This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3318367



# Unsupervised Multimodal Learning for Dependency-Free Personality Recognition

Sina Ghassemi[1], Tianyi Zhang[1*], *Member, IEEE,* Ward van Breda, Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, and Reinout E. de Vries

**Abstract**—Recent advances in AI-based learning models have significantly increased the accuracy of Automatic Personality Recognition (APR). However, these methods either require training data from the same subject or the meta-information from the training set to learn the personality-related features (i.e., subject-dependency). The variance of feature extraction for different subjects compromises the possibility of designing a dependency-free system for APR. To address this problem, we present an unsupervised multimodal learning framework to infer personality traits from audio, visual, and verbal modalities. Our method both extracts the handcraft features and transfers deep-learning based embeddings from other tasks (e.g., emotion recognition) to recognize personality traits. Since these representations are extracted locally in the time domain, we present an unsupervised temporal aggregation method to aggregate the extracted features over the temporal dimension. We evaluate our method on the ChaLearn dataset, the most widely referenced dataset for APR, using a dependency-free split of the dataset. Our results show that the proposed feature extraction and temporal aggregation modules do not require personality annotations in training but still outperform other state-of-the-art baseline methods. We also address the problem of subject-dependency in the original split of the ChaLearn dataset. The newly proposed split (i.e., data for training, validation, and testing) of the dataset can benefit the community by providing a more accurate method to validate the subject-generalizability of APR algorithms.

**Index Terms**—Personality assessment, multimodal systems, multimedia signal processing, feature fusion, transfer learning, unsupervised learning, generalization performance

## 1 Introduction

Accurately predicting personality using algorithms is a complex and challenging task that has garnered significant research focus. Particularly in recent years, the tremendous achievement of AI-based methods in addressing sophisticated tasks has inspired researchers to investigate whether it is also possible to accurately predict personality [1], [2]. Such algorithms have the potential to be integrated into a broad variety of applications such as recommendation systems, mental health consulting, and job candidate screening [2], [3], [4].

Previous work [5], [6], [7], [8] on automatic personality recognition (APR) shows that multimodal data, such as audio, visual, and textual data, can improve the performance of APR systems. For example, Li *et al.* [5] found that combining audio and facial data improved the accuracy of APR compared to using these two modalities separately. By combining data from multiple modalities, APR systems can capture a more comprehensive and nuanced representation of personalities [5], [6].

To predict personality traits from multimodal data, deep learning models can be powerful tools because they can automatically learn the non-linear and high dimensional mapping between the input data (e.g., audio, visual, and verbal features) and ground truth labels (e.g., personality traits). However, the data-hungry nature of deep learning methods requires large amounts of training data with proper annotations. The release of the *ChaLearn First Impressions* dataset (*ChaLearn*) [9], [10] has been a major contribution to this line of research by providing 10,000 short video clips with observer personality annotations following the Big Five model of personality [11]. Since then, several deep learning systems [6], [12], [13], [14] have been proposed to address personality prediction based on the *ChaLearn* dataset by using multimodal data.

While much research has focused on developing deep learning algorithms for multimodal APR, most of these have been designed for subject dependent APR [15], [16]. Specifically, to recognize the personality of a specific subject, the algorithm needs training samples from the same subject to learn the embeddings (i.e., high-dimensional features) which can represent personality. This requires prior knowledge either about the subject (e.g., their personality annotation) [17] or the meta information of the training set [15], [16], [18]. For example, Song *et al.* [16] developed a self-supervised learning algorithm of person-specific facial dynamics and used the learned features for APR. In their work, the Personalized Adaptation Layers (PALs) are trained for each subject independently. Although they implemented leave-one-subject-out cross-validation for testing, the network still required meta information about which training samples are from the same subject to train the PALs.

The additional information from subjects can result in bias for APR. Although some researchers [10] have noticed the potential issue of subject-dependency for APR, it is not well addressed by previous works [6], [12], [13], [14]. An important reason for that is the *biased split* of training, validation, and testing sets when researchers develop and validate their APR algorithms. For example, in the original data split of the *Chalearn* dataset, the training, validation, and testing sets have a high dependency be-

- *S. Ghassemi, Tianyi Zhang, W. van Breda, and R. E. de Vries are with the Department of Experimental and Applied Psychology, Vrije Universiteit, Amsterdam, Netherlands*
  *E-mails: {s.ghassemi, t.zhang, w.r.j.van.breda, re.de.vries} @vu.nl*
- *J. K. Oostrom, A. Koutsoumpis and D. Holtrop are with the Department of Social Psychology, Tilburg University, Tilburg, Netherlands*
  *Email: J.K.Oostrom,A. Koutsoumpis,D.J.Holtrop@tilburguniversity.edu*

[1] *both authors contributed equally to this work*
∗ *corresponding author*

tween samples. Specifically, 73.2% and 84% of the samples in the validation and testing sets respectively, share the same YouTube channel with the samples in the training set (i.e., contains prior knowledge about the subjects for validation and testing). Such dependency can be misleading for deep learning models: a high performance due to overfitting can lead to inaccurate confidence in model performance. The difficulty of personality assessment makes deep learning models more prone to memorize individuals by their face and audio characteristics instead of learning to detect personality cues that are generalizable to any individual.

To overcome this problem, in the present work we propose a multimodal system that fuses audio, visual, and verbal modalities for APR. For each modality, an embedding module is proposed and trained on a relatively larger dataset from other tasks (e.g., emotion recognition). After that, the features learned from this module are fine-tuned to represent personality by transfer learning. Unlike many previous studies, our approach enables the feature extraction process to be trained without personality annotations. We also propose a novel and entirely unsupervised temporal aggregation module to aggregate the features by describing their probability distribution and temporal patterns. Since the temporal aggregation and the feature extraction modules do not require personality annotations during training, the system can also be adapted for datasets with smaller sample sizes. Our work contributes to the affective computing community with both technical and empirical contributions:

- The main *technical contribution* of our work is to use unsupervised methods (i.e., the combination of different features, transfer learning, temporal aggregation) for feature extraction and aggregation. The advantage of the unsupervised method is that a) the algorithm does not need annotation at the early stage, which makes it easier for researchers to understand and analyze which features or modalities are more significant than others; b) since the feature extraction and aggregation does not need backpropagation, the algorithm can be adapted for datasets with small sample sizes (i.e., avoid the problem of overfitting for supervised learning).
- The main *empirical contribution* of our work is the proposal of a dependency-free split of the *CharLearn* dataset. To our best knowledge, we are the first to analyze and discuss the subject-dependency issue in the *ChaLearn* dataset. The newly-proposed split of the dataset can provide a more accurate evaluation of the subject-generalizability of APR algorithms.

The rest of the paper is organized as follows. We describe the related studies in Section 2. After that, Section 3 details the proposed system. Section 4 and Section 5 are dedicated to experimental results and discussion. Finally, we discuss the limitation of our work and draw our conclusions in Section 6 and Section 7 respectively.

## 2 Related Work

In this section, we first introduce the Big Five Personality model, which is used to model the ground truth labels of the study. After that, we review the state-of-the-art methods for multimodal automatic personality recognition. At last, we discuss the issue of *subject-dependency* from previous works.

### 2.1 Big Five Personality model

The Big Five Personality model [11] is one of the most widely used models to quantify personality. The Big Five Personality model distinguishes five dimensions of personality: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [11]. The five dimensions of the Big Five Personality model are interrelated but distinct. Research [19] has shown that each trait has a unique and independent contribution to personality. One of the advantages of using the Big Five Personality model is that it has been found to be valid across different cultures and languages [20], [21], making it a useful tool for understanding individual differences in diverse populations. As a consequence, numerous previous works [6], [7], [14], [22] and datasets [9], [23] in automatic personality recognition employ the Big Five Personality model as the ground truth to both train and validate their algorithms. For example, in Chalearn dataset, the Big Five Personality model has been operationalized using two antipodal adjectives for each dimension (e.g., Organized and Sloppy for Conscientiousness [10]) to collect the ground truth labels. In line with the methodology adopted by numerous previous studies, we also employ the Big Five Personality model as the ground truth labels for recognition.

### 2.2 Multimodal automatic personality recognition

APR algorithms using multimodal data can be divided into two major categories: model-specific methods and model-free methods [24]. Model-specific methods require pre-designed hand-crafted features for each modality to represent personality cues from multimodal data. In general, statistical and behavior features such as Local Phase Quantization (LPQ) [14], histogram of oriented gradients (HOG) [25], gazes, head movement [26] and human posture [27] are often extracted from the visual data. For the audio features, the pitch, energy, [28] mean of amplitude and mean of absolute value from the audio Fast Fourier transform (FFT) [29] are also extracted for recognition. The extracted features are then input to machine learning classifiers to recognize personality. One of the advantages of using model-specific methods is the interpretability and explainability of these algorithms: researchers can understand which features make the most significant contribution to the prediction. This advantage can avoid potential bias and discrimination in the developed model (e.g., whether the model recognizes personalities based on gender or ethnicity).

The model-free methods use neural networks to learn the inherent structure between input data and personality labels. Thus, they can automatically extract and fuse features from different modalities for personality recognition. Neural networks such as convolutional neural networks (CNNs) [30] and residual networks [5], [30] have achieved high accuracy for recognition. For example, Y. Güçlütürk *et al.* [30] proposed a dual-stream audiovisual CNN using a deep residual network. It takes a random subset of audio and visual data as input and generates deep learning representations in the output. After that, a global average pooling is applied for temporal aggregation. A similar approach was introduced by F. Gürpinar *et al.* [31] to predict personality on ChaLearn dataset. In addition to the audio and facial expressions, the authors take into account the scene and surrounding objects in the video as well.

Due to the significant correlation between emotions and personality traits, the features learned in the context of emotion recognition [32], [33] are frequently employed for personality recognition by using transfer learning. The benefit of applying transfer learning was recently studied in the work of Zhang *et al.* [6]. The authors demonstrated that the knowledge learned

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3318367



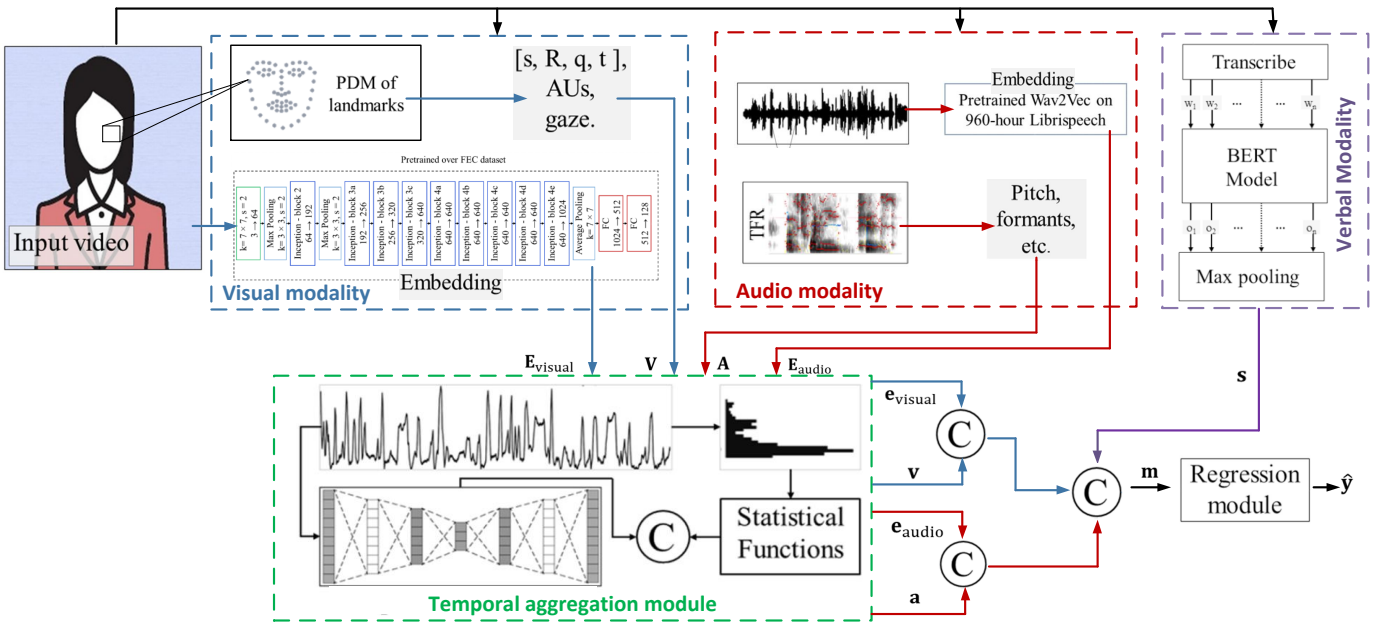


Fig. 1. Pipeline overview: audio (red), visual (blue), verbal (purple) modalities, and aggregation module (green), and 'C' stands for concatenation. During training, only the regression module requires personality annotations.

for emotion analysis can also be beneficial to the prediction of personality traits through a Siamese-like network. This network can jointly learn emotions and personality traits. The experiment results showed that their methodology resulted in the best performance for both tasks.

While the aforementioned studies offer valuable perspectives on multimodal personality recognition, the audio and visual personality cues are accomplished through either model-specific or model-free methods respectively. To take advantage of the benefits of both approaches, our method integrates handcrafted features and deep learning-based embeddings to portray personality cues. Fusion between these two kinds of features enables learning representations without personality annotations. Therefore, it facilitates automatic personality recognition in datasets with smaller sample sizes. Moreover, we introduce an unsupervised temporal aggregation mechanism to merge the obtained features in the temporal domain. The comparison between the proposed method and state-of-the-art supervised feature extraction and temporal aggregation can help us to understand whether additional information (i.e., labels) provided in supervised feature extraction and aggregation can promote recognition accuracy compared with unsupervised methods.

### 2.3 Subject-dependency in automatic personality recognition

Although there have been numerous previous works for automatic personality recognition, the issue of *subject-dependency* is not well addressed. For the vast majority of algorithms developed on Chalearn dataset [6], [7], [12], [13], [14], [22], [30], [31], the issue of subject-dependency is ignored by using the original split of the dataset. Data providers split the videos into training, validation, and testing sets using a 3:1:1 ratio. However, out of 6000 videos in the training set, 4392 videos share the same YouTube channel with the videos in the validation set or testing set. Out of 2000 videos in the testing set, 1680 videos share the same YouTube channel with 2886 videos in the training set. The individual remains the same in this portion of videos with the same YouTube channel. Therefore, the reported performance could be adulterated with unrecognized overfitting.

Compared with dependency-free personality recognition, subject-dependent personality recognition needs either training samples from one specific subject to extract his or her personality cues, or the meta-information about the training set to model the subject-specific features. For example, Rissola *et al.* [34] designed a capsule neural network to extract hidden patterns from conversations for personality recognition. Their approach requires labeled instances from the testing subject to train the network.

Fortunately, some researchers [15], [16], [18] have noticed that issue and initiated measures to tackle it by training and evaluating their method using leave-one-subject-out cross-validation. In the work of Shao *et al.* [15], a person-specific CNN architecture is learned to model the cognitive process of the target subject for personality recognition. The authors run a 7-fold subject-independent cross-validation to validate the performance. However, the model still necessitates the meta-information to discern which training samples correspond to the same subject. A similar method is also used in a more recent work from Song *et al.* [18], where the person-specific cognition is modeled by the neural architecture search learned from each subject separately.

To address the issue of subject dependency, we validate the performance of the proposed model on ChaLearn dataset by utilizing a partitioning scheme in which the videos across training, validation, and testing sets are not related to the same subject. Our analysis in Section 5.1 demonstrates a significant difference in performance between the proposed split of the dataset and its original split.

## 3 Method

**Notations**: In this study, boldface upper-case letters represent two-dimensional matrices, boldface lower-case letters represent

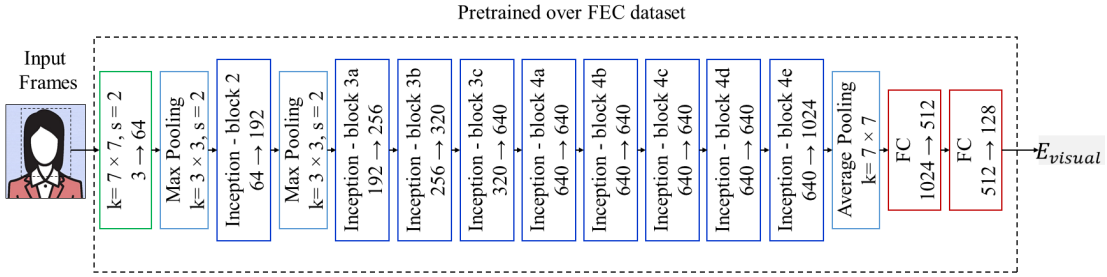


Fig. 2. Visual embedding module. The input frames are cropped around the individual's face and then resized to be 224 × 224 pixels before being fed into the network.

one-dimensional vectors, and lower-case letters represent scalar numbers. In the figures that illustrate neural networks, letters $k$, $s$, and $p$ denote the size of kernel, stride, and padding respectively.

The proposed system, as illustrated in Fig. 1, processes single-speaker video clips to predict six real numbers: the Big Five personality traits and an interview score. Hence, it performs a regression task with six dependent variables. The system's pipeline can be divided into three modalities: audio, visual, and verbal. Each modality processes the input video in two subsequent steps. First, a feature extraction module operates on the input video to compute a set of features. Second, as these features are computed locally in the time axis, a temporal aggregation module processes the extracted features to describe the entire video. Then, the output features of all modalities are fused to provide a multimodal feature vector. In the end, a regression module takes this feature vector as input and predicts the dependent variables. In the following, we detail each module along with its functionality.

### 3.1 Feature Extraction

#### 3.1.1 Visual Modality

As humans, we communicate our emotions through our facial expressions. To exploit such pronounced information, the visual modality aims to describe the individual's observable behavior in the video. We propose that this visual information can be represented by a combination of two types of features: handcrafted features and deeply learned embeddings.

***Handcrafted Features:*** In this context, the handcrafted features refer to the descriptors that are explicitly defined beforehand based on the prior assumption on what type of information is relevant for personality assessment. We extracted 56 handcrafted features using the OpenFace library [35], including 17 action units, facial landmarks, and eye gaze angle and direction.

Action units describe facial movements according to the Facial Action Coding System (FACS) [36] that can be used to code facial expressions in order to describe basic emotions [37], [38]. In the proposed system, 17 action units are detected in each frame of the input video.

Facial landmarks also convey information about the individual's facial expressions. One way to extract such information is through the actual locations of the facial landmarks. However, these locations include redundant information as they are highly correlated with one another. Moreover, due to the variations related to the movements such as head motion and rotation, standardization is required. To address this problem, we represent the facial landmarks in a more efficient and inherently standardized manner using the point distribution model's (PDM) parameters [39], [40], [41]. In PDM, each landmark location $\mathbf{x_i} = [x_i, y_i]^T$ is computed according to PDM's parameters $\mathbf{p} = [s, \mathbf{t}, \mathbf{R}, \mathbf{q}]$:

$$\mathbf{x_i} = s\mathbf{R}(\overline{\mathbf{x}}_\mathbf{i} + \mathbf{\Phi}_i\mathbf{q}) + \mathbf{t} \tag{1}$$

where $s$ is the scaling scalar, $\mathbf{R}$ is the $2 \times 3$ rotation matrix, $\mathbf{q}$ is a $m$-dimensional vector denoting the non-rigid shape parameters, and $\mathbf{t}$ is the translation vector $\mathbf{t} = [t_x, t_y]^T$. $\mathbf{\Phi}_i$ is the $3 \times m$ sub-matrix of the basis variations matrix $\mathbf{\Phi}$ corresponding to the $i^{th}$ landmark. $\overline{\mathbf{x}}_\mathbf{i} = [x_i, y_i, z_i]$ is the mean value of the $i^{th}$ landmark. Note that PDM's parameters, $\mathbf{p}$, are estimated for each frame in the video while the basis variations matrix, $\mathbf{\Phi}$, and the mean location, $\overline{\mathbf{x}}$, are constant. Therefore, PDM's parameters, $\mathbf{p}$, represent the landmarks location in an efficient and standardized manner as non-rigid variations are separate from rotation, scale, and translation.

In total, 17 action units, PDM's parameters $\mathbf{p} = [s, \mathbf{t}, \mathbf{R}, \mathbf{q}]$, and eye gaze angle and direction provide 56 visual handcrafted features. These features for an input video can be denoted by two-dimensional matrix $V \in R^{56 \times t}$ where $t$ is the number of frames in the input video (i.e., $t = video\ length \times frame\ rate$).

***Visual Embedding:*** An advantage of using handcrafted features is that they are explicitly defined. Therefore, they enable interpretation of the system's prediction. However, they cannot fully represent all available information in the input. Consequently, different emotions with subtle visual appearances but different implications may not correctly be distinguished using only handcrafted features. A solution is to represent the input by a continuous transformation via a deep CNN into an embedding space. Moreover, such CNN, if trained on a larger dataset with related annotations such as emotions, can benefit from transfer learning.

In the proposed system, to obtain this embedding space, we introduce a CNN whose architecture is inspired by *NN2* variant of *FaceNet*, proposed by F. Schroff *et al.* [42], which has delivered state-of-the-art performance in face recognition tasks. In the proposed CNN (Fig. 2), we keep the *Inception* [43] backbone up to block 4e. Then, it is followed by: a $7 \times 7$ average pooling layer, two subsequent fully connected layers with 512 and 128 outputs (embedding dimension = 128), and an L2 normalization layer. Moreover, the two fully connected layers are interleaved by batch normalization and ReLU layers. Next, to obtain the visual embedding, the proposed CNN is trained by minimizing the triplet loss on Facial Expression Comparison (FEC) dataset [44]. The FEC dataset is one of the most extensive datasets consisting of 156K face images with around 500K expression triplets. After

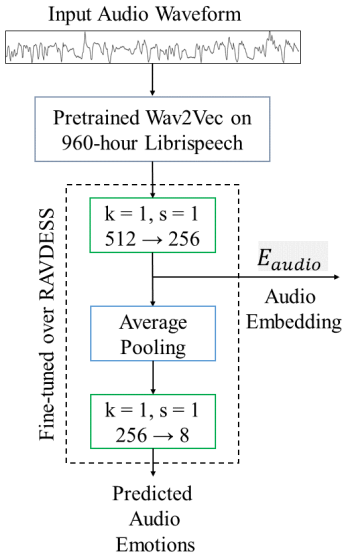


Fig. 3. Audio embedding module. The input audio is first processed by the pretrained Wav2Vec backbone, then the additional convolutional layer, which is fine-tuned on emotion dataset, generates the output embedding.

this training step, the frame is cropped around the individual's face using the OpenFace library [35] in the inference phase. Then the frames are resized to $224 \times 224$ pixels before being input to the CNN. As a result, the proposed CNN learns a function, $f : \mathbf{i} \rightarrow \mathbf{e}$, representing the individual's face, $\mathbf{i} \in R^{224^2}$, with an embedding vector $\mathbf{e} \in R^{128}$. Therefore, given an input video with $t$ frames, $\mathbf{I} \in R^{224^2 \times t}$, the visual embedding can be denoted by $\mathbf{E_{visual}} \in R^{128 \times t}$.

#### 3.1.2 Audio Modality

In addition to visually perceived behavior, audio features (i.e., voice characteristics) may carry information on both self-rated and observer-rated personality traits [45], [46], [47], [48]. Like the visual modality, the audio information is represented using handcrafted features and deeply learned embedding.

***Handcrafted Features:*** Several handcrafted audio features that are commonly used to detect paraverbal emotions [1], [49] are extracted from the individual's voice. These features, which are also referred to as low-level descriptors (LLDs), are computed using a sliding window of 25 (ms) with a step size of 10 (ms) to traverse the entire video. The window and step sizes are chosen based on the previous studies [49], [50], [51]. The LLDs extracted in our system consist of, the pitch frequency and intensity, Mel and linear scale frequency both cepstral and spectral coefficients, harmonic to noise ratio, zero crossing rate, five vocal formants represented by their frequency and bandwidth, audio intensity, chroma features, and a set of spectral features including spectral centroid, bandwidth, flatness, roll-off, and flux, resulting in 79 audio features in total. Like the visual modality, the audio features that are extracted from an input video can be denoted by a two-dimensional matrix $\mathbf{A} \in R^{79 \times t}$ where $t$ is the number of audio frames with a length of 25 (ms) and step size of 10 (ms) from each other.

***Audio Embedding:*** Although the mentioned audio features are broadly used for emotion and behavioral analysis, following the same rationale discussed for the visual modality, we argue that such features alone cannot fully represent all paraverbal information. However, an audio embedding module enables a continuous transformation of data into a space that mimics emotions. Therefore, it can enhance the representational power.

In our system, to achieve this embedding space, we adapt a CNN architecture, namely *Wav2Vec* which is initially proposed by Schneider *et al.* [52] for speech recognition. While the most common input format to audio CNNs is the time-frequency representation, Wav2Vec operates on the original audio waveform thus exploiting all available information in the audio. In a nutshell, Wav2Vec consists of two sub-modules: the encoder network and the contextual network, composed of 5 and 12 convolutional layers respectively. Analogous to the visual modality, we apply transfer learning to obtain audio embedding. Therefore, Wav2Vec is first trained by minimizing contrastive loss using the 960-hour Librispeech [53] dataset according to the proposed methodology by Schneider *et al.* [52]. Next, to make the embedding space represent emotions, two additional convolutional layers with outputs of 256 and 8 with interleaved batch normalization and ReLU are attached to Wav2Vec's last layer (see Fig. 3). Then, these layers are optimized by an emotion recognition task using 1440 speech files of Ravdess dataset [54] to predict 8 basic emotions. After this training step, the first convolutional layer's output with a dimension of 256 is used as the audio embedding.

For clarity of exposition, let $\mathbf{a} \in R^l$ denote the input audio with length of $l$, the embedding module learns a function $f$ to transform the input audio to an embedding space, $f : \mathbf{a} \rightarrow \mathbf{E_{audio}}$, where $\mathbf{E_{audio}} \in R^{256 \times t}$ is a two-dimensional matrix. Here, $t$ is a function of the input length $l$, the CNN's field of view (25 ms), and its step size (10 ms).

#### 3.1.3 Verbal Modality

Spoken and written text is one of the most studied behaviors to infer personality traits [55], [56], [57]. The most influential models of personality themselves (e.g., the Big Five and its successor, the HEXACO model) are based on the lexical tradition [58]. In the proposed system, the verbal modality processes the transcription of individuals' speech using the Bidirectional Encoder Representations from Transformers (BERT) model [59].

The BERT model is a widely deployed language representation model. In the proposed system, BERT is employed to infer verbal features. In the output, BERT generates a feature vector for each word and an aggregated contextual feature vector corresponding to the input [CLS] token. This contextual feature represents the whole sentence [59]. Our preliminary results suggest that using a max-pooling layer on top of word-level features, compared with the contextual feature, results in higher performance. We associate this with the fact that in the ChaLearn dataset videos are cropped randomly. As a result, for a large portion of videos, speech might begin or end at the middle of a sentence. Thus, spoken words have provided more clues than sentences for the raters who annotated the ChaLearn dataset. In Fig. 1, the purple box illustrates the verbal modality where the extracted feature vector is denoted by $\mathbf{s}$.

### 3.2 Temporal Aggregation

A temporal aggregation, denoted by a function $g : \mathbf{F} \rightarrow \mathbf{v}$, is required to output a fixed length feature vector, $\mathbf{v}$, for any given two-dimensional feature matrix $\mathbf{F} \in R^{n \times t}$ with a variable temporal length of $t$. Thereby, aggregating the information across

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3318367



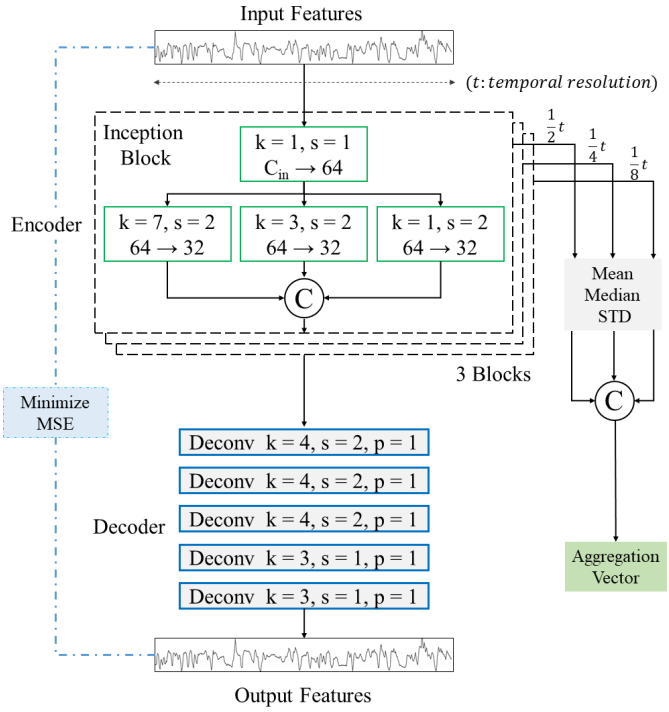


Fig. 4. Autoencoder architecture. In the training, the mean squared error between the input and the output is minimized. In the inference, the aggregation vector (in the green block) is computed. Letter $t$ represents the feature's length in the temporal axis.

$t$ timestamps and representing it in a vector $\mathbf{v} \in R^d$ where $d$ is the aggregation dimension. This section describes the baseline methods and the proposed module for temporal aggregation.

***Baseline:*** We adapt three state-of-the-art neural network architectures as baseline methods. In the training, the extracted features are used as input, personality traits and interview score, as the target output. As such, the network detects temporal patterns in the input features by aggregation over the temporal domain. Therefore, the aggregation function is optimized via the neural network by minimizing the prediction error.

The first adapted neural network is InceptionTime. InceptionTime has achieved state-of-the-art performance in several time series classification tasks [60]. It consists of Inception blocks inspired by Inception architecture [43], which is originally proposed for image classification. The second network is a Transformer [61] which relies on a *self-attention* mechanism to process sequence data. It has achieved state-of-the-art performance in tasks involving sequence data such as natural language processing [59] and speech recognition [52], [62]. The third network is a Recurrent Neural Network (RNN) and includes Gated Recurrent Units (GRU) [63]. LSTM [64] and GRU are the most common building blocks in RNNs which are commonly used in sequence learning tasks. They overcame the exploding and vanishing gradient problem that existed in the traditional RNNs. The key difference between LSTM and GRU is that LSTM has three gates (input, output, and forget) while GRU has two gates (reset and update) which makes it less prone to overfitting and more suitable for the personality assessment task.

The hyperparameters of the baseline models were optimized over the validation set. These optimized hyperparameters for each baseline model are detailed in the supplementary materials.

***Proposed aggregation module:*** The proposed aggregation module consists of two sub-modules. The first sub-module uses statistical functions to describe the features according to their probability distribution. Therefore, it disregards the temporal patterns in the features. These functions include seven aggregation functions, namely median, mean, standard deviation, skewness, kurtosis, minimum, and maximum, and also slope, offset, and curvature of fitting first and second-degree polynomials to the features.

The second sub-module detects the temporal patterns by a generative neural network, namely a *temporal* autoencoder. The autoencoder consists of convolutional layers in the encoder and deconvolutions in the decoder as shown in Fig. 4. During the training, the encoder learns to compress the input's temporal information in a high-dimensional latent vector while the decoder learns to reconstruct the input through deconvolutional layers given the encoder latent vector. Hence, the autoencoder is trained in an unsupervised manner. Inspired by InceptionTime [60], the encoder includes 3 Inception blocks. Each Inception block consists of 3 parallel convolutional layers with filter sizes of 7, 3, 1 with the stride of 2. Each convolutional layer outputs 32 feature vectors. These feature vectors are then concatenated and fed into the next Inception block. In each block, the first layer is a *Bottleneck* convolutional layer with 64 output channels which helps to avoid overfitting by reducing the number of input features.

The aggregation vector, as shown in Fig. 4, combines mean, standard deviation, and median (along the temporal axis) over the outputs of the Inception blocks. Therefore, the aggregation vector describes the probability distribution of temporal patterns detected at three scales (i.e., the output of the three inception blocks have a temporal resolution of $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ with respect to the input temporal resolution).

After computing the output vectors of these two sub-modules, the final output of the proposed aggregation module is the combination of these vectors as shown in Fig. 1. That is, the final output captures both global characteristics of the features using statistical functions and temporal patterns through the autoencoder. In the experimental section, the results show that the proposed aggregation module, while being unsupervised, outperforms the supervised state-of-the-art baselines in aggregating audio and visual features. We conjecture that this improvement is mostly because of the stronger backpropagation signal when reconstruction loss is computed for hundreds of timestamps compared with supervised learning when the prediction loss is computed for only six dependent variables. In general, the features learned in section 3.1 are with different lengths. However, the aggregation vector combines means, standard deviations, and medians (along the temporal axis) over the outputs of the Inception blocks. Therefore, no matter the different temporal sizes of the input features and – subsequently – the different lengths of aggregated outputs, the means, standard deviations, and medians over the time axis are considered, ensuring that fixed-size feature vectors are generated regardless of different input lengths.

### 3.3 Modality Fusion

After the aggregation function is applied over all extracted features in audio and visual modalities, the resulting feature vectors ($\mathbf{v}, \mathbf{a}, \mathbf{e}_{\mathbf{visual}}, \mathbf{e}_{\mathbf{audio}}$) and the verbal features, $\mathbf{s}$, are combined to form a multimodal feature vector denoted by $\mathbf{m}$ in Fig. 1. This type of fusion, i.e., combining features before the regression module, is also referred to as *early fusion*. Our preliminary results suggest

that early fusion is superior to its late variant, in which the predicted outputs are averaged across different modalities.

### 3.4 Regression

In the last block of the proposed system, the multimodal feature vector, $\mathbf{m}$, is fed into a regression module, where all of the six dependent variables, $\hat{\mathbf{y}} \in R^6$ are predicted, as shown in Fig. 1. In the system, regression is performed using the deep ensemble approach. Specifically, a deep ensemble of 32 multilayer perception (MLPs) with 32 and 8 hidden units are deployed. Each MLP is trained using random weights initialization with different seeds and for 200 epochs minimizing the mean squared error. We use early stopping after 20 epochs without any improvement larger than 0.0001 in the validation set's loss function.

## 4 Experimental Results

In this section, we first introduce the dataset and implementation details (including the dependency-free split) for the proposed algorithm. Then we report the performance of our methods using both the dependency-free and original split of the dataset. After that, we compare the result of our method with the state-of-the-art personality recognition algorithms. At last, we run several ablation studies to verify the effectiveness of each feature and modality used in our method.

### 4.1 Dataset

To evaluate the performance of our method, we test it on the First Impressions ChaLearn dataset [10]. To our knowledge, the ChaLearn dataset is the biggest publicly available audiovisual dataset with annotated personality traits. As such, it is the most widely used dataset in the context of multimodal personality prediction. The dataset includes 10,000 15-second videos extracted from more than 3,000 different YouTube channels and annotated by five dimensions of the Big Five personality model plus an interview score. This dataset also provides information on gender, perceived age, nationality, and ethnicity for each subject.

### 4.2 Implementation Details

#### 4.2.1 Dependency-free Split of Chalearn Dataset

As we mentioned in Section 2.3, the originally defined sets (training, validation, and testing) suffer from dependency between samples. This is because of a random assignment of the video clips collected from the same channel or even from the same long video for the training, validation, and testing sets (as shown in Fig. 5 (upper part)). For example, let us consider the videos collected using the YouTube channel with the ID of "-6otZ7M-Mro" as shown in Fig. 5. In the original split of the dataset, two videos namely "-6otZ7M-Mro.000" and "-6otZ7M-Mro.001" are in the training set, "-6otZ7M-Mro.003" and "-6otZ7M-Mro.005" are in the validation set, and the other two videos "-6otZ7M-Mro.002" and "-6otZ7M-Mro.004" are in the testing set. These videos are all related to one individual. Therefore, it can be inferred that the identity of the individual portrayed in this subset of videos featuring the same YouTube channel remains consistent (i.e., the issue of subject-dependency).

To define a dependency-free split for the dataset, we sorted the samples by their YouTube channel ID in ascending order. Then, using the same ratio of 3:1:1, the new split of the data

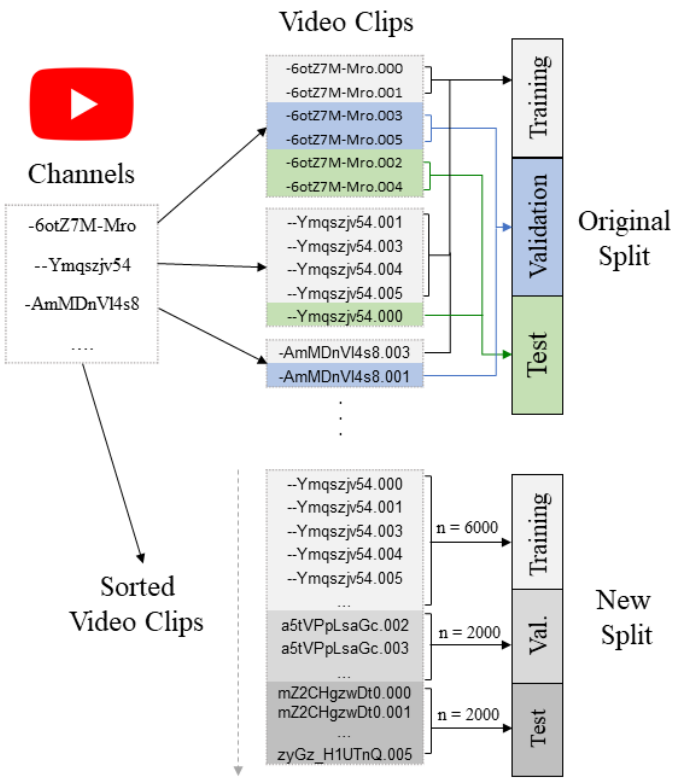


Fig. 5. The original (upper part) and the dependency-free (lower part) split of the ChaLearn dataset. The letter 'n' represents the number of videos in each set.

is defined such that the first 6000 videos used in the training set, the next 2000 videos used in the validation set, and the last 2000 videos used in the testing set (as shown in Fig. 5 (lower part)). In the new split, we also make sure that the distribution of gender and ethnicity approximately follow their distribution over all the samples in the ChaLearn dataset (as shown in Fig. 6).

In this new split, the videos across training, validation, and testing sets do not share the same YouTube channel ID. Thus, they are not related to the same individual. Therefore, the results obtained on this new split correctly measure the system's generalization in detecting personality cues. A detailed discussion and comparison between this new split and the original split are presented in Section 5.1.

#### 4.2.2 Evaluation Metrics

The evaluation metrics used to measure the performance of our method are the coefficient of determination ($R^2$) and mean accuracy ($A$), defined as following:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2}$$

$$A = 1 - \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{3}$$

where $y_i$ is the groundtruth for $i^{th}$ observation, $\hat{y}_i$ is the system prediction, and $\bar{y}$ denotes the mean across all $N$ observations. We prefer $R^2$ over $A$ for our experimental results, as it compares the unexplained variance (i.e., variance of prediction errors) with the total variance of the data. However, since $A$ is also widely used

TABLE 1
Results provided on the testing set in terms of $R^2$, and mean Accuracy ($A$), using the ChaLearn **original** sets for five personality traits (Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)), the average over personality traits, and the interview score. * Fine-tuned and trained using the newly defined and dependency-free split, '-' if not available.

| | O | | C | | E | | A | | N | | Average | | Interview | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $R^2$ | $A$ | $R^2$ | $A$ | $R^2$ | $A$ | $R^2$ | $A$ | $R^2$ | $A$ | $R^2$ | $A$ | $R^2$ | $A$ |
| Song *et al.* (2021) [16] | 0.432 | 0.918 | 0.296 | **0.926** | 0.369 | 0.908 | **0.443** | **0.918** | 0.385 | 0.926 | 0.308 | 0.917 | - | - |
| Aslan *et al.* (2021) [65] | - | 0.916 | - | 0.922 | - | 0.920 | - | 0.916 | - | 0.915 | - | 0.918 | - | - |
| Giritlioğlu *et al.* (2021) [66] | - | 0.913 | - | 0.918 | - | 0.917 | - | 0.913 | - | 0.914 | - | 0.915 | - | - |
| Li *et al.* (2020) [5] | - | **0.919** | - | 0.921 | - | 0.920 | - | 0.917 | - | 0.914 | - | 0.918 | - | **0.924** |
| Principi *et al.* (2019) [67] | - | 0.917 | - | 0.922 | - | 0.916 | - | 0.915 | - | 0.913 | - | 0.917 | - | - |
| Zhang *et al.* (2019) [6] | 0.457 | 0.915 | 0.570 | 0.921 | 0.552 | 0.920 | 0.349 | 0.914 | 0.500 | 0.914 | 0.485 | 0.917 | - | - |
| Bekhouche *et al.* (2017) [14] | - | 0.910 | - | 0.914 | - | 0.915 | - | 0.910 | - | 0.908 | - | 0.912 | - | 0.916 |
| Güçlütürk *et al.* (2017) [12] | - | 0.911 | - | 0.915 | - | 0.911 | - | 0.911 | - | 0.910 | - | 0.916 | - | 0.911 |
| Kaya *et al.* (2017) [7] | - | 0.917 | - | 0.920 | - | 0.921 | - | 0.914 | - | 0.915 | - | 0.917 | - | 0.921 |
| Zhang *et al.* (2016) [68] | 0.437 | 0.912 | 0.544 | 0.917 | 0.481 | 0.913 | 0.338 | 0.913 | 0.475 | 0.910 | 0.455 | 0.913 | - | - |
| **Proposed** | **0.495** | 0.918 | **0.576** | 0.921 | **0.582** | **0.922** | 0.371 | 0.915 | **0.546** | **0.917** | **0.514** | **0.919** | **0.572** | **0.924** |
| ***Proposed*** * | ***0.383*** | ***0.909*** | ***0.372*** | ***0.905*** | ***0.446*** | ***0.909*** | ***0.228*** | ***0.909*** | ***0.414*** | ***0.907*** | ***0.369*** | ***0.908*** | ***0.411*** | ***0.912*** |

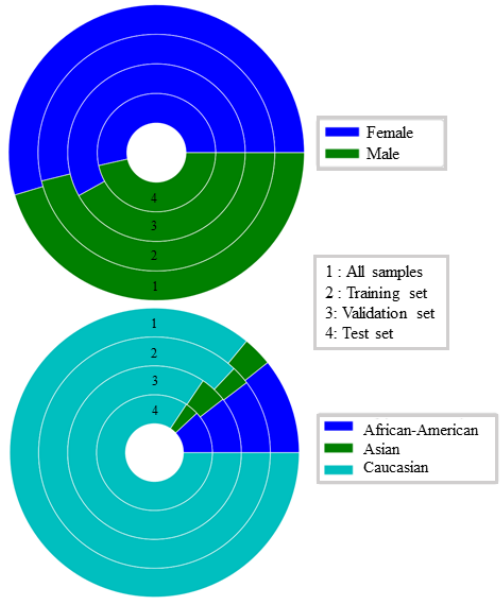


Fig. 6. The gender and ethnicity distribution over the dataset (outer ring) and newly defined training, validation, and testing sets.

by previous works, we also report it for comparison with other state-of-the-art methods.

### 4.3 Results

Table 1 summarizes the $R^2$ and $A$ of our method validated on both the new and the original split of the Chalearn dataset. Our method achieves up 0.9 accuracy for both the new and original split of the dataset, which means that our method does not overfit on one specific split. Although the accuracies for the new and original split are similar, we observe approximately 20% drop in the performance of $R^2$ from the original to the new split.

For the performance of each modality, our method achieves the lowest $A$ and $R^2$ on the dimension of agreeableness. This result is coherent with the results of numerous previous works [6], [68] that agreeableness is more difficult to recognize compared with other dimensions by using video data. In general, our method provides generalizable recognition results among all five dimensions of personality.

### 4.4 Comparison between State-of-the-art

Since most of the state-of-the-art personality recognition methods validate their results using the original split, we use the original split of the dataset to compare the system's performance with them for a fair comparison. Table 1 compares the system performance with several recent studies. As shown in Table 1, our proposed system outperforms the state-of-the-art in Extraversion, Neuroticism, interview score, and on average. However, We find that [16] outperforms (both on $R^2$ and $A$) all the baseline methods (including our method) on the dimension of Agreeableness using the original split of the ChaLearn dataset. In the work of [16], the user-specific facial dynamics are learned from each subject independently. Thus, the network has the meta information about which training samples are from the same subject. We believe this additional information can promote recognition accuracy on a specific personality scale.

However, our method still provides the best recognition results averaged on all the scales. Thus, it is more generalizable among personality scales compared with [16]. It also performs nearly as well or as the best results for the other traits. Particularly, the average $R^2$ improves significantly from 0.369 (1st row) on the newly defined testing set to 0.514 on the original testing set (last row). We believe this improvement is related to the dependency between samples rather than the system's generalization capability. Therefore, we conjecture a similar (20-30%) decline in performance will be true for other studies if the dependency between training and testing samples is removed.

### 4.5 Ablation Study for features and modalities

**Feature-level fusion:** As reported in the first three sections of Table 2, we studied the effect of feature-level fusion between handcrafted and embedding on performance for the audio and the visual modalities and their combination (Audio + Visual). As

TABLE 2
Results are provided on the testing set in terms of $R^2$ using the **newly** defined dataset split for five personality traits (Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)), the average over personality traits, and the interview score. The aggregation over the features is performed using the proposed module. For each modality, the best performance is in bold.

| Modality | Features | O | C | E | A | N | Average | Interview |
|---|---|---|---|---|---|---|---|---|
| Audio | Handcrafted | 0.299 | 0.186 | 0.278 | 0.127 | 0.289 | 0.236 | 0.251 |
| Audio | Embedding | 0.279 | 0.276 | 0.264 | 0.178 | 0.316 | 0.263 | 0.303 |
| Audio | Fusion | **0.314** | **0.278** | **0.291** | **0.182** | **0.334** | **0.280** | **0.314** |
| Visual | Handcrafted | 0.294 | 0.273 | 0.381 | 0.153 | 0.295 | 0.279 | 0.297 |
| Visual | Embedding | 0.293 | 0.238 | 0.358 | 0.157 | 0.299 | 0.269 | 0.300 |
| Visual | Fusion | **0.325** | **0.309** | **0.415** | **0.179** | **0.341** | **0.314** | **0.346** |
| Audio + Visual | Handcrafted | 0.378 | 0.333 | 0.428 | 0.214 | 0.384 | 0.347 | 0.375 |
| Audio + Visual | Embedding | 0.347 | 0.337 | 0.399 | 0.211 | 0.38 | 0.335 | 0.382 |
| Audio + Visual | Fusion | **0.383** | **0.372** | **0.446** | **0.228** | **0.414** | **0.369** | **0.411** |
| Verbal | BERT embedding | 0.034 | 0.059 | 0.033 | 0.017 | 0.042 | 0.037 | 0.041 |
| *Audio + Visual + Verbal* | | *0.387* | *0.374* | *0.445* | *0.225* | *0.411* | *0.368* | *0.411* |

shown in Table 2, for the audio modality, embedding performed better than handcrafted features. However, their combination results in better representation, improving $R^2$ from 0.236 and 0.263, for handcrafted and embedding respectively, to 0.280 for fusion on average. We also observed a similar increase in the interview scores (from 0.251 and 0.303 for handcrafted and embedding respectively, to 0.314 for fusion on average). In the case of the visual modality, handcrafted features outperformed embedding on average. However, similar to the audio modality, the fusion improved the performance considerably from 0.279 and 0.269, for handcrafted and embedding respectively, to 0.314 for fusion on average, and from 0.297 and 0.300 to 0.346 for the interview score. For the combination of these modalities, the results followed the same trend. $R^2$ was increased from 0.347 and 0.335, for handcrafted and embedding respectively, to 0.369 for fusion on average and for the interview score from 0.375 and 0.382 to 0.411. Therefore, as the results imply, the feature-level fusion brought a considerable improvement to performance. These results proved that the embedding obtained by transfer learning when combined with the handcrafted features provides more representative information.

**Verbal modality:** The fourth section of Table 2 reports the results of the verbal features to predict personality traits and the interview score. We conducted several experiments on the validation set with different methods of extracting verbal features such as bags of words, bags of n-grams, and word2vec embedding [69]. Thus far, the best performance was obtained using the pre-trained BERT model [59]. The embedding over each word was extracted by the BERT model and later the verbal feature was obtained by applying a max-pooling layer on the embedding of all words in the transcription.

As shown in Table 2, the verbal modality achieved an average $R^2$ of 0.037. The comparison between these results with other modalities implies that verbal information is less representative of perceived personality in the ChaLearn dataset. These results are in line with previous findings over the ChaLearn dataset [10], [65], [70], [71] where the verbal features resulted in the poorest performance compared with audiovisual features. We conjectured that this poor performance is related to the collection process of the ChaLearn dataset. The videos in the ChaLearn dataset were collected from YouTube channels. Thus, the individuals necessarily do not talk about personality-relevant topics. Moreover, the video clips used as input data were trimmed at random timestamps, which makes them too short to fully cover a topic. Therefore, in most cases, the speech started or ended in the middle of a sentence. In conclusion, it appears that the verbal information provides a very limited amount of clues to the Amazon Mturk workers who have labeled the dataset.

**Modality-level fusion:** Concerning the fusion between modalities, the results, as shown in Table 2, indicate that fusion between visual and audio features, denoted by "Audio + Visual", provides a significant enhancement. Specifically, audio-visual fusion improves the average $R^2$ over personality traits from 0.280 for audio and 0.314 for the visual modality to 0.369. In addition to the personality traits, the $R^2$ for the interview score has increased from 0.382 to 0.411. According to previous works [6], [7], [8], the audio and visual modalities each capture different types of personality cues. Thus, the results they provide are relatively independent. However, they can compensate for each other and provide additional information for learning personality cues, which is one of the advantages of predicting personalities using the fusion of multimodal data compared with using each modality separately.

However, there is no gain when the verbal information is added to the equation denoted by "Visual + Audio + Verbal". As noted above, we associate this lack of improvement with the weak effect of verbal information on the personality ratings for the ChaLearn dataset.

## 5 Discussion

### 5.1 Subject-dependent v.s. dependency-free split: does it make a difference?

As shown in Table 1, the performance of our methods decline by approximately 20% after we implement the dependency-free split of the dataset. Thus, it is worthwhile to discuss whether the newly defined split of the data offers advantages compared with the original split.

TABLE 3
Results are provided on the testing set in terms of $R^2$ using the **newly** defined sets for five personality traits (Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)), the average over personality traits, and the interview score. This table compares the proposed aggregation module with state-of-the-art methods (InceptionTime, Transformer, RNN-GRU) and for visual and audio modalities as well as their combination (Audio + Visual). For each modality, the best performance is in bold.

| Modality | Aggregation | O | C | E | A | N | Average | Interview |
|---|---|---|---|---|---|---|---|---|
| Audio | **Proposed** | **0.314** | **0.278** | **0.291** | **0.182** | **0.334** | **0.280** | **0.314** |
| Audio | InceptionTime | 0.302 | 0.271 | 0.274 | 0.157 | 0.307 | 0.262 | 0.287 |
| Audio | Transformer | 0.295 | 0.259 | 0.274 | 0.162 | 0.317 | 0.261 | 0.292 |
| Audio | RNN (GRU) | 0.241 | 0.209 | 0.214 | 0.147 | 0.267 | 0.216 | 0.260 |
| Visual | **Proposed** | **0.325** | **0.309** | **0.415** | **0.179** | **0.341** | **0.314** | **0.346** |
| Visual | InceptionTime | 0.293 | 0.246 | 0.389 | 0.152 | 0.300 | 0.276 | 0.310 |
| Visual | Transformer | 0.257 | 0.193 | 0.356 | 0.098 | 0.244 | 0.230 | 0.245 |
| Visual | RNN (GRU) | 0.204 | 0.162 | 0.28 | 0.096 | 0.228 | 0.194 | 0.238 |
| Audio + Visual | **Proposed** | **0.383** | **0.372** | **0.446** | **0.228** | **0.414** | **0.369** | **0.411** |
| Audio + Visual | InceptionTime | 0.370 | 0.339 | 0.434 | 0.207 | 0.387 | 0.347 | 0.387 |
| Audio + Visual | Transformer | 0.360 | 0.318 | 0.418 | 0.195 | 0.372 | 0.333 | 0.369 |
| Audio + Visual | RNN (GRU) | 0.289 | 0.244 | 0.331 | 0.164 | 0.324 | 0.270 | 0.329 |

TABLE 4
Standard deviation for each dependent variable computed within videos from the same channel (averaged over all channels, $\overline{\sigma}_{within}$ ) and for all videos in the dataset ($\sigma_{all}$).

| Dependent Variable | $\overline{\sigma}_{within}$ | $\sigma_{all}$ |
|---|---|---|
| Openness | 0.083 | 0.146 |
| Conscientiousness | 0.074 | 0.155 |
| Extraversion | 0.081 | 0.151 |
| Agreeableness | 0.084 | 0.134 |
| Neuroticism | 0.085 | 0.153 |
| Interview Score | 0.076 | 0.148 |

Escalante *et al.* [10] argued that the variations in appearance of the same individual in different videos are high such that it results in different personality scores when the videos are rated separately. However, our analysis in Table 4 shows that the mean of the standard deviations for each dependent variable within the videos from the same YouTube channel ($\overline{\sigma}_{within}$) is much lower than the standard deviation computed for all samples ($\sigma_{all}$). Thus, it indicates that the samples from the same Youtube channels have higher dependency between each other compared with the samples from all videos. When these samples from the same Youtube channels are split in the training, validation, and testing sets respectively, it carries the dependency into different sets. This dependency between samples across different sets is a type of *data leakage* that must be avoided when deep learning models are utilized.

After analyzing the distribution of variables within and across different YouTube channels, we further compared the hyperparameters of our model trained by both the original and new split of the dataset. The original split leads to a different configuration of the aggregation module as the hyperparameter tuning is carried out on the original validation set. Specifically, the best performance is achieved using Fisher vector encoding [72] in which the dimension of the aggregated vector is maximized. Initially, we consider Fisher encoding in our baseline aggregation methods inspired by its recent application in time series classification [73]. However, later we omitted its results due to its weak performance and overfitting problem when it is applied to the new split of the dataset. Our observations indicate that even during the early phase of hyperparameter tuning, the dependency-free split leads to entirely different modules in the system when compared with the original split of the dataset. This observation implies that the original dataset split is not reliable even at the hyperparameter tuning phase, as 1674 videos out of 2000 videos in the validation set share the same YouTube channel with 2948 videos out of 6000 videos in the training set.

In conclusion, the dependency-free split of the dataset makes a significant difference in the aspect of variable distribution and hyperparameter tuning. The newly proposed dependency-free split has important implications for previous and future studies that are based on the ChaLearn dataset.

### 5.2 Comparison between different temporal aggregation methods

In this section, we compare the proposed temporal aggregation method with different aggregation techniques described in Section 3. The comparison can help us understand whether the unsupervised temporal aggregation method we design offers advantages compared with sophisticated supervised neural networks.

As described in Section 3, Transformer, InceptionTime, and GRU networks are trained for 200 epochs using a minibatch size of 256 and by minimizing the mean squared error for the six dependent variables. The Adam optimizer [74] is deployed with the initial learning rate and weight decay of 0.00001. The temporal autoencoder in the proposed aggregation module is also trained using the same optimization setup, however, by minimizing the reconstruction error with an initial learning rate of 0.001. For all networks, as a preprocessing step, standardization is applied for each input dimension and over the time axis. The result on the validation set is used to optimize the hyperparameters. More details regarding the optimized hyperparameters for each network can be found in the supplementary materials.

Here, for the sake of conciseness, results are reported for the fused features (handcrafted + embedding) of audio, visual, and the combination of both modalities in terms of $R^2$ over the five personality traits, its average, and the interview score. The first row for each modality in Table 3 reports the results of applying the proposed aggregation module, while the other three rows report the results of applying the baseline models.

As shown in the upper part of Table 3, for the audio modality, among the baseline models the InceptionTime outperforms the other two baselines in Openness, Conscientiousness, and on average. Instead, the Transformer network gains the best performance for Agreeableness, Neuroticism, and interview score. However, the proposed module outperforms all the baselines in all personality traits and also the interview score. On average, the proposed module outperforms the InceptionTime by a margin of 0.018 improving $R^2$ from 0.262 to 0.280.

For the visual modality (the middle part of Table 3), among the baseline models, the InceptionTime achieves the best performance in all personality traits and the interview score. Nonetheless, the proposed module achieves the largest $R^2$ for all dependent variables. The average $R^2$ has improved from 0.276 to 0.314 by the proposed module compared with InceptionTime performance.

Finally, the bottom part of Table 3 shows the results over the combination of audio and visual modalities (i.e., "Audio + Visual"). We observe a similar phenomenon that the InceptionTime achieves the best $R^2$ performance for all dependent variables among the baseline models. The proposed aggregation has gained the best $R^2$: it is increased by a margin of 0.022 from 0.347 to 0.369 by the proposed module compared with the performance of InceptionTime on average.

Unlike datasets for image or audio recognition, the datasets for personality recognition are small due to the high amount of annotation burden for personality. Although the dataset we use in this work is relatively large, it might be possible that it is not enough for deep-learning based temporal aggregation methods (e.g., Transformer). Due to the limited availability of annotated data in the field of personality recognition, the feature extraction and temporal aggregation method we designed are all unsupervised, rendering them particularly appropriate for tasks characterized by small sample sizes (i.e., avoid the problem of overfitting for supervised learning). Thus, for the task of personality recognition, our method is more suitable compared to deep-learning based methods.

In conclusion, the above-mentioned results show the effectiveness of the proposed aggregation module in summarizing the information across the temporal domain. Therefore, its application is not limited to this specific task and can be utilized in other problems such as action recognition and video classification.

### 5.3 Ethical issue and potential misuse for multimodal-based personality recognition

While multimodal-based APR can have potential benefits such as enhancing user experience or personalizing interventions in healthcare, it can also raise several ethical concerns. As personalities are recognized as "health data" according to General Data Protection Regulation [75] (*EU GDPR, Article 29*), misuse or unauthorized access to this data can violate an individual's privacy. In addition, our method requires facial videos as input for personality recognition. Facial videos are a form of biometric data according to GDPR, which is considered highly sensitive and might lead to profiling since it is unique to each individual. The misuse of such data can lead to identity theft or other serious forms of harm. Thus, individuals must be made aware of the type of data being collected, how it will be used, who will have access to it, and any potential risks. The researchers have to implement robust security measures to protect data from unauthorized access or breaches.

In addition, AI models can also be biased [70] due to flaws in the datasets they were trained on (e.g. if the dataset overrepresents certain demographic groups). This could lead to discriminatory outcomes. Thus, it's also crucial to ensure that the training data is diverse and representative. Researchers are also supposed to check the fairness of the developed model and implement bias mitigation if the developed model is biased among certain groups.

## 6 Limitation and Future Work

Given the challenges of predicting personalities using multimodal data, there are natural limitations to our work. First, we only explored standard hand-crafted features and deep-learned features generated by pre-trained networks. In the future, we want to explore more unsupervised learning frameworks (e.g., the self-supervised learning method used in [16]) to promote recognition accuracy. Second, it is essential to compare the performance of our method with more state-of-the-art personality recognition algorithms. However, there are no benchmark classification results using the dependency-free split of the ChaLearn dataset. Thus, it is difficult to make comparisons with more advanced algorithms. In the future, we will compare our method with more state-of-the-art using open-source Benchmarks [76]. This paper also lacks insightful visualization to validate the effectiveness of the proposed method. Our task encompasses a multi-label regression problem, where conventional visualization strategies may not adequately represent the distribution of personality traits. In future research, we intend to investigate and develop visualization techniques that are specifically tailored to multi-label regression problems in the context of personality recognition. At last, although the feature extraction and temporal aggregation do not require personality annotation (i.e., unsupervised), our method still needs these annotations for regression. In the future, we will extend our algorithm for unsupervised regression and compare its performance with other supervised methods.

## 7 Conclusion

In this research, a multimodal system is proposed that incorporates various improvements, such as a novel temporal aggregation method and feature extraction strategy. Our ablation studies conducted on different features demonstrate that the embedding obtained through transfer learning can significantly enhance the representational capability when combined with handcrafted features. Our experiments on different temporal aggregation methods also show that the proposed aggregation module within the system has the ability to identify both global and local temporal information. In general, our method surpasses the current state-of-the-art supervised baselines even when trained in an unsupervised manner. Thus, the feature extraction and temporal aggregation do not require personality annotations in the training which means that they can be adapted for datasets with small sample sizes. Furthermore, we discuss an unintended dependency in one of the most referenced video datasets. Our proposition introduces

a new split of the dataset aimed at resolving the issue of subject-dependency of the Chalearn dataset. It offers a more precise means to authenticate the subject-generalizability personality recognition algorithms.

## References


[1] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[2] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, pp. 1–27, 2019.

[3] H. Ning, S. Dhelim, and N. Aung, "Personet: Friend recommendation system based on big-five personality traits and hybrid filtering," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 394–402, 2019.

[4] C. C. Liem, M. Langer, A. Demetriou, A. M. Hiemstra, A. S. Wicaksana, M. P. Born, and C. J. König, "Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening," in *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 197–253.

[5] Y. Li, J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi, and G. Guo, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," *International Journal of Computer Vision*, pp. 1–18, 2020.

[6] L. Zhang, S. Peng, and S. Winkler, "Persemon: a deep network for joint analysis of apparent personality, emotion and their relationship," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 298–305, 2019.

[7] H. Kaya, F. Gurpinar, and A. Ali Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–9.

[8] H. Kaya and A. A. Salah, "Multimodal personality trait analysis for explainable modeling of job interview decisions," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 255–275.

[9] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European Conference on Computer Vision*, 2016, pp. 400–418.

[10] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güç;lütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. Jacques, M. Madadi, S. Ayache, E. Viegas, F. Gurpinar, A. S. Wicaksana, C. Liem, M. A. J. Van Gerven, and R. Van Lier, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing (Early Access)*, pp. 1–1, 2020.

[11] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[12] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 316–329, 2017.

[13] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2017.

[14] S. Eddine Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 10–13.

[15] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Personality recognition by modelling person-specific cognitive processes using graph representation," in *proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 357–366.

[16] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2021.

[17] G. An, S. I. Levitan, J. Hirschberg, and R. Levitan, "Deep personality recognition for deception detection." in *INTERSPEECH*, 2018, pp. 421–425.

[18] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2022.

[19] J.-M. Dewaele, "Personality: Personality traits as independent and dependent variables," *Psychology for language learning: Insights from research, theory and practice*, pp. 42–57, 2012.

[20] Y. H. Poortinga, F. J. Van De Vijver, and D. A. Van Hemert, "Cross-cultural equivalence of the big five: A tentative interpretation of the evidence," *The five-factor model of personality across cultures*, pp. 281–302, 2002.

[21] L. A. Migliore, "Relation between big five personality traits and hofstede's cultural dimensions: Samples from the usa and india," *Cross Cultural Management: An International Journal*, vol. 18, no. 1, pp. 38–54, 2011.

[22] F. Gürpınar, H. Kaya, and A. A. Salah, "Combining deep facial and ambient features for first impression estimation," in *European Conference on Computer Vision*, 2016, pp. 372–385.

[23] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 350–359.

[24] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2020.

[25] J. Joshi, H. Gunes, and R. Goecke, "Automatic prediction of perceived traits using visual cues under varied situational context," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 2855–2860.

[26] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion—a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 443–455, 2012.

[27] L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez, "Multimodal analysis of body communication cues in employment interviews," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 437–444.

[28] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection," in *Proceedings of the 2007 Workshop on tagging, mining and retrieval of human related activity information*, 2007, pp. 9–14.

[29] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, p. 107715, 2022.

[30] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *European Conference on Computer Vision*, 2016, pp. 349–358.

[31] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *23rd International Conference on Pattern Recognition*, 2016, pp. 43–48.

[32] T. Zhang, A. El Ali, A. Hanjalic, and P. Cesar, "Few-shot learning for fine-grained emotion recognition using physiological signals," *IEEE Transactions on Multimedia*, 2022.

[33] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Weakly-supervised learning for fine-grained emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, 2022.

[34] E. A. Rissola, S. A. Bahrainian, and F. Crestani, "Personality recognition in conversations using capsule neural networks," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019, pp. 180–187.

[35] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *The 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66.

[36] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[37] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[38] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of Neuroscience Methods*, vol. 200, no. 2, pp. 237–256, 2011.

[39] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," in *Medical Imaging: Image Processing*, vol. 4322, 2001.

[40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[41] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[44] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5683–5692.

[45] P. Ekman, W. V. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect." *Journal of Personality and Social Psychology*, vol. 38, no. 2, p. 270, 1980.

[46] J. Stern, C. Schild, B. C. Jones, L. M. DeBruine, A. Hahn, D. A. Puts, I. Zettler, T. L. Kordsmeyer, D. Feinberg, D. Zamfir *et al.*, "Do voices carry valid information about a speaker's personality?" *Journal of Research in Personality*, vol. 92, p. 104092, 2021.

[47] A. Koutsoumpis and R. E. de Vries, "What does your voice reveal about you?" *Journal of Individual Differences*, 2021.

[48] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. van Breda, S. Ghassemi, and R. E. de Vries, "The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the big five and the linguistic inquiry and word count (liwc)." *Psychological Bulletin*, vol. 148, 2022.

[49] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2227–2231.

[50] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 7390–7394.

[51] 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, 2017.

[52] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in *The Annual Conference of the International Speech Communication Association*, 2019, pp. 3465–3469.

[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[54] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *Plos One*, vol. 13, no. 5, p. e0196391, 2018.

[55] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[56] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[57] K. Luyckx and W. Daelemans, "Using syntactic features to predict author personality from text," *Proceedings of Digital Humanities*, vol. 2008, pp. 146–9, 2008.

[58] M. C. Ashton and K. Lee, "Objections to the hexaco model of personality structure—and why those objections fail," *European Journal of Personality*, vol. 34, no. 4, pp. 492–510, 2020.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Human Language Technologies*, 2019, pp. 4171–4186.

[60] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inception-time: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[62] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Annual Conference on Neural Information Processing Systems*, 2020.

[63] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1724–1734.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[65] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, vol. 110, p. 104163, 2021.

[66] D. Giritlioğlu, B. Mandira, S. F. Yilmaz, C. U. Ertenli, B. F. Akgür, M. Kınıklıoğlu, A. G. Kurt, E. Mutlu, Ş. C. Gürel, and H. Dibeklioğlu, "Multimodal analysis of personality traits on videos of self-presentation and induced behavior," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 337–358, 2021.

[67] R. D. P. Principi, C. Palmero, J. C. Junior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 607–621, 2019.

[68] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis," in *European Conference on Computer Vision*, 2016, pp. 311–324.

[69] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, Workshop Track Proceedings*, 2013.

[70] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 268–277.

[71] L. Hickman, N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo, "Automated video interview personality assessments: Reliability, validity, and generalizability investigations." *Journal of Applied Psychology*, vol. 107, no. 8, p. 1323, 2022.

[72] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[73] W. Huang, B. Yue, Q. Chi, and J. Liang, "Integrating data-driven segmentation, local feature extraction and fisher kernel encoding to improve time series classification," *Neural Processing Letters*, vol. 49, no. 1, pp. 43–66, 2019.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[75] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "Eu general data protection regulation: Changes and implications for personal data collecting companies," *Computer Law & Security Review*, vol. 34, no. 1, pp. 134–153, 2018.

[76] R. Liao, S. Song, and H. Gunes, "An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition," *arXiv preprint arXiv:2210.09138*, 2022.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3318367



**Sina Ghassemi** is a postdoctoral researcher at Vrije Universiteit Amsterdam. He received his PhD in Telecommunication Engineering with specialization in Signal Processing and AI from Politecnico di Torino, Turin, Italy, in 2018. His research interests are in the field of deep learning, signal processing, and affective computing.

**Tianyi Zhang** is currently working as a postdoc at Vrije Universiteit Amsterdam. He got his PhD degree in the faculty of Electrical Engineering, Mathematics & Computer Science (EEMCS) in Delft University of Technology. He was also associated with the Distributed & Interactive Systems (DIS) group at Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands. His research interests lie in human-computer interaction and machine learning based affective computing and automatic personality recognition (C.A.: t.zhang@vu.nl, https://tianyi-zhang-tz.github.io/Tianyi-Zhang-TZ/).

**Ward van Breda** completed a PhD in the application of machine learning in the domain of mental health, which was conducted at Vrije Universiteit Amsterdam. Currently Ward is involved in several research projects where machine learning is used to model mental states, such as stress and anxiety, and personality traits, in different experimental settings.

**Antonis Koutsoumpis** is a PhD candidate at the Organizational Section of the Experimental and Applied Psychology Department, at the Vrije Universiteit Amsterdam, the Netherlands. He has a background in psychology and his research interests include automatic personality assessment from asynchronous video interviews as well as the verbal and non-verbal behaviors that individuals exhibit depending on their personality traits.

**Janneke Oostrom** is a Professor of Work & Organizational Psychology at Tilburg University's department of Social Psychology. Her research focuses on understanding and improving psychological assessments, with the goal to make them more predictive of future work behaviors, while reducing discrimination against marginalized groups. She received her Master (2005) and Ph.D. (2010) in Work and Organizational Psychology from the Erasmus University Rotterdam.

**Djurre Holtrop** is an assistant professor at Tilburg University's department of Social Psychology. He has worked in consulting for psychometric assessments, leading large scale online assessment projects. He completed his PhD at the VU Amsterdam studying the refinement of personality and vocational interest questionnaires. Subsequently, he worked for the University of Western Australia and Curtin University to study the recruitment, motivation, and retention of volunteers. Currently, his research focusses on personnel recruitment and selection and volunteer attraction and engagement.

**Reinout E. de Vries** is Full Professor in Organizational Psychology with a chair in 'Personality at Work' at the Vrije Universiteit Amsterdam, the Netherlands. His main areas of interest are the theoretical background, structure, measurement, and effects of personality, leadership, communication styles, and situations.