Deep Learning Techniques for Electroencephalography Analysis

Georgios Zoumpourlis

Submitted in partial fulfillment of the requirements of the Degree of
Doctor of Philosophy

Supervisor: Prof. Ioannis Patras

School of Electronic Engineering and Computer Science

Queen Mary University of London

United Kingdom

December 2023

Statement of originality

I, Georgios Zoumpourlis, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Georgios Zoumpourlis

Date: December 31, 2023

Details of publications:

- **Georgios Zoumpourlis** and Ioannis Patras. "Motor Imagery Decoding Using Ensemble Curriculum Learning and Collaborative Training", 12th IEEE International Winter Conference on Brain-Computer Interface (BCI), 2024.
- **Georgios Zoumpourlis** and Ioannis Patras. "CovMix: Covariance Mixing Regularization for Motor Imagery Decoding", 10th IEEE International Winter Conference on Brain-Computer Interface (BCI), 2022.
- **Georgios Zoumpourlis** and Ioannis Patras. "Pairwise Ranking Network for Affect Recognition", 9th IEEE International Conference on Affective Computing and Intelligent Interaction (ACII), 2021.

Related and other publications:

- Mina Bishay, **Georgios Zoumpourlis** and Ioannis Patras. "TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition", British Machine Vision Conference (BMVC), 2019.

Abstract

In this thesis we design deep learning techniques for training deep neural networks on electroencephalography (EEG) data and in particular on two problems, namely EEG-based motor imagery decoding and EEG-based affect recognition, addressing challenges associated with them. Regarding the problem of motor imagery (MI) decoding, we first consider the various kinds of domain shifts in the EEG signals, caused by inter-individual differences (e.g. brain anatomy, personality and cognitive profile). These domain shifts render multi-subject training a challenging task and impede robust cross-subject generalization. We build a two-stage model ensemble architecture and propose two objectives to train it, combining the strengths of curriculum learning and collaborative training. Our subject-independent experiments on the large datasets of Physionet and OpenBMI, verify the effectiveness of our approach. Next, we explore the utilization of the spatial covariance of EEG signals through alignment techniques, with the goal of learning domain-invariant representations. We introduce a Riemannian framework that concurrently performs covariance-based signal alignment and data augmentation, while training a convolutional neural network (CNN) on EEG time-series. Experiments on the BCI IV-2a dataset show that our method performs superiorly over traditional alignment, by inducing regularization to the weights of the CNN. We also study the problem of EEG-based affect recognition, inspired by works suggesting that emotions can be expressed in relative terms, i.e. through ordinal comparisons between different affective state levels. We propose treating data samples in a pairwise manner to infer the ordinal relation between their corresponding affective state labels, as an auxiliary training objective. We incorporate our objective in a deep network architecture which we jointly train on the tasks of sample-wise classification and pairwise ordinal ranking. We evaluate our method on the affective datasets of DEAP and SEED and obtain

performance improvements over deep networks trained without the additional ranking objective.

Acknowledgments

First of all I must thank my supervisor, Prof. Ioannis Patras, for his help and support throughout my studies, but also for trusting me and giving me the opportunity to conduct this PhD research project. I would also like to thank the rest of the members of my supervisory team, Prof. Shaogang Gong and Dr. Qianni Zhang, for their insightful guidance.

I would like to note the importance of studying and working around people that inspired me to consider doing my PhD. I would like to especially thank the supervisor of my diploma thesis at Aristotle University of Thessaloniki, Prof. Anastasios Delopoulos. I also thank Dr. Petros Daras and Dr. Nicholas Vretos, who trusted me to join the Visual Computing Lab of CERTH, where I gained valuable experience under their supervision. I must acknowledge the inspiration that I received from Dr. Christos Tzelepis, Dr. Spyridon Thermos, Dr. Nikolaos Paterakis and Nikolas Adaloglou.

I would also like to thank all the past and present colleagues I met in MMV and QMUL over these years, for the countless interesting discussions and memorable moments: Dimitris, Mina, Tingting, Aria, Ye, Wenxuan, Niki, Chen, Giannis, James, Zheng, Jingjing, Zengqun, Petar, Juan, Debin, Yeming, Zhonglin, Ioanna, Alexandros, Zhaoyang, Yibao, Yilong, Ellie, Andrej, Krishna, Maria, Silvia, Samadhi, Issa, Woody, Orestis and Thomas. I thank all of them very much and apologize for any involuntary omission in this list. Special thanks go to Komal, Tristan, Vandana and Simran.

I would like to thank my parents for supporting me throughout this challenging period. Finally, I owe the most heartfelt thank you to Stella. It has been vital to have her by my side.

Contents

List of Figures

List of Tables

List of abbreviations

BCI	Brain-Computer Interface
CNN	Convolutional Neural Network
CSP	Common Spatial Pattern
cVEP	code-modulated Visually Evoked Potential
DA	Domain Adaptation
DG	Domain Generalization
DL	Deep Learning
DNN	Deep Neural Network
EA	Euclidean Alignment
EEG	Electroencephalography
ERM	Empirical Risk Minimization
ERP	Event-Related Potential
IAPS	International Affective Picture System
ITR	Information Transfer Rate
LDA	Linear Discriminant Analysis
MI	Motor Imagery
ML	Machine Learning
RA	Riemannian Alignment
RHH	Right-Hemisphere Hypothesis
SCP	Slow Cortical Potential
SCM	Spatial Covariance Matrix
SGD	Stochastic Gradient Descent
SMR	Sensorimotor Rhythms
SPD	Symmetric Positive Definite
SSVEP	Steady-State Visually Evoked Potential
SVM	Support Vector Machine

Chapter 1

Introduction

Contents

1.1 Electroencephalography (EEG)

The brain is the most complex organ within the human body, having an estimated average of 86 billion neurons [92]. As a part of the central nervous system, the brain is involved in a multitude of functions such as motor movements, vision, hearing and emotional responses. To accomplish these processes, the neurons of the brain send electrical and chemical signals to other neurons or parts of the body. Electroencephalography (EEG) is a neuroimaging technique that measures the electrical activity of the brain through electrodes affixed to the scalp. The voltage captured by these electrodes corresponds to collective activity of populations of neurons. Research around

the existence of electric current on animal brains dates back to the nineteenth century, in the works of pioneer electrophysiologists Richard Caton [37] and Adolf Beck [23]. The first EEG recording on a human brain was done by Hans Berger in 1924, while the results were published in 1929 [25].

Nowadays, technological and research developments have enabled the widespread collection, storage and analysis of EEG data. In some cases these data need to be further analyzed by human experts, which can be a lengthy process. For example, in the healthcare domain, professionals with neurophysiological expertise are required to interpret hour-long EEG recordings, in order to detect neonatal encephalopathy or to provide sleep disorder diagnoses. The creation of automated EEG analysis systems has been beneficial not only to medical professionals, but also to individuals and research communities of several fields. Other applications of automated EEG systems include stroke patient rehabilitation, visual spellers, interactive image synthesis and personalized recommendation systems. In many of these applications, it is critical to have fast and accurate neural decoding systems. For such purposes, various statistical tools and machine learning techniques have been used [84, 76, 123, 97] since the decade of 1980.

In this thesis we focus on the problems of motor imagery decoding and affect recognition from EEG data. In this chapter we first briefly describe the limitations of traditional machine learning techniques and the potential of deep learning for EEG analysis, in Section 1.2. Next, we introduce the problems of motor imagery decoding and affect recognition, in Sections 1.3 and 1.4 respectively. Afterwards, in Section 1.5 we list the challenges and assumptions of the studied problems, while in Section 1.6 we summarize the contributions of this thesis. Finally, in Section 1.7 we detail the outline of this thesis.

1.2 Machine learning for EEG analysis

Traditional machine learning methods have shown particular limitations in their ability to effectively handle natural data in their raw form. For this reason, and concerning the modality of EEG, a plethora of handcrafted feature types has been investigated for EEG-based tasks [94, 178, 68], spanning the time, frequency and time-frequency domains. However, the lack of a consistent narrative on the suitability of each feature type for a given task, is evident in the literature [102]. The vast majority of handcrafted features are electrode-wise, i.e. they are computed using the signal of a single EEG electrode. The electrode-wise nature of these features brings into question their potential towards capturing phenomena with manifestations in multiple areas of the brain. Moreover, handcrafted features of an EEG time-series signal are computed on temporal windows with pre-specified length. This restricts their usefulness towards detecting patterns with varying duration. Additionaly, algorithms trained on handcrafted features present poor cross-subject generalization, thus in most cases they are used to build subject-specific models.

The advent of deep learning (DL) [134], which is a subset of machine learning using neural networks to learn representations, has led to significant progress on several data modalities, such as images, audio and text. Since the breakthrough year of 2012 when the DL architecture of AlexNet [127] won the ImageNet [193] large-scale visual recognition challenge, DL models have achieved state-of-the-art results on many benchmarks and are widely deployed on production-level for a multitude of tasks [240]. The emergence of DL techniques has been relatively slower for EEG data [192], partially due to the lack of large publicly available datasets and open-source code libraries. However, there is an ever-increasing interest on the development of DL algorithms that will enable bringing EEG technology in out-of-the-lab settings.

A big advantage of DL approaches for EEG data, is their ability to learn representations in an end-to-end manner, with minimal signal preprocessing needed. In this

way, the requirement of standard machine learning techniques to operate on precomputed feature representations is bypassed. Deep architectures for EEG have excelled at supervised [232] and self-supervised [16] learning scenarios, while proof-of-concepts have been demonstrated in reinforcement learning [242]. Significant steps have been made in the exploration of training methodologies and model architectures that can accomodate learning from EEG datasets with an increasingly large number of participants and learning from multiple datasets. Furthermore, DL architectures capable of learning from a varying number of EEG electrodes [125] or learning from corrupted channels [17] have been introduced.

Deep learning has undoubtedly brought dramatic improvements into the field of EEG-based learning. Despite this fact, there are still plenty of open issues that are not addressed by existing works. Whole families of deep learning techniques remain relatively unexplored on EEG data. Furthermore, EEG-based DL approaches are often motivated by unconventional rationales, that are not grounded to knowledge from the domain of neuroscience. For these reasons, we deem that further research is necessary on the direction of building more sophisticated methods for automated EEG analysis using DL.

1.3 EEG-based motor imagery decoding

The main problem that we investigate in this thesis is EEG-based motor imagery (MI) decoding. Motor imagery is a well-known paradigm for brain-computer interfaces (BCI), involving the imagination of motor acts, without overt motor execution or muscle activation [149]. In the MI decoding task, EEG signals or signal-derived features are fed to machine learning models, to predict the type of imagined movements (e.g. left-hand, right-hand, feet or tongue movement). There are several factors that influence the behaviour of a subject during a motor imagery task, including personality type, cognitive profile, neurophysiological predictors, brain anatomy and familiarity

with BCI technology [179, 104, 105]. These factors result in widely varying patterns of EEG signals, even within an individual, rendering MI decoding a difficult task. Practical applications of MI decoding include controlling a cursor or a prosthetic (i.e. artificial) limb.

We divide the problem of MI decoding in two subproblems which we investigate. The first subproblem is the development of a method that addresses the issue of domain shifts (i.e. inter-subject differences in the characteristics of EEG signals), that is inherent in multi-subject EEG datasets. Our proposed solution enables: (i) training on a large number of subjects and (ii) robust generalization on new (i.e. unseen) subjects. The second subproblem that we deal with, is the exploration of ways to learn domain-invariant representations through the combination of data alignment and data augmentation techniques. We build a method that effectively achieves regularization on CNN networks that operate on EEG time-series, by simultaneously performing data alignment and data augmentation on the EEG signals during training.

1.4 EEG-based affect recognition

The second problem that we investigate is EEG-based affect recognition, where we apply an idea originating from psychology. The role of emotions in human experience and communication is crucial [182], as they affect actions, decision-making [143] and situation awareness in human-centric environments. The increasing availability and use of multimedia in everyday life, poses the challenge of further exploring the capabilities of systems that can analyze and measure human affective responses to such multimedia content. Practical applications of EEG-based emotion recognition include profile-based video content suggestion [36] and neuromarketing [107].

Within the affective computing field, the topic of emotion recognition from EEG data has been extensively studied in several works [5]. Using music video clips or movies as stimuli that are presented to human subjects, affect recognition works aim

to estimate the affective responses of the subjects to the content of the stimuli. The problem of affect estimation is highly subjective, as the groundtruth labels of affective datasets are usually emotional ratings in a continuous range of a numerical scale, self-reported by human participants. Moreover, affective ratings are ordinal by nature, since humans assign such values by comparing forthcoming to previously encountered emotional experiences. The ordinality of human emotions has been largely ignored in the literature, as raw continuous annotations are converted into discrete classes (i.e., splitting the scale range of ratings into classes such as "low"/"high" arousal/valence, by defining a threshold value). Thus the developed machine learning models are often trained solely on classification tasks that do not enable them to accurately capture the structure of emotions. We address the problem of affect estimation by designing a neural network architecture that does not disregard the ordinality of emotional ratings. We exploit knowledge from the original continuous annotations by learning to perform pairwise comparisons of samples with respect to their annotations.

1.5 Challenges and assumptions

In this thesis we address two problems, namely motor imagery decoding and affect recognition. There are both general (i.e. problem-independent) and problem-specific issues that render these tasks challenging. We briefly present the most important issues that obstruct the development of robust EEG-based DL techniques:

- **Low Signal-to-Noise Ratio (SNR):** EEG signals measure electrical brain activity, as captured by electrodes that are placed on the scalp. Between the sources of this electrical activity (i.e. the cortical neurons) and the sensors that measure this activity (i.e. the electrodes), lie materials with different electrical conductivity, such as the scalp, the skull bone and the dura mater [51, 66]. This means that electrical activity travelling through these materials will be distorted before being picked up by the EEG sensors. This distortion inserts noise in the

recorded EEG signal, making it difficult to recover the original brain activity. Among the most well-known methods that improve the SNR of EEG signals, is stimulus synchronized signal averaging. Signal averaging works under the assumption that the signal is random (thus, averaging would yield the stable EEG response), and the limitation that the signal is time-locked to an event, which is the case in event-related potential (ERP) studies. However, a requirement of implementing real-time BCIs is to be able to decode single trials, overcoming the necessity of collecting data from multiple trials to perform signal averaging. Moreover, in affective datasets where the stimuli are entire videos, each stimulus is presented only a single time to each participant, rendering trial averaging impossible.

- **Limited data availability:** The limited availability of EEG data is reflected on several levels. First and foremost, there is a limited number of EEG datasets that are publicly available, mostly because of privacy concerns. There is also a limited number of available datasets on specific problems (e.g. affect recognition) and the existing datasets are collected using different types of annotations (e.g. discrete or continuous affective ratings) and different types of stimuli (e.g. music video clips or movies). In many cases, EEG datasets contain a small number of human subjects (usually less than fifty participants per dataset) and a small number of trials per subject (usually a few tens of trials per participant). Combining multiple datasets to combat data scarcity, is often impeded by the use of different types of EEG recording equipment/devices (i.e. research-grade or consumer-grade headsets) across datasets. Finally, the collection of EEG data is an expensive and time-consuming process that requires domain expertise and abidance to EEG data standardization protocols [177].

- **Inter-subject variability:** This is one of the biggest challenges in EEG-based learning, referring to the existence of differences in the characteristics of EEG signals acquired from different individuals (i.e., there are different data distribu-

tions for each subject) [151]. In the literature, often the data of each subject (and also of each session for a specific subject) are considered as a separate data domain [126], hence the inter-subject differences are treated as domain shifts. There are several sources of variation that lead to domain shifts, such as the personality type, cognitive profile, neurophysiological predictors and brain anatomy of each subject [104, 105]. Overcoming the issue of inter-subject variability is of utmost importance for building ML models with strong subject-independent performance. A popular learning paradigm towards this direction, is the family of domain adaptation (DA) techniques, which requires available data from the test subjects (i.e. from the target domains) during the training phase. Evidently, this requirement is a limiting factor, and there is a growing interest in the development of calibration-free BCIs and ML models that do not require any knowledge about test subjects during training.

- **BCI illiteracy:** This term has been used to describe the phenomenon of users achieving low performance when attempting to operate a BCI system [3]. In the MI paradigm, this relates to the ability of users to perform voluntary modulation of their sensorimotor rhythms, which are widely considered as the most relevant brain pattern used by MI BCIs. BCI illiteracy is more common on the MI paradigm, compared to other BCI paradigms such as ERPs or Steady State Visually Evoked Potentials (SSVEPs) [185]. To improve the performance of "BCI-illiterate" subjects, some works have employed co-adaptive learning [224], i.e. adapting a classifier to the brain signals of a user. This adaptation is done while the BCI system presents visual feedback to the user, e.g. by showing whether a cursor is moving into a previously indicated target direction. An alternative but more challenging direction of research towards reducing the issue of BCI illiteracy, is the exploration of ways to train stronger classifiers without the necessity of online adaptation.

- **Annotation subjectiveness and uncertainty:** Affective datasets let parti-

cipants rate their emotional experiences while viewing multimedia content, using the well-known Likert scales and self-assessment manikins [32]. That is, regarding a particular stimulus, every participant reports the perceived emotion in terms of arousal, valence or other attributes. The issue of label subjectiveness is inherent in the process of affect annotation, both for users that perform self-reporting of perceived emotions and for external annotators that perform data labelling [247, 246]. The reaction of each person to a stimulus depends on several factors, such as the current mood, the personality type, and familiarity with the content of the stimulus. Moreover, each person may have a varying emotional reaction across numerous trials where the same stimulus is presented. Another limitation of affect annotation schemes, is that often they do not account for the uncertainty of human perception [87], which is also reflected on the assigned label values from human annotators. Typically, even when (inherently uncertain) continuous annotations are available for a dataset, researchers transform them through quantization/thresholding to derive "hard" (i.e. discrete) labels that are used as groundtruth in classification tasks. This transformation further amplifies the existing label uncertainty. Moreover, collapsing the continuous labels into discrete classes results in information loss regarding the pairwise relationships of samples with respect to their original (i.e. continuous) labels.

- **Long trial duration:** Affective datasets that utilise music video clips or movies as stimuli, have large durations for each trial [122, 158]. The provided trial-wise labels are assumed to correspond to the entire duration of each stimulus. Each stimulus has varying temporal evolution and audiovisual content across shorter temporal windows (e.g. some moments may be more interesting/memorable than others, or may elicit stronger emotions). Thus, the single groundtruth label of each trial has fluctuating relevance to the content of the corresponding stimulus across time. A trivial approach that is usually employed to alleviate this issue, is to discard an initial segment of each trial (e.g. the first quarter of its entire

duration). This is done under the assumption that the annotated emotion of a trial is not elicited immediately after the onset of the stimulus.

1.6 Contributions

A summary of the main contributions of this thesis is provided in this section.

In the first main chapter (Chapter 3), we propose a domain generalization method for motor imagery decoding. The main contributions of that chapter are the following:

- We propose a model ensembling approach to address the issue of domain shifts in the task of motor imagery decoding. Our method is based on employing diverse feature extractors to avoid the phenomenon of negative transfer learning due to domain shifts. Specifically, we build a model ensembling architecture that consists of two stages, namely the first stage that performs feature extraction using multiple base models, and the second stage that contains a single shared classifier that operates on top of all base models. A big advantage of our architecture is its simple design. Existing ensemble architectures usually have a varying number of filters, or varying filter lengths within each base model or across base models. They also use multiple branches or multi-view input representations (e.g. EEG signals that are bandpass-filtered on different frequency ranges) per base model. By contrast, our architecture has the same input space and the same design for all base models that act as feature extractors. The total number of trainable parameters for our architecture, as well as its computational cost, are kept low even when processing EEG signals from a large number of electrodes. Moreover, our architecture is trained in an end-to-end manner in a single phase and does not require any hyperparameter tuning or model selection process.

- We design a curriculum learning scheme, such that each model of our ensemble architecture is trained on *all* the source domains (i.e. all training subjects), but

progressively specializes to a specific subset of subjects, as training proceeds. As a result, each feature extractor of our ensemble, captures patterns that are mostly specific to EEG signal characteristics of a subset of training subjects. In essence, our curriculum learning scheme equips our ensemble with local (i.e. focused on a subset of the entire training set) feature extraction power and promotes diversity across the models of the ensemble. The combination of multiple feature representations leads to strong generalization capabilities, as our architecture covers a wide range of patterns through several models that act as diverse feature extractors. To our knowledge, curriculum learning has not been previously explored for cross-subject MI decoding.

- Another contribution of our method, is that we introduce an intra-ensemble distillation loss, in order to regulate the trade-off between feature diversity and generalization performance [28]. Our distillation loss pushes the class score predictions of each individual model, close to the average of the predictions of all the other models, thereby controlling the diversity within the ensemble. This is done by using pseudolabels (obtained from the predictions of all the other models) as groundtruth targets for each model. In essence, our distillation loss materializes a collaborative training scheme, that leads to exchange of knowledge *across* the models, working complementary with the curriculum learning scheme that is designed for *each* model. The balance between diversity and generalization is controlled through a hyperparameter that weights the contribution of the distillation loss to the total loss. Our work is the first to propose a pseudolabelling scheme for EEG-based knowledge distillation.

In the following chapter (Chapter 4), we propose a regularization method for motor imagery decoding. The main contributions of that chapter are the following:

- We propose covariance mixing (CovMix), a method that acts as a regularizer

on CNNs that are trained on EEG time-series decoding. A rising trend when using CNNs for EEG signal processing, is to transform the signals by performing covariance-based alignment [89, 126], before feeding them as inputs to the CNN. Alignment has been shown to improve cross-subject generalization, by helping to learn domain-invariant representations. Typically, methods that employ alignment do so by using a symmetric positive definite (SPD) matrix that corresponds to session-wise statistics. This SPD matrix is estimated either by directly computing the session-wise covariance matrix of the EEG signals, or by averaging the trial-wise covariance matrices across the entire session. In contrast to such approaches, we propose to compute the SPD matrix that is used to perform alignment, by interpolating between session-wise and trial-wise covariance statistics with a random proportion. Considering the nature of covariance matrices, we perform interpolation on the Riemannian manifold, by traversing along the geodesic that connects the session-wise and trial-wise SPD matrices. Practically, our method simultaneously performs alignment and data augmentation, as in each training step the input data are aligned using a different transformation matrix (i.e. obtained by performing interpolation with a different proportion of session and trial statistics).

- Our proposed method is performed before feeding the data to the classification network. Thus, CovMix is agnostic to the model architecture that performs MI decoding, and can be used in any method that employs CNNs for EEG-based classification.

- We perform cross-subject evaluation on BCI Competition IV-2a dataset [215], showing that adding CovMix acts as regularization to the classification network and yields stronger generalization results compared to the standard covariance-based alignment and the domain generalization techniques of MixUp and MixStyle.

Finally, in the last main chapter of this thesis (Chapter 5), we work on the task of affect recognition and apply the idea of performing pairwise ordinal ranking of data samples with respect to their affective labels. The main contributions of that chapter are the following:

- We investigate the utilization of pairwise ranking as an auxiliary objective to improve the performance of deep neural networks on EEG-based affect classification. The kind of available affect annotations determines the expected output of an affect recognition model, hence also the type of machine learning approach that can be applied, namely regression, classification or preference learning. Typically, plain classification approaches are applied on EEG-based affective datasets, by quantizing the original ratings to obtain discrete classes that correspond to affective state levels. However, transforming ratings of ordinal nature into nominal classes results in information loss regarding the structure of ratings. A more suitable approach is preference learning, that involves comparing (i.e. ranking) emotional ratings. Despite the exciting results of deep learning methods on affective computing problems, the possibility of building deep networks that can compare samples corresponding to different affective states, has remained mostly unexplored. Refraining from using solely a sample-wise classification objective, we propose employing an additional pairwise objective, namely the emotional rating comparison. Considering a pair of data samples and their affective labels, the comparison task infers the ordinal ranking relation between the labels of the samples (i.e. higher/similar/lower arousal, higher/similar/lower valence). We build a deep architecture that is jointly trained on two tasks: i) sample-wise affect classification and ii) pairwise ordinal ranking of samples with respect to their labels. Our goal is to boost the classification performance of affect recognition models, leveraging the additional supervision of the ranking task only during training. Our experiments show that the former task benefits from the latter, as

treating the data in a pairwise manner enables better representation learning.

- We further consider affective datasets such as SEED, where the original labels are discrete, instead of continuous. In this case, affect recognition models are directly trained on the available discrete labels. The three classes of SEED (i.e. "Positive", "Neutral" and "Negative") practically correspond to three ordered levels of valence. However, existing works ignore the evident ordinality in the classes of SEED and treat them as being nominal. In an alternative approach, we adapt our proposed network architecture to perform pairwise ranking of samples with respect to their discrete labels, by inferring the ordinal relations between them. Our experiments show that our method can be beneficial even in cases where the original affective annotations are discrete, instead of continuous.

1.7 Outline of the thesis

The rest of the thesis is structured as follows. We start by discussing related works in Chapter 2. We follow by introducing our proposed ensemble learning methodology in Chapter 3, where we also present the experimental evaluation on two large MI datasets (Physionet and OpenBMI). In Chapter 4 we address the issue of regularizing CNNs that operate on EEG time-series, through our developed technique, CovMix, that uses Riemannian geometry to jointly perform EEG signal alignment and data augmentation during training. Moreover, we evaluate the proposed technique in the MI dataset of BCI IV-2a. In Chapter 5, we work on the problem of affect recognition and provide a preliminary study on employing pairwise ranking as an auxiliary task towards boosting the classification performance of deep neural networks. Evaluation is performed on two datasets, namely DEAP and SEED, where the original affective labels are continuous and discrete, respectively. Finally, in Chapter 6 we draw our conclusions, highlight the wider implications and discuss the strengths and limitations of the presented works. Moreover, we identify potential applications and suggest directions for future research.

Chapter 2

Related work

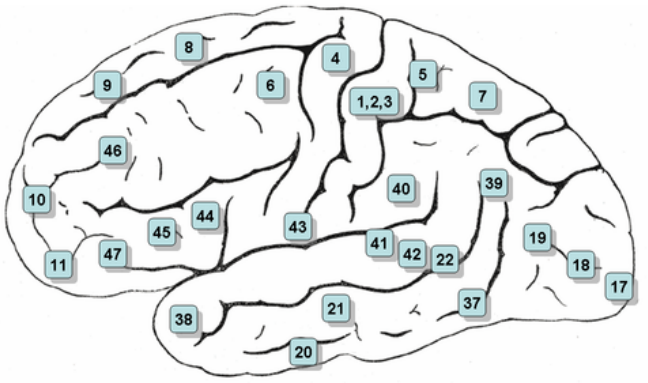
Contents

In Chapter 1 we introduced the field of electroencephalography. We begin Chapter 2 by describing various techniques for EEG signal analysis in Section 2.1. We continue by discussing the types and design principles of BCIs in Section 2.2. The state of the art methods for motor imagery decoding are discussed in Section 2.3. In order to study the problem of affect recognition, we provide an overview of affect modelling and annotation schemes in Section 2.4. Then we discuss state of the art research on affect recognition in Section 2.5. In Section 2.6, we describe the datasets as well as the evaluation metrics that are used in this thesis. Finally, we conclude the chapter in Section 2.7.

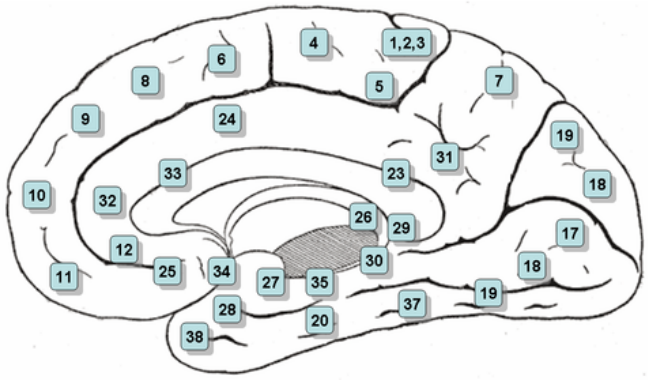
2.1 EEG signal analysis

There are several types of tasks that can be performed by analyzing EEG signals. Features derived from EEG signals can be used as a proxy towards interpreting underlying processes of the brain. Research works that attempt to gain insights into collected EEG data, are usually based on established connections between cortical functions and brain areas. Among the most widely used mappings from cortical regions to functions, is the categorization of German neuroanatomist Korbinian Brodmann, that divides the cerebral cortex in 52 brain regions [34], called "Brodmann areas". These regions are essential for specific brain functions (e.g. sensation processing, motor planning, working memory etc.), however each individual Brodmann area is not associated with a single functional role. A visualisation of Brodmann areas is shown in Fig. 2.1. These cortical areas are the sources of brain activity, however the neuroimaging technique of EEG does not directly measure brain activity in its source space. In fact, EEG indirectly measures brain activity by capturing the electrical activity on the sensor space, i.e. on the scalp surface where the electrodes are placed. To support easier comparisons and enable reproducibility across research studies that utilize EEG data, a standardization method for the placement of EEG electrodes is necessary. The first such system was presented in 1958 by Herbert Jasper [99] and was given the name "International 10–20 system". The standard electrode locations according to the 10-20 system are shown in Fig. 2.2. With the advent of more dense EEG montages that use hundreds of electrodes, newer systems have been proposed, such as the 10-10 [42] and the 10-5 [169] system.

In order to process EEG signals using any algorithm, the continuous EEG signal needs to be discretely sampled with a specific frequency using an analog to digital amplifier. According to the Nyquist–Shannon sampling theorem [204], the sampling frequency should be at least two times larger than the maximum frequency that one wants to retain from the original signal, in order to avoid the loss of information.



(a) Lateral (side) view



(b) Medial view (section between the right and left hemispheres)

Figure 2.1: Illustration of Brodmann areas. Figure taken from Wikipedia.

Typically, the upper limit of EEG frequencies that are studied lies in the range between 40-100 Hz. The reason for this, is that high frequencies are synchronously generated by smaller regions of the brain, resulting in smaller signal amplitude that cannot be captured on the scalp surface, while also because the skull filters out high-frequency signals [165]. According to the most widely known categorization of neural oscillations based on their frequency content, there are five basic frequency bands: delta (0.1-4Hz), theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz) and gamma (30-150Hz). An indicative

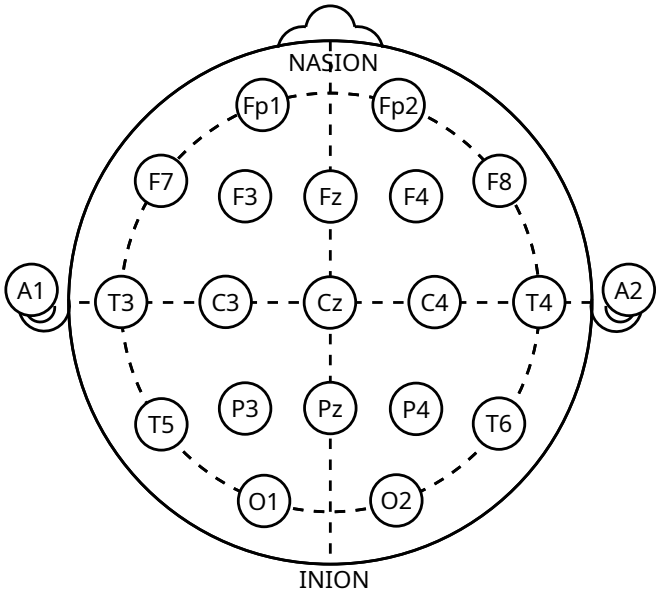


Figure 2.2: Illustration of the International 10-20 EEG electrode placement system. Figure taken from Wikipedia.

illustration of an EEG signal for each of the five bands is shown in Fig. 2.3.

Due to their nature (i.e. being collected from the scalp surface), EEG signals also carry information that does not originate from the brain. This pertains to information related to motion artifacts (e.g. saccadic eye movements, eye blinking, jaw clenching, muscle/head/body movements), heart rate and power-line interference. In order to obtain clean EEG signals that do not carry such noise, a few preprocessing steps are commonly used:

- **Notch filtering:** A bandstop filtering operation that attenuates the power-line interference appearing on 50Hz (or 60Hz for power lines on the United States of America and South Korea).
- **Bandpass filtering:** A filtering operation that keeps the frequency content of a specific range. An equivalent operation is to apply a highpass and a lowpass filter.

- **Bad channel/data rejection:** There are cases where an EEG electrode might have a faulty connection, thus the signal of this electrode cannot be further analyzed. To avoid using such data, one can reject individual EEG electrodes. Optionally, the signal of a faulty electrode can be replaced by performing interpolation from the neighbouring electrodes. There are also cases where short-time events (e.g. movements) render all EEG signals contaminated with artifacts. In such cases, one can opt to reject all data within a contaminated short-time window.

- **Artifact removal:** Individual sources of noise (e.g. ocular artifacts) can be removed through techniques such as Independent Component Analysis (ICA) [225, 235]. First, the EEG signals are decomposed into a number of source components. Then, the unwanted components are rejected (either automatically or upon visual inspection from an expert [43]), and the EEG signals are reconstructed using only the remaining ICA components.

2.2 Brain-computer interfaces

Brain-Computer Interfaces (BCIs) [245] are communication systems that translate brain activity into commands, enabling the interaction of human users with robotic limbs, computers or wheelchairs. The term "Brain-Computer Interface" (BCI) was introduced by computer scientist Jacques Vidal in 1973 [222], while in 1977 Vidal showed the first application [223] of a BCI where the user could move a cursor in a two-dimensional maze using Event-Related Potentials (ERPs) [221]. According to a definition given by Wolpaw *et al.* [236], "*a brain–computer interface is a communication system that does not depend on the brain's normal output pathways of peripheral nerves and muscles*". There is a wide spectrum of BCI applications, including post-stroke rehabilitation of limb motor impairments [15], character typing through visual spellers [244] and interactive image generation [209, 61]. A general framework for

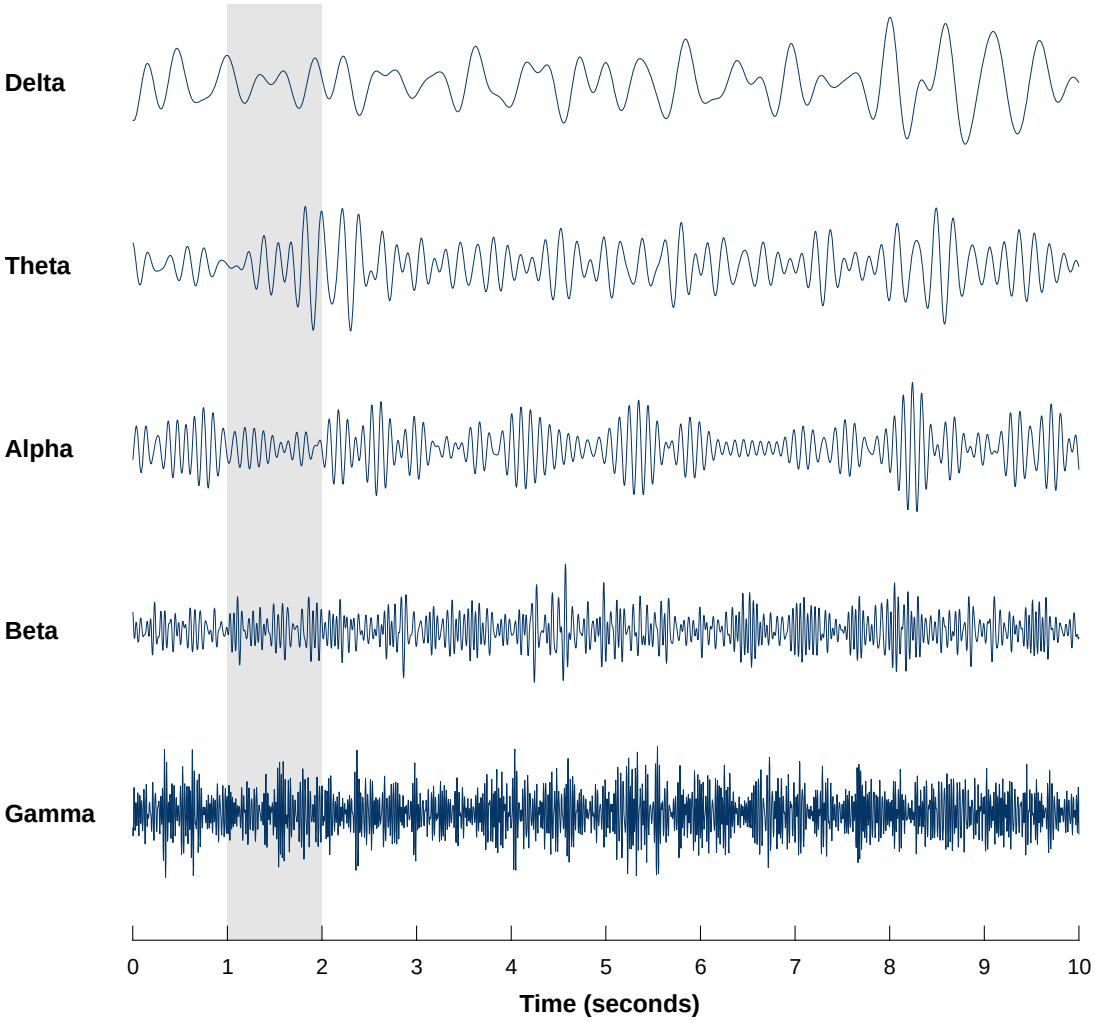


Figure 2.3: Illustration of the five frequency bands. Figure taken from Wikipedia.

designing BCI systems was proposed by Mason and Birch [154], containing six basic steps (illustrated in Fig. 2.4):

1. **Signal acquisition:** A neuroimaging technique is used in order to collect signals that measure brain activity. The most prevalent neuroimaging modality used in BCIs is EEG [1].

2. **Signal preprocessing:** Various algorithms (e.g. filtering and artifact rejection) can be used to reduce the amount of noise that is present in the obtained signals.

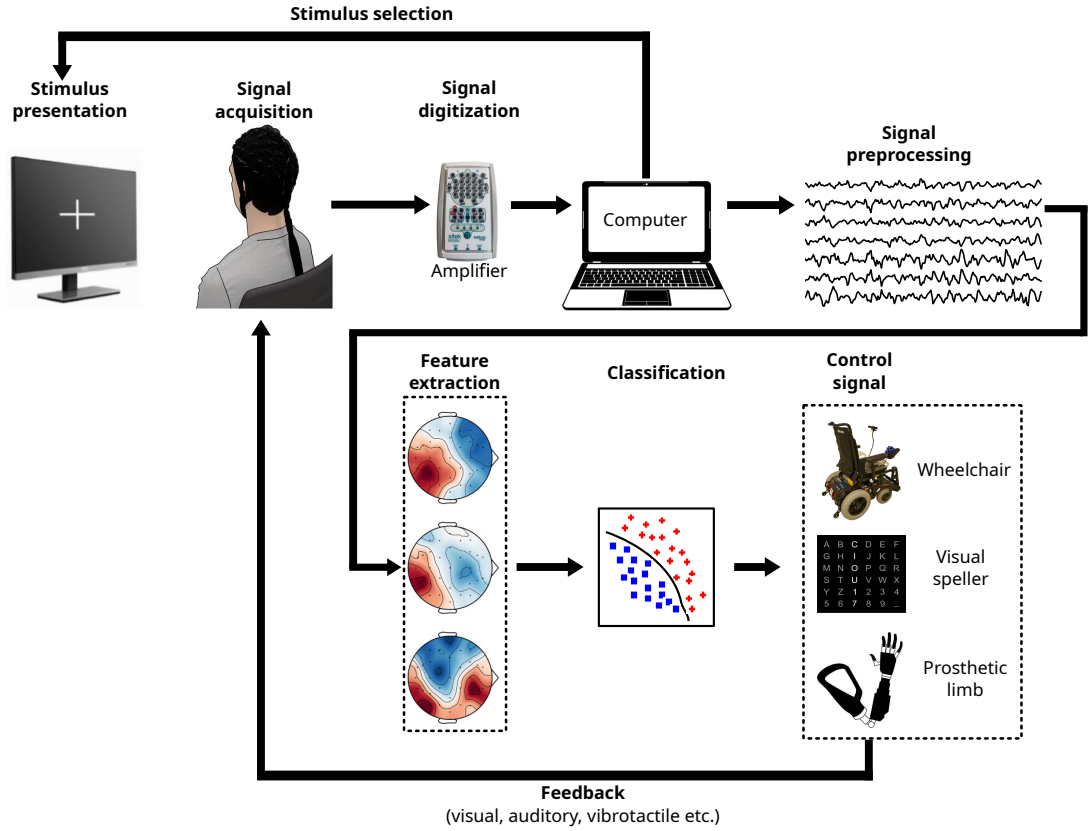


Figure 2.4: Architecture of a brain-computer interface.

3. **Feature extraction:** Several kinds of features can be extracted from the preprocessed signals. In the recent years there is a growing interest on the usage of deep neural networks for representation learning from EEG signals, while previously more focus was given on handcrafted features [102].

4. **Classification:** A classification algorithm is applied on the learned features, enabling the BCI system to infer a class as its prediction. This step is also known as "feature translation".

5. **Control signal:** A command is executed, corresponding to the predicted class of the previous step. This command controls an application (e.g. moves a wheelchair or a prosthetic limb, types a character on a visual speller, etc.).

6. **Feedback:** A feedback is provided to the users, informing them about how well

they can control the BCI [191]. Typically, this feedback can be of various forms (e.g. visual, auditory or haptic). The ultimate goal of providing feedback to the users is to train them, so as to increase their performance.

There are several types of BCIs that, based on their characteristics, can be divided in various sub-categories [40, 49], such as: (i) dependent and independent, (ii) endogenous and exogenous, (iii) invasive and non-invasive, (iv) synchronous (system-paced) and asynchronous (self-paced) and (v) evoked and spontaneous. A major criterion of categorization for BCI systems is whether they depend on evoked or spontaneous EEG signals:

- **Evoked BCIs:** The term "Evoked Potentials" (EP) is used to describe the neural response that is generated in the brain when a human perceives an external stimulus. This response is time-locked to the stimulus. Among the most well-known examples of evoked BCI systems are ERP-based BCIs (e.g. P300 [76] and code-modulated Visually-Evoked Potentials (cVEP) [29]) and BCIs that are based on Steady-State Visually-Evoked Potentials (SSVEP) [162]. Some of the advantages of evoked BCIs are their high Information Transfer Rate (ITR) [237], low calibration time and their ability to perform satisfactorily with a low number of electrodes.

- **Spontaneous BCIs:** By contrast, spontaneous BCI systems do not employ external stimulation, thus allowing the user to control them by voluntarily generating specific rhythmic patterns of neural activity. The most well-studied example is the Motor Imagery (MI) [181] paradigm, which is described in more detail in Section 2.3, while other spontaneous BCIs are based on the concept of Slow Cortical Potentials (SCP) [62, 202]. An advantage of spontaneous BCIs over evoked BCIs is that their usage causes lower fatigue and mental workload, as the user is not required to stare at stimuli [241, 110].

2.3 Motor Imagery

Motor imagery (MI) is a well-known paradigm for BCIs, involving the imagination of motor acts, without overt motor execution or muscle activation [149]. In the MI decoding task, EEG signals or signal-derived features are fed to machine learning models, to predict the labels of imagined limb movements (e.g. left-hand versus right-hand movement). The relationship between motor imagery and motor preparation/execution has been widely studied [120, 217], showing that these processes recruit common brain areas [86] such as the primary motor cortex [101]. The usage of the MI paradigm for BCIs is based on the phenomenon of sensorimotor rhythms (SMR) [250], i.e. rhythmic oscillations over the sensorimotor cortex that arise during motor imagery. Thus, a key ingredient towards recognizing motor intentions from brain activity, is the ability to perform spatial localization of SMR modulation. A systematic mapping of the human somatosensory cortex was presented in the seminal work published by Penfield and Brodley [174] in 1937, while a few years later Penfield and Rasmussen [175] introduced a mapping that associates areas of the cerebral cortex with motor and sensory functions. This mapping, which is shown in Fig. 2.5, has been known as the "Penfield Homunculus". Many works that study MI tasks (e.g. imagination of hand/foot movement) have observed clear SMR modulation on EEG electrodes that are placed in scalp areas overlying the corresponding brain areas, as described in the Penfield Homunculus. For example, the C3/C4 EEG electrode locations are positioned on the scalp areas that are closely related with right/left hand motor movements.

Following the general trend of the EEG field, there is a growing number of MI datasets with an increasingly large number of participants. The issue of inter-subject variability impedes the applicability of MI-based BCIs in real-world settings, as lengthy user training and model calibration processes remain necessary. This variability is related to spatial, spectral and temporal differences in the manifestation of sensorimotor rhythms across individuals [188]. Thus, these differences lead to domain shifts that are

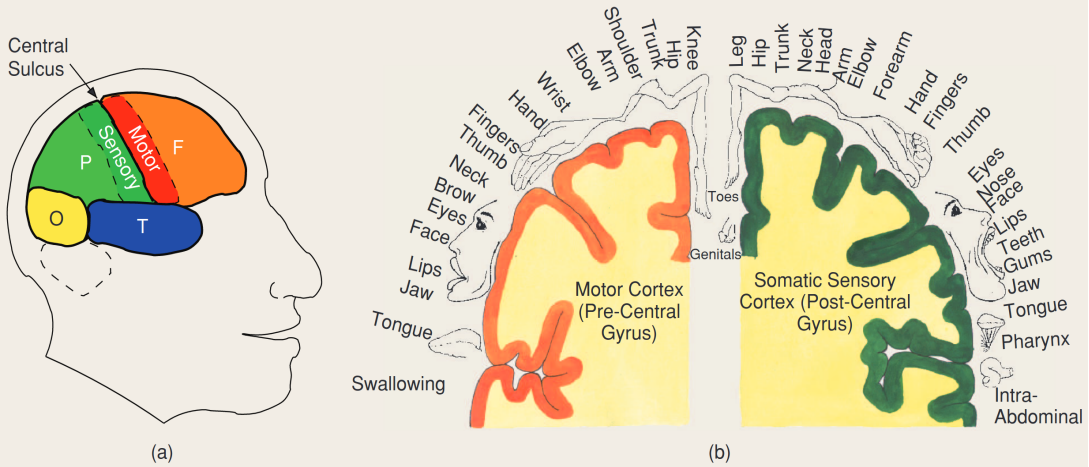


Figure 2.5: Illustration of the Penfield Homunculus: (a) The central sulcus is a groove in the cerebral cortex, separating the frontal lobe from the parietal lobe. More specifically, it separates the primary motor cortex from the primary somatosensory cortex. (b) According to the Penfield Homunculus, the motor cortex is responsible for processing motor functions, while the somatosensory cortex is responsible for processing sensory information (e.g. touch, temperature or pain) of different body parts. Figure taken from [30].

present in multi-subject MI datasets. Exploring research directions that would allow overcoming these domain shifts, is of paramount importance towards building MI-based BCIs with robust performance in subject-independent settings.

In the remaining part of this section we present an overview of previous work and state-of-the-art architectures on MI decoding.

Common Spatial Patterns (CSP): Early works on MI decoding include spatial filtering techniques, such as the method of Common Spatial Patterns (CSP) [30, 123]. In CSP, the variances of the filtered signals are maximized/minimized over certain conditions (i.e. classes). CSP methods involve computing the average spatial covariance matrix of the EEG signals for each class and jointly diagonalizing them. Handcrafted feature vectors are extracted from the spatially filtered signals, collapsing the temporal aspect by computing the variance in the dimension of time. Then, typical classifiers such as Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) [220]

are used.

One limitation of the CSP method is that while the spatial filters are learned, the spectral content is filtered either using a range that is selected manually, or using a generic broad frequency range. An alternative solution is proposed in the work of Dornhege *et al.* [65] that describes a simultaneous optimization of a spatial and a spectral filter. As the process of finding the single "best" frequency band for each subject can be time-consuming, more recent methods leverage filter banks, i.e. multiple frequency bands in the same time. The method of Sub-Band CSP (SBCSP) [163] performs CSP and trains individual classifiers in multiple sub-bands. In order to infer the final prediction, SBCSP performs fusion on the scores of each sub-band using an SVM. The method of Filter Bank CSP (FBCSP) [9] builds and extends the SBCSP technique, as it includes four stages, namely frequency filtering, spatial filtering, feature selection and classification. By investigating multiple algorithms for feature selection and classification, FBCSP shows superior results compared to SBCSP and CSP. In contrast to works that assume equal importance for all frequency sub-bands, Tang *et al.* [214] propose assigning higher weights to specific sub-bands.

Several works highlight the importance of exploiting multi-band information in the CSP framework, but less works explore the impact of the time window that is used. The work of Miao *et al.* [157] proposes a method called Common Time-Frequency-Spatial Patterns (CTFSP) that seeks to optimize the time window. CTFSP applies a sliding window approach to segment each sample in multiple time windows. Then, for all the data that correspond to each time window, CTFSP performs multi-band filtering, feature extraction and feature selection, before training a window-specific SVM classifier. The final prediction is inferred by fusing the predictions of the individual classifiers. The work of Zhang *et al.* [255] describes a technique for simultaneously optimizing the frequency bands and the appropriate time windows, using a single SVM classifier.

The CSP framework is also used for ensemble learning, by combining the predic-

tions of multiple EEG-based models. Initially this has been tried in intra-subject settings [164, 136, 172, 48, 254]. Further work has been done on using CSP for subject-independent models, by forming ensembles of CSP filters [198, 199].

Departing from subject-specific models and attempting to learn from multiple subjects, numerous CSP-based algorithms are among the early works that tackle cross-subject MI decoding. Lotte and Guan [148] propose incorporating information from multiple subjects in a CSP framework by regularizing the covariance matrix of a subject towards the average covariance matrix of other subjects. However, for each target subject the method requires selecting a subset of relevant source subjects. Another work that builds on CSP, proposes weighting source subjects based on their similarity to the target subject [46]. Alternatively from CSP works that use multiple training subjects to learn task-relevant information, Samek *et al.* [197] propose learning information related to the common non-stationarities that exist in the data. Specifically, the assumption made in [197] is that non-stationarities that are caused by differences in the stimulus presentation or feedback mode between sessions, are consistent across subjects. An example of such a case is an experimental process where the training (i.e. calibration) session is recorded without providing visual feedback to the user, while in the testing session visual cues inform the user about the predictions of the BCI system. This experimental design would yield increased activity in the occipital area, that should be considered in the computation of the spatial filters.

Azab *et al.* [13] propose a transfer learning technique, leveraging multiple classifiers that are trained on subject-specific CSP feature spaces from the source (i.e. training) subjects. In more detail, the authors introduce a measure of similarity between the feature space of each training subject and the target (i.e. test) subject. Afterwards, when computing the classification parameters of the test subject using a few labelled trials, this measure is taken into account while also using knowledge from the classification parameters of the source subjects.

However, the family of CSP methods presents poor cross-subject generalization and is restricted by the discarding of temporal information that occurs during feature extraction. Furthermore, the CSP framework supports binary (i.e. two-class) problems and extending CSP to multi-class scenarios requires solving multiple subproblems (i.e. either between pairs of classes or in an one-versus-all manner) [142, 155, 115].

Covariance-based techniques: Instead of using the spatial covariance matrix of EEG signals to compute spatial filters, another line of research inspired by Riemannian geometry, directly uses them as the input features of classification models [248, 52]. Covariance matrices lie on the Riemannian manifold of symmetric positive definite (SPD) matrices, hence should be treated accordingly. Riemannian geometry is a rich framework for implementing algorithms that manipulate covariance matrices, as explicit formulas exist for several operations (e.g. computation of the Riemannian distance between two matrices, or computation of the Riemannian mean for a set of matrices) in the Riemannian manifold. The work of Barachant *et al.* [19] is the first to propose a Riemannian classification framework for motor imagery BCIs that is based on covariance matrices. Specifically, two scenarios are proposed in [19]: (i) classification in the Riemannian manifold and (ii) classification in the Riemannian tangent space. The first scenario is implemented using the Minimum Distance to Riemannian Mean (MDRM) method. In MDRM, class-wise representations are obtained by estimating the Riemannian mean of all trial-wise covariance matrices for each class. Note that this Riemannian mean is a matrix that still belongs to the Riemannian manifold. Then, a given test sample (i.e. covariance matrix) is classified by computing its distance to all class-wise SPD matrices and assigning the index of the class that yields the minimum distance, as the predicted class label. The second scenario is implemented by mapping each sample into the tangent space that is located at the Riemannian mean of all the training samples. This tangent space is Euclidean and thus classifiers such as LDA and SVM can be employed.

The method proposed in [20] deals with the problem of cross-session variability in the feature distributions that are used in Riemannian frameworks. Evidently, when training a classifier on features obtained through tangent space mapping, the selection of the reference matrix that is used to perform the mapping is crucial. It is shown that by properly adapting the reference point, the performance of classifiers can be improved on data from sessions that are not included in the training set. Kalunga *et al.* [109] address the problem of data scarcity proposing a method that performs data augmentation on the Riemannian space, by interpolating on the log-Euclidean geodesic between trial-wise covariance matrices. Specific care is taken to reject outliers that might exist in the original covariance matrices, using the Riemannian potato [18] technique for automatic artifact detection. Despite their widespread use, Riemannian geometry frameworks are prone to outliers due to the non-stationarity of EEG signals [248] and do not allow temporal information extraction.

Deep learning: Shifting from traditional handcrafted techniques (i.e. CSP-based or Riemannian frameworks) to deep learning techniques and CNN-based architectures, has enabled the exploration of more sophisticated methods. Using CNNs as feature extractors that automatically learn spatial, temporal and spectral representations, has led to remarkable progress and state-of-the-art results in cross-subject EEG-based tasks, including motor imagery decoding. Early works that use deep learning techniques on EEG data focus on applications such as P300 detection [38] and epileptic seizure detection [10].

A major breakthrough for CNN-based BCI applications is the introduction of EEGNet architecture by Lawhern *et al.* [132, 133]. EEGNet is a compact architecture that operates on raw multi-channel EEG time-series and can be applied in various BCI paradigms for classification purposes. EEGNet first temporally convolves the signals in an electrode-wise manner, obtaining several frequency-specific signals that correspond to each electrode. Then, EEGNet performs a spatial convolution that combines

information across electrodes for each set of frequency-specific signals. Afterwards, a depthwise convolution is applied to filter the temporal dimension, followed by a pointwise convolution that combines information across the frequency dimension. Finally, a single fully-connected layer is employed as the classifier of the architecture, predicting the probability of each class. Variants of EEGNet that are inspired from the Inception architecture [213] are presented in Incep-EEGNet [189], EEG-Inception [251] and MI-EEGNet [190].

Schirrmeister *et al.* [201] propose two convolutional architectures, namely "Deep ConvNet" and "Shallow ConvNet", coupled with a training strategy that uses multiple crops from each trial both during training and inference, to improve model performance. These two architectures are compared against the handcrafted FBCSP [9] method showing their superiority. Inspired by DenseNet [98] architecture, Kostas and Rudzicz propose the architecture of TIDNet [126] that uses dilated convolutions and residual connections in the temporal and spatial feature extraction stages. The method of MIN2Net [12] employs multi-task learning in order to achieve better representation learning. Specifically, MIN2Net is trained on three tasks: (i) an autoencoding task where an encoder-decoder architecture attempts to reconstruct the original signals, (ii) a deep metric learning task on the intermediate feature space at the output of the encoder and (iii) a standard classification task using the output of the encoder. However, the data provided in [12] suggest that the reconstruction task is poorly addressed in the training process of MIN2Net, indicating the importance of carefully assessing the benefits of including each task in multi-task learning pipelines.

Other works go beyond designing neural architecture components and investigate meaningful transfer learning approaches and combinations of multiple sources of data. One such work is presented in [253] where the authors propose the technique of adaptive transfer learning (ATL). ATL attempts to improve the performance of a pretrained model, on test data from a subject that is not included in the training set. To achieve

this, labelled data from the test subject are used for further training the pretrained model, adapting its parameters to the test subject. The impact of two factors is studied: (i) the amount of labelled data from the test subject and (ii) the number of layers that are finetuned on the test subject. The experimental analysis validates that ATL provides improvements over purely subject-independent models. However, the assumption of using labelled data from a test subject, renders such scenarios as less realistic and difficult to implement.

By contrast, other more plausible transfer learning scenarios try to utilize multiple sources of data to train models with stronger robustness. Such models can be directly employed in purely subject-independent settings, without requiring finetuning on target subjects. The work of EEGSym by Perez *et al.* [176] proposes pretraining models on external datasets in order to overcome inter-subject variability. Combining multiple EEG datasets requires careful data preprocessing, to appropriately account for the differences in electrode montages and experimental conditions. Moreover, a novel architecture along with data augmentation techniques are included in the method of EEGSym. The results of [176] prove the power of transfer learning, as EEGSym achieves state-of-the-art performance in subject-independent evaluation settings, even with a small number of electrodes. In [124] the method of BENDR is described, where a transformer architecture is trained in a self-supervised manner on a large EEG dataset [166] comprising more than a thousand subjects. Then, the architecture is further finetuned in a supervised manner on various EEG-based tasks, showing the benefits of self-supervised pretraining.

Overall, the adoption of deep learning for EEG-based tasks has led to dramatic performance improvements on several tasks. An attempt to provide various taxonomies of deep learning techniques applied on MI decoding, with respect to different criteria, is shown in [7]. In the following paragraphs we discuss more extensively various families of deep learning techniques for several EEG-based tasks, yet with a particular focus on

MI decoding.

Domain generalization (DG): Training on multiple source domains with the aim of generalizing on unseen target domains, is at the heart of domain generalization techniques. A much larger volume of works has been published for DG on visual data, compared to the works on EEG data. Regarding learning from visual data, there is a line of works that claim various benefits, however upon a rigorous comparison, the authors of [83] show that the trivial approach of Empirical Risk Minimization (ERM) [219] can, if carefully tuned, outperform several state-of-the-art DG techniques. Suitable DG approaches can help building strong cross-subject algorithms for learning from EEG data. There has been very little cross-fertilization between DG algorithms for visual and EEG data, and up to now a systematic study on DG methods that work well on both of these two modalities has not been reported. One major reason is that many DG works for visual data [140] rely on models pretrained on large vision datasets (e.g. ResNet [90] pretrained on ImageNet [128]). Among other choices, the strength (i.e. the accuracy on the test set of the external dataset) of these models affects the performance on the final DG task [249]. The lack of off-the-shelf pretrained models for EEG data, means that EEG-based DG algorithms cannot equally build on the power of transfer learning. Moreover, it has been shown that when using particular DG techniques on visual data, switching from shallower to deeper pretrained vision models (e.g. ResNet-50 instead of ResNet-18 architecture) reduces the claimed gains over an ERM baseline, leading to marginal or no boost at all [83]. These findings make it difficult for researchers to draw parallels from visual to EEG data, and to be inspired from DG techniques that are tailored to visual data, adapting them on EEG data.

Considering the above factors, it is worth noting that several works building subject-independent models for EEG data (which by nature is a DG problem), do not explicitly take care of inter-subject variability by any means [12, 150, 258, 176]. Practically, this leads to ERM-based approaches that simply minimize the training loss over all

source domains (i.e. training subjects) simultaneously. Apart from the effective ERM baseline, other DG methods that have been occasionaly used for multi-subject EEG training, are MixUp [252, 126] (which was initially proposed for visual data), and Euclidean/Riemannian Alignment (EA/RA) [89, 126, 243] (proposed for EEG data). Alignment methods such as EA/RA have been proven significantly useful for learning domain-invariant representations, while in the same time providing the ability to combine them with other multi-domain learning techniques [126]. By contrast, MixUp has been found to have a rather detrimental effect on multi-subject training [126].

Ensemble learning: Model ensembling [88] is ubiquitous in the applications of machine learning on EEG data [208, 53]. A successful example of model ensembling is the work of Bakas *et al.* [14], where a k-fold cross validation process results in k trained models, with each model trained on data from all the available training subjects. The authors of [187] leverage the power of available crowdsourced algorithms for an EEG-based seizure prediction competition [129], exploring the possibility of obtaining performance improvements by combining them through model ensembling. A simple ensembling approach that builds on a neural network receiving the output probabilities of all individual algorithms as inputs, manages to achieve higher seizure prediction performance.

In [67], a deep neural network ensemble named InceptionEEG-Net (IENet) is proposed. Focus is given in the architecture design of IENet, by increasing the receptive field size of the architecture and keeping the computational cost small. In [64], an ensembling method is proposed, where an ensemble classifier is built by combining models trained on different cross-validation splits of the training data. A set of experimental runs for hyperparameter tuning is required in order to pick the best model for each cross-validation split. Thus the suggested pipeline cannot be trained in an end-to-end manner, and cannot be implemented in a single training phase. An example of ensemble learning for EEG-based cognitive state estimation is presented in [39], where

the method of FBCSP [9] is combined with a deep ensemble model. Each individual model of the ensemble is trained on data from a single subject upon a subject-specific feature selection process (i.e. the input data of each model contain different feature subsets).

Feature diversity: One of the key properties of ensemble learning, is the emergence of diverse feature representations across the individual base models of ensembles. However, feature diversity can also be obtained through alternative techniques which do not fall within the category of ensemble learning, as they explore ways to obtain diverse features through a single model. In [152], a multi-branch network architecture is proposed, where the input EEG signal is divided in four frequency bands, with a dedicated branch for each band. The authors of [8] introduce a multi-branch network based on EEGNet [133], where each branch contains a different number of temporal filters, as well as a different temporal filter length. A network capable of processing spectral-spatial representations as inputs is presented in [130].

In the work of Wei *et al.* [233], a multi-branch Separate-Common-Separate Network (SCSN) is proposed to tackle the issue of negative transfer learning. Negative learning can appear when training subject-agnostic feature extractors, i.e. when all the layers of a single model are trained on all the training subjects. As a remedy to this, SCSN has a separate feature extractor for each training subject. However such an approach leads to non-optimal solutions, as training subject-specific layers compromises their generalization capability.

2.4 Affect modelling and annotation

Research on affect recognition is theoretically grounded on particular models that define the structure and type of emotions. Thus, the process of modelling human emotions using affect modelling schemes, precedes that of recognizing emotions. In 1806, Charles Bell publishes his work titled "*Essays on the Anatomy of Expression in Painting*" [24],

stating that facial expressions reflect human emotions and studying the anatomy and physiology of facial expressions. Modelling facial expressions as combinations of facial muscular movements, is a primitive way of describing the elementary components that constitute certain emotions. Charles Darwin, inspired by the work of Bell, publishes the book "*The expression of the emotions in man and animals*" [58] in 1872, where he identifies a particular set of facial expressions that are universal, i.e. commonly perceived across cultures. The works of Bell and Darwin have been very influential to early attempts of defining emotion models [144, 216, 75].

Various schemes to model human emotions have been proposed, with one of the most important ones being the categorical scheme of Ekman, that suggests the existence of seven universal emotions [71, 70]. Ekman further studies the existence not only of universal [73] but also culture-specific emotions [74]. However, the line of works that are based on the research of Ekman, mostly focus on facial expressions as a manifestation of human emotions. This restricts not only the range of emotional experiences that can be described through such emotion models, but also the possible applications that such models can have in real life. The affective experiences that a human has in everyday life are composed of a plurality of emotions, that are not necessarily expressed through facial expressions. In fact, humans constantly process and perceive emotional stimuli and their emotional reactions can include various bodily behaviors, facial expressions, neural responses, or even concealment of emotions.

Another emotion modelling scheme is introduced by Russell and Barrett in [194, 195]. According to the dimensional model of Russell, affective states are related to each other and organized in a circular arrangement. In this circumplex model, emotions are adequately represented in two dimensions, namely arousal (i.e. intensity) and valence (i.e. pleasantness). An illustration of the dimensional model of emotions is shown in Fig 2.6 where the two axes correspond to arousal and valence. Additionaly, Fig 2.6 depicts various affective states that can be described as combinations of different arousal

and valence values. The convention of the dimensional model of emotions requires annotators to rate emotions by assigning values on a bounded continuous range of a scale. The particular lower and upper limits of this bounded range may vary across works (e.g. $[-1, 1]$ in [160], or $[1, 9]$ in [122]), yet the concept remains the same. Furthermore, Likert scales [141] and Self-Assessment Manikins (SAM) [32] can be used in the annotation process, to assist users while rating the affective dimensions. An indicative example of the picture-oriented technique of SAM is shown in Fig 2.7.

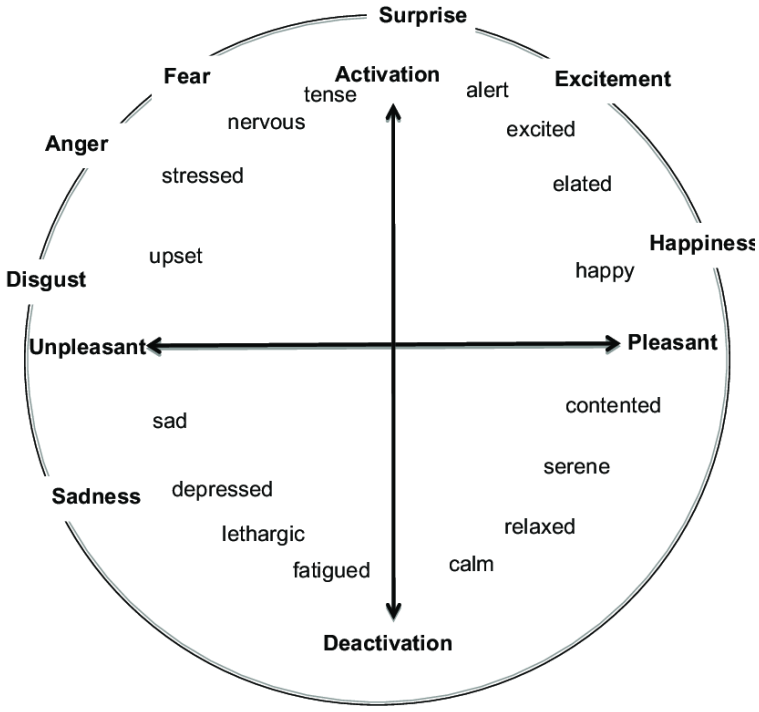


Figure 2.6: Illustration of the dimensional emotion modelling scheme proposed by Russell and Barrett [194, 195]. Figure taken from [82].

The model of Ekman is usually related to a classification task, where each emotional state is assigned to a class, while the dimensional model of Russell is related to a regression task, as it refers to the degree that an emotion is positive or intense. However, even when using the dimensional model for annotation (i.e. label collection) purposes, many evaluation protocols (thus also the developed algorithms) map the annotation values of each emotional scale into classes (e.g. mapping a continuous arousal scale from

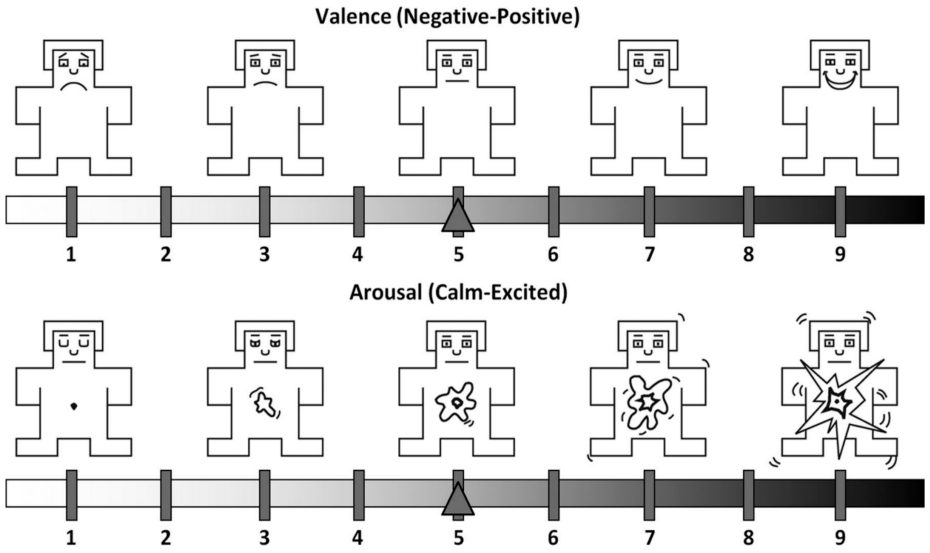


Figure 2.7: Illustration of the Self-Assessment Manikin (SAM) [32] technique for emotion annotation. SAM can be implemented either in an interactive computer program or in a paper-and-pencil version.

the range $[1.0, 9.0]$, to the classes "Low Arousal" and "High Arousal"). An evaluation bias is introduced through this convention, as the annotations are inherently subjective self-assessment ratings and the boundaries between what can be perceived as "Low Arousal" or "High Arousal" vary across individuals.

According to the taxonomy of Stevens [210], three types of scales can be used to measure emotions: nominal, ordinal and interval. The differences between these scale types are shown in Fig. 2.8. In essence, nominal emotion annotations can be used in a classification task, while ordinal ones can be used in ranking and interval ones can be used in a regression task. There is evidence that treating annotations of affective datasets as ordinal values through ranking approaches can yield more generalisable affect models, compared to treating them as nominal values through classification [153]. When dealing with an affect recognition task, be it either classification or regression, the selection of a specific modelling scheme for emotions, partly determines the limitations that will be inherent in any developed method. Hence, we claim that new proposed methods to analyze affective datasets, should be accompanied with a careful selection

of the appropriate emotion modelling scheme.

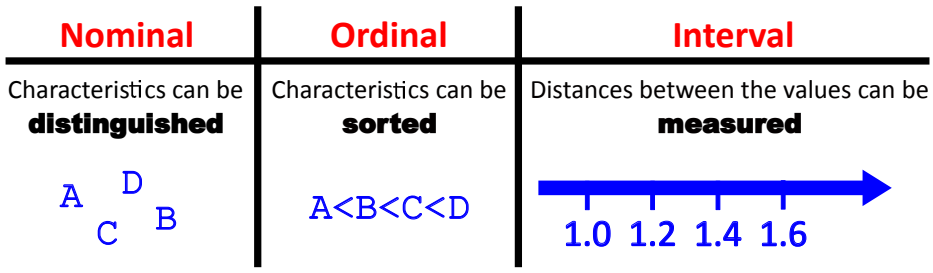


Figure 2.8: Illustration of the scale types that can be used for measuring human sensation, based on the taxonomy of Stevens [210].

2.5 Affect recognition

Upon the adoption of an emotion modelling scheme and the collection of affective annotations, one can proceed to the stage of training affect recognition models. Recognizing emotional reactions to stimuli can be achieved through numerous modalities, including visual (e.g. facial expressions, gaze, body pose, gestures), audio (e.g. speech prosody), text (e.g. dialog transcripts) and neurophysiological (e.g. brain signals, heart rate, skin conductivity, body temperature) data, to name a few. There is a vast amount of works that use visual, audio and text modalities to study the topic of affect recognition. Admittedly, less effort has been devoted to study emotions using neurophysiological signals, including modalities of brain signals [6, 59].

EEG is one of the neuroimaging techniques that can provide useful information towards affect recognition. Early works that study EEG-based affect recognition, focus on combinations of facial expressions and EEG signals during the presentation of short films as affective stimuli [60, 72]. However, the fact that the facial expressions of the participants are studied together with their EEG signals, raises questions about potential confounders (e.g. ocular or muscular artifacts) in the conducted analysis. Keil *et al.* [113] study the gamma band activity as well as the ERP components evoked by the presentation of affective pictures taken from the International Affective Picture System

(IAPS) [131] dataset. Their findings support the existence of lateralization in emotion processing between the two brain hemispheres, as expressed by the Right-Hemisphere Hypothesis (RHH) of Borod *et al.* [31, 118]. Aftanas *et al.* [2] use images from the IAPS dataset to study the inter-hemisphere distribution of evoked changes in the power of EEG signals. Their study shows that effects of both evoked synchronization and desynchronization [180] can be observed in the anterior and posterior areas of the cerebral cortex. Onton and Makeig [168] design an eyes-closed emotion imagination task, using the method of guided imagery [211] to induce emotional states through a set of prerecorded verbal suggestions to the subjects. Their work employs ICA to decompose brain activity into independent components, showing connections between emotional valence and broadband high-frequency activity in these components.

The release of publicly available datasets that provide EEG data and affective annotations, such as DEAP [122], MAHNOB-HCI [206] and DREAMER [111], has played a major role in the development of numerous works that tackle EEG-based affect recognition using machine learning techniques. In the remaining part of this section we present an overview of previous work and state-of-the-art architectures on affect recognition.

Handcrafted features: As earlier mentioned in Section 1.2, traditional machine learning methods cannot effectively handle EEG data in their raw form. A multitude of works investigate ways of effectively leveraging information from different types of handcrafted features. A systematic review on the strength of each feature type in emotion recognition is presented in the study of Jenke *et al.* [102]. An analysis of different feature types and dimensionality reduction methods is presented by Liu *et al.* [145] with the goal of efficiently fusing a variety of features. It is shown that combining supervised and unsupervised dimensionality reduction methods on the fused features, can improve the performance of both random forest and SVM classifiers. Considering datasets with EEG signals from multiple electrodes that reflect inner processes of

multiple cortical areas, it is reasonable to take this into account when developing affect recognition methods. Many works rely on simple electrode-wise EEG features and naive concatenations of them. Chen *et al.* [44] investigate shifting from electrode-wise EEG features to features that can explicitly take into consideration the interactions between different electrodes. Their idea is inspired by prior works indicating that cortical areas cooperate in certain patterns instead of working separately [167]. Through their proposed connectivity features between pairs of electrodes, interaction types such as activation/deactivation and synchronization/desynchronization are captured. Wu *et al.* [239] analyze frequency-specific brain functional connectivity patterns that are associated with emotions, identifying critical subnetworks.

Deep learning: The advent of deep learning has enabled a rapid progress in the field of representation learning for affect recognition. A comparison between traditional machine learning techniques and a CNN architecture for affect recognition is shown in [11]. A convolutional architecture for affect recognition, named TSception, is presented by Ding *et al.* [63], exploitting the spatial assymetry of brain activity during emotion elicitation. An important issue when dealing with EEG data in deep networks, is the existence of domain shifts. In subject-independent evaluation settings, the test data often have distributions that are distant from the distributions of the training data, which leads to weak test performance. An unsupervised domain adaptation (UDA) technique for visual data is originally proposed in [78]. The key idea is that the features used for classification should not be discriminative with respect to the shift between the source/target (i.e. train/test) domain. Unsupervised domain adaptation is implemented by adding a domain classifier that is trained together with a "gradient reversal layer" in order to confuse the domain classifier about the domain origin of the features. The domain-invariant features that are obtained through this approach, are the motivation for applying UDA in EEG data in the work of Jin *et al.* [106], achieving performance improvements when combined with simple MLP deep architectures. Other works that employ domain adaptation for affect recognition include [137] and [139].

Subject-specific & subject-independent models: A major division in EEG-based affect recognition is between works that build subject-dependent and subject-independent models, as this selection plays a significant role on the generalization capabilities of the trained models. Subject-dependent models utilize data originating only from one subject, to achieve better adaptation to new trials of the same subject, with the cost of limiting the available training samples. Subject-independent models are trained on data from many subjects, achieving better generalization on unseen subjects.

2.6 Datasets and evaluation metrics

In this section we provide details on the datasets and evaluation metrics that are used in the experimental analyses of this thesis. We begin by describing the datasets and metrics used on the task of motor imagery decoding and then continue by referring to the task of affect recognition.

Datasets on motor imagery:

The first dataset that we use on the motor imagery decoding problem is "**BCI Competition IV Dataset 2a**" (IV-2a[1]) [215]. The dataset contains EEG recordings of 9 participants, collected over two different days for each subject (i.e. there are two sessions per participant), having 25 electrodes (22 EEG and 3 electrooculographic channels) and a sampling frequency of 250Hz. The classes of the dataset correspond to 4 different imaginary movement types that the subjects performed, namely left hand, right hand, feet and tongue. Each session contains 72 trials of each class.

The protocol for collecting the IV-2a dataset is shown in Fig. 2.9. In the beginning of each trial (i.e. $t = 0sec$), a fixation cross appears on the black screen and a short acoustic warning tone is presented. Then, at $t = 2sec$ a visual cue appears in the

[1]More details about IV-2a dataset are provided by the `moabb` library.

screen for 1.25sec. This cue is an arrow pointing towards one out of four possible directions (left, right, down or up) corresponding to one of the four classes (left hand, right hand, foot or tongue). This cue indicates to the subjects which motor movement they should imagine. Afterwards, the subjects imagine this movement. At $t = 6sec$ the fixation cross stops appearing on the screen and the subjects stop imagining the motor movement. Each trial finishes with a short break of variable duration, where the screen remains black.

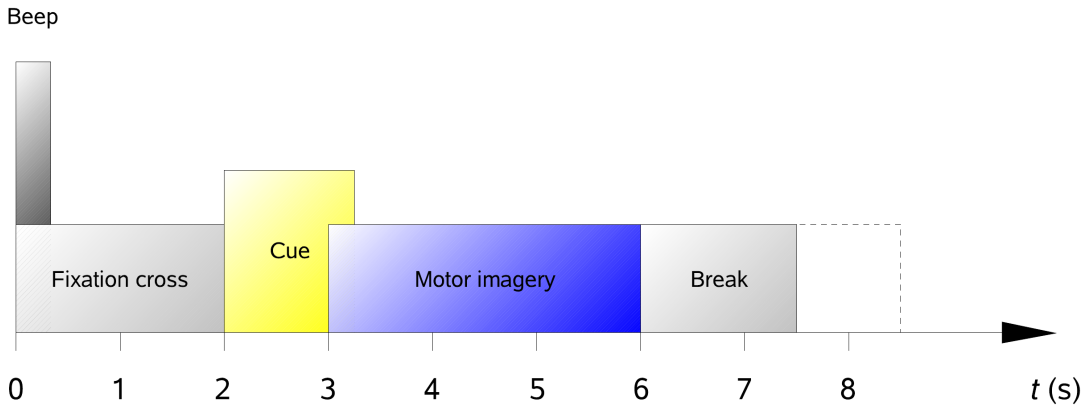


Figure 2.9: Illustration of the protocol that was employed to collect the data of the IV-2a motor imagery dataset. Figure taken from [215].

The second dataset that we use is "**Physionet EEG Motor Movement/Imagery Dataset**" (Physionet[2]) [79, 200]. The dataset contains EEG recordings from 109 participants, with trials that belong to 4 classes: left-hand, right-hand and feet imagery, as well as rest. The data are recorded with 64 EEG electrodes at a sampling frequency of 160Hz.

The third dataset that we use is **OpenBMI**[3] [135]. The data of OpenBMI correspond to trials of 2 classes (left-hand and right-hand imagery) collected from the EEG recordings of 54 participants, with 62 electrodes at a sampling frequency of 1000Hz. Each participant has data from two sessions and each session has two runs. The first

[2]More details about Physionet dataset are provided by the `moabb` library.
[3]More details about OpenBMI dataset are provided by the `moabb` library.

run of each session is done in an offline manner, i.e. without feedback. The second run is done in an online manner, providing real-time visual feedback to the user.

In all three aforementioned motor imagery datasets, we use classification accuracy as our evaluation measure. We note that the class label distributions of these datasets are not imbalanced.

Datasets on affect recognition:

The experimental protocol followed during the collection of EEG-based affective datasets, has some general directions that remain similar across datasets. The participants are initially equipped with all the necessary devices to measure their neurophysiological signals, and then the stimuli presentation process begins. There are three periods during this process, namely the resting state (also called "baseline") period, the stimuli presentation ("trial") period where the subjects are shown short videos as stimuli and the self-assessment period. The resting state serves as a period that the subject remains neutral, and can act as a reference point to be compared with the trial period where the emotions are elicited while the participant observes the stimulus. After each stimulus is presented, the participant rates the emotion felt during the presentation, either using an annotation graphical interface, or by filling an assessment form.

The annotation for each affective dimension of a video is a single label for the entire duration (i.e., one *Arousal* label, one *Valence* label, etc.), prohibiting any reliable attempt for temporally fine-grained emotion recognition. When dealing with video stimuli that last tens or hundreds of seconds, it is reasonable to consider that the emotions reported by participants for a single video, are not elicited constantly in the same level for the whole duration of the stimulus. For this reason, early moments of each trial, where the emotion-eliciting stimulus content might not have unfolded sufficiently, are usually excluded from training/testing.

The first dataset that we use on the affect recognition problem is "**DEAP**" [4]) [122]. DEAP is a database for spontaneous emotion analysis, containing neurophysiological signals of participants that watched and rated their emotional response to music video clips. The emotions were rated along the scales of arousal, valence, and dominance. The participants also rated the liking of and familiarity with the videos. The self-assessment ratings of arousal, valence, dominance and liking are on a $[1.0, 9.0]$ continuous scale, and they can be mapped to the binary low/high values by thresholding them in the middle value of 5.0. Typically, the evaluation of algorithms on DEAP is done on the problems of binary (low versus high) classification for the target attributes of arousal and valence, using the ratings of the participants as groundtruth. DEAP contains data from 32 participants and 40 music video clips as stimuli, with a fixed duration of 60 seconds for each clip. The data are recorded with 32 EEG electrodes at a sampling frequency of 512Hz.

The second dataset that we use on the affect recognition problem is "**SEED**" [5]) [256]. SEED is a dataset for EEG-based emotion recognition, having 15 participants and 15 Chinese movie videos as stimuli, with varying duration for each clip (4 minutes in average). The data collection was repeated three times, in different days, for each subject (i.e. there are three sessions per participant). The data are recorded with 62 EEG electrodes at a sampling frequency of 1000Hz. The annotations are categorical, belonging in three classes, namely "Positive", "Neutral" and "Negative". A single class label is assigned to each video stimulus from the organizers of the experiment. That is, the labels of SEED are not separately self-reported by each participant, but are rather "global", as defined by the researchers that conducted the study.

In both of the aforementioned affect recognition datasets, we use classification accuracy and F1 score as our evaluation metric. The class label distribution of DEAP is imbalanced, hence false conclusions might be drawn about the performance of the

[4] More details about DEAP dataset are provided in its official project page.
[5] More details about SEED dataset are provided in its official project page.

developed models, when solely observing the accuracy measure.

2.7 Conclusions

In this chapter we began by providing an overview of the general fields of electroencephalography analysis and brain-computer interfaces. Then we introduced the principles of motor imagery decoding and discussed state-of-the-art methods that address it. We followed by explaining the topics of emotion modelling, annotation and recognition and then presented the state-of-the-art research works on emotion recognition. Finally, we described the datasets and the evaluation metrics that are used in the experiments of this thesis. In this section, we draw several important conclusions from our literature review on the topics of motor imagery decoding and affect recognition.

Firstly, we identify gaps in the existing literature regarding the topic of motor imagery decoding, especially in cross-subject experimental settings. Existing ensemble learning works that fall within the broad family of domain generalization techniques, present three main disadvantages: (i) increased computational complexity, (ii) lengthy model selection processes and (iii) less generic feature extractors. The ensemble learning technique proposed in Chapter 3 of this thesis, is a novel method that does not inherit any of the three aforementioned negative aspects. Specifically, it is based on an ensemble curriculum learning scheme that promotes feature diversity across multiple models that act as feature extractors. Our architecture is a simplistic model ensemble without bells and whistles, yielding strong performance with a compact model size. This is in contrast to works that try to build diverse feature extractors through complex architectures [67] (e.g. using multiple inception-based branches, different number of filters per branch, different filter length per branch, etc.) or through training multiple subject-specific [233, 39] (thus also less generic) models. Our ensemble learning method is also trained in an end-to-end manner and in a single phase, thus is a more attractive alternative to works that require multiple hyperparameter tuning runs [64]

to train each base model.

The method described in Chapter 4 of this thesis, called "CovMix", is a novel regularization technique that simultaneously performs data augmentation and data alignment during training. Previous works that act as regularizers either on the input space (e.g. MixUp [252]) or on the feature space (e.g. MixStyle [257]) of CNNs, are not motivated by neuroscience-grounded principles. Applying such methods to regularize CNNs, involves mixing feature statistical distributions in an inter-domain manner (i.e. across training subjects). Due to the existence of inter-subject differences in the characteristics of EEG signals, mixing statistics across subjects might not be helpful towards learning meaningful representations. On the contrary, CovMix mixes statistics in an intra-domain manner, by fusing trial-wise and session-wise statistics for each individual subject. In essence, CovMix achieves regularization by transforming the original data with a different spatial filtering operation in each training iteration. By doing so, CovMix not only performs data alignment [89] that has been shown to help transfer learning, but also performs data augmentation through its stochastic nature. To the best of our knowledge, our work is the first to combine the steps of data alignment and data augmentation in the same operation, for neural time-series decoding.

Regarding the topic of affect recognition, we note that the majority of works transform the original affective annotations provided by datasets into classes that have a nominal nature. By doing this, such methods disregard the existing ordinality in the structure of emotions. Thus, they rely on plain classification approaches and cannot exploit fine-grained information such as the relations between training samples with respect to their original affective ratings. Our proposed affect recognition method that is presented in Chapter 5 of this thesis, introduces a training methodology that combines the tasks of classification and ranking during training. More specifically, apart from the standard sample-wise classification task, we propose the task of pairwise ranking of samples with respect to their affective ratings, as an additional training objective. In

this way, we are able to learn representations that not only capture the coarse division between classes (e.g. "high arousal" versus "low arousal"), but are also helpful towards inferring pairwise ordinal relations between samples.

Chapter 3

Motor Imagery Decoding Using Ensemble Curriculum Learning and Collaborative Training

Contents

3.1 Introduction

In this chapter, we consider the problem of cross-subject motor imagery decoding and propose a method that presents robustness against several domain shifts (i.e. variations of EEG signal characteristics [151] across individuals). As research efforts on the field of BCIs are focusing on obtaining strong cross-subject performance, two families of learning techniques have gained increasing interest: domain adaptation and domain generalization. The major drawback of domain adaptation methods has been their requirement of having available data from the test subjects during the training phase.

Domain adaptation methods may require either unlabelled test data, or even a small portion of labelled test data [93]. Domain generalization methods are rather easier to be adopted by researchers, as they can be used without accessing the testing data during training. There are several works on EEG-based domain generalization problems, that do not explicitly attempt to address inter-subject differences by any means, thus equally minimizing the loss over all training subjects simultaneously [12, 258, 176]. Hence the limitation of such works is that they practically correspond to the trivial domain generalization approach of Empirical Risk Minimization (ERM) [219]. Our proposed method falls within the umbrella of domain generalization techniques and overcomes this limitation by treating differently each training subject.

Our method is based on the concepts of model ensembling, curriculum learning and knowledge distillation and is able to learn diverse feature representations that can lead to improved cross-subject generalization. Existing works on model ensembling often require independently training several individual models [64]. Apart from being time-consuming, such training strategies impede jointly training multiple models in a single phase. Joint single-phase training is necessary in order to measure and control the diversity of feature representations across the multiple models, as training progresses. Our method effectively addresses this need, through our novel two-stage model ensemble architecture, built with multiple feature extractors (first stage) and a shared classifier (second stage), that allows to be trained in a single phase. An overview of our proposed two-stage architecture is shown in Fig. 3.1.

To promote feature diversity, we introduce an ensemble curriculum learning scheme, that enforces each feature extraction model of our ensemble architecture to focus on different subjects of the training set. This scheme is materialized through the first subject-wise loss term that we use to train our architecture. When trained using this loss term, our architecture covers a wide range of patterns through several models that act as diverse feature extractors. Previous works have shown that there is a trade-off

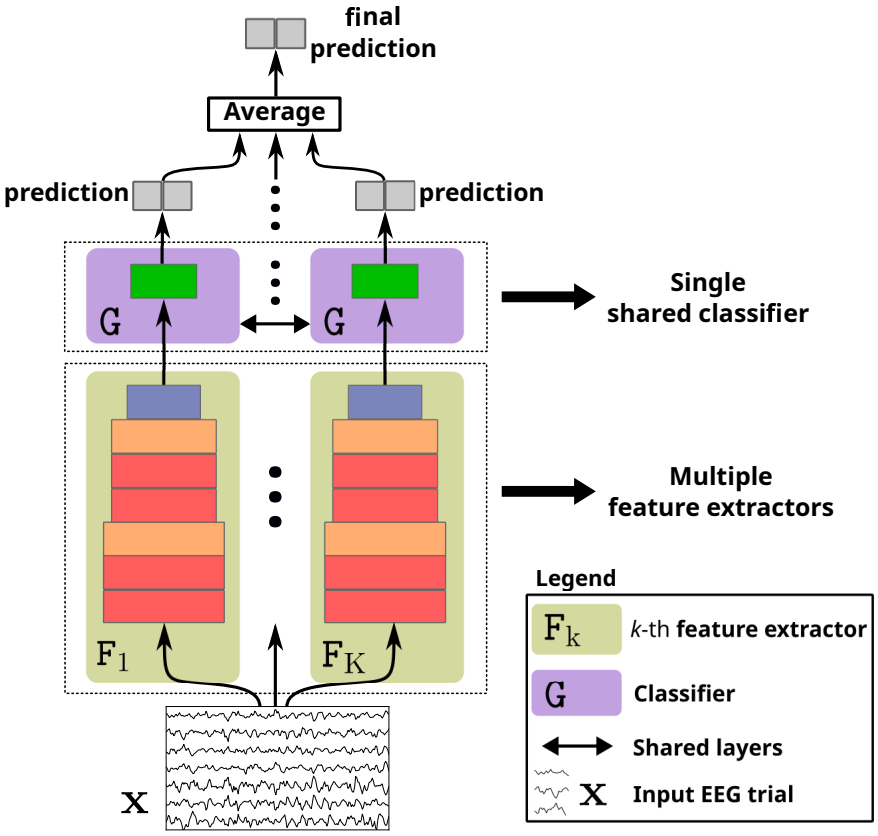


Figure 3.1: Overview of our proposed ensemble architecture during inference. First, an input EEG trial is fed to multiple feature extractors that produce diverse feature representations. Then, a single shared classifier predicts the class scores corresponding to each feature representation, and these class scores are averaged to compute the final prediction.

between diversity and generalization [28], which means that diversity is desired up to a certain extent, further than which it can have a detrimental effect on generalization. To regulate this trade-off, we introduce an intra-ensemble distillation loss term, that controls the diversity within the ensemble. The combination of the two aforementioned loss terms helps balancing diversity and generalization, which indicates that our second loss term acts complementary to the first one and leads to further performance improvements.

Our contributions are the following:

- We propose a simplistic and compact two-stage model ensemble architecture that allows to be trained in a single phase, without requiring any model selection process. All models of our ensemble are sufficiently generic feature extractors as they are trained on the entire training set.

- We pair our architecture with a novel curriculum learning scheme that promotes diversity across the models of the ensemble. Thus, each model specializes to a different subset of training subjects. To our knowledge, curriculum learning has not been previously explored for cross-subject MI decoding.

- We introduce an auxiliary loss term that is based on the concept of knowledge distillation across models. Our proposed intra-ensemble distillation loss balances the diversity-generalization trade-off of our architecture, leading to further performance improvement.

- We conduct our experimental analysis on two large motor imagery datasets (Physionet [79] and OpenBMI [135]) totalling more than 150 subjects. We compare our method against state-of-the-art and standard ensembling techniques showing superior results.

The rest of this chapter is organized as follows. in Section 3.2 we describe our method, i.e. our proposed architecture along with the loss terms that are used to train it. In Section 3.3 we present the results of our experimental analyses and ablation studies, where we compare our work with other state-of-the-art works, ensembling techniques and a single-model baseline as a reference. Lastly, in Section 3.4 we conclude the chapter.

3.2 Proposed Method

In this section we describe the proposed methodology, which consists of a model ensemble architecture, a curriculum training scheme and an intra-ensemble distillation

loss. We provide an overview of the training pipeline for our proposed architecture in Fig. 3.2 and present its individual components in the following subsections. Specifically, we begin by explaining our ensemble architecture in Subsection 3.2.1. Then, we introduce the first loss term that materializes our curriculum learning scheme in Subsection 3.2.2, as well as the second loss term that enables collaborative training across the models of the ensemble, in Subsection 3.2.3.

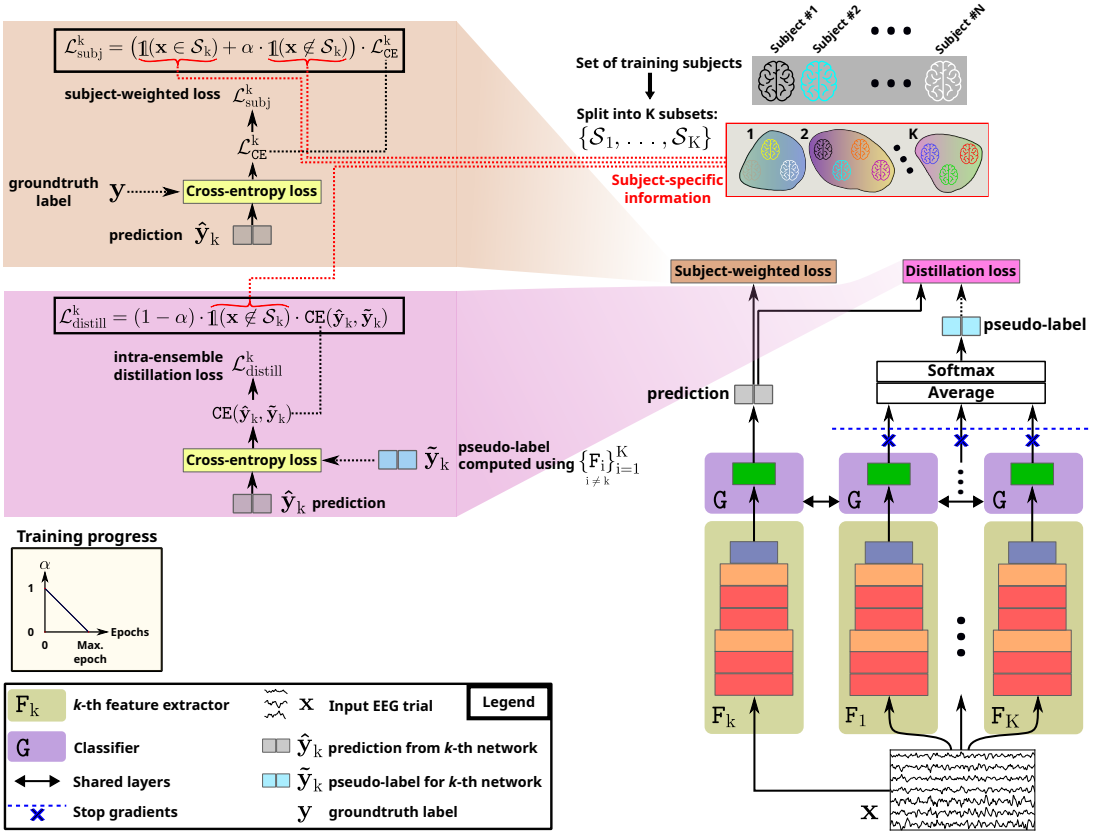


Figure 3.2: Our proposed architecture has K first stage models and a shared classifier in the second stage. The input trial $\mathbf{x}$ is separately passed to each one of the first stage models, obtaining the feature vectors $[\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_K}]$ (Eq. 3.3). For the k-th model, the class-wise scores $\hat{\mathbf{y}}_\mathbf{k}$ are computed by forwarding $\mathbf{f_k}$ to the shared classifier of the second stage (Eq. 3.4). In an ensembling scenario where the architecture is trained without curriculum learning, we compute the individual model losses $\mathcal{L}^{k}_{CE}$ (Eq. 3.5) and minimize the loss $\mathcal{L}^{total}_{CE}$ (Eq. 3.6) for all models. In the ensemble curriculum learning scenario, we compute the individual subject-weighted losses $\mathcal{L}^{k}_{subj}$ (Eq. 3.9) and minimize the loss $\mathcal{L}^{total}_{subj}$ (Eq. 3.10) for all models. When also performing collaborative training, we additionaly compute the losses $\mathcal{L}^{k}_{distill}$ (Eq. 3.12) and minimize the total loss $\mathcal{L}_{total}$ (Eq. 3.14) for all models.

3.2.1 Architecture

Single model

In this work, we use the well-established EEGNet [133] architecture as our strong single-model baseline. The selection of EEGNet is justified from the fact that it achieves compelling performance, with a reasonably small number of trainable parameters and a simple network design (e.g. without streams of varying kernel lengths, or band-wise processing streams). In the task of MI decoding, the time-series signals $\mathbf{x} \in \mathbb{R}^{\mathrm{C}\times\mathrm{T}}$ of an EEG trial with C electrodes and T samples in the temporal dimension, are fed as input to EEGNet. The class-wise scores $\hat{\mathbf{y}} \in \mathbb{R}^{\mathrm{N_C}}$ (where $\mathrm{N_C}$ is the number of classes) are obtained as output, while the groundtruth label $\mathbf{y} \in \mathbb{R}^{\mathrm{N_C}}$ is represented in the form of a one-hot vector. Thus, in the case of EEGNet the output scores are computed as

$$\hat{\mathbf{y}} = \texttt{EEGNet}(\mathbf{x}), \tag{3.1}$$

and the network is optimized by minimizing the cross-entropy (CE) loss $\mathcal{L}_{\texttt{CE}} = \texttt{CE}(\hat{\mathbf{y}}, \mathbf{y})$, given by

$$\texttt{CE}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{\mathrm{i}=1}^{\mathrm{N_C}} y_{\mathrm{i}} \log\Big(\texttt{softmax}(\hat{y}_{\mathrm{i}})\Big), \tag{3.2}$$

where y_{i} and $\hat{y}_{\mathrm{i}}$ are the i-th elements of $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively. The detailed architecture of EEGNet is shown in Table 3.1.

Model ensemble

Our proposed model ensemble architecture, shown in Fig. 3.2, uses EEGNet as its elementary component and consists of two stages. The first stage contains multiple feature extraction networks having exactly the same design, with each network producing a feature vector. Considering the EEGNet architecture that is presented in Table 3.1, each first stage network contains all the layers up to (and including) the feature flattening layer of EEGNet. We use $\texttt{F}_{\mathrm{k}}(\cdot)$ and $\mathbf{f}_{\mathrm{k}}$ to denote the k-th feature extractor and its output feature vector. The output feature vectors from the first stage,

Table 3.1: Architecture of a single EEGNet model. The input of the model has a shape of $B \times 1 \times C \times T$, where B is the batch size, C is the number of EEG electrodes and T is the number of samples in the temporal dimension. The output of the model has a shape of $B \times 2$, in the case of two output classes.

Layer	**Input shape**	**Output shape**
Dropout (p=0.4)	$B \times 1 \times C \times T$	$B \times 1 \times C \times T$
Temporal Convolution, 8 filters kernel=(1, 64), stride=(1, 1), pad=(0, 32)	$B \times 1 \times C \times T$	$B \times 8 \times C \times T$
Spatial Convolution, 16 filters kernel=(1, C), stride=(1, 1), pad=(0, 0) max. weight norm=1.0	$B \times 8 \times C \times T$	$B \times 16 \times 1 \times T$
Temporal Pooling kernel=(1, 4), stride=(1, 4), pad=(0, 0)	$B \times 16 \times 1 \times T$	$B \times 16 \times 1 \times T$
Batch Normalization 2D	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
ELU activation	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
Dropout (p=0.1)	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
Separable Convolution Depthwise, 16 filters, 16 groups kernel=(1, 16), stride=(1, 1), pad=(0, 8)	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
Separable Convolution Pointwise, 16 filters, 16 groups kernel=(1, 1), stride=(1, 1), pad=(0, 0)	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
Batch Normalization 2D	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
ReLU activation	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/4$
Temporal Pooling kernel=(1, 8), stride=(1, 8), pad=(0, 0)	$B \times 16 \times 1 \times T/4$	$B \times 16 \times 1 \times T/32$
Flatten	$B \times 16 \times 1 \times T/32$	$B \times T/2$
Fully Connected	$B \times T/2$	$B \times 2$

are computed as

$$[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_K] = [\mathtt{F}_1(\mathbf{x}), \mathtt{F}_2(\mathbf{x}), \ldots, \mathtt{F}_K(\mathbf{x})]. \tag{3.3}$$

The second stage has a single shared classification head $\mathtt{G}(\cdot)$, that computes the class-wise prediction scores for each feature vector originating from the first stage. Based on the EEGNet layers that are presented in Table 3.1, the second stage of our architecture corresponds to the last layer of EEGNet, i.e. a single fully connected layer that performs classification. We use $\hat{\mathbf{y}}_k$ to denote the scores corresponding to the k-th feature vector $\mathbf{f}_k$. The scores are computed as

$$[\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_K] = [\mathtt{G}(\mathbf{f}_1), \mathtt{G}(\mathbf{f}_2), \ldots, \mathtt{G}(\mathbf{f}_K)]. \tag{3.4}$$

In the simple scenario where no curriculum learning occurs, this architecture is trained by minimizing the sum of the individual losses for the predictions of each model. The loss $\mathcal{L}_{\mathrm{CE}}^{\mathrm{k}}$ for the predictions $\hat{\mathbf{y}}_{\mathrm{k}}$ of the k-th model, and the total loss $\mathcal{L}_{\mathrm{CE}}^{\mathrm{total}}$, are computed as

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{k}} = \mathtt{CE}(\hat{\mathbf{y}}_{\mathrm{k}}, \mathbf{y}) \tag{3.5}$$

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{total}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathcal{L}_{\mathrm{CE}}^{\mathrm{k}}. \tag{3.6}$$

In the inference phase, to classify an input sample $\mathbf{x}$ we fuse the model-wise scores through a simple average operation and obtain a final score vector $\hat{\mathbf{y}}_{\mathrm{ens}}$ as follows:

$$\hat{\mathbf{y}}_{\mathrm{ens}} = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{\mathbf{y}}_{\mathrm{k}}. \tag{3.7}$$

To this end, the described architecture is purely subject-agnostic, having no subject-specific layers (both in the first and second stage). In the following subsection we propose an ensemble curriculum learning scheme that is applied during training and changes the nature of the first stage layers. Our curriculum provides a strong alternative to the typical subject-agnostic layers, that can be adopted in ensemble learning.

3.2.2 Ensemble curriculum learning

Our goal is to make each feature extractor to specialize on a specific subset of subjects. That is, we want to induce *local* (i.e. focused on a subset of the entire training set) feature extraction power to each model in the first stage. Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{\mathrm{N}}\}$ be a dataset with the data of N subjects, where $\mathcal{D}_{\mathrm{n}}$ denotes the sub-dataset containing the trials of the n-th subject. For an ensemble with K models ($\mathrm{K} \geq 2$), we split $\mathcal{D}$ into K non-overlapping subsets $\mathcal{S}$: $\mathcal{D} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{\mathrm{K}}\}$. We do this splitting process by randomly assigning the sub-dataset of each subject to one of the K subsets, with a uniform probability for all subsets. Therefore, we have $\bigcup_{\mathrm{k}=1}^{\mathrm{K}} \mathcal{S}_{\mathrm{k}} = \mathcal{D}$ and $\mathcal{S}_{\mathrm{i}} \cap \mathcal{S}_{\mathrm{j}} = \emptyset$ for $i \neq j$. Each subset $\mathcal{S}_{\mathrm{k}}$ corresponds to the k-th model and contains the sub-datasets of the subjects on which we drive the k-th model to specialize.

To achieve this specialization, we design a *subject-weighted* loss function where we inject subject-specific coefficients to weigh the contribution of each subject to the loss of each model. Considering the subject-weighted loss $\mathcal{L}^{\mathrm{k}}_{\mathrm{subj}}$ that is used to train the k-th model, the subject-specific coefficients linearly decay over epochs the loss contribution of the subjects that *do not* belong to $\mathcal{S}_{\mathrm{k}}$. Effectively, this makes the k-th model to focus more on the subjects of $\mathcal{S}_{\mathrm{k}}$ that have a non-decaying loss contribution. We scale the contribution of a training sample $\mathbf{x}$ to the loss $\mathcal{L}^{\mathrm{k}}_{\mathrm{subj}}$ through the coefficient $\beta(\mathbf{x}, \mathrm{k})$. If trial $\mathbf{x}$ corresponds to a subject that belongs in $\mathcal{S}_{\mathrm{k}}$ (hence $\mathbf{x} \in \mathcal{S}_{\mathrm{k}}$), then we keep $\beta(\mathbf{x}, \mathrm{k}) = 1$ throughout the whole training process. Otherwise ($\mathbf{x} \notin \mathcal{S}_{\mathrm{k}}$), we decay $\beta(\mathbf{x}, \mathrm{k})$ from 1 to 0 while training progresses, that is:

$$\beta(\mathbf{x}, \mathrm{k}) = \begin{cases} 1 & , \text{if } \mathbf{x} \in \mathcal{S}_{\mathrm{k}} \\ \alpha & , \text{if } \mathbf{x} \notin \mathcal{S}_{\mathrm{k}} \end{cases}, \tag{3.8}$$

where $\alpha = 1 - \frac{\text{epoch}}{\mathrm{N}_{\text{epochs}}} \in [0, 1]$ represents the progression of training, as $\mathrm{N}_{\text{epochs}}$ is the maximum number of training epochs and epoch is the current epoch. The loss $\mathcal{L}^{\mathrm{k}}_{\mathrm{subj}}$ of the k-th model and the total subject-weighted loss $\mathcal{L}^{\mathrm{total}}_{\mathrm{subj}}$ are computed as follows:

$$\mathcal{L}^{\mathrm{k}}_{\mathrm{subj}} = \underbrace{\beta(\mathbf{x}, \mathrm{k})}_{\substack{\text{subject-specific} \\ \text{coefficient}}} \cdot \mathcal{L}^{\mathrm{k}}_{\mathrm{CE}} \tag{3.9}$$

$$\mathcal{L}^{\mathrm{total}}_{\mathrm{subj}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathcal{L}^{\mathrm{k}}_{\mathrm{subj}}. \tag{3.10}$$

To allow a better understanding of the coefficient $\beta(\mathbf{x}, \mathrm{k})$ that is involved in Eq. 3.9, we show an overview of our curriculum learning scheme in Fig. 3.3. Specifically, we consider an example where we are provided with a dataset $\mathcal{D}$ containing EEG data from 10 human subjects and our proposed model ensemble architecture consists of $\mathrm{K} = 3$ models. The dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{10}\}$ containing the sub-datasets of 10 subjects is split into $\mathrm{K} = 3$ non-overlapping subsets: $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$. Our curriculum learning scheme aims to make the k-th model to specialize on the subjects belonging to subset $\mathcal{S}_{\mathrm{k}}$ of $\mathcal{D}$, while still training on the whole dataset $\mathcal{D}$. This is done using

the coefficient $\beta(\mathbf{x}, \mathrm{k})$ that controls the loss contribution of a training sample $\mathbf{x}$ on the weight updating process for the k-th model. To achieve specialization on the samples of $\mathcal{S}_\mathrm{k}$, when $\mathbf{x} \in \mathcal{S}_\mathrm{k}$, we set $\beta(\mathbf{x}, \mathrm{k}) = 1$ throughout the training process. We also want to train the k-th model on the rest of the subjects of $\mathcal{D}$ (i.e. those that do not belong to $\mathcal{S}_\mathrm{k}$), albeit with a progressively decreasing loss contribution over time. For this reason, when $\mathbf{x} \notin \mathcal{S}_\mathrm{k}$ we set $\beta(\mathbf{x}, \mathrm{k}) = \alpha$, with α decaying from 1 to 0 while training progresses.

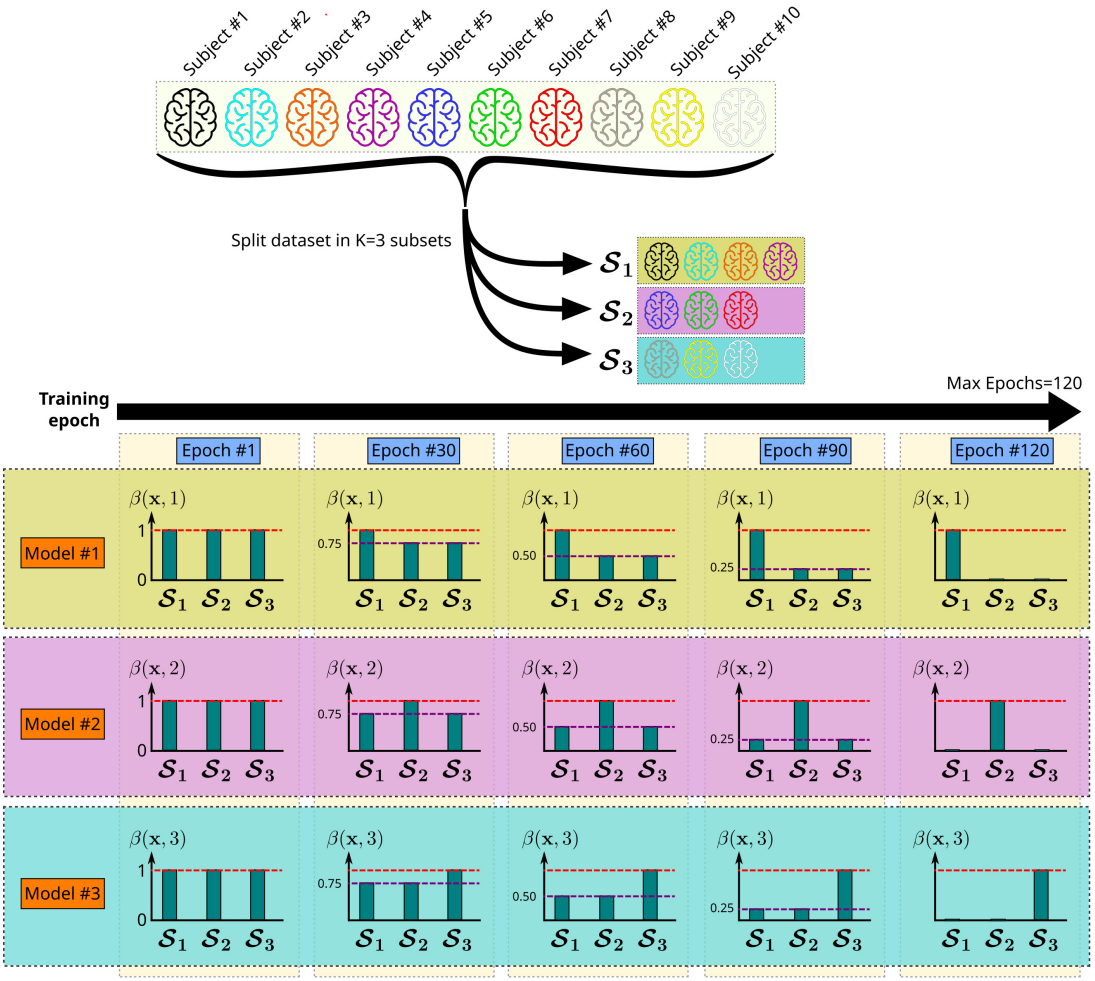


Figure 3.3: Indicative illustration of our curriculum learning scheme. In this example, we are provided with a dataset $\mathcal{D}$ containing EEG data from 10 human subjects and our proposed model ensemble architecture consists of $\mathrm{K} = 3$ models.

3.2.3 Intra-ensemble distillation for collaborative training

In this subsection we propose a collaborative training scheme which helps to regulate the diversity-generalization trade-off in our model ensemble. In order to classify a sample, we extract its first stage representations, feed them to the shared classifier of the second stage and average the individual scores across models. The diversity between the first stage representations of a sample can make the classifier to compute inconsistent class scores across models. This, in turn, can negatively affect the final prediction scores, as they will be the result of fusing multiple contradicting predictions. We observe that, although feature diversity is a desirable property of our ensemble, it can also have an adverse effect on the generalization capabilities.

To overcome this phenomenon, we introduce a loss term that promotes consistency across the multiple model predictions, in order to improve the performance of the entire ensemble. We design our proposed intra-ensemble distillation loss to operate on the predicted scores of the second stage, instead of operating on the features extracted from the first stage. An overview of our distillation loss is shown in Fig. 3.2. Considering each prediction $\hat{\mathbf{y}}_{\mathrm{k}}$ of the k-th model, our loss pushes it closer to the softmaxed average of the predictions from all the other models (which is the pseudolabel in our distillation loss). Specifically, we compute the pseudolabel $\tilde{\mathbf{y}}_{\mathrm{k}}$ for the k-th model as:

$$\tilde{\mathbf{y}}_{\mathrm{k}} = \texttt{softmax}\Bigg(\frac{1}{\mathrm{K}-1}\sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{k}}^{\mathrm{K}}\hat{\mathbf{y}}_{\mathrm{i}}\Bigg), \tag{3.11}$$

and use the cross-entropy loss between the prediction $\hat{\mathbf{y}}_{\mathrm{k}}$ and the pseudolabel $\tilde{\mathbf{y}}_{\mathrm{k}}$. We note that we apply a stop-gradient [45] operation on the pseudolabels, as shown in Fig. 3.2. We do this in order to ensure that only the weights of the k-th model are updated based on this loss term, while the other models remain unaffected. For the k-th model, we opt to *not* apply this loss on the samples of $\mathcal{S}_{\mathrm{k}}$. This is done through a binary mask that zeroes out the distillation loss of these samples. We do so, as the k-th model is sufficiently trained on the samples of $\mathcal{S}_{\mathrm{k}}$ through their groundtruth labels $\mathbf{y}$.

We note that it is necessary to scale the contribution of the intra-ensemble distillation loss to the total loss of the architecture, in accordance with the progress of training. In the beginning of the training process, the weights of the architecture are randomly initialized. Hence, penalizing the distance of model predictions from the derived pseudolabels is not so meaningful in the early epochs. As training proceeds, each feature extractor progressively focuses on a subset of subjects and feature diversity increases. As shown later in the experiments, our distillation loss indirectly controls this emerging feature diversity by bringing closer the class scores computed from various first stage features. We linearly increase the contribution of the distillation loss to the total loss, across training epochs, by multiplying it with the scalar $1-\alpha$ that quantifies the training progress.

The distillation loss $\mathcal{L}^{\mathrm{k}}_{\mathrm{distill}}$ of the k-th model, and the total distillation loss $\mathcal{L}^{\mathrm{total}}_{\mathrm{distill}}$ are computed as follows:

$$\mathcal{L}^{\mathrm{k}}_{\mathrm{distill}} = (1-\alpha)\cdot \underbrace{\mathbb{1}(\mathbf{x} \notin \mathcal{S}_{\mathrm{k}})}_{\substack{\text{subject-specific} \\ \text{mask}}} \cdot \mathtt{CE}(\hat{\mathbf{y}}_{\mathrm{k}}, \tilde{\mathbf{y}}_{\mathrm{k}}) \tag{3.12}$$

$$\mathcal{L}^{\mathrm{total}}_{\mathrm{distill}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathcal{L}^{\mathrm{k}}_{\mathrm{distill}}. \tag{3.13}$$

We compute the total loss $\mathcal{L}_{\mathrm{total}}$ of our architecture as:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{subj}} \cdot \mathcal{L}^{\mathrm{total}}_{\mathrm{subj}} + \lambda_{\mathrm{distill}} \cdot \mathcal{L}^{\mathrm{total}}_{\mathrm{distill}}, \tag{3.14}$$

where we empirically set $\lambda_{\mathrm{subj}} = \mathrm{K}$ and $\lambda_{\mathrm{distill}} = 0.7$.

3.3 Experimental results

3.3.1 Datasets

We apply our method on the problem of motor imagery decoding and work on two large datasets: Physionet [79, 200] and OpenBMI [135]. A brief description of these

datasets was provided in Chapter 2.6. In the experiments that we conduct on Physionet dataset, we use the signals from all 64 electrodes. Regarding our experiments on OpenBMI dataset, when not otherwise stated, we use a subset of 20 electrodes and we use the data from the two offline runs (i.e. the first run of the first and second session) for each participant, following the default settings of the MOABB [100] benchmark. Our preprocessing steps for both datasets are the following: (i) we remove the 50Hz component through notch filtering (60Hz for OpenBMI) (ii) we perform bandpass filtering (4Hz-38Hz) (iii) we resample the signals to 100Hz and (iv) for each trial, we crop a temporal window of 4 seconds, starting from its onset event. Upon obtaining the cropped trials, we use the session-wise covariance matrices of the EEG signals and perform Riemannian Alignment on the time-series of each trial, as in [259].

3.3.2 Comparison with other works and baseline

We compare our proposed method with four state-of-the-art techniques that provide their source code, namely Adaptive Transfer Learning (ATL) [253], EEGSym [176], TIDNet [126] and MIN2Net [12]. In order to fairly judge the impact of our proposed methodology, we also implement three additional methods: a single model baseline and two ensembling techniques using the EEGNet architecture.

EEGNet-Single: The baseline method (mentioned as "EEGNet-Single") is a single EEGNet model, that serves as a reference for the performance of an EEGNet architecture without ensembling.

EEGNet-Ensemble: We implement the first ensembling method by training multiple individual EEGNet models in the entire training set. During inference, we fuse their predictions through a simple averaging operation to obtain the final prediction. In essence, this ensembling method (mentioned as "EEGNet-Ensemble") represents a post-training model ensemble.

EEGNet-Bagging: We implement the second ensembling method by training mul-

tiple individual EEGNet models in random subsets of the training set. Specifically, we train each individual EEGNet model on 85% of all the available training subjects. We choose the subjects to be kept for training in each experiment, by simply performing random subsampling. During inference, we fuse the predictions of all models through an averaging operation to obtain the final prediction. This method (mentioned as "EEGNet-Bagging") represents the well-known ensembling technique of bootstrap aggregating [33].

3.3.3 Evaluation settings

We perform evaluation in two ways: (i) in a 5-fold cross-validation (CV) manner and (ii) in a Leave-One-Subject-Out (LOSO) manner. In the 5-fold CV scenario, we split the subjects of our dataset into 5 disjoint folds and run 5 experiments. In each experiment, we use a different fold as our test set and then assign 3 folds to our training set and the 1 remaining fold to our validation set. In the LOSO scenario for a dataset with N subjects, we run N experiments where in the n-th experiment we use the data of the n-th subject as our test set. In each experiment, we split the remaining $\mathrm{N}-1$ subjects into our training and validation set. Specifically, we assign 80% of these subjects to the training set and the rest 20% to the validation set of the experiment. In both CV and LOSO scenarios, the reported accuracy is the average of the test accuracies across all experiments.

3.3.4 Training details

We train all models (i.e. our proposed method, the single model baseline and the model ensembling methods) for 120 epochs with a batch size of 64. We use a Stochastic Gradient Descent (SGD) optimizer, setting the momentum to 0.9 and weight decay to 0.01. We initialize the learning rate at 0.01 for the first 60 epochs and then decrease it to 0.002 for the remaining 60 epochs.

3.3.5 Results (5-fold cross-validation)

In the first part of our experimental analysis we evaluate against methods that provide source code, under a 5-fold cross-validation scenario, without any model adaptation on test data or pretraining on external datasets. We note that these experiments are performed using exactly the same train, validation and test splits, the same trial length and the same number of electrodes for all methods (except for the method of EEGSym [176] that has an architectural requirement of 16 electrodes). Having the same experimental settings enables us to fairly judge the performance of all methods.

Table 3.2 shows the results of the methods trained on Physionet dataset with 5-fold cross-validation. Concerning our baseline method, we observe that a single EEGNet model achieves an accuracy of 82.09%. The standard model ensembling technique of EEGNet-Ensemble reaches an accuracy of 84.56% when fusing eight individual EEGNet models, while EEGNet-Bagging performs slightly better reaching an accuracy of 84.81%. Our proposed method presents a substantial boost of +1.80% over the standard ensemble scenario, reaching an accuracy of 86.36% when we use seven models at the first stage of our architecture. The model of EEGSym [176] achieves an accuracy of 83.91%, using $\sim 10\times$ more trainable parameters than the best performing architecture of our proposed method. EEGSym without pretraining on external data, performs worse than both our method and the standard model ensemble. Regarding the method of TIDNet [126], the accuracy of 82.19% is similar to that of our EEGNet-Single baseline model.

Table 3.3 shows the results of our 5-fold cross-validation experiment on OpenBMI for various methods. The baseline model of EEGNet-Single achieves an accuracy of 78.31% and the method of EEGNet-Ensemble provides a small boost of +0.67% when using eight individual EEGNet models, leading to an accuracy of 78.98%. The ensembling method of EEGNet-Bagging performs slightly better than the standard ensembling. Our proposed method performs superiorly, yielding an accuracy of 79.73% when using

three first stage networks. The method of MIN2Net [12] has a low performance, with an accuracy of 69.44%. Regarding the method of ATL [253], the accuracy of 77.52% falls behind the results of both our proposed method and our baseline, using $\sim 60\times$ more trainable parameters than our proposed method. Our results show that a simple ensemble architecture trained with a curriculum learning scheme and an auxiliary loss can achieve high cross-subject generalization, without any adaptation on test data or complex model architecture.

Regarding the reported results of our proposed method in Table 3.2 and Table 3.3, we note that the optimal number of first stage feature extractors K is inferred from the accuracy on the validation set. We provide more details about the impact of K on the performance of our architecture in Subsection 3.3.7 (Table 3.6 and Table 3.7). Similarly, regarding the reported results of the EEGNet-Ensemble and EEGNet-Bagging methods in Table 3.2 and Table 3.3, the optimal number of individual EEGNet models within an ensemble is chosen based on the validation accuracy.

Table 3.2: Performance of various methods on Physionet dataset, under 5-fold CV evaluation settings. The best accuracy is highlighted with bold.

Method	**Parameters**	**Accuracy (%)**
EEGNet-Single	2.5K	82.09
EEGNet-Ensemble, 8 models	20.0K	84.56
EEGNet-Bagging, 8 models	20.0K	84.81
EEGSym [176]	147.8K	83.91
TIDNet [126]	694.2K	82.19
Ours, K=**7**	15.7K	**86.36**

Table 3.3: Performance of various methods on OpenBMI dataset, under 5-fold CV evaluation settings. The best accuracy is highlighted with bold.

Method	**Parameters**	**Accuracy (%)**
EEGNet-Single	1.8K	78.31
EEGNet-Ensemble, 8 models	14.3K	78.98
EEGNet-Bagging, 7 models	12.6K	79.28
MIN2Net [12]	37.1K	69.44
ATL [253]	278.8K	77.52
Ours, K=**3**	4.6K	**79.73**

3.3.6 Results (LOSO)

In this experiment, we compare our method against other state-of-the-art works that report LOSO results on Physionet and OpenBMI. We note that we mention the results of these methods as reported in their original works, ensuring that they do not utilise labelled data from the test subjects.

The results on Physionet dataset are shown in Table 3.4. The method of EEGSym achieves state-of-the-art performance reaching an accuracy of 88.56%. EEGSym performs transfer learning by pretraining on four external datasets, which proves to be highly valuable. Our proposed method is the best performing model among the works that do not train on external data. We outperform the method of [21] that trains separate convolutional layers for each training subject. This further validates the existence of more efficient and accurate alternatives to complex deep architectures and the incorporation of subject-specific components.

The results on OpenBMI dataset are shown in Table 3.5. Our method presents state-of-the-art performance, scoring an accuracy of 85.07% when using all 62 electrodes of OpenBMI and having K = 4 first stage models. We outperform all other techniques, including the method of EEGSym that employs pretraining on external data. The geometric deep learning technique of TSMNet [121] and the algorithm of [130] that trains a convolutional architecture on spectral-spatial inputs, present an accuracy gap of more than $\sim$ 10% from the methods of ATL, EEGSym and our technique. This indicates that deep architectures operating on covariance matrices of EEG time-series (e.g. [121] and [130]), are generally less suitable for cross-subject MI decoding, compared to architectures that operate on raw EEG time-series.

3.3.7 Ablation studies

In our ablation studies we conduct experiments to observe the performance of both our proposed architecture and the two model ensembling methods against which we

Table 3.4: Comparison with other state-of-the-art methods on Physionet dataset with LOSO evaluation settings. The best accuracy is highlighted with bold.
*: pretrained on external data

Method	Parameters	Accuracy (%)
OPS [243] - Human Neurosc. 2020	N/A	67.00
Causal Viewpoint [21] - ICLRW 2022	N/A	83.90
EEGSym* [176] - TNSRE 2022	147K	**88.56**
Ours, K=**7**	15.7K	85.82

Table 3.5: Comparison with other state-of-the-art methods on OpenBMI dataset with LOSO evaluation settings. The best accuracy is highlighted with bold.
*: pretrained on external data

Method	Parameters	Accuracy (%)
MIN2Net [12] - TBME 2022	37.1K	72.03
Mutual Inf. [103] - TNNLS 2021	N/A	73.32
Spectral-Spatial [130] - TNNLS 2019	77M	74.15
TSMNet [121] - NeurIPS 2022	4.5K	74.60
ATL [253] - Neural Networks 2021	305K	84.19
EEGSym* [176] - TNSRE 2022	147K	84.72
Ours, K=**4**	8.7K	**85.07**

compare.

Proposed architecture

We investigate the impact of three factors on the performance of our ensemble architecture. The first factor is the number of first stage models K in the architecture. The second component is the loss $\mathcal{L}_{\text{subj}}^{\text{total}}$, that materializes our curriculum learning scheme. The third component is the distillation loss $\mathcal{L}_{\text{distill}}^{\text{total}}$ that enables collaborative training. We concurrently explore the effects of all these component choices, performing a sweep over the hyperparameter K and trying combinations of our loss terms.

Our first set of experiments (denoted as "$\mathcal{L}_{\text{CE}}$") corresponds to the scenario of training a model ensemble architecture (as described in Section 3.2.1), i.e. without curriculum learning and without our distillation loss. In our second set of experiments (denoted as "$\mathcal{L}_{\text{subj}}$") we train our architecture with ensemble curriculum learning, as described in Section 3.2.2, i.e. without our distillation loss. In the third experimental run (denoted

as "$\mathcal{L}_{\text{total}}$") we apply our entire method (i.e. using both subject-weighted loss and intra-ensemble distillation loss), training our architecture with the loss of Eq. 3.14. All experiments are performed with a 5-fold cross-validation setting.

The results of our ablation study on Physionet dataset are shown in Table 3.6. We observe a general trend of increasing accuracy for all our experimental sets, as K increases up to the value of 7 (further increasing K does not yield performance improvements). The only exception is the case where we train our architecture without curriculum learning (i.e. first row in Table 3.6), where the accuracy saturates at K = 6. This indicates that training multiple feature extractors by equally fitting them to the entire training set, is a suboptimal approach of training on multiple source domains. Thus, applying our curriculum learning scheme through $\mathcal{L}_{\text{subj}}$ to induce diversity in the feature extractors, is a straightforward step. The results of the second row in Table 3.6 verify the positive impact of curriculum learning in our ensemble architecture. In some cases (i.e. when K = 3, K = 4 and K = 6) curriculum learning provides negligible accuracy gains. When further incorporating our distillation loss in the total optimization objective of our architecture (i.e. third row in Table 3.6), we get additional accuracy boosts, except for the cases of K = 2 and K = 4. The beneficial effect of regulating the balance between feature diversity and model generalization through our distillation loss, is higher in the cases of K = 6 and K = 7 where the accuracy boosts are +0.42% and +0.68% respectively. This finding is particularly interesting, showing that the combination of our two loss terms can increase the performance of model ensembles, even when using many feature extraction models. On the contrary, an ensemble architecture trained solely with the standard cross-entropy loss, is more prone to performance saturation.

The results of our ablation study on OpenBMI dataset are shown in Table 3.7. The standard ensemble architecture trained without curriculum learning (i.e. first row in Table 3.7) achieves a maximum accuracy of 79.24% when K = 5. By using our

Table 3.6: Ablation study on Physionet dataset with 5-fold CV evaluation settings. Rows correspond to experiment sets done with different optimization objectives. Columns correspond to the number of first stage models (K) in our architecture. The best accuracy of each row is highlighted with bold.

Loss terms	**Accuracy (%)**					
	K=2	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**
$\mathcal{L}_{\text{CE}}$	83.34	84.70	84.97	84.93	**85.53**	85.34
$\mathcal{L}_{\text{subj}}$	84.38	84.72	85.10	85.40	85.62	**85.68**
$\mathcal{L}_{\text{total}}$	83.76	84.78	85.02	85.48	86.04	**86.36**

Table 3.7: Ablation study on OpenBMI dataset with 5-fold CV evaluation settings. Rows correspond to experiment sets done with different optimization objectives. Columns correspond to the number of first stage models (K) in our architecture. The best accuracy of each row is highlighted with bold.

Loss terms	**Accuracy (%)**					
	K=2	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**
$\mathcal{L}_{\text{CE}}$	79.15	79.08	78.96	**79.24**	78.94	79.20
$\mathcal{L}_{\text{subj}}$	79.02	**79.58**	79.13	79.15	79.01	79.31
$\mathcal{L}_{\text{total}}$	79.25	**79.73**	79.53	79.46	79.10	79.66

curriculum learning scheme, we improve the accuracy of our architecture in four out of six cases, achieving a maximum accuracy of 79.58% when K = 3. The incorporation of our distillation loss term in the total loss of our architecture (i.e. third row in Table 3.7) provides consistent improvements in all cases. Our best model has an accuracy of 79.73% when K = 3, with a boost of 0.65% over its corresponding standard ensemble model.

Model ensembling methods

In the experiments of Section 3.3.5 we compared our proposed method against our single model baseline as well as two model ensembling methods, under a 5-fold cross-validation scenario. We note that these experiments are performed using exactly the same train, validation and test splits. Here we provide additional results, presenting the performance of the EEGNet-Ensemble and EEGNet-Bagging methods as the number M of individual EEGNet models varies.

EEGNet-Ensemble: In Table 3.8 we show the cross-subject performance of EEGNet-

Ensemble on Physionet dataset under a 5-fold cross-validation scenario, when using from 2 to 9 EEGNet models. The best accuracy (84.56%) is achieved when fusing the predictions from 8 EEGNet models. In Table 3.9 we show the cross-subject performance of EEGNet-Ensemble on OpenBMI dataset under a 5-fold cross-validation scenario, when using from 2 to 9 EEGNet models. The best accuracy (78.98%) is achieved when fusing the predictions from 8 EEGNet models.

Table 3.8: Performance of EEGNet-Ensemble on Physionet dataset with 5-fold cross-validation evaluation settings. Columns correspond to the number of individual EEGNet models (M) that we use. The best accuracy is highlighted with bold.

Accuracy (%)							
M=2	**M=3**	**M=4**	**M=5**	**M=6**	**M=7**	**M=8**	**M=9**
82.99	84.01	84.10	84.21	84.04	84.26	**84.56**	84.47

Table 3.9: Performance of EEGNet-Ensemble on OpenBMI dataset with 5-fold cross-validation evaluation settings. Columns correspond to the number of individual EEGNet models (M) that we use. The best accuracy is highlighted with bold.

Accuracy (%)							
M=2	**M=3**	**M=4**	**M=5**	**M=6**	**M=7**	**M=8**	**M=9**
78.95	78.91	78.86	78.93	78.91	78.95	**78.98**	78.97

EEGNet-Bagging: In Table 3.10 we show the cross-subject performance of EEGNet-Bagging on Physionet dataset under a 5-fold cross-validation scenario, when using from 2 to 9 EEGNet models. The best accuracy (84.81%) is achieved when using the predictions from 8 EEGNet models. In Table 3.11 we show the cross-subject performance of EEGNet-Bagging on OpenBMI dataset under a 5-fold cross-validation scenario, when using from 2 to 9 EEGNet models. The best accuracy (79.28%) is achieved when using the predictions from 7 EEGNet models.

Table 3.10: Performance of EEGNet-Bagging on Physionet dataset with 5-fold cross-validation evaluation settings. Columns correspond to the number of individual EEGNet models (M) that we use. The best accuracy is highlighted with bold.

Accuracy (%)							
M=2	**M=3**	**M=4**	**M=5**	**M=6**	**M=7**	**M=8**	**M=9**
83.05	84.22	84.17	84.49	84.68	84.73	**84.81**	84.74

Table 3.11: Performance of EEGNet-Bagging on OpenBMI dataset with 5-fold cross-validation evaluation settings. Columns correspond to the number of individual EEGNet models (M) that we use. The best accuracy is highlighted with bold.

Accuracy (%)							
M=2	**M=3**	**M=4**	**M=5**	**M=6**	**M=7**	**M=8**	**M=9**
78.99	79.00	79.18	79.26	79.20	**79.28**	79.07	79.11

3.4 Conclusions

In this chapter we propose a method for cross-subject motor imagery decoding that leverages the combined strengths of model ensembling, curriculum learning and collaborative training. We design an ensemble architecture that is trained end-to-end in a single phase. We show that our curriculum training scheme can induce diversity to the feature extraction models of our architecture, improving its performance over standard ensembling. Our method also benefits from the exchange of knowledge between the models of our ensemble, that occurs through our auxiliary distillation loss. We conduct experiments on the datasets of Physionet [79] and OpenBMI [135], totalling more than 150 participants, and demonstrate state-of-the-art results. Our proposed method outperforms other approaches that try to tackle MI decoding using complex networks [253, 126], multi-task learning [12], geometric deep learning [121], subject-specific layers [21] or pretraining on multiple external datasets [176].

Chapter 4

Covariance Mixing Regularization for Motor Imagery Decoding

Contents

4.1 Introduction

In the previous chapter, we proposed an ensemble learning method for cross-subject motor imagery decoding that belongs in the broad family of domain generalization techniques. In this chapter, we further study the problem of motor imagery decoding and explore a combination of two other approaches for domain generalization, namely domain-invariant representation learning and data augmentation [229].

Approaches that enable domain-invariant representation learning have led to remarkable improvements in the field of EEG-based BCIs, with the most common such

technique being data alignment using covariance statistics [89]. Initially, data alignment was intended to be applied on covariance matrices, to improve the performance of Riemannian classification frameworks (e.g. MDRM classifiers [19]). However, recently data alignment has also been directly applied on EEG time-series [126], improving the performance of CNN-based classification models. As explained in [126], data alignment can be seen as a way of learning domain-invariant representations. This is supported by the fact that the aligned EEG signals of each subject have close to identity covariance matrix, enabling better transfer learning across subjects.

Alignment is typically performed using covariance statistics that are session-wise, i.e. computed based on an entire set of EEG trials that may have a duration of several minutes. Specifically, in both training and testing phases, each trial is transformed using its corresponding session-wise covariance matrix. A basic property of EEG signals is their non-stationarity [248], hence their statistics vary across time. That is, the covariance statistics within shorter time windows are different from the session-wise statistics. Exploiting information from trial-wise covariance statistics within alignment approaches has remained unexplored, thus we focus on this issue and propose adapting the standard alignment process to take into account trial-wise statistics. A straightforward way of doing this adaptation, is to perform interpolation between the session-wise and trial-wise covariance matrices. Due to the nature of covariance matrices (i.e. lying on the Riemannian manifold of SPD matrices), this is done through Riemannian interpolation, avoiding treating them as having a Euclidean nature.

Building on the benefits of CNN-based learning and covariance-based alignment, we propose a method called "CovMix", that mixes session-wise and trial-wise covariances to concurrently perform data alignment and augmentation on EEG time-series. Instead of the standard alignment during training, we suggest aligning each trial choosing *randomly* an SPD matrix that lies on the geodesic between the session-wise and the trial-wise covariance matrices. Afterwards, the aligned trials are fed to a CNN model

that is trained to classify them. Along the training process, the CNN model will receive each trial as input multiple times, yet aligned using different SPD matrices. However, all the augmented versions of each original trial, will be lying on the same geodesic. Effectively, this regularizes the CNN model by feeding it with various transformations of each training sample. Inference is performed using the standard alignment, to keep the process being deterministic.

Our contributions are summarized as follows:

- We propose CovMix, a method that mixes session-wise and trial-wise covariance matrices to jointly perform EEG signal alignment and data augmentation during training.
- CovMix is performed before feeding the data to the classification network, thus it can be used in any method that employs CNNs for motor imagery decoding.
- We evaluate networks trained on BCI Competition IV-2a dataset [215] with cross-subject settings, showing that adding CovMix acts as regularization to the classification network, yielding stronger generalization results compared to the standard covariance-based alignment and other techniques.

The rest of this chapter is organized as follows. In Section 4.2 we briefly refer to the standard alignment technique and in Section 4.3 we describe our proposed approach. In Section 4.4 we provide details about our evaluation setup and present results of our method on cross-subject experiments, comparing it with other techniques. Finally, in Section 4.5 we conclude the chapter.

4.2 Preliminaries

Before delving into the description of our proposed method, we first provide some general details on the concept of covariance-based time-series alignment. Let $\mathbf{S_X} =$

$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be the set of n band-pass filtered EEG trials that a recording session contains. Let $\mathbf{X}_i \in \mathbb{R}^{C \times T}$ be the i-th EEG trial of the session, where C is the number of EEG channels and T is the number of samples in the dimension of time. The covariance matrix $\mathbf{P}_i$ of trial $\mathbf{X}_i$ is calculated as $\mathbf{P}_i = \mathbf{X}_i \mathbf{X}_i^T$. The set of all trial-wise covariance matrices of the session is $\mathbf{S_P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$.

To compute the session-wise SPD matrix, i.e. the Riemannian mean of all trial-wise covariance matrices in the set $\mathbf{S_P}$, we need to define the concept of Riemannian distance. Covariance matrices lie on the space of SPD matrices with dimension C, denoted as $\mathcal{P}(C)$, which is a Riemannian manifold. Considering two points (i.e. two SPD matrices) $\mathbf{P}_1$ and $\mathbf{P}_2$ on $\mathcal{P}(C)$, the Riemannian distance metric [26] is defined in Eq. 4.1:

$$\delta(\mathbf{P}_1, \mathbf{P}_2) = \Big(\sum_{i=1}^{n} \log^2 \lambda_i \Big)^{\frac{1}{2}}, \tag{4.1}$$

where λ_i are the eigenvalues of $\mathbf{P}_1^{-1} \mathbf{P}_2$.

The session-wise covariance matrix is computed as the Riemannian mean $\overline{\mathbf{P}}$ of the set $\{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$ of trial-wise covariance matrices. There is no closed-form solution for the computation of the SPD matrix $\overline{\mathbf{P}}$, thus it is solved as an optimization problem, i.e. minimizing the Riemannian distance between $\overline{\mathbf{P}}$ and all the trial-wise covariances, according to Eq. 4.2:

$$\overline{\mathbf{P}} = \underset{\mathbf{P} \in \mathcal{P}(C)}{\operatorname{argmin}} \sum_{i=1}^{n} \delta(\mathbf{P}_i, \mathbf{P}) \tag{4.2}$$

Riemannian Alignment: Typically, alignment [89, 243] on EEG signals is performed separately within each session, applying the same session-specific transformation to all trials. Considering the session-wise SPD matrix $\overline{\mathbf{P}}$ and the trial-wise signals $\mathbf{X}_i$, the aligned signals $\hat{\mathbf{X}}_i$ in this method are computed as:

$$\hat{\mathbf{X}}_i = \overline{\mathbf{P}}^{-\frac{1}{2}} \mathbf{X}_i \tag{4.3}$$

4.3 Proposed method

CovMix Alignment: We propose an alternative method of alignment, that transforms each trial differently over multiple training steps. We use various transformation matrices to align each trial, that are obtained by traversing the geodesic connecting the reference state (i.e. the session-wise SPD matrix $\overline{\mathbf{P}}$) and the trial-wise covariance matrix $\mathbf{P}_{\mathrm{i}}$ (instead of deriving it directly from $\overline{\mathbf{P}}$). Having the reference state as starting point, ensures that the finally computed SPD matrices are still relevant to the session-wise statistics. The computation of the transformation matrices by traversing geodesics, ensures that our method respects the Riemannian nature of covariance matrix space.

We mix (i.e. interpolate) the session-wise SPD matrix $\overline{\mathbf{P}}$ with that of the i-th trial $\mathbf{P}_{\mathrm{i}}$, obtaining the mixed SPD matrix $\mathbf{P}_{\mathrm{mix}}$. An illustration of performing interpolation on the Riemannian manifold is provided in Fig. 4.1. We use a scalar $\alpha, 0 \leq \alpha \leq 1$, which we call covariance mixing coefficient, to control the distance between $\mathbf{P}_{\mathrm{mix}}$ and $\overline{\mathbf{P}}$. More specifically, we sample α from a uniform distribution $\mathcal{U} \sim [0, 1]$, and compute the weighted Riemannian average [26] between matrices $\overline{\mathbf{P}}$ and $\mathbf{P}_{\mathrm{i}}$ as shown in Eq. 4.4, so that $\delta(\mathbf{P}_{\mathrm{mix}}, \overline{\mathbf{P}}) = \alpha \cdot \delta(\mathbf{P}_{\mathrm{i}}, \overline{\mathbf{P}})$.

$$\mathbf{P}_{\mathrm{mix}} = \overline{\mathbf{P}}^{\frac{1}{2}} \left(\overline{\mathbf{P}}^{-\frac{1}{2}} \mathbf{P}_{\mathrm{i}} \overline{\mathbf{P}}^{-\frac{1}{2}} \right)^{\alpha} \overline{\mathbf{P}}^{\frac{1}{2}} \tag{4.4}$$

We can control the regularization induced to the classification network by CovMix, by restricting the range of values that are sampled for α. To do so, we use the hyperparameter α_{max} to set the maximum value of α, and sample from the distribution $\mathcal{U} \sim [0, \alpha_{\mathrm{max}}]$. CovMix is applied *only* during the training phase, similarly to data augmentation methods, aligning the signals as follows:

$$\hat{\mathbf{X}}_{\mathrm{i}} = \mathbf{P}_{\mathrm{mix}}^{-\frac{1}{2}} \mathbf{X}_{\mathrm{i}} \tag{4.5}$$

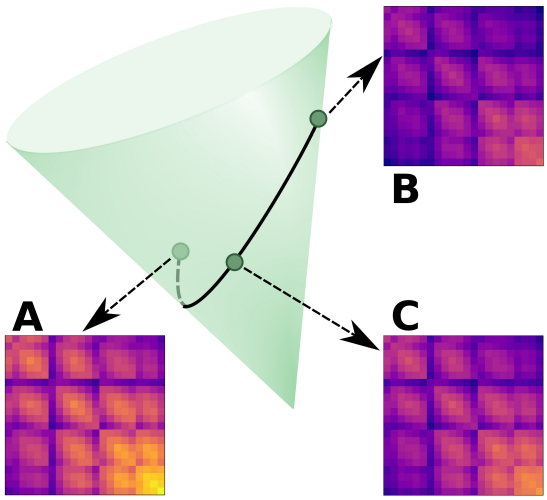


Figure 4.1: Mixing two covariance matrices, by traversing their geodesic on the Riemannian manifold. The point corresponding to matrix **C**, lies on the shortest path that connects **A** and **B**.

During the inference phase, we apply the Riemannian alignment and transform the signals using Eq. 4.3.

Let us note that our transformation matrix is not an arbitrary deep learned matrix, but an SPD matrix obtained by traversing on particular geodesics of the Riemannian manifold, that connect session-wise and trial-wise covariances. This ensures that the transformation matrix is *close* to the reference matrix for each domain, to facilitate training on multiple source domains as in [126]. Moreover, our transformation matrix does not aim to suppress noise on EEG trials, as [17] does. Having done the covariance mixing in the Riemannian space, we perform data augmentation in the Euclidean space of EEG time-series. This allows us to use classifiers such as CNN models, unlike the method of [109] that generates augmented data points that require Riemannian classifiers.

4.3.1 CNN architecture

We opt to use a modified version of EEGNet [133] as our CNN architecture for motor imagery classification. Specifically, we remove the batch normalization layer at the

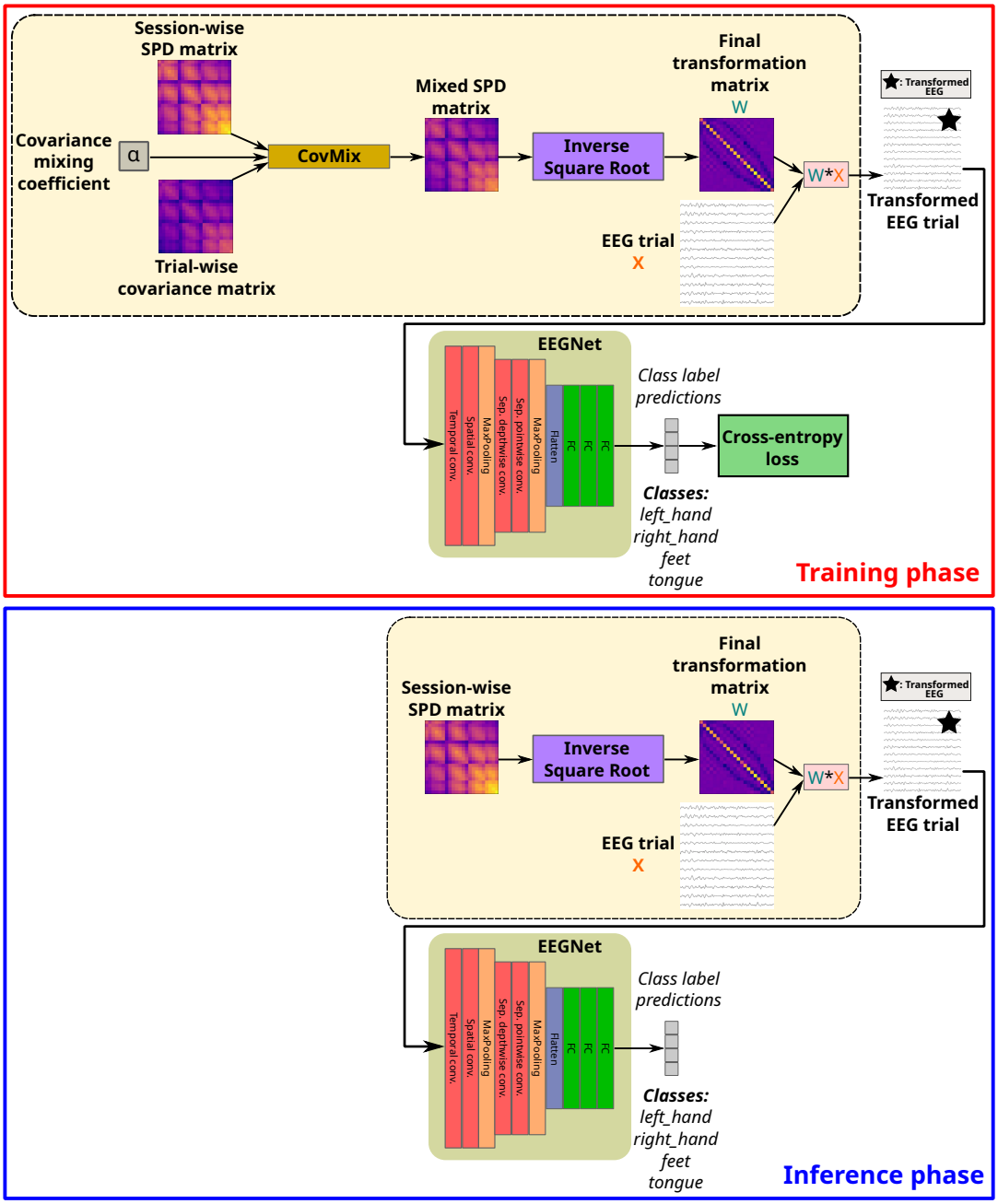


Figure 4.2: Overview of our proposed method. CovMix mixes session-wise and trial-wise covariance statistics following Eq. 4.4 and performs alignment by multiplying the EEG signals with the inverse square root of the mixed matrix using Eq. 4.5. In the inference phase we do not mix covariance statistics and alignment is performed using Eq. 4.3. Finally, the transformed EEG signals are fed to EEGNet to be classified.

temporal convolution stage, and use 3 fully connected (FC) layers at the classification head.

The pipeline of performing CovMix, along with the components of the CNN architecture, are shown in Fig. 4.2.

4.4 Experimental results

Dataset: Our EEG preprocessing pipeline has the following steps: 1) bringing the EEG signals into the measurement unit of uV (microvolts) 2) keeping only the channels of 22 EEG electrodes, discarding the EOG electrodes 3) notch filtering to remove the 50Hz component 4) bandpass filtering in the range 4-38 Hz and 5) resampling signals to 100Hz. We crop the temporal window $[0.0, 4.0]$ seconds of each trial, where $t = 0$ is the event onset. The size of each input sample is C $\times$ T, where C=22 (number of EEG channels) and T=400 (number of time samples). Evaluation is performed in a Leave-One-Subject-Out (LOSO) manner, using both sessions for all subjects.

Comparison to other methods: We compare CovMix with three other methods, namely Riemannian Alignment (RA) [89], MixUp [252] and MixStyle [257]. We implement RA as in [243], using the Riemannian mean of covariances and transforming the EEG signals instead of re-centering the covariances. For MixUp, considering two data samples $\mathbf{x}_i$, $\mathbf{x}_j$ and their labels encoded as one-hot vectors $\mathbf{y}_i$, $\mathbf{y}_j$ we create the augmented samples as $\mathbf{x}' = \lambda\mathbf{x}_i + (1-\lambda)\mathbf{x}_j$ and $\mathbf{y}' = \lambda\mathbf{y}_i + (1-\lambda)\mathbf{y}_j$, where $\lambda \sim \text{Beta}(2.0, 2.0)$. We also evaluate the method of MixStyle, which is a state-of-the-art domain generalization technique that can be plugged in between CNN layers. For MixStyle, let $\mathbf{x}$ be a batch of samples, and $\tilde{\mathbf{x}}$ be the randomly shuffled version of $\mathbf{x}$ across the batch dimension. First, the feature statistics $\gamma_{\text{mix}} = \lambda\sigma(\mathbf{x}) + (1-\lambda)\sigma(\mathbf{x})$ and $\beta_{\text{mix}} = \lambda\mu(\mathbf{x}) + (1-\lambda)\mu(\mathbf{x})$ are computed, using the operators of $\mu(\cdot)$ (mean value) and $\sigma(\cdot)$ (standard deviation) along the temporal dimension, with $\lambda \sim \text{Beta}(0.1, 0.1)$. Then, MixStyle is performed with a probability of 50% on training batches, computing

$\mathbf{x}' = \gamma_{\text{mix}} \frac{\mathbf{x}-\mu(\mathbf{x})}{\sigma(\mathbf{x})} + \beta_{\text{mix}}$ and detaching the operators of $\mu(\cdot)$ and $\sigma(\cdot)$ from the gradient computation. Upon attempting to plug MixStyle in several convolutional stages of EEGNet, we find that using it after the third convolutional layer is the most effective choice, and report results with this setting. We also report results of a baseline EEGNet model trained without any signal aligning (mentioned as "Baseline"), to serve as a reference for evaluating the benefits of alignment.

Training hyperparameters: Batch size is set to 64 and training is conducted for 120 epochs. The cross-entropy loss is used as the criterion for MI classification, where the targets are 4 classes. Stochastic Gradient Descent is selected as the optimizer (momentum=0.9, weight decay=0.01). We set the dropout probability of EEGNet to 0.1.

Results: Table 4.1 shows the results of all methods on the IV-2a dataset. Compared to the baseline model that is trained without any alignment on the EEG signals, the Riemannian Alignment (RA) method provides an accuracy boost of +4.48% (from 53.45% to 57.93%). Training EEGNet with CovMix using the default setting (i.e. $\alpha_{\max} = 1.0$, denoted as "CovMix" in Table 4.1), further increases the performance by +3%. The regularization method of MixUp improves the accuracy only by +0.70% compared to RA, while the domain generalization approach of MixStyle gives a more considerable increase of +2.83% over RA. Overall, CovMix yields the highest accuracy of 60.93% outperforming all other methods.

Considering the regularization effect of MixUp, we find it to be insufficient. One drawback of applying MixUp in EEG signals, is the existence of large inter-subject differences on the channel-wise statistics. Thus, directly mixing signals from different domains on the input space (i.e. the time-series) may be detrimental for multi-source training. An indirect solution to this issue, is to scale the signal values of each trial in the range $[-1, +1]$ as in [126], before applying MixUp. Nevertheless this scaling is not consistent within each session, as it depends on trial statistics. In contrast to MixUp

Table 4.1: Evaluation of several methods on the motor imagery classification problem, using the dataset of BCI Competition IV-2a. The dataset contains 9 participants, and columns S01-S09 correspond to the accuracy of LOSO evaluation on one participant each time. All the numbers reported in this table are the average of 3 runs.

Method	**S01**	**S02**	**S03**	**S04**	**S05**	**S06**	**S07**	**S08**	**S09**	**Avg.**
Baseline	68.98	38.02	71.47	45.02	36.34	38.54	45.19	67.99	69.50	53.45
RA	75.80	35.36	81.25	49.13	45.08	44.09	49.24	70.08	71.35	57.93
MixUp	75.63	34.14	79.11	48.67	46.70	46.81	52.54	74.24	69.79	58.63
MixStyle	76.09	42.36	80.20	55.84	45.71	47.97	58.39	67.18	73.09	60.76
CovMix	76.21	41.72	82.46	54.40	44.73	47.16	57.17	73.20	71.35	60.93
CovMix*	78.12	43.86	83.68	58.33	45.54	47.51	60.53	75.34	72.91	**62.87**

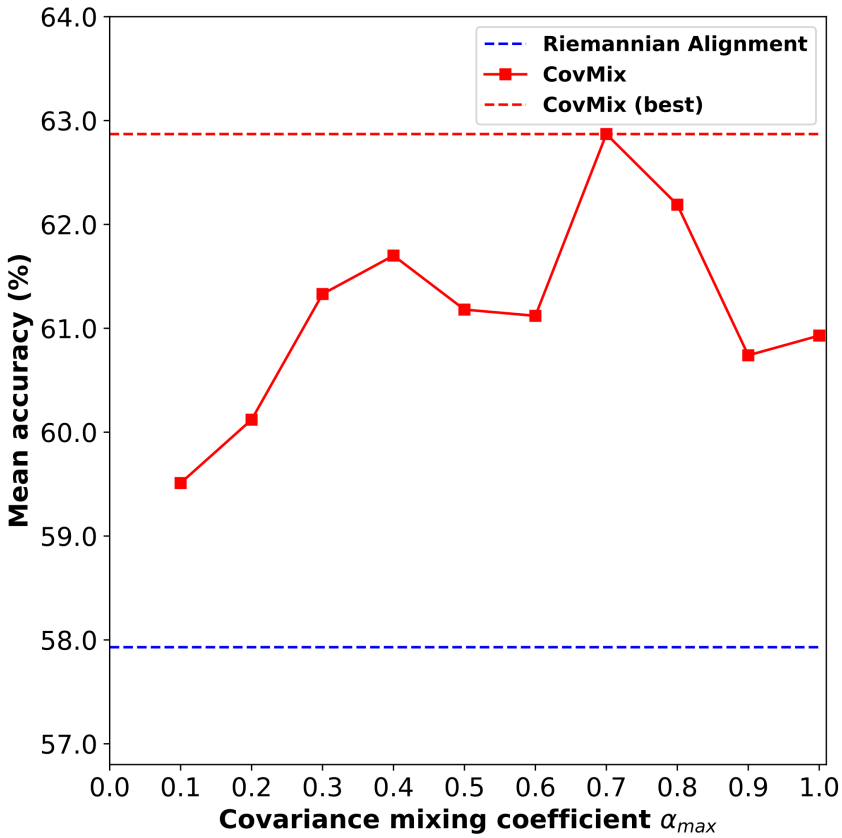


Figure 4.3: The amount of regularization induced to the network by CovMix, is controlled through the hyperparameter $\alpha_{\max}$. The achieved performance obtained using CovMix, improves as we increase $\alpha_{\max}$ from 0.1 up to 0.7.

that transforms batch samples on the input space, MixStyle is done on the space of intermediate layer features. Thus, it is less prone to cross-domain differences on the

input space. Yet for MixStyle to be effective, it needs to be plugged in to early layers of CNNs, where features mostly reflect domain-related information (while late layers are expected to be increasingly related to class label information). Differently from MixUp and MixStyle, CovMix does not involve mixing information stemming from different domains. The results show that regularization can be effectively induced in an intra-domain manner.

Ablation study: To examine the impact of hyperparameter $\alpha_{\max}$ on the performance of CovMix, we do an ablation study and run experiments setting $\alpha_{\max}$ from 0.1 to 1.0 with a step of 0.1. Smaller values of $\alpha_{\max}$ induce smaller regularization to the network. The results of our sweep are shown in Fig. 4.3. We observe that the performance of CovMix does not fall below that of RA, for any value of $\alpha_{\max}$. Tuning $\alpha_{\max}$ leads to even better performance compared to the default setting of $\alpha_{\max} = 1.0$, reaching a maximum accuracy of 62.87% when $\alpha_{\max} = 0.7$ (denoted as "CovMix*" in Table 4.1). The test subjects benefit differently from the values of $\alpha_{\max}$. In six out of nine subjects (specifically subjects 1, 3, 5, 7, 8 and 9), we achieve the highest test accuracies when setting $\alpha_{\max}$ between 0.6 and 0.8. However, for the rest three subjects (2, 4 and 6) the highest test accuracies occur when $\alpha_{\max}$ is in the range of 0.1 to 0.3.

Visualization of augmented SPD matrices: In Fig. 4.4, we provide a t-SNE [218] visualization of the covariances corresponding to trials, and the SPD matrices generated using CovMix with randomly sampled values of α. We can see that the points corresponding to interpolated SPD matrices mainly occupy the space between the barycenter of the entire session (i.e. all trials from all classes) and the points of trial-wise covariance matrices.

4.5 Conclusions

In this chapter we discuss the problem of EEG-based MI decoding in transfer learning scenarios. Alternatively to methods that extract handcrafted features from EEG

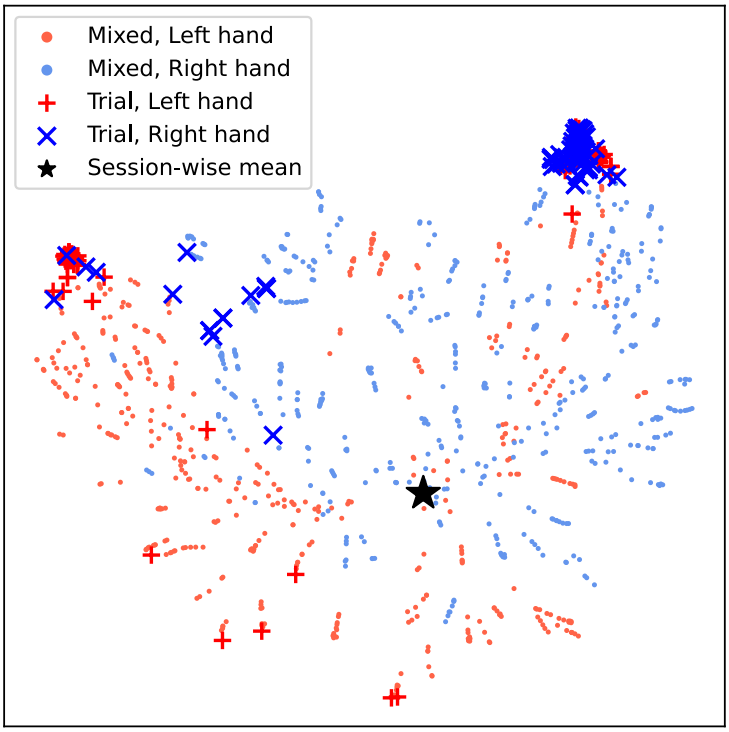


Figure 4.4: Visualization of t-SNE embeddings from the trial-wise covariance matrices and the mixed SPD matrices that were obtained by performing random interpolations with CovMix. We use EEG signals from the second session of subject 9. Notice also the Riemannian barycenter of *all* trials (plotted with marker "★").

signal time-series, we use a CNN model as feature extractor. Through our proposed method, we concurrently perform alignment on the EEG signals and regularization on the CNN, applying different signal transformations during the training phase. We use a Riemannian framework to derive the transformation matrices, mixing trial-level and session-level covariance statistics. We conduct experiments on BCI IV-2a dataset for MI classification, showing that CovMix performs superiorly against the traditional Riemannian Alignment, the regularization method of MixUp and the domain generalization method of MixStyle. Our results indicate the potential of leveraging covariance-based alignment as a means towards regularization of deep neural networks.

Chapter 5

Pairwise Ranking Network for Affect Recognition

Contents

5.1 Introduction

In the previous two chapters we explored the topic of EEG-based motor imagery decoding, where the groundtruth labels are given as discrete classes, representing imagined movements of human limbs. Two uncontrolled factors affecting the correspondence between the inputs (i.e. the EEG data) and the targets (i.e. the groundtruth class) when training a DNN on motor imagery decoding, are inter-subject variability and the so-called "BCI illiteracy" phenomenon, that we briefly mentioned in Section 1.5. That is, the inability of subjects to operate an MI-based BCI system, as well as their individual characteristics, are reflected on the EEG data distribution of each subject, in the form of covariate shift [184].

In this chapter we study the topic of EEG-based affect recognition, where several stimuli are presented to each subject in order to be subjectively rated in terms of their affective content. In the majority of affect recognition works, the issue of "BCI illiteracy" is not a concern, as the task simply demands from the subjects to observe the audiovisual stimuli, without depending on the manifestation of particular oscillatory rhythms. This is a major difference between the motor imagery BCI paradigm and affect recognition studies. Unarguably, inter-subject variability not only remains an issue in EEG-based affective research, but also in many cases it impedes applying cross-subject and cross-dataset transfer learning approaches [183].

An additional factor that renders emotion recognition a challenging problem, is the presence of label subjectiveness. Each subject assigns affective ratings to multimedia content in a different way, that depends on several things, including personal biases (e.g. music "taste"), personality traits [112], current mood, familiarity with the stimuli and affective content of previously encountered stimuli, among others. This subjectiveness is rarely taken into consideration when designing machine learning approaches for affect recognition. One common option to reduce the impact of label subjectiveness, is to train subject-specific models, thus overcoming the need to account for inter-subject differences in the annotation of affective experiences. However, even in the case of intra-subject affect recognition models, problems related to the subjectiveness of emotion annotations persist.

During emotion data labelling, typically humans assign a value in a continuous range, for each emotional experience. These values are assumed to be on an absolute scale (i.e. as opposed to being on a relative scale), however even for a single annotator the perception of the rating scale may change across time [153]. Works inspired from the adaptation level theory of Helson [91], suggest that human judgments of presented stimuli are relative to the context [196], including previously encountered stimuli, rather than absolute. Therefore emotions can be expressed in relative terms, i.e. through

comparisons between different affective state levels. Labelling emotions by assigning relative values has been an alternative path to the traditional scheme of absolute labels [246, 147]. This means that annotating emotions involves comparison of the human affective states between past and forthcoming experiences. Therefore, one possible way of inferring such ordinal relations between affective states, is through machine learning models that can explicitly compare them.

We study the problem of affect recognition on datasets where annotations are provided in the form of sample-wise labels. Typically, plain regression or classification approaches are applied on such datasets. In the case of regression, the inherent biases of continuous affect annotations described above, are harmful for the training process thus also for model performance [247]. Other problems arise when adopting classification approaches as a remedy to the shortcomings of regression. Discrete classes cannot express the compoundness of emotions. Transforming ratings of ordinal nature into nominal classes results in information loss regarding the structure of ratings. Furthermore, the class splitting criteria defined by researchers, do not always accurately reflect the manifestations of affect [153]. Hence, a more suitable approach is preference learning [77], that involves comparing emotions. The superiority of preference learning methods over classification algorithms for affect recognition, has been previously studied by Melhart *et al.* [156]. We follow an alternative direction, investigating the utilization of preference learning as an auxiliary objective to improve the performance of deep neural networks on classification.

Despite the exciting results of deep learning methods on affective computing problems, the possibility of building deep networks that can compare samples corresponding to different affective states, has remained mostly unexplored. Refraining from using solely a sample-wise classification objective, we propose employing an additional pairwise objective, namely the emotional rating comparison. Considering a pair of data samples and their affective labels, the comparison task infers the ordinal

ranking relation between the labels of the samples (i.e. higher/similar/lower arousal, higher/similar/lower valence). We use a shared deep feature extractor along with separate network heads that infer the affective state level of each sample and perform pairwise ranking between samples. Our experiments show that the former task benefits from the latter, as treating the data in a pairwise manner enables better representation learning. The main contributions of our method are the following:

- We propose a deep architecture that is jointly trained on sample-wise classification and pairwise ordinal ranking.
- We conduct experiments on two EEG-based affect recognition datasets, showing performance gains from the incorporation of a ranking objective in the training process.

5.2 Proposed method

The main motivation of our work is to investigate meaningful combinations of classification and ordinal ranking through deep neural networks, in the field of affective computing. In contrast to typical network architectures that operate on affective data solely in a sample-wise manner, we aim to additionaly perform pairwise operations between samples, learning the ordinal relation between their corresponding affective ratings. Our goal is to boost the classification performance of emotion recognition models, leveraging the additional supervision of a ranking task only during training. Traditional preference learning systems such as RankNet [35] train a function that maintains a higher score for the preferred option. The preference decision is a fixed operation on sample-wise preference scores, without involving any trainable parameter. Our method differs from such systems, as it learns the ordinal relation through a trainable module. Considering that the emotion label space has an inherently ordinal structure, we avoid disregarding such knowledge, by further exploiting it through the ranking task. To achieve this, we utilise the provided affective state annotations to

form rank-based labels, and construct a deep architecture that can handle both the end-goal task of classification, as well as the additional task of pairwise ranking. In the following paragraphs, we explain various aspects of our method.

5.2.1 Methodology

Pairwise ordinal ranking: The proposed methodology that derives pairwise ranking labels is applicable on datasets having as annotations either *continuous* affective ratings or *categorical* labels of ordinal nature. Firstly, we explain its functionality on continuous labels that are defined on a bounded scale. Considering a pair of samples x_1 and x_2 (with corresponding affective rating labels y_1 and y_2), the goal of the ranking task is to infer the ordinal relation between the labels y_1 and y_2. In previous works this is addressed by establishing a preference of the sample with the higher rating over the other sample, i.e. $x_1 \succ x_2$ or $x_1 \prec x_2$. The symbols of "$\prec$"/"$\succ$" denote preceding/succeeding order of the samples with respect to their ratings y_1 and y_2, i.e. by using these symbols we do not imply a comparison on the raw feature values of x_1 and x_2. A minimum difference value between the compared ratings is often used to discard unclear comparisons. To avoid posing very strict constraints over pairs of ratings with small difference, we opt to add a third case of rank, namely the case $x_1 \sim x_2$, if x_1 and x_2 have similar ratings [173]. We define a hyperparameter $\epsilon > 0$, called "rank tolerance", such that $x_1 \sim x_2$ holds true when $|y_1 - y_2| \leq \epsilon$. Thus, $x_1 \succ x_2$ when x_1 has a higher rating than x_2 under the condition $y_1 > (y_2 + \epsilon)$, and $x_1 \prec x_2$ when $y_1 < (y_2 - \epsilon)$. The ordinal relations for continuous ratings are shown in Table 5.1 as well as in Fig. 5.1.

Relation	**Condition**
$x_1 \succ x_2$	$y_1 > (y_2 + \epsilon)$
$x_1 \sim x_2$	$\|y_1 - y_2\| \leq \epsilon$
$x_1 \prec x_2$	$y_1 < (y_2 - \epsilon)$

Table 5.1: List of ordinal ranking relations and their corresponding conditions, when performing a comparison operation over continuous ratings.

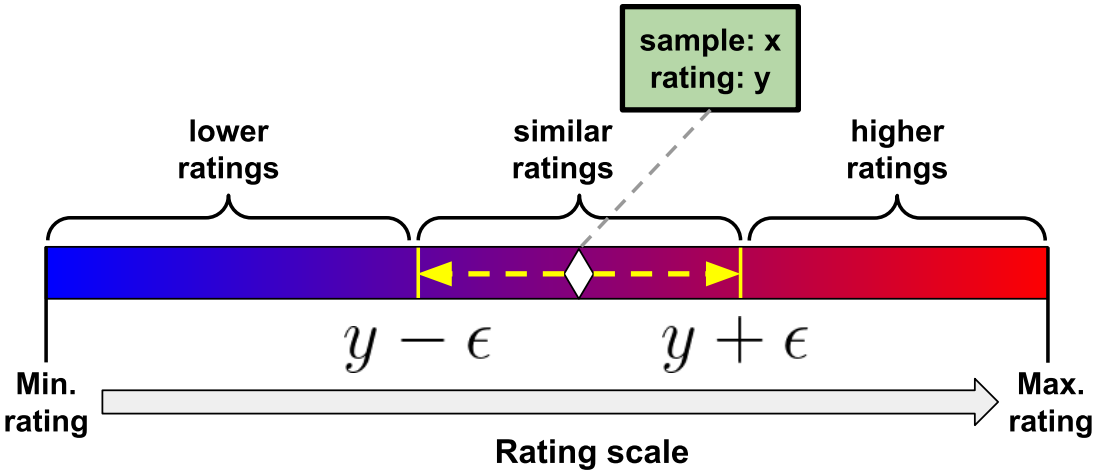


Figure 5.1: Illustration of the ordinal relations defined over a bounded continuous rating scale.

Joint training - combining classification and ranking: Our method is simple and can be integrated into existing affect recognition architectures. In essence, every deep neural network operating on an end-goal task of affect classification, consists of a backbone that extracts feature representations and a classification head that maps these representations to class scores. We suggest adding an extra supervisory signal by imposing a pairwise ranking objective on the intermediate representations learned by the backbone, leveraging the knowledge around the ordinal nature of emotions. The ranking task is performed by a ranking head that is stacked on top of the backbone network, receiving two feature representations and inferring their ordinal relation with respect to the affective ratings of their corresponding samples. The processing pipeline for classification remains intact and the total architecture is trained in an end-to-end manner. We fully backpropagate the gradients of both the classification loss and ranking loss to the backbone, updating its weights based on both loss terms. The backbone network benefits from the additional ranking supervision, extracting features that enable better generalisation on the end-goal task. Both the classification and ranking loss are computed using a cross-entropy criterion.

5.2.2 Network architecture

We aim to build an architecture that operates on affective data inputs to perform sample-wise classification of emotions, as well as a pairwise comparison operation (ordinal ranking) with respect to the emotional ratings for a pair of samples. Regarding the implementation of deep networks that accomodate pairwise operations, our work builds on the concept of Relation Networks [212] that have been used for few-shot image recognition. In the context of Relation Networks, a *relation* module refers to a mechanism that learns to compare feature embeddings for a pair of samples, to determine whether they have the same class label or not. We adapt the framework of Relation Networks to suit the purposes of pairwise ranking. We propose using a *ranking* module that learns to perform ordinal ranking on the feature embeddings of a pair of samples, by inferring the ordinal relation between their affective ratings. Note that the inputs of our ranking module are pairwise feature embeddings, formed by concatenating the sample-wise embeddings obtained from a backbone feature extractor, for each pair of samples. Our architecture, named Pairwise Ranking Network ("PRNet"), is shown in Fig. 5.2. A detailed explanation of its consisting modules is provided below.

Embedding module: The embedding module is the backbone of our architecture, serving as a feature extractor. The batch samples are fed as inputs to the embedding module and a feature embedding is computed for each sample. The produced embeddings are to be further processed for the tasks of classification and ranking by the corresponding modules.

Classification module: The classification module receives as input the features produced by the embedding module, and predicts the affective state for each sample. The groundtruth targets are discrete emotion classes (e.g. "low"/"high" arousal, "low"/"high" valence). We denote the classification predictions as $\hat{\mathbf{y}}_{\text{cls}}$ and the corresponding groundtruth values as $\mathbf{y}_{\text{cls}}$. Note that $\hat{\mathbf{y}}_{\text{cls}}$ contains probabilities obtained by passing the outputs of the classification module through a softmax layer, while $\mathbf{y}_{\text{cls}}$ contains one-hot encodings

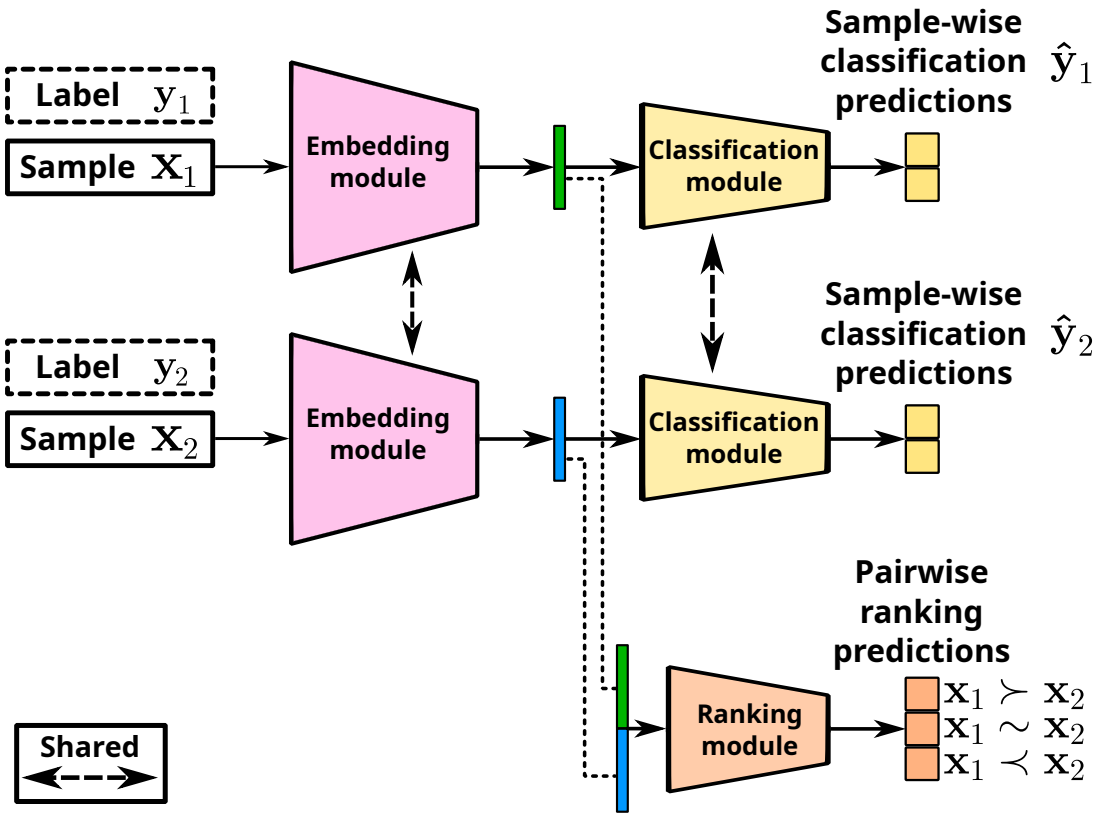


Figure 5.2: The architecture of a Pairwise Ranking Network that accomodates joint training on classification and ranking tasks.

of the labels and k_{cls} denotes the number of classes. The classification loss term $\mathcal{L}_{cls}$ is computed as follows:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{k_{cls}} y_{cls}^{i} \log(\hat{y}_{cls}^{i}). \tag{5.1}$$

Ranking module: The ranking module operates on pairwise feature representations that correspond to sample pairs, and infers their ordinal relation with respect to their affective ratings. To form the pairwise feature representation of two samples, we get the feature vectors extracted from the embedding module for both samples, and we concatenate them across the channel dimension. To form multiple pairs of sample embeddings during training with a batch size of N_b, we split each batch into two sub-batches of size $N_{sub} = \frac{N_b}{2}$. Every sample of each sub-batch is compared against all samples of the other sub-batch, yielding $(N_{sub})^2$ pairs in total. Denoting the softmaxed

ranking predictions and one-hot groundtruth values as $\hat{\mathbf{y}}_{\text{rank}}$ and $\mathbf{y}_{\text{rank}}$ respectively, the ranking loss term $\mathcal{L}_{\text{rank}}$ is defined as follows:

$$\mathcal{L}_{\text{rank}} = -\sum_{\text{i}=1}^{\text{k}_{\text{rank}}} \mathbf{y}_{\text{rank}}^{\text{i}} \log(\hat{\mathbf{y}}_{\text{rank}}^{\text{i}}), \tag{5.2}$$

where $\text{k}_{\text{rank}} = 3$ denotes the number of possible ordinal relations for a pair of samples.

The total loss that is used to optimize the Pairwise Ranking Network is the sum of the loss on the end-goal task and the ranking loss. We use a coefficient α to weight the contribution of the ranking loss to the total loss, i.e. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{rank}}$, setting the value of α equal to 1 when not stated otherwise.

Architecture details: In our proposed architecture, the embedding module consists of two fully-connected (FC) layers with 128 nodes each, receiving 100-dimensional feature vectors as inputs. The classification module consists of one FC layer for each of the targets (i.e, Arousal, Valence), with k_{cls} output nodes, where k_{cls} is the number of classes. The ranking module consists of one FC layer for each of the targets, having $\text{k}_{\text{rank}} = 3$ output nodes. The baseline model for our experiments is the composition of the embedding and classification modules, i.e. a simple model with a feature extractor and a classifier.

5.3 Experimental results

We apply our method on the problem of affect recognition, aiming to exploit the ordinality of emotions through our analysis. Specifically, we study the datasets of DEAP [122] and SEED [256] where the original affective annotations are inherently ordinal. A brief description of these datasets was provided in Chapter 2.6. Each dataset has been annotated through a different process and is evaluated on a different end-goal task. An overview of the datasets used in our study is shown in Table 5.2.

Dataset	Annotation process	Annotation values	End-goal task
DEAP [122]	-Self-assessment reports -Varying per participant	Arousal, Valence Range: [1.0, 9.0]	Classification: Low/High Arousal Low/High Valence
SEED [256]	-Determined from the study's authors -Fixed for all participants	3 discrete classes: Negative Neutral Positive	Classification: Negative Neutral Positive

Table 5.2: Details regarding the affective annotations and evaluation tasks on the datasets used in our work.

5.3.1 Dataset details

DEAP dataset: DEAP [122] is a dataset for EEG-based emotion recognition, having 32 participants and 40 music video clips as stimuli, with a fixed duration of 60 seconds for each clip. Groundtruth labels for arousal and valence are given as self-assessment ratings in the continuous range of $[1.0, 9.0]$. The end-goal task on DEAP is the classification of "Low"/"High" affective states, defined by thresholding the rating scale in the middlepoint of 5.0. The classification head of our deep architecture predicts class scores for these two outputs on arousal and valence.

In the case of DEAP dataset, the original labels are continuous ratings that are quantized to obtain the final classification labels, thus their initial ordinality is lost. Moreover, collapsing entire ranges of the rating scale into single classes leads to models that cannot reason about intra-class sample differences. The application of a ranking approach on the original ratings is straight-forward, following the ordinal relations that are shown in Table 5.1.

SEED dataset: SEED [256] is a dataset for EEG-based emotion recognition, having 15 participants and 15 Chinese movie videos as stimuli, with varying duration for each clip (4 minutes in average). The labels are categorical, belonging in three classes, namely "Positive", "Neutral" and "Negative". The end-goal task of SEED is the classification of these three states. The discrete class annotations of SEED are traditionally

treated as being nominal, ignoring the evident ordinality. The classes of SEED practically correspond to three ordered levels of valence, therefore inferring ordinal relations between them is plausible. We adopt the convention that the "Positive" class corresponds to higher valence compared to "Neutral" and "Negative", and that the "Neutral" class corresponds to higher valence compared to "Negative". These ordinal relations that are used on SEED dataset are shown in Table 5.3.

y_1 \ y_2	**Negative**	**Neutral**	**Positive**
Negative	$x_1 \sim x_2$	$x_1 \prec x_2$	$x_1 \prec x_2$
Neutral	$x_1 \succ x_2$	$x_1 \sim x_2$	$x_1 \prec x_2$
Positive	$x_1 \succ x_2$	$x_1 \succ x_2$	$x_1 \sim x_2$

Table 5.3: The ordinal relations that are adopted in our work, for the categorical labels of SEED dataset to be rendered useful in the pairwise ranking task.

5.3.2 Experiments

EEG data preparation: To perform training on DEAP and SEED, we represent each input sample in the form of a feature vector. Among the most well-established EEG signal features for emotion recognition, are Power Spectral Density (PSD), Power Spectral Asymmetry (PSA) and Differential Entropy (DE). For the signal of each electrode, these features are computed in a specific frequency band and for a short time window (2 seconds on DEAP, 1 second on SEED). We use 5 frequency bands for feature extraction, namely theta band ($4-8$ Hz), alpha band ($8-12$ Hz), slow alpha band ($8-10$ Hz), beta band ($12-30$ Hz) and gamma band ($30-45$ Hz). PSD features characterize the spectral content of each signal, while PSA features measure the asymmetric hemisphere activation occuring in the brain through pairs of laterally corresponding/symmetric electrodes. We compute PSD and PSA as in [122], using the method of Welch [234]. The DE features measure the complexity of the signal across time [68]. On DEAP dataset, we use the PSD and PSA features, concatenating their feature vectors. On SEED, we use the precomputed DE features that are provided in [256]. On both datasets, to discard features of negligible discriminability, feature

selection is applied using Fisher's linear discriminant, similarly to [122], keeping the 100 most discriminative features. Afterwards, a zero-mean and unit-variance normalization procedure is applied on each of the remaining features, using the statistics of the train set.

Training details: Training is done for 20 epochs with a batch size of 40, using a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001, a momentum of 0.9 and weight decay set equal to 0.0005. For DEAP dataset, the ordinal ranking operation is performed setting $\epsilon = 0.25$ and following Table 5.1. The training process is a subject-dependent 10-fold cross validation. For each subject the 40 available trials are split into 10 folds (each fold containing 4 trials), keeping 9 folds as the train set and 1 fold as the test set. For SEED dataset, the ordinal ranking operation is performed following Table 5.3. The training process is subject-dependent and the train-test splits are done in the same way with [256]. On both datasets, evaluation is done by computing the classification accuracy and F1 score. Especially on DEAP where there is significant class imbalance, the F1 score is a more representative measure of model performance.

Our experiments explore the impact of joint training on the model classification performance. As a baseline method, a plain MLP network (with 2 FC layers in its embedding module and 1 FC classification layer) is trained only on the classification task. We train our proposed architecture jointly on the classification and ranking tasks. From the results of Table 5.4 and Table 5.5 we can see that joint training improves the accuracy and F1 score both on DEAP and SEED. Considering the F1 scores, the performance improvement of the proposed method over the baseline is statistically significant on DEAP ($p < 0.01$ for both arousal and valence), but not on SEED ($p = 0.058$).

The results verify our motivation of forming and learning pairwise relations utilising the available affective annotations. On DEAP, we notice that collapsing fine-grained

Model	Arousal		Valence	
	Acc.	**F1**	**Acc.**	**F1**
Classification loss	60.49	51.94	57.69	54.61
Proposed method: Classification + ranking loss	**60.60**	**53.25***	**58.42**	**55.57***

Table 5.4: Accuracy (%) and F1 score on DEAP dataset. Stars indicate statistical significance of the F1-score distribution over subjects, according to Student's t-test ($* = p < 0.05$)

Model	3-class problem	
	Acc.	**F1**
Classification loss	74.80	72.79
Proposed method: Classification + ranking loss	**76.98**	**75.51**

Table 5.5: Accuracy (%) and F1 score on SEED dataset.

affective rating information into discrete classes, is harmful for the training process. Incorporating the ranking supervision through the ordinal relation labels derived by the original continuous ratings, we boost the performance of our model. Similarly, the fact that our approach considers the ordinality of the classes on SEED, shows that our method can be beneficial even in cases where the original annotations are discrete.

5.4 Conclusions

The findings of our method highlight that exploring the ordinality of emotions through deep neural networks that accomodate pairwise ranking comparisons, is beneficial for affect recognition models. The proposed method is evaluated on two EEG datasets with different affective annotation processes, showing consistent performance gains. We believe that our study provides a promising direction on training robust emotion recognition models, through tasks that abide to the ordinal nature of emotions.

Chapter 6

Conclusions

Contents

6.1 Results and contributions

In this thesis we explored the development of deep learning techniques for the EEG-based problems of MI decoding and affect recognition. We started by studying the issue of domain shifts in cross-subject MI decoding, showing that the combination of model ensembling, curriculum learning and collaborative training can effectively increase model robustness. Furthermore, we investigated a data augmentation technique for CNN regularization and proposed an adaptation of the covariance-based data alignment process for EEG time-series. Finally, regarding the problem of affect recognition, we developed a training method that can exploit the inherent ordinality of emotions, towards learning better representations.

In the first main chapter of this thesis we started by considering the challenges that arise when training on multi-subject EEG datasets for MI decoding. Specifically, we identified the several factors of variation that lead to domain shifts across subjects and described the inter-subject differences in terms of spatial, temporal and spectral EEG characteristics. We discussed the possible directions of research that can be used to tackle the issue of domain shifts, and opted to focus on the development of a domain generalization technique. Specifically, we aimed to build a method employing diverse feature extractors to avoid the issue of negative transfer learning due to domain shifts. The category of ensemble learning techniques has been well-known for accomodating diverse feature learning, however existing ensemble learning works presented several drawbacks. These included high computational costs, length model selection processes, requirement for hyperparameter tuning and dependence on subject-specific components, among others.

We framed our proposed approach as a model ensembling method combined with an ensemble curriculum learning strategy and a collaborative training scheme through intra-ensemble distillation. Both curriculum learning and knowledge distillation have been largely unexplored in deep learning methods for EEG-based tasks. We showed that curriculum learning promotes feature diversity across the multiple feature extractors of a model ensemble. Nevertheless, controlling the extent of diversity within our ensemble architecture was necessary, as we observed a trade-off between the properties of diversity and generalization. We also showed that this trade-off can be effectively regulated via collaborative training, that is materialized through an intra-ensemble distillation objective in the training process.

We evaluated our proposed approach on two large MI datasets comprising more than 150 subjects, achieving superior results against state-of-the-art works and standard ensembling methods. Additionaly, we conducted ablation studies to investigate the impact of each component of our method. Our findings indicated the strong potential

of domain generalization techniques as a tool for overcoming inter-subject differences in motor imagery decoding.

Further studying the problem of MI decoding, in the second main chapter of this thesis we focused on EEG data alignment techniques that have presented remarkable benefits towards learning domain-invariant representations. Considering the non-stationarity of EEG signals across time, we investigated the incorporation of short-time (i.e. trial-wise) statistics in the alignment process, instead of solely using the session-wise statistics. Our goal was to obtain multiple alignment transformations of each sample (i.e. trial) during training, in order to regularize a CNN model that performs MI decoding. We adapted the standard alignment process by performing Riemannian interpolation between SPD matrices that contained trial-wise and session-wise statistics. Moreover, we inserted a stochastic component in our method, by opting to mix trial-wise and session-wise statistics with a random proportion in each training iteration. The adapted alignment framework was employed during training, while standard alignment was used during testing to keep the process being deterministic. Building on the benefits of CNN-based learning and covariance-based alignment, our developed approach concurrently performed data alignment and augmentation on EEG time-series. We conducted experiments on a MI dataset and compared our method against standard alignment as well as other domain generalization techniques, showing superior results.

Finally, in the last main chapter of the thesis, we explored the topic of affect recognition. Affective datasets typically contain emotion annotations that are obtained through self-assessment ratings. Firstly, considering such ratings as a form of annotation, we discussed related works that describe their ordinality as well as the factors that render them highly subjective. We also explained why we considered human emotional judgments as being relative to the context, rather than absolute. Moreover, we provided details about the biases and the loss of information that occur when transforming originally continuous annotations into discrete classes, from works that adopt

plain classification approaches for affect recognition. Based on this analysis, we suggested expressing emotions in relative terms, i.e. through pairwise comparisons between affective states. We derived the pairwise ordinal relations than can be inferred using the original annotations of affective datasets and we formulated a ranking task. This task was then additionaly employed during training, along with the standard classification task. To achieve this, we designed a neural network architecture consisting of a shared backbone feature extractor and two task-specific heads, namely a classification head and a ranking head. The goal of our proposed training methodology was to obtain feature representations capturing not only the nominal labels of the classification task, but also the ordinal structure of emotions. In this way, we exploited knowledge from the original continuous annotations that led to feature representations of higher quality. We conducted experiments on two datasets, comparing models that were jointly trained on the tasks of classification and ranking, against models that were trained only on the classification task. Our results showed that incorporating the ranking task in the training process is beneficial to model performance. Thus, developing deep learning techniques that abide to the ordinal nature of emotions is a plausible direction of research towards obtaining robust affect recognition models.

6.2 Wider implications

The study presented in Chapter 3 of this thesis provided valuable insights on how to design and train robust deep neural networks in multi-domain datasets, exploiting the power of learning diverse representations through model ensembling. The effectiveness of this approach has broader repercussions on the pursued directions in the field of BCIs. The rapid adoption of deep learning techniques within several BCIs involving decoding tasks, has been a result of their capability to: i) accomodate multi-subject training and ii) generalize on unseen subjects. Yet the exploration of deep learning methods that can operate under various domain shifts, has been restricted to ERM-based approaches that equally minimize the training loss across all source domains. While such approaches

have been proven to yield satisfactory cross-subject performance, they also come with a cost. Specifically, the phenomenon of negative transfer learning that has been observed in several neuroimaging modalities [233, 95], partially limits the generalization potential of ERM-based techniques. Our proposed method was able to alleviate to an extent this effect, due to its unique training methodology where each base model within an ensemble, is trained so as to specialize to a different subset of the source domains (i.e. training subjects). Hence the models comprising the ensemble are less exposed to the detrimental effects of domain shifts. This finding sheds light on the importance of the optimization objectives when training deep networks on multi-domain neuroimaging datasets.

Departing from the traditional idea of ERM-based training that leads to pure subject-agnostic layers, we introduced a training methodology that leads to layers combining both subject-agnostic and subject-specific properties. CSP-based approaches [148, 46] have explored the feasibility of selecting multiple relevant source subjects that could provide good subject-to-subject transferability, through appropriately formed optimization problems. These early works have served as inspiring examples of handcrafted multi-domain training techniques that differ from standard ERM-based methods. However, since the adoption of deep learning approaches for BCIs, similar research directions have remained relatively unexplored, with most efforts focusing on alternative deep network architecture designs. Existing works trying to design deep architectures that are capable of extracting diverse representations, have followed several approaches, apart from the standard model ensembling. These include using multi-branch networks with a varying number of filters [8] or varying filter length [67], feeding multi-view representations as inputs through band-specific filtering [152] and subject-specific branches [233]. Such details of the network architectures do not suffice to account for the multiple aspects of domain shifts and inter-subject variability.

We argued that further unleashing the potential of deep learninng methods for rep-

resentation learning on neural signals, requires to jointly consider the design of network architectures and training objectives. Subsequently, we built an ensemble architecture and paired its training methodology with two ensemble-oriented loss terms that promote both feature diversity and generalization at the same time. The ablation studies on the contribution of each one among the two novel proposed training objectives, validated that both are helpful towards better cross-subject decoding and proved their usefulness for our architecture. This finding can serve as a starting point for further research on several modalities beyond EEG, where there are no well-established domain generalization techniques.

Previous research on domain generalization has mostly focused on the visual modality, where it has been shown that strong backbone models have a beneficial role [83] in generalization. There has been very little cross-fertilization between domain generalization algorithms for visual and EEG data, and up to now a systematic study on the methods that work well on both of these two modalities has not been reported. One major reason is that many domain generalization works for visual data [140] rely on models pretrained on large vision datasets (e.g. ResNet [90] pretrained on ImageNet [128]). Among other choices, the strength (i.e. the accuracy on the test set of the external dataset) of these models affects the performance on the final task [249]. The lack of off-the-shelf pretrained models for EEG data, means that domain generalization algorithms benefit less from the power of transfer learning. This makes it difficult for researchers to draw parallels from visual to EEG data, and to be inspired from techniques that are tailored to visual data, adapting them on EEG data. We hope that our findings will prove of great interest to researchers studying cross-subject decoding techniques for various neural signal modalities.

In Chapter 4 we introduced a novel approach for MI decoding, called "CovMix", that draws inspiration from domain-invariant representation learning and data augmentation techniques. One implication of our method is that it highlighted the importance of

designing neuroscientifically plausible transformations for data augmentation. Previous works on data augmentation for EEG time-series signals have adopted transformations that yield non-interpretable changes in the content of EEG [119]. On the contrary, our work explored the application of spatial transformations on EEG signal alignment and augmentation, where the employed spatial filters have a direct association to the statistics of a particular trial. This is an important aspect of data augmentation techniques for various biosignal modalities where often it is desirable that the applied transformations result in data which maintain the physical properties of these biosignals. In the case of CovMix, we showed that it is possible to exploit subject-specific trial-wise information in order to generate new EEG signals that maintain a spatial covariance matrix corresponding to an individual. A second implication of our method is that common knowledge about the non-stationarity of EEG signals [248] should be considered when designing novel alignment approaches. The statistics (e.g. inter-channel spatial correlations) of EEG signals change constantly and experimental conditions such as task difficulty, audio/visual stimuli and feedback type can cause further changes [197]. Exploring alignment techniques that take into account this non-stationarity can yield data-driven methods that are more robust both to intra-session and cross-session variations of EEG signal statistics. Thus, a favourable ability of CovMix is that it can lead to trained models which present increased robustness when deployed in test sessions where the conditions are different than those in the training data. This aspect has remained unexplored in previous works that studied the design of alignment or data augmentation techniques without considering the non-stationarity of EEG signals.

In Chapter 5 we introduced a method for EEG-based affect recognition, where the task of pairwise ordinal ranking with respect to affective ratings was proposed as an auxiliary training objective, along with the standard cross-entropy loss for the task of classification. Previous works have focused on directly employing typical classification objectives, collapsing continuous affective attributes such as arousal and valence into discrete classes. A broader consequence of this work is the importance of exploiting

fine-grained affective label information, which often comes in abundance in affective datasets.

6.3 Strengths and limitations

Sample sizes: The size of the datasets that were used in Chapter 3 is a strength of the presented work. Specifically, PhysioNet [79, 200] dataset contains 109 subjects while OpenBMI [135] dataset contains 54 subjects, being among the largest datasets on the task of motor imagery decoding. The dataset of BCI Competition IV-2a [215] that was used in Chapter 4 has a small size, containing only 9 subjects. This limits the reliability of the findings, requiring further investigation using datasets with larger sample sizes. Regarding the task of affect recognition, the utilized datasets had a moderate size, with DEAP [122] containing 32 subjects and SEED [256] containing 15 subjects. Apart from the datasets of DEAP and SEED that were used in this thesis, other existing datasets for EEG-based affect recognition [207, 228, 159, 4, 50, 171] contain from 10 to 40 subjects, restricting the ability to conduct thorough studies.

Cross-subject generalization: The works presented in Chapter 3 and Chapter 4 were evaluated in cross-subject scenarios, either in Leave-One-Subject-Out or in k-fold cross-validation (with each fold containing multiple subjects) experimental settings. This is a strength of these works, since we demonstrated that achieving robust cross-subject generalization is possible without utilizing data from the test subjects during training. The work presented in Chapter 5 was restricted to an intra-subject analysis, hence the findings may not apply to cross-subject affect recognition scenarios.

Aging effects: The effects of aging on the ability of humans to effectively perform motor imagery, have been investigated in a plethora of works [108, 161, 138]. One important aspect of building EEG-based models for BCIs, is the validation of their functionality across age groups. The motor imagery datasets of PhysioNet [200, 79] and BCI IV-2a [215], that were used in Chapter 3 and Chapter 4 respectively, did not

provide the age of each subject as accompanying information. This was a limiting factor of the presented analyses, as without knowing the age distributions of the subjects in the training and test sets, no certain conclusions could be drawn regarding generalization across age groups. On the contrary, the motor imagery dataset of OpenBMI [135] that was used in Chapter 3, provided the age distribution of its subjects ($24-35$ years old). Regarding the datasets that were used in Chapter 5, DEAP [122] had an age range of $19-37$ years old (mean age 26.9), while SEED [256] dataset reported that the mean age was 23.2 ± 2.3 years old. The age ranges of OpenBMI, DEAP and SEED were limited to young people, which on the one hand might affect the potential for generalization on elderly people, but on the other hand allowed for coherent performance analyses on this specific age group.

EEG data form: Another strength of the works presented in Chapter 3 and Chapter 4 is that they operated on minimally preprocessed EEG time-series signals, building on the capability of CNNs to learn representations from raw waveforms. This provided interpretability to the behaviour of the trained CNN models, as their temporal filters define the spectral content of the filtered signals and their spatial filters define the importance of each EEG electrode in the spatial mixing process. On the contrary, the architecture presented in Chapter 5 operated on handcrafted features of EEG signals, such as power spectral density and differential entropy. These features were separately computed for each EEG electrode and each frequency band. Hence two limitations of this work are the following: i) there was no ability to capture spectral content in frequency ranges different than those of the pre-specified bands (delta, theta, alpha, beta and gamma) and ii) there was no ability to jointly learn the spatial mixing along with the temporal filtering process. This highlights the imminent need for EEG-based affect recognition methods that are capable of processing raw EEG waveforms and providing interpretable results.

EEG montage density: The density of an EEG montage is a factor that has a

significant impact in the performance that can be obtained while operating an EEG-based BCI system. Dense EEG montages generally provide higher accuracies, due to their high spatial resolution and wide spatial coverage, that allow inferring a more complete estimate of the neural activity inside the brain [205]. Investigating the feasibility of low-density EEG montages is of great value for bringing BCIs into real-world scenarios using portable, consumer-grade EEG devices that typically have low-density montages. The works presented in this thesis kept the full EEG montage of each dataset while analyzing the data of each task. In more detail, regarding the MI datasets of PhysioNet, OpenBMI and BCI IV-2a, we used their full montages which contained 64, 62 and 22 electrodes respectively. Additionaly, a part of our experimental analysis on the dataset of OpenBMI involved a reduced montage with a subset of 20 electrodes, which can still be considered as a medium-density montage. Furthermore, regarding the affective datasets of DEAP and SEED, we used their full montages containing 32 and 62 electrodes respectively. Thus, the fact that low-density EEG montages were not explored as potential options for MI decoding and affect recognition, is a limitation of the presented works.

Label subjectiveness and uncertainty: It is widely known that affective experiences are highly subjective [55] and also that cross-cultural differences have an impact on the perception of affective stimuli across individuals [54]. Affective datasets often employ self-reporting methods [122, 158], letting the participants to annotate their perceived emotions. While self-reported labels are valuable when considering personalized experiences, they also reflect the annotation biases of each subject [153]. For example, music/movie genre preferences can lead to highly different affective ratings across subjects for a given stimulus. Moreover, several affective datasets employ videos as stimuli, with their duration usually ranging between 1-5 minutes [122, 256]. Each stimulus video may contain arbitrary scenes, without restrictions on the diversity of the audiovisual content within shorter temporal segments of the video. In such cases, assigning a single affective rating that corresponds to the entire duration of each stimu-

lus, may not suffice to describe the emotional experience of an individual. Determining the particular moments within a stimulus that led a subject to assign a given label, is a challenging problem due to the inherent uncertainty of human judgments. Hence, a limitation of the work presented in Chapter 5, is that it does not study the subjective biases and the label uncertainty that are task of affect recognition.

6.4 Potential applications

The power of deep learning techniques for EEG brainwave decoding, along with the portability of EEG devices, enable the widespread usage of EEG-based BCI systems in out-of-the-lab settings. Indicative applications of brain-computer interfaces that are relevant to the tasks studied in this thesis encompass the sectors of healthcare, gaming and multimedia services.

The methods that were developed in Chapter 3 and Chapter 4 of this thesis, are related to the area of neuro-rehabilitation [186], where EEG-based BCIs have found a variety of applications. Neuro-rehabilitation is based on the principle of neuroplasticity [114], i.e. the potential of the brain to reorganise, building new neural pathways that have a positive impact in the restoration of motor skills. EEG-based neurofeedback has been incorporated as a component of rehabilitation systems, helping people who experience motor weakness following a stroke, brain injury or other neurological conditions such as Parkinson's disease and multiple sclerosis. Motor imagery exercises are an essential part of neuro-rehabilitation systems for motor skill training, suitable for patients in all rehabilitation phases, i.e. acute, subacute and chronic [170]. Real-time feedback [57] can be given to the patients while they practise motor imagery training, based on the output of machine learning models trained on the task of MI decoding. This feedback can be beneficial by closing the loop between motor intentions and virtual reality (VR) environments [226, 227]. Furthermore, the BCI-derived feedback can be utilized to apply functional electrical stimulation (FES) [231, 27] in

order to engage weakened muscles of the extremities. Other application scenarios of neuro-rehabilitation involve BCIs that provide vibrotactile feedback [81] and BCIs that control soft robotic gloves [47] or digital therapy devices [117].

The task of EEG-based affect recognition that was studied in Chapter 5 of this thesis, is related to the area of affective BCIs (aBCIs) [238]. A recent application of aBCIs involves monitoring the emotional states of users during multimedia presentation. For example, considering the case of audio-based multimedia content, the BCI-derived feedback of music-induced emotions can be used to create personalized song playlists [41] or to adaptively select the songs of a music stream in an online manner [69]. Moreover, in the case of video content such as advertisements, aBCIs can be used for neuromarketing [116] applications, e.g. for evaluating preferences of products and commercials [230, 85]. Another interesting application of aBCIs is that of discovering correlates of mental health disorder biomarkers from EEG responses to affective stimuli. An indicative application relates to anhedonia [146], a symptom of depression, where lack of pleasure can be observed in the responses to presented pleasurable stimuli. Machine learning techniques that can quantify the affective experiences of individuals during multimedia stimuli [96] can serve as powerful tools in the exploration of EEG-based correlates of mental health biomarkers.

6.5 Directions for future research

The results presented in this thesis open up new research directions that can be pursued to further build deep learning techniques for EEG-based brain-computer interfaces. The family of domain generalization techniques that are tailored to EEG data is still under-explored, with several works resorting to domain adaptation methods [106] that require access to data from the target domain during training. Building foundation models [56] for EEG data is an essential part of this process, similarly to the field of visual data analysis where strong backbone models have been proven to be

highly beneficial within domain generalization frameworks [83]. Moreover, the framework developed in Chapter 3 only investigated the scenario of randomly splitting a dataset into subsets of subjects. Exploring other options for dataset splitting is an interesting research direction. For example, using measures that are related to the brain anatomy of each subject (e.g. functional connectivity of EEG signals) or measures that quantify the "BCI illiteracy" or data quality of each subject (e.g. the separability between classes for each subject), are some indicative and meaningful alternatives for dataset splitting. Another useful direction is that of adapting generic (i.e. subject-agnostic) models to personalized ones [22] by calibrating with a minimal number of trials from a candidate target subject. Furthermore, including EEG data collected during executed movements (e.g. using PhysioNet [79, 200] dataset that contains EEG data from both executed and imagined movements), can inform the training process of deep learning models for decoding imagined movements. Moreover, studying the impact that EEG non-stationarity levels have on the ability of users to voluntarily control BCIs by modulating their sensorimotor rhythms, is an unexplored problem which is highly related to the study that was presented in Chapter 4 of this thesis. Recent works from the field of neuroscience that provide a more accurate description of the motor cortex [80], augmenting the long-existing "Penfield Homunculus" [175], can also inspire the development of novel EEG-based BCIs for decoding movements from brain activity.

Regarding the task of affect recognition, one future direction could be the collection and analysis of reliable datasets with a more precise and clear relation between the affective ratings and the content of lengthy audiovisual stimuli (e.g. introducing more dense annotations for short temporal segments). This would bridge the gap between works studying affect elicitation in short ERP-based data from static image stimuli [96] and works that analyze emotions elicited from video stimuli [158]. Additionaly, dense temporal annotations would render the task of continuous EEG-based tracking of affective states more realistic, compared to the currently prevalent practice of inferring

affective states corresponding to entire videos. Another important issue that deserves further research relates to the impact that internal context (e.g. the content of temporally neighboring stimuli) has on human judgments, when individuals self-report their affective states. Evidence shows that humans form an internal context [203] from previously encountered stimuli that act as reference points when evaluating forthcoming experiences [246]. Investigating whether and how this context affects the perceived emotions and the collected annotations, could help improving the design of psychology-informed experimental paradigms for affective BCIs. Finally, the topic of identifying the EEG electrode locations that are essential for affect recognition should be investigated thoroughly, as this would enable bringing affective BCIs in out-of-the-lab settings using sparse EEG montages and consumer grade devices.

Bibliography

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of neural engineering*, 16(1):011001, 2019. 22

[2] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev. Analysis of evoked eeg synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neuroscience and behavioral physiology*, 34(8):859–867, 2004. 40

[3] M. Ahn, H. Cho, S. Ahn, and S. C. Jun. High theta and low alpha powers may be indicative of bci-illiteracy in motor imagery. *PloS one*, 8(11):e80886, 2013. 10

[4] T. B. Alakus, M. Gonen, and I. Turkoglu. Database for an emotion recognition system based on eeg signals and various computer games–gameemo. *Biomedical Signal Processing and Control*, 60:101951, 2020. 103

[5] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017. 7

[6] L. B. Alford. Localization of consciousness and emotion. *American Journal of Psychiatry*, 89(4):789–799, 1933. 39

[7] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: a review. *Neural Computing and Applications*, pages 1–42, 2021. 32

[8] G. A. Altuwaijri, G. Muhammad, H. Altaheri, and M. Alsulaiman. A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for eeg-based motor imagery signals classification. *Diagnostics*, 12(4):995, 2022. 35, 100

[9] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 2390–2397. IEEE, 2008. 27, 31, 35

[10] A. Antoniades, L. Spyrou, C. C. Took, and S. Sanei. Deep learning for epileptic intracranial eeg data. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016. 30

[11] A. Appriou, A. Cichocki, and F. Lotte. Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals. *IEEE Systems, Man, and Cybernetics Magazine*, 6(3):29–38, 2020. 41

[12] P. Autthasan, R. Chaisaen, T. Sudhawiyangkul, P. Rangpong, S. Kiatthaveephong, N. Dilokthanakul, G. Bhakdisongkhram, H. Phan, C. Guan, and T. Wilaiprasitporn. Min2net: End-to-end multi-task learning for subject-independent motor imagery eeg classification. *IEEE Transactions on Biomedical Engineering*, 69(6):2105–2118, 2021. 31, 33, 50, 61, 64, 66, 70

[13] A. M. Azab, L. Mihaylova, K. K. Ang, and M. Arvaneh. Weighted transfer learning for improving motor imagery-based brain–computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1352–1359, 2019. 28

[14] S. Bakas, S. Ludwig, K. Barmpas, M. Bahri, Y. Panagakis, N. Laskaris, D. A. Adamos, and S. Zafeiriou. Team cogitat at neurips 2021: Benchmarks for eeg transfer learning competition. *arXiv preprint arXiv:2202.03267*, 2022. 34

[15] P. D. E. Baniqued, E. C. Stanyer, M. Awais, A. Alazmani, A. E. Jackson, M. A. Mon-Williams, F. Mushtaq, and R. J. Holt. Brain–computer interface robotics for hand rehabilitation after stroke: A systematic review. *Journal of NeuroEngineering and Rehabilitation*, 18(1):1–25, 2021. 21

[16] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. 6

[17] H. Banville, S. U. Wood, C. Aimone, D.-A. Engemann, and A. Gramfort. Robust learning from corrupted eeg with dynamic spatial filtering. *NeuroImage*, 251:118994, 2022. 6, 76

[18] A. Barachant, A. Andreev, and M. Congedo. The riemannian potato: an automatic and adaptive artifact detection method for online experiments using riemannian geometry. In *TOBI Workshop lV*, pages 19–20, 2013. 30

[19] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011. 29, 72

[20] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*, 112:172–178, 2013. 30

[21] K. Barmpas, Y. Panagakis, D. Adamos, N. Laskaris, and S. Zafeiriou. A causal viewpoint on motor-imagery brainwave decoding. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 65, 66, 70

[22] K. Barmpas, Y. Panagakis, S. Bakas, D. A. Adamos, N. Laskaris, and S. Zafeiriou. Improving generalization of cnn-based motor-imagery eeg decoders via dynamic convolutions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 108

[23] A. Beck. Oznaczenie lokalizacyi w mózgu i rdzeniu za pomocą zjawisk elektrycznych [the determination of localizations in the brain and spinal cord with the aid of electrical phenomena]. *Rozprawy Akademii Umiejętności, Wydział Matematyczno-Przyrodniczy*, 1:187–232, 1891. 4

[24] C. Bell. *Essays on the Anatomy of Expression in Painting.* Longman, Hurst, Rees, and Orme, 1806. 35

[25] H. Berger. Über das elektroenkephalogramm des menschen [on the electroencephalogram of man]. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929. 4

[26] R. Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009. 74, 75

[27] S. Bhattacharyya, M. Clerc, and M. Hayashibe. Augmenting motor imagery learning for brain–computer interfacing using electrical stimulation as feedback. *IEEE Transactions on Medical Robotics and Bionics*, 1(4):247–255, 2019. 106

[28] Y. Bian and H. Chen. When does diversity help generalization in classification ensembles? *IEEE Transactions on Cybernetics*, 2021. 13, 51

[29] G. Bin, X. Gao, Y. Wang, B. Hong, and S. Gao. Vep-based brain-computer interfaces: time, frequency, and code modulations [research frontier]. *IEEE Computational Intelligence Magazine*, 4(4):22–26, 2009. 24

[30] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2007. 8, 26

[31] J. C. Borod, B. A. Cicero, L. K. Obler, J. Welkowitz, H. M. Erhan, C. Santschi, I. S. Grunwald, R. M. Agosti, and J. R. Whalen. Right hemisphere emotional perception: evidence across multiple channels. *Neuropsychology*, 12(3):446, 1998. 40

[32] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. 8, 11, 37, 38

[33] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 62

[34] K. Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues [Comparative localization theory of the cerebral cortex represented in its principles on the basis of cell structure].* Barth, 1909. 18

[35] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 86

[36] L. Canini, S. Benini, and R. Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):636–647, 2013. 7

[37] R. Caton. The electric currents of the brain. *British Medical Journal*, 2:278, 1875. 4

[38] H. Cecotti and A. Graser. Convolutional neural networks for p300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):433–445, 2010. 30

[39] D. D. Chakladar, S. Dey, P. P. Roy, and M. Iwamura. Eeg-based cognitive state assessment using deep ensemble model and filter bank common spatial pattern. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4107–4114. IEEE, 2021. 34, 46

[40] A. T. Chan, J. C. Quiroz, S. Dascalu, and F. C. Harris. An overview of brain computer interfaces. In *Proc. 30th Int. Conf. on Computers and Their Applications*, 2015. 24

[41] H.-Y. Chang, S.-C. Huang, and J.-H. Wu. A personalized music recommendation system based on electroencephalography feedback. *Multimedia Tools and Applications*, 76:19523–19542, 2017. 107

[42] G. E. Chatrian, E. Lettich, and P. L. Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities. *American Journal of EEG technology*, 25(2):83–92, 1985. 18

[43] M. Chaumon, D. V. Bishop, and N. A. Busch. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods*, 250:47–63, 2015. 21

[44] M. Chen, J. Han, L. Guo, J. Wang, and I. Patras. Identifying valence and arousal levels via connectivity between eeg channels. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 63–69. IEEE, 2015. 41

[45] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 59

[46] M. Cheng, Z. Lu, and H. Wang. Regularized common spatial patterns with subject-to-subject transfer of eeg signals. *Cognitive neurodynamics*, 11(2):173–181, 2017. 28, 100

[47] N. Cheng, K. S. Phua, H. S. Lai, P. K. Tam, K. Y. Tang, K. K. Cheng, R. C.-H. Yeow, K. K. Ang, C. Guan, and J. H. Lim. Brain-computer interface-based soft robotic glove rehabilitation for stroke. *IEEE Transactions on Biomedical Engineering*, 67(12):3339–3351, 2020. 107

[48] M. N. Cherloo, H. K. Amiri, and M. R. Daliri. Ensemble regularized common spatio-spectral pattern (ensemble rcssp) model for motor imagery-based eeg signal classification. *Computers in Biology and Medicine*, 135:104546, 2021. 28

[49] I. Choi, I. Rhiu, Y. Lee, M. H. Yun, and C. S. Nam. A systematic review of hybrid brain-computer interfaces: Taxonomy and usability perspectives. *PloS one*, 12(4):e0176674, 2017. 24

[50] Y. Cimtay and E. Ekmekcioglu. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition. *Sensors*, 20(7):2034, 2020. 103

[51] M. X. Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014. 8

[52] M. Congedo, A. Barachant, and R. Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. 29

[53] M.-C. Corsi, S. Chevallier, F. D. V. Fallani, and F. Yger. Functional connectivity ensemble method to enhance bci performance (fucone). *IEEE Transactions on Biomedical Engineering*, 2022. 34

[54] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117(4):1924–1934, 2020. 105

[55] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017. 105

[56] W. Cui, W. Jeong, P. Thölke, T. Medani, K. Jerbi, A. A. Joshi, and R. M. Leahy. Neuro-gpt: Developing a foundation model for eeg. *arXiv preprint arXiv:2311.03764*, 2023. 107

[57] S. Darvishi, M. C. Ridding, B. Hordacre, D. Abbott, and M. Baumert. Investigating the impact of feedback update interval on the efficacy of restorative brain–computer interfaces. *Royal Society open science*, 4(8):170660, 2017. 106

[58] C. Darwin. The expression of the emotions in man and animals. *London: J. Murray*, 1872. 36

[59] R. J. Davidson. Affect, cognition, and hemispheric specialization. In *Emotions, cognition, and behavior*, pages 320–365. Cambridge Univ. Press, 1984. 39

[60] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology: I. *Journal of personality and social psychology*, 58(2):330, 1990. 39

[61] K. M. Davis, C. de la Torre-Ortiz, and T. Ruotsalo. Brain-supervised image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18480–18489, 2022. 21

[62] L. Deecke, B. Grözinger, and H. Kornhuber. Voluntary finger movement in man: Cerebral potentials and theory. *Biological cybernetics*, 23(2):99–119, 1976. 24

[63] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, pages 1–1, 2022. 41

[64] I. Dolzhikova, B. Abibullaev, R. Sameni, and A. Zollanvari. An ensemble cnn for subject-independent classification of motor imagery-based eeg. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 319–324. IEEE, 2021. 34, 46, 50

[65] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Muller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE transactions on biomedical engineering*, 53(11):2274–2281, 2006. 27

[66] P. K. Douglas and D. B. Douglas. Reconsidering spatial priors in eeg source estimation: does white matter contribute to eeg rhythms? In *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–12. IEEE, 2019. 8

[67] Y. Du and J. Liu. Ienet: a robust convolutional neural network for eeg based brain-computer interfaces. *Journal of Neural Engineering*, 2022. 34, 46, 100

[68] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013. 5, 93

[69] E. Dutta, A. Bothra, T. Chaspari, T. Ioerger, and B. J. Mortazavi. Reinforcement learning using eeg signals for therapeutic use of music in emotion management. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5553–5556. IEEE, 2020. 107

[70] P. Ekman. Universal facial expressions in emotion. *Studia Psychologica*, 15(2):140, 1973. 36

[71] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 36

[72] P. Ekman, R. J. Davidson, and W. V. Friesen. The duchenne smile: Emotional expression and brain physiology: Ii. *Journal of personality and social psychology*, 58(2):342, 1990. 39

[73] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 36

[74] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987. 36

[75] P. Ekman, W. V. Friesen, and S. S. Tomkins. Facial affect scoring technique: A first validity study. 1971. 36

[76] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988. 4, 24

[77] J. Fürnkranz and E. Hüllermeier. *Preference learning*. Springer, 2010. 85

[78] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 41

[79] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 43, 52, 60, 70, 103, 108

[80] E. M. Gordon, R. J. Chauvin, A. N. Van, A. Rajesh, A. Nielsen, D. J. Newbold, C. J. Lynch, N. A. Seider, S. R. Krimmel, K. M. Scheidter, et al. A somato-cognitive action network alternates with effector regions in motor cortex. *Nature*, pages 1–9, 2023. 108

[81] N. A. Grigorev, A. O. Savosenkov, M. V. Lukoyanov, A. Udoratina, N. N. Shusharina, A. Y. Kaplan, A. E. Hramov, V. B. Kazantsev, and S. Gordleeva. A bci-based vibrotactile neurofeedback training improves motor cortical excitability during motor imagery. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1583–1592, 2021. 107

[82] S. Grodal and N. Granqvist. Great expectations: Discourse and affect during field emergence. In *Emotions and the organizational fabric*, volume 10, pages 139–166. Emerald Group Publishing Limited, 2014. 8, 37

[83] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 33, 101, 108

[84] D. E. Gustafson, J. S. Eterno, and W. Jarisch. Signal analysis techniques for interpreting electroencephalograms. Technical report, Scientific Systems Inc. Cambridge MA, 1980. 4

[85] A. Hakim, S. Klorfeld, T. Sela, D. Friedman, M. Shabat-Simon, and D. J. Levy. Machines learn neuromarketing: Improving preference prediction from self-reports using multiple eeg measures and machine learning. *International Journal of Research in Marketing*, 38(3):770–791, 2021. 107

[86] M. Hallett, J. Fieldman, L. G. Cohen, N. Sadato, and A. Pascual-Leone. Involvement of primary motor cortex in motor imagery and mental practice. *Behavioral and Brain Sciences*, 17(2):210–210, 1994. 25

[87] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 890–897, 2017. 11

[88] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990. 34

[89] H. He and D. Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019. 14, 34, 47, 72, 74, 78

[90] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 33, 101

[91] H. Helson. Adaptation-level theory: an experimental and systematic approach to behavior. 1964. 84

[92] S. Herculano-Houzel, K. Catania, P. R. Manger, and J. H. Kaas. Mammalian brains are made of these: a dataset of the numbers and densities of neuronal and

nonneuronal cells in the brain of glires, primates, scandentia, eulipotyphlans, afrotherians and artiodactyls, and their relationship with body mass. *Brain, Behavior and Evolution*, 86(3-4):145–163, 2015. 3

[93] E. R. Heremans, H. Phan, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *Journal of Neural Engineering*, 19(3):036044, 2022. 50

[94] B. Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970. 5

[95] P. Holderrieth, S. Smith, and H. Peng. Transfer learning for neuroimaging via re-use of deep neural network features. *medRxiv*, pages 2022–12, 2022. 100

[96] G. Honke, I. Higgins, N. Thigpen, V. Miskovic, K. Link, S. Duan, P. Gupta, J. Klawohn, and G. Hajcak. Representation learning for improved interpretability and classification accuracy of clinical factors from EEG. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 107, 108

[97] M.-P. Hosseini, A. Hosseini, and K. Ahi. A review on machine learning for eeg signal processing in bioengineering. *IEEE reviews in biomedical engineering*, 14:204–218, 2020. 4

[98] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 31

[99] H. Jasper. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*, 10:370–375, 1958. 18

[100] V. Jayaram and A. Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018. 61

[101] M. Jeannerod. Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage*, 14(1):S103–S109, 2001. 25

[102] R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014. 5, 23, 40

[103] E. Jeon, W. Ko, J. S. Yoon, and H.-I. Suk. Mutual information-driven subject-invariant and class-relevant deep representation learning in bci. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 66

[104] C. Jeunet, F. Lotte, M. Hachet, and B. N'Kaoua. Impact of cognitive and personality profiles on motor-imagery based brain-computer interface-controlling performance. In *17th World Congress of Psychophysiology (IOP2014)*, 2014. 7, 10

[105] C. Jeunet, B. N'Kaoua, S. Subramanian, M. Hachet, and F. Lotte. Predicting mental imagery-based bci performance from personality, cognitive profile and neurophysiological patterns. *PloS one*, 10(12):e0143962, 2015. 7, 10

[106] Y.-M. Jin, Y.-D. Luo, W.-L. Zheng, and B.-L. Lu. Eeg-based emotion recognition using domain adaptation network. In *2017 international conference on orange technologies (ICOT)*, pages 222–225. IEEE, 2017. 41, 107

[107] F. P. Kalaganis, K. Georgiadis, V. P. Oikonomou, N. A. Laskaris, S. Nikolopoulos, and I. Kompatsiaris. Unlocking the subconscious consumer bias: A survey on the past, present, and future of hybrid eeg schemes in neuromarketing. *Frontiers in Neuroergonomics*, 2:672982, 2021. 7

[108] M. Kalicinski, M. Kempe, and O. Bock. Motor imagery: effects of age, task complexity, and task setting. *Experimental aging research*, 41(1):25–38, 2015. 103

[109] E. Kalunga, S. Chevallier, and Q. Barthélemy. Data augmentation in riemannian space for brain-computer interfaces. In *STAMLINS*, 2015. 30, 76

[110] I. Käthner, S. C. Wriessnegger, G. R. Müller-Putz, A. Kübler, and S. Halder. Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain–computer interface. *Biological psychology*, 102:118–129, 2014. 24

[111] S. Katsigiannis and N. Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017. 40

[112] E. G. Kehoe, J. M. Toomey, J. H. Balsters, and A. L. Bokde. Personality modulates the effects of emotional arousal and valence on brain activation. *Social cognitive and affective neuroscience*, 7(7):858–870, 2012. 84

[113] A. Keil, M. M. Müller, T. Gruber, C. Wienbruch, M. Stolarova, and T. Elbert. Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical neurophysiology*, 112(11):2057–2068, 2001. 39

[114] F. Khan, B. Amatya, M. P. Galea, R. Gonzenbach, and J. Kesselring. Neurorehabilitation: applied neuroplasticity. *Journal of neurology*, 264:603–615, 2017. 106

[115] J. Khan, M. H. Bhatti, U. G. Khan, and R. Iqbal. Multiclass eeg motor-imagery classification with sub-band common spatial patterns. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):1–9, 2019. 29

[116] V. Khurana, M. Gahalawat, P. Kumar, P. P. Roy, D. P. Dogra, E. Scheme, and M. Soleymani. A survey on neuromarketing using eeg signals. *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):732–749, 2021. 107

[117] C. Kilbride, D. J. Scott, T. Butcher, M. Norris, A. Warland, N. Anokye, E. Cassidy, K. Baker, D. A. Athanasiou, G. Singla-Buxarrais, et al. Safety, feasibility, acceptability and preliminary effects of the neurofenix platform for rehabilitation

via home based gaming exercise for the upper-limb post stroke (rhombus): results of a feasibility intervention study. *BMJ open*, 12(2):e052555, 2022. 107

[118] W. D. Killgore and D. A. Yurgelun-Todd. The right-hemisphere and valence hypotheses: could they both be right (and sometimes left)? *Social cognitive and affective neuroscience*, 2(3):240–250, 2007. 40

[119] S.-J. Kim, D.-H. Lee, and Y.-W. Choi. Cropcat: Data augmentation for smoothing the feature distribution of eeg signals. In *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–4. IEEE, 2023. 102

[120] T. J. Kimberley, G. Khandekar, L. L. Skraba, J. A. Spencer, E. A. Van Gorp, and S. R. Walker. Neural substrates for motor imagery in severe hemiparesis. *Neurorehabilitation and Neural Repair*, 20(2):268–277, 2006. 25

[121] R. J. Kobler, J.-i. Hirayama, Q. Zhao, and M. Kawanabe. Spd domain-specific batch normalization to crack interpretable unsupervised domain adaptation in eeg. *arXiv preprint arXiv:2206.01323*, 2022. 65, 66, 70

[122] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011. 11, 37, 40, 45, 91, 92, 93, 94, 103, 104, 105

[123] Z. J. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg. *Electroencephalography and clinical Neurophysiology*, 79(6):440–447, 1991. 4, 26

[124] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, page 253, 2021. 32

[125] D. Kostas and F. Rudzicz. Dn3: An open-source python library for large-scale raw neurophysiology data assimilation for more flexible and standardized deep learning. *bioRxiv*, 2020. 6

[126] D. Kostas and F. Rudzicz. Thinker invariance: enabling deep neural networks for bci across more people. *Journal of Neural Engineering*, 17(5):056008, 2020. 10, 14, 31, 34, 61, 63, 64, 70, 72, 76, 79

[127] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. 5

[128] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 33, 101

[129] L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang, D. Lavery, et al. Epilepsyecosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg. *Brain*, 141(9):2619–2630, 2018. 34

[130] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee. Subject-independent brain–computer interfaces based on deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 31(10):3839–3852, 2019. 35, 65, 66

[131] P. J. Lang, M. M. Bradley, B. N. Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997. 40

[132] V. Lawhern, A. Solon, N. Waytowich, S. Gordon, C. Hung, and B. Lance. Eegnet: a compact convolutional network for eeg-based brain-computer interfaces. arxiv. *arXiv preprint arXiv:1611.08024*, 2016. 30

[133] 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*, 15(5):056013, 2018. 30, 35, 54, 76

[134] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 5

[135] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019. 43, 52, 60, 70, 103, 104

[136] X. Lei, P. Yang, P. Xu, T.-J. Liu, and D.-Z. Yao. Common spatial pattern ensemble classifier and its application in brain-computer interface. *Journal of Electronic Science and Technology*, 7(1):17–21, 2009. 28

[137] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu. Cross-subject emotion recognition using deep adaptation networks. In *International conference on neural information processing*, pages 403–413. Springer, 2018. 41

[138] X. Li, P. Chen, X. Yu, and N. Jiang. Analysis of the relationship between motor imagery and age-related fatigue for cnn classification of the eeg data. *Frontiers in Aging Neuroscience*, 14:909571, 2022. 103

[139] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong. A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition. In *IJCAI*, pages 1561–1567, 2018. 41

[140] Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, and D. Li. Domain generalization using pretrained models without fine-tuning. *arXiv preprint arXiv:2203.04600*, 2022. 33, 101

[141] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 37

[142] C. Lindig-León, S. Rimbert, and L. Bougrain. Multiclass classification based on combined motor imageries. *Frontiers in neuroscience*, 14:559858, 2020. 29

[143] A. Litt, C. Eliasmith, and P. Thagard. Neural affective decision theory: Choices, brains, and emotions. *Cognitive Systems Research*, 9(4):252–273, 2008. 7

[144] H. Liu. *Data and the Development of Research Methods in the Science of Human Emotional Expression from Darwin to Klineberg*. PhD thesis, University of Leeds, 2016. 36

[145] J. Liu, H. Meng, M. Li, F. Zhang, R. Qin, and A. K. Nandi. Emotion detection from eeg recordings based on supervised and unsupervised dimension reduction. *Concurrency and Computation: Practice and Experience*, 30(23):e4446, 2018. 40

[146] W.-h. Liu, L.-z. Wang, H.-r. Shang, Y. Shen, Z. Li, E. F. Cheung, and R. C. Chan. The influence of anhedonia on feedback negativity in major depressive disorder. *Neuropsychologia*, 53:213–220, 2014. 107

[147] P. Lopes, G. N. Yannakakis, and A. Liapis. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE, 2017. 85

[148] F. Lotte and C. Guan. Learning from other subjects helps reducing brain-computer interface calibration time. In *2010 IEEE International conference on acoustics, speech and signal processing*, pages 614–617. IEEE, 2010. 28, 100

[149] M. Lotze and L. G. Cohen. Volition and imagery in neurorehabilitation. *Cognitive and behavioral neurology*, 19(3):135–140, 2006. 6, 25

[150] S. Ludwig, S. Bakas, D. A. Adamos, N. Laskaris, Y. Panagakis, and S. Zafeiriou. Eegminer: Discovering interpretable features of brain activity with learnable filters. *arXiv preprint arXiv:2110.10009*, 2021. 33

[151] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu. Reducing the subject variability of eeg signals with adversarial domain generalization. In *International Conference on Neural Information Processing*, pages 30–42. Springer, 2019. 10, 49

[152] W. Ma, H. Xue, X. Sun, S. Mao, L. Wang, Y. Liu, Y. Wang, and X. Lin. A novel multi-branch hybrid neural network for motor imagery eeg signal classification. *Biomedical Signal Processing and Control*, 77:103718, 2022. 35, 100

[153] H. P. Martinez, G. N. Yannakakis, and J. Hallam. Don't classify ratings of affect; rank them! *IEEE transactions on affective computing*, 5(3):314–326, 2014. 38, 84, 85, 105

[154] S. G. Mason and G. E. Birch. A general framework for brain-computer interface design. *IEEE transactions on neural systems and rehabilitation engineering*, 11(1):70–85, 2003. 22

[155] H. Meisheri, N. Ramrao, and S. Mitra. Multiclass common spatial pattern for eeg based brain computer interface with adaptive learning classifier. *arXiv preprint arXiv:1802.09046*, 2018. 29

[156] D. Melhart, K. Sfikas, G. Giannakakis, and G. Y. A. Liapis. A study on affect model validity: Nominal vs ordinal labels. In *Workshop on Artificial Intelligence in Affective Computing*, pages 27–34. PMLR, 2020. 85

[157] Y. Miao, J. Jin, I. Daly, C. Zuo, X. Wang, A. Cichocki, and T.-P. Jung. Learning common time-frequency-spatial patterns for motor imagery classification.

IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29:699–707, 2021. 27

[158] 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*, 12(2):479–493, 2018. 11, 105, 108

[159] S. Mishra, M. Asif, and U. S. Tiway. Dataset on emotions using naturalistic stimuli (dens). 2021. 103

[160] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 37

[161] T. Mulder, J. Hochstenbach, M. Van Heuvelen, and A. Den Otter. Motor imagery: the relation between age and imagery capacity. *Human movement science*, 26(2):203–211, 2007. 103

[162] A. M. Norcia, L. G. Appelbaum, J. M. Ales, B. R. Cottereau, and B. Rossion. The steady-state visual evoked potential in vision research: A review. *Journal of vision*, 15(6):4–4, 2015. 24

[163] Q. Novi, C. Guan, T. H. Dat, and P. Xue. Sub-band common spatial pattern (sbcsp) for brain-computer interface. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pages 204–207. IEEE, 2007. 27

[164] Q. Novi, C. Guan, T. H. Dat, and P. Xue. Sub-band common spatial pattern (sbcsp) for brain-computer interface. In *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pages 204–207. IEEE, 2007. 28

[165] P. L. Nunez, R. Srinivasan, et al. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006. 19

[166] I. Obeid and J. Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016. 32

[167] K. N. Ochsner and J. J. Gross. The cognitive control of emotion. *Trends in cognitive sciences*, 9(5):242–249, 2005. 41

[168] J. A. Onton and S. Makeig. High-frequency broadband modulation of electroencephalographic spectra. *Frontiers in human neuroscience*, page 61, 2009. 40

[169] R. Oostenveld and P. Praamstra. The five percent electrode system for high-resolution eeg and erp measurements. *Clinical neurophysiology*, 112(4):713–719, 2001. 18

[170] R. Ortner, D.-C. Irimia, J. Scharinger, and C. Guger. A motor imagery based brain-computer interface for stroke rehabilitation. *Annual Review of Cybertherapy and Telemedicine*, 181:319–323, 2012. 106

[171] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020. 103

[172] S.-H. Park, D. Lee, and S.-G. Lee. Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):498–505, 2017. 28

[173] S. Parthasarathy, R. Cowie, and C. Busso. Using agreement on direction of change to build rank-based emotion classifiers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2108–2121, 2016. 87

[174] W. Penfield and E. Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443, 1937. 25

[175] W. Penfield and T. Rasmussen. The cerebral cortex of man; a clinical study of localization of function. 1950. 25, 108

[176] S. Pérez-Velasco, E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, and R. Hornero. Eegsym: Overcoming inter-subject variability in motor imagery based bcis with deep learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1766–1775, 2022. 32, 33, 50, 61, 63, 64, 66, 70

[177] C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld. Eeg-bids, an extension to the brain imaging data structure for electroencephalography. *Scientific data*, 6(1):1–5, 2019. 9

[178] P. C. Petrantonakis and L. J. Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on information Technology in Biomedicine*, 14(2):186–197, 2009. 5

[179] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. Da Silva. Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006. 7

[180] G. Pfurtscheller and F. L. Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999. 40

[181] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer. Eeg-based discrimination between imagination of right and left hand movement. *Electroencephalography and clinical Neurophysiology*, 103(6):642–651, 1997. 24

[182] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. 7

[183] S. Rayatdoost and M. Soleymani. Cross-corpus eeg-based emotion recognition. In *2018 IEEE 28th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2018. 84

[184] H. Raza, H. Cecotti, Y. Li, and G. Prasad. Adaptive learning with covariate shift-detection for motor imagery-based brain–computer interface. *Soft Computing*, 20(8):3085–3096, 2016. 83

[185] D. Regan. Some characteristics of average steady-state and transient responses evoked by modulated light. *Electroencephalography and clinical neurophysiology*, 20(3):238–248, 1966. 10

[186] D. J. Reinkensmeyer, V. Dietz, et al. *Neurorehabilitation technology*, volume 798. Springer, 2016. 106

[187] C. Reuben, P. Karoly, D. R. Freestone, A. Temko, A. Barachant, F. Li, G. Titericz Jr, B. W. Lang, D. Lavery, K. Roman, et al. Ensembling crowdsourced seizure prediction algorithms using long-term human intracranial eeg. *Epilepsia*, 61(2):e7–e12, 2020. 34

[188] S. Rimbert, D. Trocellier, and F. Lotte. Is event-related desynchronization variability correlated with bci performance? In *2022 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering*, 2022. 25

[189] M. Riyad, M. Khalil, and A. Adib. Incep-eegnet: a convnet for motor imagery decoding. In *International Conference on Image and Signal Processing*, pages 103–111. Springer, 2020. 31

[190] M. Riyad, M. Khalil, and A. Adib. Mi-eegnet: A novel convolutional neural network for motor imagery classification. *Journal of Neuroscience Methods*, 353:109037, 2021. 31

[191] A. Roc, L. Pillette, J. Mladenovic, C. Benaroch, B. N'Kaoua, C. Jeunet, and F. Lotte. A review of user training methods in brain computer interfaces based on mental tasks. *Journal of Neural Engineering*, 18(1):011002, 2021. 24

[192] 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*, 16(5):051001, 2019. 5

[193] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[194] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 8, 36, 37

[195] J. A. Russell and L. F. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999. 8, 36, 37

[196] J. A. Russell and U. F. Lanius. Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology*, 4(2):119–135, 1984. 84

[197] W. Samek, F. C. Meinecke, and K.-R. Müller. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Transactions on Biomedical Engineering*, 60(8):2289–2298, 2013. 28, 102

[198] C. Sannelli, C. Vidaurre, K.-R. Müller, and B. Blankertz. Csp patches: an ensemble of optimized spatial filters. an evaluation study. *Journal of Neural Engineering*, 8(2):025012, 2011. 28

[199] C. Sannelli, C. Vidaurre, K.-R. Müller, and B. Blankertz. Ensembles of adaptive spatial filters increase bci performance: an online evaluation. *Journal of neural engineering*, 13(4):046003, 2016. 28

[200] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004. 43, 60, 103, 108

[201] 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*, 38(11):5391–5420, 2017. 31

[202] S. Schmidt, H.-G. Jo, M. Wittmann, and T. Hinterberger. 'catching the waves'–slow cortical potentials as moderator of voluntary action. *Neuroscience & Biobehavioral Reviews*, 68:639–650, 2016. 24

[203] B. Seymour and S. M. McClure. Anchors, scales and the relative coding of value in the brain. *Current opinion in neurobiology*, 18(2):173–178, 2008. 109

[204] C. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, jan 1949. 18

[205] A. Soler, L. A. Moctezuma, E. Giraldo, and M. Molinas. Automated methodology for optimal selection of minimum electrode subsets for accurate eeg source estimation based on genetic algorithm optimization. *Scientific Reports*, 12(1):11221, 2022. 105

[206] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011. 40

[207] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui. Mped: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, 7:12177–12191, 2019. 103

[208] A. Soria-Frisch. A critical review on the usage of ensembles for bci. *Towards Practical Brain-Computer Interfaces*, pages 41–65, 2012. 34

[209] M. Spape, K. Davis, L. Kangassalo, N. Ravaja, Z. Sovijarvi-Spape, and T. Ruotsalo. Brain-computer interface for generating personally attractive images. *IEEE Transactions on Affective Computing*, 1(1), 2021. 21

[210] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946. 8, 38, 39

[211] L. Summer and H. L. Bonny. *Music consciousness: The evolution of guided imagery and music*. Barcelona Publishers, 2002. 40

[212] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 89

[213] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 31

[214] R. Tang, Z. Li, and X. Xie. Motor imagery eeg signal classification using upper triangle filter bank auto-encode method. *Biomedical Signal Processing and Control*, 68:102608, 2021. 27

[215] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012. 9, 14, 42, 43, 73, 103

[216] S. Tomkins. Affect, imagery, and consciousness: Vol. 1. the positive affects, 1962. 36

[217] R. H. Van der Lubbe, J. Sobierajewicz, M. L. Jongsma, W. B. Verwey, and A. Przekoracka-Krawczyk. Frontal brain areas are more involved during motor

imagery than during motor execution/preparation of a response sequence. *International journal of psychophysiology*, 164:71–86, 2021. 25

[218] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 81

[219] V. Vapnik. *Statistical learning theory*. Wiley, 1998. 33, 50

[220] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 26

[221] H. G. Vaughan Jr. The relationship of brain activity to scalp recordings of event-related potentials. In *National Aeronautics and Space Administration and the American Institute for Biological Sciences Conference, Sep, 1968, San Francisco, CA, US*. US National Aeronautics and Space Administration, 1969. 21

[222] J. J. Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157–180, 1973. 21

[223] J. J. Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65(5):633–641, 1977. 21

[224] C. Vidaurre and B. Blankertz. Towards a cure for bci illiteracy. *Brain Topography*, 23(2):194–198, 2010. 10

[225] F. C. Viola, S. Debener, J. Thorne, and T. R. Schneider. Using ica for the analysis of multi-channel eeg data. *Simultaneous EEG and fMRI: Recording, Analysis, and Application: Recording, Analysis, and Application*, pages 121–133, 2010. 21

[226] A. Vourvopoulos, D. A. Blanco-Mora, A. Aldridge, C. Jorge, P. Figueiredo, and S. B. i Badia. Enhancing motor-imagery brain-computer interface training with embodied virtual reality: A pilot study with older adults. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 157–162. IEEE, 2022. 106

[227] A. Vourvopoulos, C. Jorge, R. Abreu, P. Figueiredo, J.-C. Fernandes, and S. Bermudez i Badia. Efficacy and brain imaging correlates of an immersive motor imagery bci-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in human neuroscience*, 13:244, 2019. 106

[228] H. Wang, X. Wu, and L. Yao. Identifying cortical brain directed connectivity networks from high-density eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1489–1500, 2020. 103

[229] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 71

[230] R. W. Wang, Y.-C. Chang, and S.-W. Chuang. Eeg spectral dynamics of video commercials: impact of the narrative on the branding product preference. *Scientific reports*, 6(1):36487, 2016. 107

[231] Z. Wang, L. Chen, W. Yi, B. Gu, S. Liu, X. An, M. Xu, H. Qi, F. He, B. Wan, et al. Enhancement of cortical activation for motor imagery during bci-fes training. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2527–2530. IEEE, 2018. 106

[232] X. Wei, A. A. Faisal, M. Grosse-Wentrup, A. Gramfort, S. Chevallier, V. Jayaram, C. Jeunet, S. Bakas, S. Ludwig, K. Barmpas, M. Bahri, Y. Panagakis, N. Laskaris, D. A. Adamos, S. Zafeiriou, W. C. Duong, S. M. Gordon, V. J. Lawhern, M. Śliwowski, V. Rouanne, and P. Tempczyk. 2021 beetl competition: Advancing transfer learning for subject independence & heterogenous eeg data sets. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 205–219. PMLR, 06–14 Dec 2022. 6

[233] X. Wei, P. Ortega, and A. A. Faisal. Inter-subject deep transfer learning for motor imagery eeg decoding. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 21–24. IEEE, 2021. 35, 46, 100

[234] P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. 93

[235] I. Winkler, S. Haufe, and M. Tangermann. Automatic classification of artifactual ica-components for artifact removal in eeg signals. *Behavioral and brain functions*, 7(1):1–15, 2011. 21

[236] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, T. M. Vaughan, et al. Brain-computer interface technology: a review of the first international meeting. *IEEE transactions on rehabilitation engineering*, 8(2):164–173, 2000. 21

[237] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. 24

[238] D. Wu, B.-L. Lu, B. Hu, and Z. Zeng. Affective brain–computer interfaces (abcis): A tutorial. *Proceedings of the IEEE*, 2023. 107

[239] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu. Investigating eeg-based functional connectivity patterns for multimodal emotion recognition. *Journal of neural engineering*, 19(1):016012, 2022. 41

[240] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation

system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. 5

[241] J. Xie, G. Xu, J. Wang, M. Li, C. Han, and Y. Jia. Effects of mental load and fatigue on steady-state evoked potential based brain computer interface tasks: a comparison of periodic flickering and motion-reversal based visual attention. *PloS one*, 11(9):e0163426, 2016. 24

[242] D. Xu, M. Agarwal, F. Fekri, and R. Sivakumar. Playing games with implicit human feedback. In *Workshop on Reinforcement Learning in Games, AAAI*, 2020. 6

[243] L. Xu, M. Xu, Y. Ke, X. An, S. Liu, and D. Ming. Cross-dataset variability problem in eeg decoding with deep learning. *Frontiers in human neuroscience*, 14:103, 2020. 34, 66, 74, 78

[244] M. Xu, J. Han, Y. Wang, T.-P. Jung, and D. Ming. Implementing over 100 command codes for a high-speed hybrid brain-computer interface using concurrent p300 and ssvep features. *IEEE Transactions on Biomedical Engineering*, 67(11):3073–3082, 2020. 21

[245] D. Yadav, S. Yadav, and K. Veer. A comprehensive assessment of brain computer interfaces: Recent trends and challenges. *Journal of Neuroscience Methods*, 346:108918, 2020. 21

[246] G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 12(1):16–35, 2018. 11, 85, 109

[247] G. N. Yannakakis and H. P. Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015. 11, 85

[248] F. Yger, M. Berar, and F. Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1753–1762, 2016. 29, 30, 72, 102

[249] Y. Yu, H. Jiang, D. Bahri, H. Mobahi, S. Kim, A. S. Rawat, A. Veit, and Y. Ma. An empirical study of pre-trained vision models on out-of-distribution generalization. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 33, 101

[250] H. Yuan and B. He. Brain–computer interfaces using sensorimotor rhythms: current state and future perspectives. *IEEE Transactions on Biomedical Engineering*, 61(5):1425–1435, 2014. 25

[251] C. Zhang, Y.-K. Kim, and A. Eskandarian. Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014, 2021. 31

[252] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 34, 47, 78

[253] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan. Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network. *Neural Networks*, 136:1–10, 2021. 31, 61, 64, 66, 70

[254] S. Zhang, Z. Zhu, B. Zhang, B. Feng, T. Yu, Z. Li, Z. Zhang, G. Huang, and Z. Liang. Overall optimization of csp based on ensemble learning for motor imagery eeg decoding. *Biomedical Signal Processing and Control*, 77:103825, 2022. 28

[255] Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki. Temporally constrained sparse group spatial patterns for motor imagery bci. *IEEE transactions on cybernetics*, 49(9):3322–3332, 2018. 27

[256] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. 45, 91, 92, 93, 94, 103, 104, 105

[257] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 47, 78

[258] H. Zhu, D. Forenzo, and B. He. On the deep learning models for eeg-based brain-computer interface using motor imagery. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022. 33, 50

[259] G. Zoumpourlis and I. Patras. Covmix: Covariance mixing regularization for motor imagery decoding. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–7. IEEE, 2022. 61