ORIGINAL ARTICLE

Personality in 3D: multimodal deep learning framework for big five trait prediction

Devraj Patel · Sunita V. Dhavale · Bhushan B. Mhetre

Received: 29 July 2025 / Accepted: 3 February 2026

Abstract
Automatic Personality Prediction (APP) is a key area in affective computing, aiming to infer human personality traits from behavioural cues. This study proposes a multimodal deep learning framework for predicting Big Five personality traits using text, audio, and video data. We employ modality-specific and multimodal datasets annotated for the Big Five traits and explore a range of ardy proposes a multimodal deep learning framework for predicting Big Five personality traits using text, audio, and video data. We employ modality-specific and multimodal datasets annotated for the Big Five traits and explore a range of architectures, including transformers, CNNs, and recurrent models (LSTM, CRNN). Results show that audio features, especially MFCC-1 and MFCC-2, offer high predictive power, while sentiment-aware textual embeddings enhance linguistic modelling. Visual features capture non-verbal cues vital for comprehensive trait assessment. We implement both early and late fusion strategies to integrate affective, linguistic, and visual signals, improving robustness and generalisation. To ensure transparency, we incorporate Explainable AI (XAI) techniques, including SHAP and Grad-CAM, to identify influential features across modalities. This enables human-centred analysis and builds trust in model predictions. Our findings highlight the effectiveness of deep multimodal learning for personality modelling and demonstrate how combining behavioural signals with interpretability tools leads to more adaptive and transparent personality-aware AI systems.

Keywords Big Five · Deep Learning · Explainable AI (XAI) · Multimodal Fusion · Personality 3D

1 Introduction

Human personality plays a pivotal role in shaping individual behaviours, communication styles, and social interactions. As artificial intelligence systems become increasingly integrated into daily life, the ability to automatically infer personality traits from behavioural signals–referred to as Automatic Personality Prediction (APP)–has garnered growing interest across affective computing, human-computer interaction, and social robotics.

Among the various psychological models, the Big Five Personality Traits – Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism–offer a robust, widely accepted framework for personality assessment [1]. In parallel, the rapid growth of multimodal digital data, text from social media, audio from speech,

These authors contributed equally to this work: Sunita V.Dhavale and Bhushan B.Mhetre.

Published online: 24 March 2026

and facial cues from videos has opened up new opportunities to assess personality in a holistic and data-driven manner.

This paper presents a comprehensive multimodal deep learning framework for predicting Big Five traits, utilising three complementary modalities: text, audio, and video. By designing modality-specific pipelines and multimodal fusion strategies, we aim to capture the full spectrum of behavioural signals relevant to personality assessment. We employ a suite of deep learning architectures, including transformer-based models for text, convolutional and recurrent models for audio and video, and fusion networks that integrate affective, linguistic, and visual cues.

We structure our investigation around the following research questions:

- How do different deep learning architectures perform across individual modalities for Big Five personality prediction?
- Which modality or combination of modalities provides the most robust and generalizable performance?
- How do early and late fusion strategies impact the overall accuracy and stability of personality prediction?
- Can explainability techniques (e.g., SHAP, Grad-CAM) reveal modality-specific contributions and model decision patterns?

The key contributions of this work include:

- A unified, end-to-end deep learning framework for Big Five trait prediction across text, audio, and video modalities.
- A comparative analysis of modality-specific models, highlighting performance trends and strengths of each modality.
- Exploration of multimodal fusion techniques, demonstrating improved prediction accuracy and robustness.
- Integration of Explainable AI (XAI) tools to enhance the interpretability of model predictions across all three modalities.

By modelling personality "in 3D," this study not only advances the state of the art in multimodal APP but also lays the foundation for the development of transparent, adaptive, and human-centred personality-aware AI systems applicable to personalised education, mental health diagnostics, virtual assistants, and beyond.

2 Related work

Personality computing has evolved at the intersection of psychology, machine learning, and multimodal signal processing. This section first explores key psychological theories of personality that underpin computational models of personality. We then review recent advancements in deep learning techniques applied to text, audio, and visual modalities for Automatic Personality Prediction (APP). Subsequent subsections examine strategies for multimodal fusion, the integration of explainable AI (XAI) to enhance model transparency, and identify current gaps that motivate our proposed framework.

2.1 Personality psychology theories

Understanding personality structure has been a central goal in psychology for over a century, resulting in the development of multiple theoretical frameworks. These models form the backbone of APP systems by offering trait definitions, measurement constructs, and psychological grounding.

Big five (OCEAN) model: The Five-Factor Model (FFM), also known as the Big Five, is the most widely accepted trait-based model in personality psychology [1]. It defines five core personality dimensions: Openness (creativity and curiosity), Conscientiousness (organisation and discipline), Extraversion (sociability and energy), Agreeableness (compassion and cooperation), and Neuroticism (emotional instability). The model has

demonstrated strong empirical validity across diverse cultures and populations [2]. Due to its dimensional and continuous nature, the Big Five is especially well-suited to machine learning applications in APP, including predicting traits from social media data [3], video interviews [4], resumes [5], and audio streams [6].

Myers-Briggs type indicator (MBTI): The MBTI classifies individuals into 16 personality types based on four dichotomous dimensions: Introversion–Extraversion, Sensing–Intuition, Thinking–Feeling, and Judging–Perceiving [7]. Despite criticisms regarding psychometric reliability and validity [8], MBTI remains popular in career guidance, team dynamics, and NLP-based classification tasks [9, 10]. Several APP studies have applied deep learning techniques to classify MBTI types using blog posts, tweets, and essays, revealing language-use patterns aligned with MBTI types.

PEN model (Eysenck's Three-Factor Model): Proposed by Hans Eysenck, the PEN model identifies three biological dimensions of personality: Psychoticism, Extraversion, and Neuroticism [11]. Though less granular than the Big Five, PEN has been applied in early computational psychology research and remains relevant in neuropsychological and behavioural biometrics studies [12]. In APP, its focus on emotional and impulsive behaviour has inspired acoustic and EEG-based personality estimation systems.

16PF (Sixteen personality factor questionnaire): Developed by Raymond Cattell, 16PF assesses 16 primary factors derived through factor analysis [13]. While it offers a comprehensive understanding of personality, the high dimensionality and complexity of the assessment make it less practical for APP. However, it has influenced feature design in early machine learning models for psychological profiling and occupational screening.

DISC model: DISC theory, originally introduced by William Marston, categorises behaviour into four types: Dominance, Influence, Steadiness, and Conscientiousness [14, 15]. DISC is widely used in leadership development and organisational behaviour applications. In computational contexts, DISC has been applied to personality-adaptive chatbot design and job screening systems, although less frequently than trait-based models.

HEXACO model: The HEXACO model extends the Big Five by adding a sixth factor, Honesty-Humility, designed to capture ethical and prosocial behavior [16]. HEXACO has demonstrated superior predictive power in contexts involving trust, integrity, and cooperation [17]. While not yet widely adopted in APP due to the limited availability of labelled datasets, its inclusion is growing in moral computing and recommender systems.

Applications across models: While each model has distinct theoretical underpinnings, their application in AI varies depending on the use case. Recent work by Pletzer and Abrahams [18] provides a comprehensive review of how personality traits, particularly those defined by trait-based models like the Big Five, influence job performance across various occupational roles. Their study in Current Opinion in Psychology highlights the predictive value of traits such as Conscientiousness and Emotional Stability (the inverse of Neuroticism) in determining task performance, teamwork, and leadership potential. The paper also discusses emerging trends, including the integration of AI-based personality assessments in hiring and development contexts. This reinforces the practical significance of accurate and interpretable personality prediction models, like the one proposed in our work, especially when applied to domains involving human evaluation and decision-making.

The MBTI has been widely adopted in classification tasks and personality-driven dialogue systems. The Big Five and HEXACO are more suitable for regression tasks and multimodal modelling due to their continuous structure. DISC and PEN models have niche relevance in emotion-aware computing and behavioural biometrics. These models inform the design of APP systems and guide the construction of datasets and the development of evaluation metrics. The Big Five, in particular, remains dominant in computational research due to its strong psychometric foundation and widespread use in benchmark datasets, such as ChaLearn First Impressions [4] and Essays.csv [19].

2.2 Multimodal deep learning for personality prediction

Recent years have seen rapid advances in deep learning models that infer personality traits from text, audio, and video, collectively advancing the field of affective and social signal computing. This multimodal direction is motivated by the fact that personality is expressed not only through language but also through vocal prosody,

facial micro-expressions, and behavioural dynamics. Deep learning enables end-to-end modelling of such complex signals, and this section outlines the current state of research across modalities.

Text modality: Text remains one of the most extensively studied modalities for personality inference. Foundational work by Pennebaker and King [19] highlighted the correlation between linguistic style and personality traits, providing early evidence for feature-based personality detection. Subsequent efforts integrated NLP techniques such as part-of-speech tagging, stemming, n-grams, and parse trees to infer personality from authorial writing styles, as demonstrated in [20] using the Big Five Factor Model.

With the advent of deep learning, transformer-based architectures like BERT, RoBERTa, and XLNet have been fine-tuned to capture contextual and affective semantics from text [9]. For instance, Ren et al.[21] proposed a sentiment-aware BERT-based model for multi-label personality detection, significantly enhancing performance by incorporating affective cues. Similarly, Sun et al.[10] introduced a CNN-LSTM hybrid architecture for MBTI classification using social media posts, achieving an average F1-score of 0.82. In another study, Konakalla et al. [22] employed RCNN and MLP classifiers on tweets from the Russo-Ukrainian conflict, achieving up to 85% accuracy for Neuroticism and Agreeableness traits.

The predictive utility of structured professional text was also demonstrated by Agarwal et al. [5], who applied logistic regression to resume datasets, achieving approximately 76.9% accuracy in personality classification. These efforts have been further supported by standardised datasets, such as IPIP[23–26], which enable supervised model training across diverse personality tasks.

Despite these advances, single-modality models often lack nuanced affective context. To address this, recent methods integrate sentiment-aware embeddings and hybrid attention mechanisms [27], thereby improving alignment with psycholinguistic theory and enhancing the robustness of trait-level prediction.

Audio modality: Speech encodes prosodic, phonetic, and emotional characteristics that are closely related to traits like Extraversion and Neuroticism. Gokul and Lalitha [6] applied auditory nerve modelling to extract psychoacoustic features, achieving classification accuracy up to 78% on selected Big Five traits using SVMs. Shilpa et al. [28] proposed a DNN-based system to assess human stress levels via speech, which also correlates with emotional instability.

More recent work [29] employed MFCC (Mel-Frequency Cepstral Coefficient) features combined with a CRNN model, achieving F1-scores exceeding 0.84 for Extraversion and Conscientiousness. In [30], transformer encoders were integrated with paralinguistic cues, such as pitch, formants, and speaking rate, to improve interpretability. These models, however, still face challenges such as sensitivity to environmental noise and the need for robust pre-processing pipelines.

Visual modality: Facial expressions, head pose, gaze direction, and microgestures provide powerful non-verbal cues for inferring personality. The ChaLearn First Impressions Challenge [4] established a benchmark dataset of short interview clips annotated for Big Five traits. Participants used 3D CNNs, LSTMs, and fusion networks, with the best models achieving up to 91% accuracy in trait prediction.

In [31], a CLIP-based transformer was fine-tuned on personality-labelled video datasets, demonstrating state-of-the-art performance in predicting Openness and Agreeableness, but struggled with Neuroticism due to the subtlety of visual cues. Wang et al. [32] introduced a self-supervised visual personality encoder using temporal attention, achieving Pearson correlations above 0.75 for multiple traits. In [33], the author proposed a video-based personality recognition framework that leverages a temporal emotion-based LSTM model to predict apparent first impression personality traits. The model extracts sequential emotion cues from video frames and integrates them into a single-modality architecture, demonstrating that dynamic emotional transitions play a crucial role in inferring Big Five traits.

While visual networks show promise, they often require large-scale annotated data and are computationally expensive, limiting real-time deployment. Additionally, generalising across cultural and lighting conditions remains a challenge.

Multimodal insights: In the multimodal domain, combining text, audio, and video has consistently enhanced prediction performance. Gucluturk et al.[34] proposed a deep residual network for apparent personality inference

using visual and auditory streams. Escalante et al. [35] established multimodal fusion baselines, highlighting the complementary cues across modalities. CR-Net [36] presented a classification-regression network with modality-specific branches and fusion layers tailored to Big Five trait prediction. Santhosh et al. [37] extended this work by introducing a deep cross-modal hierarchical attention model, achieving competitive performance on benchmark datasets.

Mishra and Sagnika [38] surveyed multimodal approaches and reported a 10–15% improvement in accuracy over unimodal baselines, emphasising the benefits of combining verbal and non-verbal cues. Patil et al. [39] demonstrated that late fusion of CV/resume text and facial videos led to more robust trait predictions than early concatenation. These studies collectively affirm that integrating multimodal cues yields a richer understanding of personality, particularly for complex traits like Conscientiousness and Neuroticism.

Despite these advances, several research gaps remain:

- Lack of unified architectures tested across all three modalities under consistent evaluation protocols.
- Inadequate integration of explainability tools in personality prediction pipelines.
- Limited exploration of cross-modality trait transfer (e.g., predicting visual traits from audio).
- Sparse availability of high-quality, annotated multimodal datasets aligned with personality theories.

These gaps motivate our work, which proposes a 3D personality modelling framework using modality-specific encoders, fusion networks, and Explainable AI tools to holistically and transparently predict Big Five traits from behavioural signals.

2.3 Fusion strategies and 3D personality modelling

Incorporating multiple modalities: text, audio, and video, offers a more comprehensive understanding of personality, often referred to as "3D personality modelling." Each modality captures unique behavioural signals: linguistic cues from text, prosodic and tonal variations from audio, and micro-expressions from visual input. Fusion strategies aim to unify these heterogeneous signals into a single predictive model.

Broadly, fusion techniques are categorised as follows:

- *Early fusion*: Combines raw or intermediate features from each modality before feeding them into a shared model. This approach enables the learning of cross-modal dependencies and synergistic representations. For instance, concatenating MFCC embeddings, visual frame encodings, and contextual textual embeddings before classification allows the model to discover interdependencies that mirror human multimodal perception.
- *Late fusion*: Aggregates outputs or confidence scores from independently trained modality-specific models. This strategy often proves more robust in noisy environments or when certain modalities are missing or corrupted, as the final prediction benefits from redundancy across channels.

Studies such as Zhang et al. [40] and Wu et al. [41] compared these fusion techniques in personality and emotion recognition tasks. They observed that while early fusion enhances performance in clean, well-aligned datasets, late fusion offers superior generalizability in real-world scenarios with asynchronous or incomplete data streams. Notably, Zhou et al. [42] demonstrated that hybrid strategies–incorporating both early and late fusion elements–can further optimise performance by balancing cross-modal learning with decision-level robustness.

Recent works have also explored adaptive and attention-based fusion mechanisms. Bao et al.[43] proposed an Adaptive Information Fusion Network (AIFN) that dynamically integrates multi-modal features via learned attention weights, showing strong performance gains across diverse user profiles. Similarly, Wang et al.[44] introduced an emotion-guided fusion model with contrastive learning objectives, effectively aligning cross-modal embeddings while preserving trait-specific distinctions. Waqas et al. [45] employed a data fusion approach utilising a BERT-GCN hybrid, demonstrating that structural alignment between modalities improves both interpretability and classification accuracy.

Cognitive load-aware fusion mechanisms are also emerging. In [46], an attention-based fusion mechanism was introduced that dynamically weights modalities based on task relevance and contextual salience. Their system, evaluated on a multimodal personality corpus, achieved state-of-the-art accuracy while reducing computational overhead.

Beyond accuracy, 3D personality modelling facilitates more context-aware and fair predictions. Foundational surveys by Vinciarelli and Mohammadi [47] and Ilmini and Fernando [48] provide a historical and contemporary perspective on apparent personality detection, highlighting the evolution from handcrafted to deep learning-based methods, as well as the growing emphasis on explainability and modality-specific inference. Recent reviews by Mehta et al. [49], Zhao et al. [50], and Mostafa et al. [51] further underscore that multimodal integration enhances trait resilience, fairness, and adaptability. However, these studies consistently point to unresolved challenges, including the scarcity of fully annotated, large-scale multimodal datasets and the absence of standardised fusion protocols across personality prediction benchmarks.

To address these gaps, our framework systematically compares early, late, and hybrid fusion strategies using both hand-crafted features (e.g., MFCCs, sentiment scores) and deep embeddings (e.g., BERT, 3D CNN). This design enables robust identification of trait–modality pairings and allows us to evaluate whether certain traits (e.g., Extraversion) are more accurately predicted through individual modalities or fused representations.

2.4 Explainable AI for personality prediction

Interpretability is crucial in personality-aware AI systems, especially given the psychological and ethical implications associated with profiling human traits. Whether applied in educational personalisation, recruitment decisions, or clinical settings, stakeholders must understand not only what personality trait a system predicts, but also why it makes that decision. To this end, our framework integrates a suite of Explainable AI (XAI) techniques, focusing on both local (instance-specific) and modality-specific interpretability.

We focus on two widely adopted and complementary explainability methods: SHAP (SHapley Additive Explanations) and Grad-CAM (Gradient-weighted Class Activation Mapping), selected for their effectiveness across structured (text/audio features) and unstructured (images/video frames) data, respectively. These choices were made after evaluating several alternatives, balancing explainability fidelity, modality compatibility, and computational cost.

SHAP (SHapley additive explanations): SHAP provides feature-level attribution using cooperative game theory, quantifying the contribution of each feature to a model's prediction. It is particularly effective in identifying token-, frequency-, or pixel-level importance in transformer-based models or audio spectrograms. Liu et al. [52] used SHAP to reveal that facial expressions like smile duration and gaze strongly influenced Extraversion predictions, while head motion variance impacted Openness. Similarly, Mishra [53] emphasised SHAP's utility in trust-sensitive AI, advocating for clustering and summarisation to handle high-dimensional personality data.

In our framework, SHAP is employed across all three modalities–highlighting emotionally charged or socially significant words in text (e.g., "rude", "honesty"), emphasising spectral regions in MFCCs for speech, and identifying facial landmarks in video frames. SHAP's model-agnostic and locally faithful explanations make it especially suitable for complex, multimodal deep learning models like ours. However, SHAP's computational cost can be high for large models and datasets, and it assumes feature independence, which may not always hold in sequential data.

Grad-CAM and class activation mapping variants: Originally proposed by Zhou et al. [54] and extended into Grad-CAM [55], this family of techniques generates class-specific heatmaps by computing gradients of the output with respect to intermediate convolutional layers. Grad-CAM highlights spatial regions in images (or spectrograms) that contribute most to a prediction. This method has become a standard for CNN-based interpretability due to its ease of integration and visual intuitiveness.

In our model, Grad-CAM is applied to the visual modalities. The CAM maps allow us to understand which facial regions (e.g., smiles, eyes, head tilt) or which time-frequency patches in speech were most influential in predicting personality traits. The fidelity of these explanations is measured using metrics like:

- *Drop in Confidence:* Expected drop in prediction confidence when the saliency region is occluded (smaller is better).
- *Increase in Confidence:* Number of cases where explanation enhances prediction confidence (larger is better).

To further enhance the fidelity of saliency maps, we also explored advanced CAM variants, including Grad-CAM++, EigenGrad-CAM, Score-CAM, Ablation-CAM, and Random-CAM. Each of these techniques offers unique advantages: Grad-CAM++ handles multiple object occurrences [56]; EigenGrad-CAM improves spatial localisation by focusing on eigenvectors; Score-CAM is gradient-free and uses class scores for weighting; Ablation-CAM ablates channels to measure contribution; and Random-CAM serves as a control baseline. The 2020 CVPR Workshop paper [57] applied Score-CAM for personality prediction and demonstrated that such saliency techniques can effectively reveal psychologically meaningful features (e.g., eye contact, posture) that correspond to human-perceived traits.

Although many CAM variants exist, we prioritised standard Grad-CAM for our framework due to its interpretability, computational efficiency, and compatibility with both video and audio spectrogram inputs. We also validated that it produced coherent saliency maps aligned with domain expectations. However, future extensions may incorporate Score-CAM or Ablation-CAM for improved localisation and robustness in temporal models.

Broader trends in multimodal explainability: Explainable AI (XAI) is increasingly influencing domains such as affective computing and personality prediction, moving beyond traditional tools like SHAP and CAM. Haz et al.[27] leveraged SHAP to uncover psycholinguistic indicators of narcissism, illustrating the potential of XAI for trait-specific insights. Similarly,[58] proposed an attention-based interpretability framework for emotion detection in text, aligning linguistic structure with psychological reasoning. Recent studies [59, 60] further advance this trend by integrating explainability directly into multimodal emotion recognition systems, utilising situational knowledge and reasoning mechanisms to enhance both model performance and human trust in feedback-driven contexts.

However, many existing models incorporate XAI only in post-hoc visualisations, lacking formal integration into the system architecture. In contrast, our approach embeds explainability into the core of both training and evaluation. Each modality is paired with dedicated interpretability tools, ensuring predictions are not only traceable and psychologically grounded but also transparent and accountable–qualities critical for responsible personality recognition.

3 Methodology framework

An overview of the proposed multimodal personality prediction framework is illustrated in Fig. 1. The framework is designed to leverage complementary information from visual, audio, and textual modalities extracted from video-based inputs. Each modality undergoes modality-specific preprocessing and is subsequently passed through a dedicated subnetwork for high-level feature representation.

For the visual modality, representative image frames are extracted from the input video and processed through a Residual-Dilated-Separable Convolutional (RDSC) based visual subnetwork. The audio modality involves extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the raw audio stream, which are then passed to an audio subnetwork to capture relevant temporal and spectral features. The textual modality, comprising speech transcripts, is preprocessed through standard NLP techniques, including tokenisation and embedding, before being processed by a transcript subnetwork. Outputs from these modality-specific subnetworks are then integrated using fusion strategies. We explore early fusion, late fusion, and attention-based fusion mechanisms

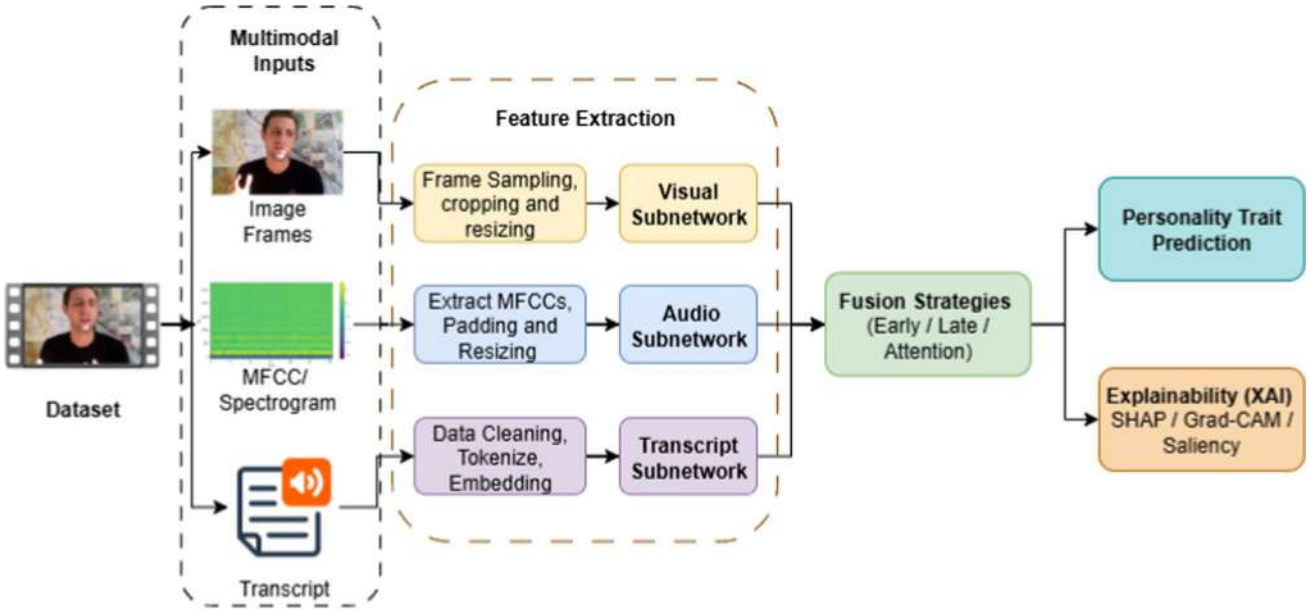

Fig. 1 Overall system architecture for 3D multimodal personality recognition

to effectively combine multimodal representations. The fused feature vector is subsequently passed through fully connected layers for final personality trait classification.

In addition to prediction, the framework incorporates explainability mechanisms based on eXplainable AI (XAI) techniques such as SHAP, Grad-CAM, and saliency maps. These tools offer interpretability at both the feature and modality levels, thereby increasing transparency and providing insights into the decision-making process of the model.

3.1 Data collection and preprocessing

This study utilises the ChaLearn first impressions V2 dataset [4], a widely recognised benchmark for automatic personality perception. The dataset comprises approximately 10,000 short video clips (each lasting around 15 s), wherein individuals speak directly to the camera. Each video is annotated with continuous scores for the Big Five personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness, ranging from 0 to 1. The dataset provides three synchronised modalities: raw audio, video frames, and corresponding transcripts, thus enabling a rich and multimodal analysis of behavioural signals.

3.1.1 Audio preprocessing

Audio tracks are extracted from each video using the FFmpeg backend and converted to mono. To ensure consistent sampling conditions and suppress potential aliasing artefacts above the human-audible range, the raw waveform $x(t)$ is first low-pass filtered at 20 kHz using an 8^{th}-order Butterworth filter and then resampled to 44.1 kHz:

$$x_f(t) = \text{LPF}_{20\text{kHz}}(x(t)), \tag{1}$$

$$x_r(t) = \text{Resample}(x_f(t), 44{,}100\ \text{Hz}). \tag{2}$$

The filtered and resampled signal $x_r(t)$ is then transformed using the Short-Time Fourier Transform (STFT). A Hann window of 1024 samples with a hop length of 512 samples is used, providing a balanced trade-off between temporal and frequency resolution.

From the STFT magnitude spectra, Mel-Frequency Cepstral Coefficients (MFCCs) are computed using 24 triangular Mel filterbanks:

$$\mathrm{MFCC}_{i,j} = \mathrm{DCT}\left(\log\left(\sum_{k=1}^{K} |X_j[k]|^2 \cdot H_i[k]\right)\right), \tag{3}$$

where $X_j[k]$ denotes the STFT magnitude of frame j, and $H_i[k]$ is the i-th Mel filter.

The resulting MFCC matrix is zero-padded to a fixed temporal width of 1319 frames:

$$\mathrm{MFCC}_p = \mathrm{Pad}(\mathrm{MFCC}, [24 \times 1319]). \tag{4}$$

Coefficient-wise z-score normalisation is then applied:

$$\mathrm{MFCC}_z(i,j) = \frac{\mathrm{MFCC}_p(i,j) - \mu_i}{\sigma_i}, \tag{5}$$

where μ_i and σ_i denote the mean and standard deviation of the i-th MFCC dimension computed over the training set.

Finally, the normalised MFCC tensor is reshaped into a 3D array suitable for convolutional processing:

$$A \in \mathbb{R}^{24\times1319\times1}. \tag{6}$$

3.1.2 STFT window size ablation

We conducted an ablation study to examine the sensitivity of the audio feature extraction stage to the choice of STFT window size. Four configurations were evaluated: $n_{\mathrm{fft}} \in \{256, 512, 1024, 2048\}$. For each setting, Mel-Frequency Cepstral Coefficients (MFCCs) were computed following the preprocessing pipeline described in Section 3.1. To quantify feature stability, we computed the mean cosine similarity between the MFCC representations obtained with a given window size and those obtained using the default $n_{\mathrm{fft}} = 1024$.

Table 1 reports the results. As expected, the 1024-point STFT serves as a reliable baseline ($\mathrm{similarity} = 1.0$). Smaller windows (256 and 512) introduce slightly higher temporal resolution at the expense of spectral resolution, whereas the largest window (2048) provides finer frequency detail but reduced temporal precision. In all cases, the similarity remains high (> 0.95), demonstrating that the proposed audio preprocessing pipeline is robust to moderate variations in STFT parameters and that downstream model predictions are unlikely to be significantly affected.

These results empirically validate our choice of a 1024-point STFT window as a balanced configuration, while also confirming the robustness of our MFCC-based audio representation.

3.1.3 Visual preprocessing

To capture visual and non-verbal cues, six random frames are extracted from each video. Each frame undergoes a series of transformations: colour space conversion from BGR to RGB, resizing to 248×140 pixels, random cropping to 128×128, and normalisation of pixel values to the $[0, 1]$ range.

Let V denote a video with F frames. The preprocessing steps are as follows:

$$\{f_k\}_{k=1}^{6} \sim \mathrm{UniformSample}(V, 6) \tag{7}$$

Table 1 STFT ablation results measured via mean cosine similarity with respect to the baseline configuration ($n_{\mathrm{fft}} = 1024$)

Window size	Window duration (ms)	Mean cosine similarity
256	5.8	0.9524
512	11.6	0.9850
1024	23.2	1.0000
2048	46.4	0.9968

$$f_k^{\mathrm{RGB}} = \mathrm{BGR2RGB}(f_k) \tag{8}$$

$$f_k^{\mathrm{resized}} = \mathrm{Resize}(f_k^{\mathrm{RGB}}, 248 \times 140) \tag{9}$$

$$f_k^{\mathrm{crop}} = \mathrm{Crop}(f_k^{\mathrm{resized}}, 128 \times 128) \tag{10}$$

$$f_k^{\mathrm{norm}}(i,j,c) = \frac{f_k^{\mathrm{crop}}(i,j,c)}{255} \tag{11}$$

$$V' \in \mathbb{R}^{6\times 128\times 128\times 3} \tag{12}$$

Although only a limited number of frames are sampled from each video, temporal dynamics are explicitly modelled at the feature level. Frame-wise embeddings are treated as a temporal sequence and processed using a multi-head self-attention module, allowing the network to capture inter-frame dependencies and temporal facial dynamics relevant to personality traits.

3.1.4 Text preprocessing

The transcripts corresponding to each video are stored in a dictionary keyed by video ID. These raw text strings are tokenised using a pretrained transformer tokeniser (e.g., BERT), which handles punctuation, casing, and subword units internally. Minimal preprocessing is applied to retain contextual richness.

Let T denote the transcript string. The following operations are performed:

$$\mathrm{tok} = \mathcal{T}(T) = \{w_1, w_2, \ldots, w_n\} \tag{13}$$

$$E = [\mathrm{Embed}(w_1), \mathrm{Embed}(w_2), \ldots, \mathrm{Embed}(w_n)] \in \mathbb{R}^{n\times d} \tag{14}$$

where d is the embedding dimension (typically $d = 768$ for BERT-base).

3.1.5 Ground truth annotation

Trait scores are loaded from pre-defined dictionaries for each video. For every sample, a 5-dimensional vector is extracted:

$$Y \in \mathbb{R}^{5\times 1} \tag{15}$$

corresponding to the Big Five traits: {EXT, AGR, CON, NEU, OPN}.

3.1.6 Dataset assembly

A multi-threaded pipeline is implemented to preprocess all videos in parallel. For each video instance, the audio tensor A, video tensor V', token embedding E, and label vector Y are aggregated into a structured tuple:

$$\mathcal{D} = \{(A, V', E, Y)_i\}_{i=1}^{N} \tag{16}$$

The processed dataset $\mathcal{D}$ is serialised into binary '.pkl' files for efficient disk loading during model training and evaluation.

This structured preprocessing pipeline ensures temporal alignment and modality consistency, facilitating robust training and evaluation of both unimodal and multimodal personality prediction models.

3.2 Model architectures

To capture the diversity of behavioural signals exhibited through spoken language, vocal prosody, and facial expressions, we design a modular deep learning architecture consisting of three dedicated subnetworks; each responsible for a specific modality: text, audio, or video. These subnetworks extract modality-specific embeddings, which are then fused and passed through a shared regression head to predict the Big Five trait scores. The overall design supports flexible inclusion or exclusion of input modalities and is optimised using the Mean Squared Error (MSE) loss. To formally describe the system design, we provide pseudo-algorithms detailing the core operations of each subnetwork and the fusion step.

Computational efficiency was a key design consideration in the proposed multimodal framework. Instead of employing computationally expensive 3D convolutional networks for video processing, we adopt residual dilated separable convolutions combined with sparse frame sampling and attention-based temporal modelling. This design significantly reduces the number of parameters and floating-point operations while preserving temporal sensitivity. Similarly, modality-specific subnetworks operate independently, allowing for selective activation based on the availability of input modalities. These architectural choices allow the framework to scale efficiently without compromising prediction accuracy.

3.2.1 Transcription subnetwork: transcriptionSubNet

To comprehensively model behavioural patterns encoded in linguistic expressions, we design a dedicated text-based subnetwork termed *TranscriptionSubNet*. This subnetwork operates as a core component of the broader multimodal architecture, which includes parallel branches for audio and visual modalities. Each subnetwork independently extracts salient features from its respective modality, culminating in a unified multimodal representation for personality trait prediction.

As illustrated in Fig. 2, TranscriptionSubNet processes raw textual transcripts derived from speech through a hierarchical series of neural components. The input transcript undergoes tokenisation and embedding using a pre-trained BERT encoder, followed by a bidirectional LSTM layer to capture contextual dependencies. Multi-head self-attention is then applied to emphasise semantically relevant parts of the sequence. The resulting attention-enhanced representations are pooled and passed through a projection layer to obtain a compact 128-dimensional feature vector, which encapsulates the linguistic characteristics pertinent to personality inference. The complete computational flow is formalised in Algorithm 1.

3.2.2 Audio subnetwork: audioSubNet

To extract affective and personality-relevant cues from the auditory channel, we employ a dedicated audio subnetwork referred to as AudioSubNet. This component is designed to capture both local spectral features and global temporal dependencies inherent in speech signals. As depicted in Fig. 3, the audio input is represented

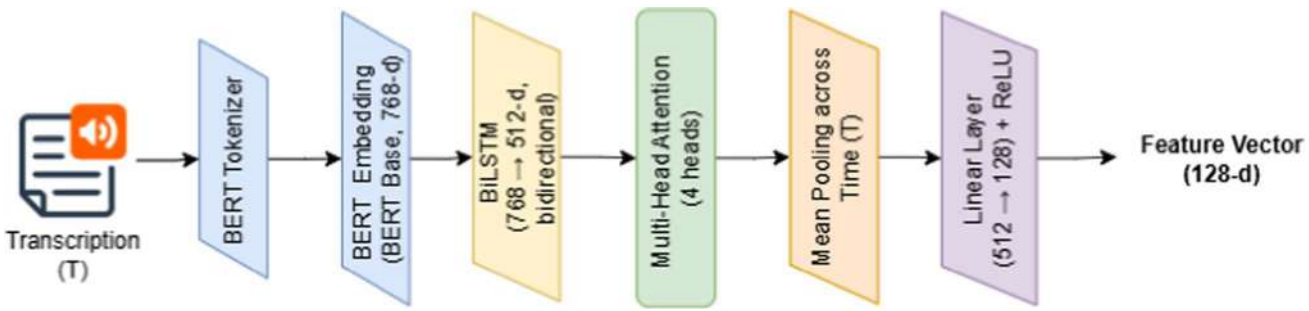

Fig. 2 Workflow of transcription sub network.

Algorithm 1 Transcription subnetwork: TranscriptionSubNet

Input: Transcript T
Tokenization: $(\textbf{input\_ids}, \textbf{attention\_mask}) = \text{BERT\_Tokenizer}(T)$
Embedding: $E = \text{BERT\_Embedding}(\textbf{input\_ids}, \textbf{attention\_mask})$ ▷ BERT is frozen
Sequence Encoding: $H = \text{BiLSTM}(E)$
Self-Attention: $A = \text{MultiHeadAttention}(H, H, H)$
Pooling: $P = \text{Mean}(A)$
Projection: $F_{\text{text}} = \text{ReLU}(\text{Linear}(P))$
Output: Feature vector $F_{\text{text}} \in \mathbb{R}^{128}$

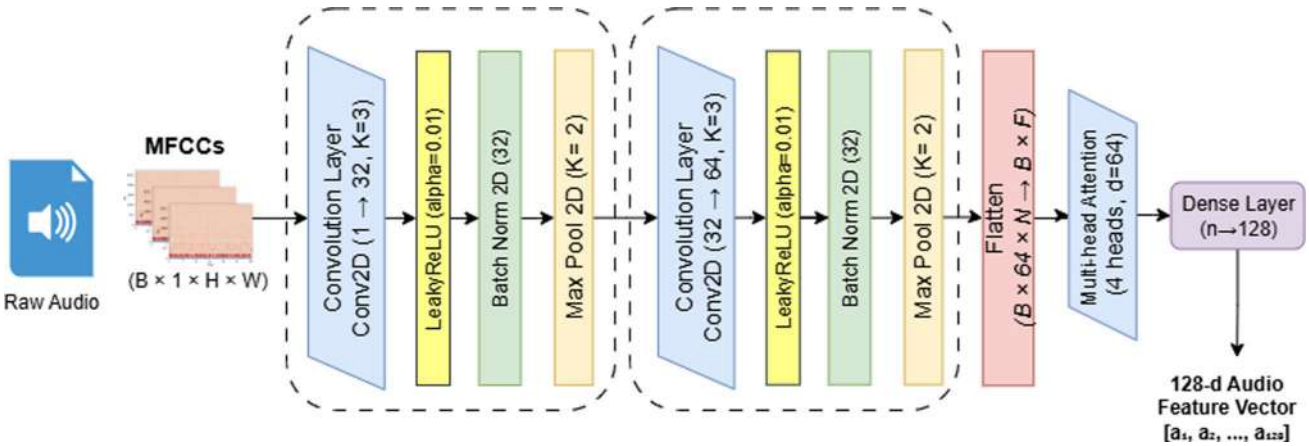

Fig. 3 Workflow of audio sub network

Algorithm 2 Audio Subnetwork: AudioSubNet

Input: MFCC tensor $A \in \mathbb{R}^{24\times1319\times1}$
Conv Block 1: $C_1 = \text{MaxPool2D}(\text{LeakyReLU}(\text{BN}(\text{Conv2D}_{3\times3,16}(A))))$
Conv Block 2: $C_2 = \text{MaxPool2D}(\text{LeakyReLU}(\text{BN}(\text{Conv2D}_{3\times3,32}(C_1))))$
Conv Block 3: $C_3 = \text{MaxPool2D}(\text{LeakyReLU}(\text{BN}(\text{Conv2D}_{3\times3,64}(C_2))))$
Flatten: $F = \text{Reshape}(C_3) \in \mathbb{R}^{L\times64}$ ▷ L is the flattened temporal dimension
Temporal Encoding: $H = \text{MultiHeadAttention}(F, F, F)$
Pooling: $P = \text{Mean}(H)$
Projection: $F_{\text{audio}} = \text{ReLU}(\text{Linear}(P))$
Output: Feature vector $F_{\text{audio}} \in \mathbb{R}^{128}$

using Mel-Frequency Cepstral Coefficients (MFCCs), which are widely adopted in speech-based affective computing. These MFCC features are processed through a series of convolutional layers to encode spatial hierarchies, followed by a Multi-Head Attention (MHA) mechanism to model long-range temporal interactions in the speech signal. The final output is a compact feature vector encoding the speaker's acoustic behaviour. The computational workflow of the audio subnetwork is detailed in Algorithm 2.

3.2.3 Visual subnetwork: VisualSubNet

The visual subnetwork, illustrated in Fig. 4 and detailed in Algorithm 3, extracts expressive spatiotemporal features from input video sequences. Each frame is independently processed by a Residual Dilated Separable

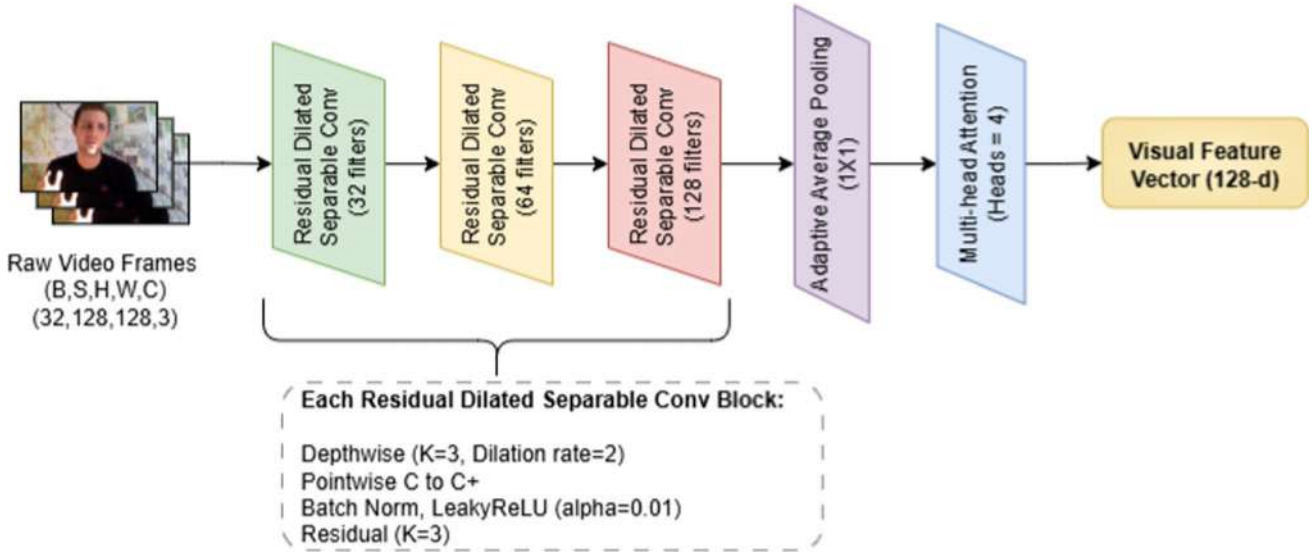

Fig. 4 Work flow for video sub network.

Algorithm 3 Visual Subnetwork: VisualSubNet

Input: Video tensor $V \in \mathbb{R}^{T \times H \times W \times 3}$ ▷ $T = 6$, $H = W = 128$
for $t = 1$ to T **do**
  $F_t = V[t]$ ▷ Extract frame t
  $R_t = \text{RDSC\_Block}(F_t)$
end for
Stack: $R = \text{Stack}(R_1, R_2, \ldots, R_T) \in \mathbb{R}^{T \times h \times w \times c}$
Reshape: $R_{\text{flat}} = \text{Reshape}(R) \in \mathbb{R}^{T \times (h \cdot w \cdot c)}$
Temporal Encoding: $H = \text{MultiHeadAttention}(R_{\text{flat}}, R_{\text{flat}}, R_{\text{flat}})$
Pooling: $P = \text{Mean}(H)$
Projection: $F_{\text{video}} = \text{ReLU}(\text{Linear}(P))$
Output: Feature vector $F_{\text{video}} \in \mathbb{R}^{128}$

Algorithm 4 Fusion and Regression Head

Input: Feature vectors $F_{\text{text}}, F_{\text{audio}}, F_{\text{video}}$
Concatenation: $\mathcal{F} = \text{Concat}(F_{\text{text}}, F_{\text{audio}}, F_{\text{video}})$
Dense Layer: $H_1 = \text{Dense}(256)(\mathcal{F})$
Dropout: $H_2 = \text{Dropout}(H_1)$
Prediction: $\hat{y} = \text{Sigmoid}(\text{Dense}(5)(H_2))$
Output: Personality trait vector $\hat{y} \in \mathbb{R}^5$

Convolution (RDSC) block to capture multiscale spatial information with low computational overhead. The resulting frame-level embeddings are stacked and reshaped before being passed to a Multi-Head Self-Attention (MHA) module, which models temporal dependencies and highlights salient facial dynamics across frames. By applying self-attention over the stacked embeddings, the network effectively captures temporal relationships while alleviating the limitations of sparse frame sampling. Finally, the attention outputs are temporally pooled and linearly projected to produce a fixed-length visual feature vector representing the temporally aware facial characteristics of the input sequence.

Algorithm 5 Training Pipeline

```
for epoch = 1 to 50 do
    for each mini-batch (X, Y) in training set do
        Ŷ = ForwardPass(X)
        ℒ = MSE(Ŷ, Y)
        Backpropagate ℒ
        Update using Adam optimizer
        ClipGradients (max_norm = 1.0)
    end for
    Save model if validation loss improves
end for
```

3.2.4 Fusion and regression head

To integrate the diverse information captured by the modality-specific subnetworks, we employ a fusion strategy followed by a regression head to predict the five personality trait scores. As shown in Algorithm 4, the standard approach involves concatenating the high-level feature vectors extracted from the text, audio, and visual subnetworks. This unified representation is then processed through fully connected layers with non-linearity and regularisation, culminating in a sigmoid-activated output layer for multi-trait prediction. We also experimented with multiple fusion paradigms–namely, early fusion, late fusion, and tensor-based attention fusion to evaluate their comparative effectiveness in capturing cross-modal dependencies.

3.2.5 Training pipeline

To optimise the multimodal personality prediction model, we adopt a supervised training pipeline grounded in robust deep learning practices. The model is trained for a fixed number of 10 epochs using the Adam optimiser, which is well-suited for handling sparse gradients and dynamically adjusting learning rates. Mean Squared Error (MSE) is employed as the loss function to minimise the deviation between predicted and ground truth personality scores across the Big Five dimensions.

To enhance training stability and prevent exploding gradients, gradient clipping with a maximum norm of 1.0 is applied. Furthermore, a validation-based checkpointing mechanism is incorporated to preserve the best-performing model weights, ensuring generalizability. The entire pipeline operates on mini-batches to leverage parallelism and improve convergence efficiency. The training procedure is summarised in Algorithm 5.

3.3 Multimodal fusion

Multimodal fusion lies at the core of our proposed framework, enabling the integration of complementary information derived from text, audio, and visual streams. While each modality offers valuable cues about human behaviour and personality, their combination provides a more complete and nuanced representation of the subject, consistent with real-world human perception. In this study, we plan to implement and evaluate three primary fusion strategies: early fusion, late fusion, and attention-based fusion.

Early fusion: Early fusion involves concatenating the modality-specific feature representations before feeding them into the final prediction layers. In our architecture, 128-dimensional embeddings from the text (BiLSTM), audio (CRNN), and video (RDSN-MHA) subnetworks will be concatenated into a single 384-dimensional vector when all modalities are active. This combined representation will then be passed through fully connected layers for joint trait prediction. Early fusion allows the model to learn interdependencies across modalities during training, often leading to better generalisation when modalities are strongly correlated.

Late fusion: In the late fusion framework, modality-specific features are independently processed through their respective deep sub-networks and prediction heads, each optimised to estimate the Big Five trait scores from text, audio, and visual inputs. Each modality produces its own set of trait-wise predictions using separate regression layers. These unimodal outputs are subsequently aggregated using soft averaging to produce the final personality prediction. This strategy emphasises independent learning per modality while maintaining simplicity and robustness in the fusion step. It is particularly effective when the modalities contribute complementary information and are variably informative across different traits.

Attention-based fusion: We implement an attention-driven intermediate fusion strategy that integrates modality-specific embeddings from text, audio, and visual branches. Features extracted from each unimodal backbone are first projected into a common embedding space and then concatenated as a modality sequence. A multi-head self-attention mechanism is applied across this stacked representation, allowing the model to dynamically learn inter-modal dependencies and context-aware weighting of each modality. The attention output is subsequently aggregated through mean pooling to form a unified representation, which is passed through a regression head to predict the Big Five personality scores. This approach enables the model to capture nuanced interactions between modalities and selectively attend to salient features relevant for each personality trait.

To further control computational overhead during multimodal fusion, we restrict cross-modal interactions to compact feature embeddings rather than raw feature maps. Fusion is performed on low-dimensional representations (128-d per modality), thereby avoiding the high cost of tensor fusion. Additionally, late fusion enables independent unimodal inference, reducing redundant computation when certain modalities are unavailable. This design ensures that the framework remains suitable for deployment in resource-constrained environments.

4 Design considerations and evaluation outcomes

This section presents the empirical evaluation of our proposed multimodal framework for predicting the Big Five personality traits. We analyse the contribution of individual modalities, assess the impact of different fusion strategies, and examine interpretability using XAI methods. All models were trained for 10 epochs on two NVIDIA DGX H100 GPU nodes using the ChaLearn dataset under consistent hyperparameter settings.

4.1 Unimodal and multimodal modality performance

To understand the role of each input modality and their combinations, we trained the model using different modality configurations: audio, visual, text (transcription), and their bimodal and trimodal combinations. Table 2 summarises the train, validation, and test MSE losses. The unimodal results in Table 2 also reflect scenarios where other modalities are absent, effectively serving as modality-missing cases. Owing to the late fusion strategy, the proposed framework degrades gracefully when one or more modalities are unavailable.

The results reveal several important findings:

Table 2 Model performance across different input modalities (MSE Loss)

Input modalities	Train loss	Validation loss	Test loss
Audio	0.0159	0.0182	0.0179
Visual	0.0197	0.0219	0.0194
Transcription (Text)	0.0152	0.0198	0.0199
Audio + Transcription	0.0126	0.0177	0.0178
Video + Transcription	0.0145	0.0185	0.0182
Audio + Video	0.0155	0.0173	0.0170
Audio + Video + Transcription	0.0122	0.0168	0.0161

- *Unimodal insights:* Among the individual modalities, audio yields the best test performance (MSE = 0.0179), followed by visual (0.0194) and transcription (0.0199). This underscores the importance of prosodic and acoustic features in conveying personality cues such as Extraversion and Neuroticism.
- *Bimodal trends:* Fusion of two modalities consistently improves performance. For instance, combining audio with video or text significantly reduces test loss. Notably, the audio+video configuration outperforms the audio+text configuration, suggesting that complementary non-verbal cues strengthen prediction accuracy.
- *Trimodal advantage:* The combination of all three modalities achieves the lowest test loss (MSE = 0.0161). This supports our hypothesis that "3D personality modelling" effectively captures linguistic, vocal, and visual traits holistically.
- *Generalisation capacity:* Across all configurations, training and validation/test losses are well-aligned, suggesting strong generalisation and minimal overfitting.

4.2 Fusion strategy performance across big five traits

To assess the efficacy of multimodal fusion strategies in predicting the Big Five traits, we evaluated three distinct fusion approaches: early fusion (feature-level concatenation), late fusion (ensemble of unimodal regressors), and attention-based fusion (transformer-enabled cross-modal attention with late stacking). The models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Table 3 presents a comprehensive comparison across these metrics.

- *Trait-wise modality insights:* Performance variation across traits highlights modality sensitivities. Extraversion and Neuroticism achieve better R^2 scores with Late Fusion, suggesting stronger individual modality contributions, particularly from audio and visual channels. Text-rich traits, such as Openness, show marginal gains under Attention Fusion, reflecting improved representation learning.
- *Fusion strategy comparison:* All three strategies yield comparable average MSEs (0.0161), but Attention Fusion offers the best overall MAE (0.1014) and highest average R^2 (0.2528), indicating slightly better generalisation and predictive fidelity across traits.
- *Attention fusion advantage:* The attention-based mechanism combines transformer-driven cross-modal integration with ensemble stacking, allowing the model to dynamically prioritise salient modalities per trait. This

Table 3 Performance comparison of fusion strategies across big five traits

Trait	Metric	Early fusion	Late fusion	Attention fusion
Extraversion	MSE	0.0170	0.0168	0.0168
	MAE	0.1049	0.1039	0.1045
	R^2	0.2458	0.2559	0.2554
Neuroticism	MSE	0.0166	0.0171	0.0166
	MAE	0.1025	0.1043	0.1025
	R^2	0.2929	0.2727	0.2915
Agreeableness	MSE	0.0148	0.0149	0.0148
	MAE	0.0970	0.0976	0.0968
	R^2	0.1660	0.1593	0.1687
Conscientiousness	MSE	0.0171	0.0166	0.0169
	MAE	0.1040	0.1042	0.1044
	R^2	0.2673	0.2855	0.2749
Openness	MSE	0.0150	0.0153	0.0152
	MAE	0.0983	0.0991	0.0987
	R^2	0.2815	0.2691	0.2737
Average	MSE	0.0161	0.0161	0.0161
	MAE	0.1013	0.1018	0.1014
	R^2	0.2507	0.2485	0.2528

adaptability subtly improves performance and justifies the inclusion of such architectures in personality trait regression tasks.

4.3 Model eXplainability through saliency analysis

To improve the interpretability of our multimodal personality recognition framework, we employed saliency-based attribution techniques across the text, audio, and visual modalities. These visualisations offer insight into which input features most influenced the model's trait-specific predictions, facilitating transparency and human-aligned explanations.

4.3.1 Text modality

For the textual stream, we computed token-level saliency scores by backpropagating gradients through the BERT+BiLSTM encoder. As depicted in Fig. 5, tokens with emotional or social significance–such as *"rude"*, *"honesty"*, and *"opinion"*, showed strong activations. These salient tokens aligned closely with predictions for traits like Openness and Conscientiousness, suggesting that the model effectively attends to semantically rich linguistic cues when reasoning about personality.

4.3.2 Audio modality

Saliency maps for the audio stream were generated over Mel-Frequency Cepstral Coefficients (MFCCs), highlighting the temporal and spectral features most influential to trait inference. As shown in Fig. 6, MFCC channel 2 exhibited consistently high activation across several time steps, which corresponds to variations in vocal pitch, tone, and energy–acoustic attributes known to relate to emotional expressiveness. These patterns were particularly relevant for traits like Neuroticism and Extraversion, reinforcing the model's ability to localise prosodic indicators of personality.

4.3.3 Visual modality: Grad-CAM analysis

To interpret the visual stream, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the outputs of the visual subnetwork. These heatmaps reveal spatial regions of input video frames that were most influential for trait-specific predictions.

Figure 7 presents Grad-CAM visualisations for the Agreeableness (AGR) trait across three representative samples. Sample 0 shows high activation around the eyes and jawline, often associated with empathetic expressions. In Sample 1, the model focuses on the mouth and cheeks–regions tied to expressive speech and emotional cues.

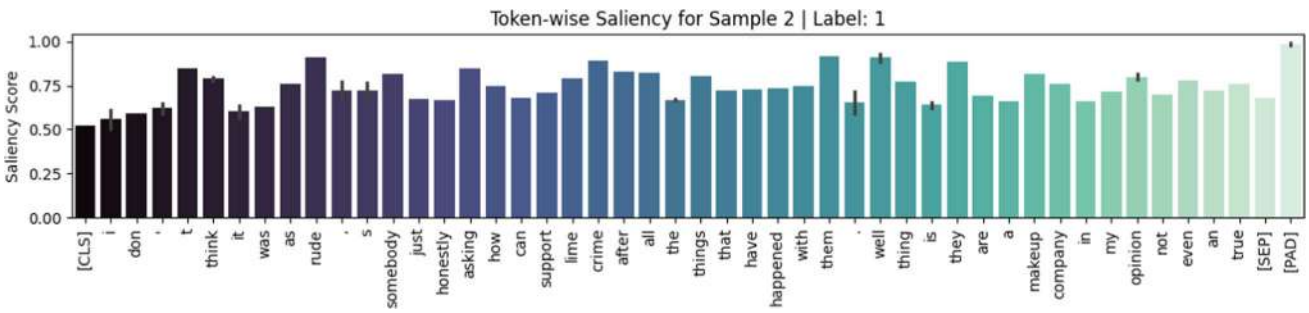

Fig. 5 Token-wise saliency heatmap for transcription input. High-activation words correlate with psychologically relevant features.

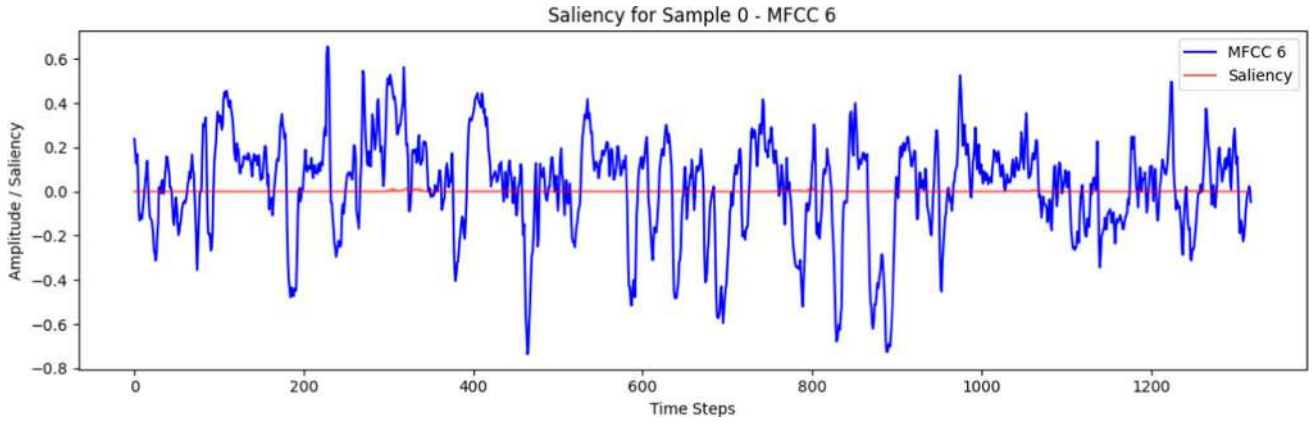

Fig. 6 Saliency visualisation over MFCC spectrogram (Audio input). Warm regions represent time-frequency segments with strong predictive influence.

Fig. 7 Grad-CAM visualisations for the AGR (Agreeableness) trait across three samples. Warm regions indicate high visual importance in the prediction.

Fig. 8 Visual saliency maps for Sample 4 across the Big Five traits. Highlighted regions denote the spatial focus of the model's prediction for each personality dimension.

Despite a neutral appearance in Sample 2, activations are concentrated on the forehead and mid-face, indicating the model's sensitivity to subtle facial dynamics.

All three samples correspond to a ground-truth AGR label of 0. The relatively diffuse activation across the face reflects the model's low confidence in predicting high Agreeableness, reinforcing its discriminative capacity. These results suggest that the visual stream can capture fine-grained, psychologically valid indicators of personality traits.

4.3.4 Visual modality: trait-wise saliency comparison

To further investigate how facial regions contribute differently to each trait, we computed gradient-based saliency maps across the Big Five traits for a single individual. Figure 8 displays the results for Sample 4. Each map highlights areas the model relied upon when predicting a specific personality dimension.

We observe that the AGR and EXT maps emphasise the mouth and cheek regions, likely reflecting the model's focus on expressiveness and sociability. The NEU map shows concentrated attention around the eyes and lower face–areas associated with emotional strain or tension. In contrast, the OPN map displays widespread activation around the eyes and brows, possibly indicating curiosity or reflective cues. These distinctions support the model's capability to learn nuanced, trait-specific facial representations.

These saliency-based analyses validate that the model's decisions are grounded in modality-specific, psychologically interpretable features. Across text, audio, and video, the model demonstrates coherent attention

to emotionally and socially salient cues–supporting the trustworthiness and transparency of our system. These findings not only affirm the internal reasoning of the model but also highlight its potential utility in real-world, human-centric personality assessment applications.

5 Conclusion

This study introduces a modular, multimodal deep learning framework for predicting Big Five personality traits using text, audio, and visual cues. We systematically evaluated three fusion strategies–early, late, and attention-based–and found that attention-based fusion achieved the best overall performance, with the lowest average MSE (0.0161), MAE (0.1014), and highest R^2 score (0.2528). Attention fusion effectively captured cross-modal interactions and showed superior adaptability across traits, particularly for Neuroticism and Openness.

Our results further reveal that different traits are predominantly expressed through different modalities: vocal cues for Extraversion and Neuroticism, linguistic patterns for Openness and Conscientiousness, and visual signals for Agreeableness. The trimodal setup outperformed all unimodal and bimodal baselines, affirming the importance of holistic 3D personality modelling.

We also incorporated interpretability through saliency-based analyses, uncovering psychologically meaningful features influencing predictions–enhancing both model transparency and user trust. Overall, the attention-based approach, supported by explainability tools, provides a balanced solution that combines performance and interpretability.

The ChaLearn First Impressions V2 dataset is predominantly composed of samples from Europe and North America, which may introduce cultural bias and limit cross-cultural generalisation. While the present study focuses on architectural design and multimodal fusion strategies, future work will evaluate the proposed framework on more culturally diverse datasets to enhance fairness and generalisability. In addition, future extensions will investigate time-aligned temporal modelling, cross-cultural validation, and the integration of additional behavioural signals–such as gaze dynamics and physiological cues–to develop more human-aware and culturally robust personality recognition systems. Finally, advanced model compression techniques, including structured pruning, quantisation-aware training, and cross-modal knowledge distillation, will be explored to improve scalability for edge and embedded deployments.

Acknowledgements This research utilized the resources of the Paramshakti supercomputing facility at IIT Kharagpur, established under the National Supercomputing Mission (NSM) of the Government of India and supported by CDAC, Pune. The authors sincerely appreciate the computational resources and support provided, which significantly contributed to the successful completion of this study.

Funding This work was not supported by any external funding sources.

Data availability The data supporting the findings of this study are available upon reasonable request from the corresponding author.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work presented in this paper.

Ethical approval Not applicable.

Consent for publication All authors have given their consent for the publication of this work.

Materials availability Not applicable.

Code availability The code used in this study is available upon reasonable request from the corresponding author.

References

1. Digman J (1990) Personality structure: emergence of the five-factor model. Ann Rev Psychol 41:417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221
2. John OP, Naumann LP, Soto CJ (2008) Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In: John OP, Robins RW, Pervin LA (eds) Handbook of Personality: Theory and Research, 3rd edn. Guilford Press, New York, NY, pp 114–158
3. Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. In: CHI '11 Extended Abstracts on Human Factors in Computing Systems. CHI EA '11, pp. 253–262. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1979742.1979614
4. Ponce-Lopez V, Chen B, Oliu M, Corneanu C, Clapés A, Guyon I, Baró X, Escalante HJ, Escalera S (2016) Chalearn lap 2016: First round challenge on first impressions -dataset and results. European Conference on Computer Vision, 400–418 https://doi.org/10.1007/978-3-319-49409-8_32
5. Agarwal N, Justina Akshaya MJ, Shetty N, Kumar S (2024) Machine learning driven personality prediction system using the concept of cv analysis. In: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), vol. 1, pp. 1–6.https://doi.org/10.1109/ICEECT61758.2024.10738967
6. Gokul K, Lalitha S (2018) Personality identification using auditory nerve modelling of human speech. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1731–1737. https://doi.org/10.1109/ICACCI.2018.8554815
7. Michael J (2003) Using the Myers-Briggs type indicator as a tool for leadership development? Apply with caution. J Leadersh Organ Stud 10(1):68–81. https://doi.org/10.1177/107179190301000106
8. Furnham A (1990) The fakeability of the 16 pf, myers-briggs and firo-b personality measures. Personality Individ Differ 11(7):711–716. https://doi.org/10.1016/0191-8869(90)90256-Q
9. Vasquez RL, Ochoa-Luna J (2021) Transformer-based approaches for personality detection using the mbti model. In: 2021 XLVII Latin American Computing Conference (CLEI), pp. 1–7. https://doi.org/10.1109/CLEI53233.2021.9640012
10. Sun X, Liu B, Cao J, Luo J, Shen X (2018) Who am i? personality detection based on deep learning for texts. In: 2018 IEEE International Conference on Communications (ICC), pp. 1–6. https://doi.org/10.1109/ICC.2018.8422105
11. William Revelle, D.M.C.: A model for personality at three levels. Journal of Research in Personality 56, 70–81 (2015) https://doi.org/10.1016/j.jrp.2014.12.006
12. Eysenck HJ (1991) Dimensions of personality: 16, 5 or 3?-criteria for a taxonomic paradigm. Personality Individ Differ 12(8):773–790. https://doi.org/10.1016/0191-8869(91)90144-Z
13. Cattell HEP, M.A.D (2008) The sixteen personality factor questionnaire (16pf). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.) The SAGE handbook of personality theory and assessment 2, 135–159 https://doi.org/10.4135/9781849200479.n7
14. Marston WM (1928) Emotions of Normal People, 1st edn. Routledge
15. Utami E, Hartanto AD, Adi S, Oyong I, Raharjo S (2022) Profiling analysis of disc personality traits based on twitter posts in Bahasa Indonesia. J King Saud Univ Comput Inform Sci 34(2):264–269. https://doi.org/10.1016/j.jksuci.2019.10.008
16. Lee K, Ashton MC (2020) HEXACO Model of Personality Structure. Springer, Cham
17. Ashton MC, Lee K (2007) Empirical, theoretical, and practical advantages of the hexaco model of personality structure. Pers Soc Psychol Rev 11(2):150–166. https://doi.org/10.1177/1088868306294907.
18. Pletzer JL, Abrahams L (2025) Personality and job performance: A review of trait models and recent trends. Curr Opin Psychol 65:102054. https://doi.org/10.1016/j.copsyc.2025.102054
19. Pennebaker J, King L (2000) Linguistic styles: Language use as an individual difference. J Pers Soc Psychol 77:1296–312
20. Pramodh KC, Vijayalata Y (2016) Automatic personality recognition of authors using big five factor model. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 32–37 . https://doi.org/10.1109/ICACA.2016.7887919
21. Ren Z, Shen Q, Diao X, Xu H (2021) A sentiment-aware deep learning approach for personality detection from text. Inform Process Manag 58(3):102532. https://doi.org/10.1016/j.ipm.2021.102532
22. Konakalla VSKC, Kona CNSM, Chintamaneni THP, Boddu V, Dhuli VS (2023) Personality prediction based on tweets of russo-ukrainian conflict in social networks. In: 2023 IEEE 20th India Council International Conference (INDICON), pp. 1143–1148. https://doi.org/10.1109/INDICON59947.2023.10440884

23. Goldberg, L.R.: A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., Fruyt, F.D., Ostendorf, F. (eds.) European Conference on Personality, vol. 7, pp. 7–28. Tilburg University Press, Tilburg, The Netherlands (1999). https://www.tib.eu/de/suchen/id/BLCP
24. Johnson JA (2014) Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. J Res Pers 51:78–89. https://doi.org/10.1016/j.jrp.2014.05.003
25. Cupani M, Lorenzo-Seva U (2016) The development of an alternative IPIP inventory measuring the big-five factor markers in an Argentine sample. Personality Individ Differ 91:40–46. https://doi.org/10.1016/j.paid.2015.11.051
26. Mairesse F, Walker M (2007) PERSONAGE: Personality generation for dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 496–503. Association for Computational Linguistics, Prague, Czech Republic. https://aclanthology.org/P07-1063
27. Haz L, Rodriguez-Garcia MI, Fernandez A (2025) Using deep neural networks architectures to identify narcissistic personality traits. Expert Syst 42(6):70056
28. Perera BTN, Jayarathne BGDN, Dharmakeerthi TGGM, Thanthilage KTDDK, Priyadarshana YHPP (2020) Shilpa: A novel neural based approach for measuring human stress level. In: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0308–0313. https://doi.org/10.1109/IEMCON51383.2020.9284866
29. Wang J, Saleem N, Gunawan T (2024) Towards efficient recurrent architectures: a deep LSTM neural network applied to speech enhancement and recognition. Cogn Comput 16:1–16. https://doi.org/10.1007/s12559-024-10288-y
30. Natarajan S, Rahman Al-Haddad SA, Ahmad FA, Kamil R, Hassan MK, Azrad S, Macleans JF, Abdulhussain SH, Mahmmod BM, Saparkhojayev N, Dauitbayeva A (2025) Deep neural networks for speech enhancement and speech recognition: a systematic review. Ain Shams Eng J 16(7):103405. https://doi.org/10.1016/j.asej.2025.103405
31. Gan PZ, Sowmya A, Mohammadi G (2023) Clip-based model for effective and explainable apparent personality perception. In: Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing. MRAC '23, pp. 29–37. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3607865.3613178
32. Wang K, Ye C, Zhang H, Xu L, Liu S (2025) Graph-driven multimodal feature learning framework for apparent personality assessment. IECE Trans Emerg Top Artif Intell 2(2):57–67. https://doi.org/10.62762/tetai.2025.279350
33. Wang J, Li H, Woo WL, Shan S (2025) A single modality apparent first impression personality recognition model with temporal emotion based lstm. Expert Syst Appl 259:125114. https://doi.org/10.1016/j.eswa.2024.125114
34. Gucluturk Y, Guclu U, Baro X, Escalante HJ, Guyon I, Escalera S, Gerven MAJ, Lier R (2018) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9(3):316–329. https://doi.org/10.1109/TAFFC.2017.2751469
35. Escalante HJ, Kaya H, Salah AA, Escalera S, Güçlütürk Y, Güçlü U, Baró X, Guyon I, Junior JCJ, Madadi M (2020) Modeling, recognizing, and explaining apparent personality from videos. IEEE Trans Affect Comput 13(2):894–911
36. Li Y, Wan J, Miao Q, Escalera S, Fang H, Chen H, Ali M, Guo G (2020) Cr-net: A deep classification-regression network for multimodal apparent personality analysis. Int J Comput Vision. https://doi.org/10.1007/s11263-020-01309-y
37. Santhosh R, Nagaraja G, Azam F (2025) Deep cross-modal integration with hierarchical multi-head attention for big five personality prediction. Multimed Tools Appl. https://doi.org/10.1007/s11042-025-21006-7
38. Mishra S, Sagnika S (2021) A review on personality prediction. In: Priyadarshi N, Padmanaban S, Ghadai RK, Panda AR, Patel R (eds) Adv Power Syst Energy Manag. Springer, Singapore, pp 61–70
39. Patil P, Goyal S, Dwivedi T, Bhat S (2022) Personality recognition for candidate screening. In: Singh PK, Wierzchoń ST, Chhabra JK, Tanwar S (eds) Futuristic Trends in Networks and Computing Technologies. Springer, Singapore, pp 907–919
40. Ouarka A, Ait Baha T, Es-saady Y, El Hajji M (2024) A deep multimodal fusion method for personality traits prediction. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20356-y
41. Suman C, Saha S, Gupta A, Pandey SK, Bhattacharyya P (2022) A multi-modal personality prediction system. Knowl-Based Syst 236:107715. https://doi.org/10.1016/j.knosys.2021.107715
42. Giritlioglu D, Mandıra B, Yilmaz S, Ertenli U, Akgur B, Kurt AG, Mutlu E, Gürel SC, Dibeklioglu H (2020) Multimodal analysis of personality traits on videos of self-presentation and induced behavior. J Multimodal User Interfaces. https://doi.org/10.1007/s12193-020-00347-7
43. Bao Y, Liu X, Qi Y, Liu R, Li H (2024) Adaptive information fusion network for multi-modal personality recognition. Comput Anim V Worlds 35(3):2268. https://doi.org/10.1002/cav.2268
44. Wang Y, Li D, Funakoshi K, Okumura M (2023) Emp: Emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. ICMR '23, pp. 243–252. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3591106.3592243
45. Waqas M, Zhang F, Laghari AA (2025) Traitbertgcn: Personality trait prediction using bertgcn with data fusion technique. Int J Comput Intell Syst 18:64. https://doi.org/10.1007/s44196-025-00792-w

46. Liu S, Wang K (2025) Comprehensive review: Advancing cognitive computing through theory of mind integration and deep learning in artificial intelligence. In: Proceedings of the 8th International Conference on Computer Science and Application Engineering. CSAE '24, pp. 31–35. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3704814.3704822
47. Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Trans Affect Comput 5(3):273–291. https://doi.org/10.1109/TAFFC.2014.2330816
48. Ilmini W, Fernando T (2023) Detection and explanation of apparent personality using deep learning: a short review of current approaches and future directions. Computing 106(1):275–294. https://doi.org/10.1007/s00607-023-01221-6
49. Mehta Y, Majumder N, Gelbukh A, Cambria E (2020) Recent trends in deep learning based personality detection. Artif Intell Rev. https://doi.org/10.1007/s10462-019-09770-z
50. Zhao X, Tang Z, Zhang S (2022) Deep personality trait recognition: a survey. Front Psychol 13:839619
51. Hashemi M, Rezvani M, Khounsiavash M (2025) AI methods for personality traits recognition: a systematic review. Neurocomputing 640:130301. https://doi.org/10.1016/j.neucom.2025.130301
52. Liu Y, Zhu W, Dong L, Zhang Y, Guo X (2025) Enhancing interpretability in video-based personality trait recognition using Shap analysis. Multimed Syst. https://doi.org/10.1007/s00530-025-01690-z
53. Mishra P (2022) Practical Explainable AI Using Python, 1st edn., p. 344. Apress, Berkeley, CA.
54. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Learning Deep Features for Discriminative Localization . https://arxiv.org/abs/1512.04150
55. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. https://doi.org/10.1109/ICCV.2017.74
56. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. https://doi.org/10.1109/WACV.2018.00097
57. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel P, Hu X (2020) Score-cam: Score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 111–119. IEEE Computer Society, Los Alamitos, CA, USA . https://doi.org/10.1109/CVPRW50498.2020.00020
58. Ahmed O (2024) Exploring explainable nlp techniques for trait extraction and personality inference. PhD thesis, Media Engineering and Technology Faculty, German University in Cairo. https://doi.org/10.13140/RG.2.2.16499.13607
59. Lian Z, Sun H, Sun L, Gu H, Wen Z, Zhang S, Chen S, Xu M, Xu K, Chen K, Chen L, Liang S, Li Y, Yi J, Liu B, Tao J (2024) Explainable Multimodal Emotion Recognition. arXiv:https://arxiv.org/abs/2306.15401
60. Palash M, Bhargava B (2024) Emersk -explainable multimodal emotion recognition with situational knowledge. IEEE Trans Multimed 26:2785–2794. https://doi.org/10.1109/TMM.2023.3304015

Authors and Affiliations

Devraj Patel[1] · Sunita V. Dhavale[1] · Bhushan B. Mhetre[2]

✉ Devraj Patel
devraj_pcse19@diat.ac.in

Sunita V. Dhavale
sunitadhavale@diat.ac.in

Bhushan B. Mhetre
bhushanbmhetre@gmail.com

[1] Department of Computer Science & Engineering, Defence Institute of Advanced Technology (DU), Girinagar, Pune, Maharashtra 411025, India

[2] Department of Psychiatry, Smt. Kashibai Navale Medical College and General Hospital, Narhe, Pune, Maharashtra 411041, India