using dimensionality reduction methods like PCA and regression techniques such as ordinary least squares, ridge regression, and random forest. The framework employed a leave-one-interview-out cross-validation approach for training and testing.

The performance of the automatic prediction models was evaluated using root-mean-square error (RMSE) and the coefficient of determination (R²). The baseline model predicted the average hirability score, and the regression models aimed to improve upon this baseline. Results indicated that applicant fluency, visual expressiveness, and the interviewer's adaptive behaviour played significant roles in forming hirability impressions.

In summary, this study highlighted the importance of nonverbal cues in job interviews and demonstrated that both applicant and interviewer behaviours significantly impact hirability ratings. The proposed computational framework showed promise in automatically inferring hirability scores, suggesting potential applications in automated interview systems and further research in social computing and organizational psychology.

Validity : Ordinary Least Squares (OLS) regression consistently performed worse than the baseline-average model due to over-fitting, so its performance values were not reported. For the hiring decision variable, ridge regression and random forest outperformed the baseline-average model. Ridge regression achieved the best prediction with all features (R^2 = 0.362) and PCA (R^2 = 0.360), both significantly better than the baseline (p < 0.05). Random forest also yielded significantly better results with all features (R^2 = 0.274) and marginally better with low p-value features (R^2 = 0.289).

For stress resistance, random forest and ridge regression using low p-value features outperformed the baseline-average model with marginal significance (R^2 = 0.272 and R^2 = 0.208, respectively). Ridge regression with all features (R^2 = 0.124) and PCA (R^2 = 0.127) also showed positive, albeit not statistically significant, results.

For persuasion, random forest with all features performed marginally better than the baseline (R^2 = 0.118). However, for communication and conscience, no method surpassed the baseline-average model.

In summary, ridge regression and random forest showed the most promise for predicting hiring decisions and stress resistance, especially when using all features or PCA for dimensionality reduction. Other hirability variables, such as communication and conscience, did not see improved prediction accuracy with the tested methods.