POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Deep learning classification of Big Five personality traits starting from EEG signals

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Veronika Guleva

Advisor: Prof. Anna Maria Bianchi

Co-advisor: Alessandra Calcagno

Academic year: 2021-2022

1. Introduction

Personality represents the individual differences in behavior, emotion, and cognition that are a consequence of genetics and environmental influences. Personality psychology has developed a series of attributes, called traits, that successfully sum up these individual differences. The most influential and widely accepted paradigm in personality research is the Big Five model [1] which identifies five stable traits: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness. Personality traits are traditionally assessed using scaled self-reported questionnaires, which are intrinsically subjective and prone to bias. For this reason, in recent years, the interest for an automatic assessment of personality has emerged. Promising opportunities for objective personality inference, are found in brain imaging techniques such as magnetic resonance imaging (MRI) and electroencephalography (EEG). Most of the neuroscientific research has however concentrated on trying to find the biological bases of personality and few attempts have been made to develop practical methods for its actual assessment. Concerning personality classification, EEG signals seem to be the optimal choice compared to other techniques, due mainly to their high temporal resolution and especially to the low cost and little invasive instrumentation required for their acquisition. Literature suggests that, rather than from EEG baseline activity, personality might be inferred from situation-dependent responsiveness. Thus, recent research has concentrated on personality assessment starting from EEG signals recorded in response to different stimuli, such as videos inducing affective emotions. Specifically, most studies available in the literature have approached the EEG-based classification task using traditional processing and feature extraction methods. The main limitation of this approach is related to the need of a priori selecting the features to be employed for classification. In this scenario, deep learning methods, such as convolutional neural networks (CNN), could be adopted in order to automatically identify the more representative features, even starting from raw signals. To the best of our knowledge, literature lacks studies focused on deep learning techniques on EEG signals for the classification of personality.

The purpose of the present work is to develop a deep learning-based binary personality classification method starting from EEG data collected

in the public dataset AMIGOS [2]. EEGNet [3], a state-of-the-art CNN model designed for EEG decoding, was adopted and 5 independent EEGNet models were trained, one for each trait. The optimal structure and hyperparameters of the model were validated and different levels of preprocessing of the EEG data are tested to evaluate the model's performance on noisy signals. Finally, an evaluation of the automatically extracted features was performed.

2. Materials and methods

2.1. AMIGOS dataset

For the present study, the public AMIGOS dataset [2] was used. It provides EEG signals acquired on 40 participants (male = 27, female = 13, aged 21-40 years, mean age = 28.3), after stimulating their affective response through selected videos, 16 short emotional videos with duration of from 1 to 2 minutes, and 4 long videos with duration from 15 to 20 minutes. The EEG signals were recorded with the EPOC neuroheadset (Emotiv, U.S.A), using 14 channels and a sampling rate of 128 Hz. Personality traits were assessed through an online questionnaire filled in by each participant. The considered personality model was the Big-Five personality traits model while the form used was the Big-Five Marker Scale (BFMS) questionnaire. For each of the five personality classes (i.e., Extroversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (ES) and Openness (O)), a 1 to 7 score was attributed to each trait.

2.2. Data processing

2.2.1 EEG pre-processing

The present work was focused on the short-video data. Out of the 40 subjects, 2 were discarded due to missing self-assessed personality information. To test the model performances on different types of input data, three datasets (D1, D2 and D3) were obtained by applying different pre-processing approaches on the EEG traces. For D1, a full pre-processing pipeline was applied. Specifically, EEG traces were firstly bandpass filtered between 0.1 and 45 Hz. Then, bad channels were visually inspected and removed. Eye blinks and other artefactual sources were removed by means of the Independent Component Analysis (ICA). Finally, the eliminated channels were interpolated, and the signal was re-referenced to the common average. On D2, only bandpass filtering in the frequency range 0.1 - 45 Hz was applied. D3, instead, was bandpass filtered between 4 and 45 Hz, to remove the Delta band and get rid of most of the low-frequency noise. Both D2 and D3 were standardized by subtracting the mean and scaling to unit variance. These two datasets were poorly pre-processed in order to test the model performance on raw data.

Class	Personality Trait				
	(E)	**(A)**	**(C)**	**(ES)**	**(O)**
0	21	19	17	18	18
1	17	19	21	20	20

Table 1: Binary class counts for each personality trait.

2.2.2 EEG segmentation

Each EEG trial refers to an EEG signal acquired during the presentation of a single short video. To be able to train the CNN model, all trials were segmented with sliding windows of 3 seconds length with no overlap. This kind of cropped strategy has been found to be the most effective for CNN-EEG applications since it allows to produce more training samples and it forces the network to learn more generalized features.

2.2.3 Personality binarization

A binarization of the personality trait score was performed using the mean as separating threshold to form two classes: class 0, encompassing a low expression of the trait below threshold, and class 1, encompassing a high expression of the trait above threshold. The resulting classes counts are reported in Table 1. Agreeableness is the only trait that results perfectly balanced between the two classes, while the other classes present a slight imbalance.

2.3. EEGNeet model

The standard structure of EEGNet [3] is organized in three main blocks (Figure 1):

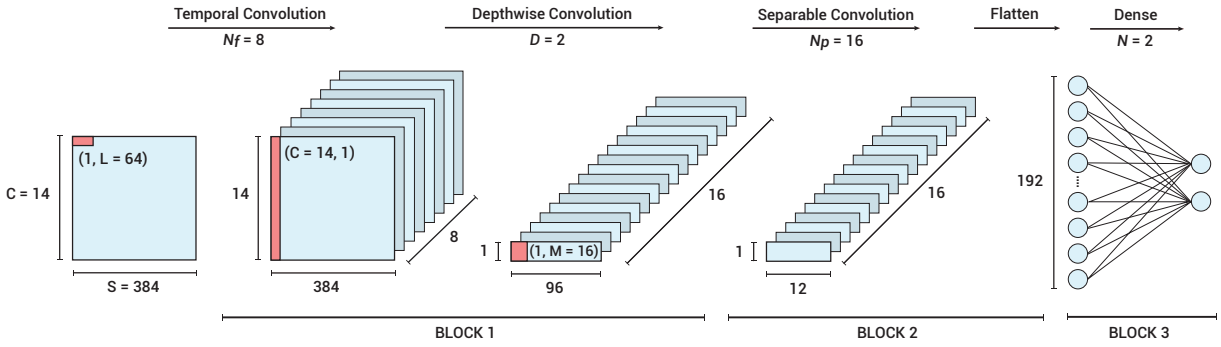

Figure 1: EEGNet standard structure.

1. N_f temporal filters of size (1, L) are convolved with the EEG traces, where L is the time length of the filter. Afterwards, the output of each temporal filter is convolved in space with D depthwise convolution filters of size (C, 1), where C is the number of channels of the EEG signal. Finally, batch normalization is applied before the exponential linear unit (ELU) nonlinearity activation. Dropout is also enabled to regularize the model. Finally, an average pooling layer is applied in order to reduce the sampling rate of the signal.
2. A depthwise convolution of size (1, M), where M is the length of the filter, followed by N_p pointwise convolutions of size (1, 1) are performed. As in block 1, batch normalization, ELU activation, dropout and average pooling are then performed.
3. A classification softmax layer with N units, corresponding to the number of classes, is applied.

The default number of temporal filters N_f is 8 and the number of spatial filters per feature map D is 2. The input shape (1, C, S) of the model was set to fit the data, where C = 14 is the number of EEG channels and S = 384 is the number of samples per window (3 seconds window x 128 Hz sampling rate).

2.4. Model validation

In order to find the optimal model configuration, the EEG and personality datasets were split into training, validation and test sets with a 70-15-15 proportion, respectively. Five random stratified splits, one for each personality trait, were implemented to keep a balanced representation of all the subjects within the sets. Model tuning was carried out on the training set and tested on the validation set, with the test set held as a holdout set for the classification task.

2.4.1 Hyperparameter tuning

The hyperparameters of the network were tuned using Keras-Tuner with the hyperband algorithm by maximizing the validation accuracy. All three datasets were used, while the personality trait used was Agreeableness because of its perfectly balanced classes. The following hyperparameters and value ranges were chosen:

- Dropout rate. Values between 0 and 0.7, with a step of 0.1, were evaluated.
- Dropout type. Choice between the two possible dropout layers: “Dropout” and “SpatialDropout2D”.
- Learning rate. Values between 10-3 and 10-5 with log sampling, were considered.

For the batch size, i.e., the number of EEG windows to pass to the network at once, practical evaluations were made by testing mini batches of 16, 32, 64, 128, 256 and 512 samples on D3.

2.4.2 Structure optimization

To test the optimal structure, the hyperband algorithm was applied on dataset D3, for the temporal and spatial filters parameters after fixing the hyperparameters to the optimal found in the previous step. The search space defined was:

- Temporal filters. Ranging from 2 to 12.
- Spatial filters. Ranging from 1 to 8.

Afterwards, a simple grid search was performed by fixing the number of spatial filters to 2 and testing the performance of the model by varying the number of temporal filters from 1 to 12 in order to compare the performance of the standard

structure with other similar structures.

2.5. Training strategy

For classification, the models were trained using the labeled EEG 3-second windows while a five-fold cross-validation strategy for assessing the performance and for splitting the dataset was performed. Specifically, a window-wise classification was implemented. The model classifies single EEG windows, and its performance is evaluated based on how well it can predict the class of the windows.
All models were trained on an NVIDIA GeForce RTX 2070 GPU in Tensorflow for 1000 epochs with an early stopping rule with patience of 20 epochs on the validation loss. The model weights that produced the best validation accuracy were saved and used to evaluate the models on the corresponding test set. Each EEGNet model was fit using the Adam optimizer.

2.6. Feature interpretability

The features extracted by the model were analyzed by visualizing: i) the frequency bands extracted by the temporal filters and ii) the topographical maps of the spatial filters weights. For this task, EEGNet models with 4 temporal filters and 2 spatial filters were chosen for an easier interpretation. The most relevant temporal filters were identified with an ablation study: different filters were deactivated one at a time and the resulting model was tested in its performance. The filters that alone accounted for a an accuracy and F1 score above chance classification, were considered as relevant.
Finally, an attribution algorithm, DeepLIFT [4], was applied on the best performing standard EEGNet model. DeepLIFT assigns a contribution value to each input EEG channel, based on how much that channel affects the final prediction. The channels with highest attribution values should be the ones identifying the brain areas that most represent the specific personality trait.

3. Results and discussion

3.1. Model validation

3.1.1 Hyperparameter tuning

For the dropout rate, the highest average accuracy was obtained on all three datasets for dropout values of 0.1. For the dropout type, the standard "Dropout" layer was selected unanimously for datasets D2 and D3, while for dataset D1, the "SpatialDropout2D" layer was selected about half of the times. The learning rate assumes several different values over the trials, but always in the order of $10^{-}5$. As for the batch size, it was found that the higher batch sizes of 256 and 512 samples converged to a more stable model and had lower variance in classification accuracy compared to lower batch sizes (Figure 2). This is most likely due to the fact that the EEG data of D3 are noisy, and the cropped windows are small. A batch size of 256 was chosen since it is both stable and achieves better accuracy than the model trained with a batch size of 512. The other hyperparameters chosen were a dropout rate of 0.1, dropout type "Dropout", and a learning rate of 0.0001, which maintains the $10^{-}5$ order.

3.1.2 Structure optimization

The first search among temporal filters in the range 2-12 and spatial filters in the range 1-8, showed that the models with the highest possible number of filters obtained the highest average accuracy. This result was expected since increasing the number of filters increases the number of trainable parameters and thus, the complexity of the model and its ability to fit the input data. However, increasing the number of parameters also increases the computational cost.
The second grid search performed by fixing the number of spatial filters to 2, showed that the performance tends to increases with the increase of the temporal filters until a plateau is reached. As such, the performance of a model with 4 temporal filters is comparable to the one with 8 (i.e., standard EEGNet structure) or more temporal filters as it can be seen in Figure 3.
Since the performance of EEGNet with number of temporal filters higher than 8 does not improve drastically, the standard structure was chosen for the final classification analysis.

3.2. Classification

Average five-fold test accuracy (Acc) and F1 results are reported in Table 2. The best performing models were the ones trained on dataset D3. The traits with higher accuracy are Agreeableness (0.93) and Extraversion (0.92) while the

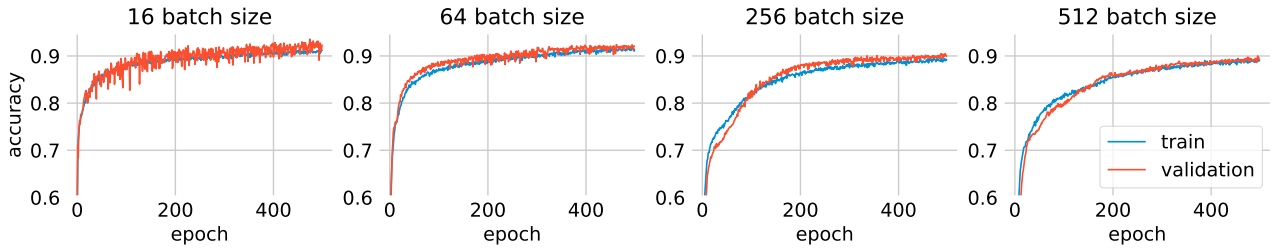

Figure 2: Training and validation curves comparison for batch sizes of 16, 64, 256 and 512.

Trait	D1		D2		D3	
	Acc	F1	Acc	F1	Acc	F1
Extraversion	0.76	0.75	0.86	0.86	**0.92**	**0.91**
Agreeableness	0.76	0.76	0.89	0.91	**0.93**	**0.93**
Conscientiousness	0.79	0.82	0.87	0.89	**0.90**	**0.91**
Emotional Stability	0.78	0.80	0.89	0.89	**0.89**	**0.90**
Openness	0.77	0.78	0.83	0.84	**0.89**	**0.89**

Table 2: Classification results for EEGNet-8,2.

rest of the traits report slightly lower accuracies. Interestingly, for D1, the pre-processed dataset, the accuracies are remarkably lower, in the range between 0.75 and 0.79. Compared to another study [5] on the same AMIGOS dataset, which classified personality by using brain connectivity features, the present study reports higher accuracies for all traits.

3.3. Feature interpretability

By the deactivation procedure, the temporal filters that resulted most relevant for personality classification are reported in Table 3. Two temporal filters for Extraversion, one for Agreeableness and one for Emotional Stability were identified as relevant. No relevant filters were identified for Conscientiousness and Openness. The interpretation of the frequency response of the temporal filters and of the topographic maps of the two corresponding spatial filter resulted difficult. The temporal filters do not extract clear frequency bands and, albeit some observed behavior was in line with some personality-EEG correlation studies, drawing a definitive conclusion on the most relevant features for personality prediction is not trivial.

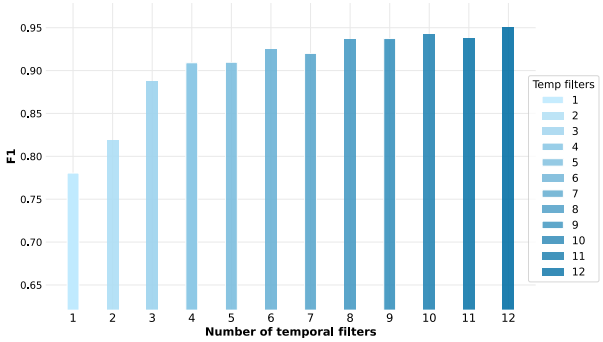

Figure 3: F1 score performance of EEGNet models with fixed $D = 2$ and varying number of N_f.

The attribution maps obtained with DeepLIFT [4], averaged for each trait over all subjects, are reported in Figure 4. It can be observed that the frontal area has high positive (red color) or negative (blue color) contribution for the prediction of all traits. This result is in line with neuroscientific studies on personality relating prefrontal brain activity to personality. Another relevant observation regards the occipital area that results as a negative contributor for the prediction of all traits. A possible explanation could be found in the fact that during the experimental procedure, subjects are intent in watching videos. Since the occipital lobe is associated to the visual cortex, activity is detected from the occipital electrodes and the network might have automatically learned to ignore its contribution as it is not relevant to personality.

4. Conclusions

In this work personality traits were classified starting from EEG signals. It was demonstrated that deep learning EEG-based classification can be a valid alternative to traditional classification methods based on manual feature extraction strategies as the main advantage of this approach stands in the ability of automatically extracting features. The state-of-the-art CNN-EEG model EEGNet was successfully employed for the classification of personality traits, obtaining the best classification results

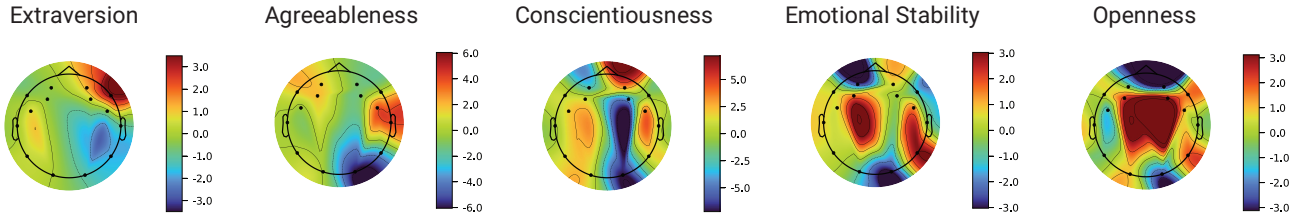

Figure 4: Average attribution maps for the five personality traits

Trait	Active filter	Acc	F1
Extraversion	All	0.898	0.884
	1	0.673	0.658
	2	0.617	0.648
Agreeableness	All	0.887	0.891
	4	0.663	0.708
Emotional Stability	All	0.882	0.888
	2	0.687	0.757

Table 3: Relevant filters identified by the ablation study on EEGNet-4,2 and their corresponding baseline performance with all filters active.

on the EEG traces of D3, on which only minimal pre-processing was performed, demonstrating that DL-based models trained on raw or minimally pre-processed signals can perform better than models trained on fully pre-processed ones. Moreover, the best structure and hyperparameters of the model were identified for the personality classification purpose. In terms of pure classification performance, EEGNet outperformed other comparable personality classification studies found in literature. Specifically, the best accuracies were obtained for Agreeableness (0.93) and Extraversion (0.92). Finally, it was shown how the learned filters can be analyzed for feature interpretability purposes.

A limitation of this study resides in the binary classification formulation. As future step, a multi-class classifier could be implemented. The feature interpretability task remains however the main limitation. Despite EEGNet allowing to isolate relevant filters and areas for the classification of the different traits, an association of specific EEG-based features to personality is not trivial. The temporal filters in fact, do not extract clear-cut frequency bands that could be directly associated to the trait under examination based on the filter's contribution to classification. For this reason, further investigation looking into more robust feature interpretability methods is needed.

References

[1] R. R. McCrae, "The Five-Factor Model of Personality: Consensus and Controversy," in *The Cambridge Handbook of Personality Psychology*, pp. 129–141, Cambridge University Press, oct 2020.

[2] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, vol. 12, pp. 479–493, apr 2021.

[3] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 2018.

[4] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *34th International Conference on Machine Learning, ICML 2017*, vol. 7, pp. 4844–4866, 2017.

[5] M. A. Klados, P. Konstantinidi, R. Dacosta-Aguayo, V. D. Kostaridou, A. Vinciarelli, and M. Zervakis, "Automatic recognition of personality profiles using EEG functional connectivity during emotional processing," *Brain Sciences*, vol. 10, may 2020.