POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Deep learning classification of Big Five personality traits from EEG signals

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Author: **Veronika Guleva**

Student ID: 943895
Advisor: Prof. Anna Maria Bianchi
Co-advisors: Alessandra Calcagno
Academic Year: 2021-2022

Abstract

Personality refers to a set of characteristics that influence the behavior and cognition of different individuals. Personality psychology has developed the Big Five model, which identifies five main attributes of personality, called traits, able to capture personality differences among individuals. Nevertheless, the five personality traits are generally assessed using self-report questionnaires which are particularly prone to bias. As a result, the demand for an automatic and more objective personality assessment approach has arisen.

In this context, the application of machine learning (ML) techniques to Electroencephalography (EEG) could be a valid classification approach. Indeed, several studies in literature successfully applied ML to EEG for different classification purposes, such as subjective emotion assessment. However, from the few existing studies that attempted to classify personality from EEG signals, it emerged that the a priori selection of EEG features able to differentiate different traits is a major limitation. In this context, the use of deep learning (DL) models that can automatically extract features could represent a promising approach. To the best of our knowledge, no studies that apply DL to EEG for personality classification are present in literature.

The aim of this thesis is to develop a DL-based binary personality classification method starting from EEG data, with a focus on model validation and interpretation of the extracted features. EEGNet, a state-of-the-art Convolutional Neural Network (CNN) model specifically designed for EEG decoding, was adopted. Classification was performed on the AMIGOS public dataset, which provides personality data and EEG traces from 38 subjects acquired during the visualization of emotional videos.

Specifically, a binary classifier for each of the five traits was implemented. To do this, a binarization of the personality scores was performed to generate a class representing low expression of the trait (class 0) and one for high expression of the trait (class 1). Moreover, in order to assess the model's ability to handle raw, minimally pre-processed, and fully pre-processed data, three different levels of preprocessing were applied to the EEG signals. The optimal EEGNet structure was assessed by means of a full validation of its hyperparameters. In the end, a five-fold cross-validation training strategy was used to assess classification performance on all the three differently pre-processed datasets.

Furthermore, the automatically extracted features were analyzed by directly visualizing the learned filters and hidden layer outputs in the frequency domain and by using DeepLIFT, a novel algorithm that assigns a contribution value to each input channel based on how much it affects the final prediction.
The best classification performance was achieved by the models trained on the minimally pre-processed data with an average accuracy and F1 scores > 0.89 for all five personality traits, while some preliminary relevant features were identified for three out of the five traits.
Keywords: EEG classification, personality classification, big five, personality traits, deep learning, convolutional neural network

Abstract in lingua italiana

La personalità comprende quell'insieme di caratteristiche che influenzano la cognizione e il comportamento di diversi individui. La psicologia della personalità ha sviluppato il modello Big Five, che identifica cinque attributi principali della personalità, chiamati tratti, in grado di cogliere le differenze di personalità tra gli individui. I cinque tratti di personalità sono generalmente quantificati utilizzando questionari autocompilativi, i quali, tuttavia, sono particolarmente soggetti a imprecisioni dovute all'autovalutazione. Di conseguenza, è sorta l'esigenza di un approccio automatico e più oggettivo per la valutazione della personalità.

In questo contesto, l'applicazione di tecniche di *machine learning* (ML) all'elettroencefalografia (EEG) potrebbe essere un valido approccio di classificazione. Effettivamente, diversi studi in letteratura hanno applicato con successo il ML sull'EEG per diversi scopi di classificazione, come ad esempio la valutazione delle emozioni soggettive.

Tuttavia, dai pochi studi esistenti che hanno tentato di classificare la personalità dai segnali EEG, è emerso che la selezione a priori delle caratteristiche EEG in grado di differenziare i diversi tratti è una grande limitazione. In questo contesto, l'uso di modelli di *deep learning* (DL) che possono estrarre automaticamente le feature potrebbe rappresentare un approccio promettente. Per quanto ne sappiamo non sono presenti in letteratura studi che applicano DL all'EEG per la classificazione della personalità.

Lo scopo di questa tesi è quello di sviluppare un metodo di classificazione binaria della personalità basato sul DL a partire da dati EEG, riservando particolare attenzione alla validazione del modello e all'interpretazione delle feature estratte. Il modello adottato per la classificazione è EEGNet, una *convolutional neural network* (CNN) progettata specificamente per la decodifica di segnali EEG. La classificazione è stata svolta sul dataset pubblico AMIGOS, il quale fornisce i dati della personalità e le tracce EEG di 38 soggetti acquisiti durante la visione di video emozionali.

In particolare, è stato implementato un classificatore binario per ciascuno dei cinque tratti. Per fare ciò, è stata eseguita una binarizzazione dei valori di personalità per generare una classe che rappresenta la bassa espressione del tratto (classe 0) e una per l'alta espressione del tratto (classe 1).

Inoltre, per valutare la capacità del modello di gestire dati grezzi, dati minimamente pre-

processati e dati completamente pre-processati, tre diversi livelli di pre-processing sono stati applicati ai segnali EEG. La struttura finale di EEGNet utilizzata per la classificazione è stata determinata mediante l'ottimizzazione dei suoi iperparametri. Infine, la performance di classificazione del modello ottenuto sono state valutate tramite una strategia di valutazione incrociata basata sulla suddivisione dei dati in cinque parti.
Inoltre, le feature estratte automaticamente dal classificatore sono state analizzate visualizzando direttamente i filtri appresi e gli output degli strati nascosti nel dominio della frequenza e utilizzando DeepLIFT, un algoritmo che assegna un valore di contribuzione a ciascun canale di input in base a quanto questo influisce sulla classificazione finale.
Le migliori prestazioni di classificazione sono state ottenute dai modelli allenati sui dati minimamente pre-processati, con un'accuratezza e F1 score medi $> 0{,}89$ per tutti e cinque i tratti della personalità, mentre a livello preliminare sono state identificate feature rilevanti per tre dei cinque tratti.
Parole chiave: classificazione EEG, classificazione personalità, Big Five, tratti della personalità, deep learning, convolutional neural network

Contents

Introduction

The possibility to automatically assess personality, starting from the individual's physiological signals, and in particular electroencephalographic (EEG) signals, has emerged in recent years thanks to advances in classification algorithms and neuroscientific studies on personality. Few studies aiming at classifying personality traits from EEG signals have been conducted and the classification methods used have relied on machine learning models that require manually extracted signal features for training. However, selecting the most representative features is a difficult challenge, particularly in emerging application areas such as personality classification, where feature-wise investigative studies are lacking. Deep learning is a promising alternative in this regard. Since deep learning models can generally extract features automatically, a priori feature selection is not required.

The aim of this work is to develop a deep learning-based personality classification method starting from EEG data acquired in response to emotional video stimuli. The public dataset AMIGOS [1] is used, which provides both EEG traces and personality data of 38 subjects. In this study, the problem of automatic personality assessment is formulated as a binary classification task, dividing personality trait scores into a high and a low class, each representing respectively a high and low expression of the specific trait. The deep learning approach is chosen with the aim of exploiting its automatic feature extraction capabilities as well as its potential for handling raw data. Specifically, a convolutional neural network model called EEGNet [2], designed for EEG decoding, is selected for this study. The model is fully validated in its hyperparameters and structure to provide a comprehensive practical assessment for the problem at hand. Different levels of pre-processing of the EEG data are tested to evaluate the model's performance on noisy signals. EEGNet models are trained separately for each personality trait and their classification performance is evaluated. A final examination of the automatically extracted features is performed to assess both the capability of the model to select relevant features, and the possible correlation of these features with personality.

1 | Background

This first introductory chapter aims to provide a comprehensive theoretical background of all necessary topics related to the overall work of the present thesis. In particular, a theoretical framework of personality is presented, followed by a minimal background on the EEG and its main features, and a comparison between machine learning and deep learning classification with a focus on EEG applications. Finally, a brief overview of related works on EEG-based personality classification, is presented.

1.1. Personality theory

Personality represents those characteristics that are unique to an individual and that distinguish them from others at a cognitive and behavioral level. Indeed, a personality profile can be identified based on how a person behaves, reacts to situations, and processes emotions. Individual personality characteristics are usually described in everyday language using a set of adjectives (e.g., extroverted/introverted, organized/scattered, emotional/stoic) that are, however, not strictly defined. In this context, researchers in the last century have tried to find the fundamental and independent principles of personality and to develop a personality model that could describe them successfully. Different approaches to the study and definition of personality have been developed in literature, and among the more relevant, the psychoanalytic perspective and the trait perspective can be cited [3]. The common core premise of these approaches is that personality is based on innate biological characteristics that are molded throughout the life course by a variety of factors, such as family, cultural background, and other experiences. The resulting pattern that characterizes an individual's behavior, cognition, and emotions forms personality [3].

The psychoanalytic approach was born in the early twentieth century and holds popularity to this day. The central theme of the psychoanalytic personality theory is the *unconscious*, a force that guides human thought and behavior. The most notable theory belonging to this framework was conceived by Sigmund Freud, credited as the founder of psychoanalysis. He identified three structures of personality in: i) the *Id*, an unconscious primitive force of biological drives; ii) the *Ego*, a rational side of personality, and iii) the *Superego*,

representing the societal and cultural rules that a person follows both consciously and unconsciously [4]. Another major exponent of the psychoanalytic approach to personality is Carl Jung, who identified personality in the *Self* and believed that conscious and unconscious forces coexist in a complementary way. His major contribution to personality theory consists in the definition of psychological types, which are based on the three dimensions of personality he identified (i.e., extroversion/introversion, thinking/feeling, and sensation/intuition). Indeed, the most popular categorization of personality types outside of the psychology field is the Myers-Briggs Type Indicator (MBTI), a psychometric test developed in the 1990s that tries to identify Jung's psychological *types* through a self-report questionnaire. MBTI has been used in research to investigate how the different personality *types* approach studying, decision making or how they deal with stress, among other things [5].

The main limitation of psychoanalytic theories of personality is their poor verifiability. A self-report questionnaire, for instance, appears inconclusive since psychoanalysis, by definition, assumes that the core of personality resides in the unconscious, which can't be accessed by the individual [4]. Moreover, the belief that personality exists on a continuum and that most people fall in the middle of a defined variable, rules out the existence of strongly defined distinct types. The MBTI test, for instance, classifies individuals in sixteen unique personality types based on four dichotomies. Empirical evidence has however shown how this kind of classification is not constant over time, making the test mostly unreliable [6]. For this reason, the trait approach has gained more relevance and credibility.

The trait approach to personality is distinguished by empirically identifying the unique traits characterizing personality, validating them scientifically and developing a measure scale. Several personality psychologists have studied and developed their own list of personality traits. The systematized and empirical approach to personality research was aided by the advent of factor analysis, a statistical procedure that examines the correlation between variables to determine a lower number of unobserved underlying variables, called factors, that could describe their variability. The first psychologist to apply factor analysis was Cattell (1945), who was able to identify 16 primary factors of personality and developed a questionnaire aimed at directly measuring these traits, the Sixteen Personality Factors Questionnaire (16 PF). Following Cattell, several studies by Fiske (1949), Tupes and Christal (1961), Norman (1963), among others, demonstrated that five factors were enough to account for the variability of personality [7].Then, several studies by Goldberg (1981), Digman and Takemoto-Chock (1981), McCrae and Costa (1985) validated the five-factor proposal and demonstrated its robustness. This consistent body of research converged into the Five-Factor Model (FFM) [7–9], the most widely accepted model of

personality today.

1.1.1. Five Factor Model

The five dimensions of the FFM [7–9], better known as the Big Five, are Extraversion, Agreeableness, Conscientiousness, Neuroticism or Emotional Stability, and Openness. These specific naming conventions became predominant with the publication of the NEO Personality Inventory (NEO-PI) by McCrae and Costa in 1985 for the assessment of the traits [7]. A revised inventory, NEO-PI-R, published in 1992, was translated in several languages and remains one of the most widely used methods of assessment. Since then, several other Big Five inventories have been published, such as the Big-Five Inventory (BFI), or the Big Five Marker Scales (BFMS) [10]. Each inventory develops its own questionnaire for the assessment of the traits. In general, a Big Five questionnaire consists of a series of questions, or adjectives or definitions, that are rated on a scale based on how well they apply to the person self-reporting their answers.

The five dimensions of the FFM, despite having been conceived as independent traits, are not entirely uncorrelated. Some traits have a shared variance which leads to higher order traits. Specifically, it's been found that Agreeableness, Conscientiousness and Emotional Stability co-vary, forming the higher order trait of Stability. While the co-variation of Extraversion and Openness forms the trait of Plasticity [11].

Similarly, the Big Five traits have a lower-order hierarchy. Each trait, or dimension, has several facets. Facets are lower-level traits that account for the variance of the higher-order trait. There is no consensus on the number of facets for each Big Five trait, but each facet has been shown to have a unique genetic contribution [11]. It's important to note that each inventory (e.g., the NEO-PI-R, the BFMS, and others) defines their own facets for the five dimensions but a general overlap and correlation is observed between the different questionnaires.

An intermediate-level layer exists between the Big Five traits and their facets. It was found that two genetic factors were necessary to explain the shared genetic variance between facets within each of the Big Five traits [11]. These middle-level factors, referred to as aspects, were also defined with factor analysis, which confirmed that each Big Five trait comprises two separable and correlated aspects. The hierarchical structure of the Big Five model is depicted in Figure 1.1. A description of each Big Five trait, and their relative aspects, is reported below [11–13].

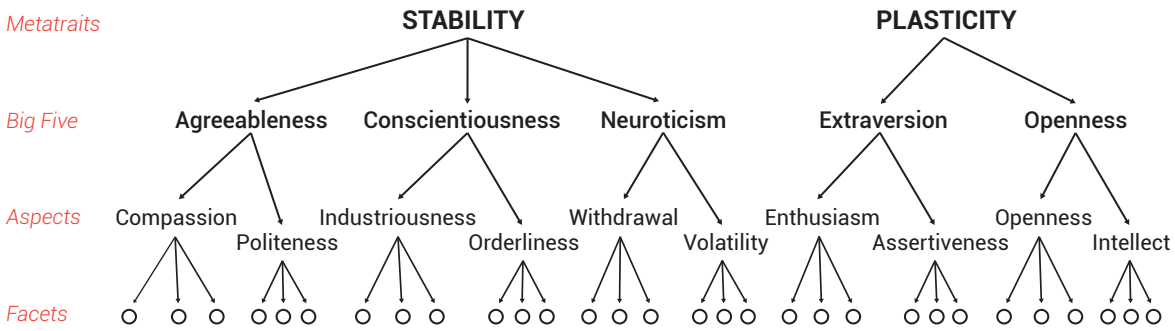

Figure 1.1: The hierarchical structure of the Big Five with its metatraits and its aspects and facets subtraits.

Extraversion encompasses traits related to sociability as well as to assertiveness and warmth. Its two aspects are Assertiveness and Enthusiasm. Assertiveness encompasses the facets related to the drive towards exciting and rewarding experiences, like gregariousness, activity, and vivacity. Enthusiasm on the other hand is related to the enjoyment of the experience, its main facet being the experience of positive emotions. Low scores of this trait represent people that are more reserved, quiet, shy, and not expressive.
Agreeableness identifies traits related to altruism and other prosocial traits. Its two aspects are Compassion, which reflects emotional attachment and concern for others, and Politeness, which represents the suppression and avoidance of aggressive or norm-violating impulses. Low scores of this trait represent people that are more cynical, egoistic, and egocentric.
Conscientiousness reflects the ability to inhibit impulsiveness and to follow strategies and abstract goals successfully. This trait is highly linked to academic success, health, and longevity because people who score highly are more likely to exhibit traits such as self-discipline, duty, and competence. Its two aspects are Industriousness, which reflects the prioritization of nonimmediate goals, and Orderliness, which reflects the avoidance of entropy by following rules set by self or others. Low scores on this trait characterize people who are impulsive and disorganized.
Neuroticism is linked to the tendency to experience negative emotions. Its two aspects are Withdrawal and Volatility. Withdrawal consists in passive avoidance, which is the tendency to inhibit behavior to avoid punishment and error, and is reflected in the facets of anxiety, depression, and self-consciousness. Volatility consists in active defensive responses and is reflected in the facets of anger, irritability, and impulsiveness. Some inventories, such as the BFMS, use Emotional Stability instead of Neuroticism, as it is its direct inverse. High scores in Neuroticism should reflect low scores in Emotional Stability and vice versa. An emotionally stable person is therefore someone who is calm, impassive,

and self-assured.
Openness reflects the tendency to effectively process abstract and perceptual information and encompasses traits such as imagination, intellectual curiosity, and aesthetic interest. Its two aspects are very distinct, with Intellect reflecting cognitive engagement with abstract information and ideas, and Openness to Experience reflecting cognitive engagement with sensory and perceptual information.

1.1.2. Neuroscience of personality

The Big Five model holds a primarily descriptive function and does not provide explanations of the underlying causes of the five personality dimensions. The causal components of personality are believed to have a largely biological basis that are not yet fully understood. Therefore, the study and identification of the biological principles underlying personality is an area of growing interest [12]. The premise of personality neuroscience is that the core individual differences in cognition, emotion, or motivation depend on consistent functional patterns in the brain [13]. The underlying systems are present in every human brain, but what characterizes the expression of a trait are the parameters varying from person to person [11]. The goal of personality neuroscience is therefore to understand the brain systems and mechanisms associated with and the cause of personality traits, and to identify related bio-markers able to capture differences in personality among individuals.
For the investigation of brain structure and functioning, neuroscience relies on brain imaging techniques. Neuroimaging techniques encompass methods such as magnetic resonance imaging (MRI), that provides structural brain images, and functional MRI (fMRI) and electroencephalography (EEG) that provide, instead, functional information, indicating, for example, which brain regions are more or less active in specific conditions [13]. A major issue in the neuroscience of personality field is the inconsistency in the findings to date, which are due in large part to the very small number of samples in neuroimaging research. In fact, brain imaging techniques are rather expensive, making the studies limited to few samples which lack statistical power and increase the number of false positive cases [11].
Although there is still no consistent and well-supported evidence in neuroscience on personality, research in this field has increased in recent years and the quality of studies has improved and some links to specific brain structure and function have found support.
In general, personality has been strongly associated to the frontal region, specifically the prefrontal cortex (PFC), which is linked to cognitive functions such as attention, working memory and decision making, and to emotional, social, and perceptual processing

[14]. Historically, this association was made following personality changes caused by traumatic brain injuries in the area [15]. Indeed, the different types of acquired personality disturbances can be intuitively linked to the Big Five traits. For instance, a social behavioral disturbance due to damage in the orbitofrontal cortex that leads to disinhibition, impulsivity, aggression, and selfishness [16], can be connected to changes in the traits of Extraversion (high) and Agreeableness (low). An executive function disturbance in the ventrolateral and dorsolateral PFC that leads to impairments in organization, planning, and perseverance [16], can be linked to Conscientiousness (low). Other disturbances that can be associated to specific personality traits are lack of empathy [16] (Agreeableness, low), emotional dysregulation [15] (Emotional Stability, low), and hypo-emotionality characterized by apathy and social withdrawal [15] (Extraversion, low).

The advent of neuroimaging techniques allowed neuroscientists to test some of these hypothesized links between personality and brain areas. Research has proven the strong correlation of all personality traits with the PFC. Specifically, correlations were found between Extraversion and activity in the orbitofrontal cortex, Agreeableness and Conscientiousness and activity in the dorsolateral PFC, Emotional Stability and activity in the medial prefrontal cortex [12], and between Openness and function of the PFC in general [13].

Another area of interest is the amygdala, an affective region associated to Extraversion and Emotional Stability, traits generally related to positive and negative emotions respectively [12]. The temporal region involved in the interpretation of other individuals' actions and intentions, is instead associated to Agreeableness [12] and to Emotional Stability [17], traits that influence the way one perceives emotional expression. In Figure 1.2, some brain regions whose structure (volume) was correlated to personality traits are represented.

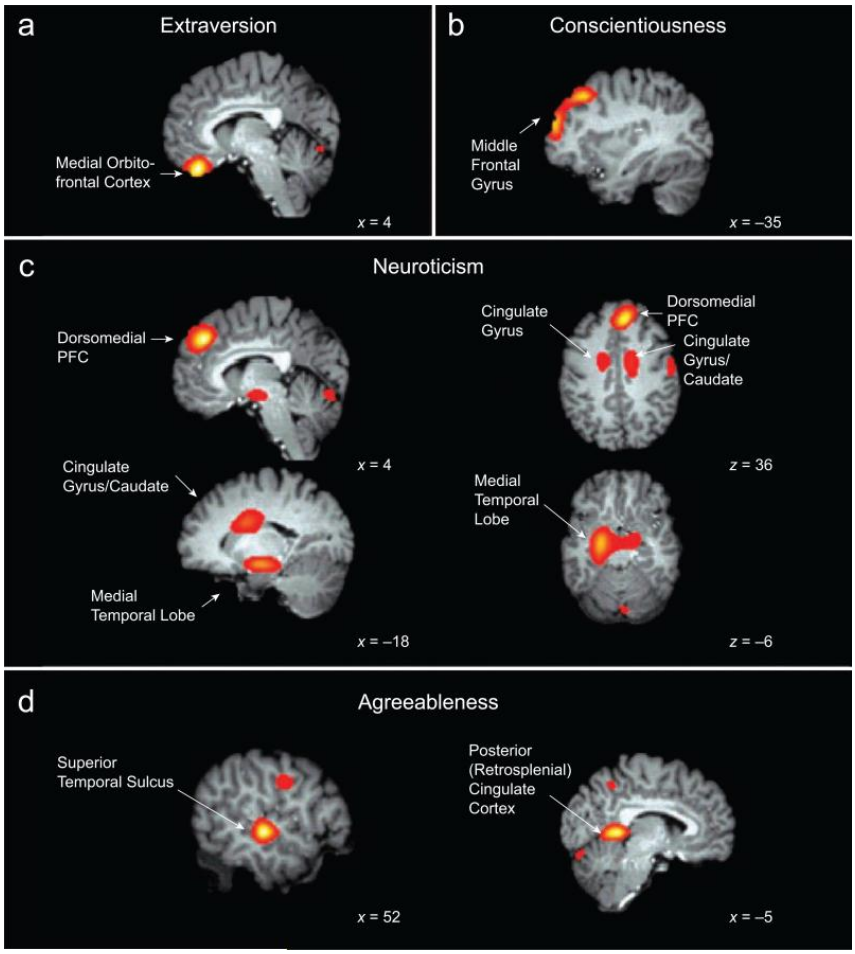

Figure 1.2: Brain regions in which local volume was significantly associated with (a) Extraversion, (b) Conscientiousness, (c) Neuroticism, and (d) Agreeableness. Coordinates indicate the locations of the brain slices [12].

1.2. Electroencephalography (EEG)

Electroencephalography (EEG) is the most widely used noninvasive technique for the measurement of the electrical fields produced by the brain. EEG picks up electrical potential differences on the scalp using electrodes. The signal measured is the sum of tiny excitatory post-synaptic potentials produced by pyramidal neurons in the cortical layers of the brain [18]. Due to the speed of propagation of the electrical fields, EEG has great temporal resolution which allows the detection of events in the order of milliseconds. However, EEG has low spatial resolution since electrodes placed on the scalp capture electrical fields that are smeared by the tissues between the sources and the sensors [19].

The electrodes used to measure the EEG signal are noninvasive. They are placed on the

scalp with a conductive bridge established between the electrodes and the skin surface. The placement of the electrodes on the scalp, and their relative name, follows an international convention that guarantees that methodology of acquisition is consistent. In Figure 1.3, the electrodes placements in the 10-10 system are depicted. The conventional names for the electrodes represent an identifying letter based on the brain area they cover — F for frontal lobe, C for central, P for parietal, T for temporal, and O for occipital. Even numbered electrodes refer to placement on the right side of the head, while odd numbered electrodes refer to those on the left. The number of electrodes used depends on the type of headset used.

EEG has a wide frequency range that goes from 0.5 Hz to approximately 70 Hz but is measurable up to 100 Hz. In this bandwidth five frequency ranges characteristic of brain electrical activity can be identified: i) the Delta band (0.5 - 4 Hz), associated to high amplitude waves and most prominent during deep sleep, ii) the Theta band (4 - 8 Hz), associated to drowsiness and memory formation and navigation, iii) the Alpha band (8 - 13 Hz), prominent in the occipital region during relaxed wakefulness phases, iv) the Beta band (13 - 30 Hz), which is a low amplitude rhythm that characterizes various mental states such as concentration, excitement, anxiety, and task engagement, and v) the gamma band (30 - 100 Hz) generally associated to sensory perception and conscious attention [19].

A critical step related to the analysis of EEG signals is related to the pre-processing phase, which has the purpose of reducing the signal-to-noise ratio. Artifactual sources can be instrumental, such as the power line noise or biological, such as muscle-electromyogram activity, eye blinks, and eye movements [19]. Generally, the EEG processing pipeline is complex and follows several steps aimed at cleaning the signal from artifacts, extracting relevant features and eventually classifying them. Pre-processing usually includes downsampling, bandpass filtering to limit the band of the signal, removal of noisy channels, removal of ocular and muscular artifacts either manually or by using independent component analysis [20], and re-referencing.

The pre-processing step is then followed by feature extraction. Generally, different approaches can be undertaken: a frequency-based approach aimed at quantifying the relative power of each frequency band, a time-based approach in case an event related response is being investigated, or a time-frequency-based approach that relates the power in time [18].

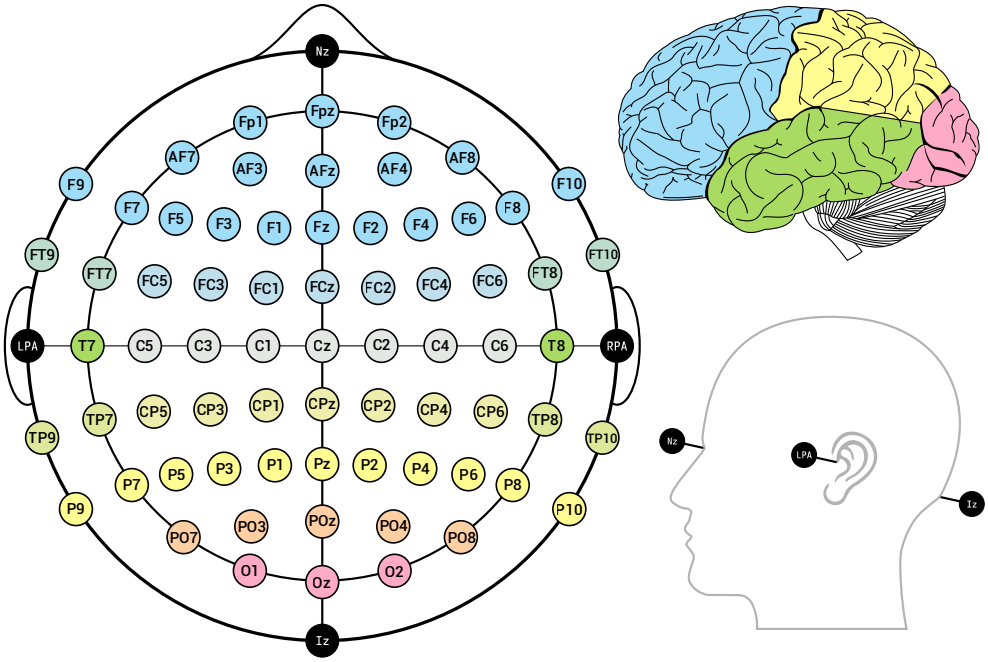

Figure 1.3: Electrodes placement in the 10-10 system and their corresponding color-coded brain area.

1.3. Deep Learning

Machine learning (ML) is a field that aims to develop algorithms able to learn from data and experience. As such, it lends itself to solve the classification problem requiring the algorithm to learn a function that maps a vector to its predefined class label [21]. The classification task is usually approached through supervised learning which relies on a labeled dataset. Specifically, each training example paired with its known classification label, is fed to the selected classification model. The model's adjustable parameters, called weights, are updated during the training of the learning algorithm. The updating method is generally focused on the minimization of an objective error function based on the prediction error (i.e., difference between the predicted label and the real label). The minimization of the error function allows the model to learn the correct representation of the data, and to predict new data reliably [21].

Several ML models can be used for classification, such as K-nearest-neighbors [22] or Support Vector Machines [23]. ML-based classification has been successfully applied in several fields, for instance for brain-computer interface (BCI) classification of EEG signals [24]. The main characteristic of ML-based classification (Figure 1.4A) is the feature extraction stage, i.e., the extraction of representative features from the input training

data. This stage depends on a priori knowledge of the type of data and the specific application, needed to select the most useful features for the classification. Selecting the right features constitutes a challenge, especially for novel problems with little a priori knowledge.

For this reason, the need for new automatic feature extraction techniques has emerged. The models that can automatically extract features fall under the domain of Deep learning (DL) (Figure 1.4B), a subfield of ML [21]. This new approach can potentially allow to learn features on raw or minimally processed data, reducing the need for application-specific processing and feature extraction strategies. Additionally, since the features learned through DL methods are not dependent on a-priori knowledge and are not application-specific, they might be more representative of the data and bring to better classification results. Just like for ML, several DL models exist, such as Convolutional Neural Networks [25], Recurrent Neural Networks [26], Autoencoders [27], Deep Belief Networks [28], among others.

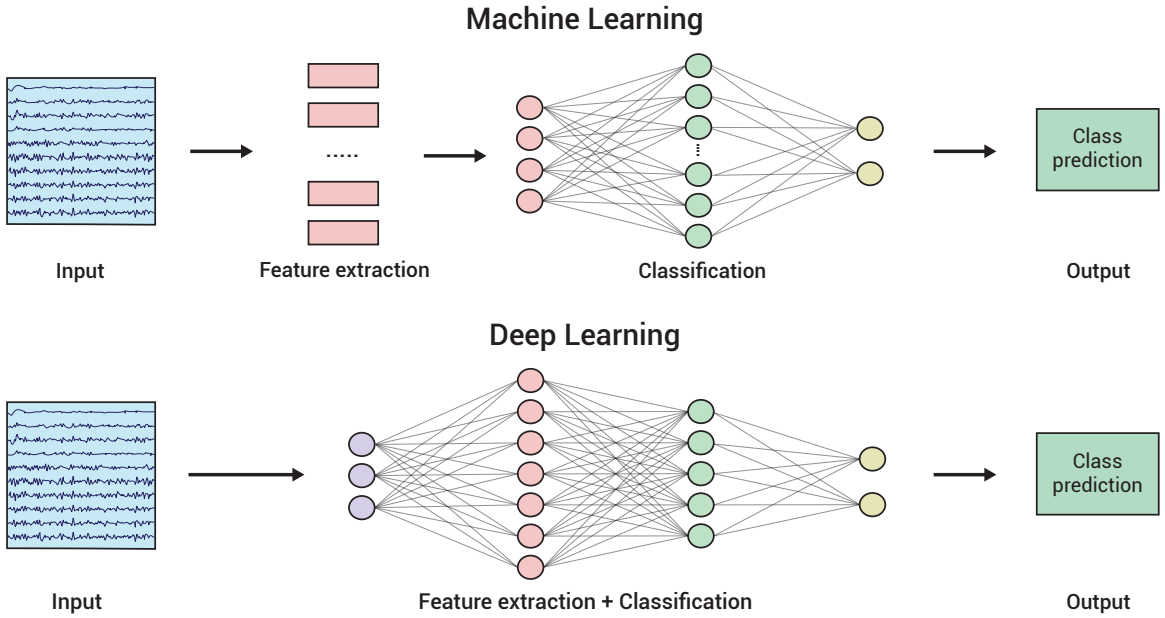

Figure 1.4: Machine learning vs. deep learning-based approach for a classification task.

1.3.1. EEG Applications

DL methods have been successfully implemented in challenging image classification problems where they have outperformed state-of-the-art ML methods relying on hand crafted features, as well as in speech recognition, text classification and other applications [29]. This success has led to the increasing interest in trying to apply the new findings in DL

to EEG decoding. A recent review of DL-EEG applications [18] lists that DL has been applied for the classification of EEG data for sleep staging [30], visual evoked potentials [2, 31], seizure detection [32], brain-computer interfaces (BCIs) [33], and emotion recognition [34]. The absolute number of DL-EEG studies is still relatively small but has exponentially increased in recent years reflecting the growing interest in the field [18].

Several DL models have been employed for EEG classification tasks with their specific architecture varying from study to study. Despite DL's main advantage being its automatic feature learning capability, more than half of the DL-EEG classification studies identified in [18] still used hand-crafted EEG features, mostly obtained in the frequency-domain. For example, obtaining spectrograms, i.e., time series of topographically organized images representing the voltage distributions across the flattened scalp surface, or the functional connectivity maps from the raw signals, and feeding these high-level extracted features to the network. The EEG signals pre-processing also varies, with most studies preferring a full pre-processing rather than using raw data directly [18]. The studies that leveraged the end-to-end learning (i.e., training feature extraction and classification simultaneously) capabilities of DL most often used CNN models [18, 30, 35]. In Figure 1.5, an example of hand-crafted features, specifically functional connectivity features, used to train a convolutional neural network for the classification of brain disorders [36] is represented.

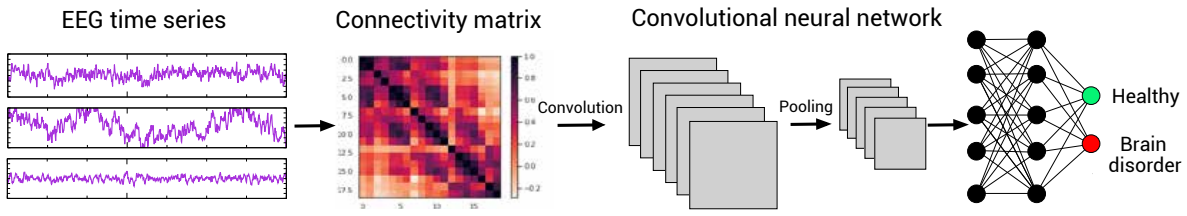

Figure 1.5: Example of DL-EEG application for the classification of brain disorders by means of connectivity features and convolutional neural network [36].

Despite the promising characteristics of DL methods, they are not always the best choice for EEG-based classification. One main limitation is related to the dataset dimension: DL classifiers, indeed, require a large amount of data to reach good performances, condition not always satisfied when dealing with EEG traces. Moreover, EEG signals, differently from other kinds of inputs usually used in DL approaches (i.e., images, sound), are generally characterized by a low signal-to-noise ratio, characteristic that can affect classification performances [18]. Another fundamental limitation of DL models is the poor interpretability of the extracted features. This presents a problem in applications where there is a need to know the type of feature that contributes most to classification.

1.3.2. Convolutional Neural Networks

Convolutional neural networks (CNNs) are artificial neural networks that can learn local non-linear features through convolutions and nonlinearities [25]. They represent higher-level features as compositions of lower-level features, through multiple layers of processing. CNNs are designed to take as input data in the form of arrays. Typically, the most common type of data processed using CNNs are images, which are 2D arrays. Other data that comes in the form of arrays are 1D signals and sequences like language, 2D audio spectrograms and 3D videos and volumetric images [21]. The typical architecture of a CNN is organized in sequential layers.

The fundamental layer of CNNs, is the convolutional layer, which maps its input to an output through a convolution operator. This operation can be performed in one, two and three dimensions (i.e., 1D, 2D, and 3D). As an example, supposing to have a 1D input x_n with N samples and the 1D convolution filter h_m of size M, then the output of the 1D convolutional layer is given by Equation (1.1) [24].

$$y(n) = \sum_{i=0}^{M-1} h_i x_{n-1} \qquad \forall n = 0, \ldots, N-1 \tag{1.1}$$

The output of convolutional layers is called feature map. Usually, by employing multiple filters in one convolutional layer, multiple feature maps, equal to the number of filters, are obtained. The convolutional layer is followed then by a linear or non-linear activation layer, which control the values of the output, and possibly by pooling layers, which aggregate a local patch of units into a single value by using an average or max operator [24]. For classification tasks CNNs are generally followed by a fully connected layer, which allows to directly map the extracted features to an output, after the flattening of the feature maps. An example of a CNN architecture is represented in Figure 1.6.

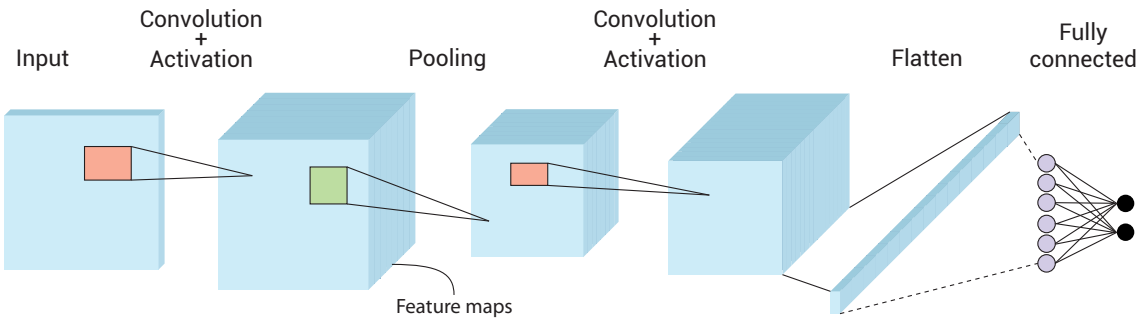

Figure 1.6: A basic CNN architecture for classification.

CNNs are usually trained using a supervised learning strategy with the cost error function

(Equation (1.2)) to be minimized.

$$\hat{\boldsymbol{w}} = \arg\min_{w} \frac{1}{\ell} \sum_{i} L(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i)) + \Omega(\boldsymbol{w}) \tag{1.2}$$

where $\{\boldsymbol{x}_i, y_i\}$ are the training data, $f_{\boldsymbol{w}}$ is the prediction function related to the CNN, $L(.,.)$ is a loss function that measures any discrepancy between the true labels of $\boldsymbol{x}_i$ and the predicted labels $f_{\boldsymbol{w}}(\boldsymbol{x}_i)$, and Ω is a regularization function for the weights $\boldsymbol{w}$. The most popular algorithm for the training of CNNs is stochastic gradient descent (SGD) [24].

Design choices for EEG applications

CNNs are the most widely used DL models for EEG applications [18]. This can be attributed to their capabilities of end-to-end learning and of exploiting hierarchical structure on the data, as well as to their success and popularity in computer vision tasks [18]. CNN design choices specifically built for EEG decoding, such as the input representation, different type of architectures and training strategies, have been investigated [18, 37].

The representation of the input EEG signal can take different forms. For example, the power spectra of the signal in time, the spectrogram, have been used as an input "image" and then state-of-the-art CNN architectures from computer vision were directly employed to try to decode "image" signal without any other adjustment. The optimal input representation of EEG signals for CNNs, as stated in [37], is a 2D array with the number of time steps as the width and the number of electrode channels as the height. This allows to design the network with the proper layers, enabling the model to mimic the process of extracting EEG features [18].

The layer design choices for the CNNs might force the model to process temporal and spatial information separately. Convolutional layers can be set up such that they replicate spatial filters used to unmix the global spatial patterns or temporal filters used to unmix the local modulations in time [37]. As for the depth of the models, it can't be concluded if deeper or shallower architectures perform better in EEG classification tasks since the performance is dependent on factors outside the architecture itself, such as the amount of available training data, the hyperparameter tuning strategy and the computational power [18].

Additional to the input and layer design choices, two different training strategies were also investigated [37]: trial-wise training, that uses whole trials as input with per-trial labels as targets, and cropped training, that uses crops, that is, sliding windows within the trial as input with per-crop labels as targets. The cropped training strategy naturally produces

way more training samples for the network compared to the trial-wise one. Crops can be of about 2 or 3 seconds, to a minimum of one crop per sample, which generates the maximum possible number of samples. The cropping technique has the aim to force the CNN into using features that are present in all crops of the trial, instead of using the differences between single crops or the global temporal structure of the features in the complete trial [37].
The aforementioned concepts were used in the design of EEGNet [2], a state-of-the-art CNN-EEG model. In the first two convolutional layers of the model, the temporal and spatial filtering techniques usually applied to EEG signals, are mimicked, and applied sequentially. EEGNet has been validated with comparable or better performance than state-of-the-art approaches for brain-computer interface (BCI) paradigms. EEG features of interest classified for BCIs are either event related or oscillatory, such as visual-evoked potentials, sensory motor rhythms, and movement-related cortical potentials [2]. As such, EEGNet has not been tested on datasets in which the EEG features of interest are not event related, or have no known specific frequency bands, as is the case for the classification of subject-based personality traits.

1.4. Personality classification with EEG

Neuroscience has linked personality traits to brain function by means of brain imaging techniques. Among brain imaging techniques, EEG is the most directly accessible measure of electrical brain activity. Unlike other burdensome and costly neuroimaging acquisition methods such as MRI and fMRI, EEG is rather inexpensive and portable, thanks to the availability of new wearable devices. Due to these characteristics, it has become the candidate technique for the development of an automatic personality assessment method. Currently however, the studies that have tackled this task are few.
Individual differences in electrical brain activity can either be stable (i.e., situationally independent) or appear in response to some stimuli (i.e., situationally dependent). To test the feasibility of using EEG to measure personality, Korjus and colleagues [38] tried to classify each trait in two classes starting from resting-state EEG (i.e., situation-independent condition). Specifically, they used power-based quantitative measures as predicting features but did not obtain promising classification results. The authors hypothesize that personality may involve more situation-dependent brain activity [38] even though it must be noted that more recent resting-state EEG-based personality studies [39–41] have found correlations between personality traits and brain arousal or brain connectivity graph measures.
Indeed, most of the subsequent studies attempted to predict personality from EEG signals

recorded in response to different types of stimuli, rather than baseline recordings. The majority of these studies [42–46] have focused on eliciting an affective response in the subjects by displaying emotional videos extracted from movies, the same method usually employed in emotion classification research. This choice seems coherent with the fact that some personality traits are intrinsically associated with the tendency to experience positive or negative emotions [11]. The main hypothesis behind this kind of experimental protocol is that the affective stimuli would produce a response in the recorded EEG signal which can capture the situation-dependent characteristics of personality traits.
All the EEG-based personality classification studies identified [42–47] made use of hand-crafted EEG features and ML methods for a binary classification. Specifically, most studies relied on frequency-based features, such as power spectral densities (PSDs) of the characteristic EEG bands (Delta, Theta, Alpha, Beta, Gamma) [43–45, 47]. Others extracted statistics-based features [42] or connectivity-based features such as brain networks [46]. The classification was performed with classical ML models such as SVMs and kNNs with varying performances. A couple of studies [43, 44] performed a regression instead, i.e., assigning a continuous value to each trait instead of a class. Moreover, all studies fully pre-processed the EEG data before extracting the features.
To the best of our knowledge, literature lacks studies on personality classification from EEG signals based on DL methods for end-to-end learning. Nevertheless, as seen in section 1.3.1, other studies demonstrated how DL applied on EEG signals can lead to promising classification results and, therefore, it could be reasonable to test its performances also in the personality classification field. Indeed, literature studies have shown how the neurological bases of personality are still not completely understood, and the identification of meaningful EEG-based features able to discriminate different traits is generally not trivial. Therefore, the ability of DL of automatically extracting features, as well as the possibility of using raw or minimally pre-processed EEG signals, could allow to overcome these limitations.

1.5. Aim of thesis

It was seen how the need for an objective method for measuring personality stems from the inherent subjectivity of self-report questionnaires currently used for assessment. The core motivation behind the possibility of using neurophysiological signals for this task lies in the neuroscientific studies linking personality to brain function. The fact that there is still insufficient well-supported evidence on the mechanisms underlying personality makes the selection of EEG-based features for classification difficult. DL is a promising alternative to a ML-based approach in this case as it can automatically extract the features

necessary for classification. DL models such as EEGNet have been successfully applied to other types of EEG-based applications, mostly aimed at classifying event-related signal components or oscillatory rhythms for BCI applications. To the best of our knowledge, no known personality classification studies have addressed the problem using DL models. In addition to automatic feature extraction, DL also has the capability of handling raw data, a feature that could greatly simplify EEG pre-processing pipelines.
The aim of the present thesis is twofold:

1. **Develop an automatic personality classification method using DL models**. This first objective is carried out by training CNN models separately for each trait, with a binary classification task. Each trait is classified in a low or high expression of the trait. The public AMIGOS dataset [1] is used as it collects both Big Five personality data and EEG signals recorded in response to emotional video stimuli. The model chosen for this task is EEGNet [2], a state-of-the-art CNN-EEG decoding model. The optimal structure of the model is assessed by a full validation of its hyperparameters. Three different types of pre-processing are applied to the EEG signals and tested to assess the ability of the model to handle noisy data. The classification performance is then evaluated using a cross-validation strategy.
2. **Analyze the automatically extracted features and their relative contribution to the prediction of each trait**. This second objective is aimed at overcoming the main limitation of DL applications, i.e., their poor interpretability. By using visualization techniques or by simply looking at the outputs of hidden layers in the trained models, the automatically extracted features can be investigated. In this regard, the features extracted by the first convolutional layer, i.e., the learned temporal filters, are analyzed. A first level of analysis is made by deactivating combinations of filters and evaluating the performance of the modified model in order to identify the features that contribute most to the prediction of personality. Then these features are further analyzed with the aim of drawing parallels between the features most relevant for the prediction of each of the five traits, and the findings in personality neuroscience research.

The rest of this work is organized as follows:

- In Chapter 2, the AMIGOS dataset and the EEGNet model used are thoroughly described. Then, the data processing and the hyperparameter tuning, model training and evaluation strategies are presented. Finally, methods used for feature interpretability, such as the visualization of filters and algorithms used for assigning an importance value to input features, are introduced.
- In Chapter 3, the results are presented divided in model validation, classification,

and feature interpretability sections. Specifically, in the model validation section the results of the optimization of the hyperparameters and structure of the model are presented. In the classification section, the final performances of the trained classifiers are described and discussed. The feature interpretability section reports some selected results regarding the learned filters and hidden layer outputs, as well as attribution maps assigning a prediction-contribution value to the input.

– In Chapter 4, the conclusions of this work are drawn, and the limitations and possible feature developments are discussed.

2 | Materials and methods

2.1. Dataset

For the present study, the public AMIGOS dataset was used. This dataset was acquired for a research on affect, personality traits and mood on Individuals and GrOupS by means of neuro-physiological signals [1]. Affect was elicited with emotional clips extracted from movies as stimuli, shown to the subjects both individually and in a group setting. The group setting strategy was set up to mimic a real-life affective response, which is usually experienced in a social context. Multiple neuro-physiological signals were acquired simultaneously during the visualization of emotional videos. Specifically, Electroencephalography (EEG), Electrocardiogram (ECG) and Galvanic Skin Response (GSR) were recorded. Moreover, frontal face and full body videos were acquired. The dataset also provides information about: i) participants' emotions, self-assessed by means of questionnaires, ii) the levels of valence and arousal [48] attributable to each video, iii) participants' personality, assessed through the Big Five Marker Scales (BFMS) [10] questionnaire and, iv) participants' mood, assessed by means of the Positive and Negative Affect Schedules (PANAS) questionnaire [49].

2.1.1. Experimental protocol

Forty healthy participants (male = 27, female = 13, aged 21-40 years, mean age = 28.3), took part of two experimental settings. Both experiments elicited affect using emotional videos as stimuli.

The first experiment, called the *short videos experiment*, was carried out in an individual setting. Each participant watched 16 short videos of duration of less than 5 minutes, presented in random order. Each trial started with a 5 second baseline acquisition period during which a fixation cross was showed, followed by the visualization of a short video. Finally, participants were asked to self-assess their affective state felt during the video.

The second experiment, called the*long videos experiment*, was carried out either in group or in individual setting. Specifically, 17 participants watched 4 long videos, of duration

greater than 14 minutes, individually, while 20 participants watched the videos in 5 groups of 4 persons each. In order to maximize social interactions, individuals that already knew each other were inserted in the same group. Videos were shown randomly, and each trial consisted of initial self-assessment, followed by the display of two long videos, and a final self-assessment.

Personality and mood related data was acquired after the long video experiment through an online form implementing the BFMS [10] and PANAS [49] questionnaires, respectively.

The videos selected as stimuli were chosen from movies aiming to elicit an emotional response. They were annotated on the valance and arousal dimensions and classified in four classes corresponding to the four quadrants of the valence-arousal (VA) space [50] (Figure 2.1): high valence-high arousal (HVHA), high valence-low arousal (HVLA), low valance-high arousal (LVHA), and low valence-low arousal (LVLA).

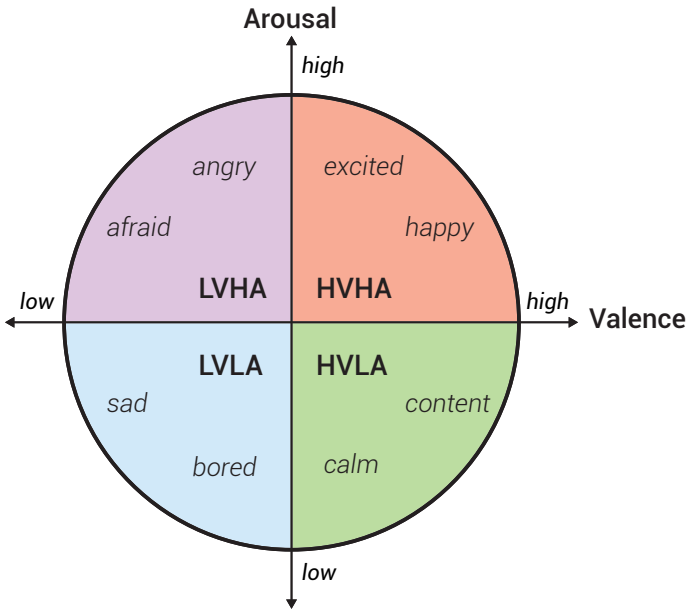

Figure 2.1: The four quadrants of the valence-arousal space and their associated emotions.

2.1.2. EEG signal

The EEG signals were recorded using the EPOC Neuroheadset (EMOTIV, U.S.A). This wireless wearable sensor has 14 channels, an internal sampling rate of 2048 Hz downsampled to 128 Hz, and a 14-bit resolution. The 14 EEG channels, positioned according to the 10-10 system are: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 (Figure 2.2).

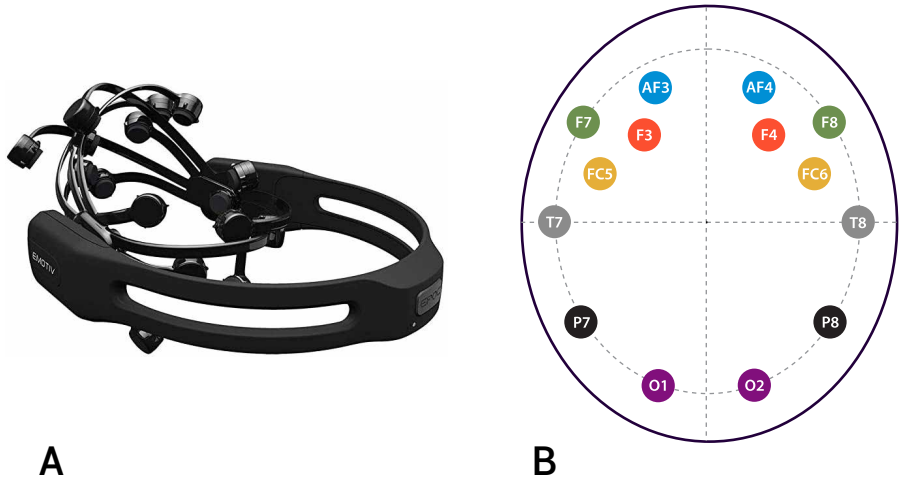

Figure 2.2: A. Emotiv EPOC headset. B. Positions of the 14 electrodes according to the 10-10 system.

2.1.3. Emotion assessment

Self-assessment

Self-assessment of the subject's levels of valence, arousal, dominance, liking, familiarity, and the basic emotions felt (i.e., neutral, disgust, happiness, surprise, anger, fear, and sadness) was carried out both before the short and long videos experiments and right at the end of the experimental protocol.

The levels of valence, arousal, and dominance were assessed using the Self-Assessment Manikin (SAM) scale ranging from 1 to 9 [48]: the valence, representing the pleasantness of a stimulus, ranged from "very negative" to "very positive" extremes; the arousal, representing the intensity of the provoked emotion, ranged from "very calm" to "very excited"; and the dominance, representing the degree of control exerted by a stimulus, ranged from "overwhelmed with emotions" to "in full control of emotions". The liking and familiarity were also assessed using a 1-9 scale. The participants were asked to select at least one or as many basic emotions they felt before and after watching the videos.

External assessment

The recorded frontal videos of each participant during both experiments, were annotated by three experts. All videos were split into 20 seconds clips resulting in 340 clips per participant, for a total of 12580 clips. The annotators evaluated the clips in random order, assigning a valence/arousal value on a continuous scale ranging from -1 (low valence/arousal) to +1 (high valence/arousal).

Big Five Marker Scale	
Positive adjectives	*Negative adjectives*
I. Extraversion	
Extroverted	Reserved
Warm-hearted	Shy
Open	Silent
Exuberant	Introverted
Vivacious	Closed off
II. Agreeableness	
Altruistic	Egoistic
Agreeable	Revengeful
Generous	Cynical
Sympathetic	Egocentric
Hospitable	Suspicious
III. Conscientiousness	
Precise	Untidy
Orderly	Inconstant
Diligent	Imprecise
Methodical	Careless
Conscientious	Rash
IV. Emotional Stability	
Self-assured	Nervous
Serene	Anxious
Calm	Emotional
Impassive	Susceptible
Jealous	Touchy
V. Openness	
Creative	Superficial
Imaginative	Obtuse
Original	
Ingenious	
Poetic	
Intuitive	
Intelligent	
Rebellious	

Table 2.1: Big Five Marker Scale facets assigned to each personality trait. The left column lists all the positive adjectives correlated with the given trait, while the right column lists the adjectives negatively correlated with the trait.

2.1.4. Personality assessment

The personality data was obtained through an online questionnaire filled in by each participant. The personality model used was the Big Five personality traits model while the form used was the Big Five Marker Scale (BFMS) [10] questionnaire. Each participant was asked to rate 50 descriptive adjectives with the prompt "I see myself as a _____ person", assigning a value within a 7-point Likert scale, with 1 meaning they did not identify with the given adjective, and 7 meaning they identified strongly with the given adjective. For each of the five personality traits (i.e., Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness), ten descriptive adjectives are assigned by the BFMS scale: 5 positive adjectives that characterize the given trait, and 5 negative adjectives that do not characterize the given trait. The final score for each trait is obtained by calculating the mean of all the adjectives' scores within the trait. Specifically, the negative adjectives' scores are adjusted by inverting their values on the scale (e.g., a negative adjective rated 7, becomes a 1 adjusted for the final score calculation, since it contributes negatively to the overall trait score). In Table 2.1, the positive and negative adjectives assigned to each trait by the BFMS questionnaire are reported.
The personality trait scores of the 38 subjects who filled in the questionnaire were collected. The mean, median and standard deviation (SD) values for each personality trait, and their Spearman inter-correlations were calculated and are reported in Table 2.2. A significant positive correlation was found between the traits Extraversion and Agreeableness (0.42), Agreeableness and Conscientiousness (0.34), and Conscientiousness and Emotional Stability (0.35). The distributions of the scores of each trait with their respective mean and median values are plotted in Figure 2.3.

	Mean	Median	SD	(A)	(C)	(ES)	(O)
Extraversion (E)	4.06	3.90	0.98	**0.42**	0.09	0.22	0.13
Agreeableness (A)	5.02	5.05	0.94		**0.34**	0.12	0.23
Conscientiousness (C)	4.88	5.00	0.94			**0.35**	-0.01
Emotional Stability (ES)	4.38	4.50	0.85				0.24
Openness (O)	4.86	4.90	0.66				

Table 2.2: Mean, median, and SD of the five personality traits scores, and their Spearman inter-correlation, for all 38 subjects. Significant correlations (p-value < 0.05) are evidenced in bold.

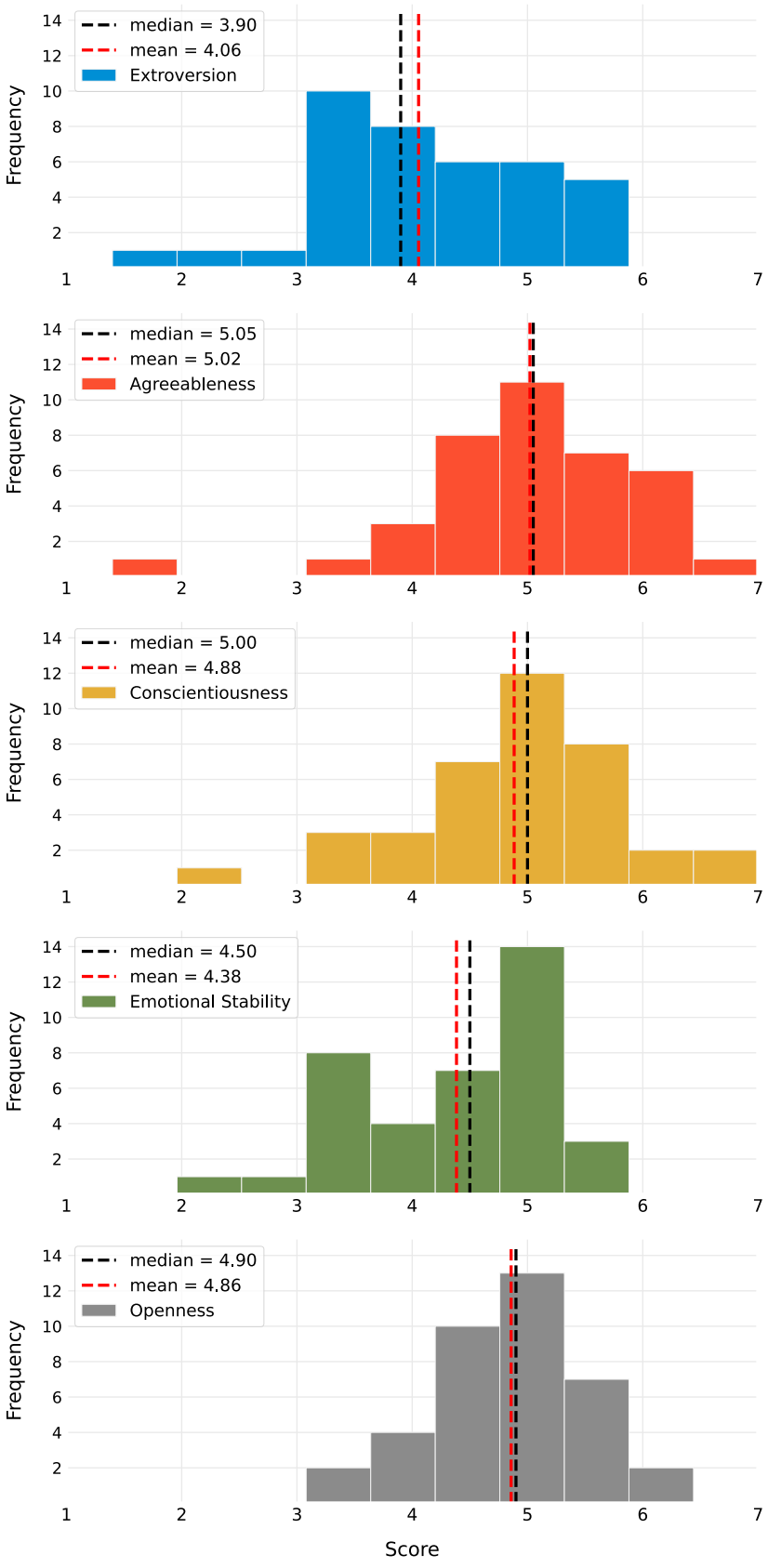

Figure 2.3: Histograms of the scores of each personality trait with their respective mean and median values.

2.1.5. Mood assessment

Mood was assessed on the positive affect (PA) and negative affect (NA) schedules (PANAS) model [49], using an online form. The questionnaire consisted of two 10-questions sets, to assess the PA and NA respectively. Each participant rated their general feelings in a 5-point intensity scale. The final score was obtained by calculating the mean of the 10-questions scores for the PA and NA respectively.

2.2. General processing and classification pipeline

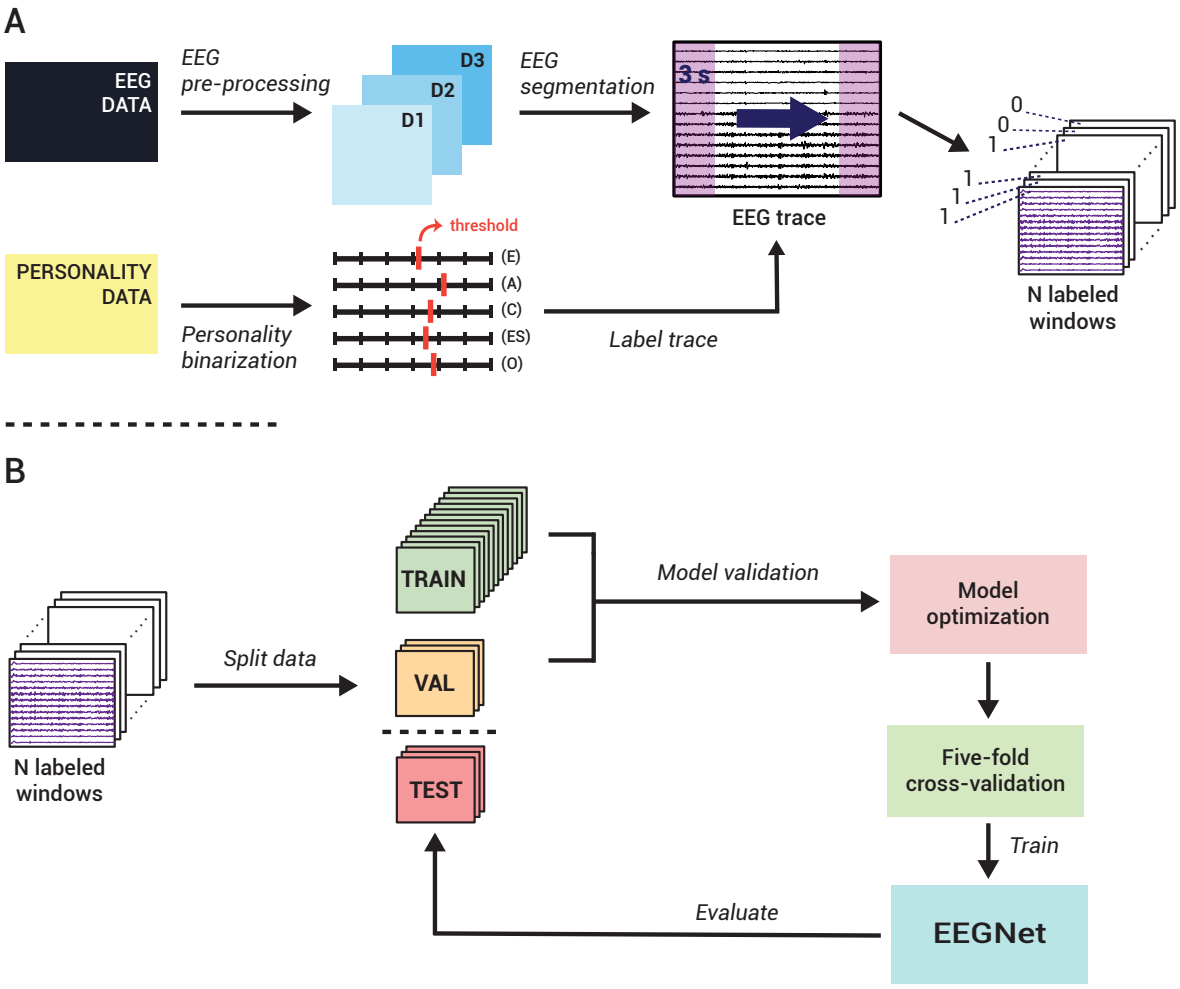

Figure 2.4: General pipeline for A. Data processing and B. Classification task.

A binary classification task is implemented as it's one of the simplest classification formulations for any predictive problem. To this end, from the data provided by the AMIGOS dataset (Section 2.1), the personality trait scores and the recorded EEG signals for each

participant were chosen. Specifically, only EEG data collected during the short videos experiment was selected given that it was recorded in an individual setting and no other interferences, such as social interactions, that represent an additional factor of analysis, should be present. Additionally, out of the 40 subjects tested, 2 (subject 8 and 28) were discarded due to missing personality data.
The general data processing and classification pipeline is represented in Figure 2.4A and Figure 2.4B, respectively. The given EEG and personality data were appropriately processed as described in Section 2.3. The structure of the chosen model, EEGNet [2], is described in Section 2.4. The model was optimized as described in Section 2.5 and trained using a five-fold cross-validation strategy, as described in Section 2.6.

2.3. Data processing

2.3.1. EEG pre-processing

To test the classification performances on differently pre-processed EEG data, three different datasets were generated (Figure 2.4A):

- **Fully preprocessed dataset (D1)** – A standard preprocessing pipeline was applied on EEG traces by means of the Matlab toolbox EEGLAB [51]. EEG data was firstly bandpass filtered in the frequency range 0.1- 45 Hz. Then, after the removal of noisy channels, independent component analysis (ICA) was performed in order to identify and remove artifactual sources, such as eye blinks and muscular noise. Finally, the removed channels were interpolated, and the signal was re-referenced to the common average [52].
- **Raw dataset (D2)** — EEG signals were band-pass filtered in the frequency range 0.1-45 Hz.
- **Minimally preprocessed dataset (D3)** – EEG data was bandpass filtered between 4 and 45 Hz to remove the delta frequency range, which generally involves most of the ocular artifacts.

Both dataset D2 and D3 were standardized by subtracting the mean and scaling to unit variance, as it is common practice for ML and DL applications. All EEG related processing, aside from the full preprocessing pipeline of D1, was carried out by means of MNE-Python [53].

2.3.2. EEG segmentation

Each EEG trial belonging to the three datasets refers to an EEG signal acquired during the presentation of a single short video for a specific subject. To be able to train the CNN model, all trials were segmented with sliding windows of 3 seconds length with no overlap (Figure 2.4A). Each segmented window has therefore (14, 384) dimensions, where 14 is the number of channels and 384 is the number of samples, considering a sampling rate of 128 Hz. This kind of cropped strategy has been found to be the most effective for CNN-EEG applications as seen in section 1.3.2, since it allows to produce more training samples and it forces the network to learn more generalized features instead of relying on the differences between single-trials.

2.3.3. Personality binarization

To generate the binary labels for the classification task, the mean value of the scores was used as a separating threshold for each personality trait (Figure 2.3). Specifically, subjects with personality scores below threshold were attributed to the class 0, representing a low expression of the trait, while subjects presenting scores above threshold were associated to the class 1, representing a high expression of the trait. Accordingly, each 3-second EEG window was assigned a 0 or 1 label, depending on the participant the signal belonged to, and the trait considered. In Table 2.3, the obtained counts, in terms of number of participants, of the two classes for each trait are reported. Agreeableness is the only perfectly balanced trait between the two binary classes, while the other traits present a slight imbalance.

Class	**Personality Trait**				
	Extraversion	**Agreeableness**	**Conscientiousness**	**Emotional Stability**	**Openness**
0	21	19	17	18	18
1	17	19	21	20	20

Table 2.3: Binary class counts for each personality trait.

2.4. EEGNet

EEGNet is a compact CNN architecture [2], whose structure is depicted in Figure 2.5. Specifically, three blocks characterize the model. A first block implements temporal and spatial filtering of the input EEG traces through specific convolutional layers and filters. A second block is designed to optimally mix the previously extracted information, time

coded, into feature maps. In the end, a third block implements the classification stage of the model.
A detailed description of EEGNet's architecture is reported in Table 2.4.

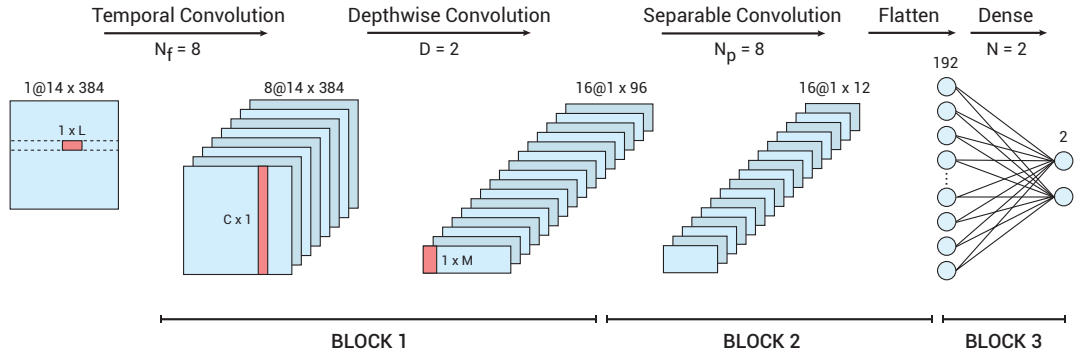

Figure 2.5: High-level structure of EEGNet. Three blocks characterize the network: block 1 representing the temporal and spatial filtering, block 2 representing the mixing of the feature maps, and block 3 representing the classification stage.

2.4.1. BLOCK 1

Block 1 (Figure 2.6) takes as an input a (C, S) vector, with C representing the number of channels of the signal and S the number of time samples. For our data, each window has respectively $C = 14$ channels and $S = 384$ samples (3 seconds window x 128 Hz sampling rate). Then, this block is organized in the following way:

- A first 2D convolutional layer ($Conv2D$). N_f temporal filters with size $(1, L = 64)$ are fitted. The size of the filter is chosen equal to half the sampling rate of 128 Hz. The output of this layer has dimensions $(14, 384, N_f)$, i.e., N_f feature maps representing the EEG windows obtained as output of the temporal filters.
- A batch normalization layer ($BatchNorm$). The output of the previous layer is standardized. This regularization is applied on a batch basis, i.e., the mean and standard deviation used for the normalization are calculated on all the windows defined by the batch size, and not on single inputs.
- A second convolutional layer ($DepthWiseConv2D$). Depthwise convolution fits a spatial filter for each temporal feature map extracted in the previous layer, i.e., for each band extracted by the convolutional layer, the network now fits D spatial filters. D is a depth parameter that specifies the number of spatial filters to be fitted for each feature map. In total, the number of filters for this layer are N_f (number of feature maps) x D (number of spatial filters per feature map). The size of the

spatial filters is $(C = 14, 1)$, to cover all channels. The weights of the spatial filters are regularized with a maximum norm constraint of 1. Since the resulting feature maps from this layer are only connected to their corresponding temporal feature maps from the previous layer, DepthWise2D reduces the number of parameters to be learned. The output of this layer reduces the number of channels C to 1, leading to an output with dimension $(1, 384, N_f \cdot D)$.

- A second batch normalization layer (*BatchNorm*).
- An activation layer (Activation). An exponential linear unit (ELU) function is used to rescale the output of the previous layer. The ELU function (Equation 2.1) performs an identity operation on the positive input values, and an exponential operation on the negative input values.

$$ELU = \begin{cases} x & \text{if } x > 0 \\ e^x - 1 & \text{if } x < 0 \end{cases} \tag{2.1}$$

- An average pooling layer (*AveragePool2D*) with size $(1, 4)$. It reduces the number of samples S by a factor of 4: $(1, 96, N_f \cdot D)$.
- A dropout layer (*Dropout*). Dropout is used to reduce overfitting by randomly dropping units during training. A unit corresponds to a neuron in the neural network. However, in CNNs units are not defined and instead dropout consists in zeroing out columns of weights in the filters.

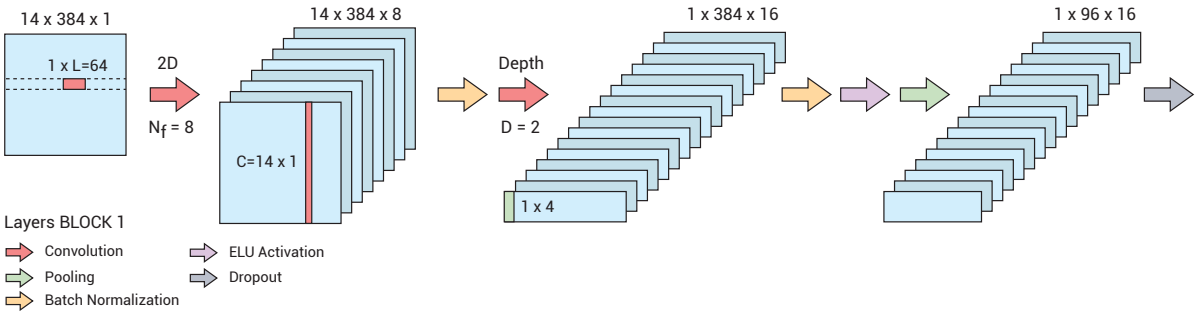

Figure 2.6: Structure of Block 1, EEGNet.

2.4.2. BLOCK 2

Block 2 (Figure 2.7) takes as input the feature maps extracted by the depthwise convolution layer after their samples were reduced by a factor 4 through average pooling. Hence, the input has shape $(1, 384/4, N_f \cdot D)$. Block 2 is then organized as follows:

- A third convolutional layer (*SeparableConv2D*). Separable convolution performs first

a depthwise convolution followed by a pointwise convolution. The depthwise convolution acting on each input channel separately (in this case, the channels are $N_f D$, the output of the dropout layer) fits filters of size $(1, M = 16)$. The depth parameter in this case is the default of 1, thus one filter is fitted for each feature map, for a total of $N_p = N_f \cdot D$ filters. The pointwise convolution mixes the output channels by fitting $N_p(1, 1)$ filters that iterate over every single point. The output of this layer has dimensions $(1, 96, N_p)$. The number of output feature maps is equal to the number of fitted pointwise filters since every feature map represents a different mix of the previous feature maps extracted.

- A batch normalization layer (*BatchNorm*).
- An activation layer (Activation) with ELU activation function.
- An average pooling layer (*AveragePool2D*) with size $(1, 8)$ is used to reduce the number of samples S by a factor of 8: $(1, 12, N_p)$.
- A dropout (*Dropout*) layer.
- A final flatten (Flatten) layer. This layer reshapes the input by flattening its dimensions: from a $(1, 12, N_p)$ 3D tensor to a $(1, N_p)$ 1D vector.

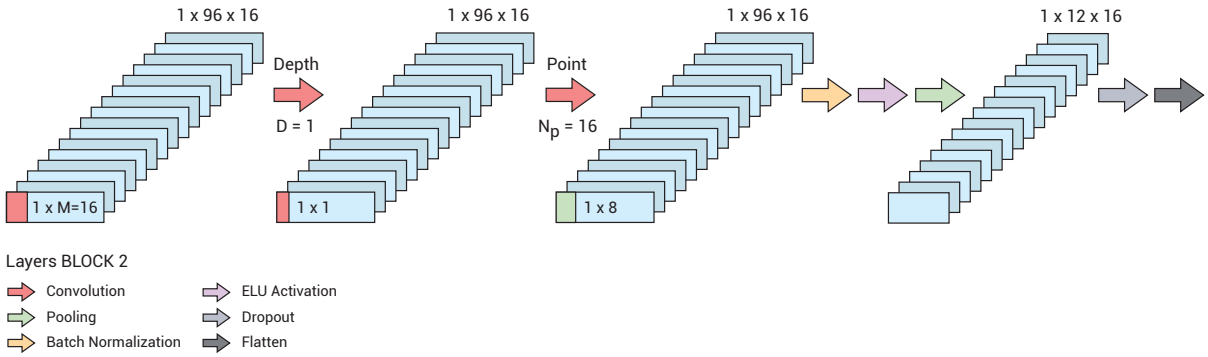

Figure 2.7: Structure of Block 2, EEGNet.

2.4.3. BLOCK 3

Block 3 (Figure 2.8) is the final block which has the function of classification. It takes as input the flattened previous feature maps and is formed by a single layer:

- A dense layer (*Dense*) ends the model's structure. This layer connects the flattened input to the N outputs, with N representing the number of classes in the classification task. A Softmax activation function (Equation 2.2) is used on the N neurons to calculate the probabilities of each class. The probabilities of all classes sum up to 1.

$$predictions = \ softmax\left(\mathbf{x}_i\right) = \frac{e^{\mathbf{x}_i}}{\sum_j^N e^{\mathbf{x}_j}} \tag{2.2}$$

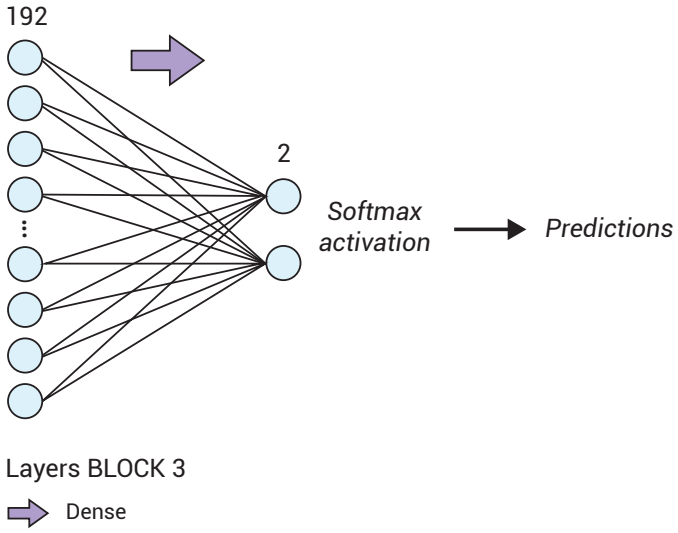

Figure 2.8: Structure of Block 3, EEGNet.

2.5. Model validation

In order to find the optimal hyperparameters and the optimal structure of the EEGNet model for the present classification task, the EEG datasets D1, D2, and D3, and personality dataset, were split into training, validation and test sets with a 70-15-15 proportion, respectively (Figure 2.9). Five random stratified splits, one for each personality trait, were implemented to keep a balanced representation of all the subjects within the sets, i.e., to have for each subject, 70% of their trials into the train set, 15% in the validation set and the remaining 15% of the trials into the test set.

Specifically, the training set is the dataset on which the model is trained, and it's used to update the weights of the model. The validation set, on the other hand, is used to optimize the model's hyperparameters by unbiasedly assessing the performance of the model during training. In fact, the model does not use the validation data to update its weights, but just to assess the performance on unseen data and accordingly tune the hyperparameters during an initial tuning phase. A model is considered to be learning optimally, with no underfitting or overfitting, if the performance on the training set is close to the one on the validation set. The test set, finally, is used to evaluate the performance of the final trained classifier on new unseen data.

Block	Layer	Filters	Size	Parameters	Output	Activation	Default values
1	Input				$(C = 14, S = 384, 1)$		
	Conv2D	N_f	$(1, L = 16)$	$64 \cdot N_f$	$(14, 384, N_f)$	Linear	$N_f = 8$ padding=’same’
	BatchNorm			$2 \cdot N_f$	$(14, 384, N_f)$		
	DepthWiseConv2D	$N_f \cdot D$	$(C = 14, 1)$	$14 \cdot D \cdot N_f$	$(1, 384, N_f \cdot D)$	Linear	D=2 padding=’valid’ max norm=1
	BatchNorm				$(1, 384, N_f \cdot D)$		
	Activation				$(1, 384, N_f \cdot D)$	ELU	
	AveragePool2D		$(1, 4)$		$(1, 96, N_f \cdot D)$		
	Dropout				$(1, 96, N_f \cdot D)$		$p = 0.5$
2	SeparableConv2D	N_p	$(1, M = 16)$	$16 \cdot D \cdot N_f + N_p \cdot (D \cdot N_f)$	$(1, 96, N_p)$	Linear	$N_p = N_f * D$ padding = ’same’
	BatchNorm			$2 \cdot N_p$	$(1, 96, N_p)$		
	Activation				$(1, 96, N_p)$	ELU	
	AveragePool2D		$(1, 8)$		$(1, 12, N_p)$		
	Dropout				$(1, 12, N_p)$		$p = 0.5$
	Flatten				$(1 \cdot 12 \cdot N_p)$		
3	Dense	$N \cdot (N_p \cdot 12)$			N	Softmax	

Table 2.4: EEGNet’s detailed structure.

For the hyperparameter tuning (Section 2.5.1) and structure optimization (Section 2.5.2) tasks, the optimization was performed on the training set and tested on the validation set, with the test set held as a holdout set for the actual classification task.

Figure 2.9: Stratified data partitioning applied on all three EEG datasets, and for all personality traits.

2.5.1. Hyperparameter tuning

Hyperparameters are those parameters that are set before training begins and are thus not learned automatically by the network. For this reason, choosing the best performing hyperparameters is crucial. For this tuning phase, the standard EEGNet structure with $N_f = 8$ temporal filters and $D = 2$ spatial filters, was initially kept as reference, while the general hyperparameters of the model were tuned to find the best performing set of configurations.

The tuning was implemented by means of Keras-Tuner [54] using the *hyperband algorithm* [55]. This latter results much faster than other hyperparameter tuning algorithms thanks to a training resource allocation strategy implemented by using the previously proposed *successive halving algorithm*. *Successive halving* first randomly chooses a subsample of parameter configurations from the search space, it then trains these configurations for a uniformly allocated amount of time, evaluating the performance of the model on each set of parameters, and, finally, it discards the bottom half of worse performing sets. This procedure is repeated by allocating exponentially more training resources to the top half of the best performing parameter sets, training them, and discarding half of the sets again, until convergence to a single optimal parameter set is reached. However, successive halving requires that the number of starting configurations and the maximum training resource time to be fixed in advance, and therefore allocates a proportional resource at each iteration. The main issue with this approach is the need of knowing a priori the amount of resources and configurations that should be considered.

The *hyperband algorithm* (Figure 2.10) solves the trade-off problem by repeating the

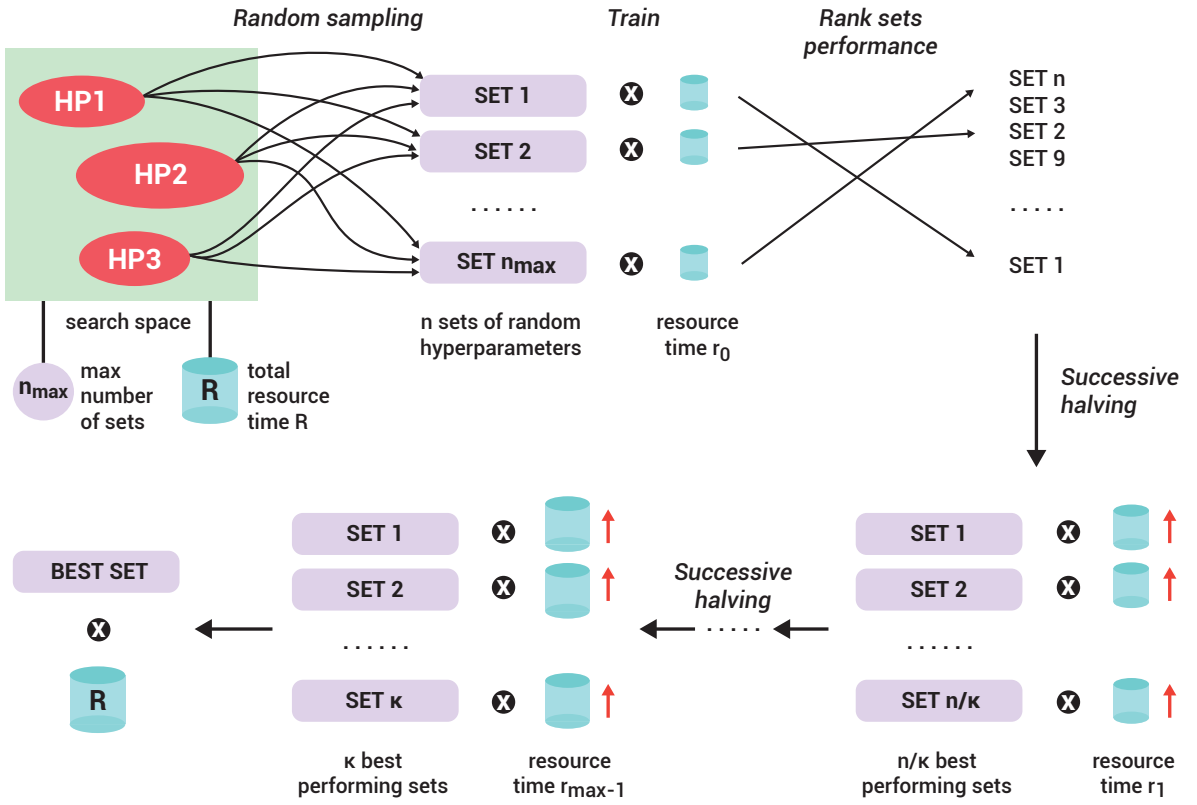

Figure 2.10: Hyperband algorithm scheme for one single bracket. The search space defines the possible values for each hyperparameter (HP). A maximum number of possible configurations n_{max}, i.e., sets of hyperparameters, and of total resource time R are defined. The successive halving loop consists in training the given sets of hyperparameters for the allocated time r_i and keeping only the n/κ best performing sets at each iteration. At each iteration, the number of sets is reduced by a factor κ while the allocated resource training time r_i is increased. At the end, one optimal set of hyperparameters is left.

successive halving procedure for s times, called brackets. Each bracket has a different resource allocation and number of configurations proportion, aimed at considering the limit cases (a) many configurations and small training time, and (b) few configurations and high training time. Given the global training resources R, representing a maximum allocated training time, and the proportion κ of configurations to be discarded in each successive halving round, the number of brackets considered is $(s_{max}+1)$ with $s_{max} = \log_\kappa R$. The algorithm loops through the brackets $s = [s_{max},\ s_{max}-1,\ \ldots, 0]$, each bracket being characterized by its own starting resource time $r_s = R\kappa^{-s}$ and number of starting configurations $n_s = (s_{max}+1)\cdot(\kappa^s/(s+1))$. Successive halving operates in each bracket by discarding the worst performing configurations and keeping the $\frac{n_i}{\kappa}$ best performing ones until one optimal configuration is reached. The best performing parameter set in the

last bracket is reported as the optimal one. A schematic representation of the algorithm on a single bracket can be seen in Figure 2.10.

Hyperparameter	Search space	Sampling
Dropout	min = 0.1 max = 0.7	step=0.1
Dropout Type	"Dropout", "SpatialDropout2D"	choice
Learning rate	min = $10e-3$ max = $10e-5$	logarithmic
Batch size	16, 32, 64, 128, 256, 512	no sampling, grid search

Table 2.5: The search space and sampling method defined for each hyperparameter.

Specifically, in the Keras-Tuner implementation of the algorithm, the global resources R are defined as the maximum number of epochs, with an epoch corresponding to one full cycle through the training dataset. The maximum number of epochs chosen was $R = 1000$, while the reduction factor for the successive halving loop was left to the default of $\kappa = 3$. The objective chosen for the training was the maximization of validation accuracy. All three datasets (i.e., D1, D2 and D3) were used for the tuning phase, while the personality trait chosen was Agreeableness because of its perfectly balanced binary classes. The hyperparameters were optimized on the training set and validated on the validation set. The hyperparameters chosen for optimization and their search space are reported in Table 2.5 and were the following:

- Dropout rate. Dropout consists in randomly dropping units of the network during training by setting the input of the units to 0, to reduce overfitting and improve regularization [56]. In a CNN this corresponds to zeroing out random weights in the filters. The dropout rate represents the fraction of units to be dropped during training. EEGNet default dropout rate is of 0.5. For the dropout search space, values between 0 and 0.7, with a step of 0.1, were evaluated.
- Dropout type. It identifies the choice between the two possible dropout layers implemented by EEGNet: the classic “Dropout” and “SpatialDropout2D”. “SpatialDropout2D”, unlike the regular dropout, drops entire feature maps at once, and not just a subset of weights.
- Learning rate. It scales the magnitude of weight updates. For the learning rate search space, values between $10e-3$ and $10e-5$ with log sampling were considered.

For the batch size, i.e., the number of samples (number of 3-second EEG windows) to pass to the network at once, a *grid search* was performed after the hyperparameter tuning

with the hyperband algorithm. Specifically, a *grid search* consists in looping through all combinations defined by the search space and selecting the best performing one. It is an exhaustive but time-consuming search algorithm, optimal for small search spaces. For the batch size, a discrete search space of 16, 32, 128, 256, and 512 samples was considered.

2.5.2. Structure optimization

To test the optimal structure of EEGNet on the training and validation sets, the hyperband algorithm was applied to the N_f temporal filters and D spatial filters parameters after fixing the hyperparameters to the optimal values identified with the procedure described in Section 2.5.1. The search space of the filters was defined as follows:

- Temporal filters. N_f values ranging from 2 to 12.
- Spatial filters. D values ranging from 1 to 8.

Afterwards, a simple grid search was performed by fixing the spatial filters $D = 2$ and testing the performance of the model by varying the temporal filters N_f from 1 to 12. This choice was made to compare the performance of the standard structure with other similar structures by limiting the number of spatial filters. This limitation is also useful for constraining the number of trainable parameters, and thus training time, and to ease subsequent interpretation of the learned features.

2.6. Training strategy

For the classification task, the standard structure of EEGNet, with $N_f = 8$ and $D = 2$ filters, was initially chosen and compiled with the optimal hyperparameters found as described Section 2.5.1. The model was trained using the labeled EEG 3-second windows as inputs while a 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. For this approach, a five-fold cross-validation training was used as explained in Section 2.6.1.

All models were trained on an NVIDIA GeForce RTX 2070 GPU in Tensorflow [43] 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 those weights were used to evaluate the models on the test set. Each EEGNet model was fit using the Adam optimizer [57] with default parameters (i.e., first and second order moments equal to 0.9 and to 0.999, respectively). This optimizer looks for the best model

parameters by minimizing the categorical cross-entropy loss function (Equation 2.3) which calculates the cross-entropy between the predicted classes $\mathbf{y}$ and the true target classes $\mathbf{t}$. For a binary classification, it is equivalent to the binary cross-entropy loss:

$$Cross - entropy = L(\mathbf{y}, \mathbf{t}) = -\sum_{i=1}^{2} \mathbf{t}_i \ln \mathbf{y}_i = -t_i \log(y_i) - (1 - t_i) \log(1 - y_i) \quad (2.3)$$

2.6.1. Five-fold cross validation

For the window-wise classification, EEGNet models were trained for each personality trait and for each EEG dataset, using the optimal parameters found during the tuning phase. The goal of this classification strategy was to obtain a robust classifier able to predict single EEG windows.

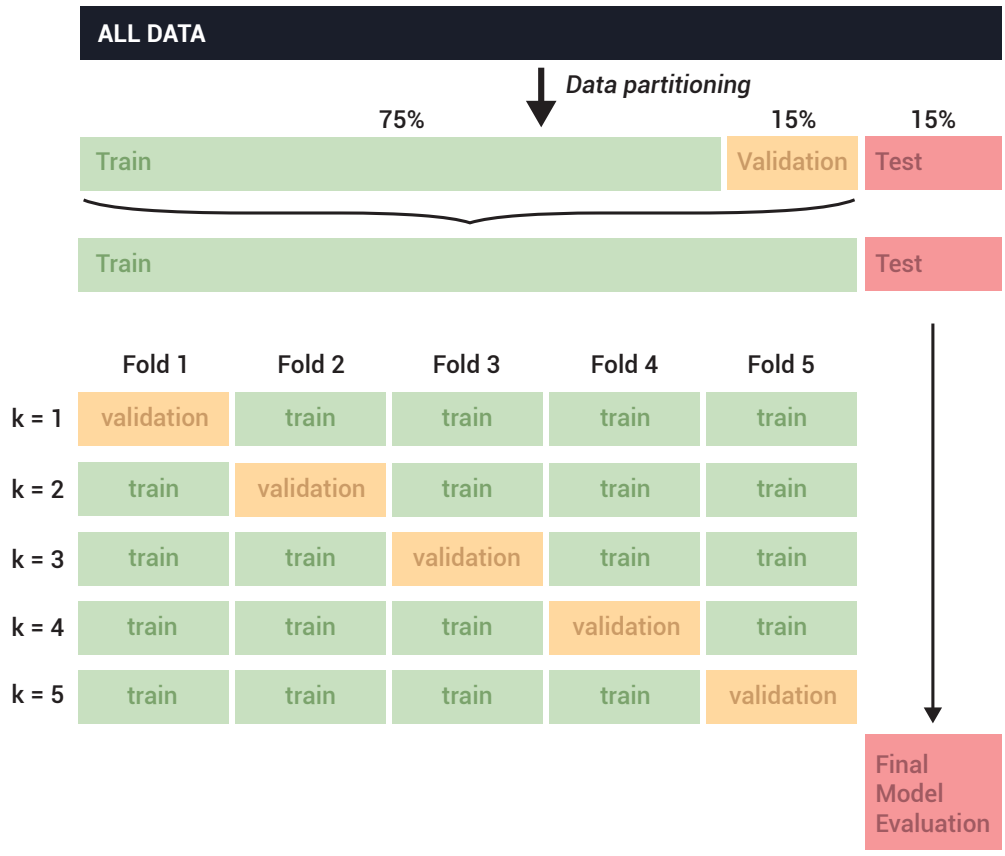

Figure 2.11: Five-fold cross-validation scheme.

For the final training, a five-fold stratified cross-validation strategy was employed. The training and validation sets, obtained as explained in Section 2.5, were combined to form the training set while the test set was kept aside for the final evaluation.

Five-fold cross-validation is a training technique used to test the performance of a model over all training data. It performs a more reliable evaluation of the model performance since it uses all the available data for both training and testing without wasting information. Instead of arbitrarily splitting the data into training, validation, and test sets, with five-fold cross-validation the training data is split randomly into five equally sized subsets with four subsets used for training, and the last holdout subset used for testing. This procedure is repeated for five iterations, and in this way all the training set is used to train and test the model. At the end of each iteration, the model is evaluated on the test set with the metric of choice. At the end of training, every model will have five different test evaluation results whose average should be a good representation of the real performance of the model over the whole range of training data.
Stratified five-fold cross-validation performs a stratified, instead of a random, split of the data into the subsets. In this way, the target class ratio is maintained across all subsets for all iterations. A representative scheme of five-fold cross-validation strategy is showed in Figure 2.11.

2.6.2. Evaluation metrics

The performance of the models was evaluated in terms of accuracy (Acc) and F1 score (F1) metrics. A classification algorithm usually yields an error due to the discrepancy between the predicted class and the real class. This error takes two forms: a false positive (FP) case when a predicted positive class is in reality negative, and a false negative (FN) case when a predicted negative class is in reality positive. True positives (TP) and true negatives (TN) on the other hand represent correct predictions of the positive and negative classes, respectively. A confusion matrix is an N x N matrix, where N is the number of predicted classes, representing the FP, FN, TP, and TN counts and it's used to evaluate the performance of a classification model. The confusion matrix for a binary classification problem is pictured in Figure 2.12.
Starting from the confusion matrix, several metrics can be calculated. Accuracy represents the proportion of correct classifications and is calculated as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

Accuracy does not consider the prediction capabilities of the model for each class, but lumps together all correct predictions. It can easily be seen how this can yield misleading results when the dataset is unbalanced, i.e., when one of the two classes weighs more than the other. For this reason, despite accuracy being the most used classification performance

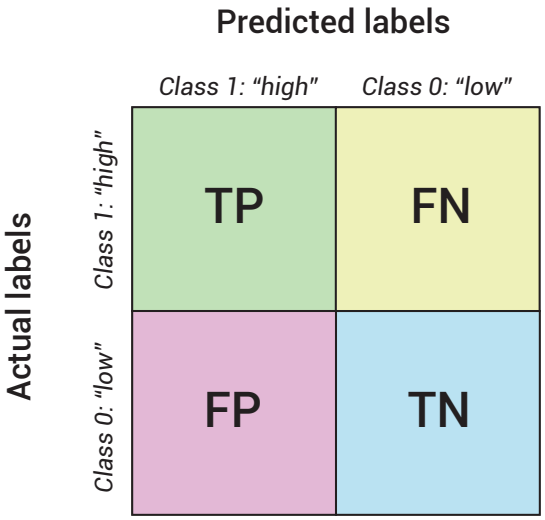

Figure 2.12: Confusion matrix for binary classification. The predicted classes counts and the actual classes counts are represented respectively on the columns and rows of the matrix.

metric, other metrics are usually preferred. To get a better understanding of the model's performance over both classes, the F1 score metric is used:

$$F1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity} = \frac{2TP}{2TP + FP + FN} \tag{2.5}$$

where precision (Equation 2.6) represents the fraction of positive samples (TP) correctly classified over all positive predictions (TP + FP), while sensitivity (or recall) (Equation 2.7) represents the fraction of positive samples (TP) correctly classified over actual positives (TP + FN):

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{2.7}$$

For the window-wise classification, the metrics were obtained on the single window classifications. Class 1 was used as the "positive label" class in all cases. Therefore, a window classified correctly as class 1, was considered a TP, a window classified incorrectly as class 1, was considered an FP, and so on. The metrics were then calculated on all the predictions of the hold-out test set as obtained in the five-fold cross-validation data partitioning strategy (Figure 2.11).

2.7. Feature interpretability

2.7.1. Visualization of learned filters

The filters learned by EEGNet, i.e., the weights assigned to each filter kernel, can be extracted and analyzed. For this analysis, the EEGNet structure with $N_f = 4$ temporal filters and $D = 2$ spatial ones was chosen for an easier interpretation of the extracted features.

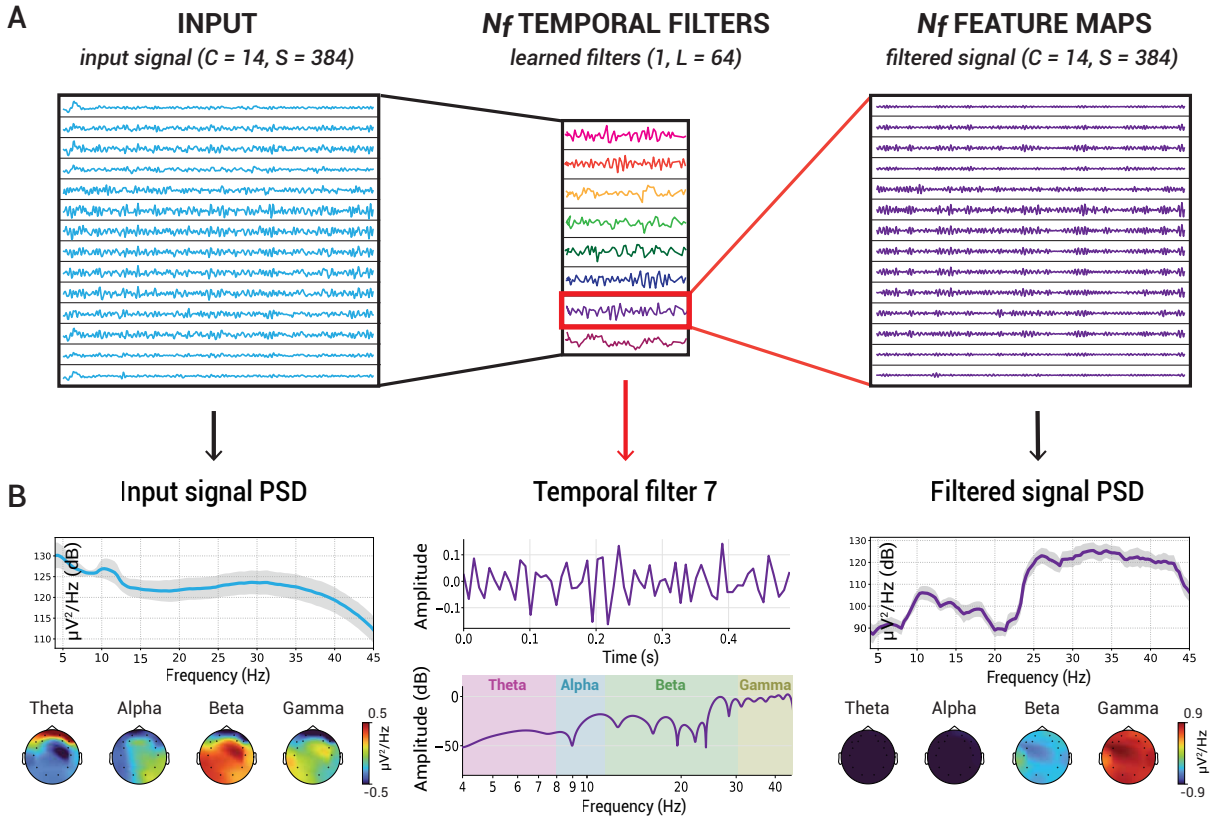

Figure 2.13: Analysis of temporal filters. A. The first convolutional layer of EEGNet: the (C = 14, S = 384) input window, the N_f (pictured 8) temporal filters with size (1, L = 64) and one output feature map, specifically the one obtained from the convolution of the input with filter number 7. B. The frequency analysis of the input and output signals. The PSD amplitude before and after filtering and the topographic maps of the four characteristic bands in the 4-45 Hz range (Theta, Alpha, Beta, Gamma) are displayed. The amplitude response of filter 7 is plotted in the time and frequency domain. In this example, filter 7 is a gamma high-pass filter and the PSD of the output shows that all frequencies below the cutoff of around 28 Hz, are filtered out.

Temporal filters

For the N_f temporal filters of the first convolutional layer, a frequency-based analysis was carried out. Since the temporal filters extract specific frequency bands from the signal, the power spectral density (PSD) of the input and output of the layer was calculated and represented both in terms of amplitude and as band specific (Theta, Alpha, Beta, and Gamma bands) topographic maps over the scalp. Both the input and output have shape ($C = 14,\ S = 384$) which lends itself to this specific type of channel-based topographic analysis (note that the output maintains the shape of the input due to padding). The N_f filters with size ($1,\ L = 64$) were considered as finite impulse response (FIR) filters and their amplitude response was plotted both in the time and frequency-domain. Specifically, in the time domain, the filter has a length of 0.5 seconds since its kernel size ($L = 64$) is half that of the sampling frequency of 128 Hz, while in the frequency domain, the covered range is limited to 45 Hz, as the last useful frequency due to the pre-processing bandpass filtering. An example analysis of the first convolutional layer and the learned temporal filters, is illustrated in Figure 2.13.

Spatial filters

For the spatial filters, D spatial filters for each N_{f_i} feature map of size ($C = 14, 1$), topographic maps were plotted to locate the electrodes characterizing the specific band extracted by the N_{f_i} temporal filter. An example of the spatial topographies of two spatial filters characterizing the same N_{f_i} feature map is shown in Figure 2.14.

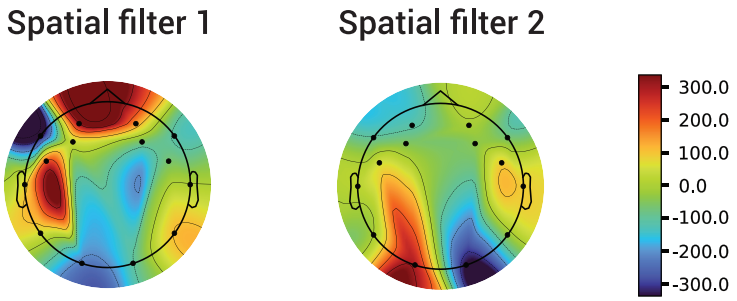

Figure 2.14: Example of two spatial filters associated to one temporal filter.

2.7.2. Deactivation of learned filters

To gather the importance for the prediction of each single temporal filter, the performance of the models was evaluated on the test set after deactivating each temporal filter one by one. Specifically, the models considered was the EEGNet-4,2 configuration. The resulting

performance is expected to be lower than the full model one. The deactivated filters that cause a higher decay in performance, are the ones carrying more information for the classification task, and are considered as relevant. Analyzing then the frequency response of the more relevant temporal filters, ideally it could be inferred which of the EEG bands are more relevant for the classification of each personality trait.

2.7.3. Attribution methods

Attribution algorithms aim at explaining how a neural network makes predictions by assigning an *attribution*, that is, a relevance or contribution, value to each element of the input [58]. Considering the input x with shape $(14,\ 384)$, representing a 3-second EEG segment, the trained network outputs $y_N(x) = [y_1(x), y_2(x)]$ predictions, i.e., one prediction for each binary class. The attribution algorithm determines the contribution $R = [R_1, \ldots, R_{CxS}]$ that each input feature x_i has on the output. If we imagine an image with $(14\ \times\ 384)$ pixels as the input, then each pixel will be assigned a contribution value based on how much it influences the output prediction. In the same way, for our input signal, $(14\ \times\ 384)$ attribution values are assigned.
Rearranging the attribution values in the shape of the input, an attribution map is created, and it can be best displayed as a heatmap. In Figure 2.15, an attribution map of the input is represented. The attribution maps visually highlight which EEG channels contribute positively (red color) or negatively (blue color) to the prediction.
Two main approaches to the attribution problem exist: a perturbation-based and a backpropagation-based approach [58]. Perturbation-based methods make perturbations to features of the input, by masking or removing them, and run a forward pass simulation to observe the impact that perturbation has on the output. The difference between the non-perturbed output and the perturbed output represents the attribution value of the input features altered. Perturbation-based methods are however computationally expensive since they require a new forward propagation for each perturbation. Backpropagation-based methods compute the attributions for all input features in a single forward-backward pass and are generally faster.
DeepLIFT [59] is a backpropagation-based attribution algorithm. It backpropagates attributions from the output to every unit i until it reaches the input. Each unit i is assigned an attribution that depends on the relative difference between that same unit activated at the original network with input x and at the modified network with reference input $\bar{x} = 0$. The general rule for determining the input attributions $R(x)$ (Equation 2.8) [58].

$$R_i(x) = (x_i - \bar{x}_i) \cdot \frac{\partial^g y_N(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}} \tag{2.8}$$

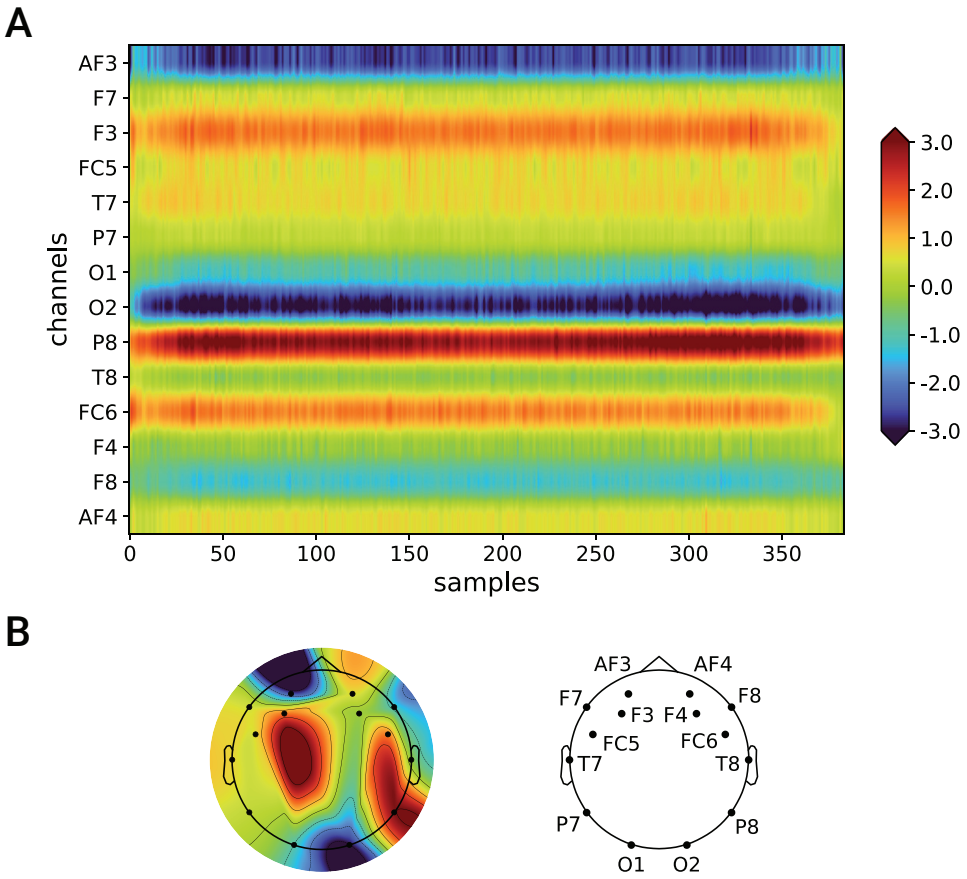

Figure 2.15: Attribution map visualization. A. Example of an attribution map of the (14 channels, 384 samples) input obtained with the DeepLIFT algorithm. Red color highlights the input channels that contribute positively to the prediction, while blue color highlights the input channels that contribute negatively instead. B. The same attribution map represented topographically on the scalp, averaged in time, and the corresponding Emotiv EPOC electrode layout.

i.e., DeepLIFT is equivalent to a feature-wise product between the difference of the input and reference input $(x_i - \bar{x}_i)$ and the partial derivative $\frac{\partial^g y_N(x)}{\partial x_i}$. With $g = \frac{f(z)-f(\bar{z})}{z-\bar{z}}$ being the ratio between the difference in output and the difference in input at each nonlinearity $f()$, for a network fed with the real input x and the reference input $\bar{x}$.
For each EEGNet model, trained to classify a specific trait,the input attributions were computed with the DeepLIFT algorithm implemented in the Deep Explain [58] framework in Python. Characteristic attribution maps were then obtained and plotted as spatial topographies, namely a general attribution map for the specific trait, obtained by averaging the attribution maps over all subjects and in time. The obtained attribution maps were then visually inspected to determine which channels contributed positively or negatively to the prediction.

3 Results and discussion

In this chapter, the results obtained are reported and discussed. For the classification task and the EEGNet model validation, the hyperparameter tuning results are presented in section 3.1, the while the five-fold cross-validation classification performance is reported in section 3.2. The feature interpretation and visualization results are reported in section 3.3. To refer concisely to the structure of EEGNet considered, the notation $\mathbf{EEGNet - N_f, D}$ will be used, where N_f is the number of temporal filters and D is the number of spatial filters.

3.1. Model validation

3.1.1. Hyperparameter tuning

For the validation of the hyperparameters of the standard EEGNet-8,2 structure, all models were trained using the Agreeableness labels as it's the only trait with balanced class counts among the subjects.

Optimization of dropout, dropout type and learning rate

An initial tuning was performed on the hyperparameters dropout, dropout type, and learning rate for all three datasets D1, D2, and D3. A total of 2072 hyperparameter combinations, or trials, were tested by the tuning algorithm for each dataset and their performance was evaluated by the validation set accuracy metric. The top 10 best performing trials are reported in Table 3.1.

For the *dropout* parameter, its values were varied between 0.1 and 0.7, with a step of 0.1. It can be noted that, for all datasets, the top 10 trials were obtained for low dropout values in the range of 0.1-0.4, with 0.1 being the most frequently selected one. Indeed, by expanding the number of trials considered and counting the number of selections made by the tuning algorithm for each dropout value evaluated, the same pattern is observed. In Figure 3.1, the counts of the dropout values selected in the top 100 best performing trials are reported. The frequency of selection of 0.1 dropout value is almost the double of the one observed for 0.2, 0.3, and 0.4. Dropout values of 0.5 and 0.6 are selected less than 10

D1: Preprocessed			
Dropout	**Dropout Type**	**Learning Rate**	**Accuracy**
0.1	SpatialDropout2D	0.00034	0.7771
0.3	Dropout	0.00085	0.7751
0.3	Dropout	0.00052	0.7568
0.4	Dropout	0.00050	0.7568
0.1	Dropout	0.00022	0.7536
0.1	SpatialDropout2D	0.00072	0.7529
0.1	SpatialDropout2D	0.00035	0.7445
0.2	SpatialDropout2D	0.00023	0.7432
0.2	SpatialDropout2D	0.00056	0.7425
0.1	SpatialDropout2D	0.00031	0.7425
D2: Bandpass 1 - 45 Hz			
Dropout	**Dropout Type**	**Learning Rate**	**Accuracy**
0.1	Dropout	0.00029	0.9257
0.1	Dropout	0.00029	0.9195
0.3	Dropout	0.00049	0.9167
0.4	Dropout	0.00076	0.9127
0.2	Dropout	0.00081	0.9099
0.1	Dropout	0.00069	0.9099
0.1	Dropout	0.00058	0.9099
0.2	Dropout	0.00081	0.9071
0.1	Dropout	0.00069	0.9065
0.1	Dropout	0.00058	0.9048
D3: Bandpass 4 - 45 Hz			
Dropout	**Dropout Type**	**Learning Rate**	**Accuracy**
0.2	Dropout	0.00086	0.9606
0.1	Dropout	0.00044	0.9543
0.3	Dropout	0.00096	0.9465
0.1	Dropout	0.00100	0.9460
0.1	Dropout	0.00048	0.9450
0.1	Dropout	0.00075	0.9431
0.1	Dropout	0.00055	0.9416
0.2	Dropout	0.00034	0.9411
0.4	Dropout	0.00076	0.9411
0.4	Dropout	0.00076	0.9411

Table 3.1: Top 10 best performing hyperparameter trials on the datasets D1, D2, and D3 for dropout, dropout type, and learning rate.

times, while the dropout value of 0.7 was never selected. As for the dropout-dependent performance, in Figure 3.2 the average accuracy and the corresponding standard deviation for all dropout values on all trials is shown. The highest average accuracy on all three datasets is obtained for the dropout value of 0.1. The average accuracy decreases as the dropout value increases, with an exception found for dataset D3, where the dropout value of 0.3 has an average better performance than the dropout value of 0.2.

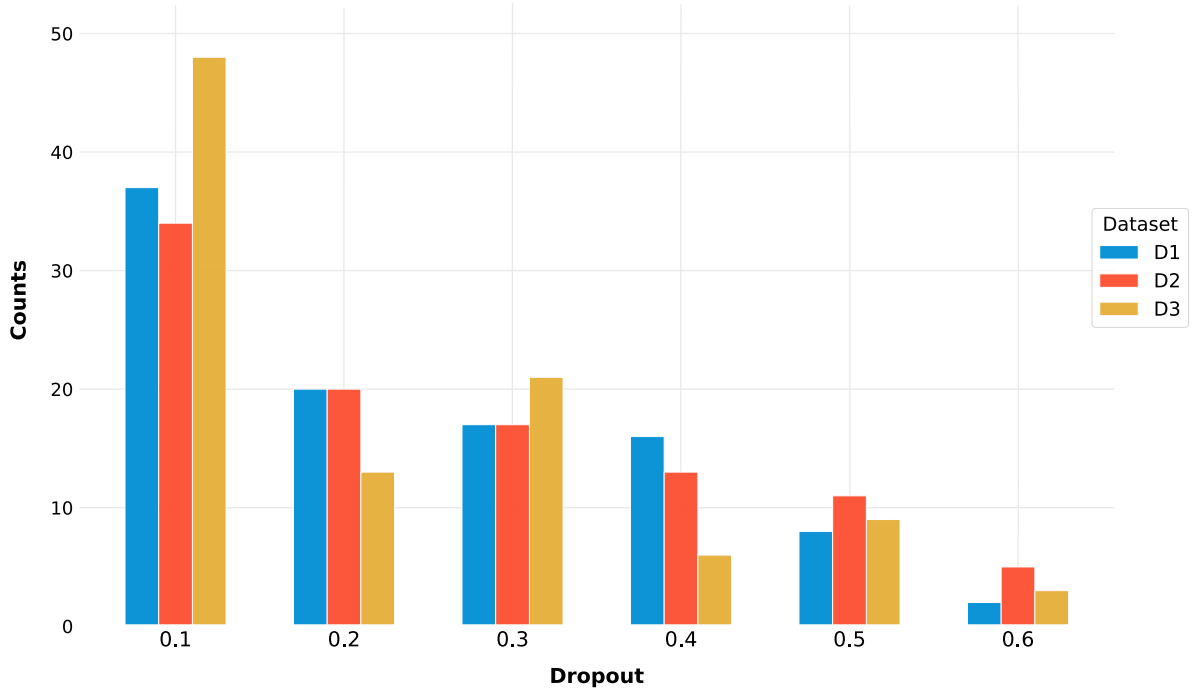

Figure 3.1: Dropout counts in the top 100 best performing trials on all datasets.

For the *dropout type*, it can be observed that for datasets D2 and D3, the standard "Dropout" layer is selected unanimously, while for dataset D1, the "SpatialDropout2D" layer is selected about half of the times. Indeed, in the top 100 best forming trials for D1, the "Dropout" layer is selected 54 times and the "SpatialDropout2D" layer is selected 46 times.

Concerning the *learning rate*, it assumes several different values over the trials, but always in the order of $10e-5$. In the top 100 best performing trials, an average learning rate of 0.00051 is obtained for dataset D1, 0.00050 for dataset D2, and 0.00069 for dataset D3. The distributions of the learning rate selections made in the top 100 best performing trials and their mean values is represented in the boxplot in Figure 3.3.

Finally, it can be noted that the accuracy performance on dataset D3 is higher than the one on datasets D1 and D3 for all hyperparameter combinations (Figure 3.2).

The final hyperparameter configuration chosen for EEGNet-8,2 is reported in Table 3.2.

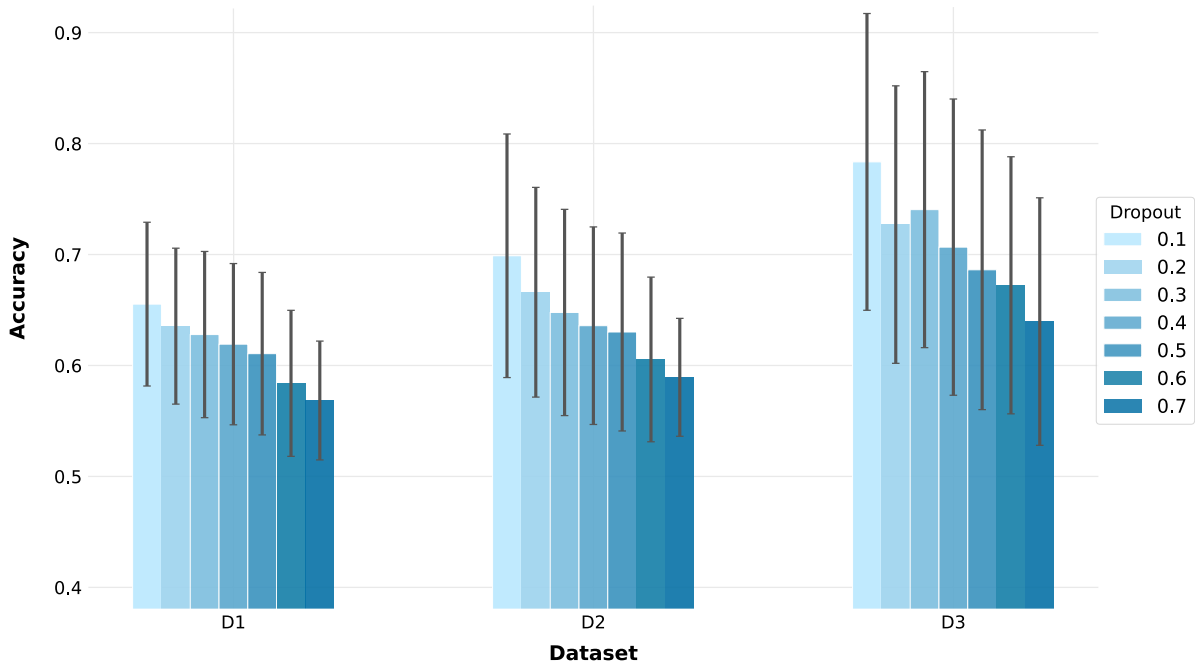

Figure 3.2: Average accuracy and standard deviation (black bars) for the different dropout rates for all 2072 trials.

Specifically for the dropout, a value of 0.1 was chosen since it resulted the best overall value for all three datasets, in terms of model performances. Concerning the dropout type, the "Dropout" layer was selected, since it was unanimously selected for dataset D2 and D3 and represented more than 50% of selections in the top 100 best trials for dataset D1. For the learning rate, an average of 0.00057 is obtained for the top 100 trails over all three datasets. Therefore, a learning rate of 0.0001 that maintains the same scale was chosen for conventional reasons.

EEGNet-8,2	
Dropout	0.1
Dropout Type	Dropout
Learning Rate	0.0001

Table 3.2: Final hyperparameter configuration selected for EEGNet-8,2.

Optimization of batch size

The *batch size* was optimized on D3 since it's the dataset returning the best accuracy performance with the selected hyperparameters (Table 3.2) for EEGNet-8,2. A qualitative

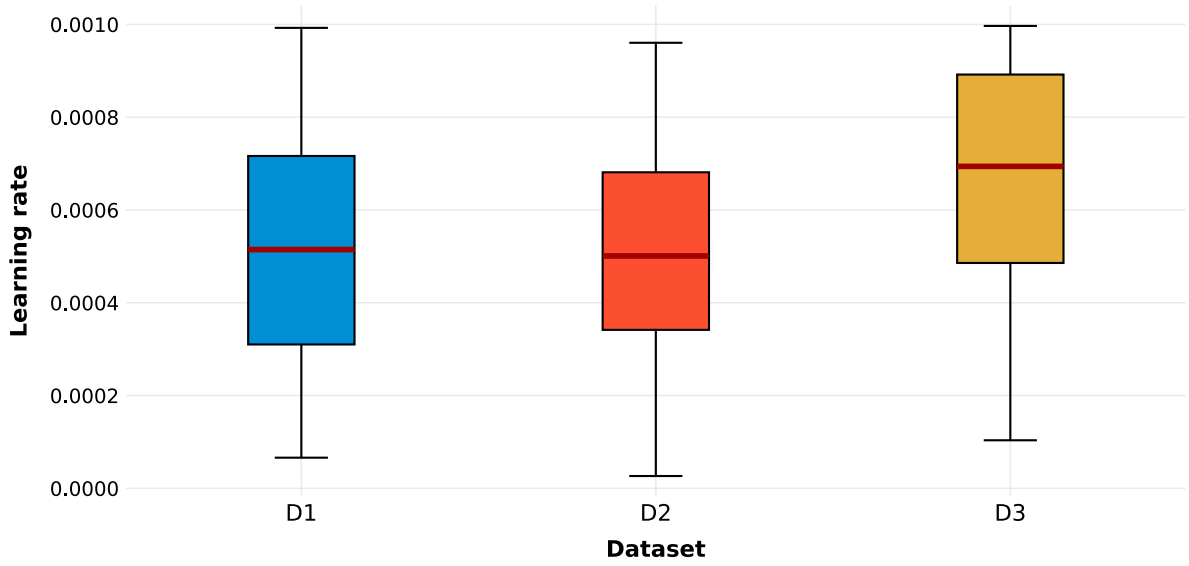

Figure 3.3: Distribution of the learning rate values chosen in the top 100 best performing trials on all datasets and their respective means.

rather than quantitative evaluation was carried out by comparing the accuracy and loss training and validation learning curves, for batch sizes of 16, 32, 64, 128, 256, and 512 sample windows, with the model trained for 1000 epochs (Figure 3.4).
Usually, low batch sizes are selected with low learning rates in order to get the best performance out of model [60]. Here, having the learning rate already fixed at 0.0001, which is relatively low, a lower batch size is expected to give a better classification performance. Indeed, as can be seen by the accuracy learning curves in Figure 3.4, batch sizes of 16, 32, and 64 present higher accuracy. However, from the same learning curves, both for the accuracy and loss, it can be observed that the smaller batch sizes present also a rather noisy behavior, while higher batch sizes of 128, 256, and 512 samples converge to a more stable model. The higher batch sizes, presenting less noise in the training and validation, have lower variance in classification accuracy and loss compared to lower batch sizes. This behavior is most likely due to the fact that the EEG signals in D3 are noisy, and the sample windows are small. Therefore, if the model updates its weights on a low batch of windows, it's most likely to pick up noise, while if it updates its weights on a high batch of windows, the present noise gets canceled out thanks to the high sample size.
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. This choice is motivated by the fact that both datasets D2 and D3 are composed of either raw or minimally pre-processed signals, making the windows noisy and higher batch sizes more stable.

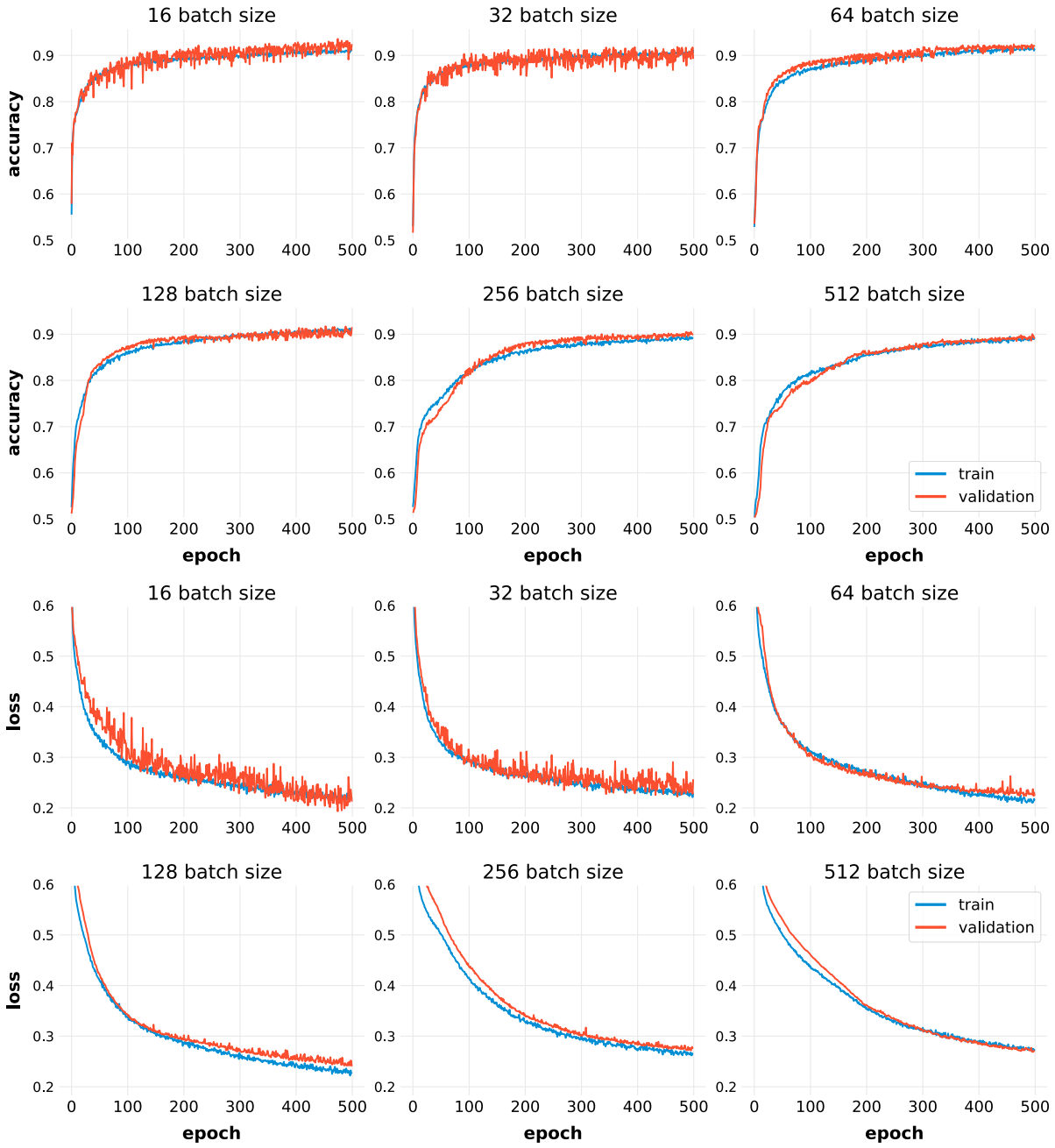

Figure 3.4: Learning curves of the EEGNet-8,2 model trained with 16, 32, 64, 128, 256, and 512 batch size on the Agreeableness trait of dataset D3.

3.1.2. Validation of EEGNet structures

Optimization of number of filters

To evaluate the performance of the model, other EEGNet-N_f, D structures with $N_f \neq 8$ temporal filters and $D \neq 2$ spatial filters, were evaluated. In Table 3.3, the top 5 best

performing trials from a hyperparameter search among N_f values ranging from 2 to 12 and D ranging from 1 to 8, while keeping a fixed dropout rate of 0.1, learning rate of 0.0001 and batch size of 256, are reported.
It can be observed that the best performing trials are obtained by using the highest number of filters available in the search space, i.e., $N_f = 12$ and $D = 8$, as evidenced in bold in Table 3.3. This result is 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 and makes a possible interpretability of the extracted features more difficult. Therefore, a reasonable trade-off between number of filters/parameters and performance should be identified. In Table 3.4, a comparison among four different EEGNet structures, and their corresponding number of trainable parameters is reported.

D1			D2			D3		
Nf	**D**	**Accuracy**	**Nf**	**D**	**Accuracy**	**Nf**	**D**	**Accuracy**
4	6	0.739	**12**	**8**	0.944	9	**8**	0.961
12	7	0.736	**12**	**8**	0.939	9	7	0.955
9	5	0.734	**12**	**8**	0.938	8	6	0.953
7	3	0.723	10	6	0.930	**12**	**8**	0.948
12	5	0.716	**12**	3	0.918	**12**	4	0.942

Table 3.3: Top 10 best performing hyperparameter trials on the datasets D1, D2, and D3 for the number of filters N_f and D.

Structure	# Parameters
EEGNet-4,2	794
EEGNet-8,2	1714
EEGNet-9,8	9956
EEGNet-12,8	15578

Table 3.4: Number of trainable parameters for EEGNet-4,2, EEGNet-8,2, EEGNet-9,8 and EEGNet-12,8 structures.

Optimization of number of temporal filters

To evaluate the model performances in function of the number of temporal filters by fixing the number of spatial filters $D = 2$, N_f was varied from 1 to 12, with the hyperparameters

set as in Table 3.2. All models were trained on dataset D3 as it is the dataset returning the best classification performance. The number of spatial filters is limited to 2 arbitrarily, since it is both the optimal standard number of filters for EEGNet, and it limits the number of parameters a priori, easing the interpretability. In Figure 3.5, the performance of each model is reported. It can be noted that for N_f values higher than 4 the performance tends to improve as the number of temporal filters is increased reaching almost a plateau. Therefore, since the performance of EEGNet with number of temporal filters higher than 8 does not improve drastically, the standard EEGNet-8,2 model, on which the hyperparameters were optimized, was chosen for the final classification analysis.

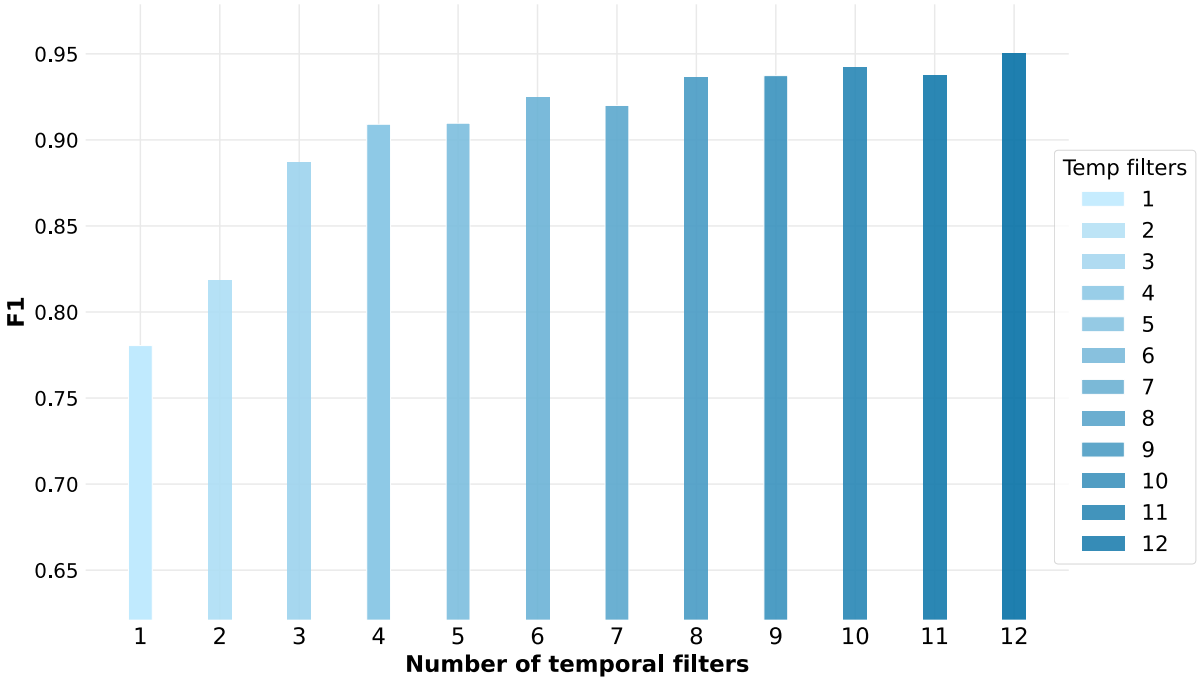

Figure 3.5: EEGNet-N_f, 2 models performance in terms of F1 score.

3.2. Classification

For the classification, the results of the five-fold cross-validation on the test set for EEGNet-8,2 with the optimized hyperparameters (Table 3.2), for each trait and for each dataset D1, D2, and D3 are reported in Table 3.5. The performance was evaluated in terms of accuracy, precision, sensitivity and F1 in each fold. The final score for each metric was obtained by averaging the scores over the five folds.

As already observed from hyperparameter tuning results on the Agreeableness trait (Table 3.1), the highest performance is achieved on dataset D3 for all traits with an average accuracy ranging from 0.89 to 0.92, while the lowest performance is found for dataset D1,

with a much lower accuracy ranging from 0.75 to 0.79. Dataset D2 gives an intermediate performance between the two instead, with accuracy ranging from 0.83 to 0.92 for the five traits.

More specifically, the traits presenting the highest performance both in terms of average accuracy and F1, are Agreeableness (Acc = 0.93, F1 = 0.93) and Extraversion (Acc = 0.92, F1 = 0.91) for dataset D3. Those two traits also present the lowest standard deviation among the five folds, as it can be seen in Figure 3.6. The other three traits for dataset D3, present a slightly lower performance: Conscientiousness (Acc = 0.90, F1 = 0.91), Emotional Stability (Acc = 0.89, F1 = 0.90), and Openness (Acc = 0.89, F1 = 0.89).

Dataset D1, aside from having the lowest performance, also presents the highest standard deviation between the folds (Figure 3.6). The best performing trait for D1, is Conscientiousness (Acc = 0.79, F1 = 0.82) but it's still substantially lower than the performance of the same trait for datasets D2 and D3.

Dataset D2 presents comparable performance to D3 for Agreeableness (Acc = 0.89, F1 = 0.91), Conscientiousness (Acc = 0.87, F1 = 0.89), and Emotional Stability (Acc = 0.89, F1 = 0.89), while it gives a lower performance for the traits Extraversion and Openness.

An example of accuracy and loss train and validation learning curves for the Agreeableness trait on all three datasets, is pictured in Figure 3.7. Dataset D2, which is composed of raw data, presents the most noise for its validation curve. On the contrary, dataset D1 presents the smoothest curves and dataset D3 shows little noise as well. Since in dataset D3 the Delta band is filtered out, we can assume that most of the noise is concentrated in the 0.1 - 4 Hz range. Other studies have compared the performance of deep learning models on EEG data with different levels of pre-processing [61] and a similar result is found, i.e., the deep learning model performs better on minimally pre-processed data. Studies applying deep learning on minimally pre-processed EEG data for EEG-based classification tasks, have reported higher performance than state-of-the-art classification [62–65].

The precision and sensitivity values are also reported in Table 3.5 to check for imbalance in the predictions that the accuracy metric alone cannot catch. For instance, a low sensitivity would indicate that the classification is unable to detect high expressions of the given trait. Both the precision and sensitivity scores are comparable, and their combined balanced score, F1, is also comparable to the accuracy score.

Only one other study [46] used the same AMIGOS dataset for a personality trait binary classification task. The present study reports a higher overall classification performance for all the five traits - Extraversion (0.92 compared to 0.84), Agreeableness (0.93 compared to 0.87), Conscientiousness (0.90 compared to 0.84), Emotional Stability (0.89 compared to 0.84) and Openness (0.89 compared to 0.73). Moreover, in [46] low scores for sensitivity

D1: Preprocessed				
Trait	**Accuracy**	**Precision**	**Sensitivity**	**F1**
Extraversion	0.759	0.736	0.754	0.745
Agreeableness	0.755	0.774	0.749	0.761
Conscientiousness	0.791	0.810	0.831	0.820
Emotional Stability	0.783	0.815	0.784	0.799
Openness	0.771	0.797	0.771	0.784
D2: Bandpass 1 - 45 Hz				
Trait	**Accuracy**	**Precision**	**Sensitivity**	**F1**
Extraversion	0.856	0.834	0.893	0.862
Agreeableness	0.891	0.880	0.939	0.908
Conscientiousness	0.865	0.888	0.886	0.887
Emotional Stability	0.886	0.918	0.859	0.887
Openness	0.834	0.890	0.802	0.844
D3: Bandpass 4 - 45 Hz				
Trait	**Accuracy**	**Precision**	**Sensitivity**	**F1**
Extraversion	0.921	0.906	0.917	0.911
Agreeableness	0.930	0.917	0.948	0.932
Conscientiousness	0.897	0.905	0.908	0.906
Emotional Stability	0.892	0.905	0.887	0.895
Openness	0.889	0.888	0.899	0.893

Table 3.5: Five-fold cross-validation results for EEGNet-8,2. Average accuracy, precision, sensitivity and F1 for each trait and each personality trait.

are reported for Conscientiousness and Openness, differently from what observed in the present study.

3.3. Feature interpretability

In this paragraph some preliminary results related to the visualization of the learned features are reported. In order to facilitate the interpretability of the analyzed learned filters and hidden layer outputs, a EEGNet-4,2 structure was considered since, as seen in Figure 3.5, its performance is comparable to EEGNet-8,2. Moreover, the model was trained on dataset D3 since it is the dataset resulting in the highest classification performance.

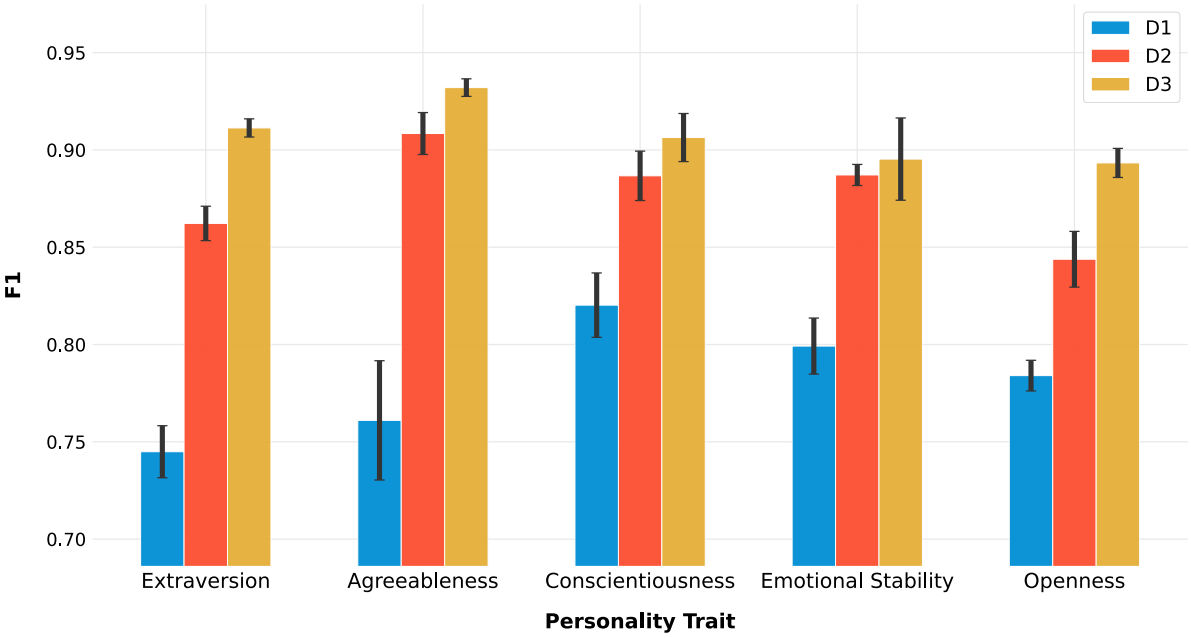

Figure 3.6: Average five-fold cross-validation F1 and standard deviation (black bars) for all five traits for each dataset.

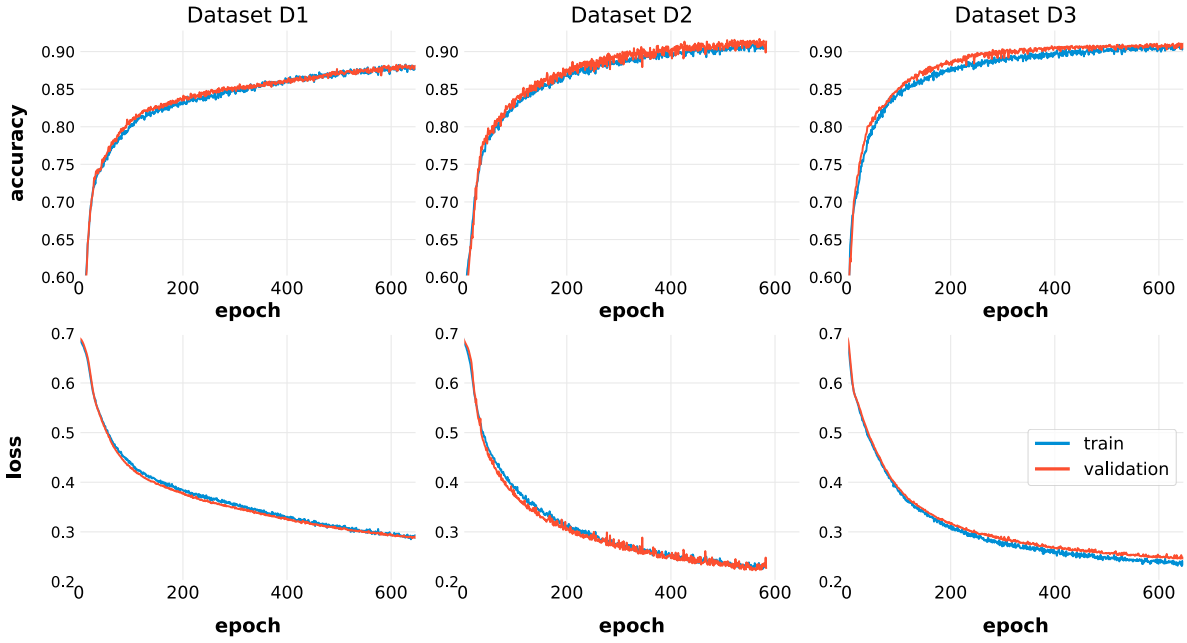

Figure 3.7: Train and validation learning curves of the accuracy and loss for the three datasets on the Agreeableness trait.

3.3.1. Performance with deactivated temporal filters

In order to evaluate the relative importance of the four learned temporal filters on the classification performances, three of them were deactivated at a time by setting their

weights to zero, while keeping only one temporal filter active. The resulting classification performances obtained on the test set for traits Extraversion, Agreeableness, and Emotional Stability are reported in Table 3.6. The results for the remaining traits are reported in Appendix A.1.
For the Extraversion trait, two temporal filters give a relevant classification result when kept active on their own. Temporal filter 1 (Extraversion-1) alone accounts for 0.67 accuracy and 0.66 F1 score, while temporal filter 2 (Extraversion-2) accounts for 0.62 accuracy and 0.65 F1 score but has lower precision of 0.55. For the Agreeableness trait, only one filter, temporal filter 4 (Agreeableness-4), accounts for a 0.66 accuracy and 0.71 F1 score. For the Emotional Stability trait, temporal filter 2 (Emotional Stability-2) accounts for 0.69 accuracy and 0.76 F1 score, but with very high sensitivity of 0.93. The representation of the relevant filters and their outputs is reported in section 3.3.2. For the remaining traits (Appendix A.1), no single active temporal filter gave any relevant result (Table A.1). This result could be explained by the fact that all four filters are somewhat relevant for the prediction and a combination of two or more filters is needed to gain any relevant accuracy. It is likely that the model needs information for a wider range of frequencies than the ones extracted by the single temporal learned filters.

3.3.2. Visualization of relevant filters and their outputs

The relevant filters identified in section 3.3.1 and their relative outputs, are described and analyzed in this section.

Extraversion

The response amplitude in the time and frequency domain of the temporal filters Extraversion-1 and Extraversion-2 , and the PSD of their average output, averaged over the 14 channels and divided by the two classes based on the inputs labels, are represented in Figure 3.8. Extraversion-1 attenuates frequencies below 15 Hz, approximatively. Thus, it can be associated to a poorly selective high-pass filter in the beta and the gamma frequency ranges. The average PSD of the output of this filter shows a slight difference between the outputs of inputs labeled as class 0 and inputs labeled as class 1, in the 15 - 45 Hz range.
Extraversion-2 behaves as a low-pass filter, preserving information in the theta and alpha bands and attenuating the beta and gamma bands. The average output shows a slight difference between the two classes in the theta-alpha range.
The two spatial filters associated to both Extraversion-1 and Extraversion-2 are represented in Figure 3.9 and Figure 3.10, respectively. The spatial filters localize the bands extracted by the temporal filters in specific areas of the brain based on the channels that

Extraversion				
Active filter	**Accuracy**	**Precision**	**Sensitivity**	**F1**
All	0.898	0.895	0.874	0.884
1	**0.673**	**0.614**	**0.708**	**0.658**
2	**0.617**	**0.548**	**0.792**	**0.648**
3	0.444	0.444	1.000	0.615
4	0.556	0.000	0.000	0.000
Agreeableness				
Active filter	**Accuracy**	**Precision**	**Sensitivity**	**F1**
All	0.887	0.871	0.911	0.891
1	0.505	0.505	1.000	0.671
2	0.505	0.505	1.000	0.671
3	0.581	0.927	0.184	0.306
4	**0.663**	**0.629**	**0.809**	**0.708**
Emotional Stability				
Active filter	**Accuracy**	**Precision**	**Sensitivity**	**F1**
All	0.882	0.879	0.898	0.888
1	0.525	0.525	1.000	0.688
2	**0.687**	**0.639**	**0.929**	**0.757**
3	0.525	0.525	1.000	0.688
4	0.475	0.000	0.000	0.000

Table 3.6: Model performance on the traits Extraversion, Agreeableness, and Emotional Stability by keeping one filter active at a time. Relevant filters highlighted in bold.

are filtered out and on the ones that are preserved. Just like for the temporal filters, the learned filters have an approximative behavior and for this reason it is not always possible to isolate the bands, for the temporal filters, or the channels and areas, for the spatial filters, making the interpretation difficult.

The spatial filters associated to Extraversion-1 localize the extracted beta and gamma bands in the left occipatal area and right frontal one (spatial filter 1) and in the area identified by channels F3, FC6 and P8 (spatial filter 2). The spatial filters associated to Extraversion-2 localize the theta-alpha band in the right frontal, central and right parietal area (spatial filter 1) and in the central FC6 area (spatial filter 2).

EEG-based studies have correlated Extraversion to increased posterior versus frontal delta and theta activity at centerline electrode sites [66] which reflects activity in the rostral anterior cingulate cortex and is linked to dopaminergic function [13]. This finding seems in part reflected by the filter Extraversion-2 which extracts the theta band (Figure 3.8)

and by spatial filter 1 which localizes the extracted band in the central to frontal centerline area (Figure 3.10).

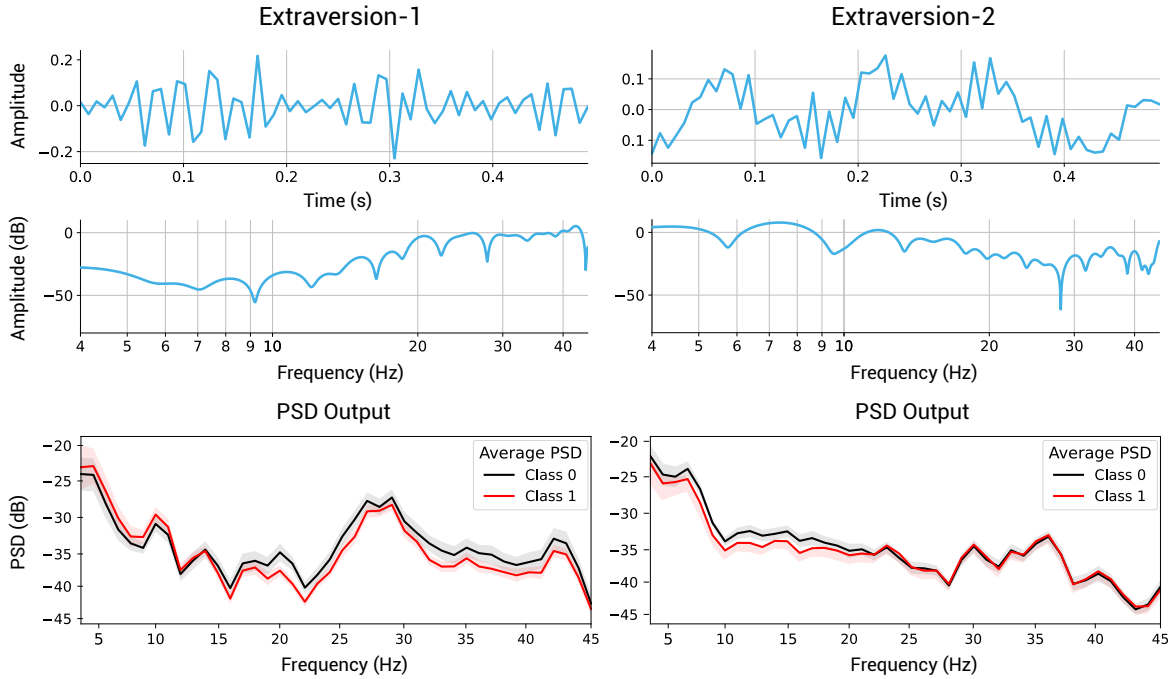

Figure 3.8: Temporal filters Extraversion-1 and Extraversion-2 and their relative average PSD output.

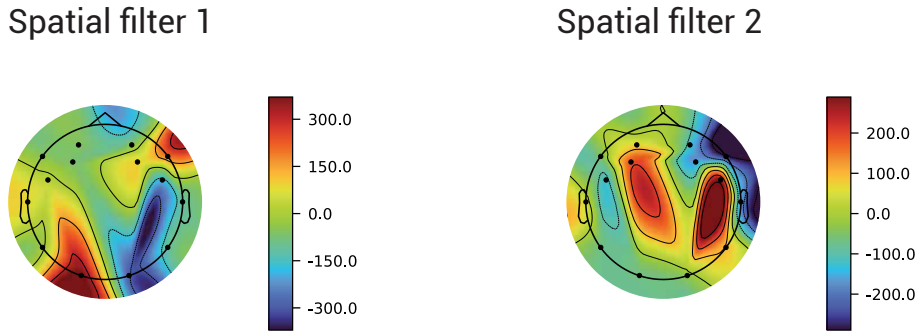

Figure 3.9: Spatial filters associated to Extraversion-1.

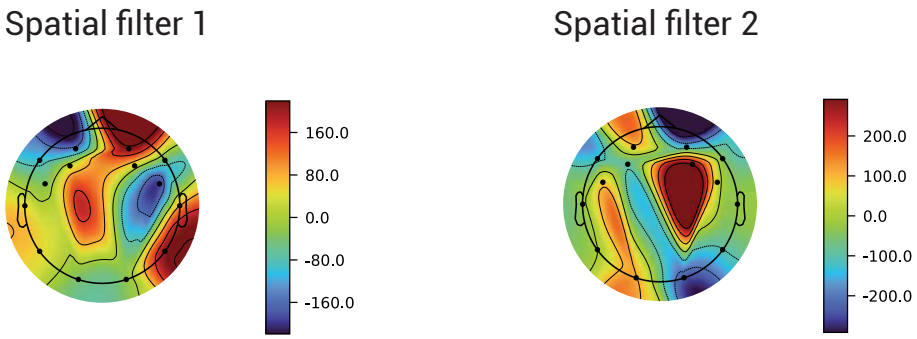

Figure 3.10: Spatial filters associated to Extraversion-2.

Agreeableness

The temporal filter Agreeableness-4 and its average PSD output are represented in Figure 3.11.

Agreeableness-4 behaves as a notch filter, attenuating frequencies in the high theta band in the 6 - 7 Hz range, and as a low-pass filter attenuating frequencies above 20 Hz in the beta band. The output of the filter does not show evident differences in PSD between class 0 and class 1.

The two spatial filters associated to Agreeableness-4 are represented in Figure 3.12. Spatial filter 1 localizes the extracted theta, alpha and beta band frequencies in the temporal and occipital region. Spatial filter 2 localizes the extracted band in the frontal F7 and F8 area and in the right occipital area.

Few studies on Agreeableness exist, but its facet of empathy has been correlated to EEG theta and alpha band oscillations related to emotional processing and mirroring [67, 68]. This finding appears to be reflected by filter Agreeableness-4 (Figure 3.12) which indeed extracts the theta and alpha bands and results as the most relevant filter found for the Agreeableness trait prediction. Another association, as seen in Section 1.1.2, linking Agreeableness to the temporal region, specifically the left temporal superior sulcus [12] is supported by spatial filter 1 (Figure 3.12) which filters that specific region.

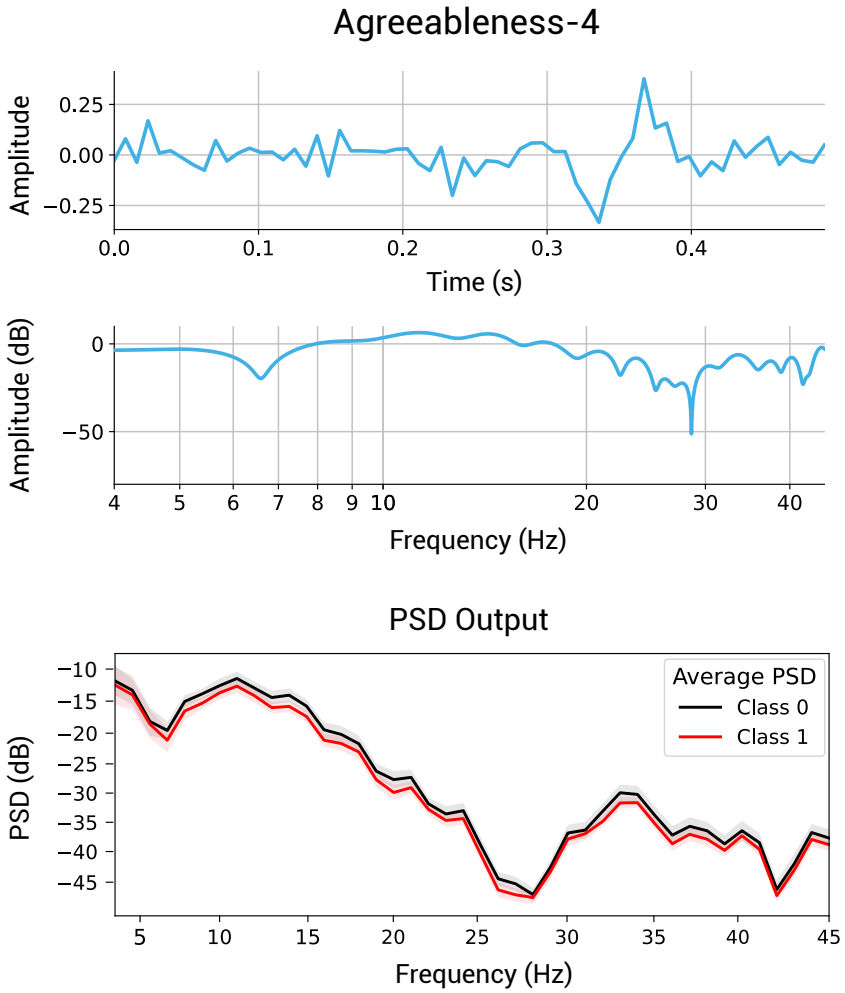

Figure 3.11: Temporal filters Agreeableness-4 and its relative average PSD output.

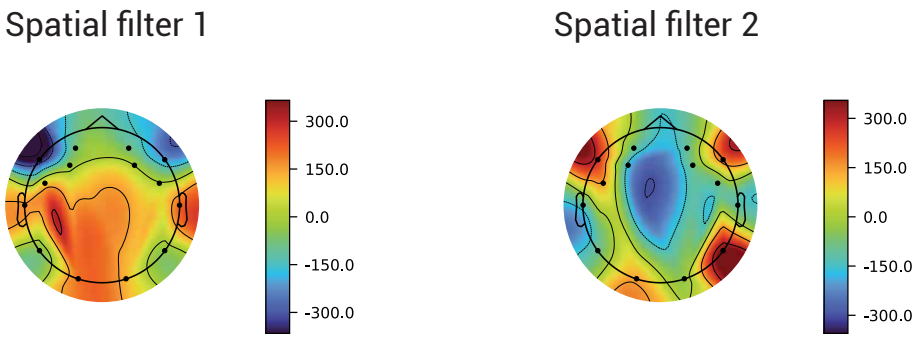

Figure 3.12: Spatial filters associated to Agreeableness-4.

Emotional Stability

The temporal filter Emotional Stability-2 and its average PSD output is represented in Figure 3.13.

Emotional Stability-2, unlike the previous temporal filters, does not extract specific frequency ranges and preserves almost the full EEG spectrum. Indeed, out of the identified relevant filters, Emotional Stability-2 accounts for the highest accuracy of 0.69 and F1 score of 0.76 and this is most likely due to the fact that classification-relevant information is carried by the full spectrum of frequencies, making a filter that just attenuates but does not filter out any band, more relevant overall. The output of the filter does not show any relevant difference in PSD between class 0 and class 1.

The two spatial filters associated to Emotional Stability-2 and their average outputs are represented in Figure 3.14. Spatial filter 1 is focused on the frontal region and left temporal-parietal region, and filters out the central and right temporal contributions. Spatial filter 2 localizes the extracted frequencies from the filter in the left temporal-parietal area.

Neuroticism, the polar opposite of Emotional Stability, as seen in Section 1.1.2 has been linked to activity in the temporal-parietal region, during a recognition task of other people's mental states [17]. The temporoparietal junction is associated indeed to emotional functions and is implicated in the perception of emotional expressions, empathy, and affective memories [17] which aligns with the negative emotionality associated with Neuroticism. In this study, a temporal-parietal spatial filtering can be observed for both spatial filters associated to the most relevant filter Emotional Stability-2, suggesting that the trained model is able to localize and extract spatially coded information relevant for the prediction of the trait.

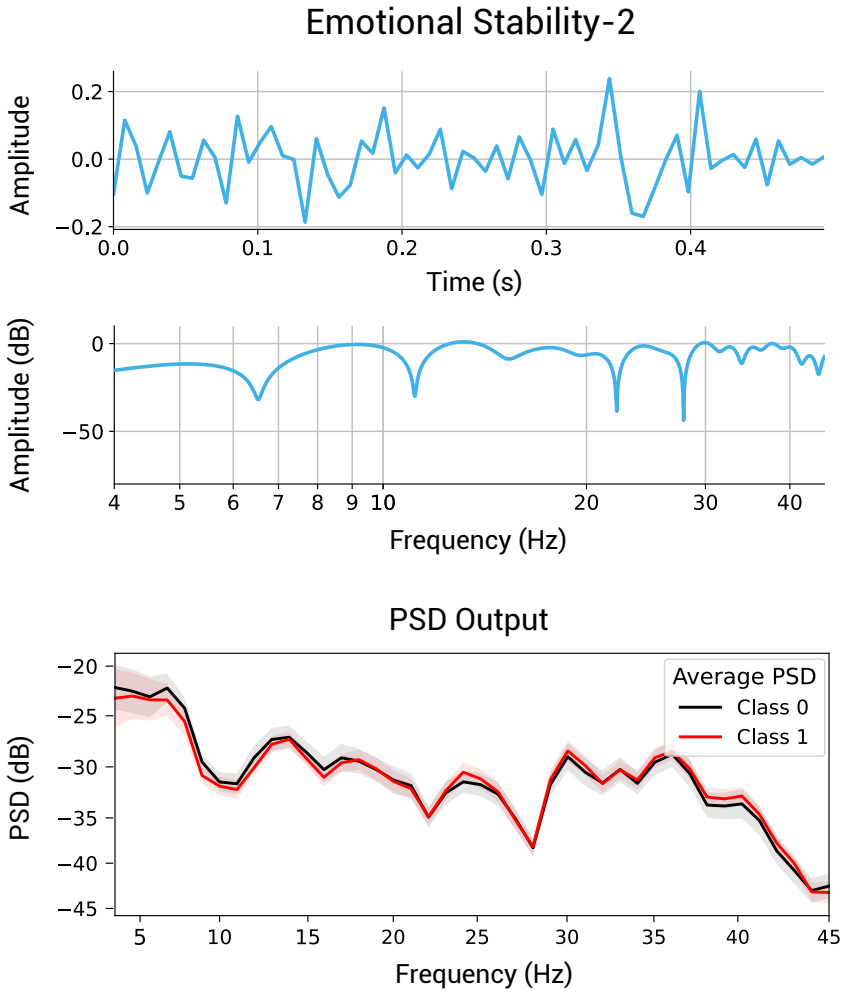

Figure 3.13: Temporal filters Emotional Stability-2 and its relative average PSD output.

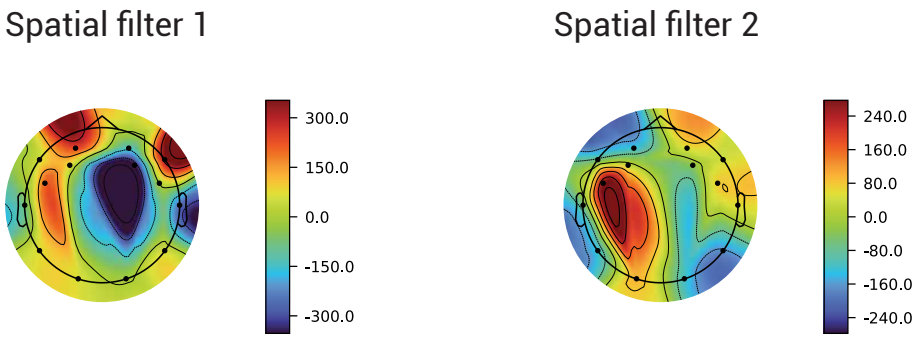

Figure 3.14: Spatial filters associated to Emotional Stability-2.

3.3.3. Attribution maps

The average attribution maps over all subjects for each trait are reported in Figure 3.15.

For Extraversion, the right prefrontal area, specifically channel F8, is identified as having the highest positive contribution for the prediction of the trait. For Agreeableness, the right temporal area, channels T8 and FC6, have the highest positive contribution, while the left frontal area has a lower but still positive contribution. For Conscientiousness, the positive contributions areas aren't so clearly defined, with only the frontal AF4 area having a high contribution, while the left temporal-parietal area and the right temporal area have a lower positive contribution. Emotional Stability has positive contribution areas corresponding to the right temporal-parietal region delimited by channels P8, F3 and FC5. The trait Openness has positive contribution identified in the posterior frontal area identified by channels F3 and F4.

For all traits, a negative contribution of the occipital area (specifically channel O2) to the classification task is observed. This behavior could be related to the attempt of the network to neglect a pronounced activity at the occipital lobe, which is though not relevant for the classification of personality traits [12]. Indeed, since the occipital lobe is associated to the visual cortex [69] and EEG signals are acquired while the subjects are intent in watching videos, it is reasonable that an enhanced activity in this area is captured in the signal. However, it's likely that that activation does not have a correlation with personality as a whole and thus the network learns to ignore the contribution given by the occipital area for all traits. Other negative contributions are given by the left frontal area for trait Emotional Stability and the frontal area for Openness.

The predictive contribution of the frontal PFC region for the traits Extraversion, Agreeableness and Conscientiousness (positive) and for Emotional Stability and Openness (negative) is in line with neuroscientific studies of personality as seen in Section 1.1.2. This result suggests that electrical brain activity in the frontal region may be informative enough for personality classification and may lend itself to portable applications implemented with wearable commercial EEG devices.

Other positive attribution regions are also in line with correlations found between brain areas and personality traits. Temporoparietal junction is correlated to Emotional Stability [17], as seen in Section 3.3.2, and the attribution map shows indeed a positive contribution for that region. Openness has been associated to activity in the posterior medial frontal cortex [13] and its attribution map shows a positive predictive for the posterior frontal region at channels F3 and F4. Finally, the activity of subcortical regions might be picked up by the EEG signal recording and it may be reflected in the obtained attributions.

For instance, the amygdala which has been associated with Extraversion and Emotional Stability (Section 1.1.2). Activity in the amygdala has been associated to frontal EEG asymmetry [70], suggesting that the EEG-based attributions found in the frontal region are not exclusively caused by activity in the PFC.

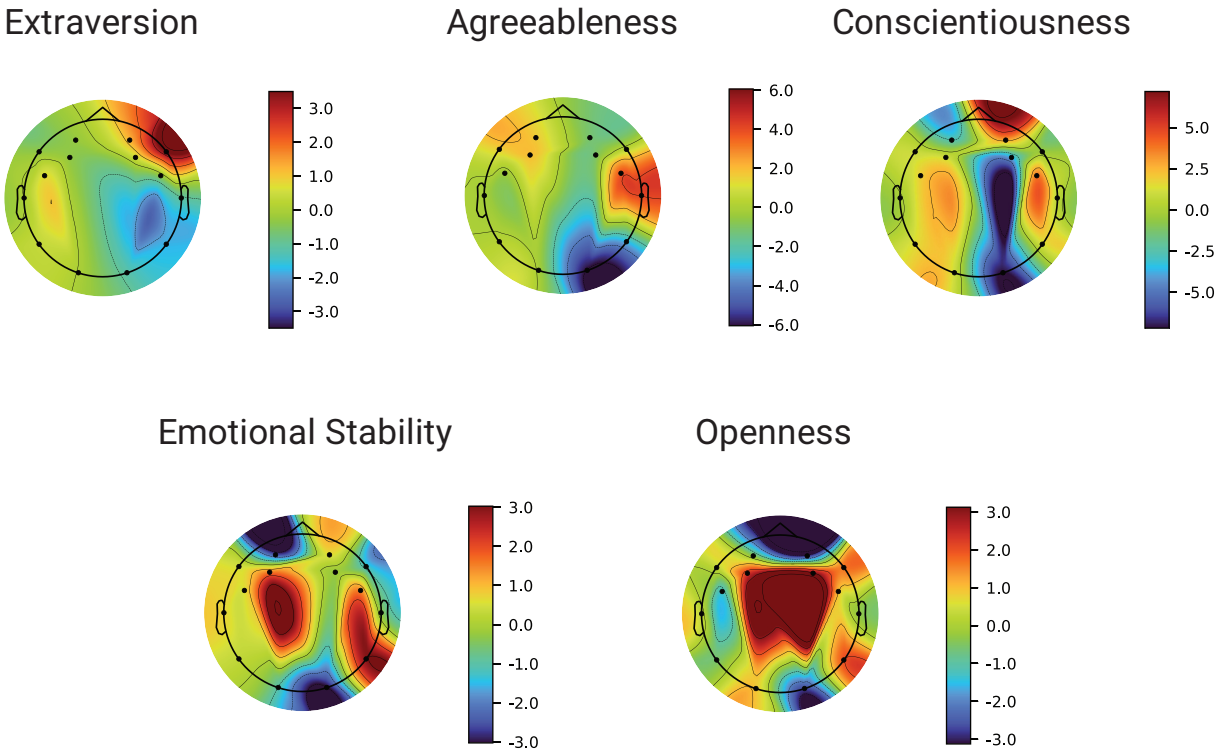

Figure 3.15: Average attribution maps for the five personality traits.

4 | Conclusions and future developments

In this work, a deep learning-based binary classification of personality traits starting from EEG signals was developed. For this purpose, EEGNet, a state-of-the-art CNN-EEG model, was adopted. Particular attention was focused on three aspects: i) the optimization of the structure and parameters of the model, ii) the identification of the best EEG pre-processing pipeline in terms of final classification performances and, iii) a preliminary interpretation of the more relevant EEG-based features automatically extracted by the network.

Concerning the structure of the model, the present study demonstrated that EEGNet could achieve good classification performances even by reducing the complexity of the first convolutional layer, composed of a bank of temporal filters. Specifically, it was shown that with the number of filters ranging from 4 to 12 the classification performances are comparable. This may constitute an advantage in terms of computational cost, since decreasing the number of trainable parameters by reducing the number of filters, also lowers the training time. Moreover, a lower complexity of the model also facilitates the interpretation of the automatically extracted features. In this work, the performances of the network on three differently pre-processed datasets (i.e., D1, D2, and D3) were also evaluated. Specifically, it was shown that the selected EEGNet model obtains its best classification scores on dataset D3, which was minimally preprocessed by the removal of the Delta band by a band-pass filtering and standardization. Comparable performances were obtained on raw data (dataset D2), while a classification accuracy decrease was observed on full pre-processed data (dataset D1). Such result is in line with the studies that have also tested the potential of DL models to directly learn from raw data [62–65].

In terms of pure classification performance, good results were obtained both in terms of accuracy and F1 score. The best performing models were obtained on traits Agreeableness and Extraversion, with test accuracies of 0.93 and 0.92, respectively. The models trained on the other traits also report good performance with accuracy of 0.90 for Conscientiousness and 0.89 for Emotional Stability and Openness. The performance achieved is higher

than most other attempts at classification of personality traits from EEG signals found in literature, all studies based on manual feature extraction strategies. Comparable studies report best classification accuracies in the range of 0.73 - 0.87 [46], 0.64 - 0.96 [47] and 0.66 - 0.74 [44] for most traits.

Another limitation could reside in the type of training employed. The training and test datasets used for the five-fold cross-validation, as they were formulated in this work, proportionally contain EEG traces from all subjects equally. This approach could introduce a bias in classification performances since the network is able to capture all the subject-specific variability during training. In order to assure a higher generalizability of the trained model, a leave-one-subject-out training strategy would be more suitable as it completely excludes one subject at a time from the training.

Finally, a preliminary interpretation of the automatically extracted features is presented. The feature interpretability, indeed, remains one of the main limitations in deep learning-based applications. Specifically, in this work different visualization techniques for the learned temporal and spatial filters of the first two blocks of EEGNet and their outputs are adopted for each personality trait. Moreover, in order to study the relative importance of the temporal filters in the classification task for each trait, different combinations of filters were simultaneously deactivated, and the resulting performance was evaluated. This approach allowed to isolate one or two most relevant temporal filters for the classification of each trait.

Nevertheless, the interpretation of the obtained results remains difficult. The learned temporal filters, for example, most times do not selectively preserve specific EEG frequency bands. Therefore, the association between different frequency traits and the relative power of EEG traces in its standard rhythms is not trivial. Concerning spatial information, an attribution-based visualization was adopted in order to highlight the most relevant EEG electrodes for the classification of each trait. Obtained results enhance the importance of frontal regions and are in line with neuroscientific-based studies on personality.

In conclusion, a personality trait classifier that outperforms other known applications was developed by means of EEGNet. The model was further tested on differently pre-processed EEG data, and it was shown that DL-based models trained on raw and minimally pre-processed signals perform better than models trained on fully pre-processed ones, a finding could simplify the time-consuming EEG pre-processing pipelines. The structure and hyperparameters of the model were also analyzed in depth and limits and consideration on EEGNet's performance were drawn. Finally, given the architecture of the model, it was shown how the hidden layers and learned filters can be analyzed for feature interpretability purposes.

Despite these promising results, the adopted classification approach presents some limi-

tations, such as the binary classification of personality. Although a binary classification approach reduces the level of complexity of the model, it does not take into account the complexity of personality. Indeed, classified traits represent a spectrum on which most people fall in the middle. Therefore, for a more precise personality assessment it could be more suitable to reformulate the problem as a multi-class classification or a regression task. Moreover, in this study the classifiers were trained separately for each trait, not taking into account the higher-level correlations between the traits, as they have been identified in the Big Five theory. Implementing a model that could classify all the five traits at once could be helpful to gain more insight into the relationship between the traits.

Concerning the interpretation of learned features, though EEGNet allows to isolate most relevant filters and areas for the classification of the different traits, a direct association of specific EEG-based features to personality is not possible from our results. This limitation is both due to the insufficient well-supported evidence on the neurological bases of personality, and to the fact that classification is performed on windows extracted from continuous EEG data. Therefore, no information about stimulus and expected response are available. Moreover, the types of stimuli were not taken into consideration. This was a conscious choice since the main focus of the present work was on classification and on testing the ability of EEGNet to classify data with different levels of pre-processing and recorded in response to different types of stimuli. However, differentiating between the stimuli, as they are classified in the dataset on the arousal-valence scale, for example by training on category-based subset of data and comparing the performance between categories, could be a valuable future step for understanding the features extracted by the network for the various personality traits.

Among other possible future developments for the classification, aside from adopting more conservative training strategies as already mentioned, other DL models could be tested using the obtained results as reference. As for the interpretability of the features, it may be helpful to take into consideration other type of features that cannot be extracted from EEGNet, such as connectivity-based features that have already been correlated in part to personality traits. A final chapter containing the main conclusions of your research/study and possible future developments of your work have to be inserted in this chapter.

Bibliography

[1] 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.

[2] 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.

[3] S. Cloninger, "Conceptual issues in personality theory," in *The Cambridge Handbook of Personality Psychology*, pp. 3–26, Cambridge University Press, jun 2012.

[4] S. Cloninger, "Freud: Classic Psychoanalysis," in *Theories of Personalities: Understanding Persons*, ch. 2, pp. 28–63, 2004.

[5] S. Cloninger, "Jung: Analytical Psychology," in *Theories of Personalities: Understanding Persons*, ch. 3, pp. 66–93, 2004.

[6] J. Feist, G. J. Feist, and T.-A. Roberts, *Theories of personality*. 2017.

[7] J. M. Digman, "Personality Structure: Emergence of the Five-Factor Model," *Annual Review of Psychology*, vol. 41, pp. 417–440, jan 1990.

[8] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications," *Journal of Personality*, vol. 60, pp. 175–215, jun 1992.

[9] L. R. Goldberg, "The structure of phenotypic personality traits," *American Psychologist*, vol. 48, no. 12, pp. 1303–1304, 1993.

[10] M. Perugini and L. D. Blas, "Big Five Marker Scales (BFMS) and the Italian AB5C taxonomy: Analyses from an etic–emic perspective," in *Big Five Assessment*, no. January, ch. 12, pp. 281–304, Hogrefe and Huber Publishers, 2002.

[11] C. G. DeYoung, "Personality Neuroscience and the Biology of Traits," *Social and Personality Psychology Compass*, vol. 4, pp. 1165–1180, dec 2010.

[12] C. G. DeYoung, J. B. Hirsh, M. S. Shane, X. Papademetris, N. Rajeevan, and J. R. Gray, "Testing predictions from personality neuroscience. Brain structure and the big five.," *Psychological science : a journal of the American Psychological Society / APS*, vol. 21, no. 6, pp. 820–828, 2010.

[13] T. A. Allen and C. G. Deyoung, "Personality Neuroscience and the Five Factor Model," *Oxford Handbook of the Five Factor Model*, vol. 1, pp. 1–62, may 2015.

[14] M. Carlén, "What constitutes the prefrontal cortex?," oct 2017.

[15] J. Barrash, D. T. Stuss, N. Aksan, S. W. Anderson, R. D. Jones, K. Manzel, and D. Tranel, ""Frontal lobe syndrome"? Subtypes of acquired personality disturbances in patients with focal brain damage," *Cortex*, vol. 106, pp. 65–80, sep 2018.

[16] D. T. Stuss, "Traumatic brain injury: Relation to executive dysfunction and the frontal lobes," *Current Opinion in Neurology*, vol. 24, no. 6, pp. 584–589, 2011.

[17] K. Jimura, S. Konishi, T. Asari, and Y. Miyashita, "Temporal pole activity during understanding other persons' mental states correlates with neuroticism trait," *Brain Research*, vol. 1328, pp. 104–112, apr 2010.

[18] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, 2019.

[19] G. R. Müller-Putz, "Electroencephalography," *Handbook of Clinical Neurology*, vol. 168, no. 2007, pp. 249–262, 2020.

[20] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287–314, apr 1994.

[21] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[23] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, pp. 1565–1567, dec 2006.

[24] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," 2018.

[25] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6, IEEE, aug 2017.

[26] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Computation*, vol. 31, pp. 1235–1270, jul 2019.

[27] W. H. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study," *Human Brain Mapping*, vol. 40, pp. 944–954, feb 2019.

[28] G. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[29] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *Journal of Neural Engineering*, vol. 16, no. 3, 2019.

[30] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, pp. 758–769, apr 2018.

[31] N. S. Kwak, K. R. Müller, and S. W. Lee, "A convolutional neural network for steady state visual evoked potential classification under ambulatory environment," *PLoS ONE*, vol. 12, feb 2017.

[32] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad, A. Khosravi, A. F. Atiya, D. Aminshahidi, S. Hussain, M. Rouhani, S. Nahavandi, and U. R. Acharya, "Epileptic Seizures Detection Using Deep Learning Techniques: A Review," *International Journal of Environmental Research and Public Health*, vol. 18, p. 5780, may 2021.

[33] R. Manor and A. B. Geva, "Convolutional Neural Network for Multi-Category Rapid Serial Visual Presentation BCI," *Frontiers in Computational Neuroscience*, vol. 9, dec 2015.

[34] Y. Zhang, J. Chen, J. H. Tan, Y. Chen, Y. Chen, D. Li, L. Yang, J. Su, X. Huang, and W. Che, "An Investigation of Deep Learning Models for EEG-Based Emotion Recognition," *Frontiers in Neuroscience*, vol. 14, dec 2020.

[35] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying Deep Learning for Epilepsy Seizure Detection and Brain Mapping Visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, pp. 1–17, jan 2019.

[36] C. L. Alves, A. M. Pineda, K. Roster, C. Thielemann, and F. A. Rodrigues, "EEG functional connectivity and deep learning for automatic diagnosis of brain disorders: Alzheimer's disease and schizophrenia," 2021.

[37] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[38] K. Korjus, A. Uusberg, H. Uusberg, N. Kuldkepp, K. Kreegipuu, J. Allik, R. Vicente, and J. Aru, "Personality cannot be predicted from the power of resting state EEG," *Frontiers in Human Neuroscience*, vol. 9, p. 63, feb 2015.

[39] H. K. Jach, D. Feuerriegel, and L. D. Smillie, "Decoding personality trait measures from resting EEG: An exploratory report," *Cortex*, vol. 130, pp. 158–171, 2020.

[40] A. Kabbara, V. Paban, A. Weill, J. Modolo, and M. Hassan, "Brain Network Dynamics Correlate with Personality Traits," *Brain Connectivity*, vol. 10, no. 3, pp. 108–120, 2020.

[41] P. Jawinski, S. Markett, C. Sander, J. Huang, C. Ulke, U. Hegerl, and T. Hensch, "The Big Five Personality Traits and Brain Arousal in the Resting State," *Brain Sciences*, vol. 11, p. 1272, sep 2021.

[42] J. Wache, R. Subramanian, M. K. Abadi, R. L. Vieriu, N. Sebe, and S. Winkler, "Implicit user-centric personality recognition based on physiological responses to emotional videos," *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pp. 239–246, 2015.

[43] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, no. i, 2015.

[44] G. Zhao, Y. Ge, B. Shen, X. Wei, and H. Wang, "Emotion Analysis for Personality Inference from EEG Signals," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 362–371, 2018.

[45] W. Li, X. Hu, X. Long, L. Tang, J. Chen, F. Wang, and D. Zhang, "EEG responses to emotional videos can quantitatively predict big-five personality traits," *Neurocomputing*, vol. 415, pp. 368–381, 2020.

[46] 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.

[47] A. R. Butt, A. Arsalan, and M. Majid, "Multimodal Personality Trait Recognition Using Wearable Sensors in Response to Public Speaking," *IEEE Sensors Journal*, vol. 20, no. 12, pp. 6532–6541, 2020.

[48] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, pp. 49–59, mar 1994.

[49] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales.," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988.

[50] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[51] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[52] K. A. Ludwig, R. M. Miriani, N. B. Langhals, M. D. Joseph, D. J. Anderson, and D. R. Kipke, "Using a Common Average Reference to Improve Cortical Neuron Recordings From Microelectrode Arrays," *Journal of Neurophysiology*, vol. 101,

pp. 1679–1689, mar 2009.

[53] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 0, no. 7 DEC, p. 267, 2013.

[54] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, and L. Invernizzi, "Keras-Tuner," 2019.

[55] L. Li, K. Jamieson, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," *Journal of Machine Learning Research*, vol. 18, pp. 1–52, 2018.

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[57] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2014.

[58] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, nov 2018.

[59] 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.

[60] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Express*, vol. 6, no. 4, pp. 312–315, 2020.

[61] A. Farahat, C. Reichert, C. M. Sweeney-Reed, and H. Hinrichs, "Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization," *Journal of Neural Engineering*, vol. 16, no. 6, 2019.

[62] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Systems with Applications*, vol. 114, pp. 532–542, dec 2018.

[63] M. A. Almogbel, A. H. Dang, and W. Kameyama, "EEG-signals based cognitive workload detection of vehicle driver using deep learning," *International Conference on Advanced Communication Technology, ICACT*, vol. 2018-Febru, pp. 256–259, mar 2018.

[64] F. Fahimi, Z. Zhang, W. B. Goh, T. S. Lee, K. K. Ang, and C. Guan, "Inter-subject

transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *Journal of Neural Engineering*, vol. 16, p. 026007, jan 2019.

[65] N. K. Nik Aznan, S. Bonner, J. Connolly, N. Al Moubayed, and T. Breckon, "On the Classification of SSVEP-Based Dry-EEG Signals via Convolutional Neural Networks," in *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pp. 3726–3731, Institute of Electrical and Electronics Engineers Inc., jan 2019.

[66] J. Wacker, M. L. Chavanon, and G. Stemmler, "Resting EEG signatures of agentic extraversion: New results and meta-analytic integration," *Journal of Research in Personality*, vol. 44, pp. 167–179, apr 2010.

[67] Y. Mu, Y. Fan, L. Mao, and S. Han, "Event-related theta and alpha oscillations mediate empathy for pain," *Brain Research*, vol. 1234, pp. 128–136, oct 2008.

[68] A. Moore, I. Gorodnitsky, and J. Pineda, "EEG mu component responses to viewing emotional faces," *Behavioural Brain Research*, vol. 226, pp. 309–316, jan 2012.

[69] K. Grill-Spector and R. Malach, "The human visual cortex," jul 2004.

[70] V. Zotev, H. Yuan, M. Misaki, R. Phillips, K. D. Young, M. T. Feldner, and J. Bodurka, "Correlation between amygdala BOLD activity and frontal EEG asymmetry during real-time fMRI neurofeedback training in patients with depression," *NeuroImage. Clinical*, vol. 11, pp. 224–238, 2016.

A | Appendix A

A.1. Deactivated temporal filters for Conscientiousness and Openness

Conscientiousness				
Active filter	**Accuracy**	**Precision**	**Sensitivity**	**F1**
All	0.863	0.892	0.854	0.873
1	0.466	0.952	0.031	0.060
2	0.450	0.000	0.000	0.000
3	0.550	0.550	1.000	0.710
4	0.550	0.550	1.000	0.710
Openness				
Active filter	**Accuracy**	**Precision**	**Sensitivity**	**F1**
All	0.865	0.867	0.873	0.870
1	0.552	0.537	0.987	0.696
2	0.604	0.660	0.489	0.562
3	0.519	0.519	1.000	0.683
4	0.481	0.000	0.000	0.000

Table A.1: Model performance on the traits Conscientiousness and Openness by keeping one filter active at a time.

List of Figures

List of Tables