iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71416

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Multimodal Intelligence in Recruitment: Modelling Personality and Behavioural Traits from Video Interviews

K. R. Rajput[1], A. J. Kharade[2], A. P. Pawar[3], T. S. Wakhare[4], Prof. D. D. Ahir[5]
Department of Computer Engineering, MES Wadia College of Engineering, V. K. Joag Path, Pune, 411001, Maharashtra, India.

Abstract: ***With advances in AI and deep learning, automated personality analysis from video interviews has emerged as a key area in personality computing and psychological assessment. Leveraging computer vision and pattern recognition, modern models now interpret nonverbal cues to estimate personality traits directly from visual input. The recruitment landscape often depends on manual assessments that are susceptible to bias and inconsistency, making objective candidate evaluation challenging. While asynchronous video interviews (AVIs) offer scalability and convenience, they still fall short in capturing deeper personality-related cues. This research introduces an Automatic Personality Recognition (APR) framework that leverages multimodal data—text, audio, and visuals—to assess candidates along the Big Five personality traits. By applying advanced deep learning techniques to analyze recorded interviews, the system delivers objective and scalable personality evaluations. This approach enhances the fairness and effectiveness of hiring decisions, addressing key limitations in both conventional and technology-driven recruitment practices.***

Keywords: Big Five Personality Traits, Multimodal Data, Computer Vision, Asynchronous Video Interviews (AVI), Natural Language Processing (NLP), Long Short Term Memory (LSTM), Audio Processing, Deep Learning, Attention Mechanism

I. INTRODUCTION

Asynchronous Video Interviews (AVIs) are increasingly popular in modern recruitment due to their convenience, flexibility, and scalability. However, traditional personality assessments conducted during interviews are susceptible to human judgment, introducing bias and inconsistency. While AVIs address logistical challenges, they lack the face-to-face interaction necessary for a comprehensive assessment. To overcome these limitations, this research proposes an Automated Personality Recognition (APR) system that objectively evaluates the Big Five personality traits—Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. The system employs a multimodal approach, integrating textual, auditory, and visual data to enable recruitment agencies to conduct deeper and more unbiased assessments. Leveraging advanced techniques in natural language processing, computer vision, and speech analysis, the APR system automates the extraction and analysis of multimodal data from video interviews. This effectively addresses challenges related to modality balance, bias reduction, and accuracy. By mitigating the shortcomings of traditional methods, the system provides recruiters with clear, data-driven insights into candidate personalities. Ultimately, this approach lays the groundwork for transforming recruitment processes into scalable, fair, and data-driven decision-making frameworks.

II. LITERATURE REVIEW

A. Multimodal Approach

Recent studies emphasize the efficacy of multimodal systems in automating personality recognition during interviews. One system integrated verbal, visual, and personality cues using standardized video protocols and machine learning models, improving scoring accuracy [21]. HireNet [16], trained on 7,000+ interviews, used hierarchical attention and multimodal fusion, outperforming mono-modal baselines in predicting hireability. Similarly, [24] extracted 29 multimodal features to detect Big Five traits, with SVMs showing highest accuracy in Conscientiousness and Emotional Stability. Large-scale datasets like ChaLearn and First Impressions further support deep learning architectures (e.g., BiLSTM, ViT) for multimodal analysis [18].

B. Textual Features

Text-based methods offer scalable alternatives to traditional assessments. A projective Z-test approach [12] bypasses response biases, achieving AUC-ROC of 0.85. Transformer-based models such as BERT and SBERT show strong performance in interview response analysis, enabling accurate, low-latency personality inference [6][15].

C. Visual Features

Facial expressions, captured via CERT and deep embeddings (e.g., FaceNet), reveal personality-linked patterns [23][22]. Traits like Extraversion correlate with expressive micro-behaviors, supporting hybrid text-visual models. Further work with XLNet improves contextual facial analysis for systems like chatbot-based profiling [4].

D. Key Findings

Multimodal and deep learning models enhance APR accuracy. Integration of sequential (LSTM, GRU) and contextual (Transformer) models strengthens trait detection, especially when fused in ensemble systems. This foundation advances reliable, scalable recruitment assessments.

III. METHODOLOGY

A. Preprocessing and Foundational Choices

Prior to model development, foundational efforts concentrated on selecting appropriate datasets and designing a robust architecture for automated personality recognition (APR). The Big Five personality traits, commonly known as the OCEAN model, served as the theoretical framework guiding the prediction targets. Although recurrent neural architectures such as BiLSTM have shown promise in capturing temporal dependencies in sequential data, our final model architecture integrates convolutional neural networks (CNNs), pretrained visual feature extractors, and dense layers to effectively process the multimodal inputs consisting of audio, visual, and textual data.

Among various datasets, ChaLearn First Impressions V2 was selected for its monologue-based, multimodal format with rich annotations. Unlike UDIVA and MIT datasets, which focus on dyadic interactions and lack detailed labelling, ChaLearn V2 aligns better with asynchronous interview contexts.

TABLE 1
DATASET CHARACTERISTICS

Sr. no	Name	Size	Interview type	Metadata provided
1	First Impressions v2	~10,000 clips of 5-10 seconds, across 2,000 participants	Monologues	Age, gender, ethnicity annotations
2	UDIVA	~188 sessions, 90.5 hours of interactions	Dyadic interactions	Participant demographics, Personality traits
3	MIT Interview Dataset	~2,900 interviews across 1,100 participants	Dyadic interactions	Facial expressions, Speech features, Language use

This approach employs a hybrid neural network architecture designed to leverage the strengths of multiple modalities. Specifically, the audio subnetwork applies convolutional layers to spectrogram representations to capture acoustic features, while the visual subnetwork utilizes a pretrained VGG16 model on individual video frames followed by an LSTM layer to encode temporal dynamics. Textual features extracted via BERT embeddings are processed through fully connected layers with dropout for regularization. The fusion of these modality-specific representations is achieved through concatenation, followed by dense layers to predict continuous scores for the five personality traits. This architecture balances computational efficiency and representational power, outperforming classical methods such as SVMs and purely recurrent models, and offering a practical alternative to more resource-intensive Transformer-based architectures.

B. Data Modalities and Preprocessing

The proposed system integrates multimodal data—comprising audio, video, and text—that captures unique aspects of communication that collectively contribute to the predict candidate behaviour. This study employs the First Impression V2 dataset, comprising 1,891 asynchronous video interviews collected from 260 online participants, amounting to approximately 63 hours of data. Each sample is annotated with Big Five personality trait scores and binary hiring recommendations, labeled by professional raters. The dataset was partitioned into training, validation, and test subsets, ensuring non-overlapping subjects across splits to preserve evaluation integrity.

Preprocessing ensures consistency across modalities while preserving the temporal and semantic integrity of interview responses. Audio was extracted from videos and segmented into 1,319 frames, with 24 Mel-Frequency Cepstral Coefficients (MFCCs) computed per frame to capture pitch, intonation, and rhythm—key indicators of emotional tone and vocal expressiveness relevant to personality perception. Video data were preprocessed by sampling six frames at regular intervals, resizing them to 128×128 pixels, and normalizing for scale and illumination. These frames preserve facial and behavioral cues such as gaze, micro-expressions, and head movements, which are linked to traits like agreeableness and extraversion. For the text modality, transcribed audio was cleaned to remove filler words and artifacts. The resulting transcripts were embedded using Bidirectional Encoder Representations from Transformers (BERT), chosen for its ability to model contextual word usage and sentence-level semantics reflective of personality-linked language patterns. This preprocessing pipeline ensures that each modality is represented in a structured and information-rich format, serving as the foundation for subsequent stages of feature extraction, temporal modelling, and multimodal fusion.

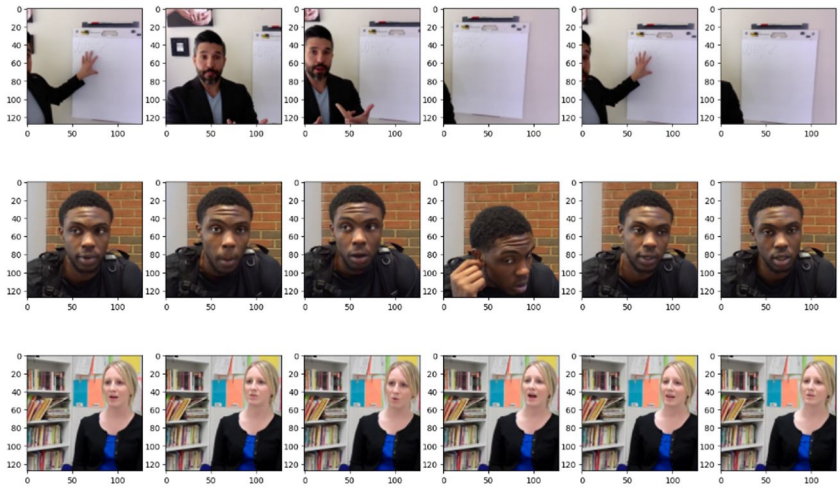


Figure 1 Preprocessing of video frames

C. Feature Extraction Architecture

Following preprocessing, specialized neural subnetworks were designed to extract salient features from each modality, aligned with the unique characteristics of audio, visual, and textual data. The objective was to obtain high-level embeddings that encapsulate both explicit and implicit cues indicative of personality traits from speech, facial behaviour, and language.

In the audio modality, Mel-frequency cepstral coefficients (MFCCs) were used as input to a series of 2D convolutional layers. These layers effectively captured vocal characteristics such as timbre, prosody, and articulation patterns associated with emotional and personality expression. Each convolutional block incorporated batch normalization to enhance training stability and max pooling to reduce dimensionality while preserving critical acoustic features.

The visual modality employed a pretrained VGG16 convolutional neural network applied frame-wise to extract spatial features from individual video frames, such as facial muscle movements, eye gaze, and micro-expressions—key indicators of nonverbal personality traits. These frame-level CNN outputs were temporally sequenced and passed through an LSTM layer to model temporal dynamics across the video segment, preserving alignment with other modalities.

For the textual modality, 768-dimensional embeddings derived from a pretrained BERT model were processed through fully connected dense layers with dropout, rather than BiLSTM layers, to capture semantic nuances of the spoken content. This design choice reflects the fixed-size embedding input and leverages dense layers to learn latent personality-related patterns embedded in language use.

The outputs from these modality-specific subnetworks were then concatenated, forming a comprehensive multimodal representation that integrates spatial, temporal, and semantic features, laying the foundation for subsequent regression to predict personality trait scores.

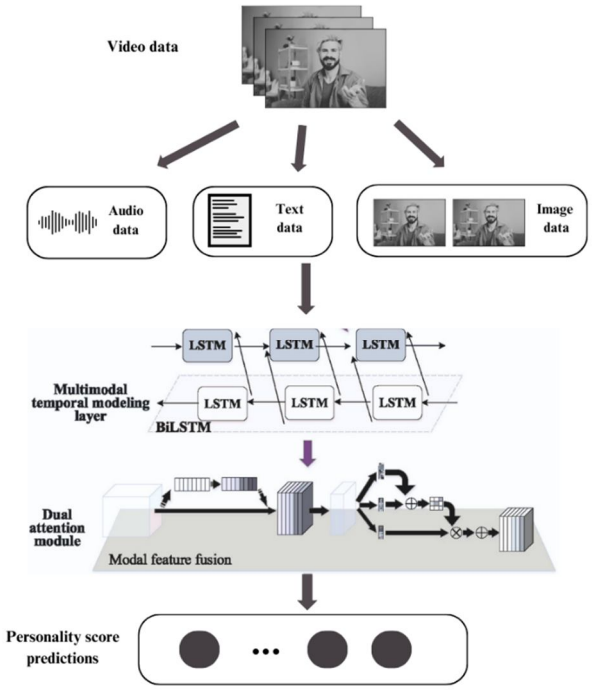


Figure 2 LSTM for multimodal video analysis

D. Temporal Modeling with Recurrent Architecture

After initial feature extraction, the framework employs recurrent architectures to capture temporal dependencies and context-sensitive patterns vital for accurate personality inference. Specifically, Long Short-Term Memory (LSTM) networks are applied to the audio and visual modalities to model the evolution of vocal and facial behaviour over time. The visual pipeline processes frame-level features through a Time Distributed layer to apply consistent transformations across frames, followed by an LSTM layer that encodes temporal dynamics such as micro-expressions and gestural cues across the video sequence.

For the textual modality, although contextual embeddings from BERT provide rich semantic information, these fixed-length vectors are further processed through fully connected layers with dropout to refine latent personality-related features. This approach acknowledges that BERT embeddings already capture bidirectional context, and dense layers are sufficient for learning from the fixed-size text representation.

The outputs of these modality-specific sequential and dense networks are combined to form temporally-informed feature vectors. This strategy preserves critical dynamic information across modalities, enabling the subsequent fusion network to leverage both temporal evolution and multimodal context, thereby improving the precision of personality trait predictions.

E. Multimodal Fusion Strategy

To construct a unified representation of personality traits, features extracted from audio, video, and text modalities are integrated. Each modality encodes complementary behavioural cues—vocal nuances, facial expressions, and linguistic patterns—that together provide a holistic view of the candidate's personality. Following modality-specific processing (LSTM for audio and video, dense layers for text embeddings), the resulting feature vectors are concatenated to form a combined representation. Although the original model does not explicitly implement an attention mechanism, the concatenation serves as a straightforward yet effective fusion strategy, preserving discriminative information across modalities. This fused embedding is subsequently passed through fully connected layers to regress the continuous scores of the Big Five personality traits. The integration of these multimodal features enables the model to leverage diverse signals, improving prediction robustness and reflecting the complex nature of personality assessment.

F. Personality Trait Inference from Multimodal Features

The consolidated multimodal embedding generated from the fusion mechanism forms the basis for predicting the Big Five personality traits. A series of fully connected (dense) layers transforms this fused representation into interpretable, continuous trait scores. These layers apply nonlinear transformations (e.g., ReLU) to model complex interactions among multimodal cues, enabling hierarchical abstraction and detection of subtle behavioral signals linked to traits like Openness and Neuroticism. The final output layer produces five normalized values corresponding to Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, aligned with standard personality measurement scales. The model is trained using a mean absolute error (MAE) loss function, minimizing deviation from ground-truth scores obtained via validated personality assessments. Framing the task as a regression problem preserves sensitivity to personality gradations, supporting nuanced candidate evaluation. This enables seamless integration with recruitment decision-support systems, offering interpretable and quantifiable personality insights.

IV. RESULTS

The multimodal personality recognition model was trained over 20 epochs, demonstrating consistent improvement in both training and validation metrics. At the final epoch, the training loss and mean absolute error (MAE) reached 0.0781, reflecting effective learning from the training data. The corresponding validation loss and MAE were slightly higher at 0.1085, indicating that the model generalized well to unseen data without significant overfitting. On the independent test set, the model achieved a test loss of 0.1105 and a complementary accuracy measure of 1 - MAE equal to 0.8895, further confirming its robustness and predictive capability. These results suggest that the model effectively captures relevant multimodal patterns correlating with the Big Five personality traits, providing reliable personality trait predictions from audio, video, and text inputs.

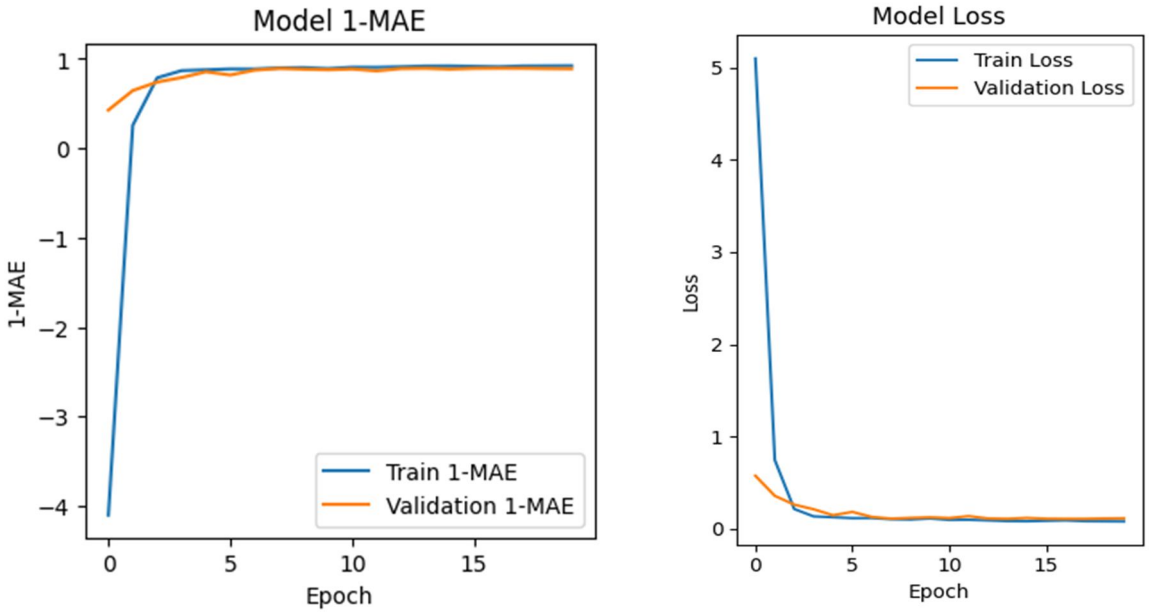


Figure 3 Training vs. Validation 1-MAE and Training vs. Validation Loss

To demonstrate practical applicability, a web-based hiring platform was developed that integrates this model. The platform facilitates asynchronous video interviews, enabling candidates to be assessed automatically on personality traits while allowing recruiters to review the results through an intuitive interface. This application highlights the model's potential for real-world recruitment and personality assessment scenarios.

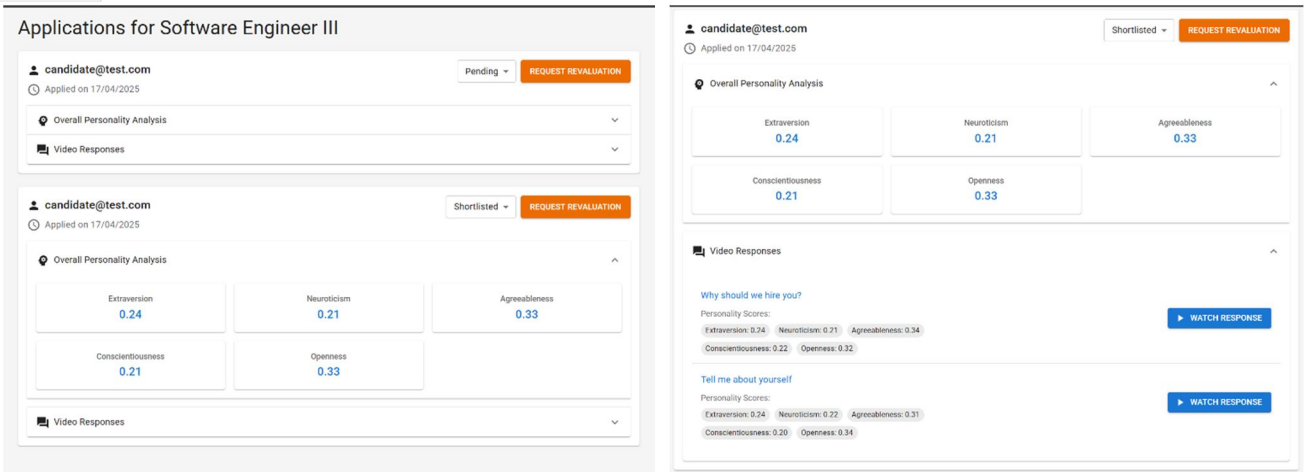


Figure 4 Displayed results on developed web application

V. CONCLUSION

This research marks a significant advancement in recruitment and selection methodologies through the development of an Automated Personality Recognition (APR) system. By leveraging multimodal data—spanning audio, visual, and textual inputs—and employing a hybrid deep learning framework, the proposed system demonstrates strong potential for delivering accurate, scalable, and unbiased personality assessments. The comparative analysis of various models, underlines the importance of balancing predictive accuracy with computational efficiency, particularly in real-world, large-scale applications. This work addresses inherent limitations of traditional assessment methods by reducing human bias and establishing an automated pipeline that aligns with the dynamic needs of modern hiring processes. Moreover, it lays the groundwork for ethical and interpretable AI solutions within recruitment and talent analytics.

Looking ahead, this research opens several promising avenues. One critical future direction is the incorporation of personalized interview feedback, wherein AI-driven insights can help candidates reflect on and improve their communication style, confidence, and engagement. Another is the development of real-time personality prediction systems, enabling live feedback during video interviews to assist recruiters with immediate behavioural insights. Furthermore, integrating APR systems into Applicant Tracking Systems (ATS) could standardize candidate evaluations across organizations, enhancing consistency and fairness in hiring. The framework also supports candidate screening based on personality fit, team building through trait complementarity analysis, and personalized training tailored to individual learning styles. There is also scope to expand trait analysis beyond the Big Five model, incorporating extended personality and behavioural metrics for richer psychological profiling. Finally, embedding principles of Explainable AI (XAI) into personality prediction models will be essential for transparency, helping stakeholders understand and trust model decisions—thereby fostering wider adoption in sensitive decision-making contexts.

Together, these directions pave the way for the evolution of ethical, intelligent, and impactful AI systems that not only enhance recruitment but also redefine how organizations understand and engage with human potential.

REFERENCES

[1] Bounab, Y., Oussalah, M., Arhab, N., Bekhouche, S. (2024). Towards job screening and personality traits estimation from video transcriptions. Expert Systems with Applications, 238(D), 122016. https://doi.org/10.1016/j.eswa.2023.122016.

[2] X. Duan, H. Li, F. Yang, B. Chen, J. Dong, and Y. Wang, "Multimodal Automatic Personality Perception Using ViT, BiLSTM and VGGish," 2024 5th International Conference on Computer Engineering and Application (ICCEA), Hangzhou, China, 2024, pp. 549-553, doi: 10.1109/ICCEA62105.2024.10604109.

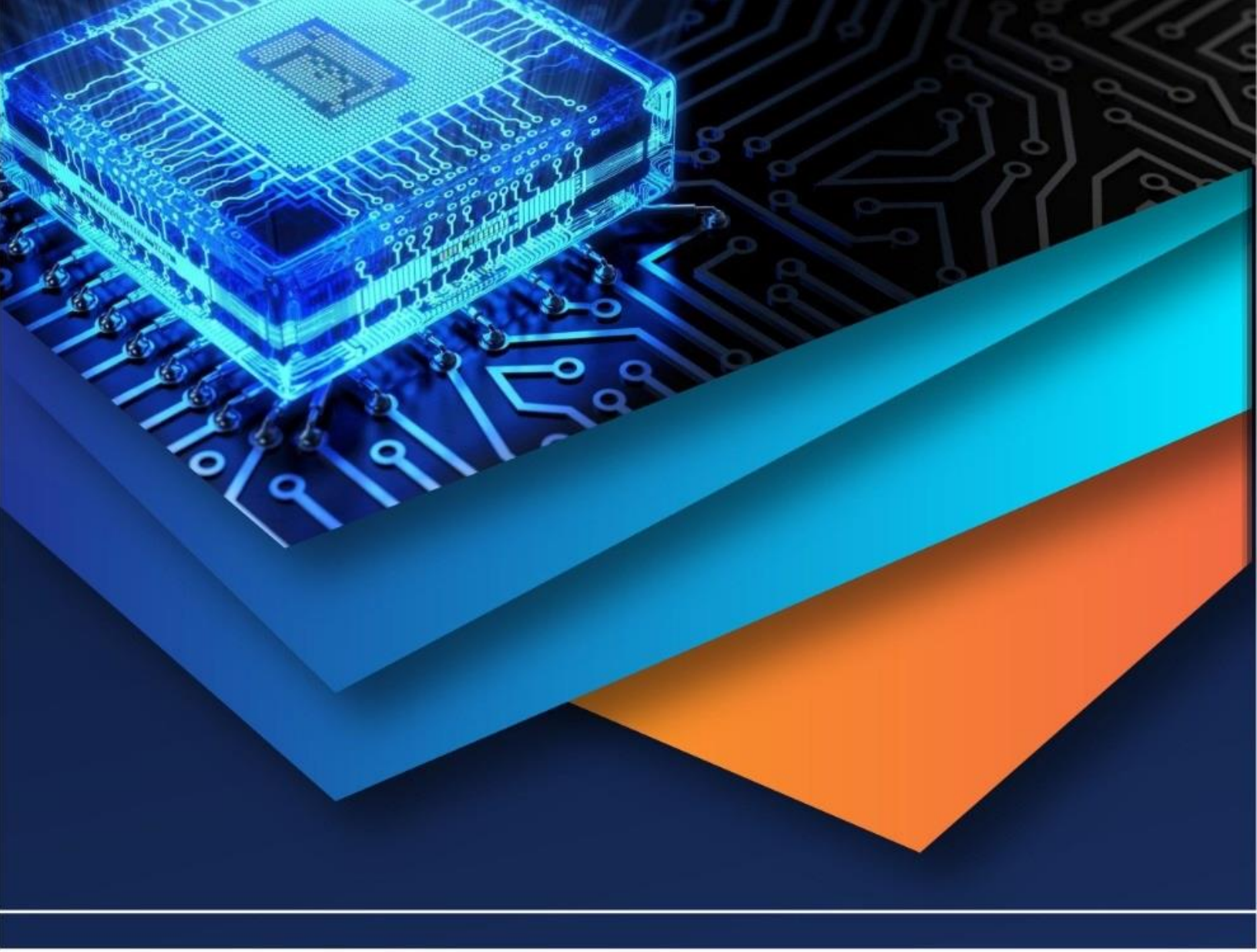
[3] K. A. Dnyaneshwar and G. Poonam, "AI-Driven Insights: Personality Evaluation in Asynchronous Video Interviews for Informed Hiring Decisions," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/ICITEICS61368.2024.10625601.

[4] O. T.-C. Chen, C.-H. Tsai, and M.-H. Ha, "Automatic Personality Recognition via XLNet with Refined Highway and Switching Module for Chatbot," 2024 IEEE International Symposium on Circuits and Systems (ISCAS), Singapore, Singapore, 2024, pp. 1-5, doi: 10.1109/ISCAS58744.2024.10558116.

[5] S. Ghassemi et al., "Unsupervised Multimodal Learning for Dependency-Free Personality Recognition," IEEE Transactions on Affective Computing, vol. 15, no. 03, pp. 1053-1066, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3318367.

[6] Zatarain Cabada, Ramón Barron Estrada, Maria Bátiz Beltrán, Víctor Sapien, Ramón Ruiz, Gerardo. (2023). Sentiment Analysis of Spanish Text for Automatic Personality Recognition in Intelligent Learning Environments. pp. 1-4. doi: 10.1109/ENC60556.2023.10508699.

[7] D. Nagajyothi, S. A. Ali, P. H. Sree, and P. Chinthapalli, "Automatic Personality Recognition In Interviews Using CNN," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-7, doi: 10.1109/GCAT59970.2023.10353423.
[8] Holtrop, Djurre, Oostrom, Janneke, Breda, Ward, Koutsoumpis, Antonis, and de Vries, Reinout. (2022). Exploring the application of a text-to-personality technique in job interviews. European Journal of Work and Organizational Psychology, 31, 1-18. doi: 10.1080/1359432X.2022.2051484.
[9] J. R. Lima, H. J. Escalante, and L. V. Pineda, "Sequential Models for Automatic Personality Recognition from Multimodal Information in Social Interactions," 2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2022, pp. 1-6, doi: 10.1109/ROPEC55836.2022.10018711.
[10] X. Duan, Y. Yu, Y. Du, H. Liu, and Y. Wang, "Personality Recognition Method Based on Facial Appearance," 2022 3rd International Conference on Computer Vision, Image and Deep Learning (CVIDL ICCEA), Changchun, China, 2022, pp. 710-715, doi: 10.1109/CVIDLICCEA56201.2022.9824658.
[11] X. Duan, Q. Zhan, S. Zhan, Y. Yu, L. Chang, and Y. Wang, "Multimodal Apparent Personality Traits Analysis of Short Video Using Swin Transformer and Bi-directional Long Short-Term Memory Network," 2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC), Qingdao, China, 2022, pp. 1003-1008, doi: 10.1109/ICFTIC57696.2022.10075178.
[12] Camati, Ricardo Enembreck, Fabrício. (2020). Text-Based Automatic Personality Recognition: a Projective Approach. pp. 218-225. doi: 10.1109/SMC42975.2020.9282859.
[13] Z. Su, Z. Lin, J. Ai, and H. Li, "Rating Prediction in Recommender Systems Based on User Behavior Probability and Complex Network Modeling," IEEE Access, vol. 9, pp. 30739-30749, 2021, doi: 10.1109/ACCESS.2021.3060016.
[14] K.Yesu, K., Shandilya, S., Rekharaj, N., Ankit, K., and Sairam, P. S. (2021). Big Five Personality Traits Inference from Five Facial Shapes Using CNN. IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, pp. 1-6. doi: 10.1109/GUCON50781.2021.9573895.
[15] Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing. doi: 10.48550/arXiv.1908.10084.
[16] L´eo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chlo´e Clavel. 2019. HireNet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19). AAAI Press, Article 71, 573–581.https://doi.org/10.1609/aaai.v33i01.3301573.
[17] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 3030-3043, Oct. 2018, doi: 10.1109/TCSVT.2017.2719043.
[18] J. Gorbova, E. Avots, I. Lüsi, M. Fishel, S. Escalera, and G. Anbarjafari, "Integrating Vision and Language for First-Impression Personality Analysis," IEEE MultiMedia, vol. 25, no. 2, pp. 24-33, Apr.-Jun. 2018, doi: 10.1109/MMUL.2018.023121162.
[19] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, pp. 59-66, doi: 10.1109/FG.2018.00019.
[20] L. Chen, R. Zhao, C. W. Leong, B. Lehman, G. Feng, and M. E. Hoque, "Automated video interview judgment on a large-sized corpus collected online," 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 2017, pp. 504-509, doi: 10.1109/ACII.2017.8273646.
[21] Chen, Lei, Feng, Gary, Leong, Chee Wee, Lehman, Blair, Martin-Raugh, Michelle, Kell, Harrison, Lee, Chong Min, Yoon, Su-Youn. (2016). Automated Scoring of Interview Videos using Doc2Vec Multimodal Feature Extraction Paradigm. doi: 10.1145/2993148.2993203.
[22] Schroff, Florian, Kalenichenko, Dmitry, Philbin, James. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. Proceedings of CVPR.
[23] Biel, J., Teijeiro-Mosquera, L., & Gatica-Pérez, D. (2012). FaceTube: predicting personality from facial expressions of emotion in online conversational video. International Conference on Multimodal Interaction.
[24] Batrinca, Ligia Maria, Mana, Nadia, Lepri, Bruno, Pianesi, Fabio, and Sebe, Nicu. (2011). Please, tell me about yourself: automatic personality assessment using short self-presentations. Proceedings of the 13th International Conference on Multimodal Interfaces. doi: 10.1145/2070481.2070528.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY