Learning Personalised Models for Automatic Self-Reported Personality Recognition

Hanan Salam HANAN.SALAM@NYU.EDU
Viswonathan Manoranjan* VM2336@NYU.EDU
SMART Laboratory, Department of Computer Science
New York University Abu Dhabi, UAE
Jian Jiang JIAN.JIANG@KCL.AC.UK
Oya Celiktutan OYA.CELIKTUTAN@KCL.AC.UK
SAIR Laboratory, Centre for Robotics Research
Department of Engineering
King's College London, UK

Abstract

Previous research has revealed differences in personality traits among different genders, age groups, and even cultures. However, existing methods have focused on one-fits-all approaches only and performed personality recognition without taking into consideration the user's profiles. In this paper, we propose to learn personalised models of self-reported big five personality traits. Our proposed approach automatically learns deep learning architectures for different user profiles using Neural Architecture Search (NAS) for predicting the Big Five personality traits from multimodal behavioural features. We experiment with two different user profiling criteria, namely, gender and age, and compare the results of our approach with the state-of-the-art methods. Overall, our results show that personalised models improve the performance as compared to the generic model. Particularly, gender-based user profiling combined with bimodal features reduces the prediction error by 0.128, achieving the state-of-the-art performance on the UDIVA dataset.

Keywords: Automatic personality recognition; Neural Architecture Search; personalised models; multimodal human behaviour analysis; personality computing

1. Introduction

Smart phones, voice assistants, and home robots are becoming more intelligent every day to support humans in their daily routines and tasks. Achieving the user acceptance and success of such technologies makes it necessary for them to be socially informed, adaptive, responsive, and responsible. They need to understand human behaviour and socio-emotional states and adapt themselves to their user's profiles (e.g., personality, gender, age) and preferences. Motivated by this, there has been a significant effort in the development of personality computing frameworks in the last decade (Vinciarelli and Mohammadi, 2014a; Silveira Jacques Junior et al., 2019). The success of machine learning, especially deep learning techniques in learning complex patterns of multimodal data as well as the availability of large-scale datasets for personality recognition (e.g., collected from YouTube) has also revamped the development of such frameworks.

* The second author and the third author contributed equally to this research.

Existing approaches in personality computing employ non-verbal behaviours as predictors of personality (Celiktutan and Gunes, 2017; Salam et al., 2016; Fang et al., 2016; Zhang et al., 2019). Most of these approaches consider one-fits-all paradigms. However previous research has revealed differences in personality traits among different user profiles, i.e., age groups, genders, and even cultures (Akyunus et al., 2021; Weisberg et al., 2011b; Hang et al., 2021). For example, females assess their personality traits differently from males. A study showed that females reported higher Big Five Extroversion, Agreeableness, and Neuroticism scores than males (Weisberg et al., 2011a). Similarly, there were also differences in self-assessment from childhood to adulthood (McCrae et al., 1999; Hang et al., 2021; Akyunus et al., 2021). A study by Roberts et al. (2006) including 92 participants investigated personality mean-level change patterns throughout the life course. The study revealed that measures on Conscientiousness and Neuroticism increased in young adulthood, whereas Openness increased in adolescence but then decreased in old age. On the other hand, personality is strongly coupled with nonverbal behaviours (Gallaher, 1992). For example, extroverted people talk more, louder, and faster and have a higher hand or facial gesture frequency. Introverted people, in contrast, avoid making eye contact and display less gestures (Jensen, 2016). In some cultures, people use gestures differently and even more or less frequently (Archer, 1997). This means that the way personality is expressed via nonverbal behaviours can differ across cultures as well (Archer, 1997). Therefore, personalised models that can take into account the individual profiles or differences seem to be the most promising approach.

Motivated by this, in this paper, we propose to learn personalised models for different user profiles (i.e., age and gender) and compare age-wise and gender-wise personalisations. For each user profile, the proposed model learns individual neural network architectures for different modalities (i.e., visual, textual, and audio) and fuses them at the decision level. The proposed approach was our response to the ChaLearn LAP Challenge on Understanding Social Behaviour in Dyadic and Small Group Interactions (DYAD) at ICCV 2021 - Automatic Self-reported Personality Recognition Track (Palmero et al., 2022). The main contributions of this work can be summarized as follows:

1. We propose to build personalised models for automatic self-reported personality. To this end, we create user profiles based on gender and age and use a Neural Architecture Search framework for designing and training separate models per user profile.

2. We evaluate our approach with a rich set of multimodal features including visual, textual, and audio features singly as well as fusing the best performing two features at the decision level.

3. Our experimental results show that personalised models improve the recognition performance in general, and achieve the state-of-the-art performance on the UDIVA dataset (Palmero et al., 2021, 2022).

2. Related Work

In this section, we review relevant work in personality computing, personalized models in the field of affective computing, and the applications of Neural Architecture Search (NAS).

2.1. Personality Computing

Personality is a set of habitual patterns of behaviours, thoughts or emotions that can characterise a certain individual (Soto and John, 2017). The predominant paradigm in personality computing research (Vinciarelli and Mohammadi, 2014b; Dhelim et al., 2021) is called the Big Five Model (McCrae and John, 1992) (also called as OCEAN). It defines personality traits along five broad dimensions: (1) extroversion (assertive, outgoing, energetic, friendly, socially active), (2) neuroticism (tendency to experience negative emotions), (3) openness (inclination towards experiences, adventure, novelty), (4) agreeableness (cooperation, compliance, trustworthiness), and (5) conscientiousness (self-discipline, high organization, consistency). Training personality recognition models requires the acquisition of the personality ground truth labels. These can be collected through subjective or objective assessment. Subjective assessment entails self-reporting or self-assessment, where the individuals report their perception of their personality by answering personality questionnaires. Objective assessment is conducted by asking external observers to provide their impressions regarding others' personality. This work is situated in the category of self-reported big five personality traits.

Existing personality computing approaches have focused on the extraction of non-verbal behaviours from different modalities including visual, audio, textual, and contextual, which are then used as predictors of the different personality trait scores. In recent years, deep learning approaches have particularly gained success in the personality computing literature due to their ability to learn complex hidden behavioural patterns from raw data or from non-verbal features.

Unimodal. Unimodal approaches based on the visual modality have been extensively reviewed in a recent survey (Junior et al., 2019). Among these approaches, Romeo et al. (2021) exploited body language cues extracted from the visual modality with state-of-the-art deep architectures such as 3D Residual Network (3DResNet), 3D Convolutional Neural Networks (3DCNN), a modified version of VGG-16 (VGG DAN+), and a combination of CNN and Long Short-Term Memory Network (CNN+LSTM). Deep learned facial dynamics in conjunction with Artificial Neural Networks were proposed by Song et al. (2021). In terms of textual modality, word-level embeddings were used as an input to deep models (Xue et al., 2021). Audio-based models exploited prosody, speech activity, voice quality, interaction features, and OpenSMILE speech features (Eyben et al., 2010).

Multimodal. Multimodal approaches fused different modalities to obtain personality trait predictions. A special emphasis on incorporating interpersonal features from interaction partners is apparent in the most recent approaches. Examples include the work by Palmero et al. (2021), where a transformer-based method was proposed to learn high level deep audiovisual and contextual features for inferring OCEAN scores. In their formulation, contextual features included features learned from the interaction partner. Various deep learning architectures were employed to learn the multimodal deep features, such as R(2+1)D network for visual features, VGGish for audio, which were then fused with contextual features and used as input to a Transformer network implementing multiheaded attention units. Similarly, Aslan et al. (2021) designed specialised sub-networks for ambient appearance (scene), facial appearance, voice, and transcribed speech modalities. The modality-specific representations were then fused via an attention mechanism. Another

approach (Curto et al., 2021) focused on implicitly learning individual and interpersonal features from the interaction partners via a multimodal multi-subject Transformer architecture using variable time windows. The proposed network included a cross-subject layer with attentional operations allowing a joint modeling of the interactants' behaviours. Although these approaches give us a positive outlook, they follow a one-fits-all approach as a fixed architecture is designed and trained on a pool of user profiles.

2.2. Personalized Models in Affective Computing

There is a recent trend to use personalised predictive models in various affective computing tasks. Personalising machine learning models can be performed at the (1) user level, or (2) user profile level.

User level. Personalising models at the user level entails learning adaptive models tailored towards each user. Using multitask learning to take into account individual differences for mood, stress, and health prediction showed improved performance over one-fits-all machine learning approaches for these tasks (Taylor et al., 2017; Jaques et al., 2016). Similarly, multitask learning was used with Gaussian process regression models for personalising self-reported pain prediction (Liu et al., 2017). The underlying motivation for multitask learning was the ability of this methodology to account for individual differences while leveraging data across the population. Supervised domain adaption with mixture of experts for personalising deep CNN models for expression recognition has also proven efficient as compared to non-personalised models (Feffer et al., 2018). User-level personalisation for facial emotion recognition was accomplished by learning and propagating individual deep facial features for each subject via a CNN architecture followed by a spatial attention map, which was then fused into a CNN (Shahabinejad et al., 2021).

User profile level. Personalisation at the user profile level entails profiling users according to a certain criteria, and then building adaptive models that takes into account these user profiles. Profile-level personalisation were less explored in the literature. In a human-robot interaction (HRI) autism therapy framework, personalised deep models were proposed for learning child-specific models of valence, arousal, and engagement (Rudovic et al., 2018a). The proposed deep architecture used specific layers to nest the children based on their culture and gender, which was followed by individual network layers for each child. Rudovic et al. (2018b), in the HRI for autism therapy setting, introduced the CultureNet, a deep learning architecture which leverages on the culture data to adapt to the target culture and child.

Although the current trend seems to be building personalised models for recognising affective states, such as mood, engagement, there has been a little effort for investigating personalisation for personality recognition. Only Shao et al. (2021) employed Neural Architecture Search (NAS) to model a subject's cognitive process. An optimal person-specific CNN architecture was learned based on the audio-visual non-verbal cues displayed by the conversational partner to predict the target subject's facial reactions. The results of the optimal CNN architecture were then used to create a person-specific graph representation for recognising the target subject's personality. However, this method was limited to interactive scenarios only, as it requires the nonverbal cues of the interacting partner. Moreover, it proposed to train a specific model for each subject, which may be computationally ex-

pensive. In contrary, our proposed approach depends on features extracted from the target subject only and proposes to assign users to pre-defined profiles for designing and training individual models for different user profiles.

2.3. Neural Architecture Search

Neural architecture search (NAS) has been proposed to automatically adjust deep neural networks, without the need for manually designing the architecture (Ren et al., 2020; Elsken et al., 2019). Existing search algorithms include NASNet (Feurer et al., 2019), PNAS (Liu et al., 2018), and Efficient NAS (ENAS) (Jin et al., 2019). NAS has been applied in the literature for training personalised models in various domains of application. These include personalised human pose estimation, efficient object recognition, and heart rate estimation from faces, among others. For instance, Xu et al. (2021) proposed a novel NAS method, called ViP-NAS, to search networks in both the spatial and temporal domains for fast online video pose estimation. The approach of Chen et al. (2021) was based on a binarised neural architecture search (BNAS) framework, with a binarised convolution search space for efficient object recognition. In the work by Lu and Han (2021), facial regions of interest (ROI) defined based on facial landmarks were used to extract RGB-ROI temporal pulse signals, which were fed into a NAS architecture for heart rate estimation. In the field of personality computing, as mentioned above, Shao et al. (2021) employed NAS to learn and train person-specific models. To the best of our knowledge, this work is the first to tackle personalisation of personality computing models using NAS based on users profiles.

3. Proposed Approach

The overview of the proposed approach is shown in Figure 1. We first cluster the participants into two profiles and then use Neural Architecture Search (NAS) to automatically design a model for each profile to recognise their self-reported personality traits. We explore profiling users based on two characteristics, namely, gender (female and male), and age ($\leq$ 30 and $>$ 30). A separate network is then designed and trained with visual, audio, and text features. The final prediction is obtained by aggregating the results of video, audio and text modalities.

3.1. Dataset

In this work, we use the UDIVA (Understanding Dyadic Interactions from Video and Audio signals) dataset (Palmero et al., 2021) as it comprises a large number of synchronised multi-sensory, multi-view recordings collected in a face-to-face dyadic interaction scenario, together with demographics data (age, gender, and ethnicity) and self-reported Big Five personality trait scores.

The dataset contains 188 dyadic human-human interactions between 147 participants, resulting in 90.5 hours of recordings. The dataset is balanced with respect to gender (55.1% male) and the age of the participants ranges from 4 to 84 years old. Participants represent 22 different nationalities, with the majority coming from Spain (68%). The dominant speaking language is Spanish, followed by Catalan and English. Recordings take place in 5 different interaction contexts: (1) *Talk*: Participants talk about any subject during 5

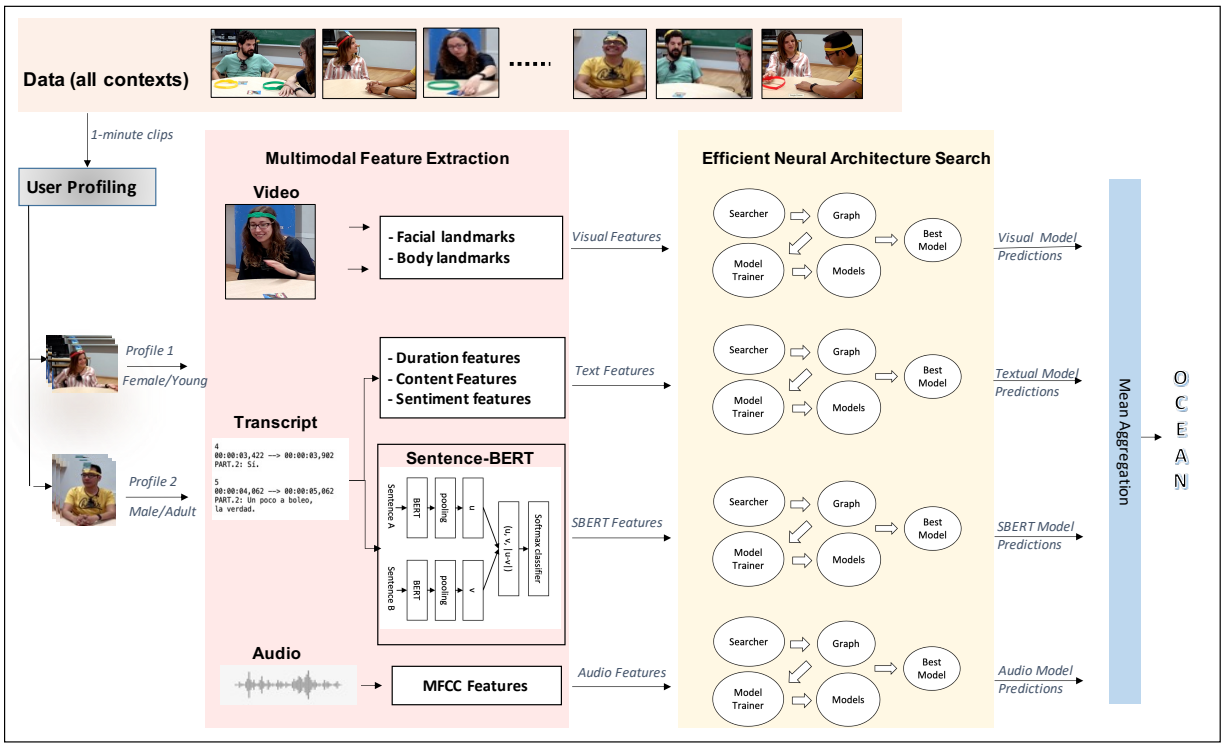

Figure 1: Overview of the proposed approach: The videos are first divided into 1 minute short clips and the participants are clustered into two profiles based on their gender or age, namely, female vs. male, or ≤ 30 vs. > 30. A set of visual, audio and text features are then extracted, which are given as input to the Neural Architecture Search (NAS) framework to automatically design a model for each profile and for each modality to recognise personality. For each user profile, the final prediction is obtained by aggregating the results of the different modalities.

minutes. (2) *Animals game*: Participants play a guessing game, where they have to ask 10 yes/no questions to guess the animal on their forehead (3 difficulty levels). (3) *Lego building*: Participants build a Lego together (4 difficulty levels). (4) *Ghost blitz card game*: Participants compete in a cards selection game. (5) *Gaze events*: Participants were directed to look at the other interactant's face, at an object, or somewhere else in the room, while performing head and eyes movements.

3.1.1. Data Pre-processing

Each participant's video and corresponding transcript and audio file were divided into 1 minute data slices. To split the videos into 1 minute video clips, the number of frames was considered. As the videos were recorded at 25 frames per second, 1500 frames corresponding to 1 minute video slices were extracted. The speech transcripts and audio files were divided into 1 minute transcripts based on the provided timestamps.

A single turn is presented in the speech transcript files as follows:

1
00:00:00,115 –¿ 00:00:01,865
PART.2: Preguntas
de "sí" o "no", ¿eh? Acuérdate.

Here, "1" in the 1^{st} line corresponds to the turn number. This is followed by timestamps of the start and end of the talk turn. The timestamp is in HH:MM:SS,SSS format. The MM value is used to split the transcripts per minute. If a dialogue spans across a minute boundary, it is assigned to the previous minute (when started). "PART.2" indicates that the talk turn corresponds to that of person 2 in the respective video. This value is used to split the dialogues in a person-specific manner.

3.2. Multimodal Feature Extraction

In this work, multimodal features are extracted from the video, audio, and text modalities. In the following, the feature extraction process is described in detail.

3.2.1. Video-based features

Both facial and body pose landmarks are considered for personality prediction. These video-based features were extracted from the annotations provided in the UDIVA dataset (Palmero et al., 2022), which were obtained in a semi-automatic manner. The features taken into consideration are:

Facial landmark statistics – 68 facial landmarks are provided for each video frame along 3 dimensions. The data was first flattened to obtain a facial landmark array of dimension 204. The mean and standard deviation is then computed for each facial landmark point over all the frames in a 1 minute video clip, resulting in a feature vector of 408.

Body landmark statistics – 24 three-dimensional body landmarks were provided for each video frame. The data was first flattened to obtain an array of dimension 72. The mean and standard deviation is then computed for each of the body pose landmark points over all the frames in each 1 minute video clip, resulting in a feature vector of 144.

Both the face and body landmarks statistics were concatenated, resulting in a feature vector of dimension 552 for each 1 minute video clip. This feature vector is referred to as FaceBody in the rest of the paper.

3.2.2. Text-based features

The transcripts of the interactions were analyzed based on each talk turn. The extracted features include talk turn duration, content, speech rate, and sentiment. Moreover, text embeddings were extracted for each talk turn using a sentence transformer model, namely, Sentence Bidirectional Encoder Representations from Transformers (SBERT).

Talk turn duration – The duration of the interaction for each person in a single minute was analyzed to generate a 5 dimensional feature set consisting of the following:

- Minimum turn duration: the minimum time (at the turn level) for which a person talked.
- Maximum turn duration: the maximum time (at the turn level) for which a person talked.
- Average turn duration: the average time across all turns for a particular person in a single minute.
- Standard deviation of turn duration: The standard deviation of the time taken in each turn for a single person over a 1 minute segment. This gives an idea of the variation in the time spent on different interactions.
- Total duration of turns: The total amount of time a person spoke in a single minute.

Talk turn content – The number of turns and the content of every dialogue were analyzed and 5 features were generated, which consist of the following:

- Turn percentage: the percentage of turns for a particular person out of the total number of turns in a single minute.
- Average words per turn: the average number of words spoken by a person in a turn across a 1 minute window.
- Longest turn: the largest number of words among all the turns over a minute for a particular person.
- Total number of words: the total number of words uttered over all the turns in a minute for a particular person.
- Standard deviation of words per turn: the standard deviation of the number of words per turn was computed to quantify the variance of the amount of vocal interaction by a particular person over a minute.

Talk turn sentiment – Each of the 1 minute transcripts was analyzed to generate 10 sentiment-based features. Since the majority of the conversations was in Spanish (68%) (Palmero et al., 2021), a Spanish sentiment recognition library was used (Bello, 2021). Moreover, to the best of our knowledge, there is no sentiment recognition system for Catalan language. The generated sentiment values ranged between 0 and 1, where 0 corresponds to fully negative and 1 corresponds to fully positive sentiment. The following sentiment-based features were computed:

- Most negative turn: The sentiment of texts across the turns over a minute for each person was computed and the smallest value was extracted.
- Most positive turn: The sentiment of texts across the turns over a minute for each person was computed and the largest value was extracted.
- Average sentiment: The average sentiment over all the turns in one minute was computed.
- Sentiment variation: The standard deviation of the sentiment values across the turns for a person over a minute was computed. This gives an idea of the variation of sentiment over successive expressions.
- Sentiment range: The sentiment range was divided into 5 equi-spaced classes corresponding to highly negative, negative, almost neutral, positive, and highly positive. The number of turns across a minute over these classes was computed and then normalized with the total number of turns by the person in that particular minute. This resulted in a 5-dimensional feature vector.
- Overall sentiment: The sentiment value was computed over the 1 minute segment per person, without dividing into turns.

Talk turn speech rate – The speech rate statistics of every talk turn was computed, resulting in 5 features:

- Minimum speech rate: the minimum speech rate (at the turn level) for which a person talked.
- Maximum speech rate: the maximum speech rate (at the turn level) for which a person talked.
- Average speech rate: the average speech rate across all turns for a particular person in a single minute.
- Standard deviation of speech rate: the standard deviation of the speech rate taken in each turn for a single person over a 1 minute segment.
- Total speech rate: the total speech rate of a person's utterance in a single minute.

SBERT – In addition to the high-level features described above, we use a Bidirectional Encoder Representations from Transformers (BERT)-based model called sentence-transformer (Reimers and Gurevych, 2019) to extract text embeddings. Sentence-BERT (SBERT) is a modification of the pretrained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings, which can be compared using cosine-similarity. This reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to approximately 5 seconds with SBERT, while maintaining the accuracy of BERT. The weights are pretrained for multi-lingual sentence embeddings (Reimers and Gurevych, 2020) with the main advantage of aligned feature space. More explicitly, vector spaces are aligned across languages, i.e., identical sentences in different languages are close. Vector space properties in the original source language from the teacher model are adopted and transferred to other languages.

3.2.3. Audio-based features

We extract Mel-frequency cepstral coefficients (MFCCs) as audio features, which are commonly used in the area of personality computing (Celiktutan and Gunes, 2017). The spectral envelope of an audio signal possesses a particular shape which models the perceived sounds by humans (Warren et al., 2005). While the actual sound is linear, the human auditory system does not perceive pitch linearly (Rao and Manjunath, 2017). MFCC features are designed to model the audio cepstrum energies in a non-linear scale known as the mel-scale. MFCC coefficients represent an estimation of the speech tone variation (Janse et al., 2014), which is highly correlated with emotional speech. The expression of emotions in speech differs among different personality traits. Using the timestamps from the speech transcripts, we separate the speech corresponding to each participant in the interaction. MFCC features are then computed over each one minute audio clip. A frame size of 256 points with 100 points overlap is used. We extract n dimensional MFCC per frame for each of the audio signals ($MFCC_n$, $n = 20$). Each dimension, $MFCC_n$, is termed as a frequency band (B). For instance, $MFCC_n$ has n frequency bands, where B_i corresponds to the i^{th} frequency band of an n dimensional MFCC. The frequency bands $B_1, B_2, \ldots, B_n$ are then analysed to find the energy distribution $E_1, E_2, \ldots, E_n$ across each of them. Energy for i^{th} band is presented below for N frames where C_{ik} corresponds to the MFCC of the k^{th} frame for the i^{th} band:

$$E_i = \sum_{k=1}^{N} C_{ik} \tag{1}$$

The resulting energy sequence $E_1, E_2, \ldots, E_n$ is used to generate the transformed order of bands $BO_1, BO_2, \ldots, BO_n$, which are used as a feature. The transformed order of bands is generated by simply sorting the energy sequence in descending order, and getting their index. In addition, the standard deviations and the median values of the MFCC values of each frequency band are computed and used as features.

3.3. Personalised Neural Architecture Search Strategy

In order to train personalised personality prediction models, we created different profiles by grouping the individuals in the dataset into different user profiles: gender-wise (females

vs. males), and age-wise (≤ 30 vs. > 30). An adaptive neural architecture was designed automatically with Neural Architecture Search (NAS) and trained for each profile.

We used the NAS framework proposed by Jin et al. (2019) due to its efficiency. The approach employs an efficient training during search via network morphism, which keeps the functionality of a neural network while changing its neural architecture through morphism operations (e.g., inserting a layer). The framework uses Bayesian optimization to guide the network morphism to enable efficient neural architecture search. To this end, a neural network kernel based on edit distance was designed and an algorithm was proposed to optimise the acquisition function in a tree-structured space. The algorithm was also implemented in an open-source AutoML system called Auto-Keras (Jin et al., 2019).

3.3.1. Implementation details

We used Auto-Keras (Jin et al., 2019) to perform the neural architecture search. For all modalities, visual, audio, and text, we used two dense layers with 32 units as the default architecture. The original training data was divided into training and validation sets using an $85-15\%$ split strategy. The best models were searched by employing network morphism operations such as inserting new layers, expanding existing layers, or adding skip connections. We use the Mean Squared Error as loss function. Each network was trained with ADAM optimiser. The number of epochs was set to 1000. The number of trials was set to 100. Finally, we used an early stopping with patience equal to 30.

3.4. Decision Fusion

We applied decision fusion to predict the personality of an individual. The scores obtained per minute were averaged over all the sessions. Thereafter, the resulting values were aggregated across different modalities using the average predictions.

4. Experimental Results

4.1. Evaluation Metric

We computed the prediction accuracy for each personality trait using the Mean Squared Error (MSE). To obtain an average accuracy over all personality traits, we used the average MSE (AMSE). Let $p_{i,j}$, $(1 \leq i \leq 5)$ be the Big-Five personality trait prediction scores and $g_{i,j}$ the associated ground truth label for each sample (j). The AMSE is defined as:

$$AMSE = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{5}\sum_{i=1}^{5}(p_{i,j} - g_{i,j})^2 \tag{2}$$

where N is the number of samples in the test set.

4.2. Performance Evaluation

Once we learn the individual neural architectures per profile and per feature modality, we train our models on the training and validation sets provided in (Palmero et al., 2022). For inference, we use the trained models on the unseen test set partition as given in (Palmero et al., 2022).

4.2.1. Performance evaluation of different modalities and their fusion

Table 1 presents the comparison of the proposed approach for different feature modalities individually as well as the fusion of feature modalities for each personality trait. Unimodal features (FaceBody, Speech Rate, Sentiment, Content, Duration, Audio, and SBERT) were used individually to train generic as well as gender-wise and age-wise personalised models. Moreover, models were trained on the early fusion of features obtained from the textual talk turn modality (Speech Rate, Sentiment, Content, Duration) with and without the sentiment feature. The evaluation of the talk turn features excluding the sentiment feature was performed since we suspected that the sentiment feature might decrease the performance. This is due to the fact that we used a Spanish sentiment recogniser to extract the talk turn sentiment of the dataset that includes English and Catalan languages. Two decision level fusion strategies were performed: (1) fusion of the predictions obtained from the two best performing modalities; and (2) fusion of all modalities where the average of the predictions from each of the modalities (FaceBody, Speech Rate, Sentiment, Content, Duration, Audio, and SBERT) was computed to obtain a final prediction. Finally, we compared the AMSE of these fusion strategies to the AMSE obtained by taking the predictions of the best performing model for each personality trait.

Looking at Table 1, the personalised models improved the performance as compared to the generic model in general. Age-wise personalisation performed slightly better than the non-personalized model for both fusion strategies, as well as compared to the best unimodal strategy. The fusion of two best performing modalities outperformed other strategies with an AMSE of 0.796. Gender-wise personalisation performed better than the age-wise models. The bimodal fusion strategy decreased the AMSE by 0.106 and 0.128 with respect to the age-wise model and the generic model, respectively.

Regarding the prediction of individual personality traits, we observed that there was no consistency among the best performing modalities across the different personalisation strategies. For instance, for the Openness trait, Duration worked best in the case of non-personalisation and gender-wise personalisation, while SBERT outperformed the other modalities for the age-wise personalisation. For Consciousness, SBERT performed the best using non-personalisation, while the Talk Turn features without sentiment performed the best in the gender-wise personalisation, and the Speech Rate yielded the smallest AMSE in the age-wise personalisation. For the remaining traits, the best performing modalities across personalisations are as follows: Extroversion (generic: Talk Turn without sentiment, gender-wise: Duration, age-wise: FaceBody), Agreeableness (generic: Content, gender-wise: SBERT, age-wise: Content), Neuroticism (generic: Speech Rate, gender-wise: Sentiment, age-wise: SBERT). In line with existing work in personality computing such as (Celiktutan and Gunes, 2017), our results showed that the most informative features changed from one trait to another but also across different user profiles.

4.2.2. Comparison with the state-of-the-art

Table 2 presents the comparison of the approaches of the teams that partook in the ChaLearn LAP Challenge on Understanding Social behaviour in Dyadic and Small Group Interactions (DYAD) at ICCV 2021 - Automatic Self-reported Personality Recognition

Table 1: Performance evaluation of the proposed approach using different modalities and their fusion. **Bold**: best performance; Underline: second best performance. WS: With Sentiment; No Sentiment: NS.

Personalisation	Feature	O	C	E	A	N	AMSE
No	**FaceBody**	0.748	1.027	1.069	2.665	1.118	1.325
	Talk Turn (NS)	0.833	0.807	**0.909**	0.689	1.130	0.874
	Talk Turn (WS)	0.854	0.809	1.014	0.658	1.373	0.942
	Speech Rate	0.864	0.912	1.039	0.664	**1.067**	0.909
	Sentiment	1.337	0.808	0.959	0.742	1.349	1.039
	Content	1.389	0.826	0.952	**0.606**	1.163	0.987
	Duration	**0.723**	0.832	1.180	0.672	1.230	0.927
	Audio	0.740	0.976	0.968	0.608	1.309	0.920
	SBERT	1.07	**0.796**	0.975	0.677	1.332	0.970
	Best of each modality	**0.723**	0.796	**0.909**	0.606	**1.067**	0.820
	Fusion of best two modalities	0.731	**0.768**	0.926	**0.575**	1.088	**0.818**
	Fusion of all modalities	0.850	0.839	0.962	0.641	1.183	0.895
Gender	**FaceBody**	0.727	0.765	0.846	0.604	0.965	0.781
	Talk Turn (NS)	1.141	**0.563**	0.887	0.624	1.005	0.844
	Talk Turn (WS)	0.712	0.691	0.877	0.688	0.953	0.784
	Speech Rate	0.732	0.666	0.886	0.605	0.923	0.762
	Sentiment	1.235	0.745	0.883	0.644	**0.874**	0.876
	Content	0.715	0.630	0.925	0.668	1.053	0.798
	Duration	**0.681**	0.745	**0.824**	0.670	1.075	0.799
	Audio	0.749	1.111	0.979	0.593	1.208	0.928
	SBERT	0.711	0.826	0.985	**0.535**	0.879	0.787
	Best of each modality	**0.681**	**0.563**	**0.824**	**0.535**	0.874	0.695
	Fusion of best two modalities	0.684	0.588	0.830	0.550	**0.796**	**0.690**
	Fusion of all modalities	0.699	0.758	0.883	0.600	0.905	0.769
Age	**FaceBody**	0.799	0.823	**0.824**	0.836	1.554	0.967
	Talk Turn (NS)	0.976	0.728	0.838	0.710	1.161	0.883
	Talk Turn (WS)	1.059	1.352	1.00	0.766	1.193	1.074
	Speech Rate	0.957	**0.711**	0.840	0.741	1.135	0.877
	Sentiment	0.926	0.719	0.878	0.726	1.075	0.865
	Content	0.952	0.988	1.031	0.782	1.210	0.993
	Duration	0.892	0.801	0.968	**0.682**	1.189	0.906
	Audio	0.937	1.18	0.957	0.800	1.143	1.003
	SBERT	**0.769**	0.758	0.873	0.731	**1.053**	0.837
	Best of each modality	0.769	0.711	0.824	**0.682**	**1.053**	0.808
	Fusion of best two modalities	**0.716**	**0.705**	**0.763**	0.695	1.100	**0.796**
	Fusion of all modalities	0.828	0.764	0.857	0.738	1.128	0.863

Table 2: Comparison of the proposed approach against other approaches presented in the challenge (Palmero et al., 2022). Our previous approach presented in the challenge was the fusion of the models trained on FaceBody features and Talk Turn features (including sentiment) only.

Team	O	C	E	A	N	AMSE
SMART-SAIR (Ours)	**0.711 (1)**	0.723 (3)	**0.867 (1)**	**0.548 (1)**	**0.997 (1)**	**0.769 (1)**
Baseline	0.744 (2)	0.794 (4)	0.886 (2)	0.653 (2)	1.012 (2)	0.818 (2)
FGM Utrecht	0.752 (3)	0.687 (2)	0.917 (3)	0.671 (3)	1.098 (3)	0.825 (3)
FGM Utrecht	0.759 (4)	**0.677 (1)**	0.955 (4)	0.677 (4)	1.163 (4)	0.846 (4)
STARS Inria	0.839 (5)	0.976 (5)	1.359 (5)	0.864 (5)	1.252 (5)	1.058 (5)
Crisie Lab	4.401 (6)	4.671 (6)	1.998 (6)	3.534 (6)	5.523 (6)	4.025 (6)
Dyadformer (Curto et al., 2021)	–	–	–	–	–	0.722
Gender-wise Bimodal NAS	**0.684**	**0.588**	**0.830**	0.550	**0.796**	**0.690**

Track[1], including our presented solution (Salam et al., 2021), which was ranked first in the challenge. For further details regarding the challenge, please refer to (Palmero et al., 2022). Our approach presented in the challenge was the fusion of the models trained on the FaceBody features and Talk Turn features (including sentiment), but not on the audio features. We also compared our best performing approach presented in Table 1 (aka Gender-wise Bimodal NAS) with the Dyadformer proposed by Curto et al. (2021).

The table shows that our approach (Gender-Wise Bimodal NAS) outperforms that of Curto et al. (2021) by reducing the AMSE from 0.722 to 0.690. Gender-wise Bimodal NAS performs better than our previous approach (winning solution in the challenge) by a margin of 0.079 in overall (i.e, the AMSE decreases from 0.769 to 0.690), as well as surpasses all other approaches for all personality traits except for Agreeableness, where our previous approach performs better slightly (by a margin of 0.002).

4.3. Analysis of Automatically Designed Architectures

Table 3 summarises the characteristics (number of parameters and layers) of the automatically designed architectures for the best performing modality for each personality trait. Even though the multimodal models performed better than the unimodal ones, they did not exhibit a unified architecture, which made the analysis of the architecture characteristics non-trivial. Nevertheless, our investigation revealed that designed personalised architectures significantly differed from each other as well as from generic models. However, we were not able to observe any trend in the automatically designed architectures considering neither the different personality traits, nor the different user profiles. We conjectured that this could be due to the fact that different network architectures may be obtained even with the same data as the network was initialised by different weights each time.

Table 4 further presents network architectures for each personality trait. For the Extroversion dimension, as the best performing modalities are both FaceBody (age-wise personal-

1. https://competitions.codalab.org/competitions/31326

Table 3: Characteristics of the automatically designed architectures for the best performing modalities for each personality dimension (number of parameters (P) & layers (L)). The best performing personalisation architecture is indicated in **bold**. NS: No Sentiment

		No		**Gender**				**Age**			
				Female		**Male**		**≤ 30**		**> 30**	
Label	**Feature**	**P**	**L**	**P**	**L**	**P**	**L**	**P**	**L**	**P**	**L**
O	**Duration**	78,348	13	**2,892**	**9**	**3,745**	**14**	36,481	10	142,092	10
C	**Talk Turn (NS)**	576	6	**3,168**	**8**	**1,057**	**8**	576	5	1,601	9
E	**Duration**	9,804	9	**1,292**	**7**	**236**	**6**	1,292	8	1,292	8
	FaceBody	302,866	11	18,962	6	142,930	7	**19,958**	**8**	**38,742**	**7**
A	**SBERT**	16,449	5	**281,281**	**9**	**17,761**	**9**	70,850	9	34,850	8
N	**Sentiment**	25,622	10	**1,718**	**9**	**51,126**	**13**	1,718	10	774	10

Table 4: Automatically designed architectures for the best performing modality for each personality trait. IL: Input Layer, ME: Multi-Category Encoding, N: Normalization, BN: Batch Normalization, RL: ReLU, D: Dense, DO: Dropout. The numbers within the brackets represent the layer's output shape.

O		**C**		**E**		**A**		**N**	
Duration		**Talk Turn (NS)**		**FaceBody**		**SBERT**		**Sentiment**	
Female	**Male**	**Female**	**Male**	**≤ 30**	**> 30**	**Female**	**Male**	**Female**	**Male**
IL (5)	IL (5)	IL (15)	IL (15)	IL (554)	IL (554)	IL (512)	IL (512)	IL (10)	IL (10)
ME (5)	ME (5)	ME (15)	ME (15)	ME (554)	ME (554)	ME (512)	ME (512)	ME (10)	ME (10)
N (5)	D (16)	N (15)	D (32)	N (554)	N (554)	D (512)	D (32)	N (10)	N (10)
D (64)	BN (16)	D (64)	RL (32)	D (32)	D (64)	BN (512)	BN (32)	D (32)	D (256)
BN (64)	RL (16)	RL (64)	DO (32)	RL (32)	RL (64)	RL (512)	RL (32)	BN (32)	BN (256)
RL (64)	DO (16)	D (32)	D (16)	D (32)	D (32)	D (32)	D (32)	RL (32)	RL (256)
D (32)	D (64)	RL (32)	RL (16)	RL (32)	RL (32)	BN (32)	BN (32)	D (32)	D (32)
BN (32)	BN (64)	DO (32)	DO (16)	DO (32)	D (1)	RL (32)	RL (32)	BN (32)	BN (32)
RL (32)	RL (64)	D (1)	D (1)	D (1)		DO (32)	DO (32)	RL (32)	RL (32)
D (1)	DO (64)					D (1)	D (1)	D (1)	D (1024)
	D (32)								BN (1024)
	BN (32)								RL (1024)
	RL (32)								DO (1024)
	DO (32)								D (1)
	D (1)								

isation) and Duration (gender-wise personalisation), we provide the FaceBody architecture only due to space restrictions. Looking at the table, one trend is that the male architectures are deeper than the female ones for the Openness and Neuroticism dimensions, which is the case for most of the features, including the worse performing ones. To gain further insights, one promising direction would be to use tools for explaining and interpreting network decisions, which has been left as a future work.

5. Discussion and Limitations

The proposed system has several advantages. It combines multiple modalities to predict the personality of an individual. The system is scalable and can adapt itself to changing trends in the data as the neural architecture search-based approach enables generation of a deep learning model depending on the user profile. Yan et al. (2020) showed that bias exists in self-reported personality recognition systems, and they demonstrated biases from different modalities and data fusion strategies. To mitigate bias, the authors proposed data balancing and adversarial learning strategies. We argue that our approach inherently deals with the unbalanced distribution of data across gender and age and takes into account the differences across different populations by creating and learning a separate model for different user profiles. However, further experiments are needed to conclude whether training personalised models is a better approach for achieving fairness as compared to traditional strategies such as data balancing, which is an interesting future research direction.

While advantageous for the above-mentioned viewpoints, the proposed approach has two main limitations. First limitation is that the system requires the meta-data regarding gender and age to assign a user to a certain profile and make inference about their personality. However, as reviewed by a recent survey (Di Mascio et al., 2022), there are many off-the-shelf-methods for age and gender recognition, which can be used as input to the proposed system. Second limitation is that our approach assumes the gender is binary. However, gender is highly complex construct, and existing gender diversity needs to be taken into account to ensure fairness (Lindqvist et al., 2021).

6. Conclusion

In this paper, we have presented an approach for personalising personality recognition models. The proposed approach automatically learns neural architectures for different user profiles using NAS to predict the Big Five personality trait scores from multimodal behavioural features. Two personalisation criteria are tested: gender-wise and age-wise. We further propose a decision level fusion strategy, which combines the prediction results of the two best performing modalities via mean aggregation. The proposed approach outperforms the-state-of-the-art approaches evaluated on the UDIVA dataset (Palmero et al., 2021).

In our current approach, the models are separately trained on 1 minute segments for each modality and after their predictions are aggregated. As a future work, we plan to apply more sophisticated techniques and explore feature level fusion strategies to train a single model that can leverage all the data pertaining to an individual for predicting their personality.

Acknowledgments

The authors would like to thank Iman Ismail and Himadri Mukherjee for their help in extracting audio and textual features. The work of Oya Celiktutan was supported by the "LISI - Learning to Imitate Nonverbal Communication Dynamics for Human-Robot Social Interaction" Project, funded by Engineering and Physical Sciences Research Council (Grant Ref.: EP/V010875/1).

References

Miray Akyunus, Tülin Gençöz, and B Türküler Aka. Age and sex differences in basic personality traits and interpersonal problems across young adulthood. *Current Psychology*, 40(5):2518–2527, 2021.

Dane Archer. Unspoken diversity: Cultural differences in gestures. *Qualitative sociology*, 20(1):79–105, 1997.

Süleyman Aslan, Uğur Güdükbay, and Hamdi Dibeklioğlu. Multimodal assessment of apparent personality using feature attention and error consistency constraint. *Image and Vision Computing*, 110:104163, 2021.

Hugo J. Bello. sentiment-analysis-spanish 0.0.25, 2021. URL https://pypi.org/project/sentiment-analysis-spanish/.

Oya Celiktutan and Hatice Gunes. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing*, 8(1):29–42, 2017. doi: 10.1109/TAFFC.2015.2513401.

Hanlin Chen, Baochang Zhang, Xiawu Zheng, Jianzhuang Liu, Rongrong Ji, David Doermann, Guodong Guo, et al. Binarized neural architecture search for efficient object recognition. *International Journal of Computer Vision*, 129(2):501–516, 2021.

David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B Moeslund, et al. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2177–2188, 2021.

Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. A survey on personality-aware recommendation systems. *arXiv preprint arXiv:2101.12153*, 2021.

Tania Di Mascio, Paolo Fantozzi, Luigi Laura, and Valerio Rughetti. Age and gender (face) recognition: A brief survey. In Fernando De la Prieta, Rosella Gennari, Marco Temperini, Tania Di Mascio, Pierpaolo Vittorini, Zuzana Kubincova, Elvira Popescu, Davide Rua Carneiro, Loreto Lancia, and Agnese Addone, editors, *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference*, pages 105–113, Cham, 2022. Springer International Publishing. ISBN 978-3-030-86618-1.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874246. URL https://doi.org/10.1145/1873951.1874246.

Sheng Fang, Catherine Achard, and Séverine Dubuisson. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 225–232, 2016.

Michael Feffer, Rosalind W Picard, et al. A mixture of personalized experts for human affect estimation. In *International conference on machine learning and data mining in pattern recognition*, pages 316–330. Springer, 2018.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Auto-sklearn: efficient and robust automated machine learning. In *Automated Machine Learning*, pages 113–134. Springer, Cham, 2019.

Peggy E Gallaher. Individual differences in nonverbal behavior: Dimensions of style. *Journal of personality and social psychology*, 63(1):133, 1992.

Yuzhan Hang, Christopher J Soto, Lydia Gabriela Speyer, Liina Haring, Billy Lee, Fritz Ostendorf, and René Mõttus. Age differences in the big five personality domains, facets and nuances: A replication across the life span. 2021.

Pooja V Janse, S Magre, P Kurzekar, and R Deshmukh. A comparative study between mfcc and dwt feature extraction technique. *International Journal of Engineering Research and Technology*, 3(1):3124–3127, 2014.

Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Multi-task learning for predicting health, stress, and happiness. In *NIPS Workshop on Machine Learning for Healthcare*, 2016.

Mikael Jensen. Personality traits and nonverbal communication patterns. *Int'l J. Soc. Sci. Stud.*, 4:57, 2016.

Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1946–1956. ACM, 2019.

Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier, and Sergio Escalera. First impressions: A survey on vision-based apparent personality trait analysis, 2019.

Anna Lindqvist, Marie Gustafsson Sendén, and Emma A. Renström. What is gender, anyway: a review of the options for operationalising gender. *Psychology & Sexuality*, 12(4):332–344, 2021. doi: 10.1080/19419899.2020.1729844. URL https://doi.org/10.1080/19419899.2020.1729844.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

Dianbo Liu, Peng Fengjiao, Rosalind Picard, et al. Deepfacelift: interpretable personalized models for automatic estimation of self-reported pain. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 1–16. PMLR, 2017.

Hao Lu and Hu Han. Nas-hr: Neural architecture search for heart rate estimation from face videos. *Virtual Reality & Intelligent Hardware*, 3(1):33–42, 2021.

Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

Robert R McCrae, Paul T Costa, Margarida Pedroso de Lima, António Simões, Fritz Ostendorf, Alois Angleitner, Iris Marušić, Denis Bratko, Gian Vittorio Caprara, Claudio Barbaranelli, et al. Age differences in personality across the adult life span: parallels in five cultures. *Developmental psychology*, 35(2):466, 1999.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–12, 2021.

Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, Albert Clapés, Johnny Núñez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, David Gallardo-Pujol, Georgina Guilera, David Leiva, Feng Han, Xiaoxue Feng, Jennifer He, Wei-Wei Tu, Thomas B. Moeslund, Isabelle Guyon, and Sergio Escalera. Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research (PMLR)*, pages 4–52, 2022.

K Sreenivasa Rao and KE Manjunath. *Speech recognition using articulatory and excitation source features*. Springer, 2017.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *arXiv preprint arXiv:2006.02903*, 2020.

B. W. Roberts, K. E. Walton, and W. Viechtbauer. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1):1–25, 2006.

Marta Romeo, Daniel Hernández García, Ting Han, Angelo Cangelosi, and Kristiina Jokinen. Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction. *Advanced Robotics*, pages 1–13, 2021.

Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19), 2018a.

Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W Picard. Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 339–346. IEEE, 2018b.

Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721, 2016.

Hanan Salam, Oya Celiktutan, Viswonathan Manoranjan, Iman Ismail, and Himadri Mukherjee. Iccv 2021 understanding social behavior in dyadic and small group interactions challenge fact sheet: Automatic self-reported personality recognition track. 2021.

Mostafa Shahabinejad, Yang Wang, Yuanhao Yu, Jin Tang, and Jiani Li. Toward personalized emotion recognition: A face recognition based attention method for facial emotion recognition. In *Proceedings of IEEE International Conference on Face & Gesture*, 2021.

Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. Personality recognition by modelling person-specific cognitive processes using graph representation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 357–366, 2021.

Julio Cezar Silveira Jacques Junior, Yağmur Güçlütürk, Marc Perez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. Van Gerven, Rob Van Lier, and Sergio Escalera. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. doi: 10.1109/TAFFC.2019.2930058.

Siyang Song, Shashank Jaiswal, Enrique Sanchez, Georgios Tzimiropoulos, Linlin Shen, and Michel Valstar. Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *IEEE Transactions on Affective Computing*, 2021.

Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.

Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014a. doi: 10.1109/TAFFC.2014.2330816.

Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014b.

Joe D Warren, AR Jennings, and Timothy D Griffiths. Analysis of the spectral envelope of sounds by the human brain. *Neuroimage*, 24(4):1052–1057, 2005.

Yanna Weisberg, Colin DeYoung, and Jacob Hirsh. Gender differences in personality across the ten aspects of the big five. *Frontiers in Psychology*, 2:178, 2011a. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00178. URL https://www.frontiersin.org/article/10.3389/fpsyg.2011.00178.

Yanna J Weisberg, Colin G DeYoung, and Jacob B Hirsh. Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 2:178, 2011b.

Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16072–16081, 2021.

Xia Xue, Jun Feng, and Xia Sun. Semantic-enhanced sequential modeling for personality trait recognition from texts. *Applied Intelligence*, pages 1–13, 2021.

Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 361–369, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375818. doi: 10.1145/3382507.3418889. URL https://doi.org/10.1145/3382507.3418889.

Le Zhang, Songyou Peng, and Stefan Winkler. Persemon: a deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing*, 2019.