Estimating learning-related mental states through brain and physiological signals.

Aurélien Appriou

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE : Mathématiques et Informatique

SPÉCIALITÉ : Informatique

Par Aurélien APPRIOU

ESTIMATING LEARNING-RELATED MENTAL STATES THROUGH BRAIN AND PHYSIOLOGICAL SIGNALS

Sous la direction de : Fabien LOTTE

Soutenue le 17 Décembre 2020

Membres du jury :

Mme. Sauzéon, Hélène	Professeur	Univ. Bordeaux	Président
M. Chevallier, Sylvain	Maître de conférence	Univ. Versailles SQY	Rapporteur
M. Zander, Thorsten	Professor	TU Brandenburg	Rapporteur
Mme ROY, Raphaëlle	Professeur associé	ISAE-SUPAERO	Examinateur
M. TANGERMANN, Michael	Associate Professor	Radboud Univ.	Examinateur

Abstract

Studying human learning is crucial: how can humans learn and what are their motivations to keep building-up knowledge? Every human is permanently learning to adapt to his environment and current generations now have to learn to use rapidly evolving technologies. As part of emerging technologies, research on brain-computer interfaces (BCIs) has become more democratic in recent decades, and experiments using electroencephalography (EEG)-based BCIs dramatically increased. This technology enables direct transfer of information from the human brain to a machine via brain signals, and can notably enable people with severe motor impairments to send commands to a wheelchair, e.g., by imagining left or right hand movements to make the wheelchair turn left or right. Such BCIs are called active BCIs since users are actively sending commands to the system, here a wheelchair, by performing mental imagery. However, the lack of robustness of BCIs limits the development of the technology outside of research laboratories, and current BCIs do not enable 10 to 30% of persons to acquire the skills required to use BCIs. If a lot of research has focused on the improvement of signal processing algorithms, the potential role of user training in BCI performance seems to be mostly neglected, and training protocols might not be suitable for all users. However, another type of BCIs recently proved particularly

promising: passive BCIs. Such BCIs are not used to directly control an application, but to monitor in real-time users' psychological states, e.g., mental workload or attention, in order to adapt an application accordingly. The goal of my PhD thesis is to to attempt to estimate learning-related psychological states such as cognitive workload, curiosity, attention, fatigue or emotions, from EEG and/or bio signals, using passive BCIs, in order to understand individual users' capabilities and motivations to learn, and therefore to adapt active BCIs training protocols accordingly.

In a first contribution, we explored recent machine learning algorithms that have shown to be promising for oscillatory-based MI-BCIs, but that have never been tested on oscillatory psychological states estimation, proposed new variants of them, and benchmarked them with classical methods to estimate both mental workload and affective states (Valence/Arousal) from oscillatory-based EEG signals. We studied these approaches with both subject-specific and subject-independent calibration, to go towards calibration-free systems. Our results suggested that a Convolutional Neural Networks (CNN) obtained the highest mean accuracy, although not significantly so, in both conditions for the mental workload study, followed by Riemannian Geometry Classifiers (RGCs). However, this same CNN underperformed in both conditions for the affective states study, when RGCs performed the best. As a second contribution, we implemented a Python library (BioPyC) to easily compare and benchmark both Signal Processing algorithms and Machine Learning algorithms for offline EEG and bio signals decoding. Based on an intuitive and well-guided graphical interface, four main modules allow the user to follow the standard steps of the BCI process without any programming skills 1) reading different neurophysiological signal data formats 2) filtering and representing EEG and biosignals 3) classifying them 4) visualizing and performing statistical tests on the results. This toolbox has been used for our 3 contributions detailed in this thesis. In a third contribution, we ran an experiment in which we used EEG, Heart Rate (HR), breathing and Electrodermal Activity (EDA) signals to measure the neurophysiologi-

cal activity of participants as they were induced into states of curiosity, using trivia question and answer chains. So far, results from the within-participant study attempting to classify EEG signals with five-fold stratified cross-validation returned classification accuracies oscillating around 60% (63.09% classification accuracy for the Filter Bank Tangent Space Classifier (FBTSC), 60.93% classification accuracy for the Filter Bank Common Spatial Pattern (FBCSP) + Linear Discriminant Analysis (LDA)). Moreover, analyses have been made concerning the classification of the bio signals (ECG, EDA and breathing): results showed interesting results since we obtained a classification accuracy of 58.45% for the classification of breathing signals by performing a LDA. Finally, a fifth contribution is underway, in which we will run an experiment in order to assess participants' cognitive load during MT-BCI training using EEG signals, in order to go towards adapting active BCI training to users cognitive workload capabilities.

Résumé étendu

L'étude de l'apprentissage humain est cruciale : comment les humains peuvent-ils apprendre et quelles sont leurs motivations pour continuer à accumuler des connaissances ? Chaque humain apprend en permanence à s'adapter à son environnement et les générations actuelles doivent maintenant apprendre à utiliser des technologies qui évoluent rapidement. Concernant les technologies émergentes, la recherche sur les interfaces cerveau-ordinateur (ICO) s'est démocratisée au cours des dernières décennies, et les expériences utilisant des ICO basées sur l'électroencéphalographie (EEG) ont considérablement augmenté. Cette technologie permet le transfert direct d'informations du cerveau humain à une machine via des signaux cérébraux, et peut notamment permettre aux personnes souffrant de graves handicaps moteurs d'envoyer des commandes à un fauteuil roulant, par exemple en imaginant des mouvements de la main gauche ou de la main droite de facon à faire tourner le fauteuil à gauche ou à droite. Ces ICO sont appelés ICO actifs car les utilisateurs envoient activement des commandes au système, ici un fauteuil roulant, en effectuant une imagerie mentale. Cependant, le manque de robustesse des ICO limite le développement de la technologie en dehors des laboratoires de recherche, et les ICO actuels ne permettent pas à environ 30 % des personnes d'acquérir les compétences requi-

ses pour utiliser les ICO. Si de nombreuses recherches se sont concentrées sur l'amélioration des algorithmes de traitement du signal, le rôle potentiel de l'entraînement des utilisateurs dans les performances des ICO semble être largement négligé, et les protocoles d'entraînement pourraient ne pas convenir à tous les utilisateurs. Cependant, un autre type de BCI s'est récemment révélé particulièrement prometteur : les ICO passifs. Ces ICO ne sont pas utilisés pour contrôler directement une application, mais pour surveiller l'état psychologique des utilisateurs en temps réel, par exemple la charge de travail mental ou l'attention, afin d'adapter une application en conséquence.

Le but de ma thèse de doctorat est d'essayer d'estimer les états psychologiques liés à l'apprentissage, tels que la charge de travail cognitif, la curiosité, l'attention, la fatigue ou les émotions, à partir de l'EEG et/ou de signaux physiologiques, en utilisant les ICO passifs, afin de comprendre les capacités et les motivations des utilisateurs à apprendre, et donc d'adapter les protocoles d'entraînement des ICO actifs en conséquence. En effet, cette thèse fait partie d'un projet de 5 ans appelé "BrainConquest", qui vise à améliorer l'entraînement des utilisateurs lors de l'apprentissage du contrôle des ICO de type "Mental-Task" (MT-ICO). Dans le cadre de cette bourse du European Research Council (ERC), plusieurs approches sont explorées, telles que l'amélioration du retour d'information qui permet aux utilisateurs d'avoir des informations sur leurs performances lors de l'exécution d'une tâche ICO en temps réel, ou la modélisation de la façon dont les utilisateurs apprennent à encoder les commandes ICO via l'EEG avec succès. L'approche que nous suivons dans cette thèse vise d'abord à identifier les états des utilisateurs qui devraient théoriquement être impliqués dans l'entraînement aux tâches ICO. Une fois que les états des utilisateur qui sont liés à l'apprentissage sont identifiés, l'étape suivante consiste à pouvoir les estimer en temps réel, afin que la tâche proposée puisse être adaptée à cet utilisateur par la suite. Certains de ces états mentaux ont été étudiés dans diverses études, comme la charge de travail cognitive, l'attention ou les émotions. Ces études ont été diverses, certaines

utilisant des mesures subjectives telles que des questionnaires. D'autres études ont estimé ces états mentaux par des mesures objectives, soit par l'activité cérébrale, par exemple les signaux EEG ou l'IRMf, ou par l'activité physiologique, par exemple le rythme cardiaque, l'EDA ou la respiration. Dans certains cas, l'utilisation d'ICO hybrides a été choisi pour combiner les signaux EEG et physiologiques afin d'estimer ces états mentaux.

Dans une première contribution, nous avons exploré les récents algorithmes d'apprentissage automatique qui se sont révélés prometteurs pour les MT-ICO basés sur l'oscillation, mais qui n'ont jamais été testés sur l'estimation des états psychologiques oscillatoires, proposé de nouvelles variantes de ceux-ci, et les avons comparés aux méthodes classiques pour estimer à la fois la charge de travail mental et les états affectifs (Valence/Arousal) à partir de signaux EEG basés sur l'oscillation. Nous avons étudié ces approches avec un calibrage à la fois spécifique au sujet et indépendant du sujet, pour aller vers des systèmes sans calibrage. Concernant la charge de travail cognitif, nos résultats suggèrent qu'un réseau de neurones convolutif (CNN) a obtenu la précision moyenne la plus élevée, bien que de manière non significative, dans les deux conditions de l'étude, suivi par les classifieurs utilisant la géométrie riemannienne (RGC). Cependant, ce même réseau de neurones a obtenu des résultats inférieurs dans les deux conditions pour l'étude sur les états affectifs, lorsque les RGC ont obtenu les meilleurs résultats. Comme deuxième contribution, nous avons mis en œuvre une bibliothèque Python (BioPyC) pour comparer et étalonner facilement les algorithmes de traitement du signal et les algorithmes d'apprentissage automatique pour l'EEG hors ligne et le décodage des biosignaux. Basés sur une interface graphique intuitive et bien guidée, quatre modules principaux permettent à l'utilisateur de suivre les étapes standard du processus ICO sans aucune compétence en programmation 1) lire différents formats de données de signaux neurophysiologiques 2) filtrer et représenter les signaux EEG et les biosignaux 3) les classer 4) visualiser et effectuer des tests statistiques sur les résultats. Cette boîte à

outils a été utilisée pour les 3 contributions détaillées dans cette thèse. Dans une troisième contribution, nous avons mené une expérience dans laquelle nous avons utilisé des signaux EEG, de fréquence cardiaque (HR), de respiration et d'activité électrodermique (EDA) pour mesurer l'activité neurophysiologique des participants lorsqu'ils sont induits dans des états de curiosité, en utilisant des chaînes de questions et réponses triviales. Jusqu'à présent, les résultats de l'étude sur les participants qui ont tenté de classer les signaux EEG avec une validation croisée stratifiée au quintuple ont donné des précisions de classification oscillant autour de 60 % (63,09 % de précision de classification pour le Filter Bank Tangent Space Classifier (FBTSC), 60,93 % de précision de classification pour le Filter Bank Common Spatial Pattern (FBCSP) + Linear Discriminant Analysis (LDA)). De plus, des analyses ont été effectuées concernant la classification des signaux biologiques (ECG, EDA et respiration) : les résultats ont montré des résultats intéressants puisque nous avons obtenu une précision de classification de 58,45 % pour la classification des signaux respiratoires en effectuant une LDA. Enfin, une cinquième contribution est en cours, dans laquelle nous allons mener une expérience afin d'évaluer la charge cognitive des participants pendant l'entraînement MT-ICO en utilisant les signaux EEG, afin d'aller vers l'adaptation de l'entraînement ICO actif aux capacités de charge cognitive des utilisateurs.

Remerciements / Acknowledgements

La légende disait donc vrai : "faire un doctorat, ce n'est pas simple tous les jours"... Mais contrairement à ce que l'on pourrait penser, cet effort, qui semble si solitaire de l'extérieur, dépend finalement d'un collectif bien huilé tout au long de ces 3 années faites de hauts et de bas. Un bon nombre de personnes m'ont accompagné dans cet effort, et je voudrais par ces mots les remercier infiniment.

Tout d'abord, je tenais à remercier les membres du jury, à savoir Thorsten Zander et Sylvain Chevallier pour avoir lu et émis des comentaires constructifs à propos du manuscrit, mais aussi Hélène Sauzéon, Raphaëlle Roy et Michael Tangermann pour avoir accepté d'être examinateurs de cette thèse.

Ensuite, je voudrais naturellement remercier mon directeur de thèse, Fabien Lotte, de m'avoir accepté sous sa tutelle et transmis tant de connaissances au cours de cette thèse. Sa vision et sa technique m'ont permis d'aborder rapidement et de manière cohérente des thématiques pluridisciplinaires, tout en me laissant beaucoup de liberté sur le choix des projets. Fabien, un grand merci pour tout !

Ces projets n'auraient pas été réalisables sans les nombreuses collaborations qui ont été engagées durant ce doctorat. I would therefore like to thank Andrzeij Cichocki, who supervised me during my visit at the Riken institute in Tokyo, Japan, Jessy Ceha and Edith Law from the University of Waterloo, Canada, with whom I shared intense brain storming sessions and more on the Curiosity project, and finally Vir-

ginia De Sa from the University of California, San Diego, for welcoming me in her lab for 3 months during my last year PhD. Je voudrais aussi souligner le travail de très grande qualité qui a été fourni par Smeety Pramij, David Trocellier et Léna Kolodzienski, avec qui j'ai pu collaborer sur différents projets.

Ensuite, je voudrais remercier toute l'équipe Potioc et l'excellente dynamique qu'elle a su émettre tout au long de mon passage à l'Inria: nul doute que les traditions se perpetueront et que nos chemins se recroiseront au SM ;) Thanks to all of you !

Puisque l'efficacité dans la sphère professionnelle passe tout d'abord par un équilibre personnel, je voudrais aussi remercier mon entourage pour m'avoir accompagné tout au long de ce doctorat, mais aussi des nombreuses années qui l'ont précédé. Ma famille a joué un rôle clef dans ce processus et m'a toujours encouragé dans mes projets de vie, qui, après un long chemin semé d'embûches, m'ont finallement conduit à ce titre de docteur. Merci à vous !

C'est aussi en toute logique que je voudrais étendre ces remerciements à mes amis, qui, sans aucun doute, ont aussi fortement contribué à cet aboutissement. Je lève mon verre aux Miaous, aux polos roses, à la CCS, aux Coronapéros, aux HSH, aux MLs, aux forains, aux Gigis, to my dear friends from SF (Bruhs, Diamonds and Butterflies) and from SD (S.) et individuellement à Pauline, Cazzo, Maxime, Arnaud et Santiago :)

Je voudrais finalement remercier ma plus grande supportrice, et saluer le soutien qu'elle m'a témoigné pour aller au bout de ce doctorat: Marion, tu m'as apporté tellement d'énergie positive en cette fin de thèse, merci pour tout !

À Clément, Marie, Joseph et Virginie

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context & Motivations

1.1.1 Brain-Computer Interfaces (BCI)

Research on brain-computer interfaces (BCIs) started in 1973 with Jacques Vidal and his concept of direct brain-computer communication (Vidal, 1973), enabling transfer of information from the human brain to a machine via brain signals, usually measured using Electroencephalography (EEG) (Clerc et al., 2016). Later, the BCIs will be redefined as a hardware and software communication and control systems that allow humans to interact with their surroundings without having to use their peripheral nerves and muscles, by using brain signals alone. This definition emerged from the reference paper entitled “Brain-Computer Interfaces for Communication and Control" (Wolpaw et al., 2002). In this case, the normal communication channels, such as speech and movement, are not used, but instead the brain activity is directly recorded and transformed into a control signal. To do so, features are extracted from the signal by applying signal processing algorithms, and are then used to feed machine learning algorithms in order to classify the signals into control commands. BCI systems have therefore two main components: (1) on the one hand, the user’s brain is used to encode commands through clear and distinct signals, so that they can be easily discriminated afterwards (2) on the other hand, the signal processing and machine learning algorithms must be as suitable as possible to the BCI task, in order to decode these commands in brain signals in an efficient way.

Today, restoring communication and control in severely paralysed patients is one of the major axes of BCI research (Ang and Guan, 2013; Pfurtscheller et al., 2008), but applications are also made for healthy individuals. Indeed, BCIs have also been developed for recreational purposes, for example when combined to gaming and virtual reality applications (Lecuyer et al., 2008; Lotte et al., 2012), or for more serious purposes, e.g., for monitoring cognitive states such as alertness (Zander and Kothe, 2011). Developing such BCI systems can be done using four main non-invasive techniques (Pfurtscheller et al., 2004): (1) functional Magnetic Resonance Imaging (fMRI), that relies on changes in the cerebral blood flow to measure brain activity (2) functional Near-Infrared Spectroscopy (fNIRS), that relies on near-infrared light to measure cerebral hemodynamic responses, and therefore brain activity (3) Electroencephalography (EEG), that relies on electrical activity recordings from the electrodes placed along the scalp to measure brain activity and (4) Magneto Encephalography (MEG), that records magnetic fields produced by electrical currents in order to map the brain activity. All these techniques have their advantages, e.g., a good spatial resolution for fMRI and fNIRS, and a good temporal resolution for EEG, but BCI systems are usually based on EEG signals (Clerc et al., 2016): as a result we choose to only focus on EEG-based BCIs in this thesis. Note that invasive BCI techniques such as Electrocorticogram (ECoG) also exist, and require to implant electrodes under the skull for extracting brain signals, and therefore imposes the user to have a surgical intervention beforehand. If the signals reading is more accurate than non-invasive systems, being a major benefit, many downsides such as side effects from the surgery have also been reported (Abdulkader et al., 2015).

Over the last forty years, BCIs have drastically evolved, and notable developments regarding the categorization of BCI systems have emerged: we now distinguish 4 types of BCIs, i.e., active, reactive, passive and hybrid. First, active BCIs allow users to send mental command to a system by using brain activity only. For example, this type of system would enable people with severe motor impairments to send commands to a wheelchair, e.g., by imagining left or right hand movements to make the wheelchair turn left or right (Millán et al., 2010). Such BCIs are called active BCIs since users are actively sending commands to the system, here a wheelchair, by performing mental imagery tasks (Zander and Kothe, 2011).

Concerning the reactive BCIs, external stimuli are used to trigger particular cerebral responses, called Event Related Potentials (ERPs), in users' brain activity. A concrete application of the use of these ERPs

is the P300 speller 1.1, which has been first introduced by Farwell and Donchin (Farwell, 1988). The P300 speller consists of a matrix of symbols (e.g., letters) displayed on a screen, and either the rows and columns of the matrix, either the symbols, light up in a random order. A positive cortical potential is induced in the user when the letter he is triggering flashes, and appears around 300ms after the occurrence, allowing the BCI system to detect the letter. This method allows users to communicate with brain activity only, but unlike active BCIs where the users send commands to the system, users expect here the system to react to their sensory inputs, e.g., through vision in the case of the P300 speller.

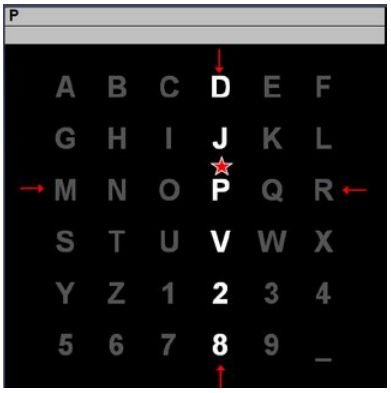

Figure 1.1: This is an example of a P300 speller interface.

Another type of BCIs proved particularly promising for Human-Computer Interaction (HCI): passive BCIs (Zander and Kothe, 2011). Such BCIs are not used to directly control some devices, but to adapt an application/interface by monitoring users' mental states, e.g., cognitive workload, emotion, attention or curiosity, in real-time. In other words, users do not voluntary interact with the BCI, i.e., they do not send commands to the system, but the system still extracts information from the users' EEG signals. For example, being able to estimate the level of a cognitive state such as the mental workload, when a user is performing a task, would allow the system to adapt the task to the user's cognitive abilities. The same would apply for adapting the task to the user's conative states (e.g., curiosity or intrinsic motivation) and/or affective states (e.g., frustration).

Finally, the concept of "hybrid BCIs" has been introduced by Pfurtscheller and al. in (Pfurtscheller et al., 2010), and describes BCI systems that combine information from different devices. Such hybrid BCI systems can be of two types: (1) two BCIs systems are combined together, i.e., the hybrid BCI uses two different brain signals, e.g., electrical and hemodynamic signals (2) one type of brain signals coupled with another input such as physiological signals, e.g., heart rate (HR), Electrodermal Activity (EDA), breathing or signals from other external device such as an eye tracking system. The goal of making such hybrid BCI systems is to reach better performances than classic single BCI systems.

1.1.2 *Passive BCI as a tool to adapt active BCI training to the user*

Although promising, non-invasive BCIs are still barely used outside laboratories due to their poor robustness with respect to noise and environmental conditions. In other words, they are sensitive to noise,

outliers and the non-stationarity of electroencephalographic (EEG) signals (Wolpaw and Wolpaw, 2012) (Erp et al., 2012). For example, based on (Blankertz et al., 2010), the average classification accuracy for a two-class MT-BCI experiment, with 80 users, is only 74.4%. In addition, it is estimated that between 10% and 30% of BCI users, depending on the type of BCI, cannot control the system at all. This phenomenon is also called BCI illiteracy/deficiency (Allison and Neuper, 2010).

So far, an important part of BCI research focused on computer machinery to address the described issues, as many signal processing and machine learning algorithms for brain signals classification have been developed (Allison and Neuper, 2010; Lotte et al., 2018a). If this research area has contributed to a slight increase in BCI performance, the classification accuracy is still relatively low and the BCI illiteracy/BCI deficiency is still high. BCI control is a skill that needs to be trained (McFarland and Wolpaw, 2018), and another axis of improvement has therefore been proposed (Jeunet et al., 2016; Lotte et al., 2013; Neuper and Pfurtscheller, 2010), consisting in focusing on the human side of the BCI systems rather than the machine side only. The idea is to ensure the user can produce clear, stable and distinct EEG signals, in order to then feed the machine learning (ML) algorithms with quality data. From this point of view, active BCI trainings take all their meanings, and like any acquisition of knowledge or know-how by humans, involves a learning phase that induces complex cognitive processes such as cognitive workload or attention. Moreover, other mental processes can also influence this learning, positively or negatively, such as the emotions or motivation that users may have during this learning.

Changes in mental states have been proposed as a cause of variation in BCI performance (Curran and Stokes, 2003; Millán et al., 2010). Myrden and Chau have been the first to formally test this hypothesis (Myrden and Chau, 2015) by investigating the effects of user mental states on BCI performance, and more particularly cognitive fatigue, frustration and attention. In particular, authors showed that there is a link between frustration and performance. They also showed that moderate fatigue is the best way to concentrate effortlessly on controlling BCI, as well as frustration would be a potential motivating factor, and attention a compensatory mechanism for growing frustration. If this study was based on subjective questionnaires, other studies have suggested that mental state fluctuations such as alertness or distraction can also influence BCI performances (Curran and Stokes, 2003; Millán et al., 2010), and proposed to estimate them using objective measures, i.e., passive BCIs. Other mental states that could influence performance have been studied using passive BCI as well, e.g. engage-

ment and cognitive workload (Berka et al., 2007; Gerjets et al., 2014), fatigue (Shen et al., 2008), and emotions (Mühl et al., 2014). For example, being able to estimate a mental state such as cognitive workload is essential to make sure the cognitive load induced by the BCI training task is adapted to the learner, i.e., that his/her working memory is never overloaded nor underloaded. Indeed, such an unsuitable cognitive load would impact the quality of learning (Sweller et al., 1998). Note that some variations in users' mental states can also impact their other mental states' estimation. For instance, it has been showed that changes in users' affective states (and more precisely variations in the stress level of the users) can impact workload estimation (Mühl et al., 2014).

1.1.3 *Passive BCIs for Human-Computer Interaction*

The use of passive BCIs, including users' mental states monitoring in real time, is not only useful for upgrading BCI training protocols, they can also be used as an evaluation method for Human-Computer Interaction (HCI). As its name indicates, HCI is a field of research that focuses on design, uses and therefore the interaction that users can have with computer technologies (Hewett et al., 1992). As computers complexity increases, it is important to analyze the impact of such interactions on the users, in order to be able to keep the potential of these technologies and adapt them to human capabilities, whether cognitive or otherwise. In order to better understand these interactions between humans and machines, numerous evaluation methods have been used in recent decades. Indeed, these evaluation methods have had to adapt as technologies have developed, and the same is true today. The methods that are currently used include subjective measures, e.g., questionnaires, as well as behavioural measures, e.g., reaction time, error rate.

More recently, new methods using physiological sensors have been developed, keeping the idea of improving ergonomics of HCI in mind (Fairclough, 2009). This field, known as "physiological computing" aims at extracting real-time information about users' states from physiological data such as heart rate, electrodermal activity or breathing. For example, emotions (Villon and Lisetti, 2006), or workload (Fairclough and Houston, 2004) are users' mental states that have been estimated through such physiological signals.

Passive BCIs can be used as an evaluation method for human-computer interaction as well (Frey et al., 2014), but also in a more global way for human factors issues in daily tasks, e.g. work performance or operational safety, that does not necessarily requires to

interact with a device. This field, examining the “brain at work", and known as “neuroergonomics" (Ayaz and Dehais, 2018; Parasuraman and Wilson, 2008), uses electroencephalography, or other neuroimaging methods, to measure users' states, possibly in real time. For example, mental workload has been widely studied to assess how cognitively difficult the manipulation of a given input device is (Frey et al., 2016; Gerjets et al., 2014). This same cognitive workload has been assessed during 3D objects manipulation tasks (Wobrock et al., 2015), navigation tasks with different input devices (Frey et al., 2016), during visualization tasks (Peck et al., 2013), during plane piloting (Gateau et al., 2015) or during the execution of tasks requiring a lot of cognitive resources by human operators in nuclear power plants (Choi et al., 2018). Workload estimation was also used to design applications that dynamically adapt to the users' states, for instance to create video games with adaptive difficulty (Fairclough, 2008), to provide an optimal sequence of teaching exercises adapted to the cognitive capabilities of each learner (Yuksel et al., 2016), or to enable users to visualize and reflect on their own mental workload levels (Gervais et al., 2016).

1.2 Approach

1.2.1 Story

This thesis is part of a 5 years project called “BrainConquest", that aims at improving users' training when learning Mental-Task BCIs (MT-BCIs) control. As part of this European Research Council (ERC) grant, several approaches are explored, such as improving the feedback that allows the users to have information about their performance when executing a BCI task in real time, or modeling how users learn to successfully encode BCI commands in the EEG. The approach we follow in this thesis aims at first identifying the users' states that should theoretically be involved in BCI tasks training (see 1.1.2). Once these learning-related user states are identified, the next step is to be able to estimate them in real time, so that the proposed task can be adapted to the user afterwards. Some of these mental states have been studied in various studies, such as cognitive workload, attention, or emotions. These studies have been diverse, some using subjective measures such as questionnaires. Other studies have estimated these mental states through objective measures, either through brain activity (see introduction - passive BCI), e.g. EEG signals, fMRI, or through physiological activity, e.g. heart rate, EDA or breathing (see introduction - physiological computing). In some cases, the use of hybrid BCIs (see

introduction - hybrid BCIs) have been used to combine EEG and physiological signals to estimate these mental states.

1.2.2 *In a nutshell*

This thesis, entitled "Estimating learning-related mental states through EEG and physiological signals", aims at estimating different types of mental states, i.e., cognitive (process of coming to know and understand), conative (personal, intentional and motivational drives to process the information) and affective states (emotional interpretation of perceptions, information, or knowledge), which are known to influence learning. To do so, both physiological and EEG data are processed in order to decode such users' states. Altogether, this PhD thesis contributes new tools and knowledge to estimate learning-related mental states through EEG and physiological signals. In particular, it contributes new software and machine learning tools, new protocols and new neurophysiological knowledge to do so. We here present the outline of this thesis step-by-step, following Figure 8.4.

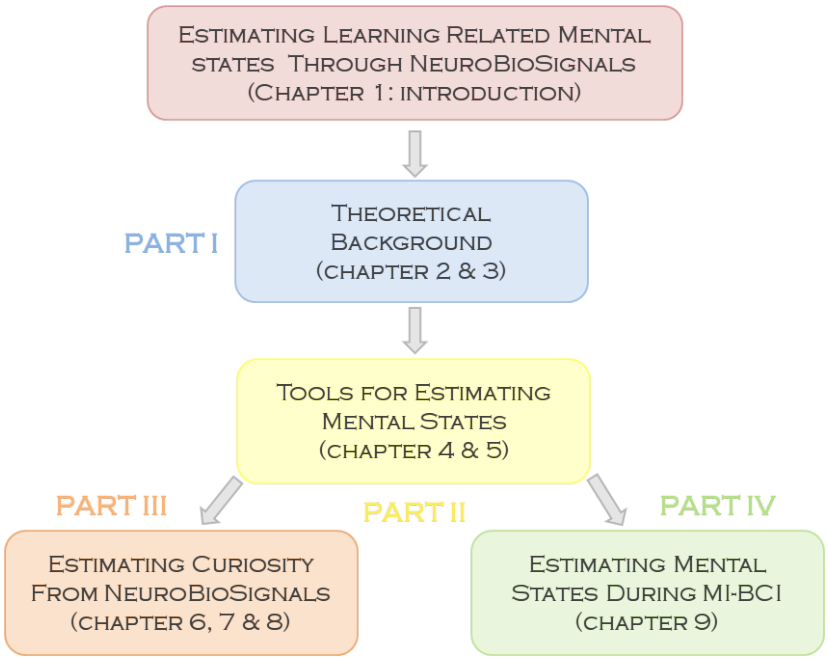

Figure 1.2: PhD thesis roadmap.

Part 1 of this thesis, i.e., the *Theoretical Background* (see Part I), is split into two chapters. The first chapter aims at first studying the historical background (see chapter 2) of the field of Psychology in order to identify which cognitive, conative and affective states are involved in human learning. Still in this part 1, we then propose a state-of-the-art (see chapter 3) of the works that have attempted to estimate these mental states through EEG signals. We notably survey the pro-

tocols that have been used to induce these mental states, e.g., N-back task for the cognitive workload, before reviewing the signal processing and machine learning algorithms that have been used, and that have proven to be effective, to estimate such mental states.

Part 2 concerns the *methods and tools used to classify learning-related mental states through EEG signals*, and is split into two chapters as well. The chapter 3 presents our ***contribution #1***, aiming at studying modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals. This study explores recent machine learning algorithms, such as Riemannian geometry based classifiers (RGC) or convolutional neural networks (CNN) that have shown to be promising for other BCI systems, proposes new variants of them, and benchmarks them with classical methods to estimate both mental workload and affective states (Valence/Arousal) from EEG signals. Then, the fourth chapter of part 2 describes out BioPyC, our ***contribution #2***, a free, open-source and easy-to-use Python platform that we designed and developed for offline EEG and biosignal processing and classification. This toolbox has been developed for many purposes, including studying the machine learning algorithms that would perform the best to estimate users' states, either through EEG or physiological signals.

Part 3 (see III), aiming at estimating a conative state that has been poorly explored through neurophysiological signals so far, but is very relevant for learning, i.e., *curiosity*. We present three chapters explaining our ***contribution #3*** in depth. In the fifth chapter, we introduce the protocol design that has been used to induce different levels of curiosity. Then, chapters 6 and 7 describe the methods that have been used for decoding curiosity from EEG signals and physiological signals, respectively.

Part 4 (see IV), corresponding to our ***contribution #4***, consists of a single chapter as it reflects the work that have been done so far on the ongoing project aiming at estimating levels of cognitive workload from EEG during MT-BCI user training, to which this thesis contributes. This eighth chapter describes out the protocol design, i.e., both materials and methods, and the current status of the project at the time of the PhD thesis submission.

Finally, a discussion and prospects are proposed in the last chapter.

Part I

Theoretical Background

PhD Thesis Roadmap

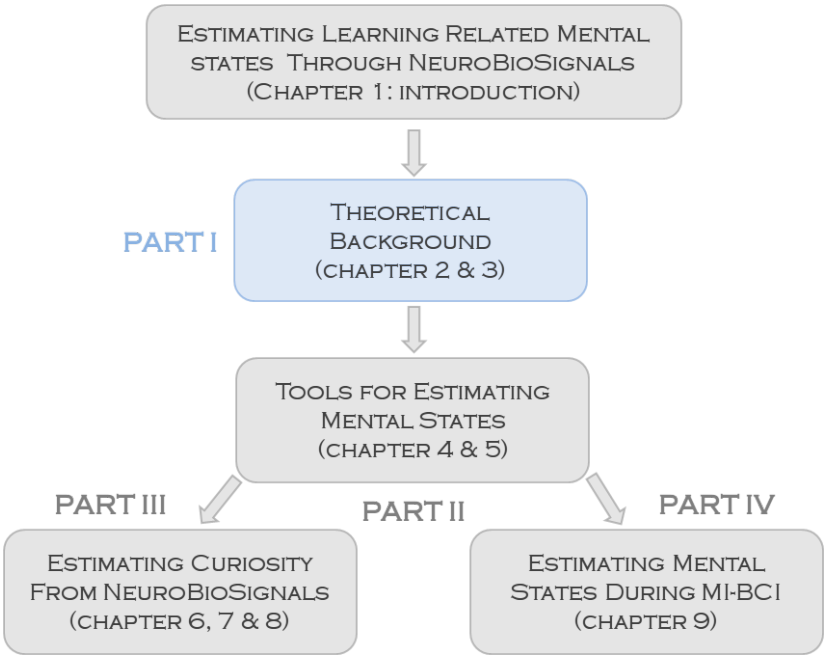

Figure 1.3: PhD thesis roadmap.

Related Papers

Peer-reviewed Journals

- Appriou, A., Pillette, L. and Lotte, F. Estimating Cognitive Workload, Attention and Affective States through EEG signals: A Review. (***Being written with the future aim to submit to the journal IEEE Transactions on Affective Computing)***

2

A survey of methods and tools to decode learning-related mental states from EEG

As explained in the introduction, the goal of this thesis is to propose tools to estimate users' states strongly related to learning through neurophysiological signals. These states, such as cognitive workload (Sweller et al., 1998) or curiosity (Kang et al., 2009), are regularly called "mental states", "psychological states" or even "psychological factors" in the literature. Moreover, the definition of "mental state" is quite vague, and is linked to several psychological theories where this very concept might sometimes differ. We therefore first attempt to study the history of psychology in order to list the users' states that are related to learning. The second step focuses on defining the mental states we decided to study in this thesis. Finally, in a third part of this chapter, we study the different existing methods that have proven to be efficient to induce such mental states.

2.1 Historical background

Psychology is a field that has been studied by many cultures over the last millennia, as the Ancient Greeks already studied it, and the

evidence of psychological thought dates back to the ancient Egypt (Wickramasekera, 2014). Psychology remained a branch of the domain of philosophy until the 1870s. This field has naturally become an independent discipline that studies human's mind, behaviors and mental processes. In 1879, Wilhelm Wundt founded the first laboratory for experimental studies in Leipzig, Germany (Rieber and Robinson, 2001). Several other individuals also made important contributions to found the field of psychology, such as Hermann Ebbinghaus (a pioneer in the study of memory) (Roediger, 1985) and Ivan Pavlov (who developed the procedures associated with classical conditioning) (Clark, 2004). Then, for decades, the field of psychology has dramatically developed with the emergence of new kinds of applied psychology such as the educational theory (John Dewey (Sikandar, 2016)), the behaviorism (Watson (Calkins, 1913) and then Skinner (Catania, 2003)), the psychological clinic (Lightner Witmer (Mertin, 2012)) or the psychoanalysis (Sigmund Freud (Frosh, 2012)).

Among those pioneers in the field of psychology, William James defined psychology as "the science of mental life" (in his Principles of Psychology (Cresswell et al., 2017)), thus classifying the field of psychology as a science. The goal of this science is therefore to describe and explain states of mind or, as more recently detailed, the mental structures and processes underlying human experience, thought, and action. The first explanations of what the mind could do were provided by the German philosopher Immanuel Kant in his book "Critique of Judgment" (Rothbart and Scherer, 1997), where he described the mental states as the thoughts, feelings, and desires that pass through our minds when we are conscious.

In the 1980th century, the psychologist Hilgard kept the idea that the mind has three main "faculties" to create the "Trilogy of Mind" (1980) (Hilgard, 1980), and described them as follow:

- ***cognition***: refers to the process of coming to know and understand; of encoding, perceiving, storing, processing, and retrieving information. It is generally associated with the question of "what" (e.g., what happened, what is going on now, what is the meaning of that information). These mental processes are sub-divided into cognitive states such as cognitive workload or attention.

- ***affect***: refers to the emotional interpretation of perceptions, information, or knowledge. It is generally associated with one's attachment (positive or negative) to people, objects, ideas, etc. and is associated with the question "How do I feel about this knowledge or information?"

- ***conation***: refers to the connection between knowledge and behavior, or between affect and behavior, and is associated with the "why." It is the personal, intentional, planful, deliberate, goal-oriented, or striving component of motivation. The conation groups together mental processes that relate to wanting, intending, or trying to do something (Militello et al., 2006). Conation is often described as a synonym of motivation.

The categorisation of mental states into three main faculties reveals the broad scope of psychology, and the complexity of how human minds work. A wide range of applications arise from this knowledge of the tripartite psychology, such as psychotherapy (prevention and treatment of mental illness), workplace (enhancing worker productivity and managerial effectiveness), but also education (enhancing teaching and learning). For each of these applications, the three faculties, i.e. cognition, affect and motivation, are involved, but also interact with each other, e.g. emotional states can serve as filters on cognition or vice versa cognitive processes can be employed in emotional self-regulation.

Many theories about learning have been developped in the history of psychology, and most of them remained assumptions until the late 1950s. Following this period, the understanding of humans and their environment became more and more complex, and a new field emerged: the cognitive science. This new discipline quickly approached learning from a multidisciplinary perspective that included anthropology, linguistics, philosophy, developmental psychology, computer science, neuroscience, and several branches of psychology (Newell and Simon, 1972; Norman, 1980). This new field has in particular made it possible to test the theories instead of just speculating about thinking and learning (Ericsson and Charness, 1994; Newell and Simon, 1972).

2.2 *Cognitive, affective and conative states & learning*

As shown previously, the trilogy of mind is a theory that comprises three main states, i.e., cognitive, conative and affective states. Early in the history of psychology, and in a stream of thought deriving from Kant's one, most learning theories focused on the cognitive dimension to learning while neglecting the emotional one. Only later, the affectional dimension, and then the conative dimension, will also be taken into account in some learning theories. The relationship between each

of these three states and learning are about to be detailed in the following steps.

2.2.1 *Cognitive states and learning*

Cognitive states are processes that enable the recognition and acquisition of information. In other words, the cognition is the accumulation of information acquired through learning or experience. To do so, information is received from different sources - i.e., perception, experience, beliefs, etc - and converted into knowledge. Among the cognitive processes, we find attention, language use, memory, learning, perception, problem solving (reasoning), and thinking (Newen, 2015). These procedures, when used together, allow to integrate knowledge and interpret the environment (Newen, 2015).

However, cognitive processes needs to be differentiated from cognitive states, even if cognitive states are also related to the acquisition of knowledge, i.e., learning, and the interpretation of the environment, and might also be processes. For example, the Cognitive Load Theory (CLT) perfectly illustrates the relationship between a cognitive state, here the cognitive load, and learning (Sweller et al., 1998), as many researchers have been using it to analyze the effect of cognitive load on learning (Paas et al., 2010). Moreover, some of these researchers developed tools to optimize the load level in various learning contexts (Paas et al., 2003). This cognitive load is thus considered to be a cognitive state in (Lohani et al., 2019), as well as fatigue, attention, distraction and stress. If three types of fatigue exists, i.e., sleep deprivation, physical fatigue and mental fatigue (Gawron et al., 2001), an increase of mental fatigue is associated to a decrease of learning (Gonzalez et al., 2011). Attention (Stadler, 1995), and particularly selective attention (Jiang and Chun, 2001), is a cognitive state that plays an important role in learning as well. Indeed, an increase of attention will be associated with learning. Indeed, "Attention" is a generic word which encompasses a set of different states. The number and characterisation of these different states differ between the different models that were developed over the years (Knudsen, 2007). In our case, the model states four types of attention, i.e., alertness, sustained attention, selective attention (related to learning though) and divided attention. An increase in distraction will have a negative impact on learning, and is therefore a cognitive state that is opposed to attention (Craik, 2014). Stress is a cognitive state that proved to enhance memory formation and therefore feed long term memory, but conversely proved to impair memory retrieval (Vogel, 2016). However, both these effects show an association between stress and learning.

Recently, in (Dirican and Göktürk, 2011), authors consider the following states to be cognitive states: attention, engagement, working memory, stress and fatigue. If we previously showed that most of these cognitive states are related to learning, the cognitive engagement is also associated to learning in educational psychology (Richardson and Newby, 2006). Note that this state of cognitive engagement might sometimes be defined as a motivational state as well, and would therefore be classified as a conative sate (Blumenfeld et al., 2006). In (O'Brien and Meister, 2002), the list is different but groups together cognitive states as well: workload and fatigue.

In summary, cognition is a group of processes that are involved in the recognition and acquisition of information, in other words, learning. Cognition encompasses several cognitive processes - attention, language, memory, learning, perception, as well as thinking - but also several cognitive states (some of which overlap with cognitive processes) related to learning - workload, fatigue, stress, engagement, distraction and attention.

2.2.2 *Affective states and learning*

In the late 20th, some theories recognized the importance of emotions in learning processes, e.g., Bloom (Bloom, 1975) who proposed two taxonomies for his theory: one for the cognitive domain and another one for the affective domain. However, there is no link or interdependence between cognition and emotion in Bloom's theory. A couple of years later, Goleman et al. developed their theory of emotional intelligence, this time drawing a parallel between cognition and emotion (Faltas, 2016). It was then shown that emotions do influence learning, either positively (facilitating learning through the development of emotional competence) or negatively (learning can be inhibited by emotional incompetence) (Shelton, 2000).

Finally, Damasio's work allowed to understand the neurological bases of emotions as well as the strong link between emotions and certain cognitive processes such as attention or memorization, that are strongly involved in learning (Damasio, 1994). Several studies have focused on the interaction between affective states and learning (Baker et al., 2010; D'Mello et al., 2013; D'Mello et al., 2008; Kort et al., 2001). Indeed, positive affective states, e.g. surprise, satisfaction, and negative affective states, e.g., frustration and disillusionment can respectively contribute towards learning or undermine learning. Another affective state, i.e., boredom, has been associated with poor learning (Baker et al., 2010). Finally, the flow, first described by Csikszentmihalyi (1990) (Csikszentmihalyi, 1990), has been proved to influence learn-

ing as well, by heightening engagement. Altogether, the affective states related to learning can be listed as follows: emotions, surprise, satisfaction, frustration, disillusionment, boredom and the flow. Note that emotions can be split into sub-states, positive or negative, e.g., valence, arousal, anxiety, anger, etc. More complex affective states such as mood or compassion could be added to the list.

2.2.3 *Conative states and learning*

Conation, often described as a synonym of motivation, is closely associated with the concepts of intrinsic motivation, volition, agency, self-direction, and self-regulation (Lemos, 2011; Mischel, 1996), all of them being associated with learning. First, Bronson proposes the idea that self-regulation corresponds to the individual learning to interact with his or her environment (Bronson, 2000). Concerning motivation, it was first defined as a psychological state or disposition that determines the initiation, vigor, or direction of cognitive behaviors or activities, and sets the value placed on various elements of the environment (Le Ny, 1994). In particular, motivation plays an key role in academic learning as well (Zimmerman and Risemberg, 1997), or in language learning (Azarnoosh and Tabatabaee, 2008). For Tardif, the transformation of information into knowledge is particularly cognitively demanding and requires a high rate of investment by the individual, i.e., a high intrinsic motivation (Tardif, 1992). To summarize, the following conative states are related to learning: the intrinsic motivation, volition, agency, self-direction, and self-regulation.

Once the mental states related to learning are identified, the next step is to understand how these mental states may vary, and how to estimate them objectively, in real time, through neurophysiological measurements.

2.3 *Defining Cognitive Workload, Emotions and Curiosity*

As indicated previously, the history of psychology allowed the birth of multiple theories, some of which have been examined more closely by cognitive science. Among them, the trilogy of the mind has notably proposed to divide the human mind into three parts, allocating three types of states to humans: cognitive states, affective states and conative states. These states are themselves subdivided into different substates, i.e., “mental states". Once again, in this thesis, we focus only

on the mental states (and thus sub-states of cognitive, affective and conative states) that are related to learning. As we have seen, many mental states related to learning have been defined, e.g., workload, fatigue, stress, engagement, distraction and attention with respect to cognitive states. The affective states include emotions, surprise, satisfaction, frustration, disillusionment, boredom and the flow, when the conative states comprise the intrinsic motivation, volition, agency, self-direction, and self-regulation.

Once again, we choose to focus on three of these mental states, strongly related to learning, in this thesis. They represent the three types of states from the trilogy of mind, i.e., cognitive, affective and conative states, and that will be described in more details in the following steps. First, we choose to study the cognitive workload among the cognitive states, as it has been widely studied through EEG signals (Antonenko et al., 2010), the literature about the estimation of the cognitive workload through EEG signals is consequent (a literature review is presented in chapter 3) and recommendations about the materials to use to continuously assess cognitive workload in real time have been proposed by Gerjets et al. (Gerjets et al., 2014). Concerning the affective states, we choose to focus on emotion as the literature about the estimation of such a mental state is important as well (Muhl et al., 2015), and is completed in chapter 3. Finally, no conative state has been studied to any great extent with EEG, which is why we have decided to focus on curiosity, a.k.a. intrinsic motivation, as it is a conative state that is strongly related to learning, and particularly on epistemic curiosity as Loewenstein described it as a desire to acquire knowledge, i.e., to learn (Loewenstein, 1994).

Now that we explained the reasons why we choose to focus on the estimation of these three mental states, i.e., cognitive workload, emotions and curiosity, in this thesis, it is important to follow strict steps to carry out these estimates successfully. The first step is to induce these mental states correctly - in order to obtain a ground truth - and multiple ways to do it are possible: concerning our three mental states, we will describe the existing methods to induce them in section 2.4. Then, the way to measure them is also important, and is usually either subjective measures (e.g., questionnaires), or objective measures (e.g., EEG signals). The type of measurement is explained in section 2.5. Finally, the estimation of the state is made by classifying data in order to discriminate between levels of this very state. The method to do it may vary based on the type of data. In this thesis, we mainly use EEG signals to estimate levels of such mental states, and both signal processing methods & machine learning algorithms are necessary. We

will detail these processes in section 2.6 and 2.7, respectively.

Before describing the steps aiming at estimating our three mental states, we first define these three mental states in details in the next section. Then comes the section that describes out the existing methods to induce such states, the section focusing on their measurements, and finally the section explaining the methods to estimate them through EEG signals.

2.3.1 *Cognitive Workload*

The concept of cognitive workload has undergone several definitions in recent decades, such as the Wickens' one, that defined it as a "relation between the (quantitative) demands for resources imposed by a task, and the ability to supply those resources by the operator" (Wickens, 2002). Other definitions have then been given, for example by Rozado et al. (Rozado and Duenser, 2015), where cognitive workload is identified as observed delays in information processing capabilities when a considerable amount of mental effort is exerted by an individual. For Haapalainen et al., the description of the cognitive workload remains close but more specific than the Wickens' one (Haapalainen et al., 2010), where cognitive workload also concerns the learning, thinking and reasoning as indicators of pressure on working memory, and mainly the associated efforts perceived by the user during the execution of a task. The measure of mental workload would therefore represent the interaction between task processing demands and human capabilities or resources (Hancock and Chignell, 1986; Valdehita et al., 2012). In 2007, Cain made a literature review about the mental workload, and gave the following definition: "a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands" (Cain, 2007).

In 2014, Durantin et al. brought a statement that docuses more on the cognitive resources in (Durantin et al., 2014): "when a task is performed, it would consume cognitive resources. According to this view, performance and the quantity of cognitive resources invested follow a linear relationship. When the quantity of cognitive resources consumed reaches an individual's maximum capacity, performance peaks (or even declines)". Moreover, research has also shown that too high levels of cognitive load condemns an individual to reach the limits of his or her cognitive abilities, leading to suboptimal decisions and human errors. Even if the cognitive demand remains reasonable for

the individual, a prolonged mental activity also leads to depletion of cognitive resources (Kamzanova et al., 2014). Low cognitive load can also lead to boredom, and thus error and distraction from surrounding factors.

In (Debie et al., 2019), authors pushed forward and defined this mental state as “a multidimensional concept that consists of four components: 1) task complexity; 2) mental workload; 3) performance; and 4) depletion factors". The first component, i.e., task complexity, refers to the difficulty of the task to be performed by the user, and has properties inherent to the task independently of the user involved in the task. Then, the mental workload component is the level of mental resources that a user is capable of giving to maintain a high-performance while performing the task. The performance component represents the interaction between users' mental workload and task complexity. Finally, depletion factors are external factors affecting the users' mental capacity, such as stress, fatigue, motivation, task importance, and attitude.

These definitions all overlap, but each of them brings more accurate details about the way to define the cognitive workload. Theories are also important to consider, particularly the on from Sweller et al. called "Cognitive Load Theory" (CLT) (Sweller et al., 1998), which is now unanimously considered as the theory to be put into practice when studying cognitive workload (Sweller et al., 2019). Following this theory, the cognitive load corresponds to the amount of cognitive/working memory resources that are necessary to process the information, and is split into 3 categories, i.e., intrinsic load, extraneous load and germane load. The intrinsic load is the resources necessary to process the inherent complexity of the content in the task, and the extraneous load the resources necessary to process the information of the irrelevant information. The use of these resources can be considered as a burden for learning. Finally, the germane load corresponds to the “good amount of resources needed to learn". Germane load occurs when information presentation is designed to encourage assimilation or accommodation of new concepts and appropriately challenge the learner. The goal of cognitive load theory is therefore to reach the germane load by playing with both the intrinsic load, i.e., by augmenting the amount of resources to process the content, and the extraneous load, i.e., by reducing the amount of resources needed to process every other information than the content. In other words, the germane load is the use of resources we wish to reach when we design training/learning material.

The CLT is therefore known to be strongly related to learning ma-

terials because it presents different recommendations for the design of instructions to facilitate humans' learning (Wiebe et al., 2010). Indeed, many studies on instruction design showed that, for similar complexity, learning will be better if the instructions are adapted and adaptive to the level of the learner (Lespiau and Tricot, 2018). Based on the CLT, the user must therefore use the cognitive resources related to the intrinsic activity, and use as few resources related to extrinsic information as possible, in order to have an optimal learning (Sweller et al., 2019).

As a result of all these statements, the cognitive workload is a very useful mental state to study as soon as human works, human performances or human interactions with devices are discussed. Moreover, it would be an important state to manipulate when we talk about human learning (Antonenko et al., 2010). Typically, the cognitive workload is measured using subjective measures (i.e., questionnaires) or behavioural measures (i.e., the actions of individuals) (Antonenko et al., 2010). However, more recent research focuses on measurement of such as user state using peripheral physiological activity (e.g., heart rate variability (Paas and Van Merrienboer, 1994), pupil dilatation (Van Gerven et al., 2004)), and more importantly brain activity (e.g., EEG) (Antonenko et al., 2010). Peripheral physiological measures, however, present a problem: as we saw previously, cognitive workload can be divided into several sub-components, and very few physiological measures can accurately distinguish the influence of each of these sub-components. These sub-components, such as mental effort, stress or fatigue, may have different implications. Thus, in a real work context, an excess of mental effort will not be regulated in the same way as an excess of fatigue.

2.3.2 *Affective States*

As we explained in the introduction of this thesis, the affective states that are related to learning include several mental states, such as emotions, but also the surprise, satisfaction, frustration, disillusionment, boredom and the flow. Defining and clusterizing each of these affective states - and more precisely the emotions dimensions we will study in more depth in this thesis (see chapter 4) - therefore remains challenging (Muhl et al., 2015). There are 2 main approaches to define emotion classes: discrete approaches, with "emotion families" such as happiness, anger or surprise (Ekman, 1992); and dimensional approaches, which group the different emotions along several dimensions. The most popular model, and the one used in our study in chapter 4, is the circumplex model of Russell (Russell, 1980), which

assumes that any affective state can be localized on a two-dimensional plane. The first axis of this plane is the valence, ranging from positive feelings to negative ones, and the second axis represents arousal, ranging from calm to excited. Note that more discrete approaches are existing, proposing different sets of emotions (Ekman, 1992), or the eight basic emotion states (anger, fear, sadness, disgust, surprise, anticipation, acceptance, and joy) proposed by Slama (Slama, 2005), as well as more dimensional models, e.g., the one including dimensions such as dominance and unpredictability (Russell and Barrett, 1999).

The fact that there is no consensus regarding the definition of emotions, or even regarding the model that represents the way to conceptualize emotions, whether discrete or dimensional, makes it difficult to measure them.

2.3.3 *Epistemic curiosity*

Philosophers and psychologists such as Cicero, Kant, and Freud characterized curiosity respectively as “innate love of learning and knowledge", “appetite for knowledge" and “thirst for knowledge", but the first to propose a definition of curiosity was William James, who said that “curiosity is an instinct that evolved to facilitate survival and adaptation through active exploration of the environment" (Gruber and Valji, 2019). Daniel Berlyne then introduced a multi-dimensional model of curiosity, characterizing this psychological state in two dimensions, i.e., perceptual/epistemic curiosity, and specific/diversive curiosity (Berlyne, 1954). Following this theory, perceptual curiosity refers to “a drive which is aroused by novel stimuli and reduced by continued exposure to these stimuli", whereas epistemic curiosity refers to “a desire to acquire knowledge, and applies mainly to humans" (Loewenstein, 1994). On the second dimension, specific curiosity is defined by the desire for a particular piece of information, whereas diversive curiosity is defined as a more general seeking for stimulation that is related to boredom. In addition to Berlyne’s theory, multiple other theories have emerged in the last half-century, e.g., incongruity and information gap theories (Loewenstein, 1994) and the optimal arousal theory (Berlyne, 1967).

Although no scientific consensus has been reached concerning the definition of curiosity, certain types of stimuli are known to trigger it, i.e., those of surprising, novel, or intermediate complexity, as well as activities that are characterized by a knowledge gap or errors in prediction (Oudeyer et al., 2016). Such triggers can lead to different momentary states of curiosity, including epistemic (Day, 1970; Litman, 2012; Litman et al., 2005; Macedo and Cardoso, 2012; Mussel, 2010) and

perceptual curiosities (Gruber and Valji, 2019). Epistemic curiosity has been particularly studied through psychological experiments in the last decades (Brod and Breitwieser, 2019), and has more recently been the object of neuroscientific experiments.

However, epistemic curiosity is an important mental state associated with spontaneous exploration, active learning, facilitated memorization and sustained engagement (Oudeyer et al., 2016). Recent research in psychology (Kidd and Hayden, 2015) and neuroscience (Gottlieb and Oudeyer, 2018) has shown its pervasive role across multiple dimensions of human cognition and learning. Curiosity also has applications in multiple disciplines, including robotics, human-computer interaction (HCI), and learning technologies. The design of robotic and computer architectures can also be guided by our understanding and conceptualizations of curiosity (Gordon et al., 2015; Oudeyer et al., 2007). Human-computer interaction (HCI) researchers have investigated how systems can elicit curiosity in order to inform the design of persuasive, engaging, and playful interactions (Tieben et al., 2011), or to motivate and incentivize crowd workers (Law et al., 2016). Finally, as curiosity-driven learning has been argued to be a crucial feature for efficient education (Freeman et al., 2014; Oudeyer et al., 2016), learning technologies are being developed to promote curiosity and motivate curiosity-driven behaviours in students (Ceha et al., 2019; Lomas et al., 2017).

2.4 *Inducing learning-related mental states*

In the literature, most of the mental states are mainly studied - and therefore induced - in laboratory set-ups. More importantly, the materials used to induce those mental states are often similar from one study to another. Indeed, only a few number of these materials are available to the experimenters, it therefore facilitates the comparison between studies dealing with the same topic. In this section, we present the materials that are commonly used in the literature to induce the mental states that are studied in this thesis - i.e., cognitive workload, emotions and curiosity - and limit them to the ones that are used in studies attempting to estimate such states through neurophysiological signals. The goal will later be to estimate these induced mental states, the materials therefore need to induce at least 2 levels of a given mental state, e.g. low workload vs high workload concerning the cognitive workload, in order to classify neurophysiological signals using machine learning algorithms.

2.4.1 *Cognitive Workload*

Multiple materials are used to induce different levels of cognitive workload. Based on the literature, we can split these tasks into two types, on the one hand the laboratory setups, e.g., N-back or Rspan tasks, and on the other hand the real-world setups, e.g., real-world car driving. In this sub-section, we only present the cognitive workload-triggering tasks that are the most commonly used in laboratory setups.

N-back task: the N-back task can be used in multiple ways, e.g., with letters or numbers (visual N-back tasks), or shapes (visuospatial N-back task), following the type of cognitive workload the experimenter wants to induce in participants (Gevins et al., 1997). The visual N-back is considered to be very strongly related to the working memory (Berka et al., 2007). The method is built as follows: users see a sequence of letters (or a single-digit number) on a screen, the letters/numbers being displayed one by one, every 2 seconds. For each letter the user has to indicate whether the displayed letter is the same one as the letter displayed N letters/numbers before. The "N" can be manipulated in order to obtain different levels of workload, usually ranging from 0-back, i.e. the easier level where the user has to identify whether the current letter is the letter 'X' (for instance), to 3-back, i.e. the most difficult level we found in the literature, where the user had to identify whether the current letter was the same letter as the one displayed 3 letters before. Between these two levels, we can find intermediate workload levels, i.e., 1 and 2-back tasks, with moderate difficulties. For each displayed letter/number, a segment of 2 seconds is usually used as a trial, and is labeled with its workload level, e.g., "low" workload for 0-back trials and "high" workload for 2-back trials.

For the visuospatial N-back task (Suchan, 2008), the same structure is kept, but with a moving square location: it consists of remembering and comparing the previous locations of a moving white square to its current location. The white square appears at a random location on the computer screen for 0.5 s then disappears, with a 2-second delay between trials. For each trial, the participant is asked to compare the box current location to its location several times before, this number being N=0 (low difficulty) to 3 (high difficulty) as in the letters-based N-back task. In other words, participants has to remember the box's last location in the 1-back condition, but they have to remember the last three locations in the 3-back task.

Rspan task: first called the Reading span task (Unsworth et al., 2005), the Rspan task is closely modeled after Unworth's version of

the task. The Rspan task is split into a primary and a secondary task, and requires from the user simultaneous processings and storages. On the one hand, participants have to process sentences by determining if they were semantically correct or not, and on the other hand they have to simultaneously maintain in memory a letter (out of nine consonants: B, F, H, J, L, M, Q, R, X) that is presented at the end of each sentence. As the whole, the task is organized by difficulty levels, then by letter group, and lastly by individual trial. Participants have a limited amount of time to read each sentence, then they are asked if the sentence makes sense or not, and then have to retain the letter that appears on the screen. The recall phase comes after a certain number N of pairs of sentences/letters, and this number N defines the difficulty level: the higher this number N, the higher the difficulty. The lower level of workload is usually 2 sentences - meaning 2 letters to retain – and the highest level of workload is usually 6 sentences/letters.

Sternberg task: the Sternberg task (Sternberg, 1969) is based on the Sternberg Memory Scanning task (Sternberg, 1966), and consists in presenting multiple series of single-digit numbers to the participant (the single-digit numbers are displayed one after the other). After each series, the participant sees a new single-digit number, i.e. the "probe", displayed on the screen, and has to determine if this new number was in the series. The difficulty of workload levels is manipulated by increasing and decreasing the length of the series, e.g. ranging from 1, i.e., the lowest level, to 6, i.e., the highest level.

go/no-go: in the go/no-go task (Redick et al., 2011), letters are stimuli, and series of stimuli are displayed one after the other to the user. There are 3 difficulty conditions in this task, i.e. low workload, medium workload and high workload. The low workload presents two letters, N as Go and X as No-Go stimuli, when the medium one has 15 letters as Go (B, C, D, F, G, H, J, K, L, M, N, P, R, S, T) and X as No-Go stimuli. Finally, the highest difficulty level consists in the same 15 Go stimuli, but X and Y as No-Go stimuli. After each letter, the participant has a short time to solve the problem, i.e., to indicate if the stimulus was Go or No-Go.

Forward Digit Span: in the Forward Digit Span (FDS) task (Kreutzer et al., 2011b), single-digit numbers are displayed one after the other to the participants, forming a series. Participants are asked to retain the whole series and then have to input the numbers in the order they were displayed. The cognitive workload difficulty levels are induce based on the length of the sequence participants have to memorize. The lowest workload level is a series of 20 sets of 3 single-digit numbers, and the highest consists in a series of 4 sets of 8 single-digit numbers, and

4 intermediate levels are defined in-between.

Backward Digit Span: the Backward Digit Span (BDS) task (Kreutzer et al., 2011a) shares similarities with the FDS, since Participants have to retain a series of single-digit numbers displayed on the screen. The difference is made in the recall phase, where participants are asked to input the single-digit numbers in the reverse order from the one displayed. The lowest difficulty level is here a series of 12 sets of 3 digits, and the highest level is made of a series of 4 sets of 8 digits. Three intermediate levels are found in-between those two levels.

Arithmetic Task: the Arithmetic Task consists in performing calculations on multiple-digit numbers. As detailed in (Zarjam et al., 2013) and (So et al., 2017), participants are told to determine the correctness of arithmetic equations displayed on a screen. The 3 difficulty levels of workload are defined as follows: the lowest one is based on additions with single-digit numbers, the medium one is based on double-digit numbers additions, and the highest one is based on equations with double-digit numbers and additions/subtractions. Note that in this task, not only the working memory is involved, i.e., by storing interim results/temporary information and doing calculation strategies, but also the long term memory, i.e., by first retrieving an arithmetical fact from it.

Other tasks: other tasks are used to induce different levels of cognitive workload, but these are used very little, which is why we do not detail them in this section. This is the case for example of the logical task, used in (Chaouachi et al., 2011), the finger tapping task used in (So et al., 2017), the mental rotation task (So et al., 2017), visuomotor tracking task (van Beurden et al., 2020) or the lexical decision task (So et al., 2017).

2.4.2 *Affective states*

Different methods for inducing affective states exist, including some for eliciting emotions, and these techniques may vary depending on the theory or model the experimenter chooses to follow. Moreover, the choice of the method will also be based on the experimenter strategy to combine the support to induce emotions and the material to measure them. Moreover, unlike techniques for inducing different levels of cognitive workload, methods for inducing emotions do not pre-calculate the impact a stimulus would have on the user, e.g., the level of valence that a video could elicit in an individual. Most often, they will be combined to a subjective rating from the user, e.g., the level of valence that the video elicited on the user from 1 to 7. In this section, we only

present common stimuli, all well known to induce different types of emotions. Note that these stimuli are considered to be mainly used in laboratory setups.

International Affective Picture System (IAPS): the IAPS (Bradley and Lang, 2017) is an image bank that makes it possible to have an extensive affective rating of various images, and thus to have several scales available. This allows the experimenters to choose the images that correspond the most to the types of emotions they want to measure, and therefore to induce them by showing participants the corresponding images (the choice depends on the model that has been chosen by the experimenter beforehand).

International Affective Digital Sounds (IADS): similar to IAPS, IADS (Soares et al., 2013) is a bank of sounds that makes it possible to have an extensive rating of various sounds, and therefore to induce emotions by playing these sounds to subjects.

Music clips: music clips are a major method to induce emotions as well: they mix two types of stimuli, i.e., auditory and visual, and proved to be effective in elicitating powerful emotions (Westermann et al., 1996). It is possible to make a bank of music clips quickly, by using affective tagging applications in order to assign tags to the these music clips.

Video clips: video clips have the same characteristics as music clips, but focus on the visual aspect of the stimuli and do not have the auditory dimension. This method is often based on extracts from famous movies, but video clips other than movie clips have been used for inducing emotions in the DEAP data set (Koelstra et al., 2011).

2.4.3 *Epistemic Curiosity*

So far, to the best of our knowledge, only Trivia questions-based materials have been presented to participant in order to induce states of epistemic curiosity. This is the case in three studies, two that have measured such states through fMRI (Gruber et al., 2014; Kang et al., 2009), and another one that has recorded EEG signals (Lima, 2019).

2.5 *Measuring learning-related mental states*

As shortly discussed in section 2.5, there are multiple existing methods for measuring the learning-related mental states that are studied

in this thesis. On the one hand, the subjective measures, i.e., user opinion, responses to questionnaires or interviews, have been used for reporting the varying status of the mental states. This is done subjectively either by the users themselves (e.g., questionnaire), or by the experimenter. On another hand, we have the behavioral measures, i.e., the actions of individuals. More recently, studies have been using objective measures as well, i.e. measuring users' mental states through peripheral physiological activity, e.g., using sensors for cardiac activity, and more importantly brain activity, e.g., EEG. In this section, we first describe out the subjective measures that are used for each of the mental states that are studied in this thesis, and then we present the objective measures that are commonly used for studying most of the mental states.

2.5.1 *Behavioral measures*

Behavioural measures (Jacob-Dazarola et al., 2016), i.e., the actions of individuals, include all actions that can be taken by observing an individual's actions. They can be physical movements, decisions, or the performance of individuals on a task. This can be measured subjectively, with what the experimenter observes and then reports, or objectively, e.g., keyboard press, reaction time or mouse tracking.

2.5.2 *Subjective measures*

Subjective measures refer to questions that seek to understand the mental state of individuals as they experience it (Annett, 2002). As seen above, subjective measures can be very structured, as with a questionnaire, but can also be more informal, as with an interview. Unlike behavioural measures, subjective measures can be taken even in the absence of physical actions. In addition, it is possible to develop questionnaires that are not specific to a particular task, making it easy to compare tasks with each other. We present the subjective measures existing for our mental states, i.e., cognitive workload, emotions and curiosity, as follows:

Cognitive Workload: cognitive workload is a mental state that has been widely studied in psychology through questionnaires and therefore subjective measures. We present here the most common subjective measures, including the most widely used questionnaire of the cognitive workload literature, i.e., the NASA-TLX (Hart, 2006):

- **NASA-TLX**: a popular method for the subjective assessment of users's mental workload is the NASA Task Load indeX (NASA-TLX) (Hart,

2006). This scale consists of six dimensions: mental demand, physical demand, temporal demand, performance, frustration level, and effort, each with 10- or 20-point scale. All components are then combined into a single score that is the weighted average of the ratings for each component. One of the limitations is that users have to wait until the end of the task to rate their level of cognitive workload.

- **Subjective Workload Assessment Technique**: through this multidimensional method, three factors are assessed, i.e., the time load, the mental effort load and the stress load (Reid and Nygren, 1988). The scoring is a subjective rating technique, and three levels are used for each of three dimensions: low, medium or high, respectively 1, 2 and 3. The three factors are then combined in order to produce an interval scale of mental workload.

- **Cooper-Harper scale**: first used for aircraft pilots while performing a handling task, this is a scale ranging from 1 to 10, with 1 being the best handling characteristics and 10 the worst (Cooper and Harper, 1969). This scale was then used in various fields.

- **ATWIT**: The Air Traffic Workload Input Technique (ATWIT) (Stein, 1985), first used for Air traffic, but then applied to many fields (Loft et al., 2014), where participants are asked to rate the level of cognitive workload on a 1 (low workload) to 7 (high workload) lickert scale. The advantage of this method is that users can report the workload as it changes, and do not have to wait until the end of the task to do so.

Affective states: to our knowledge, no widely-used questionnaires - or subjective measurements other than likert scales - are existing. Concerning the well-known likert scales, we find the self assessment Manikin (SAM) scale, and the classic Valence - Arousal - Dominance scale (Koelstra et al., 2011).

Curiosity: to our knowledge, no widely-used questionnaires - or subjective measurements other than likert scales - are existing.

2.5.3 *Neurophysiological measures*

Objective measurements (Cowley et al., 2016) are mainly derived from sensors that measure the electrical activity of some human body areas, e.g., electrocardiography or electroencephalography to collect information about the heart or the brain. Electrical signals are therefore recorded and then analyzed by signal processing algorithms. However, not all sensors are sensors of electrical activity, and other methods

are then used to process the collected data. However, we describe a large part of the sensors that can be found in the literature, for the objective measurement of mental states, without going into details. Here is the list of these sensors:

- **electroencephalography (EEG)**: electroencephalography measures changes in the electric field caused by neuronal activity. To do this, electrodes are placed on the scalp of individuals.
- **functional near-infrared spectroscopy (fNIRS)**: this technique is based on the examination of the levels of oxygenated (HbO) and deoxygenated (HbR) hemoglobin concentration in the cerebral cortex: the level of oxygenation in the brain changes the degree to which near-infrared light is reflected.
- **heart rate and heart rate variability (HRV)**: several techniques exist, but the most common one is the electrocardiography (ECG). This technique consists of placing electrodes on the individual (often at the chest level) to measure the changes in electric fields caused by the heart activity.
- **ocular activity**: ocular activities can be measures of pupil dilation, blink frequency and blink duration or saccades, and is typically measured by cameras. These cameras can be installed in different ways (e.g. fixed to the workstation, fixed to the computer screen, mounted on a bezel) and are used to record eye movements.
- **breathing**: breathing can be measured in several ways: one of the simplest and most commonly used one is to place a band around the individual's chest and measure the stretch in that band caused by breathing
- **electrodermal activity(EDA) / galvanic skin response (GSR)**: EDA uses electrodes, which are often placed on the hand or fingers of individuals. Once installed, a weak electric current flows through these electrodes and allows the measurement of electrical conductance, which varies with the level of sweating.

As seen above, in addition to these listed physiological measures we also have behavioral measures that can be considered as objective measures: they include all actions that can be taken by observing users' actions. They can be physical movements, decisions, or the performance of individuals on a task. For example, keystroke dynamics, mouse tracking, and body positioning are metrics that can be studied.

In this thesis, we mainy focus on EEG as an objective measure of

mental states, and we made a literature review of studies that have attempted to estimate our three mental states, i.e., cognitive workload, emotions and curiosity, through EEG signals only. To a smaller degree, we are also interested in physiological signals, including heart rate, breathing and EDA, as we consider them in our study on curiosity (see part 8).

2.6 *Extracting features from mental states measurement*

As we have just seen before, there are therefore 2 main types of measures to identify the different levels of mental states, i.e., subjective and objective measures. Subjective measures have long been used by researchers, particularly in the field of psychology, but have shown limitations concerning the objectivity of the measure (and hence its name): the level of a given mental state is either defined by the user himself, or by the experimenter. This leaves human beings as only judges of the variations of these mental states levels. However, more recent methods, here referred to as objective measures, have emerged with the development of new technologies, in order to obtain information on the users' neurophysiological responses. In this thesis, we focus on neurophysiolgocal measures of mental states only, in order to go towards estimation of mental states in real-time.

Among the objective measures, we find 2 subtypes, i.e. body sensors that allow to measure the physiological responses of the user, and neurotechnologies that allow to measure brain activity. While several studies have shown that it is possible to decode mental states from physiological signals alone (Debie et al., 2019), a combination of physiological signals with neural signals (Debie et al., 2019), once again, in this thesis we decided to focus on EEG only. Indeed, in the remainder of this chapter, we study the literature by focusing on papers dealing with the measurement of mental states through EEG signals exclusively.

2.6.1 *Features in the EEG*

The measurement of the brain activity due to variations in the different mental states might be complex due the low spatial resolution of the EEG. It is therefore important to focus on the strengths of EEG, i.e., a good temporal resolution, and extract features that will bring information based on both the time and frequency domains. However,

features about the spaciality of the EEG signals can be extracted as well, using spatial filters.

Time domain-based features: concerning the time domain in EEG, the Event Related Potentials (ERPs) focuses on variations in the signal over a time window of a few hundred milliseconds after a stimulus. ERP waveforms consist of a series of positive and negative voltage deflections, which are related to a set of underlying components (Kappenman and Luck, 2012). We can distinguish different types of ERPs based on the components that are referred by a letter (N/P) indicating polarity (negative/positive), followed by a number indicating either the latency in milliseconds. For instance, a positive-going peak that occurs about 300 milliseconds after a stimulus is presented is often called the P300, indicating its latency is 300 ms after the stimulus and that it is positive. The time of the trials used by the machine learning algorithms in this case is generally between 0.8 and 1.2 seconds (Mühl et al., 2014; Roy et al., 2015b). The oscillatory activity, which is the alternative to ERPs, has usually longer time windows, e.g. from 2-seconds time windows to 120-seconds time windows (Brouwer et al., 2012; Grimes et al., 2008), and is based on variations in brain rhythms, and therefore depends on the frequency domain.

Frequency domain-based features: typically, since more and more evidence shows that the oscillatory rhythm have a functional significance for the workings of the human mind, the EEG activities in human beings can be classified according to their frequency (from the slowest to the fastest: delta, theta, alpha, beta, and gamma):

- **Delta power (δ)**: frequencies from 0.5 to 4 Hz, normal in very young children, they can then characterize certain brain lesions or deep sleep (Steriade et al., 1993). This frequency range is also associated with conative states, i.e., motivational states (Knyazev, 2012), and affective states, i.e., emotions (Aftanas et al., 2002).

- **Theta power (θ)**: theta comprises oscillations between 4 and 8 Hz, and is an important oscillation to take into consideration when studying mental states, for variations in cognitive states, e.g. cognitive workload (Klimesch et al., 2008), or variations in affective states, e.g. emotions (Sammler et al., 2007). It is thought to increase as the cognitive demand of the task is important (Fernandez Rojas et al., 2020). Theta also increases when the the concentration is sustained (Gevins and Smith, 2003), or when the working memory increases as well (Borghini et al., 2012).

- **Alpha power (α)**: in general, alpha band increases in relaxed states

with eyes closed and decreases when the eyes are open (Antonenko et al., 2010). This brain wave is dominant in humans, and generally consists of oscillations in the 8–13 Hz range (Andreassi, 2007). Alpha is probably the most important oscillation to consider when mental states are studied, whether it is cognitive states (Gerjets et al., 2014), affective states (Coan and Allen, 2004), or conative states (Harmon-Jones, 2003).

- **Beta power (β)**: corresponds to frequencies above 13 Hz (and generally below 45 Hz). They generally appear in a first case of calm awakening, also called internal awakening, when the individual, with eyes closed, is in a state of diffuse attention without a specified perceptual or mental task, and the alpha rhythm is then often present with superimposed or alternating beta activities. The beta power is also subject to variations of affective states, e.g., emotions (Onton and Makeig, 2009) and cognitive states, e.g., attention (Cole and Ray, 1985). This band power can be split into two sub-bands, i.e., low beta (lβ) and high beta (hβ).

- **Gamma power (γ)**: frequencies above 35 Hz, usually about 40 Hz, up to 80 Hz. They are only very little studied when we are interested in mental states, even if some results showed they could be interesting to study, i.e., cognitive states such as attention or memory (Jensen et al., 2007), or affective states (Müller et al., 1999).

Spatial filters: There are multiple spatial filter types, but a couple of them are widely used in the EEG-based BCI literature. We quickly present them in this section, but some of them, i.e., the Common Spatial Pattern (CSP) and the Filter Bank CSP (FBCSP), that we used for our contributions, will be described in more details in chapter 2.

- **Common Spatial Pattern (CSP):** this algorithm optimizes the EEG signal-to-noise ratio: the variance of spatially filtered signals is maximized for one class and minimized for the other class (Blankertz et al., 2008).

- **Filter Bank Common Spatial Pattern (FBCSP):** EEG signals are filtered into multiple frequency bands, usually nine or ten 4Hz-wide if we refer to (Ang et al., 2012), and then N filter pairs are optimized for each frequency band using the CSP cited right above. Finally, a subset of them is selected using feature selection.

- **Principal Component Analysis (PCA):** this algorithm is employed to first obtain uncorrelated components by performing a linear and orthogonal transformation (Roy et al., 2015b). Among these components, some are then selected to finally be used as spatial filters.

- **xDAWN:** this algorithm optimizes the ratio between the signal and the signal plus noise ratio in order to obtain the spatial filters dedicated to ERP classification (Rivet et al., 2009).

In addition to these listed algorithms, the Fisher Spatial Filter (FSF) (Mühl et al., 2014) and the Canonical Correlation Analysis (CCA) (Hoffmann et al., 2006) are two other spatial filtering methods that can be found in the literature.

2.6.2 *Features in the physiological signals*

Features can also be extracted from physiological signals such as heart rate, breathing or electrodermal activity, and used to estimate mental states such as cognitive workload (Brouwer et al., 2012). These features are described in our contribution 3 - precisely in chapter 8 - where we attempted to estimate states of curiosity through EEG, but also physiological signals.

2.7 *Estimating mental states through EEG and physiological signals*

As seen in the previous sections, there are different methods for inducing mental states, measuring these mental states, and finally extracting features by processing the EEG signals that have been recorded during variations of these mental states. These three steps are critical in the process of estimating mental states through EEG signals, but several additional steps are necessary to carry out such classifications of mental states levels. Therefore, other decision-makings concerning the choice of methods/parameters have to be done in order to evaluate classification performances for a given mental state. First, the experimenters have to carefully define the trials they will use to feed the machine learning algorithms, i.e., the type signals (ERP vs oscillatory), the number of trials and the time-window into which signals are split. Second, the type of calibration, i.e. subject-specific or subject-independent, might differ from a study to another, and are defined as follows:

- **subject-specific:** the classifier needs to be built for each individual subject (Blankertz et al., 2008). First, data specific to each subject are split into two parts: the training and testing sets. Then, machine learning algorithms are trained on the first set and evaluated on the second one. These two parts can be of size relative to the number

of available trials, e.g., 50%/50% or 30%/70%.

- **subject-independent:** one of the major steps for monitoring mental states outside of the laboratories would be for users to be able to instantaneously use the system without any calibration phase. To do so, it is possible to evaluate machine learning algorithms through online & offline subject-independent studies, i.e., with a classifier built on multiple subjects and used as such on a new subject, without the need for data from this new subject. For example, the common evaluation method for this type of calibration is a leave-one-subject-out cross validation, i.e., the training phase uses all subjects expect the target subject data to train the classifier, and the testing phase applies this classifier on the target subject data only. This process is repeated with each subject used once as the target (test) subject.

Other decisions need to be made when designing the protocol, such as the number of participants to include in the study, the number of sessions per participant, or both the number and the type of sensors. These choices are important and have a direct impact on the quality of the study. For example, the number of participants must be large enough to be able to conclude on the validity of the statistical tests made on the study.

Finally, there are multiple types of algorithms to classify EEG signals, and thus mental states. The experimenters must therefore, here again, make choices about the classifier they want to use in their experiments. Among these classifiers, the linear classifiers are well known from most of the experimenters, e.g. the Linear Discriminant Analysis (LDA), the Support Vector Machine (SVM), the Gaussian Naïve Bayes (GNB) and the Logistic Regression (LR). However, more complex machine learning algorithms exist, but might not be as common as linear classifiers: this is the case of Artificial Neural Networks (ANN). Note that all the existing machine learning algorithms for the classification of EEG signals are listed and detailed in the literature review in (Lotte et al., 2018a).

3

A literature review of EEG signals-based estimation of cognitive workload & emotions

As seen previously, we focus on the estimation of three mental states in this thesis, i.e., cognitive workload, affective states and curiosity. The first step of this thesis is therefore to review the literature concerning the studies that have attempted to estimate these mental states. Indeed, the ultimate goal would be to use passive BCIs to estimate these three mental states through EEG signals.

This literature review is organized as follows: first of all, it is divided into three sub-sections, each one corresponding to a specific mental state, i.e., a sub-section is dedicated to the cognitive workload, another one to the emotions, and so on. Then, for each of these mental states, we describe out the methods that have been used in the literature to induce them, the methods for pre-processing the EEG signals, the ones for processing as well as the choices that have been made by the experimenters concerning the protocol designs (that we saw in chapter 2), and mainly the method for classifying those EEG signals.

All the criteria that have been cited right before, and which make it possible to define the methods used to estimate these mental states in literature, are grouped together in the form of tables which make it

possible to have a global overview of them.

3.1 Cognitive workload

Based on the density of the literature on this subject, the estimation of cognitive workload through EEG signals has been extensively studied over the last 30 years. Among the studies that have been carried out, the methods for estimating the cognitive workload through EEG signals differ, and therefore bring interesting perspectives concerning the step-by-step process to follow to estimate such a cognitive state, i.e. inducing, measuring, and finally processing & classifying the signals. If multiple parameters have been tested by the experimenters in the literature, only the more promising findings are reported on Table 3.1. For example, we limit our literature review to studies that sought to classify different levels of cognitive load, and do not consider the ones that just focused on features used for evaluating cognitive workload such as in (Zarjam et al., 2011). We also focus on studies that have concentrated on healthy participants, although some studies have looked at non-healthy patients (Mathan et al., 2010; Mazher et al., 2016) but we do not consider them in our literature review. We also choose to exclude studies that used unsupervised methods such as (Das et al., 2013). Finally, we limit the collected studies to those that measure variations in cognitive workload with laboratory setups, either with purely laboratory-based designs (e.g., n-back task on a screen (Brouwer et al., 2012)), but also with tasks that seek to recreate real-life situations in a simulator (Dijksterhuis et al., 2013). The data bases that have been explored are IEEE Xplore, PsyArRTICLE, Scopus as well as Google scholar, and the formulas we used were “workload + EEG" and “workload + electroencephalography".

Participants, sessions and EEG channels: most of the studies of the review have been run with at least 12 participants, going up to 35 for the study by Brouwer et al. (Brouwer et al., 2012). However, this number of participants has remained low for some of them, i.e., 5 (Heger and Schultz, 2010), 6 (Honal and Schultz, 2008) and 8 (Duraisingam et al., 2017; Gevins et al., 1998; Grimes et al., 2008). Concerning the number of sessions, all of the studies ran a single one except the study by Gevins et al. (3 sessions (Gevins et al., 1998)) and the one by Dijksterhuis et al. (multiple sessions (Dijksterhuis et al., 2013)). Finally, the number of channels on the EEG cap might differ from a study to another as well, usually 16, 32 or 64 electrode caps. In the study by So et al. (So et al., 2017), authors used a single-electrode cap, which is

Table 3.1: characteristics of many studies related to the estimation of different levels of cognitive workload using machine learning algorithms. The results column indicates the best average of the reported classification performance scores. "Participants, sessions, channels" acronyms are: participant (P), sessions (S), channels (C). "features extraction methods" acronyms are: principal component analysis (PCA), common spatial pattern (CSP), filter bank CSP (FBCSP), fisher spatial filters (FSF), canonical correlation analysis (CCA); low (l.), high (h.). "classifier" acronyms: network (net.), support vector machine (SVM), gaussian process regression (GPR), linear discriminant analysis (LDA). "calibration": subject-specific (SS), subject-independent (SI), cross-task (CT).

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score in % (calibration)**
Gevins et al., 1998	8P, 3S, 27C	continuous matching task	oscillatory 3,112 4.5s	spectral features	neural net. 2 (l. vs h.)	95.0 (SS)
Honal et al., 2008	6P, 1S, 16C	"lecture and meeting scenario" task	oscillatory 3,845 2s	spectral features short time Fourier transform	SVM 2 (l. vs h.)	92.2 (SS) 80.0 (SI)
Grimes et al., 2008	8P, 1S, 32C	n-back task (1, 3)	oscillatory 20 120s	spectral features	naïve bayes density 2 (l. vs h.)	99.0 (SS)
Heger et al., 2010	5P, 1S, 16C	flanker and switching paradigms task	oscillatory unspecified 1s	spectral features fast fourier transform PCA filtering	SVM 2 (l. vs h.)	91.0 (SS) 72.0 (SI)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Chaouachi et al., 2011	17P, 1S, 6C	FDS BDS logical task	oscillatory undefined 1s	spectral features fast Fourier transform	GPR 3 (l. vs m. vs h.)	91.0 (SS)
Brouwer et al., 2012	35P, 1S, 7C	n-back task (0, 2)	oscillatory & ERP 384 2.5s & 1s	α fast Fourier transform	SVM 2 (l. vs h.)	90.0 (SS) & 80.0 (SS)
Baldwin et al., 2012	15P, 1S, 40C	reading span n-back (1, 3) Sternberg task	oscillatory unspecified 5s	spectral features	neural net. 2 (l. vs h.)	87.1 (SS)
Roy et al., 2013	20P, 1S, 32C	Sternberg task	oscillatory 160 0.8s	spectral features electrode selection (RG) CSP	LDA 2 (l. vs h.)	65.5 (SS)
Walter et al., 2013	21P, 1S, 30C	reading span n-back (1, 3) go/no-go	oscillatory 284 1s	spectral features	SVM 2 (l. vs h.)	97.4 (SS)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Dijksterhuis et al., 2013	34P, multiple, 64C	driving simulator	oscillatory 88 1s	$\gamma 1$, $\gamma 2$ CSP	LDA 2 (l. vs h.)	95.0 (SS)
Walter et al., 2013	21P, 1S, 30C	n-back (1, 3) go/no-go reading span aritmetic	oscillatory 232 undefined	spectral features	SVM 2 (l. vs h.)	95.0-97.0 (SS) 50.0 (CT)
Hogervorst et al., 2014	14P, 1S, 7C	n-back (0, 2)	oscillatory 768 2.5s	ϑ, α	SVM 2 (l. vs h.)	86.0 (SS)
Mühl et al., 2014	22P, 1S, 28C	n-back (0, 2)	oscillatory + ERP 1440 2s(osci.) +1s(ERP)	spectral features FBCSP FSF	LDA 2 (l. vs h.)	74.0 (SS)
Gerjets et al., 2014	16P, 1S, 10C	n-back (1, 3) reading span algebra problem	oscillatory undefined undefined	spectral features	SVM 2 (l. vs h.)	73.0 (CT)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Zarjam et al., 2015	12P, 1S, 32C	aritmetic task	oscillatory 42 5.0s	spectral features entropy energy	neural net. 7 (7 levels)	98.0 (SS)
Roy et al., 2015	20P, 1S, 32C	Sternberg task	ERP 80 0.6s	spectral features PCA CCA xDAWN	LDA (10-fold CV) 2 (l. vs h.)	98.0 (SS)
Sinha et al., 2016	15P, 1S, 14C	Logical reasoning	oscillatory 10 10.0s	spectral features CSP	SVM 2 (l. vs h.)	81.0 (SS)
Bashivan et al., 2016	15P, 1S, 64C	Sternberg task	oscillatory 240 5.5s	spectral features wavelet entropy	SVM 4 (4 levels)	92.0 (SS)
Roy et al., 2016	20P, 1S, 32C	Sternberg task detection task	ERP 80 0.6s	spectral features CSP CCA	LDA 2 (l. vs h.)	91.0 (SS)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Wang et al., 2016	14P, 1S, 14C	n-back (1, 3)	ERP 98 undefined	spectral features many features	SVM 2 (l. vs h.)	84.0 (SS)
Krol et al., 2016	6P, 1S, 64C	aritmetic task words recall	oscillatory 288 10.0s	FBCSP	LDA 2 (l. vs h.)	70.0 (CT-offline) 68.0-76.0 (CT-online)
So et al., 2017	20P, 1S, 1C	aritmetic task finger tapping mental rotation lexical decision	oscillatory 240 2.5s	spectral features	SVM 2 (l. vs h.)	75.0 (SS)
Duraisingam et al., 2017	8P, 1S, 14C	Java programming task	oscillatory 4727 1.0s	spectral features energy	Naive Bayes Classifier 2 (l. vs h.)	76.6 (SS)

rather rare.

Inducing cognitive workload: in the literature, multiple methods have been used to induce workload states at different levels, and most of these methods are described in chapter 2. Among them, the N-back task and the Sternberg task are the ones that has been studied the most, with respectively 7 (Baldwin and Penaranda, 2012; Brouwer et al., 2012; Grimes et al., 2008; Hogervorst et al., 2014; Mühl et al., 2014; Walter et al., 2013; Wang et al., 2016) and 5 (Baldwin and Penaranda, 2012; Bashivan et al., 2016; Roy et al., 2013,1,1) studies out of the 20 studies that have been identified and analyzed in this review. Otherwise, the only tasks that has been used multiple times are the arithmetic task with 2 studies to its credit (So et al., 2017; Zarjam et al., 2015), and the reading span task with 2 studies as well (Baldwin and Penaranda, 2012; Walter et al., 2013). Note that 19 studies out of the 20 have been conducted under laboratory constraints, and only one has been run in a real-world-based simulator, here a driving simulator (Dijksterhuis et al., 2013). In (Duraisingam et al., 2017), authors used an uncommon task, as participants were asked to solve nine Java programs of different difficulty level.

Characterizing trials: three types of features are retained to characterize the trials that are used in the different studies: the type of trials (oscillatory vs ERP), the number of trials and the time-window. Concerning the type of trials, the literature review allowed us to count the oscillatory-based studies, i.e., 17 of them, versus the ERP-based studies, i.e., 5 of them. Then, the higher the number of trials, the better the machine learning algorithms will be trained, and therefore the better the predictions and the classification performance scores will be. In the literature, this number goes from 10 (Sinha et al., 2016) to 4727 (Duraisingam et al., 2017). Finally, the time windows vary from 0.6 to 1 second for the ERP-based trials, and from 1 (Chaouachi et al., 2011; Heger and Schultz, 2010) to 120 seconds (Grimes et al., 2008) for oscillatory-based trials. In (Roy et al., 2013), authors chose to use trials of 0.8s to avoid any confounding effect with memory encoding. Note that the tables indicate the time-windows that have been used to find the best classification performance score, but not all values of parameters are transcribed into them. For example, the Grimes et al. (Grimes et al., 2008) and the Brouwer et al. (Brouwer et al., 2012) studies have used several time windows, ranging from 2 to 120 seconds, but only the results from the use of certain time windows, here 120 seconds and 2.5 seconds respectively, are presented in the tables.

Processing EEG signals: multiple methods are used to process the EEG signals in the cognitive workload estimation literature, but the

one that is used unanimously is the band-pass filtering. As seen in chapter 2, some frequency bands are interesting when studying mental states, i.e., delta (δ), thêta (ϑ), alpha (α), beta (β) and gamma (γ). The frequency bands that are used the most for estimating cognitive workload levels are thêta and alpha, respectively in 17 and 19 out of 20 studies. Other common features have been extracted using well-known algorithms, such as spatial filters like the CSP (Dijksterhuis et al., 2013; Roy et al., 2013,1; Sinha et al., 2016), the FBCSP (Mühl et al., 2014), the FSF (Mühl et al., 2014), the CCA (Roy et al., 2015b,1), the PCA (Heger and Schultz, 2010; Roy et al., 2015b) and xDAWN (Roy et al., 2015b). In (Roy et al., 2013), authors used a Riemannian Geometry-based method in order to select the most relevant electrodes (Barachant and Bonnet, 2011). Duraisingam et al. used ratio of the different frequency bands (alpha and theta), as features that have been input in the classifier (Duraisingam et al., 2017). However, this was not the only study that used a ratio of different frequency bands as features. Note that we only indicate the frequency bands that have been used in the table, without precising the different ratio authors used.

Estimating cognitive workload from EEG signals: among the 20 studies, only 2 of them proposed a subject-independent (SI) calibration method for the classification phase (Heger and Schultz, 2010; Honal and Schultz, 2008). For the subject-specific (SS) studies, multiple machine learning algorithms have been used in the literature, starting with SVM which has been the most widely used, with a number of uses of 9 (Bashivan et al., 2016; Brouwer et al., 2012; Heger and Schultz, 2010; Hogervorst et al., 2014; Honal and Schultz, 2008; Sinha et al., 2016; So et al., 2017; Walter et al., 2013; Wang et al., 2016). Then, the LDA has been used for 5 studies, the artificial neural networks for 3 studies, and otherwise the authors have opted for Naive Bayes classifiers or even an algorithm we did not describe in this thesis: the Gaussian Process Regression (GPR) (Chaouachi et al., 2011). Unfortunately, the number of such methods, the parameters they have been using (e.g., the architecture of ANN), the variety of computed features, and differences in protocol designs from a study to another make classification performances comparisons highly problematic. Finally, 3 studies propose to estimate workload using cross-task set ups (Gerjets et al., 2014; Krol et al., 2016; Walter et al., 2013).

Take away: as we have just seen, the results are difficult to compare from one study to another, but this review of the literature offers us some perspectives. First of all, we can observe that ERPs have been used very little, indicating a tendency to measure cognitive workload over longer time intervals, i.e., using oscillatory-based passive BCIs.

In (Brouwer et al., 2012), authors used the same data set to perform both ERP-based and oscillatory-based studies, and obtained classification performances of 80.0% and 90.0% respectively, confirming a better algorithms' recognition of cognitive workload levels. While only 2 studies have performed subject-independent studies, they have performed studies with subject-specific calibration in parallel on the same data set, and have therefore used the same machine learning method for both SS and SI calibrations, and both obtained better classification performances with the subject-specific one, i.e., 91.0% for SS study versus 72.0% for SI in (Heger and Schultz, 2010), and 92.2% for SS study versus 80.0% for SI in (Honal and Schultz, 2008).

Concerning the EEG power, two oscillatory components have been early recognized as being sensitive to task difficulty manipulations - alpha and theta (Gevins and Smith, 2003; Klimesch et al., 2005). Moreover, based on the N-back task with time-windows of 2 seconds, it has been proposed by Brouwer et al. that the best frequency band for distinguishing between low (0-back) and high (2-back) cognitive workload levels is the alpha band (8-12 Hz) (Brouwer et al., 2012). The second best frequency band, even if less effective and less clear than with Alpha, is the thêta band (4-8 Hz). Still based on the work by these authors, the electrode which would be the most discriminating would be the Pz one. Concerning (Baldwin and Penaranda, 2012), their results lead to the assumption that the different tasks they have been using, i.e., reading span, Sternberg task, and spatial n-back task, induce highly dissimilar features in the EEG-signal, relying on separate neural structures or types of processing.

3.2 *Affective states*

The estimation of affective states has been widely studied in the literature over the last 20 years, and studies using various methods for inducing mental states, measuring them through EEG signals to then processing these signals, have been chosen by the experimenters. Moreover, algorithms and parameters for calibrating the experiments and thus classifying multiple emotions & levels of affective states through EEG signals have differed from one study to another. It is therefore interesting to list these studies, as well as the choices that the experimenters made concerning their parameters, as we can observe in Table 3.2.

As for the cognitive workload literature review, we focus this review on studies that attempted to discriminate at least 2 levels of affec-

tive states or at least two emotions, e.g., discriminate low valence from high valence, or joy from anger. The same holds true for the restriction of studies to those based on healthy participants or within-class classification, i.e., no cross-task studies have been kept. Finally, the studies that we collected have been limited to the ones that attempted to measure such affective states following laboratory setups: there are less cofunding factors and are more controlled. Concerning the datasets, Koelstra et al. proposed a “universal" dataset, called “DEAP" (Koelstra et al., 2011), that could be used to test and compare many parameters such as methods for feature extractions, for classification, or by playing with the number of channels, the trials characteristics, etc. The data bases that have been explored are IEEE Xplore, PsyArRTICLE, Scopus as well as Google scholar, and the formulas we used were “affective + state + EEG" and “emotion + EEG", “affective + state + electroencephalography" and “emotion + electroencephalography".

Participants, sessions and EEG channels: the number of participants varied from study to study, with 18 studies over the 21 ones having 10 or more participants, and the remainder having relatively small numbers of participants, i.e., 4 (Chanel et al., 2006), 5 (Khalili and Moradi, 2009), 7 (Choppin, 2000). The number 10 is arbitrary but this reflects the importance of having a minimum number of participants in order to draw conclusions from a study. However, most of the studies ran a single-session experiment, except for the ones by Zheng et al., who chose to set up two sessions per participant (Zheng and Lu, 2015), Finally the number of channels varied a lot between studies, i.e., going from 1 (Zhou et al., 2014) to 124 (Kothe et al., 2013). Note that Abadi et al. used 306 channels, but this number is including magnetoencephalography (MEG) ones (Abadi et al., 2013).

Inducing affective states: in the literature, multiple methods have been used to induce affective states at different levels, and most of these methods are described in chapter 2. Among them, the most frequently used methods were the International Affective Picture System (IAPS) and the video clips, with both 7 uses. Other methods are mostly based on images and video clips as well, sometimes music video clips, except for the materials used in the study by Chanel et al. (Chanel et al., 2011), where authors used a game, or in the study by Daly et al. (Daly et al., 2016), where they used musical excerpts, in order to to elicite emotions in users.

Characterizing trials: concerning the type of trials authors used in the different studies, only oscillatory-based methods have been presented in the papers we reviewed from the literature, indicating very poor knowledge of ERP-based emotion estimations. The number of

Table 3.2: characteristics of many studies relating to the estimation of different affective states, or different levels of emotions, using machine learning algorithms. The results column indicates the best average of the reported classification performance scores. “Participants, sessions, channels" acronyms are: participant (P), sessions (S), channels (C). “features extraction methods" acronyms are: mutual information (MI), normalized mutual information (NMI), filter bank common spatial pattern (FBCSP); low (l.), high (h.). “classifier" acronyms: network (net.), support vector machine (SVM), linear discriminant analysis (LDA). “calibration": subject-specific (SS), subject-independent (SI).

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Choppin et al., 2000	7P, 1S, 19C	images from IAPS sounds from IADS	oscillatory unspecified 6-10s	spectral features, coherence measures	neural net. 3	64.1 (SI)
Takahachi et al., 2004	12P, 1S, 3C	video clips	oscillatory unspecified unspecified	spectral features, stats (mean, std, etc)	SVM 5 (joy, anger, sadness, fear, relaxation)	42.0 (SI)
Chanel et al., 2006	4P, 1S, 34C	images from IAPS	oscillatory unspecified 6s	Spectral features	Naïve Bayes 2 & 3	60.0 & 46.0 (SS)
Khaili et al., 2009	5P, 1S, 54C	images from IAPS	oscillatory unspecified 12.5s	stats (mean, std, etc, correlation dimension	QDA 3	77.0 (SI)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Chanel et al., 2009	10P, 1S, 64C	Recall of emotional events	oscillatory unspecified 8s	Time frequency MI between electrod pairs	SVM 3	63.0 (SS)
Frantzidis et al., 2010	28P, 1S, 3C	images from IAPS	oscillatory unspecified 2.5s	ERPs average (40 trials), spectral features	SVM 4 (l./h. V., l./h A.)	81.0 (SI)
Petrantonakis et al., 2010	16P, 1S, 4C	images of facial expressions	oscillatory unspecified 5s	signals average per participant	SVM 6 emotions	85.0 (SI)
Lin et al., 2010	26P, 1S, 32C	music	oscillatory unspecified 30s	spectral features, hemispheric asymmetry	SVM 4 emotions	82.0 (SS)
Chanel et al., 2011	20P, 1S, 19C	game	oscillatory unspecified 300s	spectral features, features ratio	LDA 3	56.0 (SI)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Zhang et al., 2011	11P, 1S, 12C	images from IAPS	unspecified unspecified 0.5s	spectral features, hemispheric asymmetry	fuzzy clustering, neuro-fuzzy inference 4	unspecified
Koelstra et al., 2012	24P, 1S, 32C	musical video clips	oscillatory unspecified 35-117s	spectral features, hemispheric asymmetry	Gaussian Naïve Bayes 6 (l./h. V., l./h A., l./h. D.)	66.0-71.0 (SS)
Soleymani et al., 2012	24P, 1S, 32C	video clips	oscillatory unspecified 81s	spectral features, hemispheric asymmetry	SVM RBF 3 (l.,m.,h.)	V.:50.0 (SI), A.:62.0(SI)
Abadi et al., 2013	18P, 1S, 306C (with MEG)	video clips, musical video clips	oscillatory unspecified 51-128s 35-117s	spectral features, compressed DCT features	SVM 6 (l./h. V., l./h A., l./h. D.)	A.:61.0 (SS), V.:57.0 (SS), D.:59.0 (SS)
Hidalgo-Munoz et al., 2013	26P, 1S, 21C	images from IAPS	oscillatory unspecified 3.5s	spectral turbulence	SVM 2 (l./h. V.)	67.0 (SS), 82.0 (SI)

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Zhou et al., 2013	24P, 1S, 1C	images from IAPS, sounds from IADS	oscillatory unspecified 6s	spectral features	LDA, KNN 6	multimodal results
Kothe et al., 2013	12P, 1S, 124C	guided imagery	oscillatory unspecified 6s	FBCSP	LR 2 (V.:l./h.)	71.0 (SS)
Zheng et al., 2015	15P, 2S, 12C	video clips	oscillatory 2*3300 1s	differential entropy	Deep Belief Net. 3	86.7 (SS)
Soleymani et al., 2015	24P, 1S, 32C	video clips	oscillatory 20 34.9-117s	spectral features	SVM 6 (l./m./h. V., l./m./h A.)	V.:68.5 (SI) A:76.4 (SI)
Daly et al., 2016	20P, 1S, 32C	musical excerpts	oscillatory unspecified 2s	spectral features	SVM 3	65.0 (SS),

authors, year	**participants, sessions, channels**	**task**	**trials (type, number, time)**	**features extraction methods**	**classifier, classes**	**classification score (%)**
Candra et al., 2017	32P, 1S, 32C	video clips	oscillatory 400 6s	spectral features MI NMI	SVM 2 * 2 (l./h. V. & l./h. A.)	V.: 76.8 (SS), A.: 74.3 (SS)
Lin et al., 2017	32P, 1S, 32C	video clips	oscillatory 400 6s	spectral features-based images	CNN 2 * 2 (l./h. V. & l./h. A.)	V.: 80.1 (SS), A.: 78.2 (SS)

trials is not always present in the Table 3.2 because we relied on the study by Muhl et al. to list the studies that aimed at estimating affective states from 2000 to 2014 (Muhl et al., 2015), and the authors did not choose to indicate these values. However, we have reported these number of trials in Table 3.2 for the studies we reviewed ourselves, i.e., the studies that appeared between 2014 and 2017 (5 studies out of 21). Concerning the "DEAP dataset", EEG signals from 40 videos of 60s have been segmented into 6-seconds time windows, resulting in 400 trials of 6s for the studies that used this dataset (Candra et al., 2017; Lin et al., 2017). Otherwise, 20 trials have been used in (Soleymani et al., 2015) and 6600 trials (over 2 sessions) have been used in (Zheng and Lu, 2015). Finally the time windows also vary a lot from one study to another, i.e., going from 1 second in (Zheng and Lu, 2015) to 300 seconds in (Chanel et al., 2011).

Processing the EEG signals: as for the EEG signals processing in the cognitive workload studies, multiple methods are used, but the one that is used unanimously is the band-pass filtering. Some frequency bands are more interesting that others as well, i.e., thêta (ϑ) and alpha (α), even if delta (δ), beta (β) and gamma (γ) are sometimes used as well. Note that these power bands are detailed in chapter 2, and the use of such frequency bands are reported as "spectral features" in Table 3.2. However, other methods are used, such as the Filter Bank Common Spatial Pattern (FBCSP) (Kothe et al., 2013), differential entropy-based features (Zheng and Lu, 2015), compressed Discrete Cosine Transform (DCT) features (Abadi et al., 2013), basic statistics such as mean or standard deviation (Khalili and Moradi, 2009; Takahashi, 2004), hemispheric asymmetry (Koelstra et al., 2011; Lin et al., 2009; Soleymani et al., 2012; Zhang and Lee, 2012) or mutual information-based features (Candra et al., 2017).

Estimating affective states from EEG signals: among the 21 studies, and in comparison with the cognitive workload studies, eight of them proposed a subject-independent (SI) calibration method for the classification phase (Chanel et al., 2011; Choppin, 2000; Frantzidis et al., 2010; Hidalgo-Muñoz et al., 2013; Khalili and Moradi, 2009; Petrantonakis and Hadjileontiadis, 2010; Soleymani et al., 2012,1; Takahashi, 2004). However, many machine learning algorithms have been used, mainly the same ones as in the cogntive worklaod literature, except the study by Zheng et al. (Zheng and Lu, 2015) that used Deep Belief Networks in their study. As for the literature review of studies that attempted to estimate levels of cognitive workload, the detailed parameters used to fit the machine learning models are not presented in Table 3.2 (e.g., the architecture of ANN). Moreover, the designs of

the protocols run by the experimenters of the these studies highly differ, making very complicated to compare the results from one study to another. Indeed, some studies, based on Russell's theory (Russell, 1980), compared two levels of valence or arousal (i.e., low versus high), others are based on Ekman's theory (Ekman, 1992), compared 6 different emotions (i.e., joy, anger, etc). The number of classes that have been presented to the machine learning algorithms therefore differs - sometimes two classes, sometimes 6 classes - and makes impossible the comparison of classification performance scores.

Take away: the spectral features-based images used in (Lin et al., 2017) is an original method that plots the signals that have been band-passed into frequency bands, and transforms these plots into pictures that will then be feed to a Convolutional Neural Network (CNN). Based on the literature review, the methods that have been the mostly used in order to induce affective states are the International Affective Picture System (IAPS) and the video clips. Note that the video clips is the method that is proposed in the open-source data set DEAP, proposed by Koestra et al. (Koelstra et al., 2011), and that aims at inducing different levels of valence, arousal, as well as dominance to users. Concerning the machine learning algorithms, mostly the SVM has been widely used by experimenters, as for the cognitive workload-based literature review. However, in contrast to the literature review on cognitive workload, a lot of studied have been run with a user-independent calibration, which is an important step to go towards calibration-free systems.

Overall, classification results are not very convincing, regardless of the methods used to induce, measure or classify EEG signals. However, as we just mentioned, Koelstra et al. proposed a data set that could be useful for future studies on EEG-based emotions classification, as all researchers could test multiple signal processing and machine learning approaches on the same material, i.e., EEG signals made available in (Koelstra et al., 2011). This very data set is used multiple time in this review, and is actually the one we chose to use for our ***contribution #1***, detailed in chapter 4, aiming at testing modern machine learning algorithms in order to classify emotions.

Part II

Methods & Tools for classifying mental states

PhD Thesis Roadmap

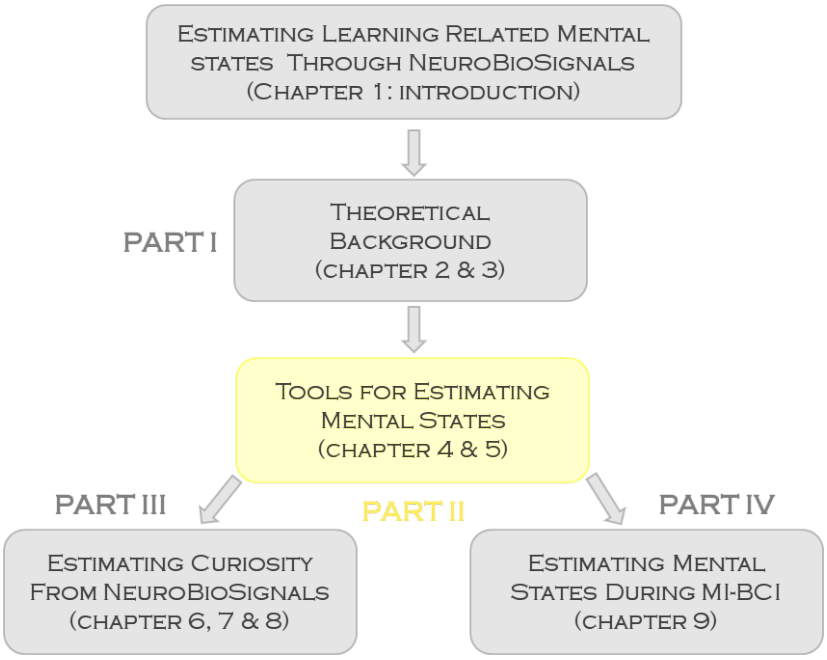

Figure 3.1: PhD thesis roadmap.

Related Papers

Peer-reviewed Journals

- Appriou, A., Cichocki, A., and Lotte, F. (2020). Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals. *IEEE SMC Magazine*, pages 1–8. (***Published)***

- Appriou, A., Pillette, L., Dutartre, D., Cichocki, A., and Lotte, F. () BioPyC, an open-source python platform for offline electroencephalographic and physiological signals classification. (***Submitted to Neuroinformatics)***

Peer-reviewed conferences

- Appriou, A., Cichocki, A., and Lotte, F. (2018). Towards robust neuroadaptive HCI: exploring modern machine learning methods to estimate mental workload from EEG signals. *CHI Conference on Human Factors in Computing*. (***Published)***

Abstract

- Appriou, A., Pillette, L., Cichocki, A. and Lotte, F. (2018) BCPy, an open-source python platform for offline EEG signals decoding and analysis. (***BCI Meeting Conference)***

- Pillette, L., Appriou, A., Cichocki, A., NKaoua, B., and Lotte, F. (2018) Classification of attention types in EEG signals. (***BCI Meeting Conference)***

4

Modern machine learning algorithms to classify workload and emotions from EEG

4.1 Research question

As explained in the introduction, estimating cognitive, affective or conative states from brain signals is a key but challenging step in the creation of passive BCI applications. Indeed, this would allow applications to monitor the users' states in real-time, and therefore adapt the interactions to individuals cognitive capabilities, i.e., optimal levels of a given cognitive state (workload, attention, etc), but also optimal levels of emotions (e.g., valence or arousal) or motivation (e.g., motivational states, curiosity). Moreover, all cognitive, affective and conative states have been shown to be involved in human learning, thus being able to estimate such states in real-time would play a major role for upgrading BCI training protocols, but this type of system could also be used as an evaluation method for HCI.

However, reliably estimating mental workload from EEG signals, over time, contexts and subjects is difficult (Mühl et al., 2014), and this observation is also valid for classifying affective states (Muhl et al., 2015) with good performance scores. Moreover, our study aiming at

reviewing papers that attempted to estimate multiple levels of cognitive workload through EEG signals (see chapter 3), and the one aiming at reviewing articles that attempted to estimate affective states (see chapter 3) - still through EEG signals - have confirmed the relatively modest performances of such passive BCIs for the time being.

In other words, based on our literature review, the classification accuracies obtained so far - mostly around 70% for workload, and around 60-65% for emotions in oscillatory-based studies - revealed the need for more robust and accurate EEG classification algorithms, in order to obtain trustable EEG-based cognitive and affective states estimators.

Therefore, the research question concerns the relevance of using modern and promising machine learning algorithms that proved efficient either in recent active BCI classification competitions (Ang et al., 2012; Yger et al., 2016), notably Riemannian geometry classifiers, or in other fields of artificial intelligence, such as Deep Learning (Lecun et al., 2015; Schirrmeister et al., 2017), in order to classify learning related mental states. Note that such algorithms have been mostly explored for EEG classification of motor tasks, but not systematically studied and compared for workload/affective states estimation. Here we formally study and compare these various algorithms as well as two new variants we propose here, for both workload, arousal and valence classification from EEG signals. We also propose guidelines about which algorithm to use in which context.

As baseline, we use two standard methods for studying workload levels/affective states classification: 1) Common Spatial Pattern (CSP) spatial filters with an LDA classifier and 2) the FBCSP (Ang et al., 2012), which is a CSP extension that won numerous active BCI competitions. Instead of using a unique specific pass-band defined by the experimenter as with CSP, FBCSP enables to optimize subject-specific frequency bands by working on a bank of band-pass filters.

Then, we study two Riemannian approaches: Minimum Distance to the Mean with Fisher geodesic filtering classifier (FgMDM) and Tangent Space Classifier (TSC). Such methods represent EEG signals as covariance matrices and classify them according to their (Riemannian) distances to prototypes of covariance matrices for each class. Note that these methods have recently won six international brain signals competitions (Yger et al., 2016).

We then propose to improve these Riemannian approaches by working on a bank of band-pass filters such as the ones used for FBCSP,

instead of using a unique band-pass filter. We name these new approaches FBFgMDM and FBTSC. Finally, we used a Convolutional Neural Network (CNN), i.e., a Deep Learning algorithm, which recently obtained promising results for many machine learning problems (Lecun et al., 2015). The strength of the CNN used is to optimize simultaneously the spatial filters, the temporal filters and the classifier, which can lead to possibly better solutions. We studied the CNN developed in (Schirrmeister et al., 2017), since it obtained promising results for motor imagery-based BCIs.

In the detailed presentation of this contribution, we first present the workload and emotion EEG data sets used, before describing each machine learning algorithm. We perform two evaluation studies: 1) a subject-specific study, with each algorithm trained on data specific to each subject, and then tested on other data from the same subject. This is the standard way current BCIs are designed, given the large between-subject variability (Blankertz et al., 2008); 2) a subject-independent study, with each algorithm trained on all data recorded from all subjects except that of the target subject, on which algorithms are tested. This is much more challenging, but if successful, would enable BCI-based monitoring without requiring any calibration for new subjects.

4.2 Methods

4.2.1 Data Sets

We propose to study two data sets, the first one focusing on cognitive workload (Mühl et al., 2014), and the second one on affective states (Koelstra et al., 2012). Both data sets are interesting for the use and validation of different types of Machine Learning algorithms.

4.2.2 Mental workload EEG data set

The data set used comes from (Mühl et al., 2014). Signals from 28 EEG electrodes (active electrodes in a 10/20 system without T7, T8, Fp1, and Fp2) were recorded from 22 users (Mühl et al., 2014). To induce mental workload variations, N-back tasks were used (explained in more details in chapter 2): in short, users had to indicate whether a letter displayed on screen was the same one as the letter displayed N letters before, in a stream of successively displayed letters. Here, 2-sec trials from a 0-back task were labeled as "low" workload, while those

from a 2-back were labelled as "high" workload. In total, 720 trials were available for each workload level and user. See the supplementary material for more information.

As introduced previously, we studied both subject-specific and subject-independent calibrations: both these methods are explained in details in chapter 2. In our study, for subject-specific calibration, the first half of each user's trials was used for training and the second half for testing. For the subject-independent calibration, the training set comprised all trials of all users except the current user used for testing, i.e., around 21*1440 = 30240 training trials. To allow the comparisons between calibration types, the testing set of each user was the same testing set as with subject-specific calibration, i.e., the second half of the trials (720 testing trials) from this user.

4.2.3 *Emotion EEG data set*

The data set used for studying emotions was the "DEAP" database (Koelstra et al., 2012). It used music-video clips to influence two types of emotion dimensions - valence and arousal, according to the circumplex model of Russell (Russell, 1980). The first step of their study consisted in making a strict selection of 40 music video clips from the Internet. The most emotional 1 min of each of these videos was automatically selected with an algorithm using informative features such as loudness, energy of the audio signals, etc.

The data set contains these 40 trials, thus corresponding to 40 music video-clips, recorded on 32 participants. EEG were recorded using 32 electrodes (placed according to the international 10-20 system). Valence and arousal levels were measured using Russell's valence-arousal scale directly after each videos, by clicking on a 1-9 continuous scale. This self-assessment system on a continuous scale makes the classes definition more complex: in DEAP (Koelstra et al., 2012) as well was in our study, 5 was kept as a threshold to split trials into two classes - low and high - for both "emotion-arousal" and "emotion-valence" data sets, making classes unbalanced. We then balances these classes by up-sampling data. Note that no artifact removal algorithm has been used in this study.

All the classifiers used were able to deal with unbalanced classes, except the CNN. We therefore up-sampled the minority class by randomly duplicating trials from this class in order to obtain balanced

classes. As with the workload data set, we study both subject-specific and subject-independent calibrations. For the subject-specific study, given the low number of trials, we performed a "leave-one-out" cross-validation. Thus, we used 40 models for each subject, each model being trained on 39 trials and tested on 1 trial. For the subject-independent study, we kept all trials of all subjects to compose the training set, except the current subject used for testing (i.e., 31*40 = 1240 trials for the training set). The testing set of each subject was composed of all trials of this subject, i.e., 40 trials.

4.2.4 *Machine learning algorithms explored*

As explained in the theorical background of this thesis, plenty of algorithms for brain signals classification are available (Lotte et al., 2018b). We propose to study algorithms that recently obtained good results for classifying motor related EEG signals, either in recent EEG classification competitions (Ang et al., 2012; Yger et al., 2016), or in independent studies (Schirrmeister et al., 2017). The existing algorithms we evaluate here were all studied on EEG-based motor imagery classification, a widely used BCI design, and obtained impressive results. Since both motor imagery, workload and emotions lead to change in EEG oscillatory activity, it is likely that methods that proved effective for motor imagery can prove effective for workload or emotion classification as well. However, to the best of our knowledge, such methods have never been tested and compared together neither on workload nor on emotions data sets nor with subject-independent calibration. We thus propose this evaluation in this chapter. We also propose some new variants of some of these algorithms. We first describe the structure of each algorithm and how we used them.

Altogether, we studied 7 algorithms. First, CSP and LDA were used as a baseline since they are widely used by the BCI community (Blankertz et al., 2008). We then explored the FFBCSP and LDA (Ang et al., 2012), a CNN (Schirrmeister et al., 2017), and four different methods based on Riemannian geometry: two existing ones, namely the Fisher geodesic Minimum Distance to the Mean classifier (FgMDM) and the Tangent Space Classifier (TSC) (Yger et al., 2016), and two new extensions we propose here to better exploit the spectral information, namely the Filter Bank FgMDM and Filter Bank TSC. For the workload data set, we assess performances using classification accuracy, i.e., the percentage of test trials correctly classified. For the emotion data set, we used balanced accuracy, i.e. the average of recall obtained on each class, since the classes were unbalanced.

(1) Common Spatial Patterns (CSP)

CSP is a widely used algorithm for binary EEG classification, for oscillatory activity-based BCI. It has been shown that changes in both workload (Brouwer et al., 2012) and emotions (Muhl et al., 2015) induce changes in EEG oscillatory activity. The CSP algorithm optimizes spatial filters, i.e., a linear combination of the original EEG signals in order to improve the EEG signal-to-noise ratio. It is done such that the variance of a spatially filtered signal, i.e. the band power of this signal, is maximized for one class and minimized for the other class. As such, CSP is particularly useful for BCI based on oscillatory activity since their most useful features are band-power features. Formally, CSP optimizes spatial filter w by either maximizing or minimizing:

$$J_{CSP}(w) = \frac{w\mathbf{X}_1\mathbf{X}_1^T w^T}{w\mathbf{X}_2\mathbf{X}_2^T w^T} = \frac{w\mathbf{C}_1 w^T}{w\mathbf{C}_2 w^T} \tag{4.1}$$

where T denotes transpose, $\mathbf{X}_i$ is the band-pass filtered training signal matrix for class i (with the samples as columns and the channels as rows) and $\mathbf{C}_i$ the spatial covariance matrix from class i. In practice, the covariance matrix $\mathbf{C}_i$ is defined as the average covariance matrix of each trial from class i (Blankertz et al., 2008).

The spatial filters w that maximize or minimize $J_{CSP}(w)$ are the eigenvectors corresponding to the largest and lowest eigenvalues, respectively, of the Generalized Eigen Value Decomposition of matrices $\mathbf{C}_1$ and $\mathbf{C}_2$. In this study, we used six filters, corresponding to the three largest and three lowest eigenvalues, as recommended in (Blankertz et al., 2008). Once these filters are obtained, we use as CSP features $f = \log(w\mathbf{X}\mathbf{X}^T w^T)$, i.e., the band power of the spatially filtered signals. We used these features as input to an LDA classifier. The CSP+LDA algorithm is one of the most popular approach since it has been widely and successfully used for BCIs based on motor imagery (Blankertz et al., 2008), as well as for workload classification, although to a minor extent (Roy et al., 2016b). The CSP requires EEG signals to be band-pass filtered in a specific narrow frequency band. The Alpha rhythm (8-12Hz) being known to vary according to both workload (Mühl et al., 2014) and emotions (Koelstra et al., 2012), we applied CSP after band-pass filtering in 8-12 Hz. We selected 3 pairs of CSP spatial filters, as recommended in (Blankertz et al., 2008), to obtain 6 band-power features used to train a Linear Discriminant Analysis (LDA) classifier. This method was used as the baseline.

(2) Filter Bank Common Spatial Patterns (FBCSP)

The FBCSP is an algorithm that optimizes both spatial and spectral filters. To do so, FBCSP first filters EEG signals into multiple frequency

bands using a filter bank. Here we used nine band-pass filters in 4Hz-wide bands (in 4-8 Hz, 8-12 Hz, ..., 36-40 Hz) as in (Ang et al., 2012). Then, for each band-passed signals, CSP is used to optimize two spatial filter pairs. From the resulting 36 features (9 bands $\times$ 4 CSP filters per band), the four most relevant ones were selected using minimal Redundancy Maximal Relevance (mRMR) (Peng et al., 2005a), and used as input to an LDA. The FBCSP algorithm proved its efficiency when winning the Fifth International BCI competition (Ang et al., 2012).

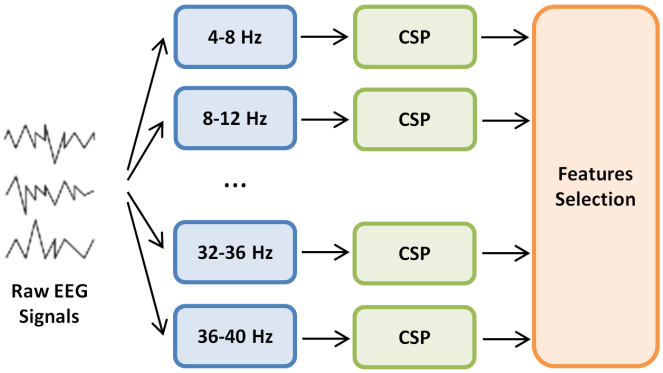

Figure 4.1: Principle of Filter Bank Common Spatial Patterns (FBCSP): 1) band-pass filtering the EEG signals in multiple frequency bands using a filter bank; 2) optimizing CSP spatial filter for each band; 3) selecting the most relevant filters (both spatial and spectral) using feature selection.

(3) Riemannian Geometry

Riemannian approaches represent EEG trials as covariance matrices, which are symmetric positive definite (SPD) matrices, and manipulate them with an appropriate geometry, the Riemannian geometry (Congedo et al., 2017; Yger et al., 2016). Classifiers based on such geometry are called Riemannian Geometry Classifiers (RGC).

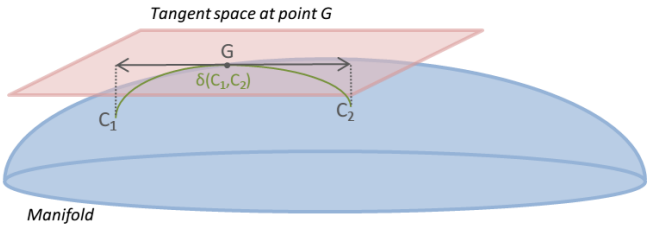

Figure 4.2: Schematic representation of a Riemannian manifold with matrix G, the Riemannian average of covariance matrices C_1 and C_2. The tangent space to the Riemannian manifold at point G is represented in red.

First, in a Riemannian $manifold$ we can estimate $intrinsic$ non-Euclidean distances between two SPD matrices, i.e. two points (here $\mathbf{C}_1$ and $\mathbf{C}_2$), using the Riemannian distance:

$$\delta^2(\mathbf{C}_1, \mathbf{C}_2) = \sum_n log^2 \lambda_n(\mathbf{C}_1^{-1}\mathbf{C}_2), \tag{4.2}$$

where $\lambda_n(\mathbf{M})$ is the n^{th} eigenvalue of matrix $\mathbf{M}$. The set of tangent vectors to point $\mathbf{G}$ on the $manifold$ defines the manifold tangent space at $\mathbf{G}$. Figure 4.2 shows the tangent space at point $\mathbf{G}$, which is the centroid (mean) of $\mathbf{C}_1$ and $\mathbf{C}_2$. More generally, any SPD matrix $\mathbf{C}_i$ can be projected onto the tangent space at point $\mathbf{G}$ using:

$$\mathbf{S}_i = Log_{\mathbf{G}}(\mathbf{C}_i) = \mathbf{G}^{1/2} logm(\mathbf{G}^{-1/2}\mathbf{C}_i\mathbf{G}^{-1/2})\mathbf{G}^{1/2}, \tag{4.3}$$

$\mathbf{S}_i$ being the projection of $\mathbf{C}_i$ onto the tangent plane, and $logm(\cdot)$ denotes the logarithm of a matrix.

Considering the principles we listed about Riemannian geometry, multiple methods for classification are possible. In this chapter, we studied two existing RGCs - the Mean classifier (FgMDM) and the Tangent Space classifier (TSC) - and introduced two new ones - the Filter Bank TSC (FBTSC) and the Filter Bank FgMDM (FBFgMDM).

Existing methods:

- ***FgMDM*** (Barachant et al., 2010): FgMDM projects training matrices $\mathbf{C}_i$ onto the tangent space at point $\mathbf{G}$ (the mean of all training data) using Eq. (4.3), to obtain matrices $\mathbf{S}_i$. Then, a Fisher geodesic filter is obtained by optimizing an LDA classifier on $\mathbf{S}_i^{vec}$, the vectorized upper-triangular elements of $\mathbf{S}_i$, to discriminate classes using such vectors. This results in a matrix of weights $\mathbf{W}$ = LDA($\mathbf{S}_i^{vec}$). The projected SPD matrices $\mathbf{S}_i$ from both the training & the testing sets are then filtered with weights $\mathbf{W}$, using $\hat{\mathbf{S}}_i = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_i^{vec}$, where $\hat{\mathbf{S}}_i$ denotes the geodesic filtered SPD matrices from $\mathbf{S}_i$. Then, these filtered matrices $\hat{\mathbf{S}}_i$ are projected back onto the manifold using equation:

$$\hat{\mathbf{C}}_i = Exp_{\mathbf{C}}(\hat{\mathbf{S}}_i) = \mathbf{G}^{1/2}expm(\mathbf{G}^{-1/2}\hat{\mathbf{S}}_i\mathbf{G}^{-1/2})\mathbf{G}^{1/2}, \tag{4.4}$$

  where $\hat{\mathbf{C}}_i$ are the filtered SPD matrices projected onto the manifold and $expm(\mathbf{M})$ denotes the exponential of matrix $\mathbf{M}$. Finally, this approach uses a Minimum Distance to the Mean classifier to classify testing geodesic filtered matrices $\hat{\mathbf{C}}_i$. To do so, during the training step, the class centroids $\mathbf{G}_k$ of each class k are computed by averaging the geodesic filtered covariance matrices $\hat{\mathbf{C}}_i^k$ from each class k. During testing, the Riemannian distances between the testing geodesic filtered matrix $\hat{\mathbf{C}}_j$ and each class centroid $\mathbf{G}_k$ are first calculated, using Eq. (4.2). The matrix $\hat{\mathbf{C}}_j$ is assigned class label k for which the centroid $\mathbf{G}_k$ is the nearest. In our study, FgMDM was applied on EEG band-pass filtered in 8-12Hz, as for the CSP.

- ***TSC***: TSC first projects all training SPD matrices $\mathbf{C}_i$ onto the tangent space at point $\mathbf{G}$ (the mean of all training matrices). Then, it uses any classifier such as LDA, SVM or Logistic Regression (LR) on the vectorized upper-triangular elements of the projected matrices (Barachant et al., 2012a). We used LR with L_2 regularization (with the default $C = 1.0$ in scikit-learn (Pedregosa et al., 2011a)). As for FgMDM, TSC used data filtered in 8-12Hz.

New methods:

- ***Filter Bank FgMDM (FBFgMDM)***: Contrary to FgMDM which exploits EEG signals in a single frequency band, this method applies FgMDM in multiple bands separately, and combines the resulting distances to exploit additional spectral information. This should possibly improve classification performances, as FBCSP did to improve CSP. To do so, FBFgMDM first filters EEG signals in multiple bands using a filter bank, as for FBCSP. Here we used the same bands as the FBCSP, i.e., 4-8 Hz, 8-12 Hz, ..., 36-40 Hz. Then for the EEG signals in each frequency band j, this method first uses a regular FgMDM, i.e., it computes the Riemannian distances $\delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij})$ between a geodesic filtered SPD matrix $\hat{\mathbf{C}}_{ij}$ and each class centroid $\mathbf{G}_{kj}$. We thus obtain nine bands $\times$ N_k classes such distances (here, $N_k = 2$). Then, from all nine bands j, the four most useful ones for classification are selected with mRMR feature selection (Peng et al., 2005a) on the Riemannian distances $\delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij})$ used as features, on the training set. For testing, we compute the squared Riemannian distances for the four bands selected using mRMR only and sum them:

$$\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i) = \sum_{j \in \Omega} \delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij}), \tag{4.5}$$

where Ω is the set of frequency bands selected with mRMR. We thus obtain k new distances $\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i)$ to each class k for each trial i, in our case $k \in \{1,2\}$. $\hat{\mathbf{C}}_i$ defines the SPD matrix filtered using FGDA algorithm in the tangent space at point $\mathbf{G}_k$ associated to a trial i. The classification prediction results in choosing the class y_i for which the summed squared distance to the centroid is the smallest:

$$y_i = \underset{k}{\operatorname{argmin}}(\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i)). \tag{4.6}$$

where k denotes the class label, $\mathbf{G}_k$ is the centroid of the class k, $\hat{\mathbf{C}}_i$ is the filtered SPD matrix for trial i and y_i the predicted class for trial i (in our case, $y_i \in \{1,2\}$).

- ***Filter Bank TSC (FBTSC)***: FBTSC also exploits more spectral information than TSC, by using a filter bank. FBTSC indeed projects matrices $\mathbf{C}_{ij}$, band-pass filtered in bands 4-8 Hz, 8-12 Hz, ..., 36-40

Hz, to the tangent space using Eq. (4.3). Then, the probabilities that the vectorized upper-triangular elements of the projected SPD matrix $\mathbf{S}_{ij}$ belongs to class k is calculated using standard classification algorithms with probabilistic outputs, such as LDA or LR. Here we used LR that directly provides such probability with its softmax function. Since we did so for nine frequency bands, in two classes k, we ended up with nine pairs of probabilities. From these pairs of probabilities, the four most relevant are selected using mRMR on the training set. Finally, we multiplied the probabilities associated to each class k, for the selected bands only, to end up with two probabilities:

$$\mathcal{P}_{ki} = \prod_{j \in \Omega} \mathcal{P}_{kij}, \tag{4.7}$$

where $\mathcal{P}_{ki}$ is the probability of trial i to be part of class k, and $\mathcal{P}_{kij}$ the probability of a projected SPD matrix $\mathbf{S}_{ij}$, band-pass filtered in frequency band j, to be part of class k. The classification prediction results in choosing the class y_i for which $\mathcal{P}_{ki}$ is the highest:

$$y_i = \underset{k}{\operatorname{argmax}}(\mathcal{P}_{ki}). \tag{4.8}$$

where k denotes the class label, and $\mathcal{P}_{ki}$ the probability of a projected SPD matrix $\mathbf{S}_i$ to be part of class k.

(4) Convolutional Neural Networks (CNN)

Deep Networks are artificial neural networks with multiple layers of artificial neurons, which makes them able to approximate efficiently any function (Lecun et al., 2015). There are different types of architecture for neural networks, such as Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN). Here, we chose to study Deep Learning with CNN, since it has already improved many fields such as computer vision (Lecun et al., 2015), and was also proved effective for motor imagery-based BCIs (Schirrmeister et al., 2017).

Shortly, a CNN is a feedforward neural network with at least one convolutional layer. This type of network flows information uni-directionally from the input to the hidden layers and finally to the output. A recent study presented a new type of CNN dedicated to motor task classification in EEG: the Shallow ConvNet (Schirrmeister et al., 2017).

The shallow ConvNet architecture consists in a 3-layer CNN with parameters that have been experimentally tested and validated by their authors (Schirrmeister et al., 2017). The first layer is a convolutional layer along the temporal dimension, while the subsequent one is a convolutional layer along the spatial dimension, i.e., over EEG electrodes. If we compare this process to the FBCSP, the first temporal convolu-

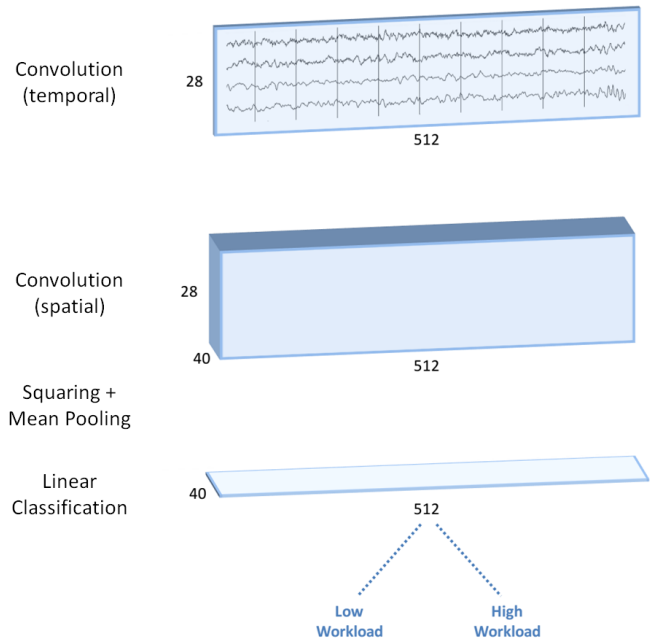

Figure 4.3: Example of the Shallow ConvNet architecture applied for the cognitive workload classification (low vs high).

tion aims at optimizing band-pass filters, and the spatial convolution aims at optimizing spatial filters. Then, signals are squared, a mean pooling is performed (to compute signals band power) and the CNN ends by a fully connected linear classification layer. Overall this CNN thus processes EEG data similarly to the FBCSP and LDA. In contrast to FBCSP, all these filters are optimized simultaneously though, which made it outperform the FBCSP on motor EEG signals (Schirrmeister et al., 2017). Here, we explored this CNN with the implementation and hyper-parameters from (Schirrmeister et al., 2017) in order to classify mental workload, valence and arousal from EEG. Note that the Shallow ConvNet uses minimally preprocessed EEG signals as input, so we filtered them in 4-40 Hz.

4.3 *Results*

We present here the classification performances of each algorithm with both subject-specific and subject-independent calibrations. Figure 4.4 summarizes the mean performance obtained on the workload data set first, then on both the valence & arousal data sets.

As a reference, the statistical chance levels (Combrisson and Jerbi, 2015) were estimated at 50.47% for the mental workload study (1440 trials and 22 subjects) and 52.27% for the affective state study (40 trials and 32 subjects). Note that for statistical tests (ANOVA), we checked

data set	study	CSP	FBCSP	FgMDM	TSC	FBFgMDM	FBTSC	CNN
emotion-valence	subject-specific	57.5904	59.1921	58.8734	59.4658	61.0144	61.0934	46.32
	subject-independant	52.5	55.2344	47.9688	49.1406	48.4375	48.75	48.0469
emotion-arousal	subject-specific	58.2586	59.1315	60.0404	60.0404	60.3008	60.5969	40.1531
	subject-independant	55.7031	55.3125	56.25	55.7813	51.6406	53.2812	47.5
workload	subject-specific	67.0041	68.5089	69.9429	68.4964	70.3385	68.7242	72.7296
	subject-independant	58.0465	60.0742	58.3072	58.3072	61.2989	60.1805	63.7357

Figure 4.4: Mean classification accuracy for each algorithm with both subject-specific and subject-independent calibrations. The best performance of each study is in green, the worst in red.

the data sphericity, and used Greenhouse-Geisser (GG) correction in ANOVA if differences were observed, with F if no difference was observed.

4.3.1 *Workload study*

Performances obtained by each algorithm on this data set are reported on Figure 4.5. We performed a 2-way ANOVA with repeated measures to evaluate the performances of factor *Algorithm* according to factor *Calibration Type* (subject-specific vs subject-independent). The normality is respected for all conditions, and the sphericity has been tested as well.

It revealed a main effect of *Algorithm* [GG(1,22)=0.517, p=0.001], and *Calibration Type* [F(1,22)=33.308, $p \leq 0.0001$], but not for *Calibration Type*Algorithm* [GG(1,22)=0.558, p=0.618].

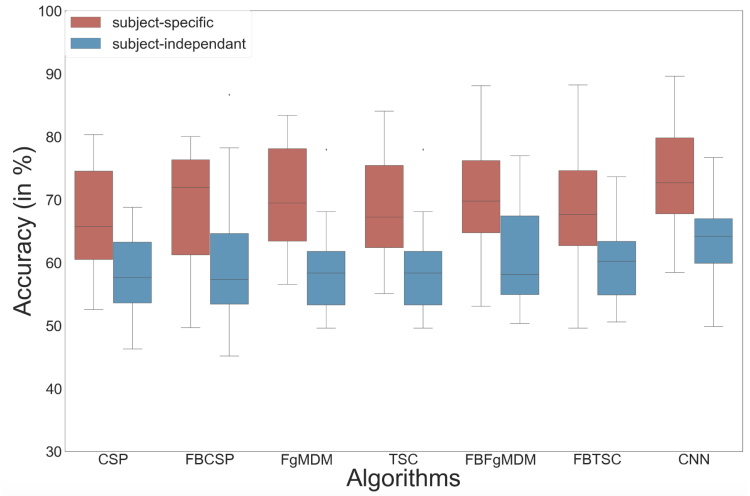

Figure 4.5: Classification accuracy of each algorithm on the workload data set.

Post-hocs analyses - Student t-test for paired samples - with Bonferroni adjustments showed no significant differences between algorithms in the subject-specific or subject-independent studies. However, performances obtained suggested better (but non-significantly so) results with the CNN compared to other algorithms, in both subject-specific and subject-independent studies. Riemannian geometry clas-

sifiers (RGC), in particular the newly proposed ones (FBfgMDM and FBTSC) provided the second best performances, just after the CNN. On the other hand, the baseline CSP+LDA obtained the worst results.

4.3.2 *Valence*

The same algorithms have been evaluated on the emotion-valence data set, and the balanced classification accuracies obtained are reported on Figure 4.6. The normality is respected for all conditions, and the sphericity has been tested as well.

We ran a 2-ways ANOVA for repeated measures to evaluate the impact of *Algorithm* on the emotion-valence data set, regarding the *Calibration Type*. The results showed significant differences in *Algorithm* [GG(1,32)=6.918, p=0.002], *Calibration Type* [F(1,32)=21.732, p<0.0001] and *Calibration Type*Algorithm* [GG(1,32)=5.374, p=0.003].

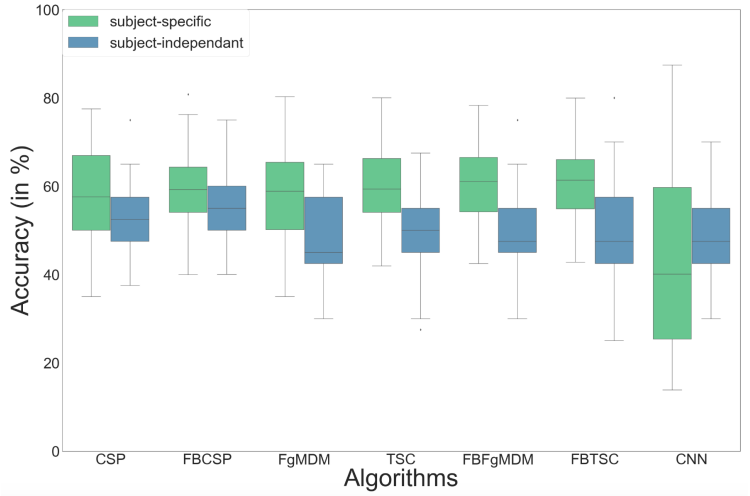

Figure 4.6: Balanced classification accuracy on the emotion-valence data set.

Post-hoc analyses - Student t-test for paired samples - with Bonferroni corrections showed a significant difference between FBTSC and CNN for subject-specific calibration [$perf_{FBTSC}$ = 61.09%, $perf_{CNN}$ = 46.32%; $p \leq 0.05$]. No algorithm showed better results than others with the subject-independent calibration. Overall, FBFgMDM and FBTSC obtained the best accuracy (both about 61%) for subject-specific calibration, while FBCSP obtained the best performances for the subject-independent one (55.2%).

4.3.3 *Arousal*

The balanced classification accuracies for the emotion-arousal data set are reported on Fig. 4.7. The normality is respected for all conditions, and the sphericity has been tested as well. We then per-

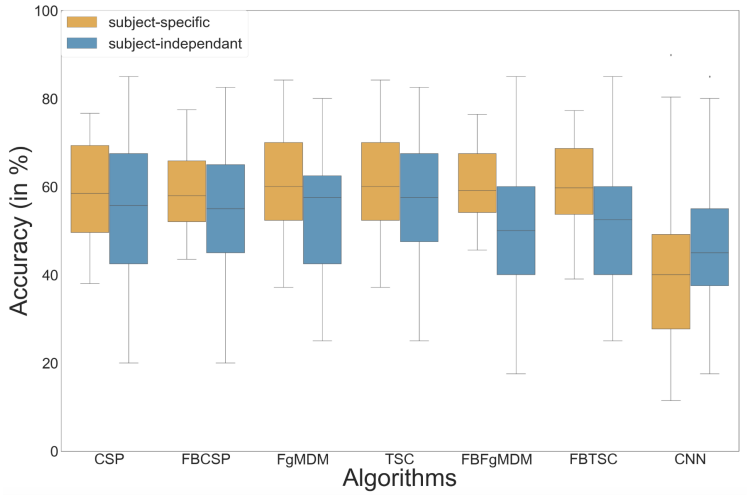

Figure 4.7: Balanced classification accuracies on the emotion-arousal data set.

formed a 2-way ANOVA with repeated measures, with factor *Algorithms* and *Calibration Type*. Results revealed significant effects for *Algorithms* [GG(1,32)=9.177, $p \leq 0.0001$], *Calibration Type* [F(1,32)=4.262, p=0.048] and *Algorithms*Calibration Type* [GG(1,32)=3.894, p=0.008].

Post-hoc analyses -Student t-test for paired samples- with Bonferroni corrections showed significant differences with the subject-specific calibration between CNN and all other classifiers (see results in the supplementary material). No algorithm showed better results than others with the subject-independent calibration. Overall the best results were all obtained by RGCs, FBFgMDM and FBTSC for the subject-specific calibration, and FgMDM for the subject-independent one.

4.4 Discussion, Conclusion and Future Work

In this chapter, we explored promising classification algorithms, both existing and new ones, to classify mental workload and emotions (valence and arousal) from EEG signals, with both subject-specific and subject-independent calibration. Altogether we studied CSP+LDA, FBCSP+LDA, four RGCs (FgMDM, TSC and two new variants proposed here: FBFgMDM and FBTSC), and CNN.

We chose two data sets 1) the first one from the paper (Mühl et al., 2014), that induced two cognitive workload levels, i.e. classification of trials labeled as high workload vs trials labeled as low workload, with 2-seconds time windows 2) the second one from (Koelstra et al., 2012) where affective states - valence and arousal - were both studied

with unbalanced subjective labeling high/low valence and high/low arousal, with 60-seconds time windows.

On the workload data set, the original authors used the FBCSP and obtained a classification accuracy of 69,4% with subject-specific calibration. This result is above the chance level (50.46%). With the same calibration and same algorithm, we obtained 68,5% classification accuracy: the difference can be explained by the use of different frequency bands and feature selection methods. In (Mühl et al., 2014), no subject-independent calibration was studied. Concerning the emotion data sets (Koelstra et al., 2012), the original authors used a NBC and obtained an accuracy of 57.6% for valence, and 62% for arousal with subject-specific calibration (chance level of 52.27%). They did not run a subject-independent calibration study.

The first results to highlight are the CNN classification performances we obtained across the different conditions and data sets. Indeed, this algorithm has a higher mean accuracy (although non-significantly so) than the original authors' results, the baseline CSP+LDA, and more importantly than both FBCSP and Riemannian methods, with both subject-specific and subject-independent calibrations on the workload data set. Moreover, obtaining reasonable performances in a subject-independent calibration from only two seconds of EEG data and only 21 users for calibration, makes the CNN particularly interesting to design calibration-free neuroadaptive technologies in the future. By contrast, this algorithm significantly under-performed with both subject-specific and subject-independent calibrations on both the valence and arousal data sets. All algorithms indeed outperformed this CNN in all conditions on the emotion data sets.

Multiple factors could explain the observed algorithm performances. First, the number of trials that are used for training models is important. In (Schirrmeister et al., 2017), authors tested the Shallow ConvNet on multiple motor-imagery data sets (from 288 to 1168 trials), and often obtained significantly better performances with the CNN than with FBCSP. In our study, the workload data set contained 720 training trials whereas both valence and arousal data sets contained 39 training trials only (with cross validation calibration). This might suggest that the CNN could be useful for mental state classification, but only when large amount of training trials are available (around 700 in our study), which is not always possible. However, other factors also differ between both data sets studied and could also explain differences in CNN performances, including the EEG epochs length (2s epochs for workload and 60s epochs for emotions), and the nature

of the mental states studied (workload vs emotions). Indeed, emotions are thought to originate from deep brain areas (Muhl et al., 2015) and are thus known as being difficult to estimate from EEG. In the future, deeper analyses would thus be needed to fully disentangle these factors, by systematically varying the types (e.g., motor-imagery, workload, emotions, curiosity, etc), time-windows (2sec, 5, 10, 30, 1min, 2min, 15, 30) and variations of the number of training trials (50, 100, 200, 500, 1000, etc).

Another relevant result is the promising classification performances of the proposed RGCs. Indeed, FBTSC and FBFgMDM outperformed the results from the data sets' authors in most conditions/data sets. Moreover, FBFgMDM with subject-specific calibration, and FBFgMDM and FBTSC with subject-independent calibration, reached higher mean accuracies than all other algorithms, except the CNN on the workload data set. More interestingly, the low number of trials in the emotion data sets did not seem to affect their performances since they also reached the highest mean accuracies on both the emotion-valence and emotion-arousal data set, both with subject-specific calibration. These promising results compared to standard RGCs (TSC and FgMDM), are probably due to the extra spectral information extracted with the filter bank, and our study enabled us to quantify this gain. If it seems natural to observe higher mean accuracies by looking at multiple frequency bands, it would be interesting to look at which frequency bands have brought more information in deeper analyses.

Finally, FBCSP+LDA obtained a higher mean accuracy than CSP+LDA, although not significantly so, in all conditions/data sets, and the higher overall mean accuracy for valence classification with subject-independent calibration. However, it did not obtain higher mean accuracies than others in any other condition. It should be noted that such results reflect the performances obtained in offline evaluation. As such they are likely to be similar to performances obtained in offline or open-loop mental state monitoring, e.g., for noeuroergonomics (ex: mental workload monitoring) or neuromarketing (ex: emotion monitoring). The performances are likely to change in closed-loop applications, with neuroadaptive technologies, and will thus need to be evaluated in this context as well.

Such results enable us to suggest guidelines about which algorithm to use for mental states classification from EEG. First, the CNN is recommended for mental workload classification with both subject-specific and subject-independent calibration, but seems to need a large amount of training trials (at least several hundreds). It should thus

probably be avoided for data sets with little training data (i.e., a few dozens). Second, Filter Bank RGCs (FBTSC and FBFgMFM) should also be recommended to obtain good classification performances notably with subject-specific calibration, for both workload and emotion classification, whatever the amount of training data. However, such methods do not seem suitable for subject-independent classification with little training data and/or for emotion classification. They seem suitable for subject-independent classification of workload with large amount of training data though.

Our results also confirmed that passive BCIs with subject-independent calibration is possible but very challenging and with much lower accuracies. Similarly, affective state classification in EEG is possible but much more challenging than workload estimation. However, those suggestions imply computational costs that will differ from an algorithm to another. Indeed, using the FB RGCs or the CNN will require a long calibration time, when the testing phase might also be time consuming and has to be considered before to go towards online uses.

For the emotion data set, we labelled trials as in the original paper to allow comparisons, i.e., with a global – subject-independent – partition between low/high valence/arousal trials, based on the SAM ratings. Note that better methods for partitioning low/high trials in a per-subject basis can also be used (Clerico et al., 2018) in the future, to limit the class imbalance. Still in the future, other deep learning architectures, notably Recurrent Neural Networks (RNN) (Lecun et al., 2015) may prove promising for EEG classification and passive BCIs as well. For example, Recurrent have obtained promising performances in Natural Language Processing and videos classification (Yang et al., 2017), and may thus prove useful for passive BCI as well. It would also be interesting to study whether CNN and RGCs can be used to estimate robustly other cognitive states such as fatigue, curiosity or engagement, and how well the proposed RGCs perform on motor imagery data for active BCIs. Similarly, it would be interesting to study the new RGCs (FBfgMDM and FBTSC) on Motor Imagery-based BCIs, since TSC and FgMDM already performed well on such data (Yger et al., 2016), as well as to estimate other mental states, e.g., attention or fatigue. Later in this thesis (in chapter 5), we will see that we tested both FBFgMFM and FBTSC on Motor Imagery-based BCIs, and obtained good results. Altogether, our results suggested that CNN and the proposed filter bank RGCs are valuable machine learning tools for scientists aiming at decoding cognitive and affective states from EEG signals.

5

BioPyC, a Python platform for offline neurophysiological signals classification

5.1 Research question

As seen in the introduction, although promising, non-invasive BCIs - and thus non-invasive passive BCIs - are still barely used outside laboratories due to their poor robustness with respect to noise and environmental conditions. In other words, they are sensitive to noise, outliers and the non-stationarity of EEG signals (Wolpaw and Wolpaw, 2012) (Erp et al., 2012). As highlighted in the chapter 4 of this part, the computing machinery considerably evolved in the last decades, and numerous signal processing and machine learning algorithms for brain signals classification have been developed (Lotte et al., 2018a). It is therefore important to be able to test many promising new algorithms, resulting from both signal processing and machine learning research, on different data sets, related to different paradigms, in order to have the most efficient tools possible for their future uses. Indeed, the chapter 4 of this part, consisting at studying tools for estimating mental states such as affective states or cognitive workload through EEG signals, showed it is essential to test such promising machine learning algorithms on studies aiming at estimating users' mental states.

However, this type of study can be complicated to set up online, because of the costs in terms of time, e.g., equipping a participant with different sensors to measure the activity in order to analyze it online, or in terms of calculation, given that some algorithms such as Deep Learning may require a lot of computing resources and would be difficult to run online. However, such studies are possible offline, by applying recent and promising classification algorithms on existing data sets, and are actually widely used in the BCI community, notably to compare various algorithms on the exact same data (Jayaram and Barachant, 2018). While offline studies are simpler to set up than online analyses, they still require specific tools and skills to do them effectively and efficiently. Indeed, using such algorithms requires expertise in programming (e.g., in MATLAB or Python), signal processing, machine learning, as well as statistics for analyzing the resulting performances of the different algorithms, whereas many BCI researchers come from diverse backgrounds such as cognitive science, neuroscience or psychology, and might not master all those skills.

If multiple BCI toolboxes are available, they all require skills such as the ones listed above, and most of them are focused on EEG and other brain signals, but usually do not include processing tools for other bio signals. It therefore highlights the need for convenient toolboxes that would be free, open source and equipped with a graphical interface that would allow users to process and classify EEG and other biological signals offline without any programming skill. Therefore, we propose here the design and implementation of such a toolbox, which would answer all these requirements.

We propose BioPyC, an open-source and easy-to-use platform for offline signal processing and classification. BioPyC is free of charge, permissively licensed (AGPL - see https://choosealicense.com/licenses/agpl-3.0/) and written in Python (Rossum, 1995), an open source programming language that is not only backed by an extensive standard library, but also by vast additional scientific computing libraries. This toolbox allows users to make offline EEG and bio signals analyses, i.e., to apply signal processing and classification algorithms to neurophysiological signals such as EEG, HR, EDA or respiratory system[1]. In order to facilitate those analyses, BioPyC offers a graphical user interface (GUI) based on Jupyter (Pérez and Granger, 2007) that allows users to handle the toolbox without any prior knowledge in computer science or machine learning. Finally, with BioPyC, users can apply and study algorithms for the main steps of bio signals analysis, i.e., pre-processing, signal processing, classification, statistical analysis and data visualiza-

[1] So far, on gitlab, the implementation of the toolbox includes more modules for EEG analysis than for physiological signals analysis, but new modules will be integrated soon, as they have been implemented for our study in chapter 8. See https://gitlab.inria.fr/potioc/BioPyC

tion, as Figure 5.1 summarizes.

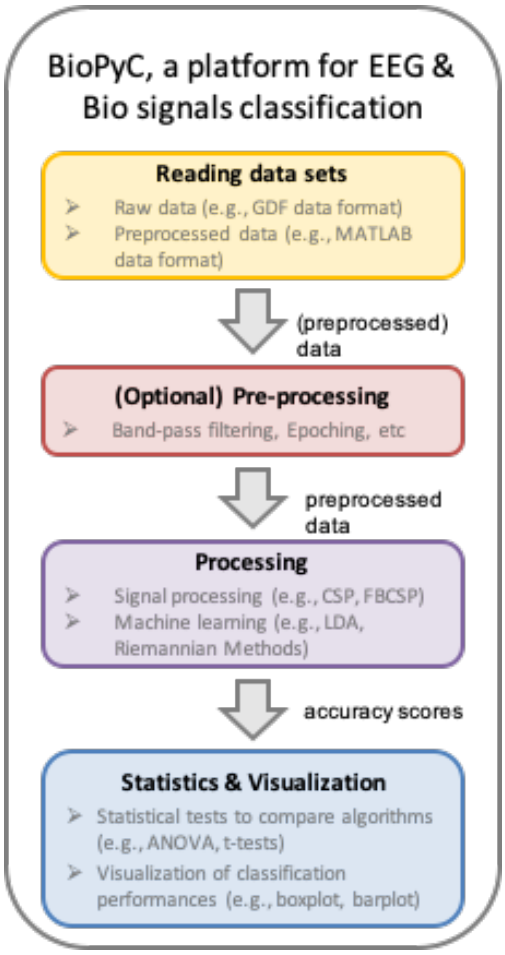

Figure 5.1: BioPyC data flow: the 4 main modules allow users to follow the standard BCI process for offline EEG and biosignal processing and classification.

BioPyC enables users to work on two types of data sets, either raw data sets which require the subsequent use of the pre-processing module (described below), or pre-processed data sets which will allow users to directly apply signal processing and machine learning algorithms on them. The pre-processing module of BioPyC offers basic features, i.e., band-pass filtering, cleaning and epoching raw EEG signals. Next, for the processing step, BioPyC offers two modules that enable users to use signal processing tools such as spatial filters, but also machine learning algorithms for the classification of neurophysiological signals. Finally, another module enables users to automatically apply appropriate statistical tests on the classification performances, to compare algorithms, and obtain visualization plots describing those performances.

While some of the existing toolboxes such as MOABB (Jayaram and Barachant, 2018), pyRiemann (Barachant and King, 2015) or MNE

(Gramfort et al., 2013), listed in the "state-of-the-art" section below, have features in common with BioPyC, e.g., programming language or supported operating system, as well as modules of the above-mentioned BCI process, e.g. signal processing or data visualisation, none of them allow users to apply and study the classification algorithms for both EEG and physiological signals. BioPyC is indeed the first one which allows to study and compare classification algorithms for neurophysiological signals (at the moment EEG, respiration and heart rate), offering modules for all the steps of this offline BCI process. Moreover, our platform enables users to run such studies by interacting with a Jupyter-based GUI, and thus does not require any skills in programming, when all other Python-based toolboxes require such skills. Concerning the statistics, both tests and visualization for comparing classification algorithms performances are done automatically by BioPyC, again facilitating users' analyses: to the best of our knowledge, no other toolbox enables automatic statistical testing.

In the detailed presentation of this contribution, we first present the state-of-the-art of BCI platforms, the different features they offer, and conclude on the distinctive features that make BioPyC unique. In Section 3, we explain BioPyC modules and the data flow in more details, including the different algorithms that have been implemented for the classification of both EEG and bio signals, as well as the statistical tests. Then, in Section 4, we present results of the multiple analyses that have been made with BioPyC so far. On the one hand, BioPyC has been applied to a widely used mental task dataset, i.e., the "BCI competition IV dataset 2a" (Ang et al., 2012). On the other hand, we used it for studying different users' mental states, e.g., cognitive workload, emotions and attention states. Finally, the discussion and conclusion come in Section 5, followed by the future works and improvements from which BioPyC could benefit in Section 6.

5.2 *State-of-the-art BCIs platforms*

So far, several platforms for online experiments - i.e., BioSig (Schölgl et al., 2011), BCI2000 (Schalk et al., 2004), OpenViBE (Renard et al., 2010b), TOBI (Müller-Putz et al., 2011), Timeflux (Clisson et al., 2019) or BCI++ (Perego et al., 2009) - and also for offline studies - i.e., MOABB, MNE, EEGLAB, PyEEG - have been developed for researchers in order to build set ups that would best suit their needs. They all have modules dedicated to the various BCI processing steps: data acquisition, signal processing, classification, statistical hypothesis testing and

	online	Graphical User Interface	statistical modelling	data visualization	programming languages	supported systems	license
BioPyC	offline	yes	yes	yes	python	Windows, Mac OS X, Linux	AfferoGPL
MOABB	offline	no	yes	yes	python	Windows, Mac OS X, Linux	BSD
MNE	offline	no	no	no	python	Windows, Mac OS X, Linux	BSD
scikit-learn	offline	no	yes	no	python	Windows, Mac OS X, Linux	BSD
BioSig	online	yes	yes	yes	matlab, opt: {C++, python}	Windows, Mac OS X, Linux	GPL
PyEEG	offline	no	yes	yes	python	Windows, Mac OS X, Linux	GPL
gumpy	online/offline	no	no	yes	python	Windows, Mac OS X, Linux	MIT license
Wyrm	online	no	no	yes	python	Windows, Mac OS X, Linux	MIT license
pyRiemann	offline	no	no	no	python	Windows, Mac OS X, Linux	BSD
Timeflux	online	yes	no	no	C++, opt: {matlab, python}	Windows, Mac OS X, Linux	MIT license
BCI2000	online	yes	no	no	C++, opt: {matlab, python, Lua}	Windows, Mac OS X	GPL
OpenViBE	online	yes	no	no	C++	Windows, Linux	AfferoGPL
TOBI	online	no	no	no	matlab	Windows, Linux	GPL
BCILAB	online	yes	yes	yes	C++, matlab	Windows, Mac OS X, Linux	GPL
BCI++	online	yes	no	no	nan	Windows	GPL

Figure 5.2: Comparison of main features of existing toolboxes having modules for EEG signal processing and classification. BioPyC values for each feature are written in black; values of features that are similar to BioPyC's ones are written in green; and finally values of features that differ from BioPyC's ones are written in grey. "opt" stands for "optional" in the figure.

visualization (Brunner et al., 2013). We synthesized in Figure 5.2 the features of each of these existing platforms, i.e., online vs offline studies, the availability of a Graphical User Interface (GUI), the existence of modules for statistical testing or data visualization, the programming language, the supported systems as well as the type of license - in order to list the strengths and weaknesses of each one. In the following, we compare the different platforms based on each of these features.

5.2.1 *Graphical User Interface (GUI)*

As explained above, it is important to develop toolboxes with GUIs for the BCIs community, when many of researchers do not come from computer science backgrounds, especially cognitive scientists, neuroscientists or psychologists. On Figure 5.2, all Python-based toolboxes, i.e., MOABB (Jayaram and Barachant, 2018), MNE (Gramfort et al., 2013), PyEEG (Bao et al., 2011), pyRiemann (Barachant and King, 2015), gumpy (Tayeb et al., 2018) and Wyrm (Venthur et al., 2015) - suffer from the lack of such interfaces.

5.2.2 *EEG signal processing*

All platforms on Figure 5.2 naturally have an EEG signal processing system for classification, more or less elaborated, based on three classical steps, i.e., pre-processing, signal processing and classification.

Pre-processing: this step consists for example in band-pass filtering raw data into specific frequency bands, epoching raw data into trials or removing artifacts, among other. While all toolboxes propose some forms of pre-processing, the ones with the most advanced preprocessing tools, including plenty of methods, are MNE amd BCILAB.

Signal Processing: this step allows users to apply spatial or temporal filters on the signals and in extracting features from them. Most existing platforms offer such filters and basic features, e.g. the Common Spatial Pattern (CSP) filter or band power features are widely used for EEG signal analysis (Blankertz et al., 2008).

Classification: all platforms also offer machine learning algorithms, from the simplest ones, such as Linear Discriminant Analysis (LDA) (Lotte et al., 2018a) for most of them, to more complex ones, such as Riemannian geometry classifiers (Yger et al., 2016) for pyRiemann (Barachant and King, 2015) and MOABB (Jayaram and Barachant, 2018), Deep Learning for gumpy (Tayeb et al., 2018) and various feature extraction methods in PyEEG (Bao et al., 2011).

5.2.3 *Statistical Modeling and Data Visualization*

The last step is divided between the statistical analysis and the visualization of performance results obtained by the machine learning algorithms. First, the visualization allows users to obtain graphs in order to have an overview of the classification results. Statistical modelling, on the other hand, consists of using statistical tests to compare the classification performances of the machine learning algorithms. This step is of primary importance when comparing classification algorithms since statistical tests allow to define, for a given study, which algorithm is the most likely to recognize patterns in neurophysiological signals. To the best of our knowledge, none of the platforms for offline signals analysis listed in Figure 5.2 have such dedicated features: they all require external toolboxes to do so when BioPyC does not.

5.2.4 *Programming Languages*

Another important criteria for defining a platform is the programming language that is used underneath, making easier/harder to de-

velop new modules for it. Concerning the main BCI platforms, the programming languages that are used are MATLAB (MATLAB, 2010), C++ (ISO, 1998) and Python (Rossum, 1995). First, the proprietary programming language MATLAB is well-known by the research community and widely used in laboratories due to its popular rapid prototyping environment. However, the license is not free of charge nor always distributed to universities and laboratories. Second, C++ is free, very efficient, but difficult to use and, therefore, generally used by computer scientists and engineers only. Finally, Python is free, simple and extendable by non-computer scientists, making the prototyping and implementation of new modules for Python-based BCI platforms easier. Moreover, Python is widely used by the scientist community, i.e., machine learning experts, engineers and neuroscientists, and many machine learning libraries have been implemented using this language, e.g., Scikit-learn (Pedregosa et al., 2011b), TensorFlow (Dignam et al., 1983) or PyTorch (Paszke et al., 2017).

5.2.5 *Supported Systems*

If BCI2000 (Schalk et al., 2004), OpenViBE (Renard et al., 2010b), TOBI (Müller-Putz et al., 2011) and BCI++ (Perego et al., 2009) do not support all operating systems, i.e., Windows, Mac OS X and Linux, all other platforms, i.e., BCILAB (Kothe, 2013), pyRiemann (Barachant and King, 2015), Wyrm (Venthur et al., 2015), gumpy (Tayeb et al., 2018), pyEEG (Bao et al., 2011), BioSig (Schölgl et al., 2011), scikit-learn (Pedregosa et al., 2011b), MNE (Gramfort et al., 2013) and MOABB (Jayaram and Barachant, 2018) do.

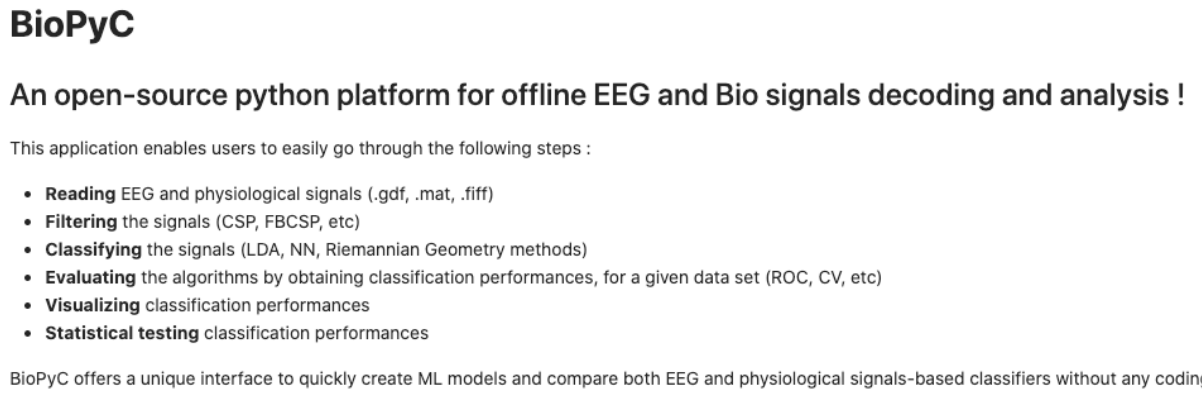

Figure 5.3: Screenshot of the Jupyter & voilà-based BioPyC's graphical user interface, allowing rich-text documentation.

5.2.6 *Licenses*

In the case of the main platforms presented in Figure 5.2, all of them are open-source and have adopted either General Public License

(GPL), Lesser General Public License (LGPL), Berkeley Software Distribution (BSD) or MIT license as their license.

5.2.7 *Distinctive Features of BioPyC*

BioPyC differs from several existing BCI platforms due to the use of Python as programming language. This difference should be highlighted, as Python is free of charge, compared to MATLAB that is not, and simple & extendable by non-computer scientists when C++ requires deep engineering skills. However, several BCI platforms such as MOABB, MNE PyEEG, pyRiemann, gumpy and Wyrm are also implemented in Python, but only MOABB follows the full offline EEG signal classification process, i.e., pre-processing, signal processing and classification. This last feature makes MOABB closely resembling BioPyC. However, those two platforms differ in three main ways: 1) BioPyC comes with a GUI based on jupyter notebook (Pérez and Granger, 2007) which acts as a tutorial and allows users to interact with the toolbox in a guided way and without requiring any programming skills, when MOABB, and other Python-based platforms do not have any GUI, and thus require programming skills for the user. Those existing platforms are thus most likely not usable by the many BCI researchers coming from diverse backgrounds such as cognitive science, neuroscience or psychology; 2) MOABB allows users to perform offline analysis only after having shared their datasets in open source, whereas BioPyC users can analyze their datasets on their own. It is an important feature to point out because most BCI researchers would want to analyze their own data before sharing them in open-source; 3) BioPyC offers modules for both statistical testing and visualization for comparing classification algorithms performances, and enables for convenient analysis since tests and plots are chosen and applied automatically by the platform, based on the distribution of the data.

More generally, BioPyC allows its users to classify physiological signals - i.e., HR, breathing, EDA - in addition to EEG signals, whereas, with the exception of Biosig, no other platform allows the classification of this type of physiological signals.

In conclusion, BioPyC distinguishes itself from other platforms through features that make it easy to use and more versatile. Indeed, it is based on Python and uses a Jupyter-based GUI. It also offers automatic statistical testing and visualization, as well as tools for classification of physiological signals such as as HR, or breathing.

5.3 Materials & Methods

BioPyC comprises four main modules, allowing users to follow the standard BCI process for offline EEG and bio signals classification: 1) reading multiple neurophysiological data formats 2) pre-processing, filtering and representing EEG and bio signals 3) classifying those signals 4) performing visualization and statistical testing on the classification performance results. Users can follow these steps through a GUI based on Jupyter (Pérez and Granger, 2007) and voilà that acts as a tutorial, explaining in a detailed way the actions to make at each step, highlighting the modularity of the platform. In this section, we detail the functionality of the platform Jupyter-based GUI, i.e., which tools we used to design it and how users are guided to interact with it. Then, we describe the modularity of BioPyC, i.e., how users can add any new module that may be necessary for their study. Finally, we present the different modules offered by BioPyC, corresponding to the major steps of the offline EEG and bio signals classification process.

5.3.1 Jupyter Notebook and Voilà as a GUI

Jupyter notebook is a scientific notebook application which allows the user to write and execute code, as well as viewing and saving the results. This tool also authorizes to write rich-text documentation using Markdown formatting, as we can see on Figure 5.3, and to display different widgets such as textbox, checkbox or “select multiple" as we can see on Figure 5.4, to make options selections easier to users. Moreover, all these features are available in a single file that is accessed via a Web browser. We then use voilà that turns Jupyter notebooks into standalone web applications in order to not have any visible code.

We designed this Jupyter interface in order to give users an intuitive path through the BCI process. Each step requires a choice from the user, and the options displayed in the following steps will be presented according to past choices. For example, if a user chooses to work on pre-processed data, only datasets where data have been previously pre-processed will be displayed.

5.3.2 BioPyC modularity

A strength of BioPyC is its modularity. Whereas the platform already comes with multiple existing modules that users can select with simple clicks, it is also possible to extend it by integrating new scripts as new modules. The kernel of the platform is made in order to allow such modifications, and make them easy to do: 1) store the new

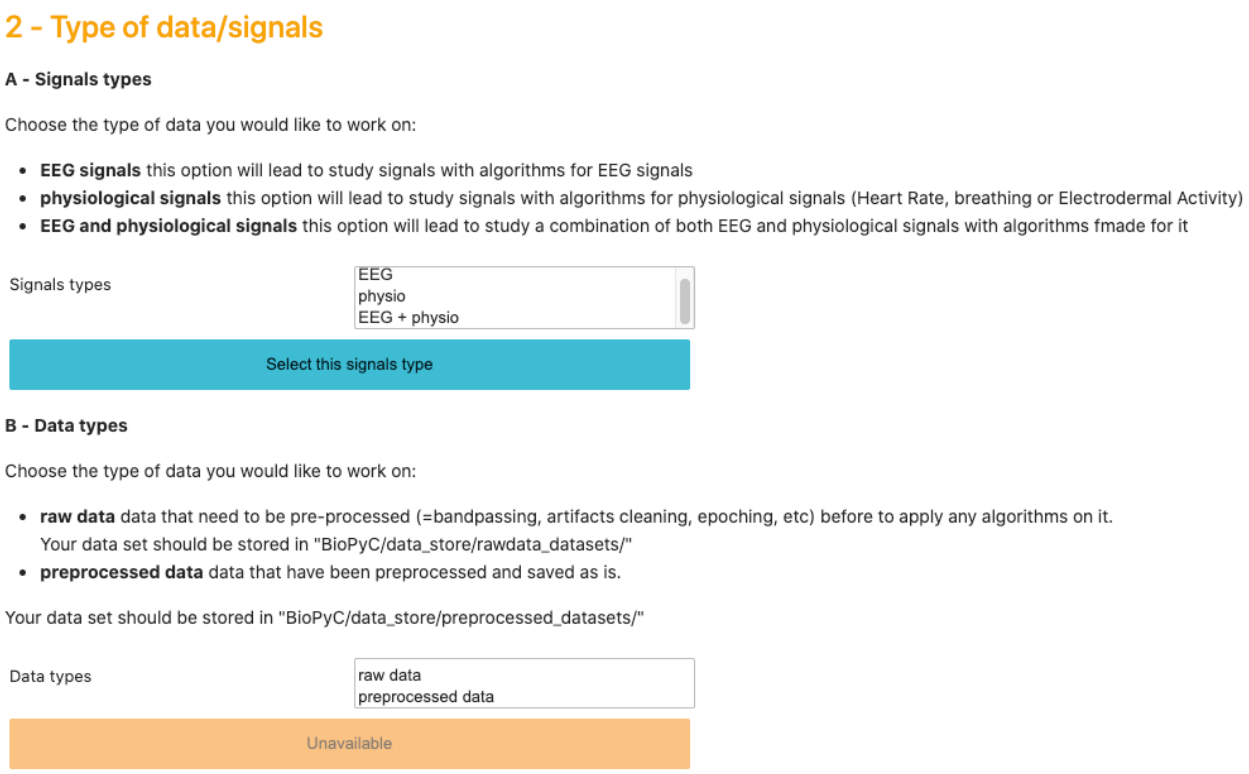

Figure 5.4: Screenshot of BioPyC's widgets, i.e., "select multiples" & buttons at the step of selecting the type of data/signals to work on. In BioPyC, a blue button stands for the action to make, when the disabled orange ones stand for future actions to make: orange buttons turn blue when the previous action is done.

script in the appropriate folder, e.g., "BioPyC/src/classifiers/" for a new classifier or "eeg_contest/src/data_readers/" for a new data format reader, corresponding to a specific format (e.g., ".gdf" or ".mat"); 2) name the python script after the classifier/data reader name with the ".py" extension; 3) follow the class and method formalism that has been used for other files of these modules.

5.3.3 *Reading data sets*

BioPyC offers users various types of data they can work with: 1) starting with raw data, directly obtained with a data acquisition software such as GDF (Schlö, 2006): this will lead to the optional pre-processing step. 2) starting with pre-processed data, where trials from various runs and sessions have already been concatenated, where the data may have been cleaned with artifact removal, band-pass filtered and epoched. Users can also choose the type of signals they want to work on, i.e., EEG signals, physiological signals or a combination of EEG and physiological signals. This step is presented on Figure 5.4.

Raw data: the raw data are read and pre-processed using the MNE python library (Gramfort et al., 2013). So far, the supported raw data format is ".gdf" (GDF - General Data Format (Schlö, 2006)), a

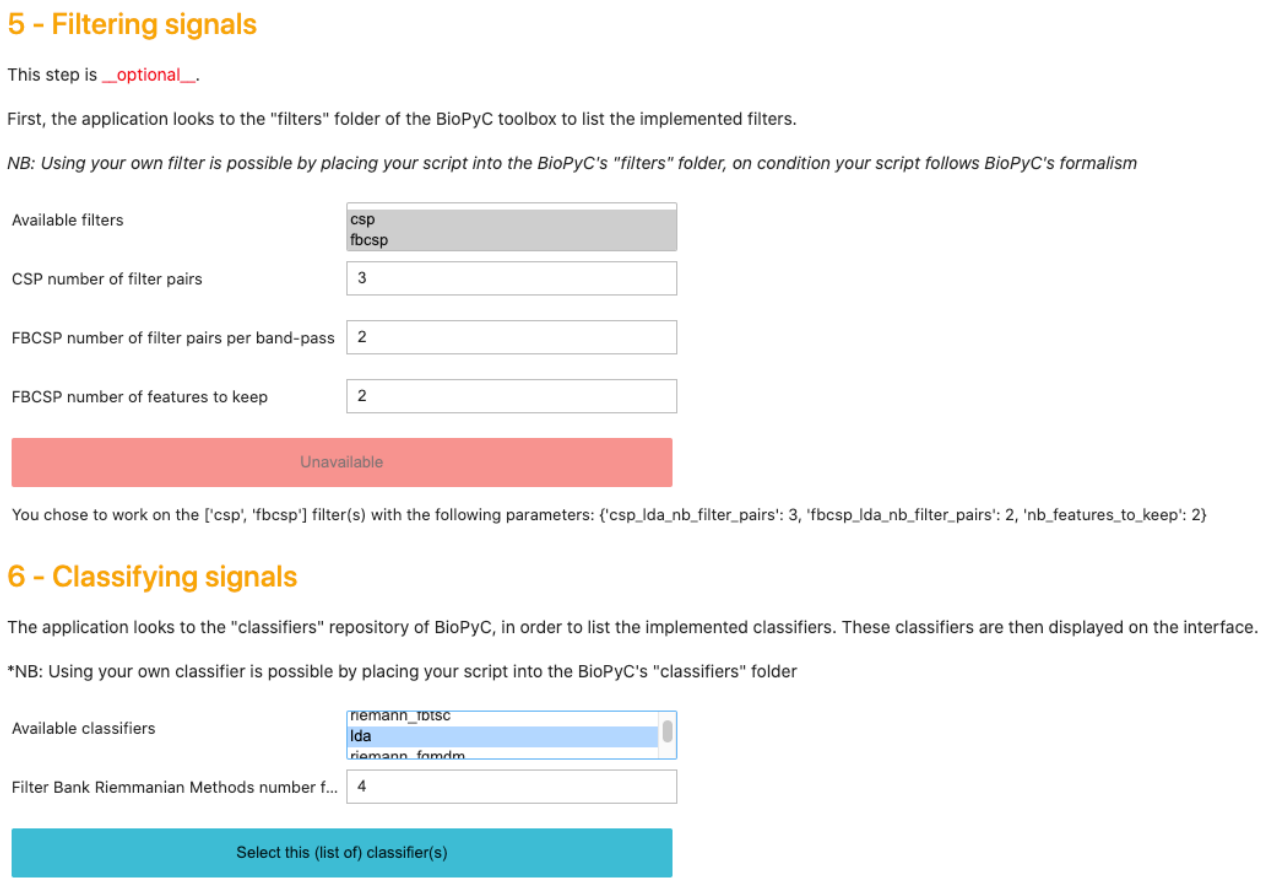

Figure 5.5: Screenshot of BioPyC filter(s) and classfier(s) selection.

standard format for EEG and biological signals. However, due to the modularity of BioPyC, python users can easily add a new data reader as explained in the Section 5.3.2 above.

Preprocessing: the pre-processing is an optional step, performed using the MNE python library as well, with multiple parameters that have to be defined through the Jupyter interface. First, users have the freedom to choose runs and sessions they want to use for each subject. Data can be cleaned from blink artifacts using ElectroOculoGraphic (EOG) channels using MNE (Gramfort et al., 2013), then band-pass filtered and finally epoched based on triggers users want to study.

Pre-processed data: in this configuration, BioPyC uses data that have been pre-processed and formatted, either coming from the pre-processing module presented above, either by reading a pre-processed data set using data readers. So far, our toolbox allows to read two types of pre-processed data formats using the python library MNE: MATLAB, i.e., ".mat" (MATLAB, 2010) and MNE, i.e., ".fiff" (Gram-

fort et al., 2013). Those two formats have been chosen since they are popular and widely use by the BCI community.

5·3·4 *Applying spatial filters and machine learning algorithms*

So far, BioPyC offers algorithms that proved effective either in BCI classification competitions, notably the Filter Bank Common Spatial Pattern (FBCSP) (Ang et al., 2012), standard Linear Discriminant Analysis (LDA) and Riemannian geometry classifiers (Yger et al., 2016) (Congedo et al., 2017)(Appriou et al., 2020), or in other rapidly increasing independent fields such of artificial intelligence, such as Deep Learning (Schirrmeister et al., 2017) (Lecun et al., 2015). We describe them below. Users can select those algorithms through the Jupyter-based GUI, as presented on Figure 5.5.

EEG spatial filters: BioPyC proposes two types of spatial filters: a) the Common Spatial Pattern (CSP) which is widely used for binary EEG classification in BCI studies, particularly for BCIs exploiting changes in brain oscillations (a.k.a. frequency band power) (Lotte et al., 2018a); b) the Filter Bank Common Spatial Pattern (FBCSP) which is an improved variant of the CSP that won numerous active BCI competitions (Ang et al., 2012). Instead of using a single frequency band, the FBCSP will explore features based on spatial filters from numerous frequency bands.

Physiological features: BioPyC offers to calculate multiple features for each type of physiological signals, so far for Heart Rate (HR) and Electrodermal activity (EDA), the HR and therefore for signals based on ElectroCardioGraphy (ECG). For each of them, statistics such as the mean or the standard deviation of RR intervals, i.e. the time elapsed between two peaks, can be calculated. Concerning the EDA, features such as the amplitude of phase peaks or inter-peaks time can be calculated using BioPyC. To see more details about the features, please refer to chapter 8 where we described out all these features.

Machine learning algorithms:

A user can select one or multiple classifier(s) in order to compare classification performances on a single dataset. Concerning EEG and physiological signals classification, we choose to integrate a Linear Discriminant Algorithm (LDA), both classic and with shrinkage, into BioPyC for classification since it is the most common used algorithms in BCI studies (Blankertz et al., 2008) (Lotte et al., 2018a). Then, con-

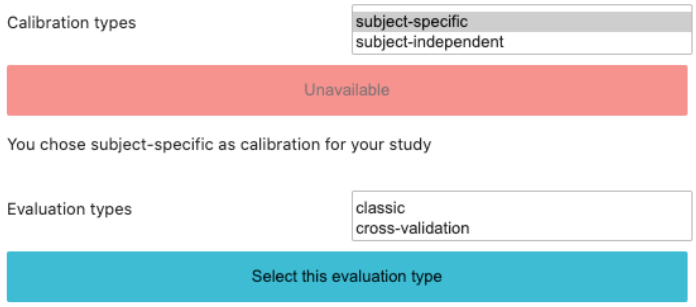

Figure 5.6: Screenshot of BioPyC's choice of both calibration and evaluation types.

cerning EEG signals only, we included four Riemannian classifiers, i.e., Minimum Distance to Mean with geodesic filtering classifier (FgMDM) and Tangent Space Classifier (TSC). Such methods represent EEG signals as covariance matrices and classify them according to their (Riemannian) distances to prototypes of covariance matrices for each class. Such methods have recently won 6 international brain signals competitions (Yger et al., 2016). In addition to those two Riemannian approaches, two new ones - Filter Bank FgMDM (FBFgMDM) and Filter Bank TSC (FBTSC) - have been introduced in chapter 4, and are also available in BioPyc. They use a bank of band-pass filters such as the ones used for FBCSP, instead of using a unique band-pass filter, and combine Riemannian classifiers from each band. Finally, BioPyC offers to use Deep Learning which recently showed promising results for many machine learning problems, with a method from (Schirrmeister et al., 2017) called ShallowConvNet, using Convolutional Neural Networks (CNN) dedicated to EEG classification. Moreover, due to the modularity of the toolbox, BioPyC users can easily add new classifiers (see Section 5.3.2), e.g., Support Vector Machine (SVM) or Logistic Regression (both available from scikit-learn (Pedregosa et al., 2011b)) to classify data previously filtered with the CSP or FBCSP.

5.3.5 *Calibration types*

BioPyC offers users to run different types of calibration approaches for studying their data, i.e., a subject-specific calibration or a subject-independent one, as we can see on Figure 5.6, depending on the motivation of their experiments.

subject-specific study: so far, due to the large between-subject variability, most of the BCI studies are subject-specific, i.e., a classifier needs to be built for each individual subject (Blankertz et al., 2008). First, data specific to each subject are split into two parts: the training and testing sets. Then, machine learning algorithms are trained on the first set and evaluated on the second one. To do so with BioPyc,

users have to set the "split ratio" (see Section 5.3.6) through a textbox displayed on the Jupyter notebook GUI.

subject-independent study: one of the major steps for using BCIs outside the laboratories would be for BCI users to be able to instantaneously use the BCI without any calibration phase. To do so, we can evaluate machine learning algorithms through offline subject-independent studies, i.e., with a classifier built on multiple subjects and used as such on a new subject, without the need for data from this new subject. In BioPyC, the evaluation method for this type of calibration is a leave-one-subject-out cross validation, i.e., the training phase uses all subjects expect the target subject data to train the classifier, and the testing phase applies this classifier on the target subject data only. This process is repeated with each subject used once as the target (test) subject.

5.3.6 Evaluation

Split ratio: the split ratio method for evaluation is the classic machine learning method for evaluating an algorithm. It consists in separating the data set in 2 parts, the first one for the training of the algorithm, and the second one for the evaluation of this algorithm. The split ratio defines the ratio of data that has to be kept for the training set. The rest of the data will be used for the evaluation.

Cross-validation: finally, due to the usually relatively small number of trials recorded during BCIs experiments, BioPyC proposes a "leave-one-out" cross-validation method based on scikit learn (Pedregosa et al., 2011b) for the evaluation if the number of trials is rather low. Each trial is used as testing set where algorithms are trained on all other trials. The number of k-folds is equal to the number of trials. The "k-fold" cross-validation method allows users to choose the number of equal segments into which data should be split. For each data segment, trials composing this segment are used as testing set when the rest of data is used as training set.

5.3.7 Statistics and Visualization

As explained in the introduction, BioPyC allows users to make basic statistics and visualization about classification performances obtained through the classification.

Performances: once algorithms have been applied on pre-processed data, classification performances scores - either the accuracy or the F1-score, depending on whether the classes are balanced or not - are

automatically calculated for each subject and for each algorithm that have been selected for the study. The accuracy score is calculated with scikit-learn "accuracy_score" method if the data set classes are balanced, or with scikit-learn "f1_score" method if they are unbalanced. Those classification performances are framed into a table using the python library pandas (McKinney, 2010), and can be directly stored into a directory that has been previously indicated by the user through a Jupyter textbox, and/or used by the module "statistical analysis", in order to make statistical testing/plotting.

Statistics: for statistical testing, BioPyC enables users to choose to make automatic appropriate tests for comparing classification performances between machine learning algorithms with the python libraries pandas (McKinney, 2010) and pingouin (Vallat, 2018):

1. Test the data normality with the Shapiro-Wilk test from the python library Scipy (Jones et al., oday)
2. Test the data sphericity with the Mauchly's test from Scipy
3. Analyze and compare means of classification performances between machine learning algorithms using, in the case data are normalized, either
    - t-test: comparing performances of two algorithms along all the subjects; comparing performances of an algorithm depending on the study type (subject-specific vs subject-independent) using pingouin
    - one-way ANOVA with repeated measures: comparing performances of multiple algorithms (more than 2) or multiple study types using pingouin
    - two-ways ANOVA with repeated measures: comparing performances with both factors (type of algorithms, type of study) using pingouin

Chance level: When measuring classification accuracy for a BCI task, given the usually small number of samples, the actual chance level should be carefully considered (Müller-Putz et al., 2008). For example, the chance level for a 2-class paradigm will not be necessarily 50%, it will depend on the number of testing trials and the confidence interval we want to work with. To solve this problem, BioPyC proposes an option for the calculation of the chance level based on (Müller-Putz et al., 2008). Moreover, users can test the difference between subjects performances and the chance level with a one-sample t-test from (Jones

et al., oday).

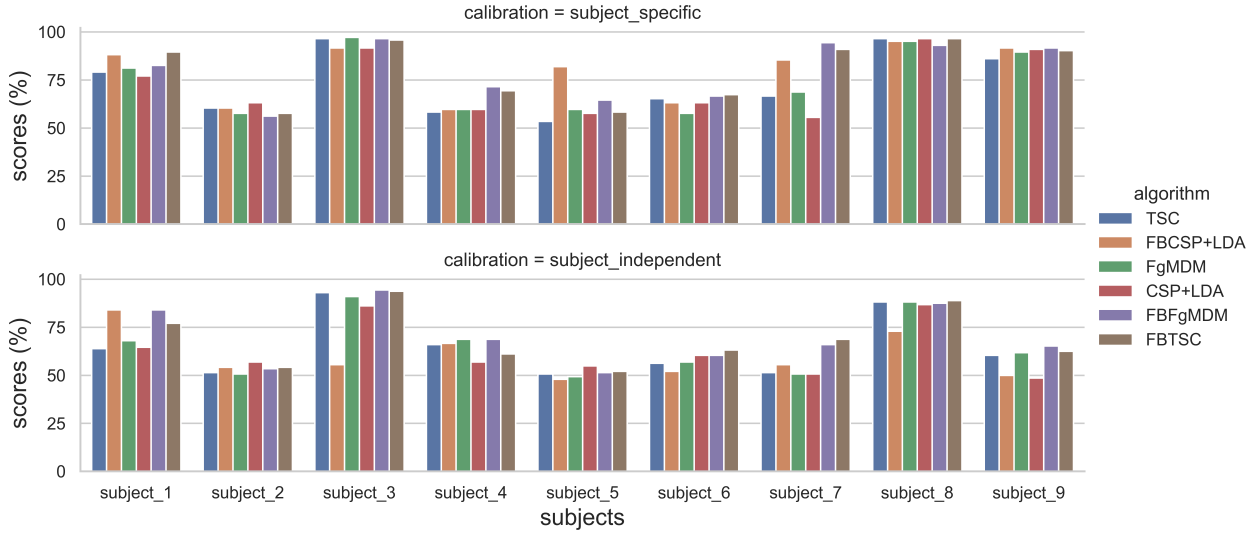

Figure 5.7: Classification accuracy of each algorithm, for each participant, on the "BCI competition IV dataset 2a", in both subject-specific and subject-independent calibrations.

Visualization: the data visualization is important in BCI studies since it allows to have an informative and explicit feedback about the obtained classification performances. BioPyC proposes both boxplots and barplots from seaborn (Hunter, 2007) and pairwise t-test visualization using scikit-learn (Pedregosa et al., 2011b). For the boxplots, the number of boxes on the plot will vary depending on the number of algorithms and the number of calibration types that have been tested, as we can see on Figure 5.8. The barplots can be used for visualizing detailed performances results of each algorithm/calibration on each participant (see Figure 5.7). BioPyC also proposes to display confusion matrices using scikit-learn (Pedregosa et al., 2011b), as presented on Figure 5.9.

5.3.8 *Evaluating BioPyC*

BioPyC has already been used to analyze 4 types of BCI data, for Mental Task BCIs and mental state decoding through passive BCIs such as Workload, Emotions and Attention. All data-sets were of different sizes (number of subjects and trials), collected in different laboratories using different EEG devices, with data stored in different formats: this led to a direct test of the robustness of BioPyC. In this section, we present the 3 datasets we studied using BioPyC.

Mental Task: first, we used the modern machine learning algorithms from BioPyC to classify mental tasks EEG signals using the dataset coming from (Brunner et al., 2008) called "BCI competition IV dataset 2a". In this dataset, EEG signals have been recorded from 22

electrodes, from 9 participants, when executing four different mental tasks. We chose to keep only two classes, namely the imagination of movement of the left hand (class 1) and right hand (class 2). Participants participated to two sessions of 6 runs, where a run consists of 24 trials (12 for each class), yielding a total of 144 trials per session.

After band-pass filtering the signals in both a single band (in 8-12 Hz) single band-based algorithms, and in 9 4Hz-wide bands for filter banks-based algorithms (in 4-8 Hz, 8-12 Hz, . . . , 36-40 Hz), we used 6 methods for classifying those 2 mental tasks, i.e., the CSP coupled with a LDA (Blankertz et al., 2008), FBCSP coupled with a LDA (Ang et al., 2012) and 4 Riemannian approaches (FgMDM, TSC, FBFgMDM and FBTSC) (Yger et al., 2016) (Appriou et al., 2020) - and compared them across 2 types of calibration, i.e., subject-specific and subject-independent. For the subject-specific calibration, classifiers were trained on the data from the first session of a participant, and testing set was the data from the second session, as it has been done in the original study (Brunner et al., 2008). Regarding the subject-independent calibration, the training set comprised all trials of all participants except the current participant used for testing, and the testing set the second session of the current participant.

Mental states: we then used the modern machine learning algorithms from BioPyC to classify mental states through EEG signals. First, we performed 7 algorithms, i.e., CSP coupled with a LDA, FBCSP coupled with a LDA, 4 Riemannian approaches (FgMDM, TSC, FBFgMDM and FBTSC) and a CNN, on 1) a cognitive workload data set and 2) an emotions data set: both subject-specific and subject-independent calibrations have been run on both data sets, following a 2-classes classification problem. We presented the methods used for these 2 studies in chapter 4.

Second, we used a Riemannian Geometry-based method (TSC) for classifying four attentional states, i.e., alertness and sustained attentions, referring to the intensity of attention (i.e., its strength), as well as selective and divided attentions, referring to its selectivity (i.e., the amount of monitored information) (Zomeren and Brouwer, 1994). This study aimed at developing a first comprehensive understanding of the different attentional states described in the model of van Zomeren and Brouwer using EEG data ((Pillette, 2019; Pillette et al., 2018)). The term "Attention" encompasses several different attentional states. Given the model of van Zomeren and Brouwer it encompasses four attentional states, i.e., alertness and sustained attentions, referring to the inten-

sity of attention (i.e., its strength), as well as selective and divided attentions, referring to its selectivity (i.e., the amount of monitored information) ((Zomeren and Brouwer, 1994)). No study provided yet a comprehensive comparison of these different attentional states.

Hence, we included 16 participants into an experiment during which they were asked to perform different tasks. Each task assessed a type of attentional state, while we recorded the participants' EEG. During each task, the participants had to react as fast as possible to the appearance of target stimuli by pressing a keyboard space bar as fast as possible. In accordance with the literature, the tasks and types of attention were differentiated by the type of sensory modality of the stimuli, number of distractors, presence of warnings tone before the stimuli and length of the task ((Francis, 2010; Schmidt, 1968; Sturm and Willmes, 2001; Van Leeuwen and Lachmann, 2004)). For each task, 80 targets stimuli were presented. We used one second prior to target presentation as the analysis window. Only data from targets that were at least one second apart from a motor response were analysed to prevent motor-related artefacts.

First, we used BioPyC to know if we could differentiate the different attentional states from one another. We used a Common Spatial Pattern filtering in the alpha range (8-12Hz) and a Linear Discriminant Analysis classifier, with 5-fold cross-validation.

Second, we used BioPyC to know if we could classify the five types of attentional states at once using only EEG data. The participant-specific discriminability (one classifier per participant) of the EEG patterns between each of the five attention tasks was assessed using the tangent-space classifier described in ((Yger et al., 2016)), with 5-fold cross-validation. We used the method from ((Barachant et al., 2012b)) to classify EEG signals into 5 classes: a linear discriminant analysis (LDA) has been performed between each pair of class, i.e., each pair of attention task, then all the resulting classifiers are combined to obtain the classification results. The 5-classes classification was performed twice with EEG data either filtered in the Theta or Alpha band. The confusion matrix, representing for each class the ratio of trials that where accurately or wrongfully associated with it over the total number of trial were then computed.

5.4 Results

In this section, we present the results obtained by the different signal processing and machine learning algorithms currently offered by BioPyC.

5.4.1 Mental tasks

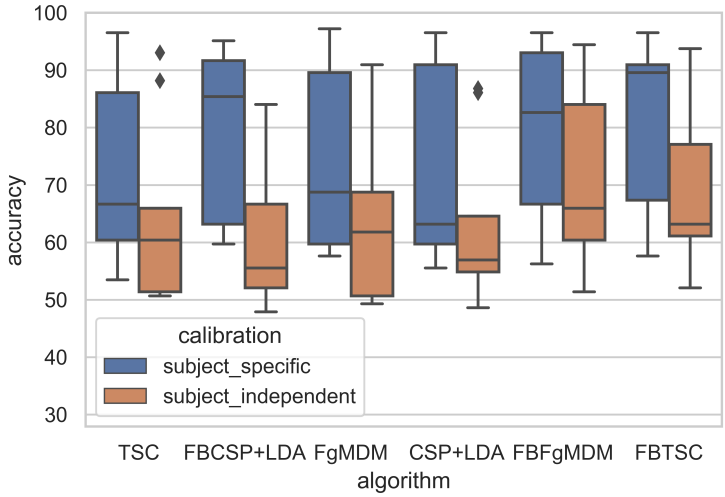

Figure 5.8: Classification accuracy of each algorithm on the "BCI competition IV dataset 2a", in both subject-specific and subject-independent calibrations.

The detailed results, i.e., classification accuracy scores obtained by each algorithm, for each participant, with both subject-specific and subject-independent calibrations, are represented on the Figure 5.7. More general results, i.e., the classification accuracy of each algorithm, with both calibrations as well, are plotted on Figure 5.8. They revealed that FBTSC and FBFgMDM obtained the highest mean accuracy, although not significantly so, with both subject-specific (mean accuracy FBTSC = 79.6%; mean accuracy FBFgMDM = 79.7%) and subject-independent calibrations (mean accuracy FBTSC = 70.1%; mean accuracy FBFgMDM = 69.1%).

5.4.2 Mental states

Results of the study concerning the estimations of mental states, i.e., cognitive workload, emotions are presented in chapter 4.

Results regarding the discrimination of attentional states from one another are promising and range from 83% accuracy (SD=0.09) to discriminate alertness (Tonic) from sustained attention to 74% accuracy (SD=0.13) to discriminate selective and divided attention.

We then classified the five types of attentional states at once. The average confusion matrices over all participants for the classification in Theta and Alpha bands are displayed in Figure 5.9.

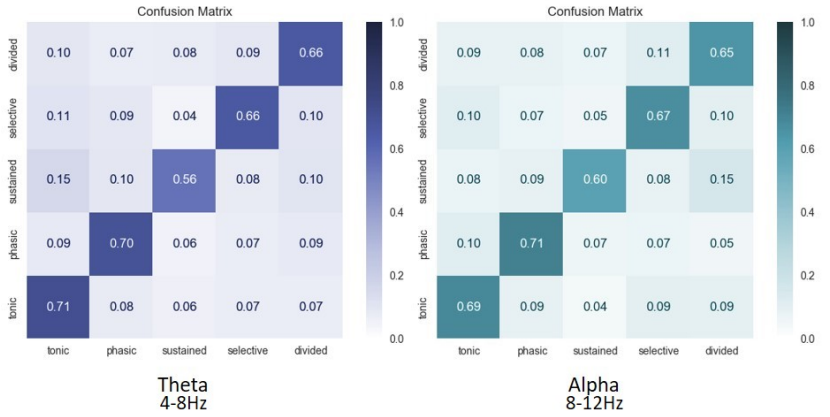

Figure 5.9: Average confusion matrices over all participants for classification of attention in Theta (4-8Hz) and Alpha (8-12Hz) frequency bands of 5 attentional states, i.e., alertness (tonic), alertness (phasic), sustained, selective, and divided.

Overall, these promising results tend to validate the model of van Zomeren and Brouwer as the different attentional state that they describe seem to have distinct electroencephalographic patterns of activation. We believe that future research assessing the learners' states during BCI user training might represent real opportunities to improve such training.

5.5 *Discussion*

Our extensive experiments from the different analyses performed for BioPyc reveal several positive aspects of the platform, i.e. the modularity, the comparison of classification algorithms, the statistical analyses, as well as the data visualization. First, the modularity of the platform is highlighted with the different data sets formats that have been used for the offline analysis presented above. Data sets for workload, emotions and attention had a Matlab (".mat") format and contained pre-processed data. The data set for the mental tasks analysis had a GDF (".gdf") format and required a pre-processing step before performing the signal processing & classification steps. Second, the data sets with a 2-class paradigm, namely mental tasks, workload, and emotions (both valence & arousal) have been used for comparing classification algorithms, which is one of the advantage of BioPyC. Those results indicated interesting information about algorithms, such as the ineffectiveness of the CNN on data-sets with a small number of trials, e.g., emotions, when the same algorithms proved to be efficient on data sets with a large number of trials, e.g., workload. Moreover, BioPyC

also proved to have efficient machine learning algorithms for multi-class classification. This is the case of the study on attention, where the data was divided into 5 classes of attention (tonic, phasic, sustained, selective and divided). Third, the statistical analyses have also multiple aspects: 1) the 2-way ANOVA with repeated measures have been used for analyzing performance results of the machine learning algorithms (factor 1) with both subject-specific and subject-independent calibrations (factor 2) in the mental tasks, workload and emotions studies; 2) post-hoc t-tests have been performed to check significant differences in performance results between algorithms. In fourth, concerning the BioPyC data visualization module, all plots that have been showed in Section 5.4 have been generated by BioPyC: 1) boxplots represented on Figure 5.8, 2) barplots represented on Figure 5.7;

5.6 Current Status and Future Work

BioPyC is currently publicly available on github at `https://gitlab.inria.fr/biopyc/BioPyC/`. Users have to clone the repository and install all dependencies (jupyter, voilà, numpy, pandas, pingouin, pyRiemann, scikit learn, scikit_posthocs, MNE==0.17 and braindecode) using pip. Then, users have to find the file BioPyC.ipynb and run "voila BioPyC.ipynb" in order to display the interface in a web browser and initialize the application. All instructions will then be given by the application that is made as an intuitive tutorial.

In its current stage, BioPyC offers the different modules that allow users to follow the standard steps of the BCI process, i.e., reading different EEG data format, filtering and cleaning EEG signals, classifying EEG signals and finally visualizing and performing statistical tests on the classification performance results.

Regarding the reading of different EEG formats, two modules are available in the current stage of the platform, namely GDF (".gdf") and Matlab(".mat"), and one is still in progress, i.e., MNE (".fiff"). Future versions on BioPyC will offer more modules for reading data, starting with Python's ones, i.e., ".pkl" and ".dat".

BioPyC currently offers tools for pre-processing the signals as well, all based on MNE (Gramfort et al., 2013): EOG-based artifacts removal, band-pass filtering and epoching. More pre-processing features are available in MNE, it would therefore be easy to add new modules for pre-processing data in the future versions of BioPyC.

Regarding the third step of the BCI process - i.e., signal processing & machine learning for EEG signals classification - so far BioPyc proposes several efficient algorithms for decoding oscillatory activity: the CSP (Blankertz et al., 2008) & FBCSP (Ang et al., 2012) are the two algorithms that allow spatial filtering, when the LDA, Riemannian Geometry methods (Yger et al., 2016) (Appriou et al., 2020), as well as the CNN (Schirrmeister et al., 2017) are the machine learning algorithms implemented in the platform. However, the ongoing works aim to integrate new signal processing and machine learning algorithms for the classification of Event Related Potentials (ERPs) into BioPyC. Among them, xDAWN (Rivet et al., 2009), which is widely used for spatial filtering and proved to be efficient for EEG-based classification of workload levels (Roy et al., 2015a). It would also be interesting to integrate other ERP spatial filtering methods into the future versions of BioPyC, e.g., principal component analysis (PCA) or canonical correlation approaches (CCA) (Noh and De Sa, 2013) that proved efficient for EEG classification of mental workload levels as well (Roy et al., 2015a). Moreover, the current works also aim at integrating machine learning methods, e.g., EEGNet (Vernon et al., 2018), for the classification of both ERPs and oscillatory activity with BioPyC. Finally, the ongoing works are also focusing on the integration of methods for the classification of bio signals on the one hand, and the integration of hybrid BCI methods for the classification of the combination of EEG signals-based and bio signals-based features on the other hand. Current work on bio signals and hybrid classification concern HR, breathing and EDA: users will soon be able to extract features such as the mean or the standard deviation of RR intervals from the ECG signals, or the amplitude of phase peaks or inter-peaks time from the EDA[2]. Current works are focusing on extracting features from the breathing signals as well. The fourth and last step of the BCI process, i.e., performing visualization and statistical tests on classification performances results, also benefits of current improvements: a new module for visualizing the percentage of use of the different frequency bands selected by the filter bank-based algorithms, i.e., FBCSP, FBFgMDM and FBTSC, is ready, and will be soon pushed on the gitlab repository.

[2] Features from physiological signals, i.e., heart rate, breathing and EDA, are already implemented, but not available on gitlab yet.

5·7 Conclusion

We presented BioPyC, an open-source and easy-to-use BCI Python platform for offline EEG and biosignal analysis. This platform allows BCI and physiological computing researchers to quickly analyze offline their data by following the classical steps of pre-processing (op-

tional), signal processing & classification, statistical analysis and data visualization. It is important to note that users do not need any programming skills to be able to process their data, since BioPyC is built in the form of a jupyter notebook with a voilà GUI that acts as tutorial: each step is described with instructions that guide users in their analyses and choices for parameters and algorithms to use. BioPyC already proved to be a comprehensive tool since it has been used for 3 extensive studies so far, with quite different aims and requirements. Moreover, since Python is free of charge, any researcher can use it for his/her experiments. Moreover, BioPyC is open-source and allows users to build new modules. For example, new signal processing or classification algorithms can be easily added to the platform, as well as data readers for new data sets formats. So far, BioPyC has still a modest number of tools, but can easily be extended in the future, and is still growing. For example, currently, the toolbox can only support one main BCI paradigm, i.e., oscillatory-based BCIs, but will soon be extended to support the evoked potentials-based BCI paradigm.

Part III

Towards measuring states of epistemic curiosity through EEG and physiological signals

PhD Thesis Roadmap

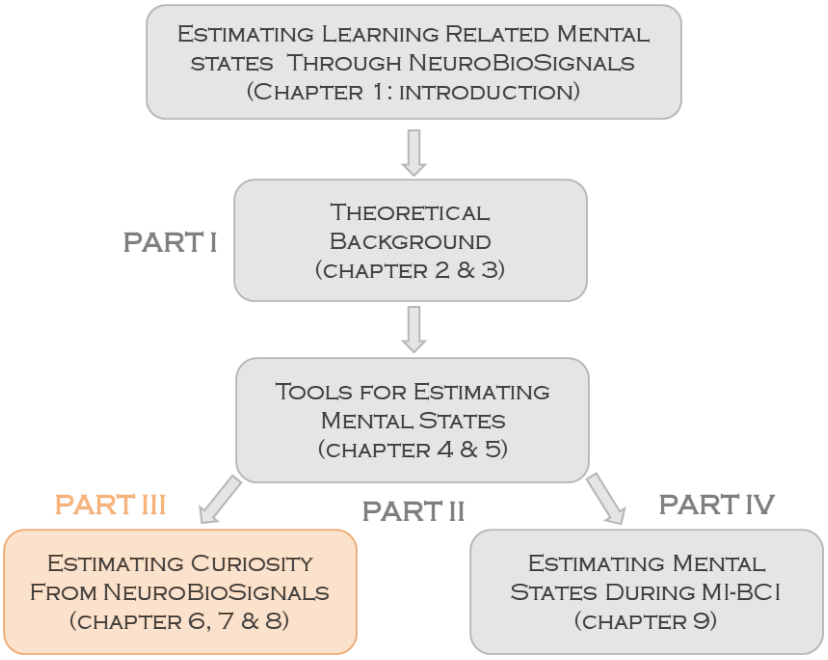

Figure 5.10: PhD thesis roadmap.

Related Papers

Peer-reviewed Journals

- Appriou, A., Ceha, J., Trocellier, D., Pramij, S., Dutartre, D., Law, E., Oudeyer, P.-Y., Lotte, F. Towards measuring states of epistemic curiosity through electroencephalographic and physiological signals. (***Being written with the future aim to submit to the journal Journal of Neural Engineering)***

Peer-reviewed conferences

- Appriou, A., Ceha, J., Pramij, S., Dutartre, D., Law, E., Oudeyer, P.-Y., Lotte, F., (2020). Towards measuring states of epistemic curiosity through electroencephalographic signals. *IEEE Systems, Man and Cybernetics Conference*. (***Published)***

6

Research Question & Protocol Design

6.1 Research question

As seen in the introduction, being able to estimate cognitive, affective or conative states from neurophysiological signals would be beneficial for the creation of passive BCI applications. Indeed, this would allow applications to monitor the users' states in real-time, and therefore adapt the interactions to individuals capabilities, for instance users' motivation (e.g., motivational states, curiosity). Moreover, conative states have been shown to be involved in human learning, thus being able to estimate such states in real-time would play a major role for upgrading BCI training protocols. This type of system could also be used as an evaluation method for HCI. In contrast to cognitive states, e.g., cognitive workload or attention, and affective states, e.g. emotions, conative states have been little studied through neurophysiological signals. Indeed, as we saw in the introduction, the literature review concerning the estimation of curiosity levels through EEG or physiological signals is very scarce. It would therefore be interesting to study curiosity through EEG - but also physiological - signals in more depth in order to obtain trustable neurophysiological-based conative state estimators.

The goal of this part III is to use passive BCIs that can monitor,

through neurophysiological signals, the level of curiosity of users. As explained, it would be both a useful tool for understanding curiosity, and beneficial to designers of interactive systems, who wish to adapt the interaction paradigm or application content to users' levels of curiosity. To do so, our objective is to explore recognition systems based on EEG signals, but also on physiological signals such as Heart Rate (HR), breathing or Electrodermal activity (EDA) - that already proved to be efficient for mental state estimations (Fairclough, 2009) - in order to classify these curiosity levels.

However, several studies have been run to better understand the neural mechanisms underlying states of curiosity. In (Kang et al., 2009), the authors scanned participants with functional magnetic resonance imaging (fMRI) in order to study the brain areas activated during the triggering of curiosity states with trivia questions. They observed that the curiosity induced by the trivia questions was correlated to brain activity in the caudate region, an area shown to be associated with anticipated rewards (Adcock et al., 2006). Moreover, they also found a correlation between surprising new information and activation of brain areas linked to memory. In (Gruber et al., 2014), the question was raised as to whether curiosity enhances long-term memory, similar to the way anticipated rewards do. They likewise conducted a fMRI study with trivia question tasks, and found a correlation between curiosity levels and variations of activation in the right hippocampus and bilateral nucleus accubens, both involved in long-term memory improvements (Gruber and Valji, 2019). Using frontal EEG asymmetry—a common tool for measuring engagement and motivation—(Lima, 2019) investigated the relation between curiosity and learning. Participants performed trivia question tasks, similar to those used in (Gruber et al., 2014; Kang et al., 2009), while the EEG signals from the frontal cortex were recorded. Here, researchers found a correlation between frontal brain asymmetry (FBA) and memory recall; however they did not observe any correlation between FBA and self-reported curiosity.

In summary, these neuroscientific studies, mainly based on fMRI, support the existence of a correlation between epistemic curiosity and memory/learning, as well as correlations between states of curiosity and activation in specific brain regions. However, these prior works did not perform continuous state monitoring. AS explained, in this work, we propose to measure states of epistemic curiosity using EEG, but also body sensors. Compared to fMRI, which can be expensive and difficult to use outside of the lab, EEG provides a usable, portable and affordable tool for measuring the temporal activation of brain states

associated with curiosity, making it suitable for applications such as BCIs.

In this chapter, we present an experiment we have conducted in which subjects were given a trivia question-based task, designed to elicit different levels of curiosity. We then show in chapter 7 that we used signal processing tools to analyze the EEG signals collected through this protocol, and assessed how well we can estimate curiosity through EEG signals using Machine Learning (ML) classifiers. Finally, in chapter 8, we present our work aiming at estimating levels of curiosity as well, but through physiological signals that have been collected through the same protocol. To do so, we extracted features from HR, breathing and EDA signals, and used ML classifiers in order to discriminate curiosity levels.

In the detailed presentation of this part III, we first introduce the protocol design, including the setup for recording both EEG and physiological signals, before doing the same with the methods we used to induce different curiosity levels. In chapter 7 we present the signal processing methods we employed, and the machine learning algorithms we used to classify two curiosity levels (low versus high) based on EEG signals, as well as the results that came out of it and discussion about them. Finally, chapter 8 presents the methods that have been used to extract features from physiological signals, as well as ML classifiers, before showing results that have been found and discussions around them. First, we thus describe in detail the experimental setup, including the participants involved and the protocol.

6.2 *Participants*

Twenty-seven participants (N=27) were recruited through ads posted on social media and the local university mailing lists (13F/14M; aged 28.7 ± 4.0). Levels of education varied between high school diploma and Ph.D. To be included in the experiment, people had to be at least 18 years old, speak French, and consent to the study[1]. Non-inclusion criteria include bad vision, heart condition, neurological or psychological diseases, and emotion-related problems. Finally, as per typical EEG studies protocol, we asked participants not to drink coffee or tea within the 2 hours prior to the experiment.

[1] This study was approved by the ethics committee of Inria Bordeaux Sud-Ouest (COERLE approval number 2019-13).

6.3 Protocol

Each participant participated in a single session that took place at Inria Bordeaux Sud-Ouest, lasting approximately 2 hours. Participants were asked to fill-in a pre-session questionnaire, assessing personal characteristics (such as gender, age and education). Next, EEG and ElectroOculoGram (EOG) electrodes were placed on the participants. For recordings, we used a BrainProduct ActiCHamp amplifier (EEG, 61 active electrodes in a 10/20 system, and EOG, 3 active electrodes to measure ocular artifacts). To our knowledge, except for (Lima, 2019) where they attempted - but did not succeed - to measure curiosity with electrodes placed onto the scalp recovering the frontal cortex, there has been no study seeking to classify curiosity levels from EEG signals. Due to the exploratory nature of our study, we covered the entire scalp with a relatively high EEG spatial resolution (61 electrodes) and took information from all brain areas. Note that we took Cz as the EEG-reference electrode. Sensors for ElectroDermal Activity (EDA) (i.e., the Galvanic Skin Response sensor), breathing (i.e., the breathing belt) and heart rate (HR) (i.e., the ginger-photoplethysmogram sensor), were also installed. All signals were recorded and visually inspected using OpenViBE (Renard et al., 2010a).

Following the setup, a 3 minute baseline, consisting of measuring EEG signals from participants at rest with opened eyes, was recorded. Finally, participants were asked to perform 4 runs of curiosity tasks (described below), around 10 minutes each, with 5 minute breaks between them. Each run consisted of a series of trivia questions and answers, inducing different levels of curiosity. Before each question, a fixation cross was presented on the screen for 3 seconds, in order to get the participant ready. Figure 6.1 illustrates the experiment flow.

6.4 Materials

Prior work using trivia questions to elicit curiosity presented participants with question/answer pairs that did not have any link to previously viewed or future questions (Gruber et al., 2014; Kang et al., 2009; Lima, 2019). In our study, we introduce a novel protocol consisting of chains of trivia question/answer pairs—i.e., if participants were curious about the answer to a certain question, the following question would follow on the same topic. The assumption is that if participants were curious about a certain question, it was likely they would be cu-

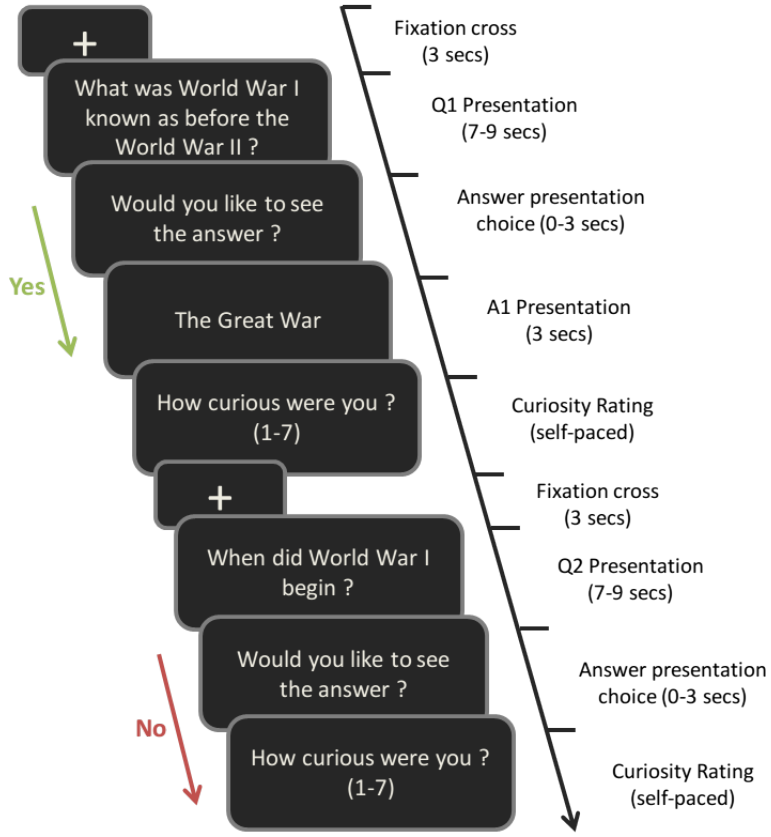

Figure 6.1: Experiment flow: 1) fixation cross 2) question presentation 3) choice to reveal the answer 4) answer presentation 5) curiosity rating.

rious about this topic in general. This new method allowed us to: 1) record a large enough amount of EEG signals from both curious and non-curious states, to then be able to train the curiosity classifiers with a balanced set of EEG examples, and 2) check the assumption that curiosity could be a mental state that increases over time, in the same way intrinsic motivation and self-directed learning increase when time spent in flow state increases (Hektner and Csikszentmihalyi, 1996).

The trivia questions used to elicit curiosity in this study came from an online trivia question dataset (`https://www.randomtriviagenerator.com`). The questions were grouped following a two-level categorization system. Questions from the website were already classified into classical trivia categories (referred to as first-order categories), such as Science, History, Geography, Arts, General Knowledge and Sports. We further classified the questions in each first-order category into groups of 4 to 20 questions based on 800 extracted keywords from the first-order category questions. For example, in the Geography category, we identified a sub-category based on the keyword "Nile". All questions from the Geography category with the word "Nile" were then grouped into a single chain of questions. Figures 6.1 and 6.2 show an-

other example of a question chain based on the keyword "World War One".

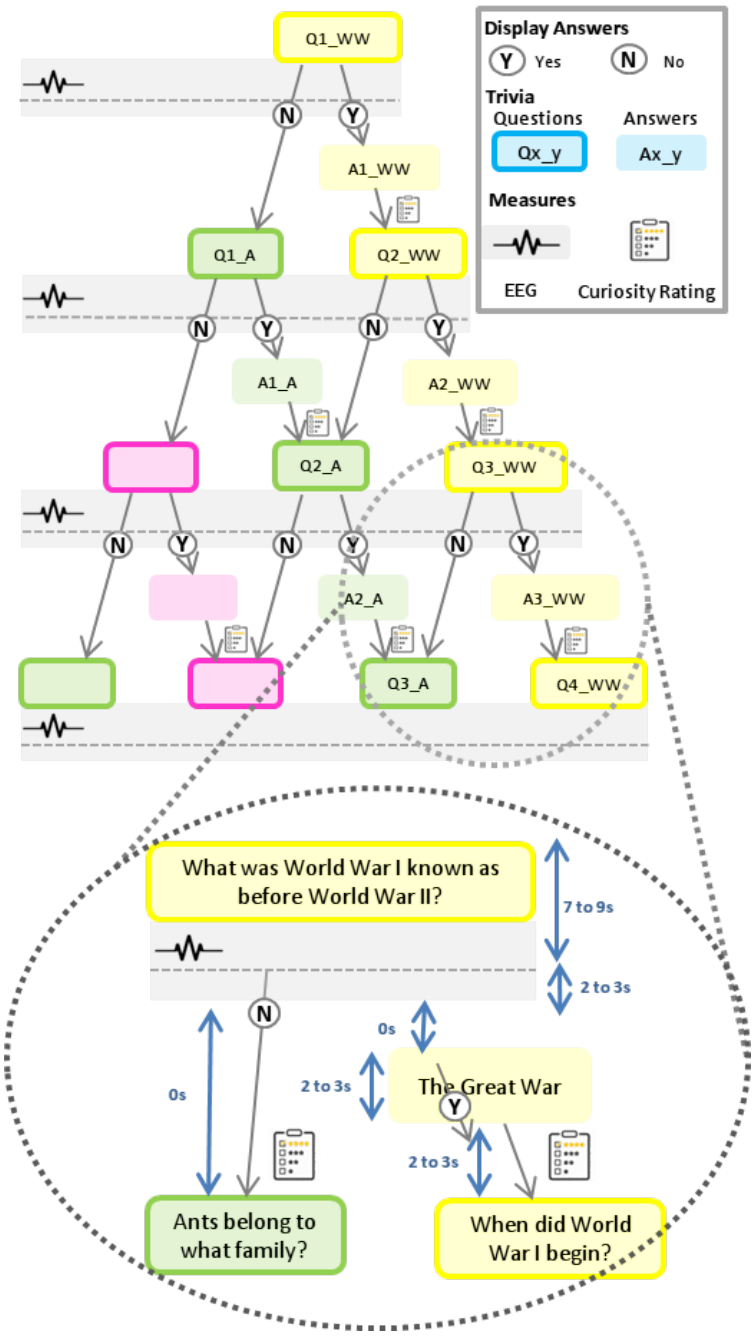

Figure 6.2: The trivia questions/answers system. In the example, a question about World War One is presented: if they choose to display the answer, they will stay on the "World War One" topic, but will continue on another topic - here scientific questions about ants - if they skip the answer.

All chains containing less than 4 or more than 20 questions were not used in this study. A minimum of 4 was chosen in order to have enough questions to define a subcategory. Conversely, we chose 20 as the maximum to ensure the subcategory was not too vague. Overall, the dataset consisted of 2000 questions/answers divided between 250

subcategories.

6.5 *Curiosity Task*

The curiosity task was presented on a computer screen using images - black background with white text. Each run was set up as follows: a first question from a random chain from a random category was displayed to the participant - for 7 to 9 seconds - right after a 3-second fixation cross. The participant then had 2-3 seconds to choose to display the answer or not, as presented in Figure 6.1. The display time was determined based on the length of the question or answer in terms of number of characters. The participant chooses to display the answer by tapping on the keyboard space bar: this question was then flagged as “curious", the concerned subcategory was considered interesting for the participant, and the answer was directly displayed on the screen for a few seconds. Participants were asked to rank their level of curiosity for the question on a 1-7 scale using a number pad right after the answer had been displayed. Following the rating, a new question randomly selected from the same subcategory was displayed.

If the participant did not choose to display the answer (by tapping on the keyboard) before the decision time elapsed, the question was labeled as “non-curious", the subcategory was not considered to be interesting to the participant. The curiosity rating scale was immediately administered without revealing the answer, followed by a fixation cross and a new random question from the next subcategory. A run ended only if at least 15 trials with questions marked as curious and 15 trials with questions marked as non-curious had been displayed. Thus, for each participant, we obtained at least 30 trials per run, i.e., a total of 120 trials in 4 runs: 60 trials per state of curiosity (curiosity & no curiosity).

7 Towards estimating states of Epistemic Curiosity through EEG signals

This chapter aims at estimating states of Epistemic Curiosity through EEG signals. In a first section, we present the signal processing and machine learning methods we perform in order to classify such states of epistemic curiosity. Then, results are presented in the next section, followed by a last section reserved to the discussion, limits and future works.

7.1 Signal Processing & Classification

Our system aims to discriminate curious from non-curious states using EEG signals. To do so, we employed machine learning approaches based on state-of-the-art algorithms developed for BCIs (Ang et al., 2012; Appriou et al., 2020; Lotte et al., 2018b) to classify EEG signals. This section describes the EEG signal preprocessing steps we used, i.e., trial epoching and labelling, the machine learning algorithms used, and finally the method used for evaluating classification performance.

7.1.1 *Pre-processing*

We first pre-processed EEG signals into N-second windows, in order to create 5 different data sets, with 1-, 2-, 3-, 4- and 5-second windows. The EEG signals for each 10 minute run were divided into approximately 30 trials, i.e., one trial per trivia question displayed. More precisely, an EEG trial was defined as ending at the time the question disappeared from the screen, and starting N-seconds earlier, as represented by the blue arrows on Figure 7.1. Note that no artifact removal algorithm has been used in this study.

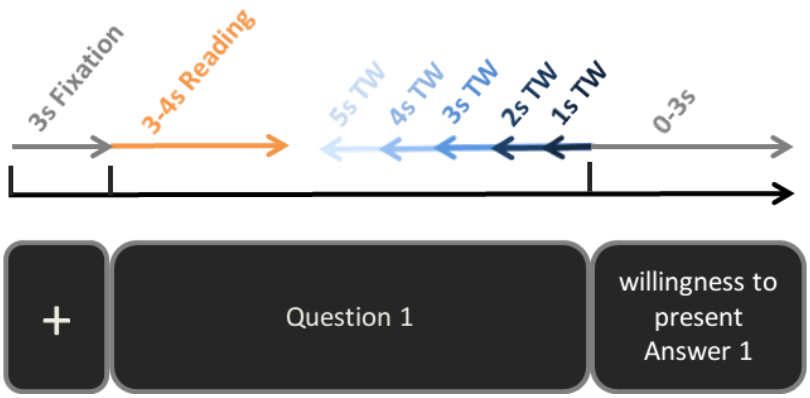

Figure 7.1: Diagram representing the way we epoched the signals into 1, 2, 3, 4 and 5-seconds time windows (TW).

7.1.2 *Classification labels*

Each trial was flagged as “answer" or “no-answer", based on the participant’s choice to reveal the answer or not, respectively. Based on the flags and participants’ self-reported curiosity ratings, we labeled each trial as “curious" if the trial was flagged as “answer" and the rating was higher than the mean of the participant’s ratings; otherwise as “non-curious", if the trial was flagged as “no-answer" and the rating was lower than the mean of the participant’s ratings. We obtained around 50 trials per curiosity level per participant (mean number of trials in class “non-curious" = 50.18 ± 12.58; mean number of trials in class “curious" = 46.85 ± 12.11).

7.1.3 *Processing & Machine Learning Algorithms*

We used two ML algorithms that are exploring multiple frequency bands: (1) a Filter Bank Common Spatial pattern (FBCSP) coupled with a Linear Discriminant Analysis (LDA) (Ang et al., 2012), and (2) a Filter Bank Tangent Space Classifier (FBTSC) (see chapter 4). Both algorithms proved effective for mental state classification from EEG (Appriou et al., 2020).

Prior work studying curiosity through EEG signals (Lima, 2019),

only used two electrodes that were placed on the frontal cortex, and the recorded signals were band-pass filtered in the alpha frequency band (8-13Hz). They did not find any correlation between activity in the frontal cortex and curiosity. It was therefore of interest to extract information from multiple electrodes (here 61) and frequency bands (here ten).

As a quick reminder, the FBCSP+LDA algorithm, works as follows: first, the training phase consists of optimally identifying and extracting both spatial and spectral features. For the spectral dimension, i.e., the frequency bands, EEG signals are filtered into ten 4Hz-wide frequency bands (in 1-4 Hz, 4-8 Hz, ..., 36-40 Hz) as in (Ang et al., 2012). Note that we included the delta band here (1-4Hz), which was not used in our contributions on cognitive workload, neither the one on emotions. For each band, the band-pass filtered EEG trials are used. Then spatial filters are built for each band using the Common Spatial Pattern (CSP) algorithm (Blankertz et al., 2008), which optimizes the EEG signal-to-noise ratio: the variance of spatially filtered signals is maximized for one class and minimized for the other class. In our study, 4 CSP filters (2 pairs) have been optimized for each frequency band, resulting in 40 features (4 CSP filters * 10 frequency bands). From those 40 features, 4 were selected using the maximum Relevance Minimum Redundancy (mRMR) feature selection algorithm (Peng et al., 2005b) to train an LDA classifier to discriminate curious from non-curious trials.

The second algorithm, FBTSC, (Appriou et al., 2020) represents EEG signals as covariance matrices and manipulates them with Riemannian geometry (Yger et al., 2016). Here, each trial is first band-pass filtered in the same ten 4Hz-wide frequency bands we used with the FBCSP (1-4Hz, 4-8Hz, ..., 36-40Hz). To design the classifier, the average spatial covariance matrix for each class (curious and non-curious) is computed for each frequency band, and all covariance matrices are then projected in tangent space at the point defined as the matrices mean. We then used the softmax function-based probabilistic output of a Logistic Regression (LR) that has been trained in the tangent space, to determine probabilities of belonging to each class. Since we have two classes and a bank of ten frequency bands, 10 pairs of probabilities were computed. From these pairs of probabilities, the four most discriminant were selected using mRMR on the training set, and then multiplied together in order to obtain 2 final probabilities, i.e., one for the “curious" class and one for the “non-curious" class. The class assigned to a test EEG trial was decided according to the highest probability.

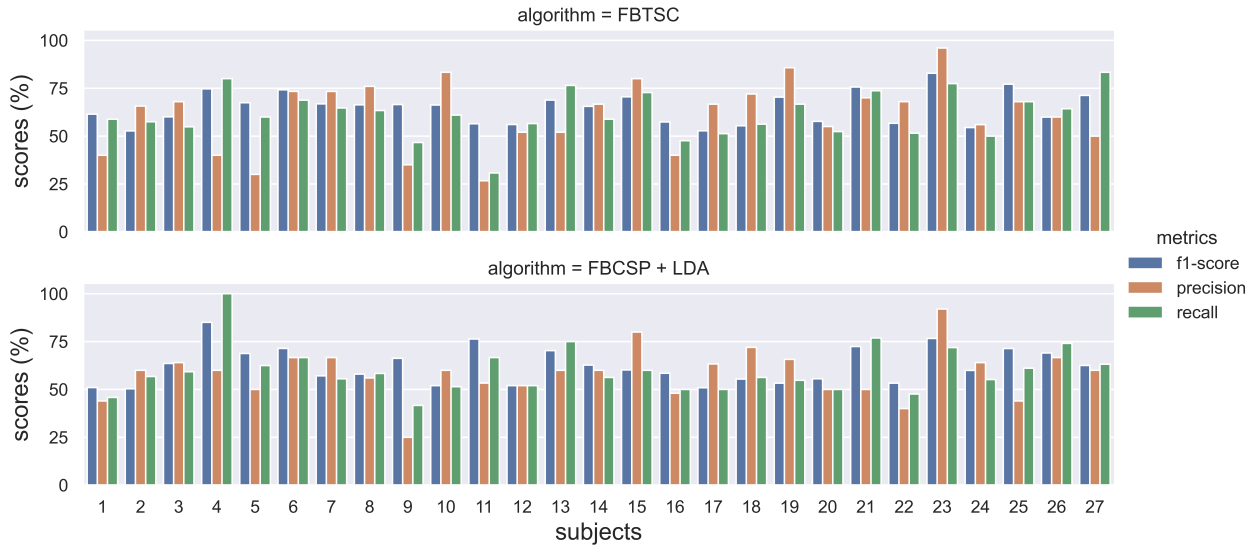

Figure 7.2: F1 score, precision and recall for the 4-seconds time window length, for each subject and for each algorithm.

7.1.4 *Evaluation*

We assessed the performances of both algorithms using a within-participant study with five-fold stratified Cross-Validation. This means the data from each participant was divided into five parts: four parts were used for training the classifier and the fifth one for testing the resulting curiosity classifier for that participant. This process was repeated five times, with each part used exactly once as the testing set.

7.2 *Results*

The F1-score, which is the weighted average of the precision and the recall, for each classifier and each time window length are reported in Figure 7.3. As a reference, the statistical chance levels using (Combrisson and Jerbi, 2015) was estimated at 51.59% (100 trials per participants on average, 27 participants).

	1 sec	2 sec	3 sec	4 sec	5 sec
FBCSP+LDA	53.40	57.34	59.87	60.93	59.82
FBTSC	57.36	55.05	62.22	63.09	62.91

Figure 7.3: Average classification performances (F1-score) across participants obtained by each algorithm with the different time window lengths.

The boxplot of the performances obtained by each algorithm with the different time window lengths are reported in Figure 7.4. We performed a 2-way ANOVA with repeated measures to evaluate the performance of the factor *Time Window* according to the factor *Algorithm* (FBCSP+LDA vs FBTSC). Note that we checked the data sphericity, as well as the normality, and used Greenhouse-Geisser (GG) correction in ANOVA if needed. The ANOVA revealed a main ef-

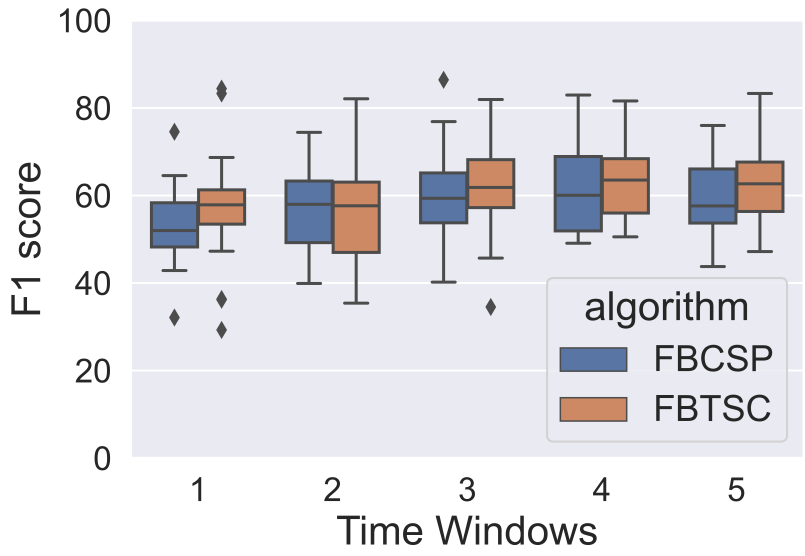

Figure 7.4: F1-score for the different time window lengths, for each algorithm.

fect of *Time Window* [GG(1,27)=0.825, p=0.00004], but not for *Algorithm* [F(1,27)=1.588, p=0.218859] nor for *Time Window*Algorithm* [GG(1,27)=0.643, p=0.285987].

Post-hoc analyses results using Student t-test with False Discovery Rate (FDR) corrections showed significant differences between the 1-second time window and 3, 4, 5-seconds time windows when using the FBCSP+LDA [$perf_{1-sec}$ = 53.4%, $perf_{3-secs}$ = 59.86%; $p \leq 0.05$, $perf_{4-secs}$ = 60.93%; $p \leq 0.05$, $perf_{5-secs}$ = 59.81%; $p \leq 0.05$]. The same method showed significant differences between the 2-second time window and 3, 4, 5-seconds time windows when using the FBTSC [$perf_{2-sec}$ = 55.05%, $perf_{3-secs}$ = 59.86%; $p \leq 0.05$, $perf_{4-secs}$ = 63.08%; $p \leq 0.05$, $perf_{5-secs}$ = 62.91%; $p \leq 0.05$]. The maximum performances for both FBCSP+LDA [$perf_{FBCSP+LDA}$ = 60.93%, *chance level* = 51.59%; $p \leq 0.05$] and FBTSC [$perf_{FBTSC}$ = 63.08%, *chance level* = 51.59%; $p \leq 0.05$] significantly outperformed the chance level for the 4-seconds time window.

Figure 7.2 shows the F1-score, precision and recall for each participant and algorithm, with the 4-seconds time window length, i.e., the time window with which both classification algorithms obtained the best performances. Still using the 4-seconds time window length, we also studied the percentage of time that each frequency band was selected (by mRMR) by each algorithm, as reported in Figure 7.5. Those results show that the ML algorithms mainly used 4 frequency bands—i.e., delta (1-4Hz), theta (4-8Hz), alpha (8-12Hz) and low beta (12-16Hz)—to classify states of epistemic curiosity.

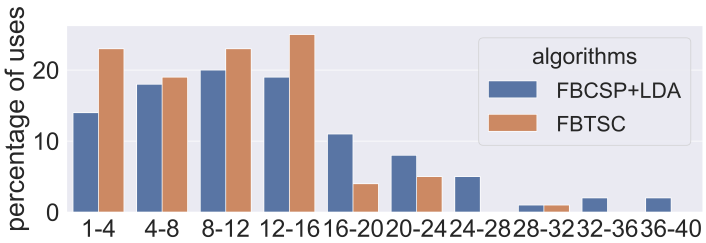

Figure 7.5: Percentage of time that each frequency band was selected by each algorithm, with the 4-seconds time window length.

7·3 *Discussion, Conclusion and Future Work*

In this contribution, we conducted an experiment aiming at collecting EEG signals during states of curiosity that were triggered using chains of trivia questions, and using ML to distinguish curious vs non-curious states. We used two ML algorithms, i.e., the FBCSP+LDA and FBTSC, to classify EEG signals at five different time-window lengths. The best results were obtained with the FBTSC, reaching about 62% of F1-score for the 3, 4 and 5s time-windows (respectively 62.22, 63.09 and 62.90%), significantly outperforming the chance level (51.59%). Results in those 3 time-window lengths also significantly outperformed the results in the 2s time window (55.05%), but not the ones in the 1s time window (57.36%). The FBCSP+LDA reached an F1-score of 59% for the 3, 4 and 5s time windows, with respectively 59.90, 60.09 and 59.81%, significantly outperforming both the chance level (51.59%) and the results in the 1s time window (53.39%), but not the ones in the 2s time window (57.34%). Overall, results indicate that both algorithms obtained better performances in the 3-to-5s time windows, suggesting a minimum time of 3 seconds to go towards curiosity state estimation based on EEG signals. Moreover, ML algorithms mostly used a range of frequency bands from delta to low beta in order to classify states of epistemic curiosity, suggesting variations of EEG activity in the low frequencies during states of epistemic curiosity.

The results also suggest that curiosity could be a mental state that increases over time, though further analysis is needed. However, these results are based on a participant-specific study, which provides a good overview of the potential of measuring states of curiosity through EEG signals, but not if we want to go towards calibration-free systems; it would be therefore interesting to run a participant-independent study as well.

As future work, we will explore ways to measure curiosity states through EEG without participant-dependent calibration. While trivia questions were used as a trigger of curiosity, new tools (e.g., social robots (Ceha et al., 2019), video games) or stimuli (e.g., videos of

magic tricks) could be used in future experiments. We will also perform deeper neurophysiological analysis to identify the EEG sensors and sources mostly modulated by curiosity levels. For example, so far, results suggested that most of the information for discriminating curiosity levels are found in theta, alpha and low beta. These frequency bands are similar to the ones used to estimate levels of workload and levels of engagement (Dehais et al., 2020). This is interesting given that both states share some common characteristics with curiosity, such as implications in long-term memory improvements. Their similarities and differences would thus need to be further studied. Further analyses can also be done to compare our results against those obtained with fMRI (Gruber et al., 2014; Kang et al., 2009), in order to gain a better understanding of the neurological markers underlying curiosity states.

8

Towards measuring states of Epistemic Curiosity through physiological signals

The previous chapter studied the classification of curiosity levels from EEG signals. To complete this first study, this chapter focuses on estimating states of Epistemic Curiosity through physiological signals. In a first section, we present the signal processing and machine learning methods we perform in order to classify such states of epistemic curiosity through heart rate, breathing and EDA signals. Then, results are presented in the next section, followed by a last section reserved to the discussion, limits and future works.

8.1 *Signal Processing & Classification*

This chapter focuses on our work that aimed at attempting to discriminate curious from non-curious states using physiological signals, i.e., Heart Rate (HR), Breathing and Electrodermal Activity (EDA). To do so, we first recorded these three types of signals using Brainproduct physiological sensors, then we extracted features from these three

types of signals using existing signal processing tools, i.e., using two Python libraries - neurokit (`https://neurokit.readthedocs.io/en/latest/index.html`) and biosspy (`https://biosppy.readthedocs.io/en/stable/`) to then apply machine learning algorithms in order to classify such physiological signals. This section describes the pre-processing steps we used, i.e., trial epoching and labelling, the machine learning algorithms performed, and finally the method used for evaluating classification performance.

8.1.1 *Pre-processing*

We first pre-processed physiological signals into 16-seconds windows. As explained in chapter 6, each 10 minutes run were divided into approximately 30 trials, i.e., one trial per trivia question displayed. More precisely, a physiological trial was defined as beginning at the time the fixation cross before the question appeared, and ending either when the willingness to present the answer or when the next fixation cross disappeared.

8.1.2 *Classification labels*

The trial labelling is the same as for the EEG trials classification, and is presented in the previous chapter in 7.1.2.

8.1.3 *Processing & Machine Learning Algorithms*

Features extraction: as explained in the introduction of this section, we used three types of physiological signals in this study, i.e., heart rate, breathing and EDA. We present the different methods that we used to extract features from each of these three types of signals.

For the ***heart rate*** signals, the entire electrocardiogram is first reduced to the R-R intervals (RRI) signals, where RRI corresponds to the interval between two successive heartbeats, or more precisely, the interval between two R peaks in the ECG. Note that we use the same methods, made for ECG, for heart rate in this thesis. We can describe the features extracted from the heart rate signal as follows:

- *sdRR:* this feature represents the standard deviation of the RRIs (Malik, 1996; Smith et al., 2013).
- *meanRR:* represents the mean of the RRI (Malik, 1996; Voss et al., 2015).
- *RMSSD:* is the Root Mean Square of the RRIs (Malik, 1996; Smith

et al., 2013; Voss et al., 2015).

- *CVSD:* is the Coefficient of Variation of Successive Differences. This corresponds to the RMSSD divided by meanRR (Malik, 1996).
- *cvRR:* is the RR coefficient of variation. This corresponds to the sdRR divided by the meanRR (Malik, 1996).
- *medianRR:* is the median of the absolute values of the RRIs' successive differences (Voss et al., 2015).
- *madRR:* RRIs' median abosolute deviation (MAD) (Malik, 1996).
- *mcvRR:* is the RRIs' median-based coefficient of variation. This corresponds to the ratio of madRR divided by medianRR (Voss et al., 2015).
- *RR50 or RR20:* successive RRIs' number of interval differences greater than 50ms or 20 ms, respectively (Malik, 1996).
- *pRR50 or pRR20:* is the proportion derived by dividing RR50 (ou RR20) by the number of RRIs (Voss et al., 2015).
- *triang:* is the HRV triangular index measurement, i.e., plotting the integral of the ratio of RRI density histogram by its height (Shaffer and Ginsberg, 2017; Smith et al., 2013).
- *Shannon_h:* Shannon entropy calculated on the basis of the class probabilities of the RRI density distribution (Voss et al., 2015).
- *VLF:* is the HRV variance in the Very Low Frequency (0.003 to 0.04 Hz) (Malik, 1996).
- *LF:* is the HRV variance in the Low Frequency (0.04 to 0.15 Hz) (Malik, 1996; Voss et al., 2015).
- *HF:* is the HRV variance in the High Frequency (0.15 to 0.40 Hz) (Malik, 1996; Voss et al., 2015).
- *Total_Power*: total power of the full density spectra (Voss et al., 2015).
- *LFHF:* is the LF/HF ratio (Malik, 1996; Voss et al., 2015).
- *LFn:* is the normalized LF power. It can be calculated using the equation "LFn = LF/(LF+HF)" (Malik, 1996).
- *HFn:* is the normalized HF power. It can be calculated using the

equation "HFn = HF/(LF+HF)" (Malik, 1996).

- *LFp:* the LF/Total_Power ratio (Voss et al., 2015).
- *HFp:* the HF/Total_Power ratio (Voss et al., 2015).
- *DFA:* is the Detrended Fluctuation Analysis (DFA) (Peng et al., 1995) of the heart rate raw signals.
- *Shannon:* is the RRIs' Shannon entropy (Voss et al., 2015).
- *sample_entropy:* is the RRIs' sample entropy (Tiwari et al., 2019).
- *correlation_Dimension:* represents the RRIs' correlation dimension (Voss et al., 2015).
- *entropy_Multiscale:* is the RRIs' entropy multiscale (Tiwari et al., 2019).
- *entropy_SVD:* RRIs' Singular Value Decomposition (SVD) entropy (Voss et al., 2015).
- *entropy_Spectral_VLF:* represents the RRIs' spectral entropy over the VLF (Voss et al., 2015).
- *entropy_Spectral_LF:* is the RRIs' spectral entropy over the LF (Voss et al., 2015).
- *entropy_Spectral_HF:* is the RRIs' spectral entropy over the HF (Voss et al., 2015).
- *Fisher_Info:* is the RRIs' Fisher information (de Geus et al., 2019).
- *Lyapunov:* if the RRIs' Lyapunov exponent (Goshvarpour and Goshvarpour, 2012).
- *FD_Petrosian:* is the RRIs' Petrosian's Fractal Dimension (Petrosian, 1995).
- *FD_Higushi:* is the Higushi's Fractal Dimension of RRIs (Gomes et al., 2017).

The study of ***the breathing signals*** mainly focuses on R-R intervals (RRI), as for the heart rate signals, and is known as the Breathing Rate Variability (BRV). RRI corresponds to the interval between two successive breathing, or more precisely, the interval between two R peaks in the breathing signals. Based on these RRIs, multiple characteristics can be extracted from the signals, and are listed as follows:

- *peak_length:* is the interval of successive peaks in the breathing pattern signal (Jaiswal et al., 2019).
- *trough_length:* is the interval of successive troughs in the breathing pattern signal (Jaiswal et al., 2019).
- *peak_amplitude:* is the amplitude calculated for each peak of the trial (Jaiswal et al., 2019).
- *trough_amplitude:* is the amplitude calculated for each trough of the trial (Jaiswal et al., 2019).
- *resp_rate:* corresponds to the breathing rate, obtained from frequency domain analysis of the breathing signals (Jaiswal et al., 2019).

Descriptive statistics are calculated on these characteristics, i.e., mean, standard deviation, min, max, first quartile, median and the third quartile, and used as features to feed machine learning algorithms.

Another type of features is also recommended when manipulating breathing signals: the frequency domain-based features. To do so, we also used as features the Power Spectral Density (PSD) calculated from 0.1 to 0.5 Hz, in 0.01 Hz steps.

The ***Electrodermal activity (EDA)*** can be described as the superposition of two distinct skin conductance responses (SCRs): on a one hand, we have the tonic activity, and on the other hand the phasic activity (Fritz et al., 2014). In the time domain, features can be extracted from these two components through descriptive statistics, characteristics based on deeper information can be obtained from the phasic component. These characteristics are described as follows:

- *phasic_peak_amplitude:* represents the amplitude of phasic peaks (Schmidt and Walach, 2000).
- *phasic_peak_longitude:* is the rise time/duration of the peaks (Schmidt and Walach, 2000).
- *phasic_peak_slope:* represents the slope of the peaks (Parent, 2019).
- *ordinate_slope* is the ordinate of the slope of the peaks, i.e., the starting point (Parent, 2019).
- *peak_peak_interval:* corresponds to the inter-peaks time (Parent, 2019).

Descriptive statistics are then calculated on all these extracted char-

acteristics, as well as on both tonic and phasic components, in order to define features representing the EDA signals. First, we have the basics statistics - i.e., mean, standard deviation, min, max, first quartile, median and the third quartile - that are calculated for each of these characteristics. Second, two statistics, i.e., skewness and kurotis, are calculated for both the tonic and the phasic components (Braithwaite et al., 2013). Finally, the "nb_peak_per_min" corresponds to the frequency of phasic peaks (Parent, 2019), and can be defined as a feature in itself (no descriptive statistics are needed).

Concerning the frequency domain, the EDA dynamics of the frequency spectrum is largely contained in frequencies below 0.4 Hz (Shimomura et al., 2008). Two types of frequency information are calculated, i.e., the Power Spectral Density (PSD) from 0.0 to 0.1 Hz, in 0.01 Hz steps (Parent, 2019) and the PSD from 0.045 to 0.25 Hz (Posada-Quintero et al., 2016).

Classification: we chose to run three studies - one for each type of signals, i.e, heart rate, breathing and EDA - in order to obtain indications about which is the best signal to use to classify levels of curiosity. Therefore, for each of our three signals, 10 features were selected using the maximum Relevance Minimum Redundancy (mRMR) feature selection algorithm (Peng et al., 2005b) in order to train an LDA classifier to discriminate curious from non-curious trials. Finally three different LDA models have been trained, one with features for heart rate signals only, a second one with breathing signals only and a last one with EDA signals only.

8.1.4 *Evaluation*

We assessed the performances of ML algorithms on the three types of signals using a within-participant study with five-fold stratified Cross-Validation. This means the data from each participant was divided into five parts: four parts were used for training the classifier and the fifth one for testing the resulting curiosity classifier for that participant. This process was repeated five times, with each part used exactly once as the testing set.

8.2 *Results*

The F1-score, which is the weighted average of the precision and the recall, for each type of signals, and for each participants, are reported on Figure 8.1. As a reference, the overall statistical chance level

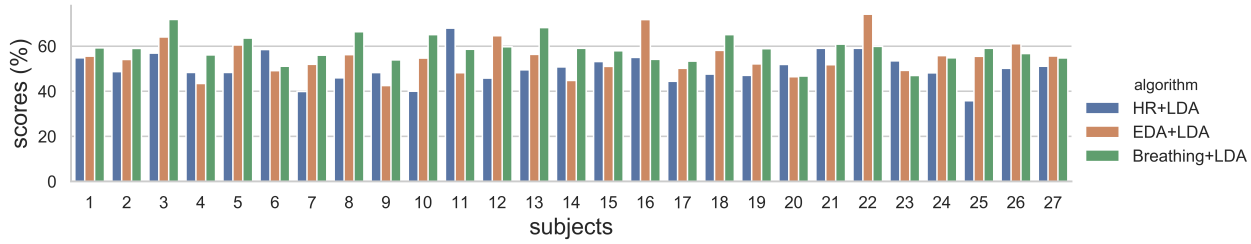

Figure 8.1: F1 score, precision and recall for each subject and for each algorithm, i.e., HR+LDA, breathing+LDA and EDA+LDA.

(across participants) using (Combrisson and Jerbi, 2015) was estimated at 51.59% (100 trials per participant on average, 27 participants).

The boxplot of the performances obtained by each types of signals are reported in Figure 8.2. We performed a 1-way ANOVA to evaluate the performance of the factor *Algorithm* where the three "types" of algorithms are HR features coupled with LDA, breathing features coupled with LDA and EDA features coupled with LDA. Note that we checked the data sphericity, as well as the normality, and used Greenhouse-Geisser (GG) correction in ANOVA if needed. The ANOVA revealed a main effect of *Algorithm* [F(1,27)=10.878, p=0.0001].

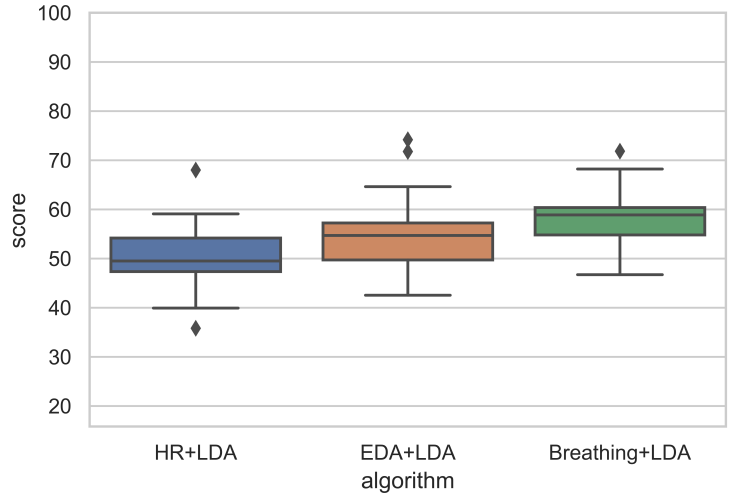

Figure 8.2: F1-score for each type of physiological signals, i.e., heart rate, breathing and EDA.

Post-hoc analyses results using Student t-test with False Discovery Rate (FDR) corrections showed significant differences between the breathing+LDA algorithm and both the HR+LDA and the EDA+LDA [$perf_{breathing+LDA}$ = 58.4%, $perf_{HR+LDA}$ = 50.1%; $p \leq 0.05$, $perf_{EDA+LDA}$ = 54.8%; $p \leq 0.05$]. Those results are presented on Figure 8.3. Note that only the breathing+LDA algorithm [$perf_{breathing+LDA}$ = 58.41%, *chance level* = 51.59%; $p \leq 0.05$] significantly outperformed the chance level.

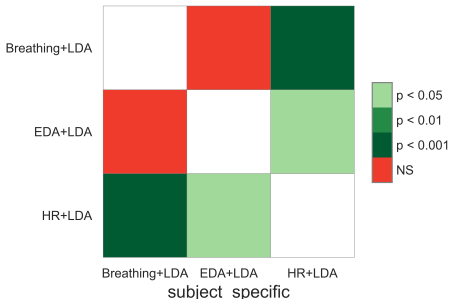

Figure 8.3: Post-hoc tests for each type of physiological signals, i.e., heart rate, breathing and EDA.

8.3 *Discussion, Conclusion & Future Work*

In this contribution, we collected physiological signals, i.e., heart rate, breathing and EDA, using the same experiment as in chapter 7, which is presented in chapter 6. In this experiment, states of curiosity were triggered using chains of trivia questions, and we used multiple signal processing methods, as well as a single ML algorithm, i.e., the LDA, in order to distinguish curious from non-curious states. We extracted between 40 and 60 features from each type of signals. We ran three parallel studies, one for each type of signal, in order to observe which type of signal could be interesting for discriminating different levels of curiosity. For each of these studies, we first selected 10 features among the total number of features for this type of signal, using the mRMR feature selection algorithm, and then applied an LDA. Results showed that only the breathing signals-based features coupled with the LDA allowed us to obtain classification performances (58.41%), significantly outperforming the chance level (51.59%). To a lesser extent, results indicated that EDA signals-based features coupled with the LDA obtained classification performances of 54.8% significantly higher than the chance level as well (51.59%). Finally, the heart rate signals coupled with the LDA only obtained 50.1% classification accuracy.

These results highlight whether these three types of signals can be used in short time windows, i.e., in 16-seconds time windows, in order to classify levels of curiosity through physiological signals. However, the experimental protocol was designed for the measurement of EEG signals mainly, meaning the installation of physiological sensors was made in addition, and only short time windows trials were thus available for classifying such states of curiosity through physiological signals. The fact that we did not have control over this factor makes our study purely exploratory. However, the results concerning the study of breathing signals are very encouraging since we obtained classification performances that are already significantly better than the chance

level. If the time windows in this study may not be optimal (they are typically too short for physiological signals analysis), they still allow us to do the classification of these physiological signals on experimental designs similar to the EEG ones, which is interesting because it means that we could get information from these different types of sensors on a common protocol design. So far, we did not have time to combine one or more types of these physiological signals with EEG signals: we therefore cannot conclude yet on the possible fruitfulness of such a combination of signals in order to classify different levels of curiosity. Therefore, as future work, it would be interesting to explore ways to measure curiosity levels through physiological signals on longer time windows. This would require to run another experiment, in order to induce curiosity states on a longer duration. While Trivia questions were used as a trigger of curiosity in our studies, new tools - e.g., social robots (Ceha et al., 2019) - or stimuli - videos of magic tricks - could be used in future experiments. Based on the data set we obtained by running the experiment described in chapter 6, and used for this contribution, further analysis should be conducted. First, we will explore ways to measure curiosity states through physiological signals with a participant-independent calibration. Second, it would be interesting to see if a combination of features from two or three types of physiological signals would allow classifiers to better discriminate curious from non-curious trials. Finally, we will run an analysis aiming at combining physiological features - at least from breathing signals, and maybe from EDA and HR as well - to EEG features in order to feed the classifiers with information from both brain and body activity.

In summary, we ran an experiment in which we used both electroencephalography (EEG) and physiological signals, i.e., heart rate, breathing and electrodermal activity (EDA), to measure the neurophysiological activity of participants as they were induced into states of curiosity, using trivia question and answer chains. Concerning the EEG signals anaylsis, we used two ML algorithms, i.e. Filter Bank Common Spatial Pattern (FBCSP) coupled with a Linear Discriminant Algorithm (LDA), as well as a Filter Bank Tangent Space Classifier (FBTSC) (that we proposed in this thesis), to classify the curious EEG signals from the non-curious ones. Concerning the physiological signals, we used multiple signal processing methods in order to extract features from the physiolgical signals, as well as a single ML algorithm, i.e., the LDA, to distinguish curious from non-curious states. We ran three parallel studies, one for each type of signal, in order to observe which type of signal could be interesting for discriminating different levels of curiosity.

To summary, global results showed that it was possible to discriminate states of epistemic curiosity through both EEG and physiological signals, with classification accuracies better than the chance level, i.e., 63.09% from EEG signals, 58.41% for breathing signals, and 54.8% for EDA signals. However, even if these differences are significant, they remain low, and future work should be conducted in order to improve these classification performances, e.g., new methods for inducing states of curiosity, new time windows for trials, etc.

Part IV

Towards measuring states of Cognitive Workload through EEG during a MT-BCI task

PhD Thesis Roadmap

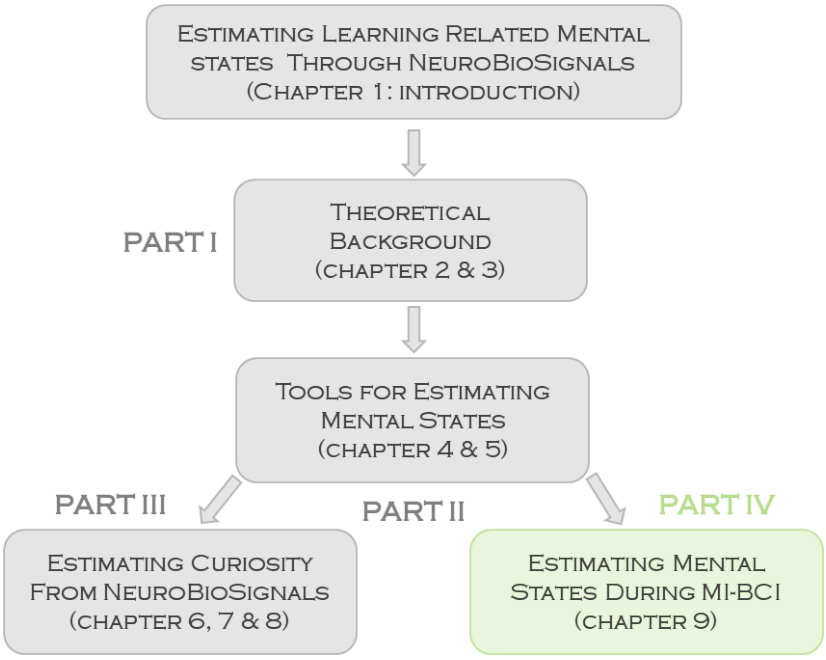

Figure 8.4: PhD thesis roadmap.

Related Papers

9 Towards estimating cognitive workload during MI-BCI training

9.1 Research question

As explained in the introduction chapter, estimating cognitive, affective or conative states from brain signals is a key but challenging step in the creation of passive BCI applications. Indeed, this would allow applications to monitor the users' states in real-time, and therefore adapt the interactions to individuals cognitive capabilities, i.e., optimal levels of a given cognitive state (workload, attention, etc), but also optimal levels of emotions (e.g., valence or arousal) or motivation (e.g., motivational states, curiosity). Moreover, all cognitive, affective and conative states have been shown to be involved in human learning, thus being able to estimate such states in real-time would play a major role for upgrading BCI training protocols, although this type of system could also be used as an evaluation method for HCI.

However, as seen in chapter 2, learning new skills requires information processing laying on Working Memory resources (Sweller et al., 2019). These resources are limited and distributed over the different processes involved by the learning material. Learning material should then be adapted to the learner's resources. It is important to understand how the cognitive state of the user evolves during the learning

process in order to predict the optimal user-training interaction that might result in an efficient learning. This should also be applied to Mental Imagery (MI)-BCI control training, which is often suboptimal because of a lack of understanding on the best way to train users, on users' states variabilities and on how to adapt it for better learning (Jeunet et al., 2017).

Thus, we propose a study aiming at estimating cognitive workload levels through EEG signals during a MI-BCI task. As indicated in the introduction chapter, one of the longer term goal of this thesis is to contribute to estimating learning-related mental states during MI-BCI user training, to later adapt the training to such states. While this PhD thesis work did not aim at completing all these objectives alone, it did contribute to go in that direction. In particular, we contributed to the design, implementation and conduction of a protocol to do so, that we report in this chapter. The study is still ongoing, and will be completed by another PhD student on the BrainConquest project. Thus no results are presented in this thesis. However, several important advances have been made in this project, i.e., the protocol has been designed, implemented validated by the ethics committee from Inria Bordeaux Sud-Ouest (COERLE) and the experiment has been run on several subjects, the data collection is not finished and no analysis have been made so far. However, the goal of this chapter is to present the work that has been conducted so far. First, we present the different tasks that users had to do in order to 1) estimate the working memory load capabilities of the participant 2) induce and measure levels of cognitive workload independently of the MI-BCI task, i.e., passive BCI tasks 3) performing the MI-BCI task, i.e., active BCI task. Second, we present the different questionnaires that have been proposed to participants in order to subjectively measure both fatigue and cognitive workload[1]. In the future, once the experiment will be fully completed, the ML tools developed in this thesis (notably FBTSC) will be used to estimate workload levels from EEG during the MI-BCI training tasks. This should enable us to assess finely the relationship between workload levels and MI-BCI performance and learning.

[1] Note that for each step, we describe the theoretical background behind any decision we took.

9.2 *Methods*

So far, we recruited 8 participants that participated to three MT-BCI training sessions, each of them lasting 2 hours and taking place on a different day. Figure 9.1 details the time of each step of the experiment.

Working memory abilities may vary for a participant to another,

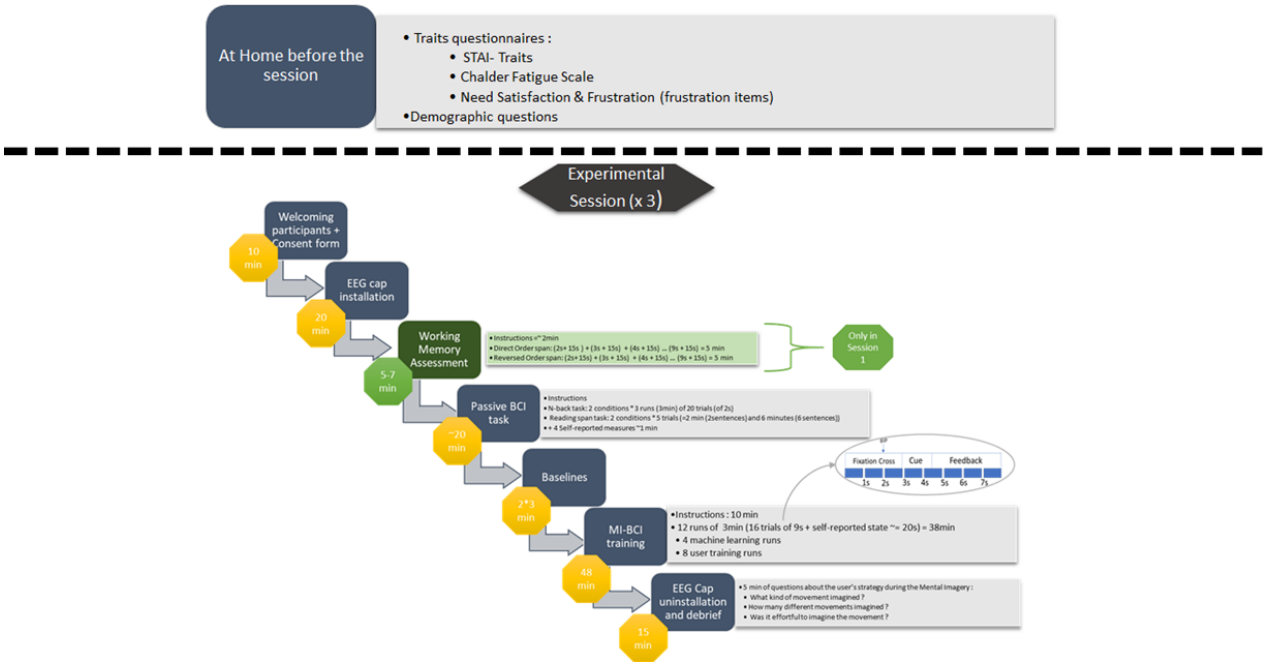

Figure 9.1: This figure summarizes our protocol with the pre-session part, the approximate time allocated to each part of the training, and details of each task. One session lasted approximately 2 hours.

this is why we first assessed participant's working memory abilities with the number span task from the Weschler Adult Intelligence Scale (Wechsler, 1955). Also, users were asked to fill in different psychometric tests, during the sessions, to assess their cognitive states. During the session they were equipped with an EEG cap in order to control the system with mental imagery (i.e. MI-BCI training). In order to later be able to monitor a subject's workload from EEG during the training, participants first completed a passive BCI task. Our hypothesis is that cognitive load will vary during the training and will be linked with learning performances. We predict that the task design will have an effect on the participant's cognitive load and learning performances.

9.2.1 *Working Memory assessment*

Working Memory (WM) processing is involved in learning and its resource limitations during a task can influence cognitive load and then learning performance. As resource limitations can vary between individuals, we assessed participants WM abilities. To do so, we used a number span task from the Weschler Adult Intelligence Scale (Wechsler, 1955). First, participants were asked to perform the Forward Digit Span task (Kreutzer et al., 2011b) (presented in chapter 2), and then the Backward Digit span task (Kreutzer et al., 2011a) (also presented in chapter 2).

9.2.2 *Passive BCI task*

In order to estimate users' cognitive load from their EEG signals during the MI-BCI training, we needed to record a ground truth corresponding to low and high cognitive load periods. To do so, we used a protocol with tasks already described in the literature as inducing high or low cognitive load to participants. Multiple methods to induce and record workload variations with EEG are available, as presented in chapter 2, and based on our literature review in chapter 3, we chose the two most used, i.e., the N-back task and the Rspan one. Note that we already presented these two tasks in chapter 2.

As seen in chapter 2, EEG is an efficient non-intrusive measure of the brain activity, yet it is overly sensitive to behaviours that might produce noise in the recording. Therefore, to record brain activity related to the cognitive task only, we had to avoid any source of noise, like movements (body, eyes, face etc.). The N-back task and the Rspan task (around 10 minutes each) have already been used in EEG studies and are then already adapted to this material. They have already been used as passive BCI techniques to observe workload variations, and crossing them will allow us to have a more general measure of workload associated brain signals (Walter et al., 2013). At the end of each workload condition in the two tasks, we assessed the participant's cognitive state using subjective measures based on the Nasa-TLX (Cegarra and Morgado, 2009). In total, the passive BCI tasks lasted approximately 20 minutes.

9.2.3 *MI-BCI training*

Once they have completed the passive BCI tasks, after a break, participants practiced a MI-BCI task in which they were trained to imagine a movement of the left or the right hand that could be recognized by the system. The training session lasted approximately 110 minutes (1 hour 50 min) in which participants completed 12 runs, lasting three minutes each, of MI-BCI task. With the user's instruction and self-reported measures at the end of each run, the MI-BCI training lasted approximately 45 minutes.

- the goal of the first 4 runs was to calibrate the system by providing examples of EEG patterns associated with each mental imagery.

- the goal of other runs was to train the user to produce a clear and stable mental imagery, in order to help the system to recognize them, and helping the user by providing a feedback paradigm. Each run included 15 trials of 9 seconds each. At the end of the run,

the user were asked to assess its state with six self-reported scales.

Description of the trial

At the beginning of each trial, a white cross was displayed. After 2 s, an auditory cue (a beep) triggered the attention of the participant towards the red arrow, which was displayed at t = 3 s for 1 s, and indicated which task the participant had to perform (imagining right- or left-hand movements upon appearance of a right or left arrow, respectively). At 4.25 s, a blue feedback bar appeared and was updated at 16 Hz for 3.75 s. This feedback direction indicated if the imagined movement was recognized by the classifier, and its length was proportional to the classifier output 9.3.2.

9.2.4 *Cognitive workload measures*

EEG signals

In previous studies, cognitive load variations have been associated with EEG oscillation changes, particularly in the theta and alpha bands (Antonenko et al., 2010; Brouwer et al., 2012; Grimes et al., 2008; Paas et al., 2003). While the MI-BCI training brain activity was already recorded with EEG, we used these features to observe objective workload variations.

subjective assessment

We wanted to assess the subjective cognitive load for the participants during all the training, and not only at the end of the session or between runs. Cognitive load is not an unidimensional measure, it depends on the task demand perceived by the subject and the efforts it takes him/her to process the task (Parent, 2019). Also, it can be influenced by other user's state factors, such as their anxiety state, their mental fatigue, or their frustration. To assess the evolution of the cognitive states during the task we asked the participant to declare, on a 10-points scale going from "low" to "high", his/her level of perceived "mental demand", "mental effort", "frustration", "anxiety" and "mental fatigue". Each participant reported his/her own state at the end of each task, i.e., after both passive BCI tasks, as well as after the active BCI task. They then reported it four times during the passive BCI task and after each run during the MI-BCI task, so 12 times in total. These scales are based on the French adaptation of the Nasa-TLX (Cegarra and Morgado, 2009).

Reaction time

Self-reported questionnaires inform us on the user's experience during the task but are not always representative of the real amount of cognitive load engaged in the task. An important mental effort can cause mental fatigue which can be characterized by increased reaction/response times, yet it is not always correlated with a self-declared feeling of fatigue. For this reason, we will estimate the reaction time (i.e., the time between the "cue" apparition and the EEG changes) of the participants. This measure will allow us to observe whether the subjective mental effort and fatigue are correlated with differences in reaction times.

Other subjective measures

While anxiety can modulate users mental states, and therefore have an impact on the brain activity recorded as EEG signals, participants had to fill-in the French version of the STAI (Spielberger et al., 1983) before MI-BCI training, in order to control anxiety state and propension before the training. The STAI trait is composed of two questionnaires, the first one assesses the anxiety trait of the participants as the second assesses its anxiety state. Each is composed of 20 questions and individuals are asked to answer on a four levels scale going from "almost never" to "almost always" for the "trait" questionnaire, and from "no" to "yes" for the "state" questionnaire.

9.2.5 *Participants' declarations*

At the end of the session, we asked participants to answer questions during a conversation with the experimenter. The questions/answers provided us qualitative information about users' strategies during the MI-BCI task and will help us to interpret and discuss the results. Because it was done at the end of the training, we did not want to add too much effort for the subjects, this was like a conversation. We asked the participants if they agree to be recorded so we can keep a trace of the discussion.

9.3 *Materials*

9.3.1 *Measure of the neurophysiological data*

The EEG signals has been recorded with 48 scalp EEG electrodes (according to the standard in 10-20 EEG system), referenced to the left ear and grounded to AFz, using a g.tec g.USBAmp amplifier and active electrodes. Such electrodes cover the whole cortex, where EEG vari-

ations due to MI and mental workload could be measured. To those electrodes, we added two EMG (ElectroMyoGraphy) electrodes (left and right hand wrist) and three EOG (ElectroOculoGraphy) electrodes to record eye artifacts (below, above and on the side of an eye).

9.3.2 *Online EEG signal processing for MI-BCI control*

To classify the two MI tasks online from EEG data, we used participant-specific spectral and spatial filters. First, from the EEG signals recorded during the calibration runs, we identified a participant-specific discriminant frequency band using the heuristic algorithm proposed by Blankertz et al. in (Blankertz et al., 2008) (Algorithm 1 in that paper). Roughly, this algorithm selects the frequency band whose power in the sensorimotor channels maximally correlates with the class labels. Here we used channels C3 & C4 after spatial filtering with a Laplacian filter as sensorimotor channels, as recommended in (Blankertz et al., 2008). We selected a discriminant frequency band in the interval from 5 Hz to 35 Hz, with 0.5Hz large bins. Once this discriminant frequency band identified, we filtered EEG signals in that band using a butterworth filter of order 5.

Then, we used the CSP algorithm (Blankertz et al., 2008) to optimize 3 pairs of spatial filters, still using the data from the two calibration runs. Such spatially filtered EEG signals should thus have a band power which is maximally different between the two MI conditions. We then computed the band power of these spatially filtered signals by squaring the EEG signals, averaging them over a 1 second sliding window (with 1/16th second between consecutive windows), and log-transforming the results. This led to 6 different features per time window, which were used as input to a LDA classifier (Lotte et al., 2018b). This LDA was calibrated on the data from the four calibration runs. These filters and classifier were then applied on the subsequent runs to provide online feedback.

9.4 *Discussion & Future work*

As explained in the introduction of this chapter, and more generally in the introduction of this PhD thesis, being able to estimate learning-related mental states such as cognitive workload, from EEG signals during a MI-BCI task, would be beneficial to the future of active BCI training as it would allow us to adapt the training task to users' capabilities, e.g., to users' cognitive capabilities. We thus took into account

lessons we have learnt from other works & contributions of this PhD thesis - i.e., from the literature review about studies aiming at estimating levels of cognitive workload through EEG signals (see chapter 3), as well as from the study aiming at decoding levels of cognitive workload using modern and promising machine learning algorithms (see chapter 4) - in order to go towards the estimation of cognitive workload during MI-BCI training. This chapter presented the protocol design that has been validated by the ethical committee of Inria, as well as the current state of the project. At the time of writing, the experiment is still running since it has been slowed down due to the Covid19-pandemics, and 8 participants have already completed their 3 sessions. The next step will therefore be to analyze these data: we will train models by using algorithms that proved to be the most efficient for cognitive workload levels classification, i.e., the algorithms we presented in the chapter 4 (and notably the FBTSC), on EEG signals recorded during the passive BCI tasks (both N-back and Rspan tasks). We will then use these models in order to attempt to classify levels of cognitive workload as well, but this time on EEG signals that have been recorded during the active BCI task, i.e., the MI-BCI task. Finally, we will verify if this estimation is correlated to the active MI-BCI performances, e.g., if a drop in users' performances would be correlated to users' cognitive workload being too high, i.e.,"overloaded". Another verification that will have to be done is the study of potential correlations between the cognitive workload variations and the MI-BCI learning.

If cross-task studies concerning the estimation of cognitive workload levels have been made (Gerjets et al., 2014; Krol et al., 2016; Walter et al., 2013), these methods have never been applied to MI-BCI tasks, making this study an important contribution in order to evaluate the feasibility of such a system. If this experiment proves to be a success, it would be a major step to go towards real-time adaptations of active BCI training to users' cognitive states. Moreover, this will encourage research studies that attempt to estimate other learning-related mental states such as affective or conative states, in order to adapt learning systems to the mood or motivation of the users.

Part V

Discussion & Prospects

10

Discussion & Perspectives

10.1 Estimating learning-related mental states through EEG signals: where are we?

As we saw in the first chapter (see chapter 2), the history of psychology includes several theories of learning. Among them, the theory entitled "Trilogy of Mind", brought by Hilgard in 1980 (Hilgard, 1980), proposes that the mind has three main faculties: cognition, affect and conation. Users' states can thus be split into three categories: cognitive, affective and conative states. Among cognitive states, we saw that the mental states such as cognitive workload, fatigue, stress, engagement, distraction and attention are related to learning. Concerning the affective states, learning-related mental states can be listed as follows: emotions, surprise, satisfaction, frustration, disillusionment, boredom and the flow. Finally, the conative states that are related to learning are the intrinsic motivation, volition, agency, self-direction, and self-regulation.

Among all the users' mental states, we decided to focus on three mental states in this PhD thesis. Among the cognitive states, we choose to focus on the estimation of cognitive workload through EEG signals, as it showed to be widely studied in the literature (see chapter 3) and recommendations about the materials to use to continuously assess cognitive workload in real time have been proposed by Gerjets et al.

(Gerjets et al., 2014). More importantly, this mental state is thought to be very relevant for learning (Sweller et al., 1998). Indeed, as recalled in the introduction to this discussion, the longer term goal of this thesis is to contribute tools to estimate levels of learning-related mental states - in real time - during an active BCI task.

Concerning the affective states, we chose to focus on emotion as the literature about the estimation of such a mental state is important as well (see chapter 3). Moreover, a public data set containing EEG signals of different emotion levels on three scales (valence, arousal and dominance), i.e., "DEAP" from Koelstra et al. (Koelstra et al., 2011), is available online. This data set has been frequently used in the literature of studies aiming at estimating emotions through EEG signals (see chapter 3), facilitating the comparisons between studies and the development of machine learning tools with a view to improve emotion level discrimination (see chapter 4). Finally, no conative state had been studied to any great extent previously, which is why we have decided to focus on curiosity as it is a conative state that is strongly related to learning, and particularly on epistemic curiosity as it has been defined as a desire to acquire knowledge, i.e., to learn (Loewenstein, 1994).

We thus focused on these three mental states in the PhD thesis. As mentioned right before, the estimations of both cognitive workload and emotions through EEG signals have been widely studied in the literature. We thus first reviewed this literature, in order to understand what has been done so far concerning the estimation of both cognitive workload and emotions, and what remains to be done to go towards real-time estimation of these two mental states. As only a very little has been done for attempting to estimate states of epistemic curiosity, based on the literature, we decided to study more in depth this conative state, as it revealed to be important for human learning (see chapter 2). We therefore built a protocol design and ran an experiment in order to go towards the estimation of epistemic curiosity through neurophysiological signals (see part III).

10.2 *Contributions of this PhD thesis*

This manuscript enabled us to describe the contributions related to the three challenges depicted in Figure. These contributions are summarised in the following paragraphs.

First, proposed a detailed state-of-the-art, which synthesizes the

protocols, measures and tools (ML) used to estimate both cognitive workload and emotions through EEG signals (see chapter 3).

Based on this literature review concerning the cognitive workload and emotions estimation from EEG (see chapter 3), the classification accuracies obtained so far - mostly around 70% for cognitive workload, and around 60-65% for emotions in oscillatory-based studies - revealed the need for more robust and accurate EEG classification algorithms, in order to obtain trustable EEG-based cognitive and affective states estimators.

Our ***contribution #1*** was therefore to study new and promising algorithms that proved efficient either in recent active BCI classification competitions (Ang et al., 2012; Yger et al., 2016), such as Riemannian geometry classifiers, or in other fields of artificial intelligence, such as Deep Learning (Lecun et al., 2015; Schirrmeister et al., 2017) (see chapter 2). Indeed, they have shown to be promising for other BCI systems, e.g., motor imagery-BCIs, but they have not been formally studied and compared together for cognitive or affective states classification. We thus explored such machine learning algorithms, proposed new variants of them (i.e., Filter Bank Tangent Space Classifier (FBTSC) and Filter Bank Minimum Distance to Mean with geodesic filtering classifier (FBFgMDM)), and benchmarked them with classical methods to estimate both cognitive workload and affective states (Valence/Arousal) from EEG signals. We studied these approaches with both subject-specific and subject-independent calibration, to go towards calibration-free systems. Our results suggested that a CNN obtained the highest mean accuracy, although not significantly so, in both conditions for the mental workload study - 72.73% in the subject-specific study, 63.74% in the subject-independent study - outperforming state-of-the-art methods on this data set, followed by RGCs. However, this same CNN underperformed in both conditions for the emotion data set, a data set with small training data. On the contrary, RGCs proved to have the highest mean accuracy with the Filter Bank Tangent Space classifier (FBTSC), for the subject-specific condition on the valence data-set (61.09%) and for the subject-specific condition on the arousal data-set (60.60%). Our results thus contributed to improve the reliability of cognitive and affective states classification from EEG. They also provide guidelines about when to use which machine learning algorithm, which will be interesting to go towards real-time estimation of these cognitive workload and emotions states.

However, the two mental states, i.e., cognitive workload and emotions, that we studied in this *contribution #1*, are far from being the only mental states of interest for learning (see chapter 2). In addition,

we studied them through two data sets only, i.e., one for cognitive workload and the other for emotions, meaning one protocol design for each of these mental states: as shown with through our literature review in chapter 3, multiple other protocols and methods to induce, measure and estimate these mental states exist. In a more general way, experiments using (EEG)-based passive BCIs dramatically increased in the last decade: the variety of protocol designs and the growing interest for physiological computing require parallel improvements in processing and classification of EEG signals, but also bio signals such as electrodermal activity (EDA), heart rate (HR) or breathing. If some EEG-based analysis tools are already available for online BCIs with a number of online BCI platforms (e.g., BCI2000 or OpenViBE), it remains crucial to perform offline analyses in order to design, select, tune, validate and test algorithms before using them online. Moreover, studying and comparing those algorithms usually requires expertise in programming, signal processing and machine learning, whereas numerous BCI researchers come from other backgrounds with limited or no training in such skills. Finally, existing BCI toolboxes are focused on EEG and other brain signals, but usually do not include processing tools for other bio signals. Therefore, in our ***contribution #2***, we designed, developed, validated and shared BioPyC, a free, open-source and easy-to-use Python platform for offline EEG and biosignal processing and classification. Based on an intuitive and well-guided graphical interface, four main modules allow the user to follow the standard steps of the BCI process without any programming skill 1) reading different neurophysiological signal data formats 2) filtering and representing EEG and bio signals 3) classifying them 4) visualizing and performing statistical tests on the results. Such a tool will allow researchers to estimate learning-related mental states with multiple methods for inducing and measuring these states, as well as multiple algorithms for processing the signals and classifying them.

While the estimation of both cognitive workload and emotions is widely studied through EEG signals, only one study attempted to run such an experiment on curiosity (Lima, 2019). However, understanding the neurophysiological mechanisms underlying curiosity, and therefore being able to identify the curiosity level of a person, would provide useful information for researchers and designers in numerous fields such as neuroscience, psychology, and computer science. Moreover, being able to estimate curiosity, which is one of the mental states related to learning (Gottlieb and Oudeyer, 2018), would enable to monitor the user's curiosity levels in real time and adapt the interaction accordingly. This would therefore be beneficial to designers of interactive systems, who wish to adapt the interaction paradigm or applica-

tion content to users' levels of curiosity. In the context of active BCIs, which are known to be notoriously difficult for novices to use (see chapter 1.1), it would be beneficial to adapt the BCI training tasks to the mental states of these users, e.g., their curiosity level, and adapt the BCI tasks to users' interests in order to prevent boredom and improve learning. A first step to uncovering the neural correlates of curiosity was to collect neurophysiological signals during states of curiosity, in order to develop signal processing and machine learning (ML) tools to recognize the curious states from the non-curious ones. Thus, in our ***contribution #3***, we ran an experiment in which we used both electroencephalography (EEG) and physiological signals, i.e., heart rate, breathing and electrodermal activity (EDA), to measure the neurophysiological activity of participants as they were induced into states of curiosity, using trivia question and answer chains. Concerning the EEG signals anaylsis, we used two ML algorithms, i.e. Filter Bank Common Spatial Pattern (FBCSP) coupled with a Linear Discriminant Algorithm (LDA), as well as a Filter Bank Tangent Space Classifier (FBTSC) (that we proposed in this thesis), to classify the curious EEG signals from the non-curious ones. Global results indicate that both algorithms obtained better performances in the 3-to-5s time windows, suggesting an optimal time window length of 4 seconds (63.09% classification accuracy for the FBTSC, 60.93% classification accuracy for the FBCSP+LDA) to go towards curiosity states estimation based on EEG signals. Concerning the physiological signals, we used multiple signal processing methods in order to extract features from the physiolgical signals, as well as a single ML algorithm, i.e., the LDA, to distinguish curious from non-curious states. We ran three parallel studies, one for each type of signal, in order to observe which type of signal could be interesting for discriminating different levels of curiosity. Results showed that breathing+LDA obtained classification performances (58.41%) that significantly outperformed the chance level (51.59%). Other results indicated that EDA signals obtained classification performances of 54.8%, when the HR signals only obtained 50.1%. To summary, global results showed that it was possible to discriminate states of epistemic curiosity through both EEG and physiological signals, with classification accuracies better than the chance level, i.e., 63.09% from EEG signals, 58.41% for breathing signals, and 54.8% for EDA signals. However, even if these differences are significant, they remain low, and future work should be conducted in order to improve these classification performances, e.g., new methods for inducing states of curiosity, new time windows for trials, etc.

In our ***contribution #4***, we proposed a study aiming at estimating cognitive workload levels through EEG signals during a motor-

imagery (MI)-BCI task. Indeed, being able to estimate learning-related mental states such as cognitive workload, from EEG signals during a MI-BCI task, would be beneficial to the future of active BCI training protocols as it would allow us to adapt the training task to users' capabilities, e.g., to users' cognitive capabilities. If major advances have been made in this project so far, i.e., the protocol has been designed and validated by the ethics committee from Inria Bordeaux Sud-Ouest (COERLE), and the experiment has been run, no analysis have been made so far. However, the goal of this chapter 9 was to present the work that has been conducted so far in this ongoing project. First, we presented the different tasks that users had to do in order to 1) estimate the working memory load capabilities of the participant 2) induce and measure levels of cognitive workload independently to the MI-BCI task, i.e., passive BCI tasks 3) performing the MI-BCI task, i.e., active BCI task. Second, we presented the different questionnaires that have been proposed to participants in order to subjectively measure both fatigue and cognitive workload.

10.3 *Limits of this PhD thesis*

This PhD thesis made four main contributions. The ***contribution #1*** is a study that proposed new algorithms and that aimed at comparing recent and promising machine learning algorithms in order to classify EEG signals from two data sets representing two mental states, i.e., cognitive workload and emotions. As a conclusion to this work, we offered a guideline concerning the machine learning algorithms to use for classifying levels of each of these mental states in the discussion part (see chapter 4). A first limitation to this guideline is the lack of flexibility on the conditions that need to be met in order for this guideline to be fully relevant. In other words, each of these data sets have been obtained through protocol designs that applied only a single conditions for each of the parameters - e.g, type of material to induce the states, number of electrodes, number of sessions, etc. It is thus difficult to fully generalize our results at this time. Moreover, our work aiming at reviewing the literature of studies that focused on the estimation of cognitive workload on the one hand, and the affective states on the other one, showed that multiple parameters and conditions have been tested so far, using many data sets (see chapter 3). To summarize, our guideline providing suggestions on which machine learning algorithm to use, and under which condition, is based on two studies only, and should thus be considered accordingly.

On more global point of view on this ***contribution #1*** and the guidelines we made out of it, there are some limitations regarding the longer-term objectives of the thesis. Indeed, the goal of this PhD thesis is to go towards measuring such cognitive workload and emotions states in real-time, when users are performing a task, e.g., MT-BCI, in order to adapt the task to users' mental states, here levels of cognitive workload or emotions. This ultimate goal would suggest machine learning algorithms would be run both in real-time and in a "real-world" condition, when our guidelines are based on data sets that have been collected in laboratory set ups conditions, and analysed with an offline calibration. Finally, these guidelines are made for oscillatory activity-based estimations of both cognitive workload and emotions levels, but did not look at ERP activity.

Concerning the ***contribution #2***, we developed BioPyC, a free, open-source and easy-to-use Python platform for offline EEG and biosignal processing and classification. Based on an intuitive and well-guided graphical interface, four main modules allow the user to follow the standard steps of the BCI process without any programming skill. However, so far, signal processing and machine learning algorithms that have been included in the toolbox are only made for oscillatory EEG signals-based studies: no algorithm allow users to test promising machine learning approaches on their ERP signals-based data sets yet. Moreover, in the current github status, BioPyC does not propose tools for classifying physiological signals such as heart rate (HR), breathing or electrodermal activity (EDA) yet, despite the fact that we found interesting results in our ***contribution #3*** aiming at classifying states of curiosity through physiological signals.

The ***contribution #3*** aimed at estimating states of epistemic curiosity through both EEG and physiological signals. Concerning the EEG signals-based classification, results showed that long time windows, i.e., 4 and 5-seconds time windows, obtained better classification performances than short ones. This may suggest that curiosity could be a mental state that increases over time. Thus a limitation of our study may be that we did not plan longer time windows in our protocol design. It was therefore not possible to verify that better performances might have been found for longer time windows, i.e., 7-seconds or even 10 or 15-seconds time windows. However, even if results are encouraging for a first study on the topic, they remain low and more studies should be done in order to better understand this mental state, as well as for developing more robust tools in order to discriminate 2 or more states. The second limit to this study would be that we focused on a user-dependent calibration, but did

not investigate the user-independent one, however essential to go towards calibration-free systems. Concerning the physiological signals, the finding concerning the size of the time windows is the same as for the EEG signals study: we applied algorithms on short-time windows only, i.e., 16-seconds times window for each trial. Finally, as mentioned in the discussion part of the chapter 8, we did not have time, so far, to combine multiple body sensor signals all together, or one or more types of physiological signals with EEG signals.

10.4 Perspectives

10.4.1 Short-term perspective

In order to answer to a main limitation of ***contribution #1*** & ***contribution #2***, which is the lack of use of recent and promising machine learning algorithms for the classification of mental states through ERP EEG signals, a new study is currently being carried out. Indeed, our literature review in chapter 3 indicated that ERP-based classifications of mental states such as cognitive workload and emotions obtained promising results (Roy et al., 2015a). As explained in our ***contribution #1***, Deep learning methods are promising approaches to investigate for the estimation of mental states through EEG signals, and therefore for ERP EEG signals classification. We therefore started to study such algorithms and apply them to ERP EEG signals. We started this study by using EEGNet for the classification of ERP-based mental states, i.e., cognitive workload and curiosity[1]. Other algorithms could be studied on these two mental states as well: among them, xDAWN (Rivet et al., 2009), which is widely used for ERP spatial filtering and proved to be efficient for ERP-based classification of workload levels (**?**). When these steps will be done, it will be interesting to integrate these algorithms into BioPyC in order to be able to widely test them on mental states data sets. It would also be interesting to integrate other ERP spatial filtering methods into the future versions of BioPyC, e.g., principal component analysis (PCA) or canonical correlation approaches (CCA) (Noh and De Sa, 2013) that proved efficient for EEG classification of mental workload levels as well (Roy et al., 2015a).

[1] This work started during my three-month scientific visit to Pr Virginia De Sa at the University of California, San Diego, this year.

There are ongoing works on BioPyC in addition to the ones that have been presented right before, focusing on the integration of methods for the classification of bio signals on the one hand, and the integration of hybrid BCI methods for the classification of the combination of EEG signals-based and bio signals-based features on the other

hand. As we saw in our contribution on curiosity, and particularly in the chapter 8 aiming at classifying states of epistemic curiosity through physiological signals, we developed tools for extracting features from body signals such as heart rate (HR), breathing and electrodermal activity (EDA). It would therefore be interesting to integrate such code into BioPyC, in order to enable BioPyC users to make analysis based on physiological signals only, or analysis based on both physiological and EEG signals. BioPyC could also propose a data visualization tool for presenting the frequency bands that have been used in the case of filter bank-based methods for EEG signals, as well as features that have been used by the classifiers in the case of HR, breathing or EDA signals-based studies.

Future work should be done based on our results from our ***contribution #3***, aiming at estimating states of epistemic curiosity trough both EEG and physiological signals. First, concerning the protocol desgin, while trivia questions were used as a trigger of curiosity, new tools (e.g., social robots (Ceha et al., 2019), video games) or stimuli (e.g., videos of magic tricks) could be used in future experiments. Then, longer time windows should be studied for both EEG signals-based classification (e.g., 8 or 10-seconds time windows) and physiological signals classification (e.g., 30-seconds or 1-minute time windows) since epistemic curiosity could be a mental state that increases over time. Still based on our results illustrating the classification of users' curiosity states through EEG signals on the one hand, and through physiological signals on the other hand, promising results were highlighted: 1) the classification performance through EEG signals (FBTSC+LDA, 63%) and 2) the classification performance through physiological signals (breathing+LDA, 58.4%). Combining both EEG signals and breathing signals in order to potentially obtain better performances could thus be a good direction to take. The combination of multiple types of physiological signals all together, as well as EEG signals with multiple types of physiological signals, would be an interesting study to make as well. To finish with the neurophysiological signals classification, concerning both types of signals, i.e., brain and physiological, studies should be run with user-independent calibration, in order to go towards calibration-free systems. Deeper neurophysiological analysis to identify the EEG sensors and sources mostly modulated by curiosity levels could be done as well. For example, so far, results suggested that most of the information for discriminating curiosity levels are found in theta, alpha and low beta. These frequency bands are similar to the ones used to estimate levels of workload and levels of engagement (Dehais et al., 2020). This is interesting given that both states share some common characteristics with curiosity, such as implications in

long-term memory improvements. Their similarities and differences would thus need to be further studied. Further analyses can also be done to compare our results against those obtained with fMRI (Gruber et al., 2014; Kang et al., 2009), in order to gain a better understanding of the neurological markers underlying curiosity states.

10.4.2 *Long-term perspectives*

As explained in the introduction of this chapter, and more generally in the introduction of this PhD thesis, being able to estimate learning-related mental states such as cognitive workload, emotions or curiosity, from EEG signals during a MI-BCI task, would be beneficial to the future of active BCI training as it would allow us to adapt the training task to users' states, e.g., to users' cognitive, affective or conative states. In this thesis, we focused the neurophysiological signal-based estimation of three mental states, i.e., cognitive workload, emotions and curiosity. This PhD thesis, with the literature review about the estimation of both cognitive workload levels (see chapter 3), as well as the study aiming at evaluating modern and promising machine learning algorithms in order to classify cognitive workload levels through EEG signals (see chapter 4), brought us to the last chapter, aiming at estimating cognitive workload during MI-BCI training. If no analysis has been done in this contribution so far, results should be an interesting contribution in order to evaluate the feasibility of a system aiming at estimating users' cognitive workload levels during a MI-BCI task. If this experiment proves to be a success, it would be a major step to go towards real-time adaptation of active BCI training to users' cognitive states. Moreover, this could also encourage more research work to attempt estimate other learning-related mental states such as affective or conative states, in order to adapt learning systems to the mood or motivation of the users.

Altogether, we hope that this thesis contributed new tools (machine learning algorithms, protocols and software) and knowledge to estimate learning-mental states in EEG and physiological signals. We also hope they open promising future research direction, notably about the study of curiosity.

References

Abadi, K. M., Kia, S. M., Subramanian, R., Avesani, P., and Sebe, N. (2013). User-centric affective video tagging from meg and peripheral physiological responses. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 582–587.

Abdulkader, S., Atia, A., and Mostafa, M.-S. (2015). Brain computer interfacing: Applications and challenges. *Egyptian Informatics Journal*, 16:213–230.

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., and Gabrieli, J. D. (2006). Reward-Motivated Learning: Mesolimbic Activation Precedes Memory Formation. *Neuron*, 50(3):507–517.

Aftanas, L., Varlamov, A., Pavlov, S., Makhnev, V., and Reva, N. (2002). Time-dependent cortical asymmetries induced by emotional arousal: Eeg analysis of event-related synchronization and desynchronization in individually defined frequency bands. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 44:67–82.

Allison, B. and Neuper, C. (2010). *Could anyone use a BCI?*, pages 35–54.

Andreassi, J. L. (2007). *Psychophysiology: Human behavior & physiological response, 5th ed.* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Ang, K., Chin, Z., Wang, C., Guan, C., and Zhang, H. (2012). Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*.

Ang, K. and Guan, C. (2013). Brain-computer interface in stroke rehabilitation. *Journal of Computer Science and Engineering*, 7:139–146.

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, 45:966–87.

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review*, 22(4):425–438.

Appriou, A., Cichocki, A., and Lotte, F. (2020). Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals. *IEEE SMC Magazine*, pages 1–8.

Ayaz, H. and Dehais, F. (2018). Neuroergonomics. page iv.

Azarnoosh, M. and Tabatabaee, S. (2008). Teacher: A key to motivating language learners.

Baker, R., D'Mello, S., Rodrigo, M., and Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive affective states during interactions with three different computer-based learning environments. *Int. J. Hum.-Comput. Stud.*, 68:223–241.

Baldwin, C. L. and Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage*, 59(1):48–56.

Bao, F. S., Liu, X., and Zhang, C. (2011). PyEEG: An open source python module for EEG/MEG feature extraction. *Computational Intelligence and Neuroscience*, 2011.

Barachant, A. and Bonnet, S. (2011). Channel selection procedure using riemannian distance for bci applications. pages 348 – 351.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2010). Riemannian geometry applied to BCI classification . *LVA/ICA*, pages 629–636.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012a). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928.

Barachant, A., Congedo, M., and Jutten, C. (2012b). Multiclass Brain-Computer Interface Classification by Riemannian Geometry To cite this version : Multi-class Brain Computer Interface Classification by Riemannian Geometry. 59(4):920–928.

Barachant, A. and King, J.-R. (2015). pyriemann 0.2.2.

Bashivan, P., Yeasin, M., and Bidelman, G. M. (2016). Single trial prediction of normal and excessive cognitive load through EEG feature fusion. *2015 IEEE Signal Processing in Medicine and Biology Symposium - Proceedings*.

Berka, C., Levendowski, D., Lumicao, M., Yau, A., Davis, G., Zivkovic, T., Olmstead, R., Tremoulet, P., and Craven, P. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78:B231–44.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of*

Psychology.

Berlyne, D. E. (1967). Arousal and reinforcement. *Nebraska Symposium on Motivation*, 15:1–110.

Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K. R., Curio, G., and Dickhaus, T. (2010). Neurophysiological predictor of SMR-based BCI performance. *NeuroImage*, 51(4):1303–1309.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K.-R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Sig Proc Magazine*.

Bloom, R. S. (1975). Stating Educational Objectives in Behavioral Terms. *Nursing Forum*, 14(1):30–42.

Blumenfeld, P., Rogat, T., and Krajcik, J. (2006). Motivation and cognitive engagement in learning environments. *The Cambridge Handbook of the Learning Sciences*.

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., Caltagirone, C., Kong, W., Wei, D., Zhou, Z., Polidori, L., Vitiello, S., and Babiloni, F. (2012). Assessment of mental fatigue during car driving by using high resolution eeg activity and neurophysiologic indices. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2012:6442–5.

Bradley, M. and Lang, P. (2017). International affective picture system. pages 1–4.

Braithwaite, J. J., Derrick, D., Watson, G., Jones, R., Rowe, M., Watson, D., and Robert, J. (2013). A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. . . . , pages 1–42.

Brod, G. and Breitwieser, J. (2019). Lighting the wick in the candle of learning: generating a prediction stimulates curiosity. *Science of Learning*.

Bronson, M. (2000). "self-regulation in early childhood: Nature and nurture," by martha b. bronson. book review. *Early Childhood Research Quarterly*.

Brouwer, A.-M., Hogervorst, M. A., van Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using {EEG} spectral power and {ERPs} in the n-back task. *Journal of Neural Engineering*, 9(4):45008.

Brunner, C., Andreoni, G., Bianchi, L., Blankertz, B., Breitwieser, C., Kanoh, S., Susila, I. P., and Venthur, B. (2013). BCI Software Platforms. *Towards Practical Brain-Computer Interfaces*, pages 303–331.

Brunner, C., Leeb, R., and Muller-Putz, G. (2008). Bci competition iv

dataset 2a: 4-class motor imagery. graz university of technology.

Cain, B. (2007). A review of the mental workload literature. *English*, page 35.

Calkins, M. (1913). Psychology and the behaviorist. *Psychological Bulletin*, 10:288–291.

Candra, H., Yuwono, M., Chai, R., Nguyen, H. T., and Su, S. (2017). EEG emotion recognition using reduced channel wavelet entropy and average wavelet coefficient features with normal Mutual Information method. *Proc. IEEE EMBC*, pages 463–466.

Catania, A. (2003). B. f. skinner's science and human behavior: Its antecedents and its consequences. *Journal of the experimental analysis of behavior*, 80:313–20.

Cegarra, J. and Morgado, N. (2009). Étude des propriétés de la version francophone du nasa-tlx. *EPIQUE 2009: 5ème Colloque de Psychologie Ergonomique*, pages 233–239.

Ceha, J., Chhibber, N., Goh, J., McDonald, C., Oudeyer, P.-Y., Kulić, D., and Law, E. (2019). Expression of curiosity in social robots: Design, perception, and effects on behaviour. *CHI*.

Chanel, G., Rebetez, C., Bétrancourt, M., and Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41:1052 – 1063.

Chanel, G.and Kronegg, J., Grandjean, D., and Pun, T. (2006). Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Multimedia Content Representation, Classification and Security*, 4105:530–537.

Chaouachi, M., Jraidi, I., and Frasson, C. (2011). Modeling mental workload using EEG features for intelligent systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6787 LNCS:50–61.

Choi, M. K., Lee, S. M., Ha, J. S., and Seong, P. H. (2018). Development of an EEG-based workload measurement method in nuclear power plants. *Annals of Nuclear Energy*, 111:595–607.

Choppin, A. (2000). Eeg-based human interface for disabled individuals: Emotion expression with neural networks. *Information Processing, Tokyo Institute of Technology, Yokohama, Japan*, PP:1–13.

Clark, R. (2004). The classical origins of pavlov's conditioning. *Integrative physiological and behavioral science : the official journal of the Pavlovian Society*, 39:279–94.

Clerc, M., Bougrain, L., and Lotte, F. (2016). *Brain-Computer Interfaces 1*. Wiley-ISTE.

Clerico, A., Tiwari, A., Gupta, R., Jayaraman, S., and Falk, T. H. (2018). Electroencephalography amplitude modulation analysis for au-

tomated affective tagging of music video clips. *Front. Comp. Neuro.*, 11.

Clisson, P., Bertrand-Lalo, R., Congedo, M., Victor-Thomas, G., and Chatel-Goldman, J. (2019). Timeflux: an Open-Source Framework for the Acquisition and Near Real-Time Processing of Signal Streams.

Coan, J. and Allen, J. (2004). Frontal eeg asymmetry as a moderator and mediator of emotion. *Biological psychology*, 67:7–49.

Cole, H. W. and Ray, W. J. (1985). EEG correlates of emotional tasks related to attentional demands. *International Journal of Psychophysiology*, 3(1):33–41.

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neur. Meth.*

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

Cooper, G. and Harper, R. (1969). The use of pilot ratings in evaluation of aircraft handling qualities. *NASA Ames Technical Report*.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M., Ravaja, N., and Jacucci, G. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction. *Foundations and Trends® in Human-Computer Interaction*, 9:151–308.

Craik, F. (2014). Effects of distraction on memory and cognition: A commentary. *Frontiers in psychology*, 5:841.

Cresswell, J., Wagoner, B., and Hayes, A. (2017). Rediscovering james' principles of psychology. *New Ideas in Psychology*, 46.

Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience.

Curran, E. and Stokes, M. (2003). Learning to control brain activity: A review of the production and control of eeg components for driving brain-computer interface (bci) systems. *Brain and cognition*, 51:326–36.

Daly, I., Williams, D., Kirke, A., Weaver, J., Malik, A., Hwang, F., Miranda, E., and Nasuto, S. J. (2016). Affective brain-computer music interfacing. *Journal of Neural Engineering*, 13(4):1–14.

Damasio, A. R. (1994). Descartes{\textquoteright} Error: Emotion, Reason and the Human Brain. *BMJ*, 310(6988):1213.

Das, D., Chatterjee, D., and Sinha, A. (2013). Unsupervised approach for measurement of cognitive load using EEG signals. *13th IEEE International Conference on BioInformatics and BioEngineering, IEEE BIBE 2013*, pages 13–18.

Day, H. (1970). *The Measurement of Specific Curiosity*. Ontario Institute for Studies in Education.

de Geus, E. J., Gianaros, P. J., Brindle, R. C., Jennings, J. R., and Berntson, G. G. (2019). Should heart rate variability be "corrected" for heart rate? Biological, quantitative, and interpretive considerations. *Psychophysiology*, 56(2):1–26.

Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M., and Abbass, H. A. (2019). Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review. *IEEE Transactions on Cybernetics*, pages 1–14.

Dehais, F., Lafont, A., Roy, R., and Fairclough, S. (2020). A Neuroergonomics Approach to Mental Workload, Engagement and Human Performance. *Frontiers in Neuroscience*.

Dignam, J. D., Martin, P. L., Shastry, B. S., and Roeder, R. G. (1983). Eukaryotic gene transcription with purified components. *Methods in Enzymology*, 101(C):582–598.

Dijksterhuis, C., de Waard, D., Brookhuis, K. A., Mulder, B. L., and de Jong, R. (2013). Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. *Frontiers in Neuroscience*, 7(7 AUG):1–11.

Dirican, A. C. and Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in Human Computer Interaction. *Procedia Computer Science*, 3:1361–1367.

D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A. (2013). Confusion can be beneficial for learning. *Learning and Instruction*, in press:X.

Duraisingam, A., Palaniappan, R., and Andrews, S. (2017). Cognitive task difficulty analysis using EEG and data mining. *2017 Conference on Emerging Devices and Smart Systems, ICEDSS 2017*, (March):52–57.

Durantin, G., Gagnon, J. F., Tremblay, S., and Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research*, 259:16–23.

D'Mello, S., Jackson, G., Craig, S., Morgan, B., Chip-Man, P., White, H., Person, N., Kort, B., Kaliouby, R., Picard, R., and Graesser, A. (2008). Autotutor detects and responds to learners affective and cognitive states. *Workshop on Emotional and Cognitive Issues at the Int. Conf. Intelligent Tutoring Systems. Montreal, Canada*.

Ekman, P. (1992). Are there basic emotions? *Psych Rev, 99(3)*.

Ericsson, K. and Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, 49:725–747.

Erp, J., Lotte, F., and Tangermann, M. (2012). Brain-computer interfaces: Beyond medical applications. *Computer*, 45:26–34.

Fairclough, S. (2008). BCI and physiological computing for computer games: Differences, similarities & intuitive control. In *Proc ACM*

CHI.

Fairclough, S. (2009). Fundamentals of physiological computing. *Interacting with Computers*, 21:133–145.

Fairclough, S. and Houston, K. (2004). A metabolic measure of mental effort. *Biological psychology*, 66:177–90.

Faltas, I. (2016). Emotional intelligence: A historical overview.

Farwell, L. (1988). Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials. *Electroencephalography Clinical. Neurophysiolgy*, 70:510–523.

Fernandez Rojas, R., Debie, E., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M., and Abbass, H. (2020). Electroencephalographic Workload Indicators During Teleoperation of an Unmanned Aerial Vehicle Shepherding a Swarm of Unmanned Ground Vehicles in Contested Environments. *Frontiers in Neuroscience*, 14(February):1–15.

Francis, A. L. (2010). Improved segregation of simultaneous talkers differentially affects perceptual and cognitive capacity demands for recognizing speech in competing speech. *Attention, Perception, & Psychophysics*, 72(2):501–516.

Frantzidis, C., Bratsas, C., Papadelis, C., Konstantinidis, E., Pappas, C., and Bamidis, P. (2010). Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Transactions on Information Technology in Biomedicine*, 14:589–597.

Freeman, S., Eddy, S. L., McDonough, M., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Academy of Sciences*, 111.

Frey, J., Daniel, M., Hachet, M., Castet, J., and Lotte, F. (2016). Framework for electroencephalography-based evaluation of user experience. In *Proc ACM CHI*.

Frey, J., Mühl, C., Lotte, F., and Hachet, M. (2014). Review of the use of electroencephalography as an evaluation method for human-computer interaction.

Fritz, T., Begel, A., Müller, S., Yigit-Elliott, S., and Züger, M. (2014). Using psycho-physiological measures to assess task difficulty in software development. *Proceedings of the 36th International Conference on Software Engineering. ACM*.

Frosh, S. (2012). *A brief introduction to psychoanalytic theory*.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., and Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PloS one*, 10(3).

Gawron, V. J., French, J., and Funke, D. (2001). An overview of fatigue. pages 581–595.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience*, 8(DEC).

Gervais, R., Frey, J., Gay, A., Lotte, F., and Hachet, M. (2016). Tobe: Tangible out-of-body experience. In *Proc TEI'16*, pages 227–235.

Gevins, A. and Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human–computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1-2):113–131.

Gevins, A., Smith, M. E., Leong, H., Mcevoy, L., Whitfield, S., Du, R., Rush, G., and Technology, S. A. M. (1998). Gevins1998-1. 40(1):79–91.

Gevins, A., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7(4):374–385.

Gomes, R., Vanderlei, L., Garner, D., Vanderlei, F., and Valenti, V. (2017). Higuchi fractal analysis of heart rate variability is sensitive during recovery from exercise in physically active men. *Medical Express*, 4.

Gonzalez, C., Best, B., Healy, A., Kole, J., and Bourne, L. (2011). A cognitive modeling account of simultaneous learning and fatigue effects. *Cognitive Systems Research*, 12:19–32.

Gordon, G., Breazeal, C., and Engel, S. (2015). Can children catch curiosity from a social robot? *Proc. IEEE ICHRI*, pages 91–98.

Goshvarpour, A. and Goshvarpour, A. (2012). Classification of heart rate signals during meditation using lyapunov exponents and entropy. *International Journal of Intelligent Systems and Applications*, 2:35–41.

Gottlieb, J. and Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758–770.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(7 DEC):1–13.

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., and Rao, R. P. N. (2008). Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. pages 835–844.

Gruber, M. J., Gelman, B. D., and Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. *Neuron*, 84(2):486–496.

Gruber, M. J. and Valji, A. (2019). 16 Curiosity and Learning.

Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. *UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing*, pages 301–310.

Hancock, P. and Chignell, M. (1986). Towards a thoery of mental workload: stress and adaptability in human-machine systems. *[No source information available]*.

Harmon-Jones, E. (2003). Early career award. clarifying the emotive functions of asymmetrical frontal cortical activity. *Psychophysiology*, 40:838–48.

Hart, S. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50.

Heger, D.and Putze, F. and Schultz, T. (2010). Online workload recognition from EEG data during cognitive tests and human-machine interaction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6359 LNAI:410–417.

Hektner, J. M. and Csikszentmihalyi, M. (1996). A Longitudinal Exploration of Flow and Intrinsic Motivation in Adolescents. *American Educational Research Association*, page 31.

Hewett, T., Baecker, R., Card, S., Gasen, J., Tremaine, M., Perlman, G., Strong, G., and Verplank, W. (1992). Acm sigchi curricula for human-computer interaction. *WWW document*.

Hidalgo-Muñoz, A., López, M., Pereira, A., Santos, I., and Tomé, A. (2013). Spectral turbulence measuring as feature extraction method from eeg on affective computing. *Biomedical Signal Processing and Control*, 8:945–950.

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, 16(2):107–117.

Hoffmann, U., Vesin, J.-M., and Ebrahimi, T. (2006). Spatial filters for the classification of event-related potentials. *European Symposium on Artificial Neural Networks (ESANN 2006)*, pages 47–52.

Hogervorst, M. A., Brouwer, A. M., and van Erp, J. B. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience*, 8(OCT):1–14.

Honal, M. and Schultz, T. (2008). Determine task demand from brain activity. *BIOSIGNALS 2008 - Proceedings of the 1st International Conference on Bio-inspired Systems and Signal Processing*, 1:100–107.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):99–104.

ISO (1998). *ISO/IEC 14882:1998: Programming languages — C++*.

Jacob-Dazarola, R., Ortiz Nicolás, J., and Cardenas, L. (2016). Behavioral measures of emotion (doi: 10.1016/b978-0-08-100508-8.00005-9). pages 101–124.

Jaiswal, D., Chowdhury, A., Banerjee, T., and Chatterjee, D. (2019). Effect of mental workload on breathing pattern and heart rate for a working memory task: A pilot study. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2019:2202–2206.

Jayaram, V. and Barachant, A. (2018). MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6).

Jensen, O., Kaiser, J., and Lachaux, J.-P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in neurosciences*, 30:317–24.

Jeunet, C., Kaoua, B., and Lotte, F. (2017). Towards a cognitive model of MT-BCI user training.

Jeunet, C., N'Kaoua, B., and Lotte, F. (2016). *Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates*, volume 228.

Jiang, Y. and Chun, M. (2001). Selective attention modulates implicit learning. *The Quarterly Journal of Experimental Psychology A*, 54.

Jones, E., Oliphant, T., and Peterson, P. (2001–today). A guide to NumPy.

Kamzanova, A., Kustubayeva, A., and Matthews, G. (2014). Diagnostic monitoring of vigilance decrement using eeg workload indices. *Human factors*, 56:1136–49.

Kang, M., Hsu, M., Krajbich, I., Loewenstein, G., McClure, S. M., Wang, J., and Camerer, C. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psych. sci.*

Kappenman, E. and Luck, S. (2012). *The Oxford Handbook of Event-Related Potential Components*.

Khalili, Z. and Moradi, M. H. (2009). Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of eeg. pages 1571–1575.

Kidd, C. and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.

Klimesch, W., Freunberger, R., Sauseng, P., and Gruber, W. (2008). A short review of slow phase synchronization and memory: Evidence for control processes in different memory systems? *Brain research*, 1235:31–44.

Klimesch, W., Schack, B., and Sauseng, P. (2005). The functional significance of theta and upper alpha oscillations. *Experimental Psychology*, 52(2):99–108.

Knudsen, E. I. (2007). Fundamental components of attention. *Annu. Rev. Neurosci.*, 30:57–78.

Knyazev, G. (2012). Eeg delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience and biobehavioral reviews*, 36:677–95.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). DEAP: A database for emotion analysis Using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2011). Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3:18–31.

Kort, B., Reilly, R., and Picard, R. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. *Proceedings IEEE International Conference on Advanced Learning Technologies 2001*, pages 43–46.

Kothe, C., Makeig, S., and Onton, J. (2013). Emotion recognition from eeg during self-paced emotional imagery. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 855–858.

Kothe, C. A.and Makeig, S. (2013). BCILAB: A platform for brain-computer interface development. *Journal of Neural Engineering*, 10(5).

Kreutzer, J. S., DeLuca, J., and Caplan, B. (2011a). Backward Digit Task. page 338.

Kreutzer, J. S., DeLuca, J., and Caplan, B. (2011b). Digit Span Test. page 849.

Krol, L., Freytag, S.-C., Fleck, M., Gramann, K., and Zander, T. (2016). A task-independent workload classifier for neuroadaptive technology: Preliminary data.

Law, E., Yin, M., Goh, J., Chen, K., Terry, M. A., and Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. *CHI*.

Le Ny, J.-F. (1994). Motivation. *Dictionnaire encyclopédique de l'éducation et de la formation*.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lecuyer, A., Lotte, F., Reilly, R., Leeb, R., Hirose, M., and Slater, M. (2008). Brain-computer interfaces, virtual reality, and videogames. *Computer*, 41:66–72.

Lemos, J. (2011). Wanting, willing, trying and kane's theory of free will. *Dialectica*, 65:31 – 48.

Lespiau, F. and Tricot, A. (2018). Primary knowledge enhances per-

formance and motivation in reasoning. *Learning and Instruction*, 56.

Lima, G. (2019). Curiosity, frontal EEG asymmetry , and learning. *41st Meeting of the CogSci Society*, pages 2161–2165.

Lin, W., Li, C., and Sun, S. (2017). *Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal*, pages 385–394.

Lin, Y.-P., Wang, C.-H., Wu, T.-L., Jeng, S.-K., and Chen, J. (2009). Eeg-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. pages 489–492.

Litman, J. A. (2012). Encyclopedia of the Sciences of Learning. *Encyclopedia of the Sciences of Learning*, (December).

Litman, J. A., Hutchins, T. L., and Russon, R. K. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition and Emotion*, 19(4):559–582.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*.

Loft, S., Bowden, V., Braithwaite, J., Morrell, D., Huf, S., and Durso, F. (2014). Situation awareness measures for simulated submarine track management. *Human Factors*.

Lohani, M., Payne, B. R., and Strayer, D. L. (2019). A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in Human Neuroscience*, 13(March):1–27.

Lomas, J. D., Koedinger, K., Patel, N., Shodhan, S., Poonwala, N., and Forlizzi, J. L. (2017). Is difficulty overrated? the effects of choice, novelty and suspense on intrinsic motivation in educational games. *CHI*.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018a). A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *Journal of Neural Engineering*, pages 0–20.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018b). A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces: A 10-year Update. *JNE*.

Lotte, F., Faller, J., Guger, C., Renard, Y., Pfurtscheller, G., Lécuyer, A., and Leeb, R. (2012). *Combining BCI with Virtual Reality: Towards New Applications and Improved BCI*, pages 197–220.

Lotte, F., Florian, L., and Mühl, C. (2013). Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. *Frontiers in human neuroscience*, 7:568.

Macedo, L. and Cardoso, A. (2012). The exploration of unknown environments populated with entities by a surprise-curiosity-based

agent. *Cognitive Systems Research*, 19-20:62–87.

Malik, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93:1043–1065.

Mathan, S., Smart, A., Ververs, T., and Feuerstein, M. (2010). Towards an index of cognitive efficacy: EEG-based estimation of cognitive load among individuals experiencing cancerrelated cognitive decline. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pages 6595–6598.

MATLAB (2010). *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

Mazher, M., Aziz, A., Malik, A. S., and Qayyum, A. (2016). A comparison of brain regions based on EEG during multimedia learning cognitive activity. *ISSBES 2015 - IEEE Student Symposium in Biomedical Engineering and Sciences: By the Student for the Student*, pages 31–35.

McFarland, D. J. and Wolpaw, J. R. (2018). Brain–computer interface use is a skill that user and system acquire together. *PLOS Biology*, 16(7):1–4.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56.

Mertin, V. (2012). Lightner witmer - origins of practical psychology. *Ceskoslovenska Psychologie*, 56:85–93.

Militello, L., Gentner, F., Swindler, S., and Beisner, G. (2006). Conation: Its historical roots and implications for future research. volume 2006, pages 240 – 247.

Millán, J. R., Rupp, R., Müller-Putz, G., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., Neuper, C., Müller, K.-R., and Mattia, D. (2010). Combining Brain-Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges. *Frontiers in Neuroprosthetics*.

Mischel, W. (1996). From good intentions to willpower. pages 197–218.

Muhl, C., Allison, B., Nijholt, A., and Chanel, G. (2015). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*.

Mühl, C., Jeunet, C., and Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Frontiers in Neuroscience*, 8(8 JUN).

Müller, M. M., Keil, A., Gruber, T., and Elbert, T. (1999). Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clinical Neurophysiology*, 110(11):1913–1920.

Müller-Putz, G. R., Breitwieser, C., Cincotti, F., Leeb, R., Schreuder,

M., Leotta, F., Tavella, M., Bianchi, L., Kreilinger, A., Ramsay, A., Rohm, M., Sagebaum, M., Tonin, L., Neuper, C., and Millán, J. R. (2011). Tools for brain-computer interaction: A general concept for a hybrid BCI. *Frontiers in Neuroinformatics*, 5(November):1–10.

Müller-Putz, G. R., Scherer, R., Brunner, C., Leeb, R., and Pfurtscheller, G. (2008). Better than random? A closer look on BCI results. *International Journal of Bioelectromagnetism*, 10(1):52–55.

Mussel, P. (2010). Epistemic curiosity and related constructs: Lacking evidence of discriminant validity. *Personality and Individual Differences*.

Myrden, A. and Chau, T. (2015). Effects of user mental state on EEG-BCI performance. *Frontiers in Human Neuroscience*, 9(JUNE):1–11.

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1:66–84.

Neuper, C. and Pfurtscheller, G. (2010). *Neurofeedback Training for BCI Control*, pages 65–78.

Newell, A. and Simon, H. A. (1972). *Human problem solving.* Prentice-Hall, Oxford, England.

Newen, A. (2015). What are cognitive processes? an example-based approach. *Synthese*, 194.

Noh, E. and De Sa, V. (2013). Canonical correlation approach to common spatial patterns. *International IEEE/EMBS Conference on Neural Engineering, NER*, (3):669–672.

Norman, D. A. (1980). Twelve Issues for Cognitive Science. *Cognitive Science*, 4(1):1–32.

O'Brien, T. G. and Meister, D. (2002). Human factors testing and evaluation: An historical perspective. pages 5–20.

Onton, J. and Makeig, S. (2009). High-frequency broadband modulations of electroencephalographic spectra. *Frontiers in human neuroscience*, 3:61.

Oudeyer, P.-Y., Gottlieb, J., and Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in Brain Research*, 229:257–284.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comp.*

Paas, F., Renkl, A., and Sweller, J. (2010). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38:1–4.

Paas, F., Tuovinen, J., Tabbers, H., and Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist - EDUC PSYCHOL*, 38:63–71.

Paas, F. and Van Merrienboer, J. G. (1994). Variability of worked

examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, 86:122–133.

Parasuraman, R. and Wilson, G. (2008). Putting the brain to work: Neuroergonomics past, present, and future. *Human Factors*, 50(3):468–474.

Parent, M. (2019). Diagnosticité des mesures physiologiques périphériques de la charge mentale. *PhD thesis*, 5:82–7.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Peck, E., Yuksel, B., Ottley, A., Jacob, R., and Chang, R. (2013). Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proc ACM CHI*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., and Dubourg, V. (2011a). Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, C.-K., Havlin, S., Stanley, H., and Goldberger, A. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos (Woodbury, N.Y.)*, 5:82–7.

Peng, H., Long, F., and Ding, C. (2005a). Feature Selection Based On Mutual Information: Criteria of Max-Dependency,Max-Relevance, and Min-Redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27.

Peng, H., Long, F., and Ding, C. (2005b). Feature Selection Based On Mutual Information: Criteria of Max-Dependency,Max-Relevance, and Min-Redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27.

Perego, P., Maggi, L., Parini, S., and Andreoni, G. (2009). BCI++: A new framework for brain computer interface application. *18th International Conference on Software Engineering and Data Engineering 2009, SEDE 2009*, pages 37–41.

Pérez, F. and Granger, B. E. (2007). {IP}ython: a System for Interactive Scientific Computing. \url{http://ipython.org} Accessed: May 10, 2014. *Computing in Science and Engineering*, 9(3):21–29.

Petrantonakis, P. and Hadjileontiadis, L. (2010). Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *Affective Computing, IEEE Transactions on*, 1:81–97.

Petrosian, A. (1995). Kolmogorov complexity of finite sequences and recognition of different preictal eeg patterns. *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*, pages 212 – 217.

Pfurtscheller, G., Allison, B., Brunner, C., Bauernfeind, G., Solis-Escalante, T., Scherer, R., Zander, T., Müller-Putz, G., Neuper, C., and Birbaumer, N. (2010). The hybrid bci. *Frontiers in neuroscience*, 4:30.

Pfurtscheller, G., Müller-Putz, G., Scherer, R., and Neuper, C. (2008). Rehabilitation with brain-computer interface systems. *Computer*, 41:58 – 65.

Pfurtscheller, G., Neuper, C., and Birbaumer, N. (2004). 4 human brain–computer interface.

Pillette, L. (2019). *Redefining and Adapting Feedback for Mental-Imagery based Brain-Computer Interface User Training to the Learners' Traits and States*. PhD thesis, Université de Bordeaux.

Pillette, L., Appriou, A., Cichocki, A., N'Kaoua, B., and Lotte, F. (2018). Classification of attention types in EEG signals. In *International BCI Meeting*.

Posada-Quintero, H., Florian, J., Orjuela-Cañón, A., Corrales, T., Charleston-Villalobos, S., and Chon, K. (2016). Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of Biomedical Engineering*, 44.

Redick, T., Calvo, A., Gay, C., and Engle, R. (2011). Working memory capacity and go/no-go task performance: Selective effects of updating, maintenance, and inhibition. *Journal of experimental psychology. Learning, memory, and cognition*, 37:308–24.

Reid, G. B. and Nygren, T. E. (1988). Human Mental Workload. *Advances in Psychology*, 52:185–218.

Renard, Y., Lotte, F., Gibert, G., Congedo, M., and Lécuyer, A. (2010a). OpenViBE: An open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. *Presence*.

Renard, Y., Lotte, F., Gibert, G., Congedo, M., Maby, E., Delannoy, V., Bertrand, O., and Lécuyer, A. (2010b). OpenViBE: An open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. *Presence: Teleoperators and Virtual Environments*, 19(1):35–53.

Richardson, J. and Newby, T. (2006). The role of students' cognitive engagement in online learning. *The American Journal of Distance Education*, 20:23–37.

Rieber, R. and Robinson, D. (2001). *Wilhelm Wundt in History: The Making of a Scientific Psychology*.

Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009).

xdawn algorithm to enhance evoked potentials: Application to brain–computer interface. *Biomedical Engineering, IEEE Transactions on*, 56:2035 – 2043.

Roediger, H. (1985). Remembering ebbinghaus. *PsycCRITIQUES*, 30:519–523.

Rossum, G. (1995). Python reference manual. Technical report, Amsterdam, The Netherlands, The Netherlands.

Rothbart, D. and Scherer, I. (1997). Kant's critique of judgment and the scientific investigation of matter. *Hyle*, 3:65–80.

Roy, R., Charbonnier, S., and Jallon, P. (2015a). A Comparison of ERP Spatial Filtering Methods for Optimal Mental Workload Estimation. pages 7254–7257.

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2013). Mental fatigue and working memory load estimation: Interaction and implications for EEG-based passive BCI. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 6607–6610.

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2015b). Enhancing single-trial mental workload estimation through xDAWN spatial filtering. *International IEEE/EMBS Conference on Neural Engineering, NER*, 2015-July:360–363.

Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016a). Efficient mental workload estimation using task-independent EEG features. *Journal of Neural Engineering*, 13(2).

Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016b). Efficient mental workload estimation using task-independent eeg features. *Journal of neural engineering*, 13(2):026019.

Rozado, D. and Duenser, A. (2015). Combining eeg with pupillometry to improve cognitive workload detection. *Computer*, 48:18–25.

Russell, J. and Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, 76:805–19.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Sammler, D., Grigutsch, M., Fritz, T., and Koelsch, S. (2007). Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44:293–304.

Schalk, G., Mcfarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000 : A General-Purpose Brain-Computer Interface (BCI) System. 51(6):1034–1043.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain*

Mapping.

Schlö, A. (2006). GDF - A general dataformat for BIOSIGNALS. *CoRR*, abs/cs/0608052.

Schmidt, R. A. (1968). Anticipation and timing in human motor performance. *Psychological Bulletin*, 70(6p1):631.

Schmidt, S. and Walach, H. (2000). Electrodermal activity (eda) - state-of-the-art measurement and techniques for parapsychological purposes. *Journal of Parapsychology*, 64:139–163.

Schölgl, A., Vidaurre, C., and Sander, T. H. (2011). BioSig: The free and open source software library for biomedical signal processing. *Computational Intelligence and Neuroscience*, 2011.

Shaffer, . and Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in public health*, 5:258.

Shelton, C. (2000). Portraits in emotional awareness. *Educational Leadership*, 58.

Shen, K.-Q., Li, X.-P., Ong, C., Shao, S.-Y., and Wilder-Smith, E. (2008). Eeg-based mental fatigue measurement using multi-class support vector machines with confidence estimate. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 119:1524–33.

Shimomura, Y., Yoda, T., Sugiura, K., Horiguchi, A., Iwanaga, K., and Katsuura, T. (2008). Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of physiological anthropology*, 27:173–7.

Sikandar, A. (2016). John dewey and his philosophy of education. *Journal of Education and Educational Development*, 2:191.

Sinha, A., Chatterjee, D., Saha, S. K., and Basu, A. (2016). Validation of stimulus for EEG signal based cognitive load analysis. *2015 5th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2015*.

Slama, M. (2005). Emotions and life: Perspectives from psychology, biology, and evolution. *Psychology & Marketing*, 22.

Smith, A.-L., Owen, H., and Reynolds, K. (2013). Heart rate variability indices for very short-term (30 beat) analysis. part 1: survey and toolbox. *J Clin Monit Comput*, 15.

So, W. K., Wong, S. W., Mak, J. N., and Chan, R. H. (2017). An evaluation of mental workload with frontal EEG. *PLoS ONE*, 12(4):1–17.

Soares, A., Pinheiro, A., Costa, A., Frade, S., Comesaña, M., and Pureza, R. (2013). Affective auditory stimuli: Adaptation of the international affective digitized sounds (iads-2) for european portuguese. *Behavior research methods*, 45.

Soleymani, M., Pantic, M., and Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3:211–223.

Soleymani, M., Pantic, M., and Pun, T. (2015). Multimodal emotion recognition in response to videos (Extended abstract). *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 3(2):491–497.

Spielberger, C., Gorsuch, R., Lushene, R., Vagg, P., and Jacobs, G. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y1 – Y2)*, volume IV.

Stadler, M. A. (1995). Role of attention in implicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(3):674–685.

Stein, E. S. (1985). Air traffic controller workload: An examination of workload probe.

Steriade, M., McCormick, D., and Sejnowski, T. (1993). Thalamocortical oscillation in the sleeping and aroused brain. *Science (New York, N.Y.)*, 262:679–85.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153:652–654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Acta Psychologica*, 30:276–315.

Sturm, W. and Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *Neuroimage*, 14(1):S76–S84.

Suchan, B. (2008). Neuroanatomical correlates of processing in visual and visuospatial working memory. *Cognitive processing*, 9:45–51.

Sweller, J., Van Merrienboer, J. J. G., and Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10:251–.

Sweller, J., Van Merrienboer, J. J. G., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31:261–292.

Takahashi, K. (2004). Remarks on svm-based emotion recognition from multi-modal bio-potential signals. pages 95 – 100.

Tardif, J. (1992). Pour un enseignement stratégique : l'apport de la psychologie cognitive. *Montrèal : Editions Logiques*.

Tayeb, Z., Waniek, N., Fedjaev, J., Ghaboosi, N., Rychly, L., Widderich, C., Richter, C., Braun, J., Saveriano, M., Cheng, G., and Conradt, J. (2018). Gumpy: A Python toolbox suitable for hybrid brain-computer interfaces. *Journal of Neural Engineering*, 15(6).

Tieben, R., Bekker, T., and Schouten, B. (2011). Curiosity and interaction: making people curious through interactive systems. In *Proc. BCS-HCI*.

Tiwari, A., Albuquerque, I., Parent, M., Gagnon, J. F., Lafond, D., Tremblay, S., and Falk, T. H. (2019). Multi-scale heart beat entropy measures for mental workload assessment of ambulant users. *Entropy*, 21(8):1–20.

Unsworth, N., Heitz, R., Schrock, C., and Engle, R. (2005). An automated version of the operation span task. *Behavior research methods*, 37:498–505.

Valdehita, S., Ramiro, E., López-Higes, R., and García, J. (2012). Effects of task load and cognitive abilities on performance and subjective mental workload in a tracking task. *Anales de Psicologia*, 28:986–995.

Vallat, R. (2018). Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

van Beurden, M., Brouwer, A.-M., Baardewijk, J., Binsch, O., Vermetten, E., and Roijendijk, L. (2020). Towards user-adapted training paradigms: Physiological responses to physical threat during cognitive task performance. *Multimedia Tools and Applications*.

Van Gerven, P., Paas, F., Van Merrienboer, J. G., and Schmidt, H. (2004). Memory load and the cognitive pupillary response. *Psychophysiology*, 41:167–74.

Van Leeuwen, C. and Lachmann, T. (2004). Negative and positive congruence effects in letters and shapes. *Attention, Perception, & Psychophysics*, 66(6):908–925.

Venthur, B., Dähne, S., Höhne, J., Heller, H., and Blankertz, B. (2015). Wyrm: A Brain-Computer Interface Toolbox in Python. *Neuroinformatics*, 13(4):471–486.

Vernon, J. L., Amelia, J. S., Nicholas, R. W., Stephen, M. G., Chou, P. H., and Brent, J. L. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013.

Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual review of biophysics and bioengineering*, 2:157–180.

Villon, O. and Lisetti, C. (2006). A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. pages 269 – 276.

Vogel, S. (2016). Learning and memory under stress: implications for the classroom. *npj Science of Learning*, 1:16011.

Voss, A., Schroeder, R., Heitmann, A., Peters, A., and Perz, S. (2015). Short-term heart rate variability - Influence of gender and age in healthy subjects. *PLoS ONE*, 10(3):1–33.

Walter, C., Schmidt, S., Rosenstiel, W., Gerjets, P., and Bogdan, M. (2013). Using cross-task classification for classifying workload levels in complex learning tasks. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 876–881.

Wang, S., Gwizdka, J., and Chaovalitwongse, W. A. (2016). Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435.

Wechsler, D. (1955). Wechsler adult intelligence scale.

Westermann, R., Spies, K., Stahl, G., and Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, 26(4):557–580.

Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science*, 3:159–177.

Wickramasekera, I. (2014). *Early Psychological Knowledge*.

Wiebe, E., Roberts, E., and Behrend, T. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, 26:474–481.

Wobrock, D., Frey, J., Graef, D., de la Rivière, J. B., Castet, J., and Lotte, F. (2015). Continuous mental effort evaluation during 3d object manipulation tasks based on brain and physiological signals. In *Proceedings of INTERACT ?15*.

Wolpaw, J. and Wolpaw, E. (2012). *Brain-Computer Interfaces: Principles and Practice*. OUP USA.

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

Yang, Y., Krompass, D., and Tresp, V. (2017). Tensor-Train Recurrent Neural Networks for Video Classification. *CoRR*.

Yger, F., Berar, M., and Lotte, F. (2016). Riemannian approaches in Brain-Computer Interfaces: a review. *IEEE TNSRE*, 25(10).

Yuksel, B., Oleson, K., Harrison, L., Peck, E., Afergan, D., Chang, R., and Jacob, R. (2016). Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proc ACM CHI*.

Zander, T. and Kothe, C. (2011). Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *J Neur Eng*, 8.

Zarjam, P., Epps, J., and Chen, F. (2011). Spectral EEG featuresfor evaluating cognitive load. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 3841–3844.

Zarjam, P., Epps, J., Chen, F., and Lovell, N. (2013). Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in biology and medicine*, 43:2186–95.

Zarjam, P., Epps, J., and Lovell, N. H. (2015). Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload. *IEEE Transactions on Autonomous Mental Development*, 7(4):301–310.

Zhang, Q. and Lee, M. (2012). Emotion development system by interacting with human EEG and natural scene understanding. *Cognitive Systems Research*, 14(1):37–49.

Zheng, W. L. and Lu, B. L. (2015). 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.

Zhou, F., Qu, X., Jiao, J., and Helander, M. (2014). Emotion prediction from physiological signals: A comparison study between visual and auditory elicitors. *Interacting with Computers*, 26:285–302.

Zimmerman, B. J. and Risemberg, R. (1997). Self-Regulatory Dimensions of Academic Learning and Motivation. *Handbook of academic learning: Construction of knowledge*, pages 105–125.

Zomeren, A. H. and Brouwer, W. H. (1994). *Clinical neuropsychology of attention*. Oxford University Press, USA.