UAB
Universitat Autònoma de Barcelona

Deep Learning Based Data Fusion Approaches for the Assessment of Cognitive States on EEG Signals

José Elías Yauri Vidalón

Universitat Autònoma
de Barcelona

Deep Learning Based Data Fusion Approaches for the Assessment of Cognitive States on EEG Signals

A dissertation submitted by **José Elías Yauri Vidalón** to the Universitat Autònoma de Barcelona in fulfilment of the degree of **Doctor of Philosophy** in the Departament de Ciències de la Computació.

Bellaterra, March 15, 2023

Director	**Dr. Aura Hernández-Sabaté** Centre de Visió per Computador Universitat Autònoma de Barcelona
Co-Director	**Dr. Debora Gil Resina** Centre de Visió per Computador Universitat Autònoma de Barcelona
Thesis committee	**Dr. Àgata Lapedriza García** MIT Medialab eHealth Center Universitat Oberta de Catalunya **Dr. Carles Sánchez** Centre de Visió per Computador Universitat Autònoma de Barcelona **Dr. Mihail Gaianu** Department of Computer Science West University of Timisoara

This document was typeset by the author using LaTeX 2$_\varepsilon$.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

ISBN 978-84-126409-2-2

Printed by Ediciones Gráficas Rey, S.L.

To Yahweh, The Unseen God.

Acknowledgements

Research is really a long journey that implies hard study, exploration, experimentation, reasoning, discovery, and learning. On the other hand, research also entails moments of fellowship, happiness, and collaboration with extraordinary people. At this moment, it is time to make a short stop to express my gratitude to the people who gave me guidance, help, and support during these last four years.

First, I would sincerely like to thank my supervisors Aura Hernández Sabaté and Debora Gil Resina, for giving me the opportunity to arrive to Barcelona-Spain and to do research under their guidance. I appreciate their patience and motivation in times of uncertainty during this research. All the meetings and discussions were productive, without their help, this thesis would not have been finished.

Second, I have a family to thank. Specially, to my loved wife Maribel and to my firstborn Daniel Matheus for their emotional support and immense love despite the time and distance. Then I thank to my parents, Félix and Ruth (my beloved mom for ever). I will never forget my mom always encouraging me in pursuing a PhD. Also my gratitude to my brothers and sisters for being aware of me, Miguel, Débora, Rousell, and Judith.

Third, thank the co-authors and collaborators who helped me through this research. I would especially like to thank Guillermo Torres, Gemma Sánchez, and personal of the Interactive and Augmented Modelling (IAM) research group. I also thank Pau Folch and Miquel Àngel Piera from the Aslogic company.

Fourth, I express my gratitude to the Centre de Visió per Computadorfor providing me of resources and materials for my research in the doctoral program in Computer Science at the Universitat Autònoma de Barcelona. I would also thank the CVC's administrative people as Kevin Salvani, Montse Culleré, and specially to Joan Masoliver and Raquel Gómez from the Technical Support Unity. In addition, I thank my fellows, Guillermo Torres, Sanket Biswas, Kai Wang, Fei Yang, Andrés Mafla, and Germán Barquero for their friendships.

Fifth, thank the *Beca Presidente de la República - 2019, Programa Nacional de Becas y Crédito Educativo (PRONABEC), Ministerio de Educación del Perú*, for its financial support throughout this doctoral study.

Finally, last but not least, I thank the Yahweh God for his love, wisdom, and comfort over the years, far from my home and family.

Abstract

For millennia, the study of the couple brain-mind has fascinated the humanity in order to understand the complex nature of cognitive states. A cognitive state is the state of the mind at a specific time and involves cognition activities to acquire and process information for making a decision, solving a problem, or achieving a goal.

While normal cognitive states assist in the successful accomplishment of tasks; on the contrary, abnormal states of the mind can lead to task failures due to a reduced cognition capability. In this thesis, we focus on the assessment of cognitive states by means of the analysis of ElectroEncephaloGrams (EEG) signals using deep learning methods. EEG records the electrical activity of the brain using a set of electrodes placed on the scalp that output a set of spatio-temporal signals that are expected to be correlated to a specific mental process.

From the point of view of artificial intelligence, any method for the assessment of cognitive states using EEG signals as input should face several challenges. On the one hand, one should determine which is the most suitable approach for the optimal combination of the multiple signals recorded by EEG electrodes. On the other hand, one should have a protocol for the collection of good quality unambiguous annotated data, and an experimental design for the assessment of the generalization and transfer of models. In order to tackle them, first, we propose several convolutional neural architectures to perform data fusion of the signals recorded by EEG electrodes, at raw signal and feature levels. Four channel fusion methods, easy to incorporate into any neural network architecture, are proposed and assessed. Second, we present a method to create an unambiguous dataset for the prediction of cognitive mental workload using serious games and an Airbus-320 flight simulator. Third, we present a validation protocol that takes into account the levels of generalization of models based on the source and amount of test data.

Finally, the approaches for the assessment of cognitive states are applied to two use cases of high social impact: the assessment of mental workload for personalized support systems in the cockpit and the detection of epileptic seizures. The results obtained from the first use case show the feasibility of task transfer of models trained to detect workload in serious games to real flight scenarios. The results from the second use case show the generalization capability of our EEG channel fusion methods at k-fold cross-validation, patient-specific, and population levels.

Keywords – Deep learning, EEG, EEG channel fusion, Cognitive states, Mental workload, Cognitive task transfer, Dataset of mental workload, Epilepsy detection.

Resum

Durant mil·lennis, l'estudi de la parella cervell-ment ha fascinat la humanitat per entendre la naturalesa complexa dels estats cognitius. Un estat cognitiu és l'estat de la ment en un moment concret i implica activitats cognitives per adquirir i processar informació per prendre una decisió, resoldre un problema o assolir un objectiu.

Mentre que els estats cognitius normals ajuden a la realització exitosa de les tasques; per contra, els estats anormals de la ment poden conduir a fracassos de tasques a causa d'una capacitat cognitiva reduïda. En aquesta tesi ens centrem en l'avaluació dels estats cognitius mitjançant l'anàlisi de senyals d'Electroencefalogrames (EEG) mitjançant mètodes d'aprenentatge profund. L'EEG registra l'activitat elèctrica del cervell mitjançant un conjunt d'elèctrodes col·locats al cuir cabellut que produeixen un conjunt de senyals espai-temporals que s'espera que estiguin correlacionats amb un procés mental específic.

Des del punt de vista de la intel·ligència artificial, qualsevol mètode per a l'avaluació d'estats cognitius utilitzant senyals EEG com a entrada hauria d'afrontar diversos desafiaments. D'una banda, cal determinar quin és l'enfocament més adequat per a la combinació òptima dels múltiples senyals enregistrats pels elèctrodes EEG. D'altra banda, s'hauria de disposar d'un protocol per a la recollida de dades anotades sense ambigüitats de bona qualitat, i d'un disseny experimental per a l'avaluació de la generalització i transferència de models. Per abordar-los, primer, proposem diverses arquitectures neuronals convolucionals per dur a terme la fusió de dades dels senyals enregistrats pels elèctrodes EEG, a nivells de senyal i característiques en brut. Es proposen i avaluen quatre mètodes de fusió de canals, fàcils d'incorporar a qualsevol arquitectura de xarxa neuronal. En segon lloc, presentem un mètode per crear un conjunt de dades inequívoc per a la predicció de la càrrega de treball mental cognitiva mitjançant jocs seriosos i un simulador de vol Airbus-320. En tercer lloc, presentem un protocol de validació que té en compte els nivells de generalització dels models basats en l'origen i la quantitat de dades de prova.

Finalment, els enfocaments per a l'avaluació dels estats cognitius s'apliquen a dos casos d'ús d'alt impacte social: l'avaluació de la càrrega de treball mental per a sistemes de suport personalitzats a la cabina i la detecció de convulsions epilèptiques. Els resultats obtinguts del primer cas d'ús mostren la viabilitat de la transferència de tasques de models entrenats per detectar càrrega de treball en jocs seriosos a escenaris de vol reals. Els resultats del segon cas d'ús mostren la capacitat de generalització dels nostres mètodes de fusió de canals EEG a nivells de validació creuada de k-fold, es-

pecífics del pacient i de població.

Paraules Clau – Aprenentatge profund, EEG, Fusió de canals d'EEG, Estats cognitius, Sobrecàrrega de treball mental, Transferència de tasques cognitives, Base de dades de sobrecàrrega mental de treball, Detecció d'epilèpsia.

Resumen

Por milenios, el estudio del par cerebro-mente ha fascinado a la humanidad con el fin de comprender la compleja naturaleza de los estados cognitivos. Un estado cognitivo es el estado de la mente en un momento específico e involucra todas las actividades cognitivas para adquirir y procesar información con el fin de tomar una decisión, resolver un problema o lograr un objetivo.

Mientras que los estados cognitivos normales contribuyen en la realización exitosa de las tareas; al contrario, los estados mentales anómalos pueden conducir al fracaso de las tareas debido a una reducción de la capacidad cognitiva. En esta tesis nos enfocamos en la evaluación de estados cognitivos mediante el análisis de señales de electroencefalogramas (EEG) utilizando métodos de aprendizaje automático. Un EEG registra la actividad eléctrica del cerebro mediante un conjunto de electrodos colocados sobre el cuero cabelludo y emiten un conjunto de señales espacio-temporales que se espera que estén correlacionadas con un proceso mental específico.

Desde el punto de vista de la inteligencia artificial, cualquier método para la evaluación de estados cognitivos utilizando señales de un EEG debe enfrentar varios desafíos. Por un lado, se debe determinar cuál es el enfoque más adecuado para la combinación óptima de las múltiples señales registradas por los electrodos del EEG. Por otro lado, se debe tener un protocolo para la recolección de datos anotados no ambiguos de buena calidad, y un diseño experimental para la evaluación de la generalización y transferencia de los modelos. Para abordar estos problemas, primero, proponemos varias arquitecturas neuronales convolucionales para realizar fusión de datos de las señales registradas por los electrodos del EEG, a niveles de señal sin procesar y a nivel de características. Se proponen y evalúan cuatro métodos de fusión, los cuales pueden ser incorporados fácilmente en cualquier arquitectura. Segundo, presentamos un método para crear base de datos no ambiguos con fines de predicción de la sobrecarga de trabajo mental utilizando juegos serios y un simulador de vuelo Airbus-320. Tercero, presentamos un protocolo de validación que tome en cuenta los niveles de generalización de los modelos en función del origen y cantidad de datos utilizados en el conjunto de test.

Finalmente, las propuestas para la evaluación de estados cognitivos se aplican a dos casos de uso de alto impacto social: la evaluación de la sobrecarga de trabajo mental para sistemas de apoyo personalizados en cabinas de aviación y la detección de epilepsia. Los resultados obtenidos del primer caso de uso demuestran la viabilidad de la transferencia de tareas para detectar la sobrecarga mental de trabajo entre tareas

realizadas en ambientes de juegos serios hacia escenarios de vuelos reales. Los resultados del segundo caso de uso muestran la capacidad de generalización de nuestros métodos de fusión de canales de EEG a nivel de validación cruzada *k-fold*, paciente específico y población.

Palabras Clave – Aprendizaje profundo, EEG, Fusión de canales de EEG, Estados cognitivos, Sobrecarga de trabajo mental, Transferencia de tareas cognitivas, Base de datos de sobrecarga mental de trabajo, Detección de epilepsia.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

According to the World Health Organization (WHO) [141], each year, more than 1.3 million of people die in road traffic accidents around the world and almost 80% of these accidents are caused by drivers' malfunctions, resulting in economic losses of near 3% of each country's gross domestic product. Besides, pilot's errors have caused almost 57% of fatal aerial accidents in the last decade with catastrophic results [91]. In both cases, drivers and pilots were strongly affected in their action and response capability due to abnormal cognitive states [19]. Other alterations in cognitive states can be produced by neurological diseases, such as dementia, Parkinson, epilepsy, ictus, etc.

A cognitive state is the state of the mind at a given time as a result of mental processes that occur within the mind [121]. Because the rational nature of humans, at each moment, the mind performs naturally mental processes, conscious and unconsciously, even while sleeping [34, 64]. These mind processes are referred to as cognition [118, 63]. Cognition involves all mental processes and thoughts to acquire information, to gain knowledge, and to perform activities of perception, attention, memory, decision making, solving problems, and actions [96].

There is a large diversity of cognitive states and they are strongly influenced by the individual's internal psychological and physiological conditions (e.g., mood, feeling, motivation, and age) and also by external stimuli (e.g., the work environment conditions and surrounding events) [52, 86, 122]. Cognitive states such as attention, awareness, and alertness are ideally normal cognitive states to successfully achieve a task. In contrast, abnormal cognitive states, such as distraction, unawareness, and fatigue, often lead to task failures with undesired results. In particular groups of people (like pilots, drivers, or decision makers in enterprises), whose daily activities demand a cognition effort to successfully develop their tasks, the presence of abnormal cognitive states, such as mental workload, mental fatigue, drowsiness, stress, may provoke subject's mistakes with catastrophic consequences, millionaires cost, and thousand of injured and dead people. Summarizing, abnormal cognitive states strongly reduce the

human performance and diminish their capability to solve tasks, retards their time of response, blocks their physical response action, and even produces other physiological and psychological disorders [30, 21].

Nowadays, cognitive states are widely studied from different disciplines: neurology aims to understand the brain, the nervous system, and its diseases; psychiatry deals with mental disorders. And, recently, computer science and engineering aims to develop automatic systems to detect and assess abnormal cognitive states. Among the wide variety of abnormal cognitive states, the most studied ones are mental workload, fatigue, distraction, and stress [19, 147, 61]. In particular, mental WorkLoad (WL) has attracted special interest because it strongly affects the human productivity and efficiency to solve tasks [114]. Indeed, WL can also produce other mental anomalies, such a fatigue and stress after continuous longtime facing WL [58, 5].

WL refers to the mental resources required to perform a task, so it depends on cognition capabilities [114, 19]. Generally, the harder the task is, the greater the mental workload results [13]. While, a too high WL might lead to mental collapse and task failure, if the WL is low, after some time, the mind becomes distracted and boredom to work, and drowsiness arises [152, 73]. Either high or low, inadequate WL can be harmful, specially in daily activities that demand a minimal cognition to have success in a task, such as piloting an airplane, driving a car, or making decision [139, 114].

In general, the assessment of cognitive states is a challenging topic, mainly due to inter-subject variability [28, 147, 77]. Particularly for WL, the classic and still most extended approach is the use of cognitive tests to which the subject undergoes while his/her self-perceived cognitive state is registered [38, 79]. Although this self-perceived strategy provides an assessment of the cognitive states of the subject, it is very subjective and strongly dependent on the psychological state of the subject. Currently, the NASA Task Load Index (NASA-TLX) [55] is the most used questionnaire to gain insight about the perceived workload levels while a subject works with various human-machine interface systems [19]. A TLX questionnaire measures the mental workload based on a weighted average of six sub-variables: mental, physical, and temporal demand, performance, effort and frustration and it is widely used in aviation to assess mental workload of pilots while interacting with plane controls [139, 88].

An alternative to self-assessment of cognitive states is the use of physiological data recorded from the subject under study. Physiological data provides a more reliable and objective measurement rather than the psychologically-dependent self-subjective reports [28, 3]. Depending on the frequency of recorded data, physiological sensors can be categorized in two main groups: a) continuous time sensors, and b) non-continuous time sensors. The first group supports continuous monitoring of physiological responses and the data recorded consists of time-varying signals. The latter consists of complex imagery devices that provide structural and functional images of the brain and hearth [71], but a low frame rate, like the functional magnetic resonance imaging (fMRI), or the near infrared spectroscopy (NIRS) or hearth ultrasound. These medical devices are useful for medical diagnosis (like Alzheimer [66], Parkinson [6], and hearth functionality [115], respectively), but cannot be used to detect WL or other punctual

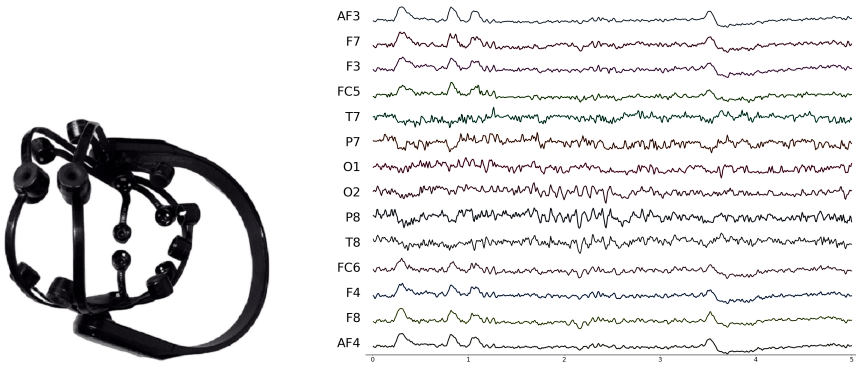
cognitive disorders (like epileptic seizures) that appear at discontinuous time intervals. Consequently, for assessment of WL and other cognitive states, usually continuous monitoring devices, portable, non-invasive, and low cost sensors are the most preferable [69].

For continuous monitoring of cognitive states, there are a broad range of physiological sensors [97], such as the electroencephalogram (EEG), the intracranial electroencephalogram (iEEG), magnetoencephalography (MEG), electrocardiogram (ECG), electrooculography (EOG), electromyography (EMG), and electrodermal activity (EDA). The EEG and iEEG records the electrical activity of the brain by means of electrodes placed over the head scalp or intracranial, respectively. The MEG also records the electrical current in the brain by means of a complex device; the ECG register the heart activity; EOG registers the activity of the eye; the EMG records the muscle responses; and the EDA registers the conductance of the skin [28, 97].

Among the above sensors, most of researchers [78, 31, 116, 93, 53, 112, 97] prefer EEG for several reasons. First, the technology is a low cost and easy to use at any age [69], unlike imaging modalities that require the cooperation of an adult. Second, it directly records the activity of the brain [101], in contrast to other data sources (like ECG, EMG, and EDA), which record complementary signals that indirectly correlate to an increase of mental WL, rather than the real WL [97, 98]. Finally, recent developments in deep learning (DL) have provided new ways to analyze EEG signals and provide objective assessment of cognitive abnormalities [49, 143, 151, 4].

Therefore, since its invention in 1929 by Hans Berger [101], EEG has become the main device to explore the brain activity and diagnose brain disorders (e.g., mental diseases, Alzheimer, Parkinson, epilepsy). In order to record the electrical activity of the brain, EEG uses a set of electrodes placed over the head scalp and moisturized by a conductive liquid to improve sensing. Each electrode records the electrical potential produced by millions of neurons while communicating themselves. Figure 1.1.a shows a commercial EEG device with 14 electrodes. The signals recorded by each electrode (also called channel) are displayed as a wave on a screen or printed in a page [45]. EEG headsets can vary in the number of electrodes depending on the purpose of the study. Taking into account the number of electrodes and the sampling frequency, named spatial and temporal resolution respectively, EEG has a relative low spatial resolution (on the order of centimeters), but a high temporal resolution (on the order of milliseconds)[22].

The signals recorded by an EEG are by nature transient and have complex waveform evolving over time [97]. EEG signals reflect the diversity of mental processes occurring in the brain at a giving time [101]. These include intellectual and cognitive processes, as well as, activity associated to body motion, sudden head movements, and eye blinking. We would like to note that this motion activity, rather than being an artifact, can also help in the detection of WL, stress, and other abnormalities. Figure 1.1.b illustrates the complexity of EEG signals. It depicts a snapshot of 5-second signal recorded by the 14 sensors of the headset shown in Figure 1.1.a. We note that EEG signals change both in frequency and amplitude, independently for each chan-



(a) (b)

Figure 1.1: The common portable electroencephalogram device Emotiv Epoc X. (a) The headset of 14 electrodes. (b) The signal patterns recorded during 5 seconds.

nel, since each electrode senses a specif region of the head scalp.

In this thesis, we use electroencephalogram (EEG) to record the electrical activity of the brain and develop deep learning (DL) based methods to detect, recognize, and assess cognitive states, particularly, mental workload.

1.1 Challenges for the assessment of cognitive states

Electroencephalogram data are time-varying signals that show complex waveforms and transient patterns [101]. Despite the recent progress achieved by machine learning/deep learning (ML/DL) methods for processing EEG signals, the assessment of WL still poses some challenges [19, 114, 97].

1. **EEG channel fusion**. The spatio-temporal signals recorded by EEG sensors should be combined to obtain a feature representation space describing the different cognitive states.

2. **Collection of unambiguous annotated data**. The collection of sufficient data from different mental states with unambiguous annotations of time signals that identify the different mental states present in each recording.

3. **Validation protocols to assess different levels of generalization of the models**. Like in other clinical applications, the across-subject generalization of models should be assessed to verify their success in predicting new unseen cases.

As follows, we give a brief description of each challenge according to the most recent state of the art (SoA) developments.

State of the Art

The EEG electrodes record the electrical activity of neurons at the same point of the brain and, thus, they provide several spatio-temporal signals [101]. The information of these signals (that can be regarded as different channels, like RGB channels in color images) should be combined in the most appropriate way to obtain an optimal output of the ML/DL method. However, the best strategy for the combination of EEG channels, also called spatial filtering in the literature, is still an open question [98, 123].

Previous research dealing with EEG signals has already pointed that working with all EEG channels is complex, so many works reduce the number of EEG channels to a single 1D signal. In order to get a single EEG channel, in some works the same experiment is repeated several times for each channel, and the channel that achieves the best result is chosen. Although this approach is trivial, it still remains to the present days, especially in the field of brain computer interface (BCI) because for motor imagery is uncomfortable to wear a multichannel EEG for a longtime [123]. Other researchers searched for a mathematical transformation to merge EEG channels. Such transformations are linear functions which project the original multichannel signal into a single new signal that maximises separation between classes [101]. This projection of EEG channels has been widely used in the detection of neurological disorders, like detection of epileptic seizures [10, 9].

More recently, the common spatial pattern (CSP) proposed in [95], has become the standard method to reduce the number of EEG channels. The CSP works maximizing the variance between two classes of signals, so it is widely used for EEG channel fusion or channel selection in two class classification problems [133, 129]. The CSP uses a projector matrix computed using the covariance matrix of signals recorded by the EEG, so CSP is highly sensitive to noise. To improve CSP, the empirical mode decomposition (EMD) algorithm could be used before to enhance the quality of signals [120], but with an additional computational burden [150, 129]. Besides, CSP is sensitive to the frequency band range, so recent improvements have involved the addition of a set of bank filters that works into multiple frequency bands, but with an additional computational cost [11].

The channels fused in a single signal are the input to a ML system trained to predict a specific cognitive state. Given that cognitive processes vary across time, signals are also cut into temporal windows for the extraction of features characterizing the cognitive states to be detected. There are two choices for feature extraction: a hand-crafted extraction using filter banks or mathematical descriptors [129, 42], or a trainable one using a DL approach [127, 68, 82, 97]. The hand-crafted approach requires a strong knowledge of signal processing methods in order to characterize the signal pattern. Examples of hand-crafted features are statistical moments (like the mean, variance, skewness, and kurtosis) and spectral features (computed in Fourier space). DL approaches are based on the use of convolutional networks able to learn the internal representation of signals from a large amount of samples. Next, for classification, any traditional machine learning (ML) algorithm (like neural networks - NN and support

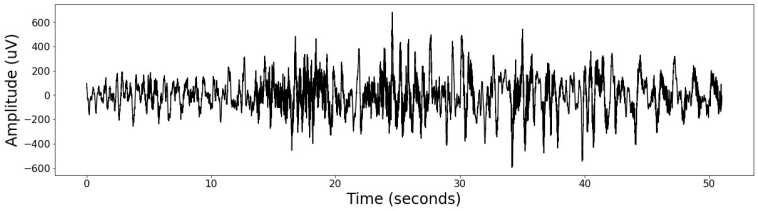
vector machine - SVM) can be trained for classification [83]. In case of using DL, it has the advantage of learning to classify during the feature extraction step [51]. To sum up, although CPS is a handcrafted spatial filtering to merge multichannel signals, it continues attracting more researches, even up to now [46, 74].

Encouraged by the recent successful results achieved by pure DL approaches, most researchers that work with EEG have focused on designing specific DL architectures for both channel fusion and temporal feature extraction. The preferred models are the convolutional neural networks (CNN) and the recurrent long short-term memory networks (LSTM) [98]. As well, few studies use DL for channel selection in order to find the most discriminative channels, so in this case, no channel fusion is performed. Like the first works, this channel selection is based on testing the performance of each electrode and selecting the best one, and thus, it is a time consuming process strongly dependent on the application [146, 81, 124].

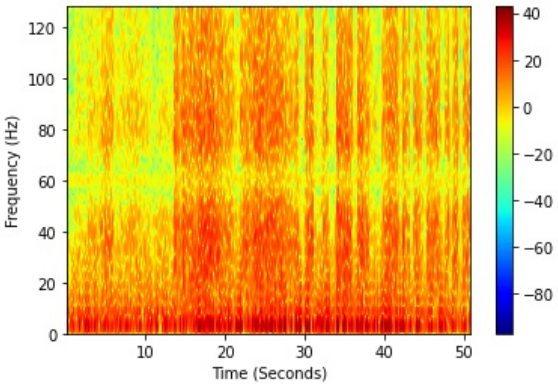
In order to process EEG data using DL, the multichannel signals recorded by EEG (see Figure 1.1.b) can be stored as a 2D matrix (like a gray scale image), where each row contains the temporal information coming from an independent EEG electrode placed over the head scalp and all rows are stacked vertically to form the 2D matrix. However, in opposite to the common sense that 2D EEG matrices can be directly processed by traditional 2D convolution it does not lead to good results [99]. The 2D kernels of these convolutions tend to inadequately combine the spatial information, so that, multichannel EEG signals still deserves special treatment.

Therefore, the idea of designing a specific layer for channel fusion has attracted a great deal of interest from researchers and it has been strongly influenced by the study presented in [99]. In this work, the authors concluded that the EEG multichannel 2D matrix should be run independently, either across time or across channels, but not simultaneously. Thus, the essence of dealing with EEG signals is to process EEG data while performing some combination of channels. There are two choices to process EEG signals: in time domain, taking directly the 2D matrix presented above or after performing a transformation of it to frequency domain. Figure 1.2 shows the representation of a signal recorded by one EEG electrode both in time and frequency domains. In time domain, the signal is a 1D signal varying only across time as shown in Figure 1.2.a. In frequency domain, the magnitudes of the Fourier Transform (FT) computed at a regular time and frequency patches are first stored in a spectrogram [59], and next, can be represented as a 2D image as shown in Figure 1.2.b. A spectrogram highlights the changes of the 1D signal both in time and frequency, simultaneously.

There are several works approaching the fusion of channels in the spatial domain. The study presented in [72] proposes to process multichannel EEG signals by steps. Firstly, to fuse spatially the multiple channels, and next, to process across the temporal dimension. In this work, the channel fusion layer consists of two CNN layers: the first layer goes across time dimension using a large kernel size and performing a rough feature extraction, whereas the second layer fuses the channels using a kernel size equal to the number of EEG electrodes. The activation functions used in both layers introduce a non-linearity in order to learn features. In order to alleviate computation bur-



(a)



(b)

Figure 1.2: Illustration of 1D and 2D representations of an EEG signal. (a) 1D EEG signal (time domain) of 51 seconds. (b) 2D spectrogram image (frequency domain) of the same signal. The spectrogram shows changes in signal intensities across time and frequencies.

den, the works presented in [104] and [107] proposed a smaller kernel size in the first convolution layer, while keeping the double convolutional layers and non-linearity on them. Finally, in [49] the authors propose to first extract temporal features with CNN and next perform a manner of spatial mapping of them using a fully connected (FC) layer. In particular, this approach does not perform EEG channel fusion explicitly, but only a combination of temporal features. In the studies presented above, convolutions go across a single dimension, either temporally or spatially.

An alternative to fusion and processing 1D EEG signals is the use of spectrogram images as input data for DL models. These studies first compute the spectrogram for each EEG signal and stack the normalized spectrograms as images along the depth dimension in such a way that the final image resembles an RGB image but with more than three channels. It follows that 2D convolutional networks can be applied to this color image and often researchers have explored the use of pretrained DL models from the image classification field [128, 98]. Before computing the spectrogram, the EEG signal should be standardized by channel, as a result, an additional computational burden is introduced which could be remarkable, specially for EEG of many electrodes.

Another issue is how temporal information should be incorporated into predictive models. In this context, LSTM has been explored due to its capability to learn temporal features in long sequences [137]. However, the training time and complexity of LSTM models increase as the number of input features increases. In order to alleviate the computation burden, researchers have introduced hybrid models involving CNN-LSTM. These approaches leverage the properties of convolution and pooling to reduce feature maps before seeding to the LSTM. The work presented in [144] uses a single EEG channel as input data for a traditional 1D CNN, and next, features extracted are sent to the LSTM. The study in [1] uses traditional 2D CNN to reduce feature maps of multichannel input data in an Encoder-Decoder architecture. In order to use 2D kernels, EEG recordings are first normalized by channels, and then, each input data of the model is scaled to 0-1 range. In this approach, the output of the encoder is fed to a LSTM. However, the reported results are low compared to the by-dimension convolution presented above. Aiming at improving performance, other variations of LSTM, like the nested-LSTM [75] and the bidirectional-LSTM [60], have been intensively explored. However, the increased performance of these LSTM models is at the cost of more trainable parameters and, thus, more training data. In the same way, a th recent study in [27] has reported that, for epilepsy detection, only LSTM-FC with large number of neurons in the FC layer outperform others complex LSTM-based approaches, though, at the cost of over-training the network with high possibility of overfitting. This study is very encouraging, but it still requires further analysis, mainly on its generalization capability.

Most of the EEG channel fusion methods combine information at the raw input level, but only few approaches have included fusion at the feature level. Furthermore, because of the absence of a conclusive study on the level in which EEG channels should be combined to provide the best results, actually many studies still use the traditional CSP method. Thereby, the searching for new methods to merge EEG channel have not been deeply explored yet; in particular, for linear methods which allow a better understanding of the combination of 1D signals due to their reversible characteristic.

The second challenge that ML/DL methods need to face is the collection of data. In order to train robust and unbiased models, the collection of a large enough dataset with unambiguous annotations is mandatory. Focusing on the goal and application of WL's studies, the majority of them are focused on assessing the level of WL in a specific group of people, such as commercial pilots to estimate their WL faced during their flights [19], car drivers to assess how WL affects their driving ability [36], air traffic controllers in high traffic conditions [94], workers during their routine office tasks [37], and other activities [29]. In these experiments, datasets are specifically collected for each experiment and often datasets are too short, restricted due to license conditions, or with raw data.

In order to record cognitive data in the domains above mentioned, there are two common choices: record data while performing the task in real life conditions or record data while performing the task in a simulator. In both cases, the subject faces a normal task and another remarkable abnormal task to collect differentiable cognitive loads.

While in some real life environments (like in office activities) could be easy to collect data, in other scenarios (like in automotive, aeronautics, industrial, military), recording of data could be almost impossible, time consuming, and expensive to collect sufficient data. For instance in aviation, datasets are collected from several professional pilots while flying in non-immersive flight simulators [93], in immersive cockpit simulators [54, 73], and during real flights [35]. As a result, datasets are mostly private, too restricted, and often insufficient to train DL models.

An alternative to simulators and real-life work conditions could be to collect data from subjects with artificially induced cognitive workload. The use of serious games is the most common method to induce different degrees of cognitive load. Unlike traditional video games which usually focus on entertainment, serious games are applications for particular goals to enhance personal skills [50]. In this way, serious games are widely used in education for different purposes [57], since training with serious games is more effective than training with conventional instruction methods [142, 153]. They are also used in social learning in favor of team opinions [130], neurorehabilitation [102], in industry, as a means of training workers to handle complex machines; in neuroscience, for understanding mental processes and flow of thoughts in the brain. There are free and commercial serious games, but they could also be designed for a specific case [103].

Among serious games, the N-Back test game is widely used because it produces cognitive load by demanding intensive use of memory in order to accomplish an asked task [62, 106]. To play the N-back test, the game should be installed in a computer and the cognitive task must be selected. In this way, many public datasets have been collected using the N-back test, although they suffer from various drawbacks [109, 16]. The recording intervals are too short providing not enough data and the asked tasks are limited to two high differentiable mental effort, which is completely different to real scenarios of WL where the mental effort has different degrees and changes along time even for the same task. Also, some researchers have designed their own serious game to incorporate an additional cognitive requirement (like multitasking, visual attention), but similarly they follow the traditional trend of the traditional N-back test [67, 76]. As ground truth, most of these datasets only provide the degree setup of the cognitive task. A few of them, give the subjective TLX-indexes. It is hard to find the scores achieved during the game which really provide an objective measurement of the cognitive workload.

In spite of the proven usefulness of serious games in other areas of applications, they are scarcely used in the context of WL assessment. Moreover, these datasets are not designed for knowledge task transfer of models, which is a topic little or nothing explored yet. Thus, in this thesis, we generate annotated data collected off-line from non specialist volunteers while performing cognitive tests. Collected data is general and sufficient to allow an effective task transfer of models trained using this data with the goal to be applied in other areas that also involves assessment of cognitive states.

Once enough annotated data has been collected, this has to be used for training and testing models using a validation protocol ensuring trustworthiness and fair as-

sessment of models. In the context of ML/DL, a validation protocol responses to two common issues: what should be validated, i.e., the ground truth, and how it should be validated, i.e., the statistical method used to assess the model performance [56, 83].

The first issue implies the assignment of the ground truth (the real or true value) to the available sample data, whereas the second issue concerns the definition of a metric score that provides a reliable measurement of the quality of a model in the test set. Nevertheless, in the context of cognitive states, analogously to the area of medical applications, a third requirement is mandatory: the assessment of levels of generalization of the models. This is because cognitive data involves data produced by various subjects performing different experiments and, thus, data can have an inter and intra subject variability that should be modelled. Therefore, this third condition requires special attention to assess the model's capability to generalize in various new unseen data scenarios, which might include new tasks and unseen subjects.

Given that the level of generalization of models should be assessed using statistical tests (like a t-test for the comparison of the model performance), it is directly associated to the experimental unit (or sampling unit) that define the samples of tests. There are three categories of sampling units: the window unit, the abnormal episode unit, and the subject unit. Each of the former units define the population samples for the computation of statistics and the ways of splitting of data into training and test sets in order to validate the model performance using an adequate metrics (e.g., accuracy, sensitivity, specificity, and F1-score).

Although we can find the three choices in the literature, most authors often only work with one of the options and specifically do not investigate the level of generalization. The first choice, the sampling unit at window level, the splitting of data does not take into account any relationships of subjects, so windows are considered interdependent [92, 49, 143, 4, 1, 27]. This approach is the most popular option and uses the classic k-fold cross validation splitting. The second choice, the sampling unit at abnormal episode level, a set of subject-specific models use both the training and test data that belongs to the same subject [110, 155, 47]. The main limitation of this strategy is that the model should be re-trained each time the subject under study changes. Finally, the third choice, the sampling unit at subject level, intends to develop models similar to cases of medical applications in which the model tries to recognize data from a subject that was completely unseen during the training, i.e., this is the population level evaluation, so it is the most realistic one from the three options proposed and it is still the most challenging [105, 107, 154]. In all three cases, the models' performance can vary considerably, so that they cannot be compared them directly, because results indicate different levels of generalization of models.

Therefore, due to the lack of a standardized validation protocol that takes into account the source and amount of cognitive data used in the training and the test, much of models proposed cannot be not fairly comparable. This may explain the limited development of real-world applications for assessment of WL in aeronautics and other related cognitive states.

1.2 Goals and contributions

The goal of this thesis is to explore, develop, and assess deep learning (DL)-based methods for the detection of cognitive states through the analysis of EEG signals. In particular, the thesis contributes to the main challenges identified in Section 1.1 in the following points:

1. **Approaches for EEG channel fusion**:

   In order to perform EEG channel fusion, we propose two neural network architectures that combines channel information at the input raw data level and after feature extraction. Besides, we introduce four fusion strategies implemented into the learning pipeline. The proposed networks are based on convolutional neural networks to process EEG data.

2. **Collection of unambiguous annotated data:**

   In order to collect an unambiguous annotated dataset for cognitive states, we propose the use of several serious games, to induce different levels of mental workload and collect the neurophysiological responses of several subjects using an EEG sensor in order to collect a sufficient data for training a model and perform knowledge transferring between tasks.

   Also, we collect a dataset from flight simulation that mimics different flights and events that jeopardize any professional pilots. The goal is to detect the workload faced by pilots during the flights by means of models pretrained in different cognitive target data and assess the capability of the models for transfer learning.

3. **Protocol to assess the generalization of models**:

   In cognitive states the use of only validation metrics (like accuracy and others) to compare the performance of models in order to select the best one for a specific application is not enough. There is a need to account the sample unit used in the experimental design in order to assess the level of generalization of models. The sample unit is the core that defines the mode of splitting of data into training and test set. Taking into account the sampling unit will provide a more insight about the performance of a model and its possibility of use in real life applications.

   In this thesis, we propose the assessment of generalization of models from the point of view of three sampling units (or unity of study). At window level (validation using the k-fold cross validation), at abnormal episode level (validation for the episodes of a specif subject), and at subject level (validation for the population). We claim that the model that performs best in all three schemes is ideally the best.

The **main contributions** of this thesis are outlined below:

1. **Neural network architectures for EEG channel fusion**

In this thesis, we present and evaluate two CNN-based neural network architectures for EEG channel fusion:

- The first neural network fuses EEG channel at the raw data level in the first step, and next, performs feature extraction of temporal features.
- On the contrary, the second neural network merges EEG channel after temporal feature extraction.

In contrast with conventional channel fusion methods based on complex mathematical techniques or non-linear transformations, we present 4 straightforward fusion methods are easy to implement and are trained into end-to-end model. The 4 fusion methods are: the average, the concatenation, the weighted average, and the multi-weighted average.

The average and concatenation are implemented by the mathematical mean and by the reshaping dimension function, respectively, whereas the both weighted average methods use a single convolution layer without any non-linear activation function.

2. **Serious games for the collection of unambiguous annotated datasets**

   In this thesis, we have collected a enough dataset in order to assess mental workload in aviation scenarios. We have collected and released a new dataset which provides unambiguous annotation like the theoretical degree level of cognitive workloads, the self-perceived workload of subject, and the achieved game scores.

   Collected data comes from two scenarios: serious games (from the N-back test and our Heat-the-Chair game) and flight simulator (from an Airbus A320 collected in an immersive cockpit and non-immersive computer desktop).

3. **Protocol to assess levels of generalization of models**

   In this thesis, we propose an objective validation protocol that allows to compare and validate different methods in order to choose the best one. Thus, from a point of view of model generalization, we propose three levels of validation:

   i) **k-Fold cross validation level**: Each time window has its own ground truth and does not matter to which subject it belongs. After shuffling the dataset, it is split into the training and test sets in order to train and test the model.

   ii) **Abnormal episode validation level**: This approach is similar to the population validation strategy; however, in this case, both the training and test data belong to the same subject.

   iii) **Population validation level**: Each time window has its own ground truth and it remains associated with the subject to belongs. The model is evaluated by choosing data from a subject each time that is used as a test set, whereas the remaining data is used for training the model.

The protocols presented to assess generalization of a model allow to validate different models in a fair fashion and in the same context where they intended to work. To assess performances, the k-fold cross validation (with a slight variations) is the best suited statistical method to validate the achieved results.

Finally, the validity of the different methods presented in this thesis has been assessed in two use cases:

1. **Mental workload assessment in flight scenarios**.

   This use case illustrates the task transfer capability of models pretrained in serious games to assess mental workload in different scenarios. Specifically, models were pretrained in the N-Back test data and validated in the flight simulator data.

2. **Epilepsy detection in pediatric patient with intractable seizures**.

   This use case illustrates the personalization levels of models. Depending of used the sampling unit, the model can report different performances, which should be in account at the moment to decide which is the best model for the desired application.

1.3 Structure of the dissertation

This dissertation is structured as follows. In Chapter 2 we present the approaches to process EEG signals, including main paradigms for EEG processing, the neural architectures for EEG channel fusion, and validation protocols to assess the levels of generalization of models. Chapter 3 provides a detailed description of the dataset collected in this research. Subsequently, Chapter 4 presents the case study of mental workload assessment in aeronautic environment. Next, Chapter 5 discloses the case study of epileptic seizure detection for medical diagnosis. Finally, Chapter 6 summarizes our conclusions and further work.

Chapter 2

Approaches to EEG processing

Any problem to be solved by a machine learning/deep learning (ML/DL) system needs to be well defined. For that, the main steps involved are the following:

1. **Data Management.** The first step is the collection of an annotated data set defining, both, the input and output data of the system. In general, the inputs, that the ML/DL method receives, are variables extracted from the raw data obtained from sensors, while the outputs are the information that the ML/DL method returns after processing the input data. This data should be unambigously annotated to constitute the ground truth (GT) used for training and testing the ML method and it usually depends on the type of problem (like classification, prediction, denoising) we are solving. In the particular case of EEG processing, given that EEG device provides a continuous recording of multiple signals that are cut in temporal windows, the input and output data can be, either a sequence or a single snap-shot. In Section 2.1 we explore different combinations of sequence/snapshots temporal units as input and output data.

2. **Selection of the most appropriate ML approach.** To solve a problem using ML/DL, a particular ML paradigm should be considered. A ML/DL model is an algorithm that is able to learn patterns from data in a supervised way and is used to make detection, classification or prediction of new unseen data. In the training step, ML-based models require a set of selected features to be trained, while DL-based approaches automatically learn features from data. As example, the support vector machine (SVM) and random forest (RF) are ML-based models and deep artificial neural networks (ANN) and convolutional neural networks (CNN) are DL-based models. In this work, we assume the model is a deep learning (DL) model. Besides, it is worth to mention that during the model definition to process EEG data, some strategy for combining EEG channels should be considered. Section 2.2 details the deep learning models designed for this work, with particular focus on the different strategies for fusion of EEG chanels.

3. **Generalization of Models.** Finally, an experimental set-up for the assessment and comparison of the developed methods should be defined. This step involves the assessment of the level of generalization of the trained models ansd should include: a strategy for splitting data, the metrics for the evaluation of models, and the statistics for their comparison. The splitting of data in training and test is directly associated to the personalization level of the model and the assessment of its generalization. A deep explanation of splitting of data and its association with the levels of generalization of models is presented in Section 2.3.

Figure 2.1 sketches and details the sequentially pipeline of the process in the particular case of processing electroencephalogram (EEG) signals. Highlighted in yellow, green, and brown we can localize the main steps previously enumerated. For the sake of sequential coherence with the pipeline shown in Figure 2.1, Section 2.1 defines some different possibilities of input and output EEG data, Section 2.2 explains the implementation of neural networks with EEG channel fusion capability, and Section 2.3 details the experimental design and data splitting for generalization of models.

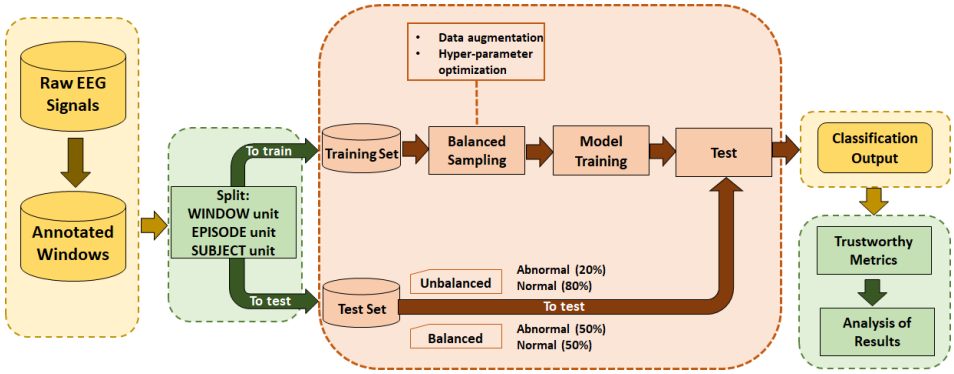


Figure 2.1: General pipeline for processing of EEG signals.

2.1 Data Management

An EEG device provides a continuous recording of multiple signals. In order to explore and analyze EEG. These continuous recording should be split into small processable chunks of data or time windows of few seconds of duration. All windows should have the same length to be properly processed by the ML/DL model. Thus, the input for a method processing EEG data is a sequence of L windows of length T and C EEG channels. Each temporal window in the sequence, namely X_i, can be represented as a 2 dimensional matrix of size [C,T] stacking the windows of the C channels. Thus, the input can be formulated as:

$$\left(X_i\right)_{i=1}^{L} \text{ where } X_i \text{ is a } C \times T \text{ matrix representing the i-th temporal window} \quad (2.1)$$

The length T is equal to the sensor sampling frequency times the number of seconds of the window. From a point of view of data structure, this input data will be an nd array of dimensions $[L, C, T]$. In case of input data $L = 1$ we are processing a single time window (named snapshot) and, thus, no temporal information is incorporated. In case that $L > 1$, models would incorporate temporal information, so we will name it sequence.

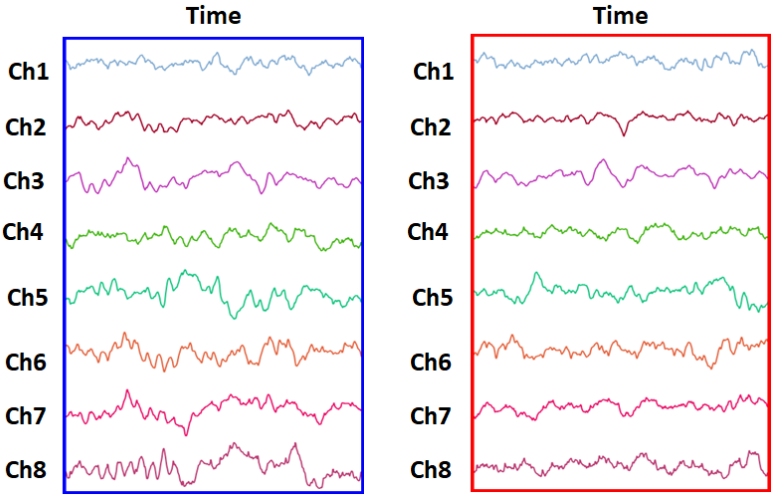


Figure 2.2: A temporal sequence of two windows of 2 seconds each for a recording with 8 channels.

Figure 2.2 shows an example of a temporal sequence of two windows of 2 seconds for a recording with 8 channels. Channels are plot along the vertical axis, while time corresponds to the horizontal axis. In this example, each time window consists of 8 EEG channels and lasts 2 seconds, which, considering a sampling frequency of 256 Hz, converts to a matrix of size [8,512] points.

In order to properly train and test a ML/DL model, it is mandatory to have data with unambiguous annotation of GT. The GT generation depends on the problem to be solved, but generally should be annotated as different labels. In analogous way to input data, the output data could be a single label or a sequence of predictions. That is:

$$\left(X_i\right)_{i=1}^{L} \longmapsto (y_i)_{i=1}^{L} \quad (2.2)$$

where y_i is the label for a the i-th temporal window X_i. For example, the signals shown in Figure 2.2 are labeled as "normal" and "abnormal" for a hypothetical case of binary classification. Figure 2.3 depicts the input data snapshots and their assigned output labels. For the case of supervised regression problems, the process of labeling is similar, assigning a numeric value instead of a categorical value.

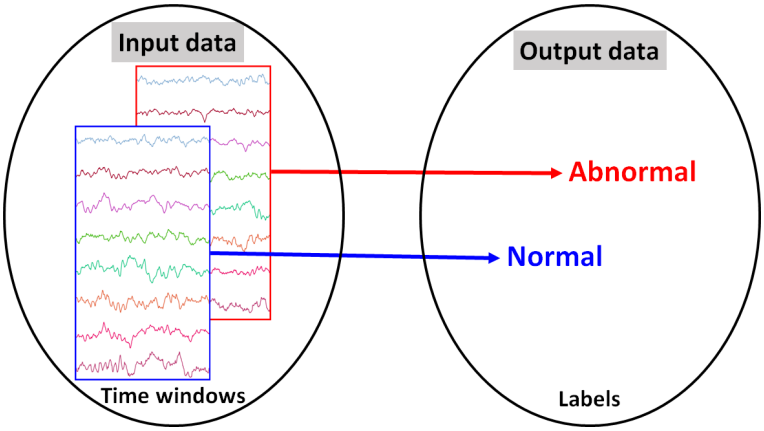


Figure 2.3: Two time windows (input data) and their assigned labels (output data).

The windows collected from all EEG electrodes are the raw input to ML/DL approaches that can output a prediction for the current time or output a future event. Given the temporal nature of EEG signals, there are several paradigms for processing these time windows and incorporating EEG temporal information, depending on the number of windows taken in input/output data. That is, focusing on the length of input/output data, according to the Equation 2.1 explained above, we can distinguish three options:

1. **(One2One) A single input maps to a single output**, where the model maps a single time window into a single one prediction. This paradigm is usually used for classification problem.

2. **(Sequence2One) A sequence of inputs maps to a single output**, where the model maps a sequence of time windows into a single one prediction. This paradigm is usually used to predict that will happen in the next snapshot.

3. **(Sequence2Sequence) A sequence of inputs maps to a sequence of outputs**, where the model maps a sequence of of time windows toward a sequence of predictions. This paradigm is usually used to predict that will happen in the next set sequence of snapshots.

2.1.1 One2One

The paradigm time window to one entails to take a single input data to achieve a single output prediction. Figure 2.4 illustrates the time window to one paradigm in which the model takes as input a single time window at the time t and provides a single prediction after processing it. This paradigm involves classification of signals that are independent of time, because it does not depend of precedent segments. Indeed, temporal information is taken into account but it is show as a snapshot of the signal,

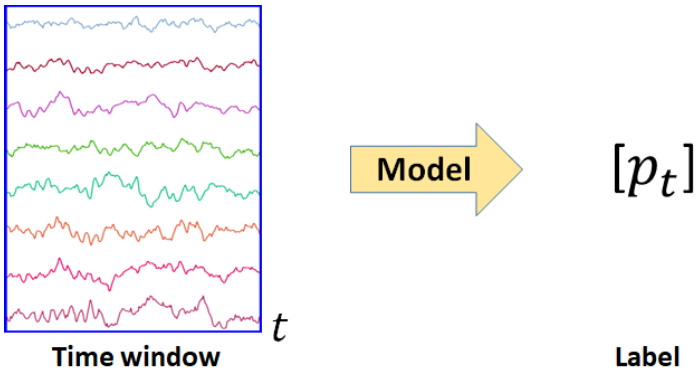


Figure 2.4: Illustration of the paradigm time window to one.

The assumption of independence of time allows to use models that not deal with time. In our case, we use convolutional neural networks (CNN) to classify a single time window signals and provide a single prediction (e.g., "normal" or "abnormal" like a medical diagnosis system). Likewise, regression is also feasible, for instance to assess the degree of sleepiness (in a scale of 1–7) given a signal segment. Both in classification an regression, the prediction consists of a single output value.

2.1.2 Sequence2One

The paradigm sequence to one encompasses to take a set of k sequentially temporal windows signals as input data to predict a single output. Figure 2.5 depicts the sequence to one paradigm, in which k consecutive input signals are fed into the model to forecast a value.

This scheme also can be used for classification and regression of the current value (time t) or for predicting the next value (time $t+1$). The main difference linked to the first paradigm is that in this schema the input data increases and the temporal information is taken into account. Because the prediction is a simple value, non-temporal DL algorithms also can be applied, while the sequence can be thought as a new time window of k concatenated signals. For instance, for $k = 4$ time windows, each of length t, the sequence becomes a new time window of length $4t$.

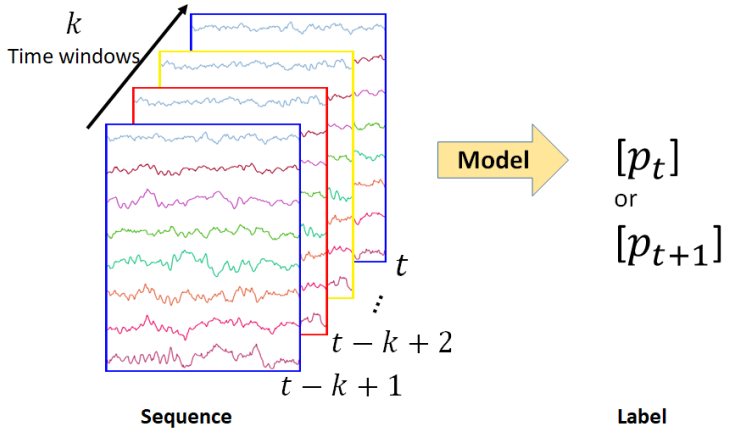


Figure 2.5: Illustration of the paradigm sequence to one.

2.1.3 Sequence2Sequence

The paradigm sequence to sequence implies to take a set of k time ordered time window signals as input data to predict k time ordered outputs. Figure 2.6 illustrates the sequence to sequence paradigm, in which k consecutive signals are fed into the model to forecast a sequence of values. Notice that both input data and output predictions are in an orderly way.

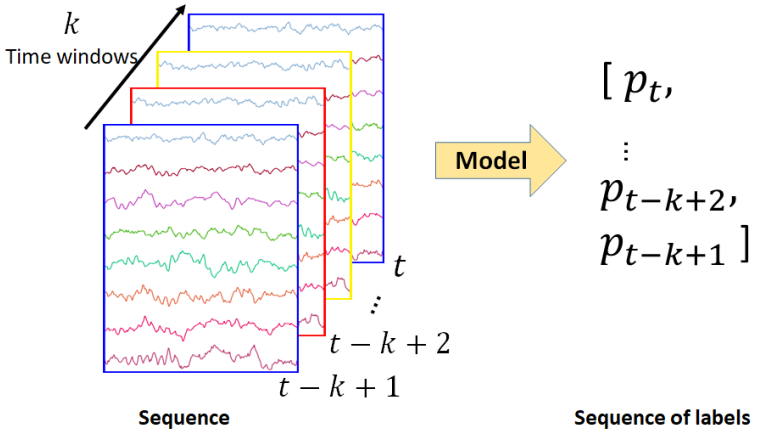


Figure 2.6: Illustration of the paradigm sequence to sequence.

Because the time matters, models that preserve the time context in long-term should be used. Thereby, long short-term memory (LSTM) and self-attention neural networks (Transformer) are more suitable for this case.

Finally, it is worth to mention that, although LSTM and Transformer were designed to deal with sequence to sequence mappings, the internal points of a single time window are also susceptible to be processed by these temporal-based models. This means that LSTM and Transformer can be applied to paradigms time window to one and sequence to one.

2.2 Neural network architectures of EEG channel fusion

From the beginnings of EEG processing, researchers have realized the importance of the simultaneous spatio-temporal nature of EEG signals for comprehension of mental processes happening in the brain. EEG records temporal data from different places of the brain, which should be optimally combined for better understanding of the brain functionality and to recognize cognitive states. The association of time data and their spatial location where they discharge would increase the performance of ML/DL systems.

In this chapter, we propose two general neural architectures based on convolutional neural networks (CNN) for the fusion of EEG signals. Both architectures consist of three sequentially main units: the Input Data Unit that takes the input data for the model, the Convolutional Unit that performs feature extraction and channel fusion, and the Output Unit that provides a combination of features coming from the Convolutional Unit by means of fully connected (FC) neurons and provide output predictions. In addition, it exists a Channel Fusion Unit, which positioning either before or after the Convolutional Unit makes the difference between these architectures. The Channel Fusion Unit placed before convolutions performs EEG channel fusion at raw input data level. On the contrary, the Channel Fusion Unit placed after convolutions furnishes EEG channel fusion at feature level. As well as, we explain the 4 different strategies proposed for fusing EEG channels, which work wherever (at input data level or at feature level).

2.2.1 Neural architecture for fusion at input data level

This neural architecture fuses EEG channels at the first step of processing, before temporal feature extraction. Next, the neural network continues with the steps of feature extraction and classification by means of the Convolutional Unit and Output Unit, respectively. Figure 2.7 depicts the proposed neural architecture for a general case of classification of EEG signals.

The proposed architecture can easily adapted for EEG headsets of any number of channels and for time windows of any length. For explanation purposes, lets consider a multi-class classification problem of n_class classes given an input data $x^{C\times T}$. The signals is processed as follows:

1. **Channel Fusion Unit**: The input data $x^{C\times T}$ becomes $x^{1\times T}$ (except for the CAT,

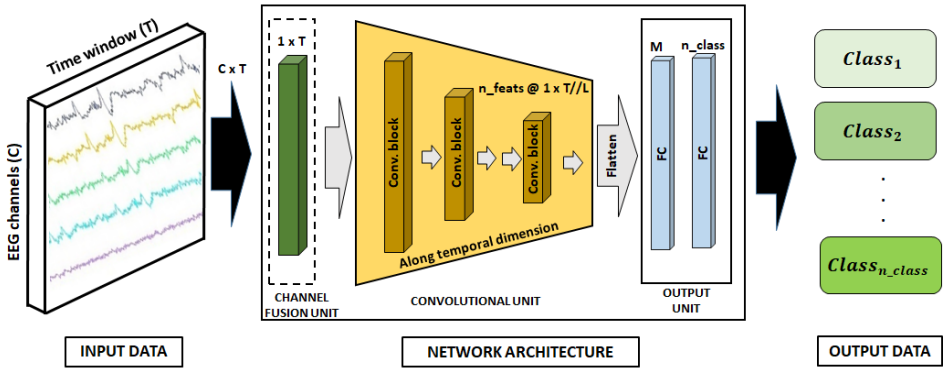


Figure 2.7: Neural network architecture for input-data level channel fusion.

that reshape it to $x^{1\times(C*T)}$).

2. **Convolutional Unit**: After L convolutional blocks, $x^{1\times T}$ becomes $x^{\text{n\_feats@}1\times(T//L)}$, where n_feats is the number of features extracted by the L^{th} convolutional block, and the $T//L$ is due to pooling that reduces the temporal dimension.

3. **Output Unit**: The outcome of the previous unit is flattened and connected to a FC layer of M neurons in order to learn the best combination of features. Next, the classification layer of n_class neurons accomplishes prediction using the Softmax function.

2.2.2 Neural architecture for fusion at feature level

Unlike to the neural network that performs fusion at the first step, this neural architecture fuses channels after temporal feature extraction. The neural architecture is similar to the previous one, except, the Channel Fusion Unit is located between the Convolutional Unit and the Output Unit. Figure 2.8 depicts the proposed neural architecture for a case of EEG signals classification.

In order to understand the data flow, lets assume a multi-class classification problem of n_class classes given an EEG signal $x^{C\times T}$. The signals is processed as follows:

1. **Convolutional Unit**: After L convolutional process, $x^{C\times T}$ becomes $x^{\text{n\_feats@}C\times(T//L)}$, where n_feats is the number of features extracted by the L^{th} convolutional block, and the $T//L$ is due to pooling operation en each convolutional block. The number of channels C keeps along convolutions.

2. **Channel Fusion Unit**: The input $x^{\text{n\_feats@}C\times(T//L)}$ becomes $x^{\text{n\_feats@}1\times(T//L)}$ (except for the CAT, that reshape it to $x^{\text{n\_feats@}1\times(C*T//L)}$).

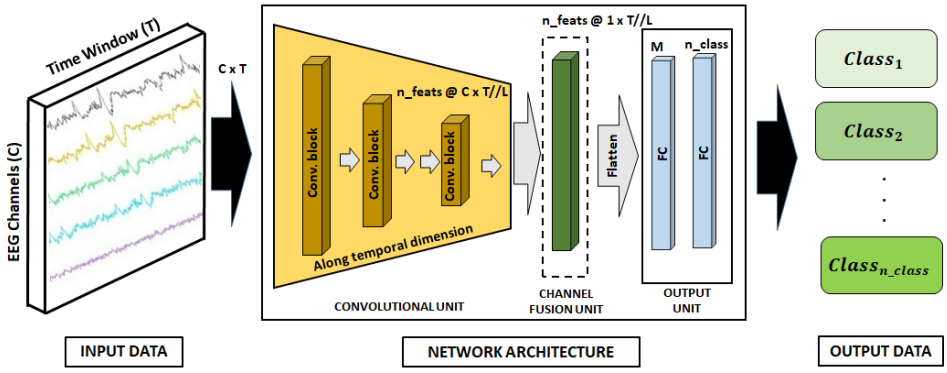


Figure 2.8: Neural network architecture for feature level fusion.

3. **Output Unit**: The outcome of the previous unit is flattened and processed by two FC layers to combine features and perform predictions.

2.2.3 Signal fusion strategies

Given an EEG time window snapshot X of shape $[1, C, T]$, for the sake of simplicity X can be rewritten as $X^{C\times T}$, where C is equal to the number of channels/electrodes and T is the time length. The problem of fusing C channels, or spatial filtering, is formalizes in Equation 2.3:

$$X^{1\times T} = \sum_{i}^{C} (x^{i\times T}) * (w_i) \tag{2.3}$$

where $x^{1\times T}$ is the resulting signal of the fusion, and $x^{i\times T}$ is a signal of the ith channel that is weighted by a weight w_i. The individual weights can be arranged in a matrix w. The weight matrix performs a linear transformation from a higher dimension to a lower dimension of EEG channels, usually a single EEG channel. As the transformation is linear, it is completely reversible.

As an example, Figure 2.9 illustrates a case of channel fusion using the average method. The topmost image shows 1D signals of 5 seconds that were recorded by an EEG of 4 channels. The bottommost image shows the result of fusion into a single channel after applying the average method. The average calculates the mean of values from the four channels at each time point in the signals. Instead of the average method, another technique might be applied in order to reduce the number of EEG channels.

In this thesis, 4 channel fusion strategies are proposed: the average, the concatenation, the weighted average, and the multiple-weighted average. The goal of each

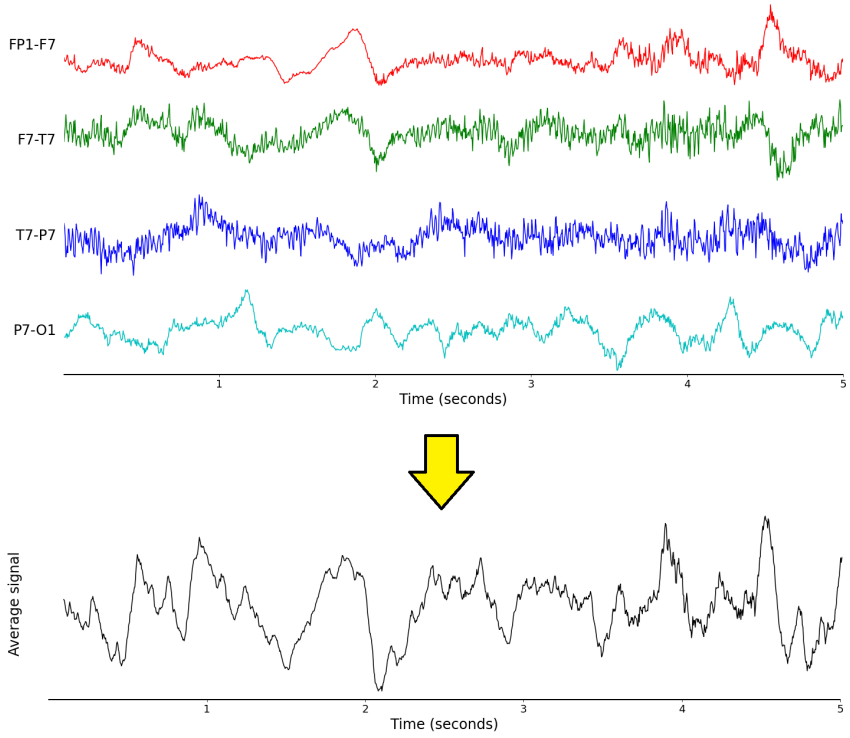


Figure 2.9: Illustration of EEG channel fusion. Four EEG signals (image on the top) are fused in one channel employing the average method (image on the bottom).

method is to fuse C EEG channels into a single one representation. As follows, we present a detailed description and implementation of each of them:

1. **Average.** The average method (AVG) performs the simple mean of channels, so $w_i = 1/C$ for each i channel. In this approach, each channel contributes equally to the output signal $x^{1\times T}$. Figure 2.9.b depicts the new fused signal after average 4 EEG signals in the illustrative example.

2. **Concatenation.** Instead of weighting, the concatenation method (CAT) does not fuse EEG channels, it just rearranges the dimension of the input data to $x^{1\times(C*T)}$, allowing the model to learn the spatio-temporal characteristics of the signals.

3. **Weighted average.** The weighted average method (W-AVG) is able to learn a matrix of weights w in order to perform channel fusion. Because, each w_i may be different, each EEG channels contribute differently to the new channel $x^{1\times T}$. The W-AVG method is implemented by a convolution with kernel size 1×1. The convolution has no activation function neither feature extraction.

4. **Multi-weighted average.** The multi-weighted average method (MW-AVG) also learns a matrix of weights w. However, in opposite to the previous W-AVG method, the MW-AVG takes into account the number of feature channels of the signals, so w becomes n_feats ×. This method is implemented by a convolution with kernel size $C \times 1$, i.e., kernel equal to the number of channels. Analogously to the previous one method, no activation function neither feature extraction is applied to ensure the linear combination of channels.

2.3 Generalization of models

As we explained above, the definition of the experimental design for training and testing of models should include: a strategy for splitting data, the metrics for the evaluation of models and the statistics for their comparison. In particular, the splitting of data before the task strongly influences in the achieved performance of models and the assessment of their capability of generalization. Generalization of models reside on the criterion of **sampling unit** used to split data into training and test set. For the assessment of cognitive states, we propose three different criteria to choose the sampling unit: time window, abnormal episode, and subject. Although these criteria have been used in different studies separately and with different names, in this thesis, we standardize them and present a general validation protocol to evaluate the capability of generalization of models.

As exposed in Section 2.1, collected EEG data must be segmented into processable temporal windows. Recordings can contain as many time windows can fit in it. Besides, now, an abnormality segment consists of temporal windows and their respective ground truth.

To illustrate the idea of sampling unit, Figure 2.10 depicts data from the signal of an EEG single electrode of N subjects. Subjects (framed in yellow) have normal and abnormal episodes (framed in green) segmented into time windows (normal episodes framed in blue and abnormal ones framed in red) of same length. The number of abnormal episodes can vary across subjects. Only for illustrative purposes, $subject_1$ has one abnormality episode, while $subject_N$ has two abnormalities.

As follows, we provide an explanation of the three levels of generalization of models and their supported sampling units:

2.3.1 Window unit: the k-fold cross validation level

A sampling in window units splits all input windows in train and test, losing the notion of belonging to subject or episode. This validation level provides the lowest assessment of generalization capability because it does not guarantee that the test and train windows do not belong to the same subject. In other words, the test set is not independent from the training samples and, thus, it is very close to evaluating with the same set of

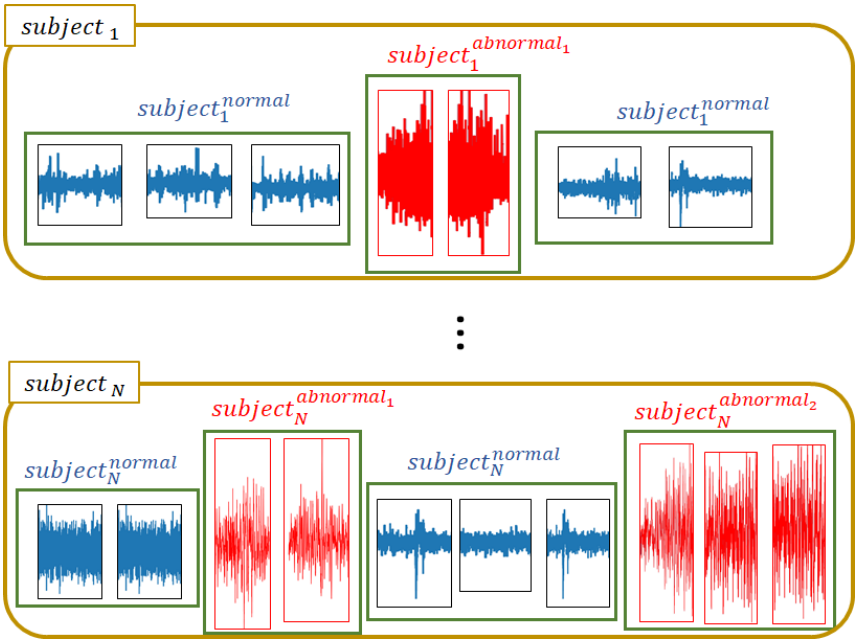


Figure 2.10: A single electrode EEG data from N subjects highlighting their episodes and temporal windows.

training samples.

The assessment of generalization is performed using the traditional k-fold cross validation (CV), so this validation protocol stands the name **the k-fold cross validation level**. Results under this scheme reports an upper bound of top performance [56]. Following the notation from equation (2.1), each $subject_j$ contains a number of L_j windows. Thus, if S is the set of N subjects,

$$S = \bigcup_{j=1}^{N} subject_j = \bigcup_{j=1}^{N} (X_i)_{i=1}^{L_j}$$

Each fold splits S in two random subsets, one for training and another for testing, and this random splitting is repeated k times. Usually, the subset for training contains the $60-70\%$ of the data, while the remainder is for testing and k = 5 or 10.

Figure 2.11 shows an example of k-folding. On the top of the figure, there is an example of a first splitting, while on the bottom of the figure there is an example of the k^{th} splitting. Because there is no notion of subject and episode, their bounding boxes were hidden and all the windows are in a row. The dashed square highlights the testing data, while continuous one are for training data. Notice that from one splitting to another some windows can be repeated in the same subset, but not all the windows. Be-

sides, a balanced distribution of unbalanced data should be taken into account, both in training and test. In Chapter 5, since the data used is highly unbalanced, we propose different ways to maintain the real distribution among training and test.

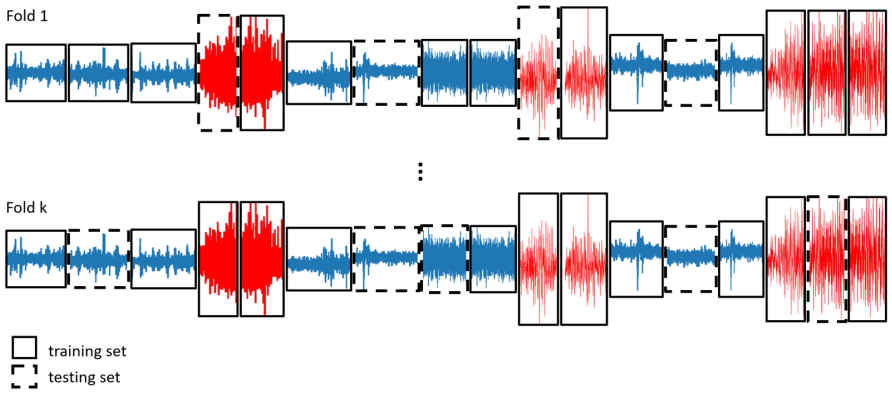


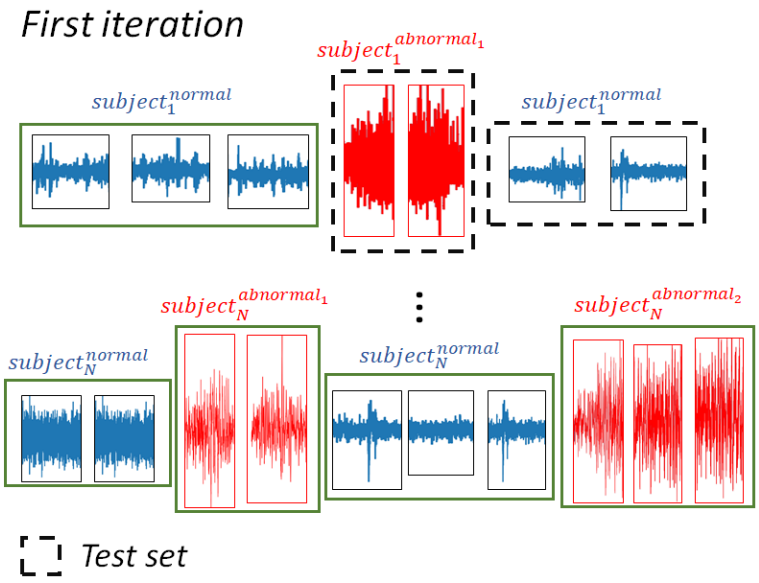
Figure 2.11: Illustration of sampling data at window level. Splitting of data for the fold 1 on top and splitting of data for the fold k on bottom.

2.3.2 Abnormal episode unit: subject-specific level

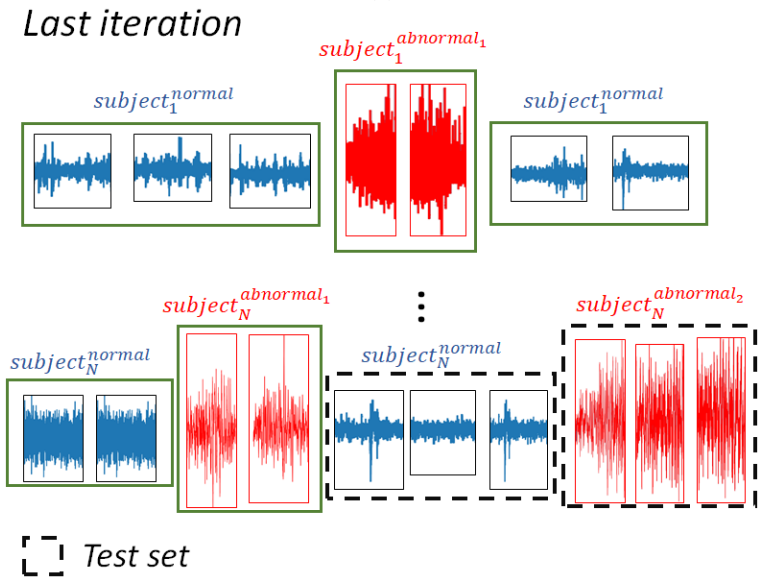
A sampling at abnormal episode would evaluate whether a model can detect new abnormal episodes in a given set of subjects. This design would evaluate models personalized for a given set of subjects. Thus, this validation protocol stands the name **the subject-specific level**.

Under this scheme, a variation of cross validation, named leave-one-out, at episode level is applied. In this way, the training set consists of all the patients minus one abnormal episode from one of them, which is the test set together with other data from normal episodes but all of them not considered in the training set. Following our mathematical notation, a $subject_j$ contains M_j abnormal episodes, $e_i^j, i = 1, \ldots, M_j$, so that the model is run $M_1 + M_2 + \ldots + M_j$ times and the iteration k corresponding to the abnormal episode i of $subject_j$ contains $S \setminus e_i^j$ for training and e_i^j plus a random normal episode for test.

Figure 2.12 illustrates the model at subject-specific level. Figure 2.12.a illustrates the first iteration while Figure 2.12.b illustrates the last one. Continuous squares correspond to training data while dashed squares correspond to test. Notice that in the first iteration, the only test data corresponds to the first subject while in the last iteration, the only data for test is the last abnormal episode together with some normal data. In cases data is very unbalanced, not all the normal episodes are used in the training set so that they can be used in the test.



(a)



(b)

Figure 2.12: Illustration of sampling data at episode level. (a) First iteration: first abnormal episode from first subject. (b) Last iteration: last abnormal episode from last subject.

2.3.3 Subject unit: Population level

For a sampling at subject level, test windows should all belong to a subject excluded from the training. That is the leave-one-out is done at patient level. This is the most generalist splitting given that the model is a population one and the test assesses its performance in a new subject never seen during training. Thus, this validation protocol stands the name **population level** because this scheme takes into account the cross-variability between subjects.

Figure 2.13.a illustrates a splitting of data for the fold 1: $subject_1$; whereas Figure 2.13.b shows the splitting of data for the fold N: $subject_N$. The dashed square highlights the test data.

Mathematically, select a subject$_i$ as test set, whereas the data of the remaining subjects composes the training set $S - subject_i$.

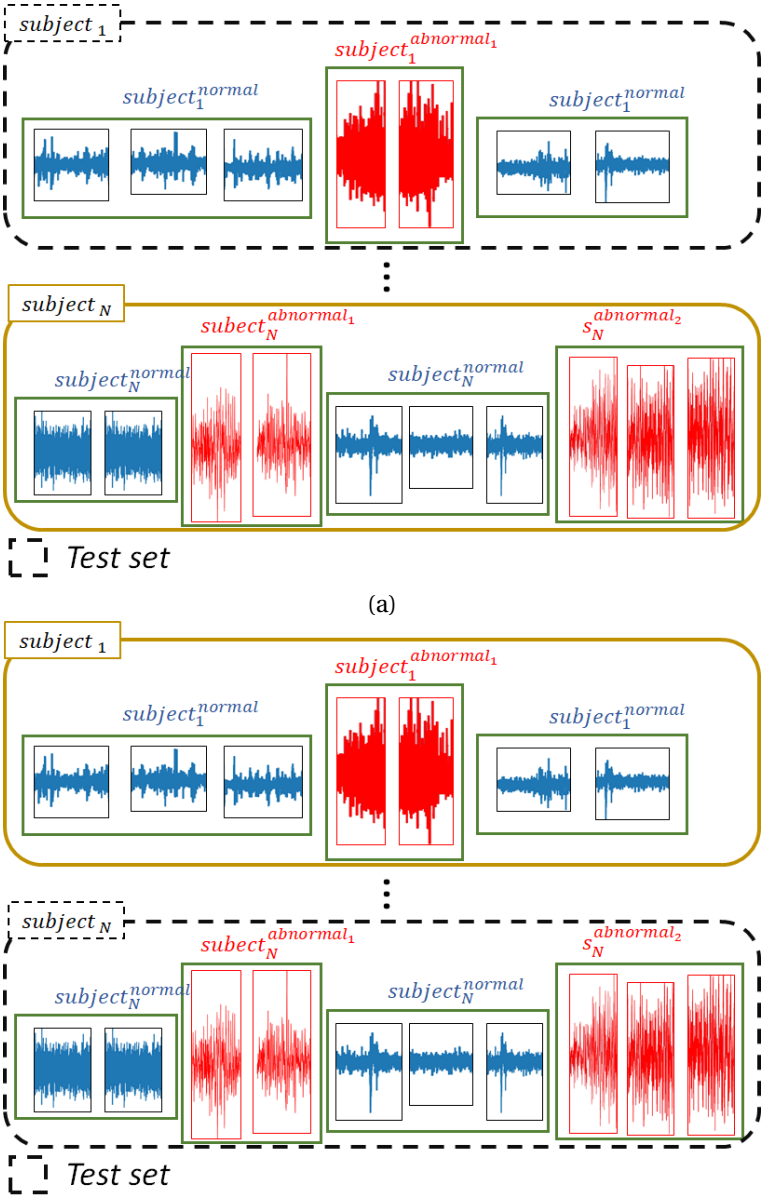


Figure 2.13: Illustration of sampling data at subject level. (a) Splitting of data to select the subject 1. (b) Splitting of data to select the subject N.

Chapter 3

Dataset for the assessment of mental workload

This chapter describes the dataset collected for mental workload assessment. The dataset contains neuro-physiological data from different experiments in which the participant faces different levels of workload.

As follows, a detailed explanation of our approach to collect the dataset is provided.

3.1 Serious games

3.1.1 The N-Back test experiment

In this experiment, we used the N-Back test game [20] to induce different levels mental workload on participants. We designed three experiments with different levels of complexity (low, medium and high) and each subject performed all the experiments using a computer desktop. The order of experiments was randomly assigned to each subject. The proposed n-back tests require the ability to manage one or two n-back tasks simultaneously, taking into account the insights shown in the n trial before, so it demands high usage of memory to successfully complete the tasks.

The three variants of the N-Back test to induce mental workload were implemented as follows:

1. **Low mental workload - position 1-back**: As Figure 3.1 shows, a square appears every few seconds in one of eight different positions on a regular grid over the screen. The player must press a key on the keyboard in case the position of the square on the current screen is the same as the position of the square appeared on the previous screen.

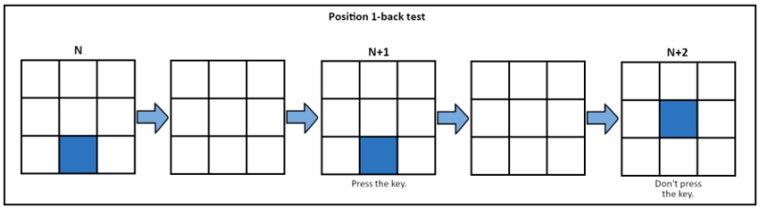


Figure 3.1: Example of position 1-back test.

2. **Medium mental workload - arithmetic 1-back**: As Figure 3.2 shows, an integer number between 0 and 9 appears every few seconds on the screen, while an arithmetic operation (plus, minus, times and divide) is audibly presented. The player has to solve this operation using the number that appeared in the previous screen and current one. Results must be typed using the numerical keys.

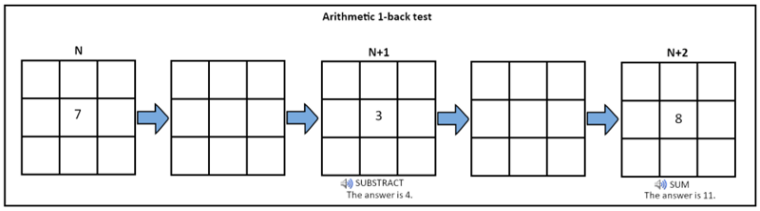


Figure 3.2: Example of arithmetic 1-back test.

3. **High mental workload - dual position and arithmetic 2-back**: This test combines the two previous ones. As Figure 3.3 shows, an integer number between 0 and 9 appears every few seconds in one of eight different positions on a regular grid. At the same time, an operation is audibly presented. As before, players have to type the solution of this operation using the number that appeared in two screens before and current one. In addition, players have to press a key in case the position of the current number is the same as the position of the number shown two screens before.

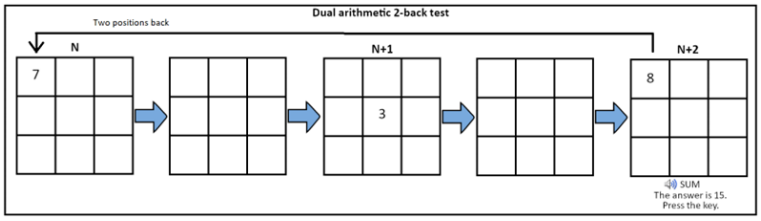


Figure 3.3: Example of dual position and arithmetic 2-back test.

Experiment Structure

Before playing and recording data, the subject was trained by playing the game during five minutes. For training, the dual position and audio 1-back mode was used, which simultaneously combines the position and audio, taking in account the 1-back step. That is, a number between 0 and 9 is audibly presented and the player must press a key if it matches with the one emitted in the previous screen, another one if its position matches and another one if both matches occur.

We assume that, in absence of any required mental effort, subjects will exhibit a baseline mental workload and their physiological responses will be accordingly in a minimum scale. We also expect that baseline levels will be different for individuals. In order to induce this baseline state, previously to the n-back tests, the participants watch a relaxing video for 10 minutes. Next, they play the game for 30 minutes. After the game, they are asked to answer the NASA-TLX questionnaire to collect their subjective perception of mental workload and effort demanded by the game. Finally, to calm down from the task, subjects take the recovery step for 10 minutes, which is likewise the baseline stage. The experimental protocol is outlined in Figure 3.4. During a session, all neuro-physiological responses are recorded; however, the dataset just contains signals from the baseline, the game task, and the recovery phase, besides the achieved scores of the player. The dataset also contains the results from the NASA-TLX questionnaire.

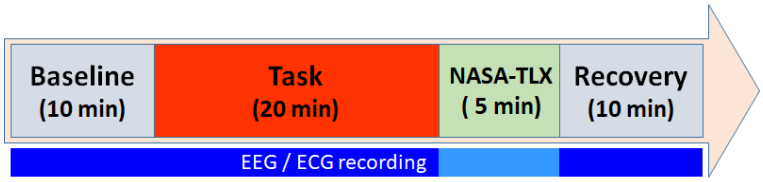


Figure 3.4: Timeline of the N-Back test experimental protocol.

3.1.2 The Heat-the-Chair experiment

This game was specifically designed to create a scenario where simultaneous tasks must be executed, replicating the demand of concentration and alertness of pilots while flying. The game consists of completing as many objectives as possible in 10 minutes. Completing an objective consists of obtaining and using the necessary pieces to form a 4-digit number, which appears on the top-left of the screen for 10 seconds and then disappears, reappearing for 5 seconds every 1 minute at the top left of the screen while the objective is not achieved. Once the correct pieces have been obtained and the target puzzle has been completed, the player increases the punctuation and a new target number to be completed automatically appears. Figure 3.5 shows the user interface of the game. The target number appears on the top left-hand panel, while the

pieces the player obtains are on the bottom right-hand panel. Notice that the bottom row is for storing the rewarded pieces (in cyan), while the top row is devoted to drag and drop the pieces and replicate the 4-digit number.

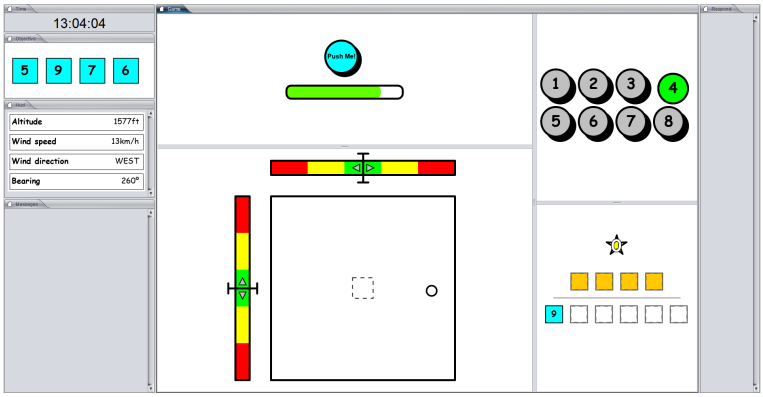


Figure 3.5: The Heat-the-Chair game user interface.

To obtain pieces, the player has to perform two main tasks:

1. **Bars with sliders**: As we can observe in figure 3.5 there are two colored bars in the bottom central-hand panel with sliders that move in the horizontal and vertical direction. The player must keep the sliders in the centre of the bars using the directional keys of the keyboard.

2. **Dots** In the same panel, there is a big square that will be filled with dots. In order to avoid this, the player must drag them to the dashed-line box shown in the center.

Located in the top central-hand panel, there is a circular button with an energy bar below that empties over time. The difficulty of tasks will increase proportionally to how empty the energy bar is: the emptier the bar is, the more difficult the game will become. Thus, the player must frequently recharge the power bar by means of the circular button.

As well, the game supports two modes of operation: with or without interruptions. Interruptions are introduced in the design of the game to emulate the interruptions that pilots receive when interacting with Air Traffic Control (ATC). In this mode, incoming events are randomly appearing to be solved. In particular, five different tasks, in randomly order, are required to be completed by the player. Tasks can be either to report a current flight parameter (altitude, wind speed, wind direction, and bearing) or to change the number of the switch box (the switch box starts randomly at each game). Flight information is shown on the left center-side of the screen and the switch box is shown on the top right-side of the screen. When an interruption arrives, an alert of

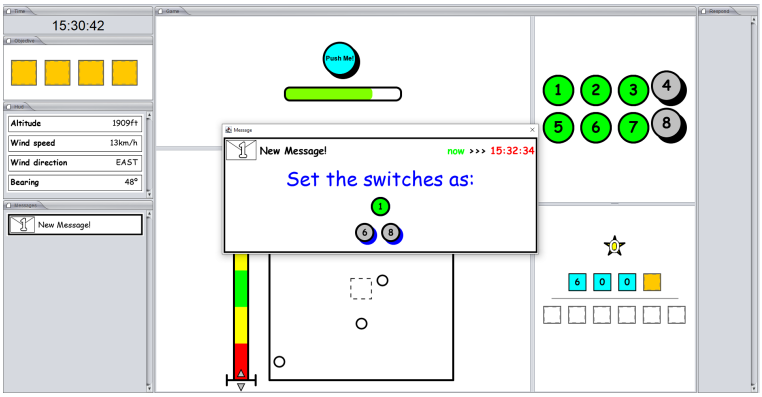


Figure 3.6: The Heat-the-Chair game with an interruption message.

messaging is shown on the bottom left-side of the screen. The player must to click and read the message. Each asked task has a starting and an ending time to be completed, out of which is the player is penalized. Figure 3.6 depicts an interruption asking for changing the current switch box. The starting and ending time to complete the task are highlighted in green and red, respectively. If the player does not conclude the task or inserts an incorrect information as answer, one rewarded piece is lost.

Experiment Structure

Before playing and without recording data, the subject was informed about the rules and trained in the Heat-the-Chair game without interruptions, during 5 minutes, in order to familiarize with controls.

Because each subject randomly faces the two modes of the game, each game is recorded in a separated session. As figure 3.7 shows, a session consists of three phases. The baseline, lasting 3 minutes, in which the subject drags balls that randomly appears on the screen and drops them to the dashed square in the centre. The game, in which the subject plays the randomly selected game mode for 10 minutes, either with interruptions or without interruptions. And finally, the questionnaire, in which the subject fills a NASA-TLX questionnaire indicating his/her subjective perceived game complexity. The neuro-physiological responses are recorded for the whole session, although the dataset just provides the signals from the baseline and the task phase, together with the achieved scores of the player. The dataset also contains the results from the NASA-TLX questionnaire.

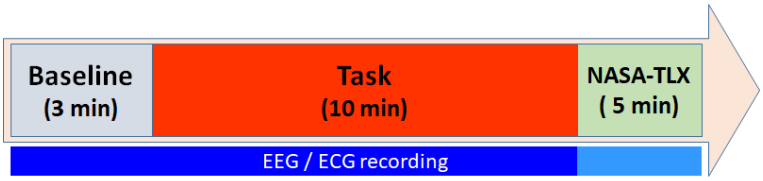


Figure 3.7: Timeline of the Heath-the-Chair experimental protocol.

3.1.3 Flight simulation with self-perceived workload estimation

The goal of this experiment was to collect experimental data useful to quantify the impact of the increment of mental workload of pilots while perform common tasks on usual flights (i.e., into the reference parameters) and while they must deal with additional unexpected phenomena, e.g., wind shears, machine failures, equipment warnings, and unusual traffics. In these situations, interaction between the crew itself and the ATC increase and pilots are more likely to make mistakes due to the mental workload.

Five flight experiments were designed to evaluate the pilots workload changes while they resolve unexpected situations. The flight simulation was carried out in an immersive Airbus-320 cockpit simulator and the chosen route to flight was from Barcelona departure airport to Lleida airport in Spain, with an approximate duration of 14 minutes. After certain interval of time, the pilot registers his self-perceived workload for that interval. Besides, all flights share the same weather, weight, and speeds conditions. Figure 3.8 illustrates the route followed by the pilot.

Two pilots participated in the experiments. However, only one pilot is selected to pilot the airplane; the non-selected pilot acts only as an observer. Every flight scenario is described as follows:

1. **Flight 1** [easy difficulty]: The pilot performs a standard flight to be used as the reference parameters.
2. **Flight 2** [medium difficulty]: During the flight, the ATC reports much traffic, so the pilot is asked to change the airplane position above the glide slope at high speed.
3. **Flight 3** [hard difficulty]: During the final stage of the flight, the airplane is hard destabilized by a strong wind shear, so the pilot must maneuver, recover the plane stability and landing it.
4. **Flight 4** [medium difficulty]: During the flight, it happens a malfunction during the approach that provokes a engine failure that increases the crew workload.
5. **Flight 5** [medium difficulty]: This flight is similar to the Flight 2 with a little variation.

Figure 3.8: Flight route of the flight simulation with self-perceived workload estimation.

As mentioned above, the pilots exchanged roles for each flight experiment. Table 3.1 outlines the pilots roles.

Table 3.1: Pilots roles.

Experiment	Pilot = Pilot 1 Observer = Pilot 2	Pilot = Pilot 2 Observer = Pilot 1
Easy	-	Flight 1
Medium	Flight 2	Flight 4, Flight 5
Hard	Flight 3	-

Experiment Structure

The experimental protocol of each flight is divided into two phases. First, the baseline phase, in which the pilot stays on the runway awaiting the order of takeoff. Second, the flight phase, which at the same time can be split in three stages: the takeoff, when the flight starts and the plane climbs; the task phase itself, and a short time for landing the plane to the ground. The task phase includes the cruise, the descent, and the approach tasks, accompanied with the usual communication with the ATC, and together with asked tasks specific of each flight simulation. Figure 3.9 outlines the timeline of flight simulation experiment. The dataset stores the neuro-physiological responses

from all phases of the experiment, as well as, the self-perceived difficulty by pilots.

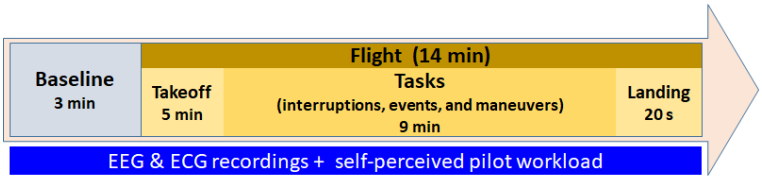


Figure 3.9: Timeline of the flight simulation self-perceived workload estimation.

3.1.4 Flight simulation with FRAM-based workload estimation

The goal of this experiment is to investigate how pilot's workload correlates to pilot's mental processes while operating flights in different circumstances. The flight is completed in an immersive simulator of the Airbus-320 aircraft. To induce workload, a set of circumstances are proposed: increase the operational interruptions of the ATC, increment the interactions of the passenger cabin crew (TCP) or flight attendant, and enforcing different warnings of the electronic centralized aircraft monitor (ECAM).

Fourteen flights were designed, which assume the pilot monitoring (PM) awkwardness to check how interruptions can overwhelm pilot flying (PF). The chosen route to flight was from Gerona departure airport to Barcelona airport in Spain, with an approximate duration of 14 minutes. Figure 3.10) illustrates the route followed by the pilot. A single pilot participated in the 14 experiments. Before each flight, the pilot received a printed description of the mission to be completed.

Each flight experiment is described as follows:

1. **Flight 1**: baseline flight. This experiment is a standard flight to be used as the reference parameters and a baseline flight. It considers the workload is acceptable and PF can attend the interruption without a negative performance impact. Three ATC instructions are asked at a given time interval as in usual flights.

2. **Flight 2**: too high and too fast. This experiment is similar to the previous one and resides on the approach phase. However, ATC instructions to be solved as soon as possible before reply.

3. **Flight 3**: ECAM interruption scenario. This experiment also resides on the approach phase and is designed to increase PF workload by means of an ECAM (ELEC GEN FAULT + APU OFF) that appears when PF workload is low and can attend the interruption without a negative impact. That ECAM requires PF attending the actions during various minutes; however, an instruction of ATC forces PF to postpone the ECAM to execute the instruction of transition level and the approach checklist before re-assuming the ECAM actions. These concurrent

Figure 3.10: Flight route of the flight simulation with FRAM-based workload estimation.

actions increased considerably the workload with an impact on the PF performance.

4. **Flight 4**: cabin crew interruption. This experiment is based on the Flight 2, and during the flight is reported a passenger with a heart attack, so it is declared an emergency on board. During the emergency the communication with the ATC is continuous, which impact the PF performance

5. **Flight 5**: baseline flight with un-opportunistic interruption. This experiment is designed to collect reference parameters likewise the baseline flight, but with a random communication fired by the TCP. In this case, different parameters are found if compared with the parameters of the baseline flight, which may be caused by PF fatigue.

6. **Flight 6**: Un-opportunistic ATC interruptions. This experiment is based on Flight 2; however, the ATC interruptions over PF are un-opportunistic because PF is attending other concurrent actions. It creates pending memory items, increasing considerably the workload impacting on the PF performance.

7. **Flight 7**: ECAM interruption scenario. This experiment is based on the Flight 2 by firing the ECAM (AUTO FLT A/THR OFF) when PF is performing the precision approach. This scenario evaluates how a malfunction of a system impact on PF performance.

8. **Flight 8**: ECAM interruption scenario. This experiment is a slight modification

of the Flight 7 by adding two ECAM (AUTO FLT A/THR OFF + AP OFF) interruptions.

9. **Flight 9**: ECAM interruption scenario. This experiment is a slight variation of the Flight 8, but with the goal to analyze fatigue on the PF.

10. **Flight 10**: Auto FLT AP OFF scenario. This experiment is a slight variation of the Flight 9, by an ECAM interruption fired at 20 NM from runway. ECAM interruption co-exist with other interruptions that increases PF workload.

11. **Flight 11**: Auto FLT A/THR OFF+ AP OFF scenario. This experiment is based on the Flight 10. It is designed to analyse the benefits of postponing a non-safety-critical interruption to a convenient time window in which the concurrent actions can be performed by PF without generating a pending memory action.

12. **Flight 12**: ELEC GEN FAULT + APU OFF scenario. This experiment is based on the Flight 3, but, in this case, the ECAM is fired in when the aircraft is in short final. The goal is also to analyse the benefits of postponing a non-safety-critical interruption to a convenient time window in which the concurrent actions can be performed by PF without generating a pending memory action.

13. **Flight 13**: ELEC GEN FAULT+ TCAS RA scenario. This experiment is designed to provoke a PF peak workload by firing an ECAM and a TCAS RA in short final. The scenario considers also the postponement of an ATC instruction to a valley workload to validate the benefits of a potential SSA.

14. **Flight 14**: Passenger heart attack scenario. This experiment is based on the Flight 5. The experiment increases the PF workload because there is an emergency on board and this emergency trigger a Direct to SOTIL. The performance of the PF with the mentioned interruptions, confirms the benefits of a CC interruption manager that could postpone non-safety-critical interruptions to convenient points.

Experiment Structure

The experimental protocol of each flight is divided into two phases. First, the baseline, in which the pilot stays on the runway awaiting the order of takeoff while checking controls and instrumentals. Second, the flight, which can be split into three stages: the takeoff, when the flight starts and the aircraft climbs; the phase of tasks; and a short time for landing the plane to the ground. The task stage includes the cruise, the descent, and the approach tasks. Most of designed events fire in the approach phase. Figure 3.11 depicts the timeline of flight simulation experiment. The dataset contains the neuro-physiological responses during all the experiment, together with the FRAM's registers, parameters, and workload estimations.

Based on the flow actions of pilot's actions during the flight, the FRAM model estimates the mental workload faced by pilots [90, 100].

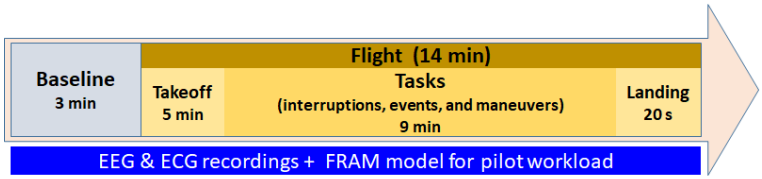


Figure 3.11: Timeline of the flight simulation with FRAM-based workload estimation.

3.2 Participants

The characteristics of all the participants are detailed as follows:

1. **The N-back test experiment**: 16 subjects, with ages ranging from 20 to 60 years, participated in the experiments. Volunteers belonged to three different university research centres and shared some scientific background with different levels of expertise, because they were either students, junior researchers, senior researchers, or professors.

2. **The Heat-the-Chair game experiment**: 17 subjects (12 male and 5 female) participated in the experiment. The volunteers share the same characteristics of participants in the previous experiment, and seven of them completed the preceding test.

3. **The flight simulation experiments**: two professional pilots with different experience level participated in the flight simulation experiment. In one hand, the flight simulation with self-perceived workload estimation has involved the two pilots, who interchanged the role of pilot and observer. On the other hand, in the flight simulation with FRAM-based workload estimation, a single pilot carried out all flight mission. Table 3.2 details the information of pilots and their participating.

Table 3.2: Pilots information.

Flight	Pilot	Gender	Age	Flight Hours
Self-perceived & FRAM-base	Pilot 1	Male	51	4000
Self-perceived	Pilot 2	Male	32	1700

Volunteers that have participated in all experiments are healthy people without any condition that might have cause an imbalance in the data recorded. Figure 3.12 illustrates some of the participants in the different experiments. All of them signed a consent to publish their images.

(a) (b)

(c) (d)

Figure 3.12: Volunteers during the experiments. (a) Performing the N-back test. (b) Performing the Heat-the-Chair game. (c) Flight simulation in a self-perceived workload estimation experiment. (d) Flight simulation in a FRAM-based workload estimation experiment.

3.3 Physiological sensors

The dataset contains both EEG and ECG signals for each subject and for each experiment. The N-Back test game and the flight simulation experiments were recorded using the EEG Epoc X [39] and the ECG Suunto Ambit3 Peak with the hearth rate belt [119], whilst the Heath-the-Chair game was recorded using the same EEG and the ECG Shimmer3 [108]. Table 3.3 presents the physiological sensors adopted for each experiment and Figure 3.13 depicts an screenshot of them.

Table 3.3: Physiological sensors used by experiment.

	Sensor		
Experiments	Emotiv Epoc X	ECG Suunto	ECG Shimmer
N-Back test	x	x	-
Heat-the-Chair	x	-	x
Flight Simulator	x	x	-

(a) (b) (c)

Figure 3.13: Sensors used for the experiments. (a) EEG Emotiv Epoc X. (b) ECG Suunto. (c) ECG Shimmer.

1. The **EEG** sensor from EMOTIV[39] (Figure 3.13.a) consists of a portable, high resolution, 14-electrodes EEG system, according to the International 10-20 System and are placed over the head scalp in order to track the electrical activity of the brain. This device provides the raw signals, in μV, at a sampling rate of 128 Hz. Besides, the sensor furnishes the power band for the major brain rhythms (beta: 4–8 Hz, alpha: 8–12 Hz, beta low: 12–18 Hz, beta high: 18–25 Hz, and Gamma: > 25 Hz). Emotiv gives 8 powers samples per second computed over the previous 2 seconds.

2. The **ECG** sensor from Suunto[119] (Figure 3.13.b) consists of a clock and a belt. To record, the clock is place on the left wrist, whilst wears the belt over the bare skin around the chest and they provide the hearth rate and respiration rate measurements.

3. The **ECG** Shimmer3 EBio Consensys Development Kit from Shimmer[108] (Figure 3.13.c) is an ECG Kit containing connectors to be patched over the chest that record the hearth rate and respiration rate measurements. Electrical impulses are specifically measured from across the chest and captured by the Shimmer sensor. These signals are transferred to Shimmers ConsensysPRO Software for real time analysis or stored onboard the SD card for post-processing.

3.4 Technical validation

3.4.1 N-Back test

In order to evaluate the technical quality of the collected data in the N-Back test, we analyzed the answers to TLX questionnaires. Since the TLX reports the self-perceived degree of workload enforced by the tasks we put them in correspondence with the performance of the players. Figure 3.14 illustrates the boxplots of the behaviour of such variables. Notice that both measures are in the range [0,100], so they can directly be compared. On the one hand, the perceived workload of participants shows a positive correlation with the theoretical workload of tasks. On the other hand, the empirical performance of subjects exhibits an negative correlation against the difficult of tasks. The higher the workload experienced the subject, the lower the performance results.

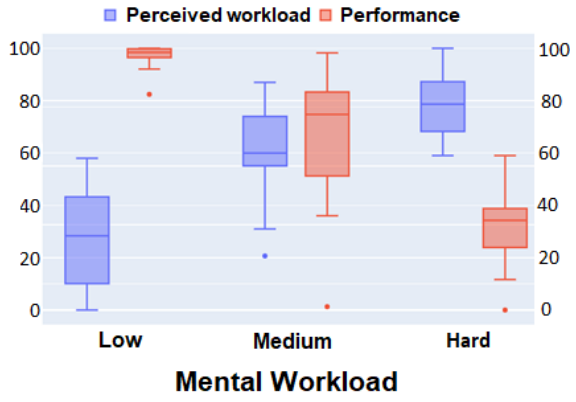


Figure 3.14: TLX analysis in the n-back test: perceived workload versus empirical performance.

3.4.2 Heat-The-Chair game

Because the Heat-the-Chair game is a slight variation of the N-Back test, to validate the quality of collected data, we also use the TLX questionnaire. In this case, the subjects' response is cross-checked with the time the player needs to obtain a piece when the play without and with interruptions. Figure 3.15 shows the boxplots of the perceived workload and the empirical performance in terms of the average time to get pieces during the game. Notice that both measures have different ranges, so that they have been normalized. The range shown in red on the left hand-side corresponds to the time needed to obtain pieces, while the range shown in blue on the right hand-side corresponds to the range of punctuation for TLX-results, between 0 and 50.

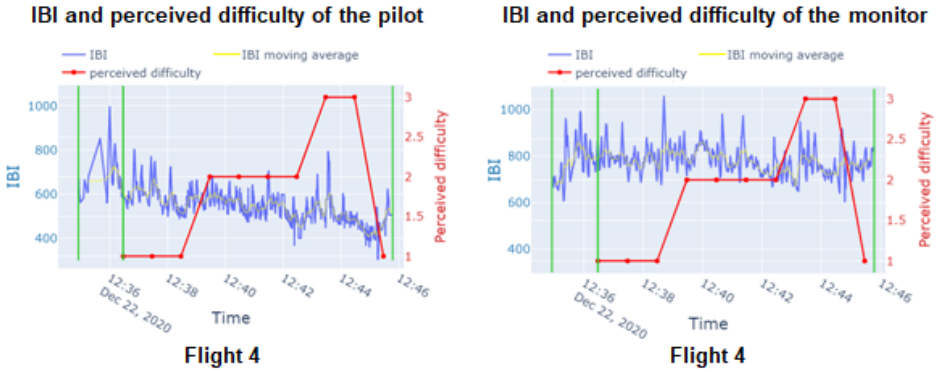
This plot supports the correctness of the proposed experimental design. That is,

on the one hand, the self-perceived mental workload has a positive correlation with the complexity of the game. On the other hand, the more complex the game is, the more longer the average time to gain a piece results and thus, the subject performance drops.

Figure 3.15: TLX analysis in the Heat-the-Chair test. TLX vs. Game performance.

3.4.3 Flight simulation with self-perceived workload estimation

To ensure the reliability of the collected data in this experiment, we correlate the task perceived difficulty against the the interbeat interval (IBI) of the heart. Figure 3.16 shows a visualization of the perceived difficulty (red line) of the pilot in the Flight 4, overlapped with the IBI (blue line) of the own pilot (a) and the observer (b). Green lines illustrate the different phases of the experiment. In Figure 3.16.a, it is noticeable that at the beginning of the flight there is no perceived difficult of the pilot because it is the baseline stage. However, it changes after the takeoff until to landing. At the beginning of the flight the IBI is stable, but after the flight starts the IBI tends to decreases as a result of the increasing difficulty in the flight. This behaviour makes sense, a lower IBI amplitude indicates that the heart is beating faster as response to the mental workload changes that the pilot is facing at a given time. On the other side, a different behaviour of IBI is noticed in the observer. Figure 3.16.b shows an IBI more stable along the flight. Again, this behaviour makes sense, the observer does not have any responsibility about the flight, so he is quite calm.



(a) (b)

Figure 3.16: The experiment on Wasim. IBI vs perceived task difficulty: (a) in the case of the pilot. (b) in the case of the copilot.

3.4.4 Flight simulation with FRAM-based workload estimation

The FRAM provides a workload estimation of pilots by using the flow actions of pilot's activities. The model is widely used in flight scenarios and for a deep understanding refers to [90, 100]. As illustration of how FRAM works, we present the case of the Flight 4 (cabin crew interruption scenario). Figure 3.17 depicts the flight route and the ATC instructions.

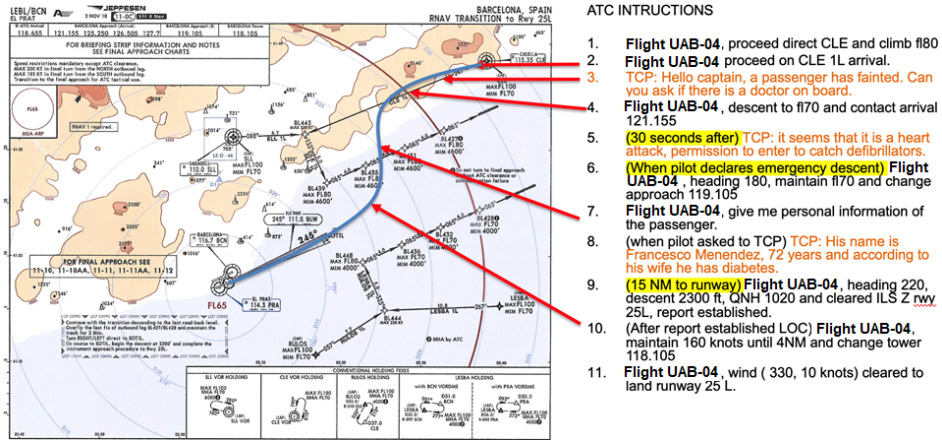
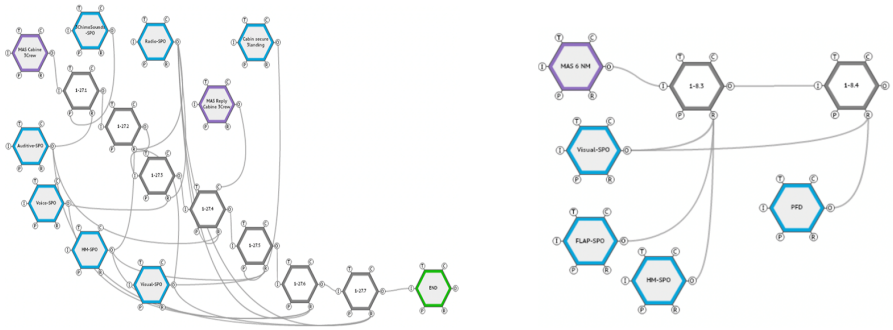


Figure 3.17: Flight plan of the Flight 5.

While the flight is ongoing, the FRAM simulation model represents the interaction of PF, ATC, cabin crew, and machine parameters. Figure 3.18 illustrates the FRAM models developed for both the cabin crew and the precision approach for the experiment.



(a) (b)

Figure 3.18: FRAM model of the Flight 4. a) The cabin crew FRAM model. b) The precision approach (Flap 2) FRAM model.

After a set of actions, FRAM can estimate the workload faced by the pilot at a given time. Figure 3.19 shows the timeline of flow of actions during the flight and the colored vertical bar depicts the guest workload. PF executes all the actions because PM is incapacitated and there are concurrent actions of 1-27.4 and 1-27.5 with 1-8.3 and 1-8.4. The green, yellow and red colours in the timeline bar, summarizes the FRAM simulation model expected workload.

For all 14 flight simulations, the dataset provides the flight plan, the chart of FRAM models, and the flow of actions along time. The workload estimated by FRAM is considered the ground-truth of the mental effort of the pilot at the given time. Researchers can use this workload for comparison against other mental workload estimators in future investigations.

3.5 Ethical Approval

The Ethics Committee on Animal and Human Research (CEEA) of the Universitat Autònoma de Barcelona, Spain, provided us with an approval letter to collect neurophysiological data in the Project "E-PILOTS (H2020)" Grant agreement ID: 831993, requested by Dr. Miquel Àngel Piera. The letter was signed on September 2, 2022, by Núria Pérez Pastor, Secretary of the CEEA.

The data collection process was carried on according to the Code of Good Practice in Research of the Autonomous University of Barcelona [12]. Also, before the experiment, each volunteer was informed about the goal of this research and a written consent was obtained. The collected data was anonymized to protect any personal information for data analyses and publications related to this research.

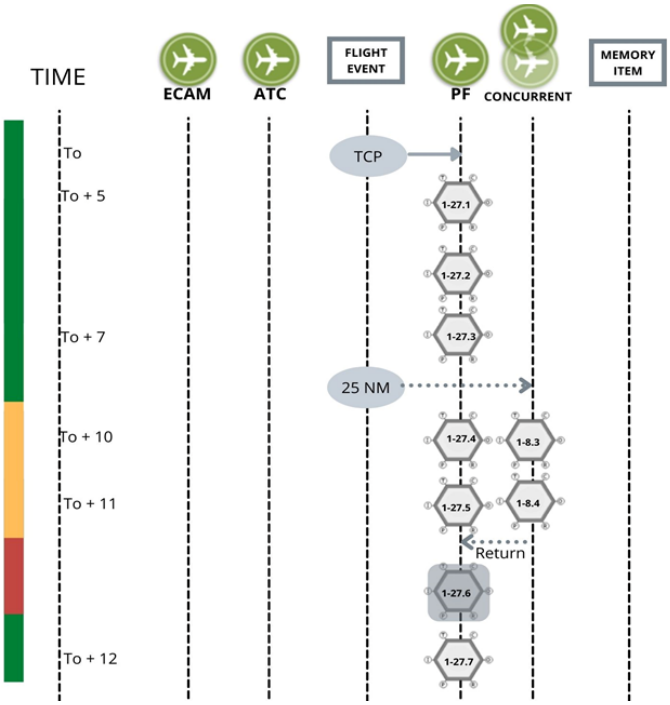


Figure 3.19: Flow of actions of the Flight 4.

3.6 Data repository

The dataset described in this paper has been partially made publicly available at the Digital Document Deposit of the Universitat Autònoma de Barcelona, downloadable in [145]. No registration is required for anyone who like to download and use this data.

The dataset is delivered in a compressed file `workload_dataset.zip`. After decompression, the dataset contains three main folders that store the collected data for the N-back test data, the Heat-the-Chair data, and the Flight Simulation with the self-perceived workload estimation, respectively.

Actually, the Flight Simulation with FRAM-based workload estimation is still not released, while the validation of FRAM models is carried out.

3.7 Usage of dataset

There are two main options to use the dataset. On the one hand, the use of a proprietary tool process the data. Since parquet files are tabular containers, MATLAB 2019b

seems the best one. On the other hand, the use of open software provides a good alternative; here, the use of Python 3.8, with the Pandas 1.4.0 dataframe library and the Pyarrow 8.0 allow to read, process and move data stored in parquets.

Aiming to prepare EEG data to be used in a classification problem, we provide a data preprocessing script that take as input the raw EEG stream and return small labeled windows of specific time (i.e., seconds) that can be used as input features of classification models. The script is available online and it can be easily applied in the N-Back test or Heat-the-Chair EEG data, without any changes.

Chapter 4

Case study 1: Mental workload assessment in flight scenarios

A fundamental aspect of multiple task management is to attend to new stimuli and integrate associated task requirements into an ongoing task set; that is, to engage in interruption management [70]. Interruptions often negatively affect human performance. Specifically, most laboratory and applied experiments demonstrate that interruptions increase post-interruption performance times [44] and error rates [87], increase perceived mental workload [65] , and motivate compensatory behavior [25].

The commercial flightdeck is a naturally multi-tasking work environment, in which interruptions are frequent and adopt various forms. Further, interruptions have been cited as a main contributing factor in many aviation incident reports. External and aircraft events, as well as interactions with other operators, compete for pilots' attention and require pilots to integrate performance requirements associated with these unexpected prompts with ongoing flightdeck tasks.

In this chapter, we apply the different architectures for EEG channel fusion, explained in chapter 2 to the recognition of mental workload (WL). Specifically, models are trained and tested on several serious games we also have assessed their transfer tasks capability in flight simulation data.

The goal of this use case is to characterize WL of flying pilots in the cockpit from the analysis of EEG signals. Results show that between the two approaches for channel fusion, projecting convolutional feature channels achieves higher performance, with 76.25% of sensitivity and 87.81% specificity in WL detection in n-back-test leave-one-out subject evaluation and good task transfer with the detected WL increasing with the number of interruptions.

The remaining of this chapter is organized as follows. Section 4.1 explains the pipeline for the assessment of workload, it includes data preprocessing, the classification neural architectures used to recognise the different levels of workload, and the

postprocessing of output predictions. Section 4.2 presents the experimental design. Section 5.3 presents the experimental results and discussion. Finally, conclusions and further studies are summarized in Chapter 6.

4.1 Strategy for workload assessment

Our strategy for the assessment of workload has three mains steps as illustrated in Figure 4.1. In the first step, the EEG signals are acquired from the EEG dataset for workload prediction described in Chapter 3. Next, EEG raw input data is preprocessed to obtain the input data to feed the models. Different multi-class classification problems are defined using the fusion approaches described in Chapter 2. Then, for classification, we present our models, able to recognize between different levels of workload. The output of the multi-class models are then binarized to recognize between be workload vs. idle information.

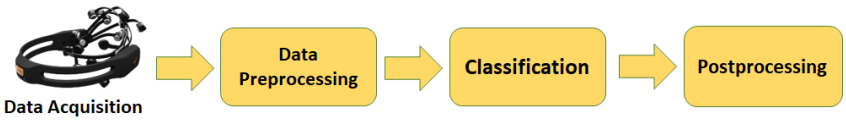


Figure 4.1: Workload assessment pipeline.

4.1.1 Data preprocessing

For EEG recording, an EMOTIV EPOC X headset [39] has been used, which has 14 electrodes placed according to the 10/20 international system. This sensor provides both raw data and power spectrum for the main brain frequencies (θ, α, β_{low}, β_{high}, and γ). Given that proposed n-back tasks are memory demanding stressing games and baseline phases consist in watching a relaxing video, the theta wave [2] is the best candidate for discriminating the different mental loads of our experimental phases. In this work, we used the power spectrum of theta wave (4–8 Hz) sampled at 8 Hz.

Eye blinking and sudden head movements introduce abrupt sharp peaks of large amplitude in the power spectra wave that should be filtered before using them as predictors of a mental state [134]. In particular, we use an Inter Quartile Range (IQR) [136] filtering strategy to detect outlier values associated to muscular movement wave peaks. Our IQR filtering is based on setting the value of the 99% percentile of the distribution to all points above it.

To ensure a high quality of signals, we further filter data according to the quality of the EEG during recordings provided by the headset itself. For each sensor and recorded sample, Emotiv reports the quality of the recording in a discrete scale with values in

the range 0—4 indicating how good the contact between sensor and head is: 4 for optimal—0 for none. For the sake of data with the highest possible quality while keeping a reasonable sample size, signals with a 25% of bad recordings were discarded (< 3). Further, since there is no evidence about what are the most discriminating sensors that best correlate with the detection of mental workload, the whole phase was discarded if the signals of two or more of the sensors were low quality. Finally, a subject was discarded if either all its base line or its workload phases were discarded, since, in this case, there were not enough data to define the binary classification. After this quality filtering, only 16 of the 20 subjects were selected for models training and testing.

In order to feed data to models, θ signals were cut in temporal windows. Notice that the size and overlap of the temporal windows might be a critical issue in order to properly include workload peaks [53]. For that we have used several window widths with different overlaps, obtaining the best results with 40 s windows overlapped 30 s. Thus, the input data of the networks were the concatenations of 40 s windows for the 14 EEG sensors ($14 \times 40 = 560$-dimensional feature space). In order to account for the difference in units and magnitudes, input data were standardized using the mean and standard deviation of the training set.

Following the approaches described in Chapter 2, we propose two architectures that differ in the moment when EEG sensor signals (channels) are projected: one projects input EEG sensors (input projector model) and the other one projects the convolutional features extracted from each EEG sensor (feature projector model). Each model has one input unit projecting EEG channels (if applicable), a convolutional unit equal for both models, and an output unit projecting the convolutional features extracted from each EEG sensor (if applicable). This output unit has a fully connected layer with Softmax activation and output the number of classes. To account for different window lengths, we apply an average pooling before the classification layer. All convolutional layers use kernels of size 3 and stride 1 and have Relu activation.

The convolutional unit has 3 blocks consisting of one convolutional layer with max pooling and having 16, 32, and 64 neurons for each convolutional layer, respectively. The classification layer has 256 neurons. For the input projector model, the projection unit has one convolutional layer with 16 neurons. For the feature projector model, the output unit has 2 blocks consisting of one convolutional layer before the classification layer. The first one has 64 neurons, the second one projects convolutional features also using 64 neurons.

Figure 4.2 shows the scheme of the architecture that combines channels at input level, while Figure 4.3 presents the scheme of the architecture that fuse channels at output feature level.

Although our main problem is a binary one, to ensure generalization capabilities of the classifier (including task transfer), we increased the diversity of the classifier by increasing the number of classes used to train the network. That is, our architecture was trained as a classifier to discriminate between a BL and WL classes using 4 different grouping of the data recorded from the 3 n-back tests:

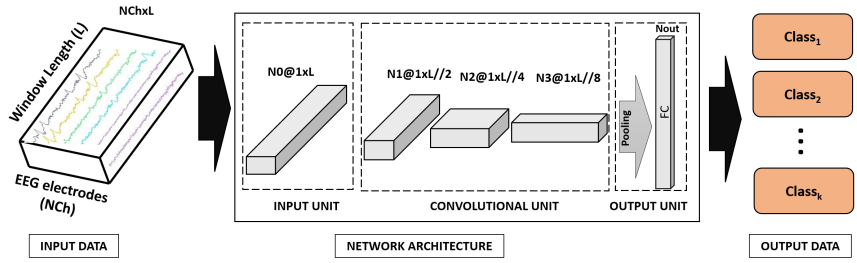


Figure 4.2: Architecture of the Input Projector Model

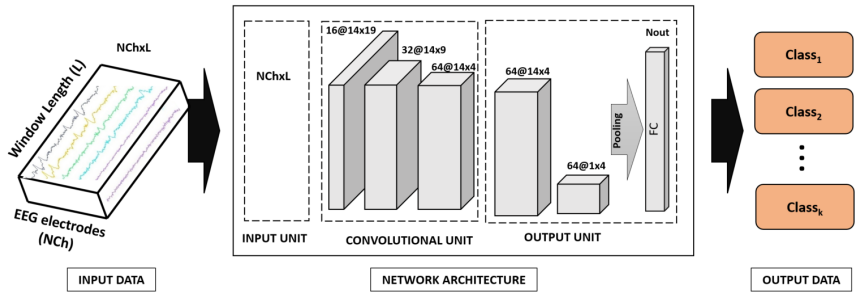


Figure 4.3: Architecture of the Feature Projector Model

1. **Binary problem** (noted BLs-WL2) given by BL = (BL1, BL2, BL3) and WL2. That is, the BL class is defined by aggregating the baselines for the 3 games and WL class defined by the workload phase of the second experiment.

2. **Three class problem 1** (noted BLs-WL2-WL3) given by BL = (BL1, BL2, BL3), WL2 and WL3. That is, a BL class defined as before and two WL classes given by the workload phase of the second and third experiments.

3. **Three class problem 2** (noted WL1-WL2-WL3) given by WL1, WL2 and WL3. That is, a BL class defined by the workload phase of the first experiment and two WL classes given by the phase 2 of the second and third experiments.

4. **Four class problem** (noted BLs-WL1-WL2-WL3) given by BL = (BL1, BL2, BL3), WL1, WL2 and WL3. That is, a BL class defined as in the first configuration and also defined by the workload phase of the first experiment and two WL classes given by the workload phase of the second and third experiments.

Unlike binary problems, in multiclass settings, the classifier does not predict the probability of belonging to each class. It rather gives a score of belongingness. It follows that the class predicted is not the one having a score above 0.5 (as is the case in binary problems), but the one having the largest value of the score predicted by the classifier. In our case, since the final class prediction is binary, we compute the binary class labels

in the multiclass settings by binarizing first the output probabilities and then taking the maximum between the two as the final class label. The transformation between classifier output and BL-WL classes scores is as follows:

1. **BLs-WL2-WL3**: The probability of BL is directly the probability of the train BL class, whereas the probability of the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.

2. **WL1-WL2-WL3**: The probability of the class BL is given the probability of the class WL1, whereas for the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.

3. **BLs-WL1-WL2-WL3**: The probability of the class BL is the maximum probability of the BL and WL1 classes, whereas for the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.

4.2 Experimental design

In order to validate the proposed models, two experiments have been conducted:

4.2.1 Training and validation using n-back-test data.

To assess to what extent a model trained over a set of individuals can successfully predict a new unseen individual, we have used a generalist population model, where a single model using all subjects was trained to assess whether inter subject variability can be properly modeled. The validation of the capability for modeling a population was tested using a leave-one-out scheme to allow statistical analysis. Models were trained using a batch size of 750, a weighted cross-entropy loss to compensate unbalances between baseline and workload phases, Adam as optimization method, 100 epochs, and a learning rate of 0.0001.

The performances of the different approaches for detection of mental workload were assessed using the accuracy (or sensitivity) for each class:

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP = number of true positives and FN = number of false negatives. Sensitivity measures the ability of the system to detect BL and WL classes. Since we have a binary classification problem with WL the positive class, the sensitivity for BL corresponds to the specificity of the model.

4.2.2 Task transfer verification using flight simulator data.

To assess the capability of our model for transfer learning, experiments were devoted to showing that the model trained to detect WL in a memory demanding task (n-back test) can detect an increase of WL associated with multitask procedures with interruptions decreasing performance.

The EEG signals of the flight dataset presented in Chapter 3 are intended to assess:

1. Correlation of WL recognition with the number of tasks carried out by the pilot. Since we expected that the proportion of samples classified by our model as medium–high WL would be higher in the intervals where the PF performed more tasks, we show the percentages of predictions for BLs and each WL in correspondence with the number of tasks demanded.

2. Correlation of WL recognition to flight complexity. Flights 2 and 4 were designed to have higher workloads than Flight 3 (Flight 1 is considered the baseline) so that the hypothesis is that the proportion of samples classified by the model as medium–high WL will be higher than in flight 3.

4.3 Results and discussion

In this section, we show and discuss the results obtained.

4.3.1 Training and validation using n-back-test data

Tables 4.1, 4.2 and 4.3 summarize the recalls of baseline (BL) and workload (WL2) for the binarized models trained on different class problems for, respectively, the input projector, the feature one and EEGNet models. Tables show ranges for WL and BL detection computed for the 16 subjects and also removing 3 outlying cases (80% of population) that all approaches fail to correctly predict.

The comparison of ranges for the two proposed approaches shows that performance is more robust for the three-class problem, although specificity is better in the 2-class and 4-class problems. Regarding projection approaches, models projecting features achieve higher performance. In particular the binary class feature projector model achieved an average detection of BL of 87.81% and WL of 76.25% for all subjects and average detection of BL of 86.65% and WL of 82.73% for 80% of the population.

Regarding EEGNet, its performance for the whole population is biased towards BL detection with a substantially low detection of WL for all classification problems, except the 3-class problem WL1_WL2_WL3, in which it achieves better WL detection. Performance in 80% of the population is more balanced between BL and WL detection, being the binary problem the best performer.

Table 4.1: Input projector model binarized

		All population	80% of population
BL-WL2	BL	85.72 ± 7.52	84.15 ± 7.50
	WL	76.22 ± 17.64	82.81 ± 11.73
BLs-WL2-WL3	BL	78.16 ± 10.83	75.5 ± 10.29
	WL	78.62 ± 16.59	84.35 ± 10.87
WL1-WL2-WL3	BL	72.94 ± 18.08	70.58 ± 19.29
	WL	77.34 ± 16.72	82.85 ± 11.48
BLs-WL1-WL2-WL3	BL	80.75 ± 9.87	79.42 ± 10.07
	WL	76.44 ± 16.81	80.96 ± 13.16

Table 4.2: Feature projector model binarized

		All population	80% of population
BL-WL2	BL	87.81 ± 7.07	86.65 ± 7.33
	WL	76.25 ± 19.27	82.73 ± 14.85
BLs-WL2-WL3	BL	79.00 ± 9.22	77.11 ± 9.13
	WL	80.94 ± 16.21	85.96 ± 11.68
WL1-WL2-WL3	BL	81.34 ± 15.76	81.27 ± 15.21
	WL	82.47 ± 15.78	86.54 ± 11.81
BLs-WL1-WL2-WL3	BL	84.75 ± 8.88	83.96 ± 9.24
	WL	76.34 ± 15.78	80.65 ± 12.49

Table 4.3: EEGNet model binarized

		All population	80% of population
BL-WL2	BL	84.44 ± 15.91	81.92 ± 16.60
	WL	67.63 ± 39.30	81.69 ± 27.71
BLs-WL2-WL3	BL	86.22 ± 11.72	84.58 ± 12.47
	WL	58.78 ± 18.99	64.73 ± 15.55
WL1-WL2-WL3	BL	62.63 ± 17.10	62.73 ± 17.99
	WL	70.22 ± 19.98	76.27 ± 15.16
BLs-WL1-WL2-WL3	BL	80.72 ± 7.87	79.54 ± 8.19
	WL	71.97 ± 17.30	78.12 ± 11.27

4.3.2 Task transfer verification using flight simulator data

Barplots in figures 4.4 and 4.5 show the percentages of WL detection as a function of the number of interruptions (0, 1, or 2). The expected pattern was the percentage of WL detection increasing with the number of interruptions. For both projection models, the 3-class problem WL1_WL2_WL3 is the only model that does not follow the expected increasing pattern. For the remaining problems, both architectures seem to behave equally.

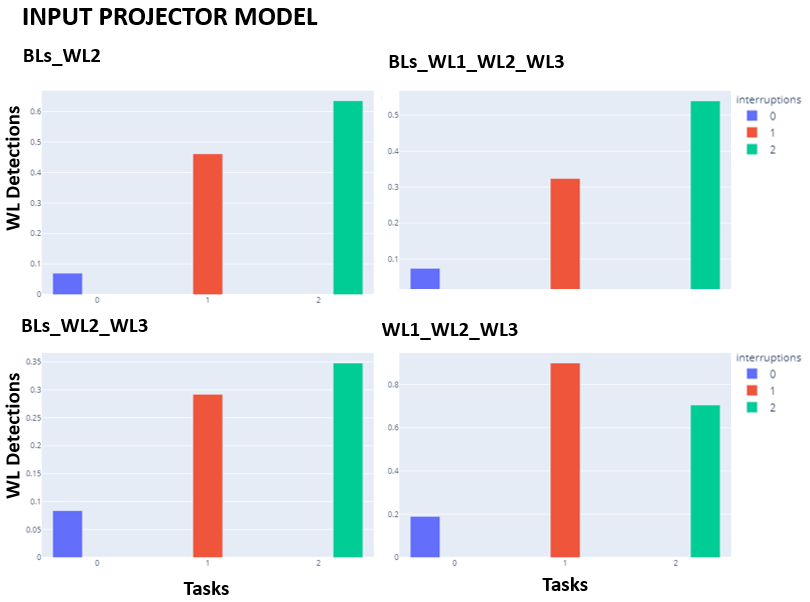


Figure 4.4: FRAM tasks barplots of WL predictions for the Input Projector model

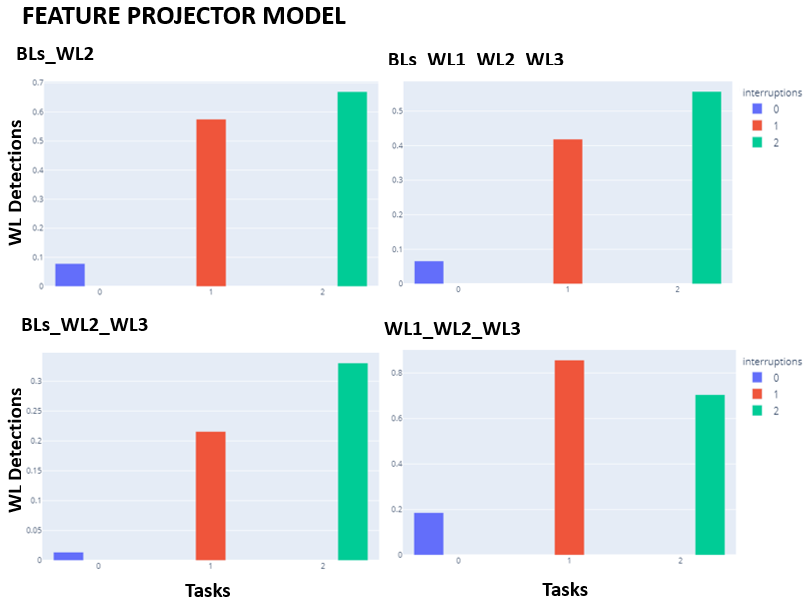


Figure 4.5: FRAM tasks barplots of WL predictions for the Feature Projector model

Figures 4.6 and 4.7 show the barplots for the number of BL and WL predictions for the four flights. The expected pattern would be to have the highest number of detection for Flight 1, Flight 2 and Flight 4 with similar amount of detected WL and Flight 3 presenting a decrease in detected WL with respect these flights. Only the feature projector model follows the pattern expected. The most significant differences between flights are evident in the 3-class problem BLs_WL2_WL3, followed by the 4-class problem. The 3-class problem WL1_WL2_WL3 does not apparently detect any difference among Flight 3 and Flight 4.

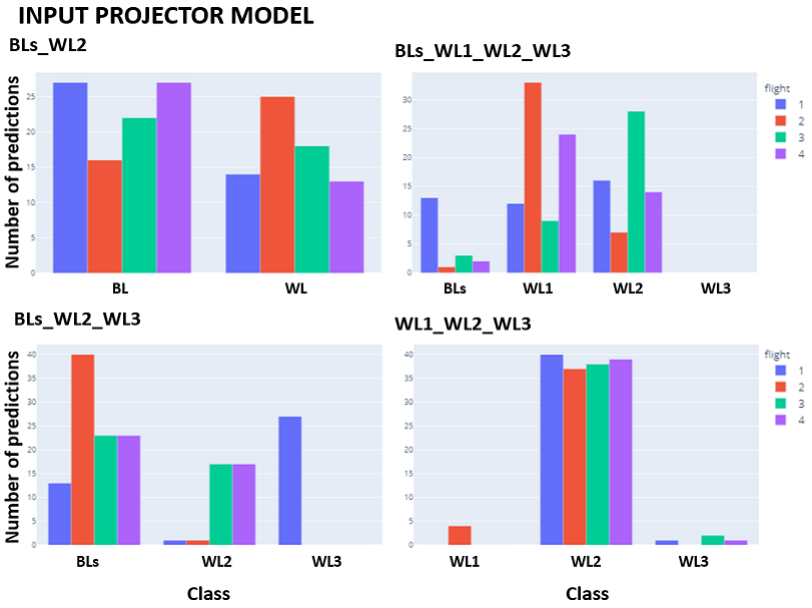


Figure 4.6: Flight test barplots of WL predictions for the Input Projector model

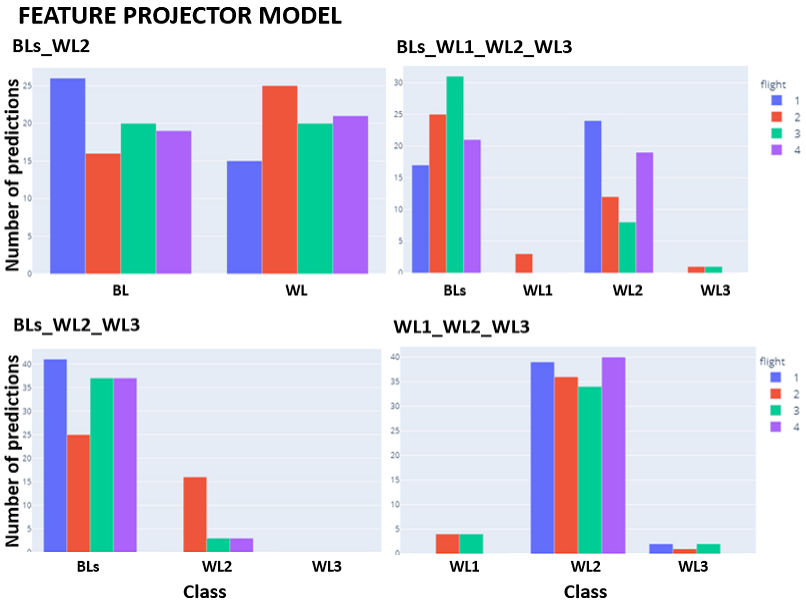


Figure 4.7: Flight test barplots of WL predictions for the Feature Projector model

Chapter 5

Case study 2: Epileptic seizure detection in pediatric patients

Epilepsy is one of the most severe neurological disorders that affects the normal functionality and activity of the brain [24]. Without a definite cause, an abnormal electrical activity of neurons starts provoking seizures, which are the observable manifestation of the epilepsy [135]. Seizures consist of subtle to strong convulsions, with movements of body, arms, and hands, and often accompanied with loss of consciousness, fainting and salivation [80]. When seizures become recurrent with certain occurrences within a day, then the quality of life of patients dramatically deteriorate, affecting their character and suffering from stigmatization of society [33, 15].

Epilepsy has no age, gender, racial group, social status, or geographic preference, and it affects both the children and adults [23]. Although there is still no cure for epilepsy, investigations are ongoing to find out what causes epilepsy, how seizure fires, and new medical treatments [32]. On the other hand, a recent estimation shows that epilepsy is affecting to more than 50 million of people around the world, and 80% of them live in poor or developing countries, who cannot access to medical treatment due to their low incomes [40]. Therefore, epilepsy has turned into a global public health problem of high social impact [140, 26]. However, not all seizures are epilepsy and there are seizures as a consequence of other diseases (e.g., Alzheimer disorder [131], stroke [85], diabetes [149], Psychogenic non-epileptic seizures [17]). Physicians and neurologists emphasize the importance of a proper diagnosis of seizures to provide the proper treatment [7], since a mistreatment of epilepsy could even worsen the disease [117].

Due to the high social impact of epilepsy in the daily life of patients, in the last decades, the research community has spent a lot of effort in researching and developing new automatic systems for epilepsy detection based on electroencephalogram (EEG). Thereby, current methods to detect and recognize epilepsy leverage recent advances in machine learning (ML) and deep learning (DL) algorithms [127, 113]. The

performance of systems is highly correlated to the ability of methods to extract discriminative features from EEG signals capable to classify those windows of signals between seizure and non-seizure. While traditional ML methods use hand-engineered features designed by an expert, DL approaches have the ability to automatically learn a set of better and rich discriminative features, which outperform the ML counterpart [112].Thus, seizure detection can be faced as a binary classification problem in a supervised learning fashion [84, 155].

One of the problems is that, since in daily life of patients, seizures last only a few seconds and most of the time are non-seizure stages, there is a high unbalance between the amount of non-seizure and seizure EEG data, hindering the training of a classifier [47, 154]. Researchers tackle imbalanced datasets by either selecting a reduced subset of the majority class, augmenting the minority class or combining both methods. That is, several studies carefully select the whole EEG seizure part and randomly select the same number of EEG non-seizure segments to construct a balanced dataset [155, 1, 27]. Other studies perform data augmentation to increase the number of seizures before making the selection of the non-seizure class. The trivial method of just repeating the same sample to increase the number of samples is useless. Instead, sliding a window with overlapping has provided greater results and has become the most used strategy in EEG data [138, 107, 132]. Another alternative is to develop a specific model to generate new samples [47]; however, generative models are still hard to train in the context of EEG [18]. While many approaches have reported high recognition rates, most results are not easily repeatable due to the freedom of researchers in selecting data.

Also, it is worthwhile to mention that in addition to the ictal stage (seizure) and interictal (non-seizure), there are two additional phases of epilepsy [8, 24]: the preictal that consists of 60–120 min before the seizure onset and the postictal, of variable duration, where the patient is recovering from his last seizure. Figure 5.1 shows the four epilepsy stages. The signal has 60 sec and comes from the same EEG electrode. Figure 5.1.a shows the ictal phase (seizure) shaded in red. The normal stage (non-seizure) is blue shaded. Figure 5.1.b shows an interictal segment extracted two hours before the seizure onset. Figure 5.1.c shows a preictal segment taken from 20 min before the seizure onset. Finally, Figure 5.1.d shows a postictal segment immediately after the end of the seizure.

Due to the freedom to select data, some studies development automatic systems to diagnose epilepsy by distinguishing preictal vs. interictal data [126, 128, 124] or ictal vs. interictal, preictal vs ictal, or ictal vs. non-ictal data. Despite the different choices of data selection, the criterion to classify a given signal into SEIZURES vs. NON-SEIZURES, or ABNORMAL vs. NORMAL seems the best option [110, 138, 132, 27]. However, if there is no consensus among physicians, there is even less consensus among scientists and there are no agreed datasets taking into account these intermediate stages.

The traditional pipeline of seizure detection based on ML/DL methods requires an input data (x_i) that is processed in order to predict its label ($\hat{y}_i$). The high temporal resolution of EEG enforces signals to be segmented into processable time windows,

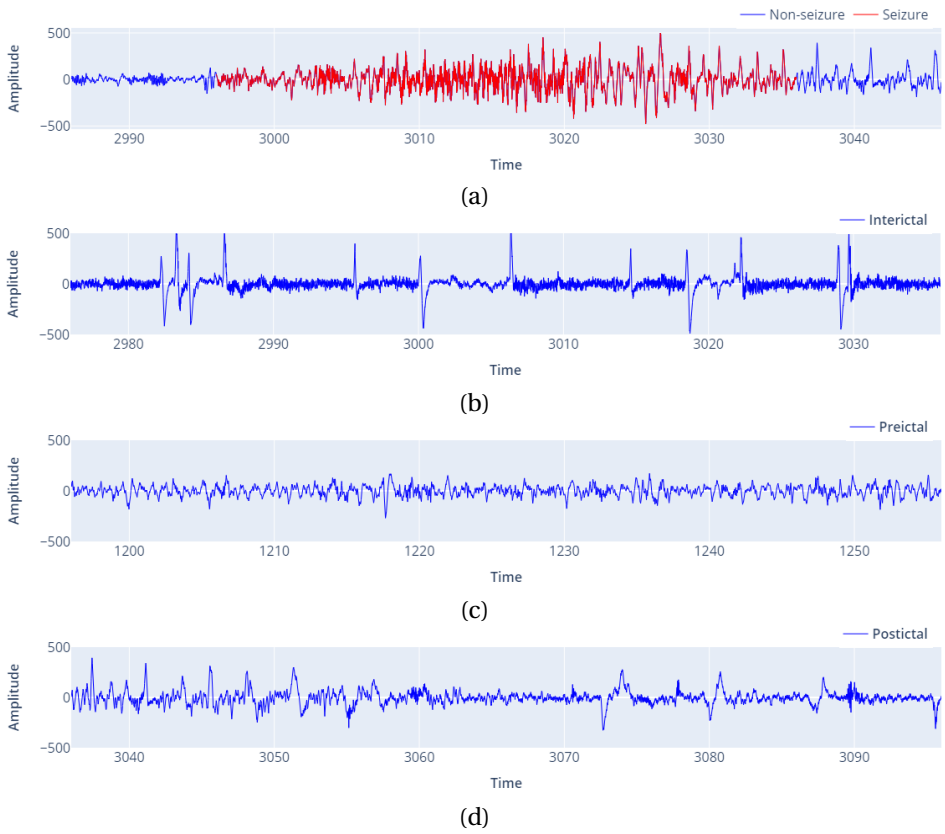


Figure 5.1: EEG signals from an epileptic episode (60 sec): (a) The seizure or ictal stage is shaded in red. (b) An interictal segment 2 hours away of the seizure. (c) A preictal segment 30 minutes before the seizure onset. (d) A postictal segment just after the seizure.

whereas its spatial resolution offers a freedom to deal with the channels.

According to the input data domain, seizure detection methods can be categorized in three main categories: time-domain, frequency-domain, and time-frequency domain. Time-domain approaches use the data that comes directly from the EEG. Frequency-domain approaches transform the time-domain data into frequency-domain data before to use it. Common new input data are the power spectrum, the 2D-spectrogram, and the main brain rhythms $(\delta, \theta, \alpha, \beta, \gamma)$ bands. Frequency-domain approaches use a mathematical transform such as the Fast Fourier Transform (FFT). And time-frequency approaches uses the Discrete Wavelet Transform (DWT). Nevertheless, using any domain of data, do not avoid of some preprocessing steps to improve the quality of data, such as the band-pass filtering and noise removal for data cleaning [127].

Also, taking in account the data source in which the proposed methods are assessed, seizure detection methods can be grouped in three main categories: k-fold CV (cross validation), patient-specific (also named intra-patient), and population (also named across-patient). The first combines and mixes the data from all patients before selecting the test set, so the test set holds the data distribution of the population. The second selects the data of a specific patient and it is used for both training and test the model. The third tries to learn a more general model in which the test set consists of a patient whose data distribution is totally unknown for the model. Finally, for evaluation, the traditional k-fold cross validation can be used directly in the k-fold CV. For patient-specific and population schemes, the leave-one-out with a slight variation is applied. The results obtained are generally reported in the form of average percentages.

As follows, a general summary of recent research advances in seizure detection using ML/DL methods in the CHB-MIT scalp dataset is presented [111]. We reduced the scope to ML/DL because these techniques provide and established the best performances and state of the art. The selected dataset is quite used in seizure detection and prediction due to its public availability and relative large size.

The seminal work of [110] presented the CHB-MIT database [111] to detect seizures. The authors use EEG time window of 6 sec to extract spatial and spectral features, together with non-EEG features, which are used as feature vector to train a support vector machine (SVM). The method was able to detect 96% of 173 seizures in a patient-specific evaluation.

Later, the study of [155] proposed to use convolutional neural networks (CNN) for seizure detection in a patient-specific scheme. First, the model consists of a single 2D convolution, an activation function, a 2D max-pooling, and a fully connected (FC) layer to combine features before classification. Next, the input data is carefully selected from the epilepsy interictal and ictal stages and next they were split in segments of 1 sec. Then, the model was trained using two data sources, separately: the time-domain EEG data and the frequency-domain 2D-spectrogram performed by the FFT. Evaluation in the CHB-MIT dataset [111], the authors found that frequency-domain increases significantly the accuracy compared to the time-domain counterpart: from an overall sensitivity and specificity of 61.2% and 63.3% to 96.9% and 98.1%, respectively. However, because the model is too simple, the high increasing could be caused by the data selection step instead of transformation itself. It could be the reason that recent approaches reside in time-domain data but increasing the model complexity. Finally, no information are provided about the selection of data neither the imbalance facing.

The study of [107] proposed a population approach for seizure detection by classifying preictal versus ictal stages. As input data, they split EEG time-domain data into segments of 2 sec, increasing the number of samples by sliding window with 80% of overlapping. The model consists of four-block CNN-based model. Each block contains a convolution, an activation function, and a max-pooling operation; except the first block that performs a convolution along time, and another convolution along channels, so the model performs one-dimensional convolution. Evaluation in the CHB-MIT

dataset [111], the method achieved an overall accuracy of 98.05%, and a sensitivity of 90%, and a specificity of 91.65%. The model achieved high performance sensitivity and establishing the state of the art related similar works [148, 138]. However, the data selection is unclear and the set of seizures is unknown.

Next, the study of [48] proposed to classify 2D-image spectrogram in a transfer learning scheme. The authors stated a multi-class classification problem using four data sources: the interictal data from two-more hours away from the seizure, the preictal I 30 min before the seizure, the preictal II 10 min before the seizure, and the ictal. First, based on the seizure length criterion, they selected 11 patients whose seizure are greater than 10 sec in the CHB-MIT dataset [111]. Next, EEG data was segmented into segments of 4 sec, which are denoised by a DWT. Finally, segments are converted into 2D-spectrogram using the FFT. To mitigate the unbalance of ictal segments, they are augmented by means of a sliding window with 50% of overlapping. The proposed model consists of three pretrained models in image classification tasks (Inception-ResNet-v2, Inception-v3, and ResNet152), whose forthcoming features are concatenated into a single one vector and feed into two FC layers of 1,024 and 512 neurons before classification. The assessment is performed in a hold-out CV, ratio 70:30, reporting a sensitivity of 92.6% and a specificity of 97.1% in detecting ictal. However, the performance is very optimistic due to data from patients are combined before splitting into the training and test set.

Then, the study of [1] proposed epilepsy detection by means of classification of interictal vs ictal stages using a 2D-CNN autoencoder (AE). First, based on the age criterion, 16 subjects were selected from the CHB-MIT dataset [111]. Next, the all dataset is standardized at once before to split it into non-overlapped segments of 1, 2, and 4 sec. To build a balanced dataset, the interictal is sub sampled randomly. The AE consists of four 2D-CNN layers that process data in both dimensions. The model simultaneously performs both the signal reconstruction and the signal classification. The learned latent vector is used as input feature for a bidirectional long short term memory (bi-LSTM) network, whose outputs are finally classified. Assessment in a 10-fold CV, the authors reported 98.72±0.77% of sensitivity and 98.86±0.53% of specificity, the best metrics in window of 4 sec. However, the high performance is too optimistic by two-fold. First the standardization of data is performed before CV. Second, CV can combine data form a patient in many folds.

Finally, the study of [27] proposed to detect epilepsy by classifying preictal and ictal data. They use a simple LSTM network, followed by a FC layer. In a hold-out CV, ratio 80:20, the model achieved the highest performance ever reported, and outperforming other methods based on LSTM or variation of it, such as the double-LSTM in [126], the CNN-LSTM [144], the nested-LSTM [75], and the bi-LSTM [60]. Assessment in the CHB-MIT dataset [111], the model achieved 99.9% of sensitivity and specificity. The study found that LSTM networks can learn features regardless of the temporal and spatial dimension of the input data. However, the LSTM has a higher number of trainable parameters that require much data to train, which cannot be fulfilled just taking preictal data of the same duration as ictal.

Subsequently, the work of [124] introduced 1D-capsule networks to classify preictal and interictal of the CHB-MIT dataset [111]. First, 30 min of data is carefully selected from 13 patients: preictal taken before 30 min the seizure onset and interictal, two hours away of the seizure. Then, data is split into time segments of 1 sec. Next, a channel selection is performed in a 5-fold CV fashion, because model just processes a single channel. Assessing the proposed model, it achieved an average accuracy of 97.74% in the channel F3-C3. However, in comparison with a 1D-CNN (even with fewer trainable parameters), the 1D-CNN reported a 97.33% of accuracy that is comparable to the capsule network, so a much study could be done.

Later, the study of [132] proposed a 1D-CNN for seizure onset detection by classifying interictal and ictal stages in a patient-specific fashion. First, ictal data is carefully prepared by joining short (<10 sec) or very consecutive seizures (<20 min), while interictal data is extracted from far away of 2 hours from seizure. Next, data is segmented into 2 sec time windows and ictal samples is augmented by windows overlapping with 50%. Then, to get a balanced dataset, the majority class is randomly sampled. Afterward, the model consist of two-1D-CNN heads, each of one of three convolutional blocks. Features are concatenated and feed to two FC layers before classification. Assessment in the CHB-MIT dataset [111], they reported an average sensitivity and specificity of 88.14% and 99.54%, respectively.

Recently, the work of [154] introduced an approach based on self-organizing fuzzy logic (SOF) for seizure detection. Evaluation of the model in the CHB-MIT dataset [111], the model achieved a geometric mean (G-Mean) of 83.35% and 92.04% for population and patient-specific detection, respectively.

Despite recent progress in research, the detection of seizures still presents challenges and open questions. As follows, we summarize the main problems and limitations of existing methods for seizure recognition:

1. Majority of studies are focused on seizure classification for k-fold CV and patient-specific schemes, and just a few studies centered on population solutions. However, there is no study of the impact of seizure variability and performances in the three schemes.

2. There is no consensus of which data should be used to train a classifier for seizure detection. Data pairs ictal vs. interictal and ictal vs preictal are used. However, a real EEG medical test could not only hold a few pairs of data.

3. The seizure detection is an unbalanced classification problem. Current methods use a random selection of samples to mitigate imbalance. After data selection a new balanced (or almost balanced) dataset is created, which is used for both training and testing. However, a real EEG medical test imbalanced with a lot of normal samples and just a few seizure samples.

4. Due to the spatial and temporal nature of EEG, current approaches process EEG data by dimensions; either along the spatial dimension and next along the temporal dimension, or vice versa. Processing EEG data along the spatial dimension

implies a fusion of EEG channels. However, there is still no final study that indicates at what stage to merge channels or what method to apply.

For epilepsy detection, we apply the different approaches proposed in Chapter 2. The proposed neural architectures implement different EEG channel fusion methods and their generalization level is assessed in different scenarios achieving comparable results in the public CHB-MIT EEG database [111].

The remaining of this Chapter is organized as follows. Section 5.1 presents our approach for epileptic seizure detection. Then, Section5.2 describes the experimental design. Next, Section 5.3 presents the achieved results and their discussion. Finally, conclusions and further studies are summarized in Chapter 6.

5.1 Epileptic seizure detection

Our approach for seizure detection is depicted in Figure 5.2. In the first step, EEG signals are acquired from the CHB-MIT EEG Scalp dataset [111]. Then, preprocessing methods are applied to get the input data for classification model. Next, classification is performed by using a set of neural network models. Finally, the postprocessing of classification predictions is executed.

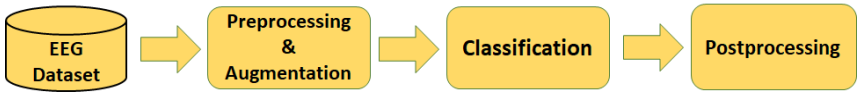


Figure 5.2: Epileptic seizure detection pipeline.

5.1.1 Dataset

In this work, the data is acquired from the public CHB-MIT dataset [111]. After a quick exploration, it has been noticed that some EEG recordings have different number of electrodes and montage, so we decided to not use them. In this way, we work with as many as possible recordings, namely, those recordings that have 23 channels. Similar selection of recordings were used in previous studies [155, 129, 1, 27, 47].

Taking into account only the recordings of 23 channels, the dataset contains around 951 hours of EEG recordings and 181 seizures. The time of all seizures totalled just 11,015 seconds, so the dataset is high unbalanced for the case of epilepsy detection. In detail, the dataset contains recordings of 23 pediatric patients, ages 1.5-22, diagnosed with intractable epilepsy [110]. Although originally there were 23 patients, the recording obtained 1.5 years later from the same patient 01 is considered as a new patient and named patient 21. Thus, the dataset has 24 patients for our experiments.

Table 5.1 summarizes the information of patients and the information of their recordings, even for the montage of 23 channels. In addition, Figure 5.3 depicts the location of electrodes according to the 10–20 positioning system and the used montage indicated by arrows between two electrodes.

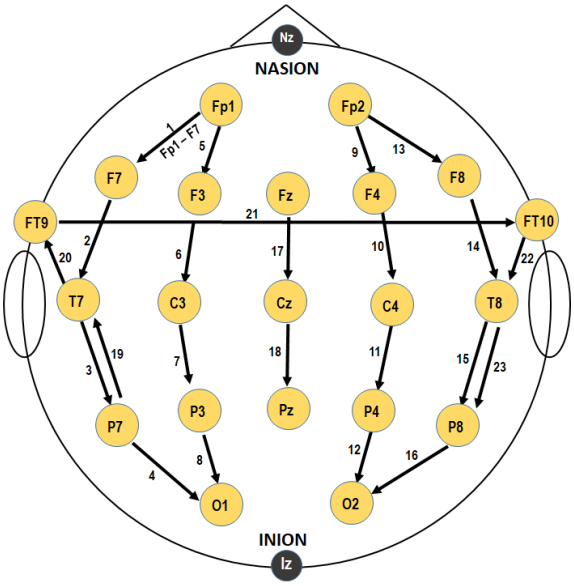


Figure 5.3: The EEG electrodes and the montage of 23 channels in the epilepsy dataset.

Moreover, after observing the timeline of recordings, it can be deduced that the collection of data was made continuously, for many consecutive hours to record the seizures experienced by the patient. The top row of Figure 5.4 illustrates the long-term EEG recording for a hypothetical patient, who have 3 seizures during the test. Seizures are red shaded, while the remaining time are yellow-green shaded. To release the dataset, continuous long-term recordings were split into one hour-long, in most cases. The dataset provides an annotation of each record, either SEIZURE-RECORD or NON-SEIZURE-RECORD depending on whether it contains epileptic seizures or not. The bottom row of Figure 5.4 illustrates the segmentation of recordings. Finally, as ground truth, the dataset furnishes the annotation of the start and end time of seizures.

5.1.2 Data preprocessing and augmentation

This work focuses on classification seizures or non-seizures in order to detect epileptic seizures, therefore a data should be selected, prepared, and labeled in order to train a binary classifier.

First, to provide repeatability of results and avoid hand-crafted data selection, a

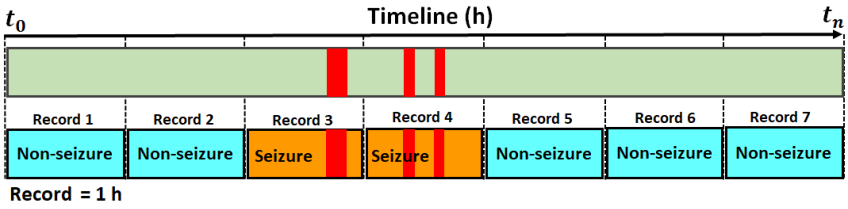


Figure 5.4: Illustrative long-term EEG recording for a patient. Epileptic seizures (red shaded) resides on SEIZURE-RECORDS, which are lemon-green shaded; whilst NON-SEIZURES-RECORDS are green shaded.

general data selection is proposed. We propose to work with recordings labeled as SEIZURE-RECORD. This approach is simple, but has two benefits. First, it reduces the amount of data allowing to work with low computational resources. Second, it preserves the richness and variability of both SEIZURE and NON-SEIZURE data. After data selection, the selected dataset contains 181 seizures and almost 185.51 (see the third column of Table 5.1).

Then, aiming to fasten the data processing, the data is down-sampled to 128 Hz, because there is no found major difference against achieved results at 256 Hz [81] and down-sampling was used previously in studies [48, 1]. Besides, no noise removal method is applied because a basic butter-worth filter has already been applied when creating the dataset [27].

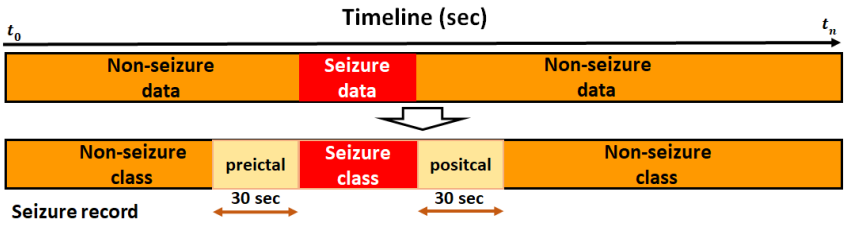


Figure 5.5: Data selection in a SEIZURE-RECORD.

Next, a visual observation of seizure boundaries indicates that signals in the transition share the same pattern for a few seconds. After a first experiment, we found that such non-seizure data in the transition hinders the classifier model. Thus, we propose to markup and exclude 30 sec before and after a seizure in order to improve the classifier and to gain insight about the behavior of signals in the frontier of seizures. The segment prior to a seizure we named preictal, while the segment posterior to a seizure we named postictal. But do not confuse these short segments with the epileptic phases of preictal and postictal, which can last longer. The 30 sec length was chosen in order

Table 5.1: Information of patients.

Patient	Gender–Age	Seizures	Seizures: Total duration (sec)	**Only SEIZURE records: Total duration (h)**	All records: Total duration (h)
01	F-11	7	442	6.65	40.55
02	M-11	3	172	2.27	35.27
03	F-14	7	402	7	38
04	M-22	4	378	10.66	156.07
05	F-7	5	558	5	39
06	F-1.5	10	153	25.89	66.74
07	F-14.5	3	325	9.04	67.05
08	M-3.5	5	919	5	20.01
09	F-10	4	276	9.58	67.87
10	M-3	7	447	14.02	50.02
11	F-12	3	806	2.79	34.79
12	F-2	27	989	9.68	20.69
13	F-3	10	440	7	11
14	F-9	8	169	7	26
15	M-16	20	1992	14.01	39.01
16	F-7	8	69	5	17
17	F-12	3	293	3.01	20.01
18	F-18	6	317	5.63	34.63
19	F-19	3	236	2.93	28.93
20	F-6	8	294	5.57	27.6
21	F-13	4	199	3.83	32.83
22	F-9	3	204	3	31
23	F-6	7	424	8.96	26.56
24	-	16	511	12	21.3
Total		181	11015	**185.51**	951.93

to avoid overlapping of such segments in two consecutive seizures. An illustration of data selection in a single EEG record is shown in Figure 5.5. The top row depicts a single seizure segment (red shaded), while the remaining data is considered as non-seizure data.The bottom row depicts the classes after marking the preictal and postictal segments. All these segments are discarded.

Then, dataset is segmented into time window input data. We propose:

- Split non-seizures data into time window of 1 sec without overlap.
- Split seizures data into time window of 1 sec with 80% of overlap.

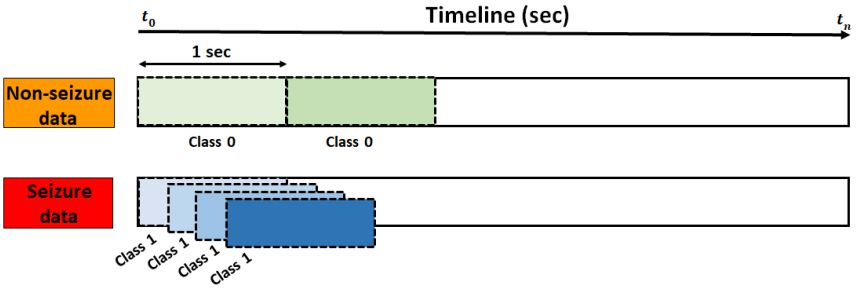


Figure 5.6: Data windowing. Non-seizure data has non-overlapping, whilst the seizure data has an overlapping. For illustrative purposes, the seizure windows were colored and slid down slightly.

The size of 1 sec is used in previous studies [155, 1, 124] and the data augmentation of seizures has result in an increasing of the classification rate [132, 127], so 1 sec time window is used in this work. Next, the ground-truth is assigned as follows: time windows that belong to the seizure are labeled as SEIZURE or Class 1; otherwise, they are labeled as NON-SEIZURE class or Class 0. Figure 5.6 illustrates the process of time windowing and ground-truth generation.

At the end, each time window is used as input data for the model. It lasts 1 sec and contains information from 23 channels. For processing purposes, it is a 2D matrix $D_{C\times T}$, where C is the number of channels and T is the sequence length. Equation 5.1 defines the EEG input data for the model.

$$\text{Let } D_{C\times T} = D_{23\times 128} \text{ an EEG input data} \tag{5.1}$$

5.1.3 Network architectures for classification

Following the approaches presented in Chapter 2, we propose two CNN network architectures for epilepsy detection. The proposed neural models are:

1. Model-1 (Early) performs channel fusion at input data level, at an early step, before feature extraction. Figure 5.7 shows the neural architecture of Model-1.
2. Model-2 (Feat) performs channel fusion at feature level, ad an intermediate step, after feature extraction. Figure 5.8 shows the neural architecture of Model-2.

The networks have the same neural components and only differ in the stage where fusion is carried out. Networks consist of three basic modules. The first module, the **Data Fusion Unit**, combines C channels into a single channel by using different data

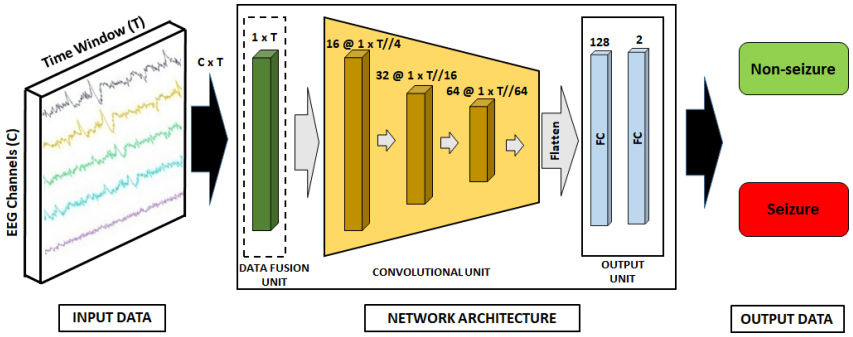


Figure 5.7: The network architecture of Model-1.

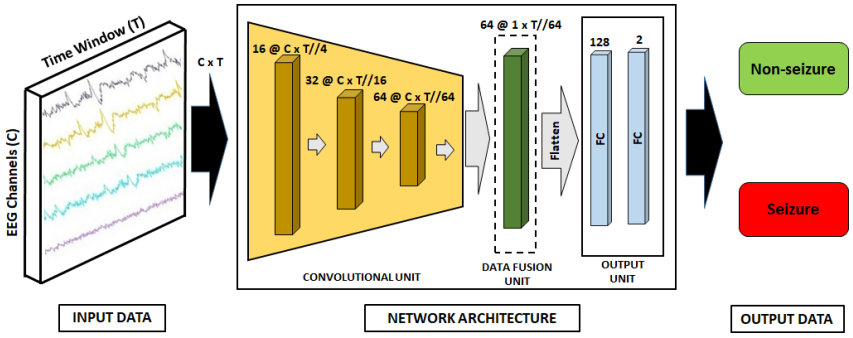


Figure 5.8: The network architecture of Model-2.

fusion methods. The four channel fusion methods are implemented accordingly to proposals of Chapter 2:

Let an input data $D_{C\times T}$:

1. The concatenation (CAT) flattens C towards a channels-wise one-dimensional vector $D_{1\times C.T}$.

2. The averaging (AVG) averages C to output $D_{1\times T}$.

3. The weighted averaging (W-AVG) performs a weighted averaging of C to output $D_{1\times T}$. The W-AVG is implemented by a convolutional layer of kernel of size 1×1 and learns C parameters.

4. The multiple weighted averaging (MW-AVG) executes a weighted averaging of C to output $D_{1\times T}$. The W-AVG is implemented by a convolutional layer of kernel

of size $C \times 1$. The name multiple is due the number of learned parameters also depends of the number of input features channels.

Next, the second module, the **Convolutional Unit**, consists of three convolutional blocks. Each block consists of a convolution layer, a Batch Normalization layer, a ReLU activation function, and a pooling operation to learn feature maps. All convolutions works along the temporal dimension T. In our implementation we used 2D convolutions, but they work along a single direction as suggested in [99].

Finally, the third module, the **Output Unit**, combines the learned features and maps them for classification purposes. This unit uses two fully connected (FC) layers. A dropout is used after the first FC layer (p=0.5), whereas the last FC predicts the output labels.

Besides, to deal alleviate imbalances during training, we use a weighted cross entropy loss function $\mathscr{L}$, whose weights are w_i for $i \in 1..c$ classes. Weights are calculated using the number of samples in the training set [89]. First, compute the inverse class frequency vector. Next, normalize between 0–1. Equation 5.2 calculates the weights w_i.

Given a binary classification problem, let n_+ and n_- the number of samples of the positive and negative class, respectively.

$$\begin{aligned} w_i &= [1/n_+, 1/n_-] \\ w_i &= w_i / \sum w_i \end{aligned} \tag{5.2}$$

In brief, there are 4 channel fusion methods that can be applied either before or after feature extraction, so 8 neural models are possible. Hence, to identify a network and the fusion that it performs, the convention network-name and fusion-name is be used (for instance, Feat CAT identifies the network that performs fusion at a feature level using the concatenation method).

5.1.4 Postprocessing

In order to improve the performance of seizure detection, a simple post-processing over predicted labels is performed. Postprocessing is feasible because the test set is a time ordered sequence, so outputs are also time ordered. We use a sliding window of length W to enhance output labels after classification, likewise a noise removal. The sliding window executes the max-voting algorithms. For seizure detection, the W length can range from 6 to 10 to detect the minimal duration of a meaningful seizure [110]). After a brief exploration of the dataset and a little experimentation, $W = 10$ gives us the highest sensitivity to detect seizures.

5.2 Experimental design

Following the approach of Chapter 2, to assess the level of generalization of models, three different experimentation levels are presented:

1. **The k-fold cross validation level**, to determine how well the model performs with as much data as possible.

2. **The population level**, to determine how well the model generalizes against an unknown patient and to indicate whether seizures share similar patterns between subjects or not.

3. **The patient-specif level**, to determine how well the model performs against an unknown seizure of the same patient and to indicate whether seizures of the same patient have commonalities or not.

As follows, a detailed description is provided:

5.2.1 The k-fold cross validation level

This experiment is designed to measure the model's performance in the all available data. Shuffling the dataset and using the traditional k-fold cross validation (CV), the models have the possibility to know the train and test sets for each running. Figure 5.9 depicts the flowchart of the k-fold CV experiment.

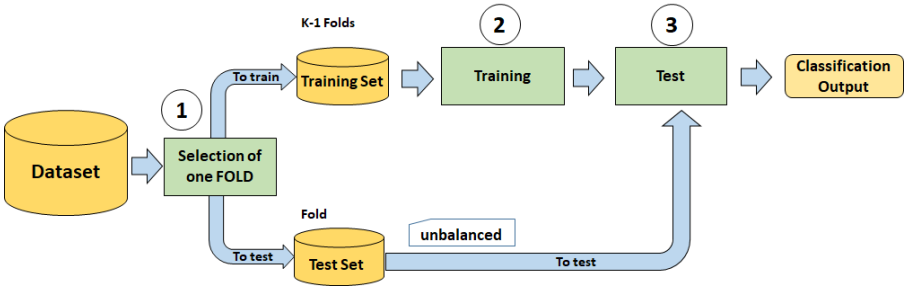


Figure 5.9: Flow chart of the k-fold cross validation experimentation level.

The k-fold CV experiment splits the dataset into k groups. Next, select a group for test set and train using the remaining k-1 groups. Repeat the process k times until the last group has been used as a test. We use k=5 and the all available data, which is unbalanced. Because data in the fold are shuffled, the samples could be unordered in time, so not postprocessing of data is applied.

5.2.2 The population level

This experiment is designed to measure the model's performance when facing a large data from an unseen patient. Moreover, this experiment can help to answer the question whether the current model could be used for seizure detection in real life applications or it still requires additional training data.

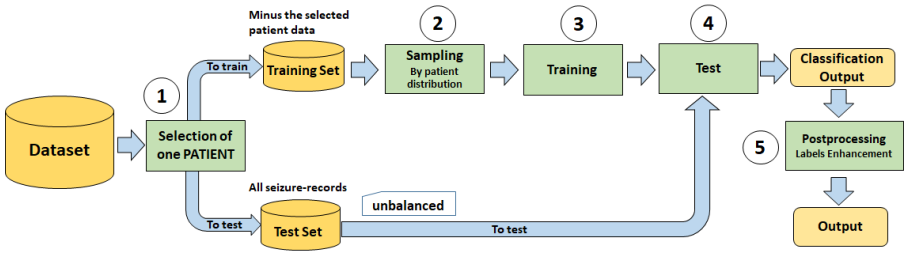


Figure 5.10: Flow chart of the population experiment level.

Figure 5.10 depicts the flowchart of the population experiment. The stages are outlined as follows:

1. Step 1: it is selected a patient for test. The data of the selected patient is separated from the dataset and composes the test set. The remaining dataset composes the training set.
2. Step 2: to reduce the computational burden due to the number of patients to be tested, a random sampling selection in the training set is used. The sampling is performed according to the data distribution of patients, so, for a given patient, the number of samples per class are equals.
3. Step 3, the training is carried out.
4. Step 4, the test data is assessed. Notice that the set is time ordered and has no augmented samples.
5. Step 5, the predicted outputs are post-processed.

5.2.3 A patient-specific level

This experiment is designed to measure the model's performance for a specific patient. The model tests a selected unseen seizure, while the rest of seizures are used as training data. The experiment provides insight about the variability of seizures in the patient.

Figure 5.11 depicts the flowchart of the patient-specific experiment. The procedure is very similar to the population experiment exposed above, except in two things. First,

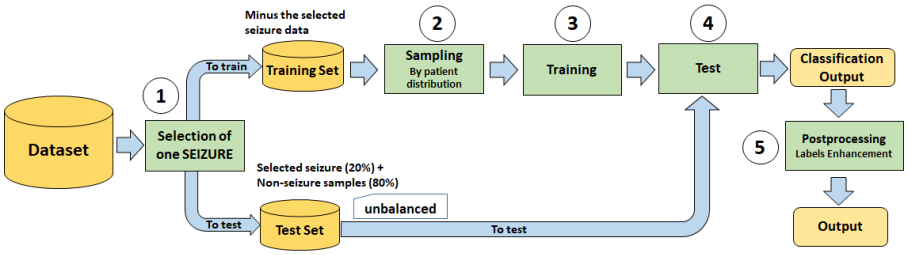


Figure 5.11: Flow chart of the patient-specific experimentation level.

instead of a patient, here a seizure of the chosen patient is selected. Second, the imbalanced test set is assembled in such a way that the 20% of the test set are the samples of the selected seizure and the following 80%, are consecutive non-seizures samples from a random non-seizure record.

5.2.4 Evaluation metrics

The seizure detection can be stated as a binary classification problem [84], in which the positive class (+) is the SEIZURE class (or 'abnormality') and the negative class (−) is the NON-SEIZURE class (or 'normality'). Once the confusion matrix has been calculated, the True positive (TP), false positive (FP), false negative (FN), and true negative (TN) are used to compute the model performance by means of:

- Recall + = TP / (TP + FN), to estimate how well the model detects seizure segments. It measures the sensitivity.
- Recall − = TN/(TN + FP), to determine how well the model detects non-seizure segments. It measures the specificity.
- Precision + = TP / (TP + FP), to evaluate the relevance of detected seizures samples.
- Precision − = TN / (TN + FN), to evaluate the relevance of detected non-seizures samples.
- F1-score = (2 * Precision * Recall) / (Precision + Recall), to estimate how well the model works in unbalanced data.
- Accuracy = (TP + TN)/(TP + FP + TN + FN)

Models were implemented in Python 3.9 environment and Pytorch 2.10 framework. Experiments were carried out in a desktop computer with Windows 10 and a NVIDIA Gforce RTX 2070 Super graphic card. Before training, the training set is standardized

by each EEG channel to guarantee uniformity of values. These scales are then applied to the test data before they are fed into the model.

5.3 Results and discussion

This section presents the results obtained from the three experiments proposed in the experimental design, together with their discussion. For the sake of understanding how the experimental setup and the exposition of data during training affect the model's performance, we present the results from the more general to the particular case.

First, the results of the k-fold CV are presented. Next, the results of the population experiment. Then, the results of the patient-specific test. Finally, the comparison of our methods against the results of three state of the art neural architectures.

5.3.1 Results of the k-fold cross validation level

Achieved results in the k-fold cross validation are summarized in Table 5.2. Note that the metric of the positive class (seizure) uses +, the negative class (non-seizure) uses −.

Table 5.2: k-fold cross validation classification results.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Early AVG	82.72±0.93	87.72±0.75	36.99±1.37	98.32±0.09	51.11±1.29	92.71±0.41	87.32±0.67
Early CAT	96.39±0.15	99.59±0.04	95.32±0.45	99.69±0.01	95.85±0.2	99.64±0.02	99.33±0.03
Early W-AVG	86.2±0.7	92.3±0.34	49.34±0.96	98.72±0.06	62.75±0.64	95.4±0.16	91.81±0.27
Early MW-AVG	85.23±0.77	93.58±0.79	53.73±2.88	98.65±0.06	65.86±1.95	96.04±0.39	92.91±0.67
Feat AVG	84.13±1.93	96.12±0.86	65.65±4.38	98.58±0.16	73.64±2.21	97.33±0.38	95.16±0.66
Feat CAT	96.22±0.37	99.6±0.02	95.39±0.21	99.67±0.03	95.8±0.14	99.63±0.01	99.32±0.02
Feat W-AVG	88.62±1.26	95.69±1.2	64.65±6.0	98.98±0.1	74.6±3.69	97.3±0.58	95.12±1.02
Feat MW-AVG	96.06±0.41	99.46±0.06	93.98±0.57	99.66±0.04	95.01±0.21	99.56±0.02	99.19±0.04

Three methods achieved the highest performance in all metrics, so they generate a great expectation for the other experiments. The selected methods are gray shaded in Table 5.2 and they consist of models that perform concatenation of input channels **Early CAT**, concatenation of feature extracted **Feat CAT**, and multiple weight averaging of features **Feat MW-AVG**. Other methods perform below our expectations and have difficulty detecting seizures (see Recall+). Also the metric F1+ score reflects the impact of imbalanced dataset.

5.3.2 Results of the population level

The performance of models against an unknown patient is summarized in Table 5.3. There are 24 patients in dataset, but the individual results were summarized by means of the mean and the standard deviation. Tanking in account Recall+, it is noticeable a high dispersion of performances. We argue that this performance's dispersion is due to the fact that seizures might vary between patients according to the type of epilepsy [43, 41].

Table 5.3: Population classification results in 100% of patients.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Early CAT	67.37±32.46	90.21±17.33	43.1±35.81	99.31±0.9	40.91±31.22	93.41±12.94	89.74±16.97
Feat CAT	68.03±30.74	90.71±17.26	45.05±36.27	99.33±0.86	42.21±30.63	93.67±13.23	90.25±16.9
Feat MW-AVG	70.02±31.58	89.19±20.75	41.45±33.55	99.34±0.97	41.14±32.24	92.25±16.47	88.8±20.22

Table 5.4: Population classification results in 80% of patients.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Early CAT	81.48±16.25	87.96±18.9	41.46±33.9	99.49±0.68	46.26±30.99	92.02±14.28	87.73±18.61
Feat CAT	81.35±15.8	88.57±18.89	42.41±33.18	99.52±0.61	47.45±30.55	92.34±14.65	88.34±18.6
Feat MW-AVG	82.59±17.0	86.82±22.8	40.53±31.03	99.53±0.72	46.73±30.89	90.64±18.25	86.68±22.32

After observation of individual results, we found that 5 patients has a high variable recall, so we decided to remove the 20% of patients (patient 14, 15, 16, 20, and 21). Here, for the sake of understanding, just the results of the best models selected in Section 5.3.1 are shown. Table 5.4 shows the performances achieved in 80% of patients. After removing hard patients, the model's performance increase noticeably. The Recall− is almost 0.83±0.17%. On the other hand, the Precision and F1-score are relative low with high variance (0.41±0.31% and 0.47±0.31%), however these values are strongly featured due to imbalance of the test set.

5.3.3 Results of the patient-specific level

For the sake to avoid useless computation time, the patient-specific experiment was performed with the three models selected in Section 5.3.1. Achieved model's performances facing an unknown seizure of a specific patient are summarized in Table 5.5. Although there 24 patients, individual results were summarized for each proposed method. In this case, 181 seizures were evaluated.

In overall, the achieved performances are high, being the method Feat Cat that provided a slightly highest metrics. Moreover, it is noticeable that Precision− and F1-score− are also quite high, and the imbalance of the set set does not seem to affect performance. We argue that this is because the training and test sets of the same patient are supposed to share the same data distribution.

Table 5.5: Patient-specific classification results in 100% of patients.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Early CAT	86.04±21.61	98.01±3.6	90.67±18.68	96.82±4.58	87.04±18.82	97.31±2.93	95.62±4.88
Feat CAT	86.8±20.68	98.1±3.39	91.61±16.95	97.0±4.41	87.81±17.3	97.45±2.7	95.84±4.5
Feat MW-AVG	85.47±21.79	97.79±3.93	90.08±18.78	96.69±4.6	86.3±18.96	97.12±2.97	95.32±4.92

On the other hand, according the standard deviation (almost 0.22) there is patients that differs in their performances. This facts indicates that seizures not always share similar patterns for the same patient[43, 41], so the classifier fails and the reported performance drops.

Table 5.6: Patient-specific classification results in 80% of patients.

Models	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Early CAT	91.32±12.34	97.89±3.76	93.15±10.62	97.93±2.82	91.36±9.25	97.84±2.33	96.58±3.61
Feat CAT	91.69±12.59	97.88±3.62	93.05±10.44	98.03±2.87	91.49±9.32	97.89±2.27	96.64±3.54
Feat MW-AVG	90.5±14.68	97.6±4.19	91.63±13.85	97.76±3.24	90.14±12.4	97.6±2.65	96.18±4.16

After observation of individual performances, patients 13, 14, 16, 17, and 20 present epilepsy recall lower than 80.00%, so we decided to remove them in order to get a more ideal model performances. Table 5.6 summarizes the patient-specific results in just 80% of patients. It is noticeable an increasing of all measurements after removing patients with variable seizures. In this case, 144 seizures were evaluated.

5.3.4 State of the art comparison

This section presents a comparison of our proposals against three state of the art neural architectures. The First, Chakrabarti et al.[27] presented a LSTM-based architecture that processes all input data without without any channel treatment to discriminate preictal vs. ictal signals in a k-fold CV scheme. The second, Gao et al.[49] proposed a neural architecture able to perform temporal followed of spatial feature learning to detect driver fatigue. Although the approach is not for epilepsy detection, it was selected because of the way they address EEG spatial channels. Finally, Hossain et al.[107], in the area of population, presented a neural architecture that combines EEG channels in the first step using a convolution that also perform feature extraction and non-linear mapping.

We present below the results of the comparison in the three proposed experimental configurations:

5.3.4.1 Comparison in the k-fold CV level

Table 5.7 summarizes the comparison of the three state of the art neural architectures versus our best methods in the k-fold CV experiment.

Table 5.7: k-fold cross validation level: comparison with the state of the art methods.

Author	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Chakrabarti et al.[27]	97.47±0.23	99.5±0.08	94.4±0.83	99.78±0.02	95.91±0.4	99.64±0.04	99.33±0.07
Gao et al.[49]	95.51±0.47	98.02±0.36	80.88±2.66	99.6±0.04	87.56±1.42	98.81±0.17	97.82±0.3
Hossain et al.[107]	88.49±1.29	93.77±1.01	55.49±3.7	98.94±0.11	68.12±2.42	96.28±0.48	93.34±0.83
Early CAT*	96.39±0.15	99.59±0.04	95.32±0.45	99.69±0.01	95.85±0.2	99.64±0.02	99.33±0.03
Feat CAT*	96.22±0.37	99.6±0.02	95.39±0.21	99.67±0.03	95.8±0.14	99.63±0.01	99.32±0.02
Feat MW-AVG*	96.06±0.41	99.46±0.06	93.98±0.57	99.66±0.04	95.01±0.21	99.56±0.02	99.19±0.04

* This work

The performance of Chakrabarti et al.[27] is similar our methods, but slightly higher in Recall+. However, our Early Cat and Feat Cat methods surpass it in Precision+. The method of Gao et al.[49] provides a near close Recall+, but its Precision+ is low. The approach of Hossain et al.[107] provides the lowest performance in all metrics.

5.3.4.2 Comparison in the population level

Table 5.8 summarizes the comparison of the three state of the art neural architectures versus our best methods models in the population experiment taking. The results have been taken in 80% of patients (19 patients of the dataset).

Table 5.8: Population level: comparison with the state of the art methods (80% of patients).

Author	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Chakrabarti et al.[27]	70.22±25.1	89.76±12.6	21.43±14.96	99.3±0.83	30.1±18.83	93.77±7.89	89.32±12.22
Gao et al.[49]	79.08±15.6	81.77±21.06	20.21±19.1	99.41±0.77	28.15±20.32	87.96±16.49	81.75±20.69
Hossain et al.[107]	76.67±21.02	86.34±15.64	19.66±14.63	99.43±0.75	28.91±19.62	91.56±10.65	86.11±15.27
Early CAT*	81.48±16.25	87.96±18.9	41.46±33.9	99.49±0.68	46.26±30.99	92.02±14.28	87.73±18.61
Feat CAT*	81.35±15.8	88.57±18.89	42.41±33.18	99.52±0.61	47.45±30.55	92.34±14.65	88.34±18.6
Feat MW-AVG*	82.59±17.0	86.82±22.8	40.53±31.03	99.53±0.72	46.73±30.89	90.64±18.25	86.68±22.32

* This work

Taking in account the Recall+, the methods of Hossain et al. and and Gao et al. achieved almost similar results, but Chakrabarti et al. got the lowest. Observing Precision+, the three methods reported similar low results. On the other hand, our proposed methods outperform them in Recall+ and doubles in Precision+.

Taking into account Recall− and Precision−, methods give high recall and precision, likewise Chakrabarti et al., who reported a higher recall than our method, but at the once a lower precision than ours. We argue that the low Recall+ and the high Recall− of Chakrabarti et al. is because the network is strongly influenced by majority class data.

5.3.4.3 Comparison in the patient-specific level

Table 5.9 summarizes the comparison of the three state of the art neural architectures versus our methods in the patient-specific experiment taking in account 80% of patients (19 patients of the dataset). For comparison, the same patients were removed from the cutting-edge models because they have poor results. It is noticeable that our methods surpass the other approaches.

Table 5.9: Patient-specific level: comparison with the state of the art methods (80% of patients).

Author	Recall+	Recall−	Precision+	Precision−	F1+	F1−	Accuracy
Chakrabarti et al.[27]	80.67±16.61	94.44±6.17	81.2±18.01	95.21±3.96	79.78±15.61	94.72±4.31	91.68±6.64
Gao et al.[49]	86.12±14.08	92.06±12.65	80.36±19.19	96.47±3.26	81.06±14.47	93.63±8.73	90.87±10.05
Hossain et al.[107]	80.13±17.85	91.1±11.83	76.44±22.23	94.96±4.19	75.88±18.34	92.53±7.56	88.91±10.02
Early CAT*	91.32±12.34	97.89±3.76	93.15±10.62	97.93±2.82	91.36±9.25	97.84±2.33	96.58±3.61
Feat CAT*	91.69±12.59	97.88±3.62	93.05±10.44	98.03±2.87	91.49±9.32	97.89±2.27	96.64±3.54
Feat MW-AVG*	90.5±14.68	97.6±4.19	91.63±13.85	97.76±3.24	90.14±12.4	97.6±2.65	96.18±4.16

* This work

5.3.5 Discussion of results

This work engages with two problems in seizure detection: the imbalanced data of long-term EEG recordings and the variability of seizures. We propose a CNN-based network architecture, which is able to generalize, push, and improve recognition of seizures by means of EEG-channel fusion strategies.

Our study analyze the capability of models in three different scenarios: k-fold CV, population, and patient-specific. However, current methods have been proposed for specific scenarios. For instance, Chakrabarti et al.[27] focus in k-fold CV, Hossasin et al.[107] engage in population, and Wang et al.[132] consider patient-specific or patient specific environment.

Moreover, current methods select EEG data in a specific way that often it is not repeatable (e.g., extracting specific seizures and discarding some of them, or selecting specifically non-seizures parts that are easily separable). For instance, Wang et al.[132] works with 145 seizures and Hossasin et al.[107] deal with 163 seizures, after discarding some seizures. Other studies discard patients, such as Abeldhameed et al.[1] that worked with 16 patients, or Gao et al.[48] that selected only 11 patients. In opposite, we propose a general EEG data selection which is simple, repeatable, and not discard any seizure. We use seizure-record with 23 EEG channels (see Table 5.1).

Taking into account the nature of the test set, most of methods use a balanced test set. In opposite, our approach uses an unbalanced test set that resembles a real life EEG data. In clinical conditions, an EEG recording might contain from minutes to hours of EEG data, often with none or only a few seconds of abnormalities. To mitigate data unbalancing, we weight the loss function according to the number of samples in the dataset (see Equation 5.2).

Finally, we propose a CNN-based neural architecture that implements 4 EEG channel fusion techniques. We found that models that perform concatenation of input channels (Early CAT) or extracted features (Feat CAT), together with the method that executes a weighted averaging of features (Feat MW-AVG), provide the best performances in the k-fold CV scheme, so they were selected for the next population and patient-specific assessment. Of the three models, Feat CAT surpasses slightly the others two, but it is not meaningful (see Table 5.2), so the three models were used in the next experiments.

In the population experiment, our approach shows a variability of performances between subjects. We argue that it mainly is because the diversity of epilepsy and their variable symptoms (see Table 5.3). Also, this low performance was perceived and reported in an early study of Tsioruis et al.[125]. In order to solve it, Hossasin et al.[107] performed a hand-crafted data selection and removal of hard seizures to report high performances. However, parsing files to select separable data might be time consuming and is not scalable. Because our proposal is based in a general data selection, we identified the hard patients and explored if the performance increases after removing them (see Table 5.4). An individual analysis of such patients might be conducted in a further study to determine if their seizures are complex or it is necessarily a more stronger noise removal.

In the patient-specific experiment, our proposal reports high performances to detect seizures (see Table 5.5). Although the variability between subjects is low, we identified who individuals have variability between their seizures itself. Five individuals show performances lower than 80.0%, which were excluded to present the Table 5.6). Among abnormal patients to be excluded, the patients 14, 16, and 20 are common for both population and patient-specific, so a further study of such individuals might be conducted.

To compare our results with the state of the art, three cutting edge neural architectures were selected (see Section 5.3.4). The networks were trained in the dataset for a fair comparison. Tables 5.7–5.9 resume the achieved performance of selected methods against our models. Chakrabarti et al.[27] and Gao et al.[49] achieved a slightly similar performance than our methods in the k-fold CV assessment, the performance of them drop in the population and patient-specific scenarios, in which our approach outperforms significantly the other methods. Our best method, the Feat CAT achieved a sensitivity of 0.96 ± 0.0 and a specificity of 1.00 ± 0.0, a sensitivity of 0.81 ± 0.16 and a specificity of 0.89 ± 0.19, and a sensitivity of 0.92 ± 0.13 and a specificity of 0.98 ± 0.04, in the k-fold CV, population, and patient-specific experiments, respectively.

Chapter 6

Conclusions and Further Lines

Deep learning methods have turned out into the key tool for the analysis of EEG neurophysiological signals for detection, diagnosis, and assessment of abnormal cognitive states. In this thesis, we have addressed the assessment of mental workload using EEG signals. Workload strongly reduces the human performance and might affect the safety of activities that demand a certain level of cognition, such as piloting and driving.

In this thesis, three main challenges for assessment of workload using EEG have been addressed:

1. **The EEG channel fusion**: Our proposal for EEG channel fusion is based on convolutional neural network. Two neural architectures were proposed and assessed. The neural models differ on the step where channel fusion is performed, either at input data level or at feature level. As well as, four channel fusion methods (simple averaging, concatenation, weighted averaging, and multi-weighted averaging) have been proposed and they are completely independent of the number of EEG electrodes and of the input size.

   Although our approach of channel fusion was designed for EEG signals, they might be used to fuse other multivariate data from other different physiological sensors like ECG, EMG, EDA. Even more, our fusion module could be directly extended for multi-source data fusion purposes, as the case of ongoing monitoring of a patient with multiple sensors in the intensive care unit of health centres.

2. **The collection of unambiguous annotated data**: Our proposal to collect unambiguous annotated datasets for cognitive states was presented and resides in the use of specialized serious games to induce a mental workload in the subject who carries out tasks that require different degree of cognition effort. Next, a new dataset was collected using the Airbus A-320 flight simulator. Two simulator environments were employed: the non-immersive simulator in a computer desktop and the immersive simulator in a cockpit. As ground truth for the serious

games, the dataset provides the theoretical difficulty of games, the game scores, and the subjective TLX reports. For the flight simulation data, it is furnished the theoretical difficulty, the self-perceived difficulty (at certain interval of time), and the estimated workload based on the FRAM agent. As well as, datasets have been validated technically for validity and reuse of data in further investigations. All cognitive data was recorded using an EEG and an ECG; the raw data was made publicly available.

3. **The levels of generalization of models**: We proposed a reliable validation protocol for comparison of models based on the their levels of generalization. Our proposal takes into account the reported performance of models, but also the mechanism of data splitting into training and test set. In this way, for testing, the source and quantity of data makes difference and the use of appropriate metrics also is recommended. Therefore, three levels of evaluation were proposed. Taking into account the ability to generalize in a new unseen data, the proposed levels are: the k-fold cross validation level, the patient-specific level, and the population level.

   The application of the proposed evaluation protocol could allow a fair comparison between models and to select the best one candidates toward to the implementation of real life applications.

Aiming to show the feasibility of our approaches, two use cases have leveraged the methods presented in this thesis achieving comparable results:

1. **Assessment of mental workload**:

   This first use case has dealt with the assessment of workload in the cockpit, where the models are trained using serious games' data and tested in the flight simulation data. Here, the models perform knowledge transfer between tasks and they continuously can sense the pilot's workload intervals during the flight.

   In detail, the models were trained and validated on self-designed games (one serious game and one flight simulator with specific scenarios) to ensure unambiguous annotations. Models were trained and validated on the serious game using the population scheme (leave-one-subject-out); and simulator data gathered from a subject not included in the training data were used to evaluate transfer capability.

   Results show that between the two architecture models, projecting convolutional feature channels achieves higher performance, with 76.25% of sensitivity and 87.81% specificity in WL detection in n-back-test evaluation and good task transfer with the detected WL increasing with the number of interruptions during the flight. By comparing our approach with a cutting-edge EEG architecture [72], our models performs better.

   Although these results provide evidence of the ability of the EEG sensor to discern between more and less demanding tasks —increasing the evidence the robustness of the EEG and its ability to transfer tasks— the fact that the 3-class

problem BLs_WL2_WL3 does not correlate with flight complexity. An improvements could consider filtering of motion artifacts by calibrating motion signals before test recording and consider the multiple aspects of a user's state when developing cognitive state detection [14].

2. **Epilepsy detection**:

   The second use case has dealt with the epileptic seizure detection in pediatric patients. Two proposed models have achieved comparable results with current methods and are very generalizable.

   In detail, two neural architectures were proposed to recognize epilepsy when carrying out the EEG multichannel fusion. As well, four channels fusion were widely explored. Results show that the architecture model that the concatenation of temporal features provides the best results in the public CHB-MIT dataset.

   In addition, the generalizability of the models were validated in three different levels: by time window (using the k-fold cross validation), by patient (using the patient validation), and by seizure episode (using the patient-specific validation). In all cases, our approach achieves promising results against three state of the art neural architectures [27, 107, 49], 96.22 ± 0.37 and 99.6 ± 0.02, 81.35 ± 15.8 and 88.57 ± 18.89, and 91.69 ± 12.59 and 97.88 ± 3.62, of sensitivity and specificity, respectively. Because our data selection of test data resembles real EEG long-term recordings and the reported sensitivity by patient is more than 80%, our approach would be feasible to implement for support clinical support in diagnosis of epilepsy.

To sum up, although the approaches presented in this thesis have been focused in assessment cognitive states using EEG, they can directly be applied and extended to recognize other mental abnormalities (like mental fatigue, drowsiness, and alertness), to medical diagnose of mental illness (like depression and anxiety) and neurodegenerative brain disorders (like dementia, Alzheimer, Parkinson). Moreover, the proposed methods can be used in different scenarios and using different continuous stream's sensors with the only purpose of continuous monitoring of anomalies/faults.

Further lines

The methods suggested in this thesis pave the way for more in-depth studies.

- **Data fusion of wave decomposition**: Although data fusion is more intuitive in multi-source data, it can also be applied to different data derived after transforming a single data source. In this way, EEG signals can be decomposed into brain rhythms (e.g., $\delta, \theta, \alpha, \beta, \gamma$) or filtered in a specific range of frequencies (e.g., 4–8 and 25-50). Depending of the application, in addition to the presented channel fusion methods, a combination of some range of frequencies could improve the performance.

- **Temporal self-attention neural architectures**: EEG signals are naturally time-dependent, so signals can be processed in many ways and paradigms of signal processing. It is still not well explored the application of temporal self-attention networks on EEG signals.
- **Brain connectivity analysis**: The spatio-temporal EEG signals can allow to build dynamic maps of the cerebrum in order to understand the flow of information between their regions. The analysis of connectivity can help to devise new methods for early detection of brain anomalies and to understand how a healthy brain works.
- **New applications**: The application of the proposed methods in fields other than cognitive states and using various physiological sensors could be assessed in future studies.

Publications

Journal papers

- Hernández-Sabaté, A., Yauri, J., Folch, P., Piera, M. À., & Gil, D. (2022). Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. Applied Sciences 2022, Vol. 12, Page 2298, 12(5), 2298. https://doi.org/10.3390/APP12052298

Conference papers

- Yauri, J., Hernández-Sabaté, A., Folch, P., & Gil, D. (2021). Mental Workload Detection Based on EEG Analysis. Frontiers in Artificial Intelligence and Applications, 339, 268–277. https://doi.org/10.3233/FAIA210144

Open dataset

- Yauri Vidalón, J. E., Folch, P., Álvarez, D., Gil, D., & Hernàndez-Sabaté, A. (2022). Dataset to Predict Mental Workload Based on Physiological Data. `https://doi.org/10.5565/DDD.UAB.CAT/259591`

Manuscripts submitted

- Hernández-Sabaté, A., Yauri, J., Folch, P., Alvarez, D., & Gil, D. EEG/ECG datasets for mental workload predictions in flightdeck environment. Journal GigaScience.
- Yauri, J., Hernández-Sabaté, A., & Gil, D. EEG channel fusion analysis for epileptic seizure detection in cross-patient and patient-specific contexts. Journal of King Saud University - Computer and Information Sciences.

Bibliography

[1] Ahmed Abdelhameed and Magdy Bayoumi. A Deep Learning Approach for Automatic Seizure Detection in Children With Epilepsy. *Frontiers in Computational Neuroscience*, 15:29, 4 2021.

[2] Richard J. Addante, Andrew J. Watrous, Andrew P. Yonelinas, Arne D. Ekstrom, and Charan Ranganath. Prestimulus Theta Activity Predicts Correct Source Memory Retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10702–10707, 6 2011.

[3] Muneeb Imtiaz Ahmad, Ingo Keller, David A. Robb, and Katrin S. Lohan. A Framework to Estimate Cognitive Load Using Physiological Data. *Personal and Ubiquitous Computing*, pages 1–15, 9 2020.

[4] Amirmasoud Ahmadi, Hanieh Bazregarzadeh, and Kamran Kazemi. Automated Detection of Driver Fatigue from Electroencephalography through Wavelet-based Connectivity. *Biocybernetics and Biomedical Engineering*, 9 2020.

[5] Shafira Karamina Alifah, Pande Bagus Widyantara, and Maya Arlini Puspasari. The Effect of Mental Workload Towards Mental Fatigue on Customer Care Agent using Electroencephalogram. *ACM International Conference Proceeding Series*, pages 173–177, 9 2019.

[6] David P. Allen and Colum D. MacKinnon. Time-frequency Analysis of Movement-Related Spectral Power in EEG During Repetitive Movements: A Comparison of Methods. *Journal of Neuroscience Methods*, 186(1):107–115, 1 2010.

[7] Ushtar Amin and Selim R. Benbadis. The Role of EEG in the Erroneous Diagnosis of Epilepsy. *Journal of Clinical Neurophysiology*, 36(4):294–297, 7 2019.

[8] Michael J. Aminoff and Robert B. Daroff. *Encyclopedia of the Neurological Sciences*. Academic Press, 2nd edition, 2014.

[9] Ralph G. Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E. Elger. Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence

on Recording Region and Brain State. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64(6 Pt 1):8, 2001.

[10] Ralph G. Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E. Elger. The Bonn EEG Time Series Database, 2001.

[11] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface. *Proceedings of the International Joint Conference on Neural Networks*, pages 2390–2397, 2008.

[12] Autonomous University of Barcelona. Code of Good Practice, 2022.

[13] Philippe Averty, Christian Collet, André Dittmar, Sylvie Athènes, and Evelyne Vernet-Maury. Mental Workload in Air Traffic Control: An Index Constructed from Field Tests. *Aviation, Space, and Environmental Medicine*, 75(4):333–341, 2004.

[14] Mahsa Bagheri and Sarah D. Power. EEG-based Detection of Mental Workload Level and Stress: The Effect of Variation in Each State on Classification of the Other. *Journal of Neural Engineering*, 17(5):056015, 10 2020.

[15] Nancy F. Bandstra, Carol S. Camfield, and Peter R. Camfield. Stigma of Epilepsy. *Canadian Journal of Neurological Sciences*, 35(4):436–440, 2008.

[16] Win-Ken Beh, Yi-Hsuan Wu, and An-Yeu (Andy) Wu. MAUS: A Dataset for Mental Workload Assessment on N-back task Using Wearable Sensor, 2021.

[17] N. M.G. Bodde, J. L. Brooks, G. A. Baker, P. A.J.M. Boon, J. G.M. Hendriksen, O. G. Mulder, and A. P. Aldenkamp. Psychogenic Non-epileptic Seizures – Definition, Etiology, Treatment and Prognostic Issues: A Critical Review. *Seizure*, 18(8):543–553, 2009.

[18] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris George Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 11 2022.

[19] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring Neurophysiological Signals in Aircraft Pilots and Car Drivers for the Assessment of Mental Workload, Fatigue and Drowsiness. *Neuroscience and Biobehavioral Reviews*, 44:58–75, 7 2014.

[20] Brain Workshop. Brain Workshop - A Dual N-Back Game, 2020.

[21] Anne Marie Brouwer, Thorsten O. Zander, Jan B.F. van Erp, Johannes E. Korteling, and Adelbert W. Bronkhorst. Using Neurophysiological Signals that Reflect Cognitive or Affective State: Six Recommendations to Avoid Common Pitfalls. *Frontiers in Neuroscience*, 9(APR):136, 2015.

[22] Borís Burle, Laure Spieser, Clémence Roger, Laurence Casini, Thierry Hasbroucq, and Franck Vidal. Spatial and Temporal Resolutions of EEG: Is It Really Black and White? A Scalp Current Density View. *International Journal of Psychophysiology*, 97(3):210, 9 2015.

[23] Jorge G. Burneo, Lorie Black, Roy Martin, Orrin Devinsky, Steve Pacia, Edward Faught, Blanca Vasquez, Robert C. Knowlton, Daniel Luciano, Werner Doyle, Sohuel Najjar, and Ruben I. Kuzniecky. Race/Ethnicity, Sex, and Socioeconomic Status as Predictors of Outcome After Surgery for Temporal Lobe Epilepsy. *Archives of Neurology*, 63(8):1106–1110, 8 2006.

[24] Gregory Cascino, Joseph I. Sirven, and William O. Tatum. *Epilepsy*. Wiley, 2nd edition, 2021.

[25] Jean-Marie Cellier and Hélène Eyrolle. Interference Between Switched Tasks. *Ergonomics*, 35(1):25–36, 1992.

[26] Center for Disease Control and Prevention. Epilepsy Is a Public Health Problem. Available on line at: https://www.cdc.gov/epilepsy/communications/infographics/cdc-epilepsy-text.htm, 9 2022.

[27] Satarupa Chakrabarti, Aleena Swetapadma, and Prasant Kumar Pattnaik. A Channel Independent Generalized Seizure Detection Method for Pediatric Epileptic Seizures. *Computer Methods and Programs in Biomedicine*, 209:106335, 9 2021.

[28] Rebecca L. Charles and Jim Nixon. Measuring Mental Workload using Physiological Measures: A Systematic Review. *Applied Ergonomics*, 74:221–232, 1 2019.

[29] Jiayu Chen, Aff M Asce, John E Taylor, M Asce, and Semra Comu. Assessing Task Mental Workload in Construction Projects: A Novel Electroencephalography Approach. *Journal of Construction Engineering and Management*, 143(8):04017053, 5 2017.

[30] Luke Clark and Barbara J. Sahakian. Cognitive Neuroscience and Brain Imaging in Bipolar Disorder. *https://doi.org/10.31887/DCNS.2008.10.2/lclark*, 10(2):153–163, 2022.

[31] Michael X. Cohen. Where Does EEG Come From and What Does It Mean? *Trends in Neurosciences*, 40(4):208–218, 4 2017.

[32] Vincent T. Cunliffe, Richard A. Baines, Carlo N.G. Giachello, Wei Hsiang Lin, Alan Morgan, Markus Reuber, Claire Russell, Matthew C. Walker, and Robin S.B. Williams. Epilepsy Research Methods Update: Understanding the Causes of Epileptic Seizures and Identifying New Treatments Using Non-Mammalian Model Organisms. *Seizure*, 24(C):44–51, 1 2015.

[33] Hanneke M. de Boer, Marco Mula, and Josemir W. Sander. The Global Burden and Stigma of Epilepsy. *Epilepsy & behavior : E&B*, 12(4):540–546, 5 2008.

[34] Maryann C. Deak and Robert Stickgold. Sleep and Cognition. *Wiley interdisciplinary reviews. Cognitive science*, 1(4):491, 2010.

[35] Frédéric Dehais, Alban Duprès, Sarah Blum, Nicolas Drougard, Sébastien Scannella, Raphaëlle N. Roy, and Fabien Lotte. Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors 2019, Vol. 19, Page 1324*, 19(6):1324, 3 2019.

[36] Gianluca Di Flumeri, Gianluca Borghini, Pietro Aricò, Nicolina Sciaraffa, Paola Lanzi, Simone Pozzi, Valeria Vignali, Claudio Lantieri, Arianna Bichicchi, Andrea Simone, and Fabio Babiloni. EEG-based Mental Workload Neurometric to Evaluate the Impact of Different Traffic and Road Conditions in Real Driving Settings. *Frontiers in Human Neuroscience*, 12:509, 12 2018.

[37] Yi Ding, Yaqin Cao, Vincent G. Duffy, Yi Wang, and Xuefeng Zhang. Measurement and Identification of Mental Workload during Simulated Computer Tasks with Multimodal Methods and Machine Learning. *Ergonomics*, 63(7):896–908, 7 2020.

[38] F. Thomas Eggemeier. Properties of Workload Assessment Techniques. *Advances in Psychology*, 52(C):41–62, 1 1988.

[39] Emotiv. Emotiv Epoc X 14-Channel Wireless EEG Headset. Available on line at: https://www.emotiv.com/epoc-x, 2021.

[40] Camilo Espinosa-Jovel, Rafael Toledano, Ángel Aledo-Serrano, Irene García-Morales, and Antonio Gil-Nagel. Epidemiological Profile of Epilepsy in Low Income Populations. *Seizure*, 56:67–72, 3 2018.

[41] Jessica J. Falco-Walter, Ingrid E. Scheffer, and Robert S. Fisher. The New Definition and Classification of Seizures and Epilepsy. *Epilepsy Research*, 139:73–79, 2018.

[42] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep Learning for Healthcare Applications Based on Physiological Signals: A Review. *Computer Methods and Programs in Biomedicine*, 161:1–13, 2018.

[43] Robert S. Fisher and Anna M. Bonner. The Revised Definition and Classification of Epilepsy for Neurodiagnostic Technologists. *Neurodiagnostic Journal*, 58(1):1–10, 2018.

[44] Cyrus K Foroughi, Nicole E Werner, Ryan McKendrick, David M Cades, and Deborah A Boehm-Davis. Individual Differences in Working-memory Capacity and Task Resumption Following Interruptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9):1480, 2016.

[45] Walter J. Freeman and Rodrigo Quian Quiroga. *Imaging Brain Function with EEG: Advanced Temporal and Spatial Analysis of Electroencephalographic Signals*, volume 9781461449. Springer, 2013.

[46] Rongrong Fu, Yongsheng Tian, Peiming Shi, and Tiantian Bao. Automatic Detection of Epileptic Seizures in EEG Using Sparse CSP and Fisher Linear Discrimination Analysis Algorithm. *Journal of Medical Systems 2020 44:2*, 44(2):1–13, 2020.

[47] Bin Gao, Jiazheng Zhou, Yuying Yang, Jinxin Chi, and Qi Yuan. Generative Adversarial Network and Convolutional Neural Network-based EEG Imbalanced Classification Model for Seizure Detection. *Biocybernetics and Biomedical Engineering*, 42(1):1–15, 1 2022.

[48] Yunyuan Gao, Bo Gao, Qiang Chen, Jia Liu, and Yingchun Zhang. Deep Convolutional Neural Network-based Epileptic Electroencephalogram (EEG) Signal Classification. *Frontiers in Neurology*, 11:375, 2020.

[49] Zhongke Gao, Xinmin Wang, Yuxuan Yang, Chaoxu Mu, Qing Cai, Weidong Dang, and Siyang Zuo. EEG-Based Spatio-Temporal Convolutional Neural Network for Driver Fatigue Evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2755–2763, 9 2019.

[50] Stefan Göbel. Serious Games Application Examples. In Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer, editors, *Serious Games: Foundations, Concepts and Practice*, chapter 12, page 437. Springer, 2016.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[52] Shi Gu, Fabio Pasqualetti, Matthew Cieslak, Qawi K. Telesford, Alfred B. Yu, Ari E. Kahn, John D. Medaglia, Jean M. Vettel, Michael B. Miller, Scott T. Grafton, and Danielle S. Bassett. Controllability of Structural Brain Networks. *Nature Communications 2015 6:1*, 6(1):1–10, 10 2015.

[53] Shankar S. Gupta, Trupti J. Taori, Mahesh Y. Ladekar, Ramchandra R. Manthalkar, Suhas S. Gajre, and Yashwant V. Joshi. Classification of Cross Task Cognitive Workload Using Deep Recurrent Network with Modelling of Temporal Dynamics. *Biomedical Signal Processing and Control*, 70:103070, 9 2021.

[54] Soo Yeon Han, No Sang Kwak, Taegeun Oh, and Seong Whan Lee. Classification of Pilots' Mental States Using a Multimodal Deep Learning Network. *Biocybernetics and Biomedical Engineering*, 40(1):324–336, 1 2020.

[55] Sandra G Hart. NASA-Task Load Index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[56] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[57] Aura Hernández-Sabaté, Meritxell Joanpere, Núria Gorgorió, and Lluís Albarracín. Mathematics learning opportunities when playing a tower defense game.

[58] Anu Holm, Kristian Lukander, Jussi Korpela, Mikael Sallinen, and Kiti M.I. Müller. Estimating Brain Load from the EEG. *The Scientific World Journal*, 9:639–651, 2009.

[59] Li Hu and Zhiguo Zhang. *EEG Signal Processing and Feature Extraction*. Springer Singapore, 1 2019.

[60] Xinmei Hu, Shasha Yuan, Fangzhou Xu, Yan Leng, Kejiang Yuan, and Qi Yuan. Scalp EEG Classification using Deep Bi-LSTM Network for Seizure Detection. *Computers in Biology and Medicine*, 124, 9 2020.

[61] Xinyun Hu and Gabriel Lodewijks. Detecting Fatigue in Car Drivers and Aircraft Pilots by Using Non-invasive Measures: The Value of Differentiation of Sleepiness and Mental Fatigue. *Journal of safety research*, 72:173–187, 2020.

[62] Susanne M. Jaeggi, Martin Buschkuehl, John Jonides, and Walter J. Perrig. Improving Fluid Intelligence with Training on Working Memory. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6829–6833, 5 2008.

[63] Kim M Kiely. Cognitive Function. In Alex C Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 974–978. Springer Netherlands, Dordrecht, 2014.

[64] J. F. Kihlstrom. Unconscious Cognition. In William P. Banks, editor, *Encyclopedia of Consciousness*, pages 411–421. Academic Press, 2009.

[65] Sandra L Kirmeyer. Coping with Competing Demands: Interruption and the Type A Pattern. *Journal of Applied Psychology*, 73(4):621, 1988.

[66] Valentina F. Kitchigina. Alterations of Coherent Theta and Gamma Network Oscillations as an Early Biomarker of Temporal Lobe Epilepsy and Alzheimer's Disease. *Frontiers in Integrative Neuroscience*, 12:36, 8 2018.

[67] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 1 2012.

[68] Nuri Korhan, Zumray Dokur, and Tamer Olmez. Motor Imagery Based EEG Classification by using Common Spatial Patterns and Convolutional Neural Networks. In *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*, pages 1–4, Istanbul, Turkey, 4 2019. Institute of Electrical and Electronics Engineers Inc.

[69] John LaRocco, Minh Dong Le, and Dong Guk Paeng. A Systemic Review of Available Low-Cost EEG Headsets Used for Drowsiness Detection. *Frontiers in Neuroinformatics*, 14:553352, 10 2020.

[70] Kara A. Latorella. Investigating Interruptions: Implications for Flightdeck Performance. *NASA*, 99, 1999.

[71] J. Street Laurence. *Introduction to Biomedical Engineering Technology*. CRC Press, 3rd edition, 2016.

[72] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A Compact Convolutional Neural Network for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, 15(5):056013, 7 2018.

[73] Dae Hyeok Lee, Ji Hoon Jeong, Kiduk Kim, Baek Woon Yu, and Seong Whan Lee. Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network. *IEEE Access*, 8:121929–121941, 2020.

[74] Chaosong Li, Weidong Zhou, Guoyang Liu, Yanli Zhang, Minxing Geng, Zhen Liu, Shang Wang, and Wei Shang. Seizure Onset Detection Using Empirical Mode Decomposition and Common Spatial Pattern. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:458–467, 2021.

[75] Yang Li, Zuyi Yu, Yang Chen, Chunfeng Yang, Yue Li, X. Allen Li, and Baosheng Li. Automatic Seizure Detection using Fully Convolutional Nested LSTM. *International journal of neural systems*, 30(4), 4 2020.

[76] W. L. Lim, O. Sourina, and L. P. Wang. STEW: Simultaneous Task EEG Workload Data Set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2106–2114, 11 2018.

[77] Ying Lin, Julian Mutz, Peter J. Clough, and Kostas A. Papageorgiou. Mental Toughness and Individual Differences in Learning, Educational and Work Performance, Psychological Well-being, and Personality: A Systematic Review. *Frontiers in Psychology*, 8(Aug):1345, 8 2017.

[78] Erik K. St. Louis, Lauren C. Frey, Jeffrey W. Britton, Lauren C. Frey, Jennifer L. Hopp, Pearce Korb, Mohamad Z. Koubeissi, William E. Lievens, Elia M. Pestana-Knight, and Erik K. St. Louis. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, 2016.

[79] A. Marinescu, S. Sharples, A. C. Ritchie, T. Sánchez López, M. McDowell, and H. Morvan. Exploring the Relationship between Mental Workload, Variation in Performance and Physiological Parameters. *IFAC-PapersOnLine*, 49(19):591–596, 1 2016.

[80] Mayo Clinic. Epilepsy - Symptoms and Causes. Available on line at: https://www.mayoclinic.org/diseases-conditions/epilepsy/symptoms-causes/syc-20350093, 2022.

[81] Luis Alfredo Moctezuma and Marta Molinas. EEG Channel-Selection Method for Epileptic-Seizure Classification Based on Multi-Objective Optimization. *Frontiers in Neuroscience*, 14:593, 6 2020.

[82] Mahta Mousavi and Virginia R. De Sa. Temporally Adaptive Common Spatial Patterns with Deep Convolutional Neural Networks. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 4533–4536. Institute of Electrical and Electronics Engineers Inc., 7 2019.

[83] Kevin Murphy. *Machine Learning : A Probabilistic Perspective*. MIT Press, Cambridge, 2012.

[84] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge, 2022.

[85] P. K. Myint, E. F.A. Staufenberg, and K. Sabanathan. Post-stroke Seizure and Post-Stroke Epilepsy. *Postgraduate Medical Journal*, 82(971):568, 9 2006.

[86] Suzanne Oosterwijk, Kristen A. Lindquist, Eric Anderson, Rebecca Dautoff, Yoshiya Moriguchi, and Lisa Feldman Barrett. States of Mind: Emotions, Body Feelings, and Thoughts Share Distributed Neural Networks. *NeuroImage*, 62(3):2110–2128, 9 2012.

[87] Antti Oulasvirta and Pertti Saariluoma. Long-term Working Memory and Interrupting Messages in Human–Computer Interaction. *Behaviour \& Information Technology*, 23(1):53–64, 2004.

[88] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of cognitive engineering and decision making*, 2(2):140–160, 2008.

[89] Trong Huy Phan and Kazuma Yamamoto. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses. 2020.

[90] Miquel Angel Piera, Juan Jose Ramos, and Jose Luis Muñoz. A Socio-technical Holistic Agent Based Model to Assess Cockpit Supporting Tools Performance Variability. *IFAC-PapersOnLine*, 52(11):122–127, 1 2019.

[91] PlaneCrashInfo.com. Causes of Fatal Accidents by Decade. Online available at: http://www.planecrashinfo.com/cause.htm, 2022.

[92] Magorzata Plechawska-Wójcik, Mikhail Tokovarov, Monika Kaczorowska, and Dariusz Zapaa. A Three-Class Classification of Cognitive Workload Based on EEG Spectral Data. *Applied Sciences*, 9(24):5340, 12 2019.

[93] Hongquan Qu, Yiping Shan, Yuzhe Liu, Liping Pang, Zhanli Fan, Jie Zhang, and Xiaoru Wanyan. Mental Workload Classification Method Based on EEG Independent Component Features. *Applied Sciences 2020, Vol. 10, Page 3036*, 10(9):3036, 4 2020.

[94] Thea Radüntz, Norbert Fürstenau, Thorsten Mühlhausen, and Beate Meffert. Indexing Mental Workload During Simulated Air Traffic Control Tasks by Means of Dual Frequency Head Maps. *Frontiers in Physiology*, 11:300, 4 2020.

[95] Herbert Ramoser, Johannes Müller-Gerking, and Gert Pfurtscheller. Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

[96] Stephen K. Reed. *Cognition: Theories and Applications*. Cengage Learning, 2012.

[97] Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep Learning in Physiological Signal Data: A Survey. *Sensors*, 20(4):969, 2 2020.

[98] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep Learning-based Electroencephalography Analysis: A Systematic Review. *Journal of Neural Engineering*, 16(5), 2019.

[99] Siavash Sakhavi, Cuntai Guan, and Shuicheng Yan. Learning Temporal Information for Brain-Computer Interface Using Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5619–5629, 11 2018.

[100] V. Salehi, T. T. Tran, B. Veitch, and D. Smith. A Reinforcement Learning Development of the FRAM for Functional Reward-based Assessments of Complex Systems Performance. *International Journal of Industrial Ergonomics*, 88:103271, 3 2022.

[101] Saeid Sanei and Jonathon A. Chambers. *EEG Signal Processing and Machine Learning*. Wiley, 2nd edition, 2021.

[102] Federica Savazzi, Sara Isernia, Johanna Jonsdottir, Sonia Di Tella, Stefania Pazzi, and Francesca Baglio. Engaged in learning neurorehabilitation: Development and validation of a serious game with user-centered design. *Computers & Education*, 125:53–61, 2018.

[103] Jesse Schell. *The Art of Game Design : A Book of Lenses*. CRC Press, 3rd edition, 2019.

[104] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Human Brain Mapping*, 38(11):5391–5420, 11 2017.

[105] Nicolina Sciaraffa, Daniele Germano, Andrea Giorgi, Vincenzo Ronca, Alessia Vozzi, Gianluca Borghini, Gianluca Di Flumeri, Fabio Babiloni, and Pietro Aricò. Mental Effort Estimation by Passive BCI: A Cross-Subject Analysis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine \& Biology Society (EMBC)*, pages 906–909. IEEE, 2021.

[106] Natalia Sevcenko, Manuel Ninaus, Franz Wortha, Korbinian Moeller, and Peter Gerjets. Measuring Cognitive Load Using In-Game Metrics of a Serious Simulation Game. *Frontiers in Psychology*, 12:906, 2021.

[107] M. Shamim Hossain, Syed Umar Amin, Mansour Alsulaiman, and Ghulam Muhammad. Applying Deep Learning for Epilepsy Seizure Detection and Brain Mapping Visualization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 15(1s), 2 2019.

[108] Shimmer. Shimmer3 Ebio Unit.

[109] Jaeyoung Shin, Alexander Von Luhmann, Benjamin Blankertz, Do Won Kim, Jan Mehnert, Jichai Jeong, Han Jeong Hwang, and Klaus Robert Muller. Open Access Repository for Hybrid EEG-NIRS data. *2018 6th International Conference on Brain-Computer Interface, BCI 2018*, 2018-Janua:1–4, 3 2018.

[110] Ali Shoeb and John Guttag. Application of Machine Learning to Epileptic Seizure Detection. In *ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 975–982, 2010.

[111] Ali Shoeb and John Guttag. CHB-MIT Scalp EEG Database, 2010.

[112] Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, Abbas Khosravi, Amir F. Atiya, Diba Aminshahidi, Sadiq Hussain, Modjtaba Rouhani, Saeid Nahavandi, and U. Rajendra Acharya. Epileptic Seizures Detection Using Deep Learning Techniques: A Review. *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 5780*, 18(11):5780, 5 2021.

[113] Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Xiaodi Huang, and Nasir Hussain. A Review of Epileptic Seizure Detection using Machine Learning Classifiers. *Brain Informatics*, 7(1):1–18, 12 2020.

[114] Fátima Pereira da Silva. Mental Workload, Task Demand and Driving Performance: What Relation? *Procedia - Social and Behavioral Sciences*, 162:310–319, 12 2014.

[115] Alessandro Silvani, Giovanna Calandra-Buonaura, Roger A.L. Dampney, and Pietro Cortelli. Brain-Heart Interactions: Physiology and Clinical Implications. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2067), 5 2016.

[116] Winnie K.Y. So, Savio W.H. Wong, Joseph N. Mak, and Rosa H.M. Chan. An Evaluation of Mental Workload with Frontal EEG. *PLoS ONE*, 12(4), 4 2017.

[117] Ernest R. Somerville. Some Treatments Cause Seizure Aggravation in Idiopathic Epilepsies (especially absence epilepsy). *Epilepsia*, 50(SUPPL. 8):31–36, 9 2009.

[118] Dan Sperber and Deirdre Wilson. *Relevance - Communication and Cognition*. Wiley-Blackwell, 2nd edition, 1995.

[119] Suunto. Suunto Ambit3.

[120] Catherine M. Sweeney-Reed, Slawomir J. Nasuto, Marcus F. Vieira, and Adriano O. Andrade. Empirical Mode Decomposition and its Extensions Applied to EEG Analysis: A Review. In Nii O. Attoh-Okine, editor, *Advances in Data Science and Adaptive Analysis*, volume 10, page 1840001. World Scientific Publishing Company, 8 2018.

[121] Makoto Takahashi, Arai Tsuyoshi, Osamu Kubo, and Hidekazu Yoshikawa. Experimental Study Toward Mutual Adaptive Interface. In *Robot and Human Communication - Proceedings of the IEEE International Workshop*, pages 271–276. IEEE, 1994.

[122] Lee Taylor, Samuel L. Watkins, Hannah Marshall, Ben J. Dascombe, and Josh Foster. The Impact of Different Environmental Conditions on Cognitive Function: A Focused Review. *Frontiers in Physiology*, 6(Jan):372, 1 2015.

[123] Neha Tiwari, Damodar Reddy Edla, Shubham Dodia, and Annushree Bablani. Brain Computer Interface: A Comprehensive Survey. *Biologically Inspired Cognitive Architectures*, 26:118–129, 2018.

[124] Suat Toraman. Automatic Recognition of Preictal and Interictal EEG Signals using 1D-Capsule Networks. *Computers & Electrical Engineering*, 91:107033, 5 2021.

[125] Kostas M. Tsiouris, Vasileios C. Pezoulas, Dimitrios D. Koutsouris, Michalis Zervakis, and Dimitrios I. Fotiadis. Discrimination of Preictal and Interictal Brain States from Long-Term EEG Data. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 2017-June:318–323, 11 2017.

[126] Kostas M. Tsiouris, Vasileios C. Pezoulas, Michalis Zervakis, Spiros Konitsiotis, Dimitrios D. Koutsouris, and Dimitrios I. Fotiadis. A Long Short-Term Memory Deep Learning Network for the Prediction of Epileptic Seizures using EEG Signals. *Computers in Biology and Medicine*, 99:24–37, 8 2018.

[127] Syed Muhammad Usman, Shehzad Khalid, Rizwan Akhtar, Zuner Bortolotto, Zafar Bashir, and Haiyang Qiu. Using Scalp EEG and Intracranial EEG Signals for Predicting Epileptic Seizures: Review of Available Methodologies. *Seizure*, 71:258–269, 10 2019.

[128] Syed Muhammad Usman, Shehzad Khalid, and Zafar Bashir. Epileptic Seizure Prediction using Scalp Electroencephalogram Signals. *Biocybernetics and Biomedical Engineering*, 41(1):211–220, 1 2021.

[129] Syed Muhammad Usman, Muhammad Usman, and Simon Fong. Epileptic Seizures Prediction Using Machine Learning Methods. *Computational and Mathematical Methods in Medicine*, 2017:9074759, 2017.

[130] Merel M Van der Wal, Joop De Kraker, Carolien Kroeze, Paul A Kirschner, and Pieter Valkering. Can computer models be used for social learning? a serious game in water management. *Environmental modelling & software*, 75:119–132, 2016.

[131] Jonathan Vöglein, Ingrid Ricard, Soheyl Noachtar, Walter A. Kukull, Marianne Dieterich, Johannes Levin, and Adrian Danek. Seizures in Alzheimer's Disease Are Highly Recurrent and Associated with a Poor Disease Course. *Journal of Neurology*, 267(10):2941, 10 2020.

[132] Xiaoshuang Wang, Xiulin Wang, Wenya Liu, Zheng Chang, Tommi Kärkkäinen, and Fengyu Cong. One Dimensional Convolutional Neural Networks for Seizure Onset Detection using Long-term Scalp and Intracranial EEG. *Neurocomputing*, 459:212–222, 10 2021.

[133] Yijun Wang, Shangkai Gao, and Xiaorong Gao. Common Spatial Pattern Method for Channel Selelction in Motor Imagery Based Brain-computer Interface. In *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, volume 2005, pages 5392–5395. Conf Proc IEEE Eng Med Biol Soc, 2005.

[134] Z L Wang, Lei Yang, and J S Ding. Application of Heart Rate Variability in Evaluation of Mental Workload. *Chinese journal of industrial hygiene and occupational diseases*, 23(3):182–184, 2005.

[135] Vibhangini S. Wasade and Marianna V. Spanaki. *Understanding Epilepsy: A Study Guide for the Boards*. Cambridge University Press, 2019.

[136] Larry Wasserman. *All of Statistics : A Concise Course in Statistical Inference*. Springer, 2010.

[137] William W. S. Wei. *Multivariate Time Series Analysis and Applications*. Wiley, 2019.

[138] Zuochen Wei, Junzhong Zou, Jian Zhang, and Jianqiang Xu. Automatic Epileptic EEG Detection using Convolutional Neural Network with Improvements in Time-domain. *Biomedical Signal Processing and Control*, 53:101551, 8 2019.

[139] Christopher D. Wickens. Situation Awareness and Workload in Aviation. *Current Directions in Psychological Science*, 11(4):128–133, 8 2002.

[140] World Health Organization. Epilepsy. Available on line at: https://www.who.int/health-topics/epilepsy, 2022.

[141] World Health Organization. Road Traffic Injuries. Available on line at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, 2022.

[142] Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D Van Der Spek. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of educational psychology*, 105(2):249, 2013.

[143] Edmond Q. Wu, X. Y. Peng, Caizhi Z. Zhang, J. X. Lin, and Richard S.F. Sheng. Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network. *IEEE Transactions on Instrumentation and Measurement*, 68(10):3907–3919, 10 2019.

[144] Gaowei Xu, Tianhe Ren, Yu Chen, and Wenliang Che. A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis. *Frontiers in Neuroscience*, 14:1253, 12 2020.

[145] José Elías Yauri Vidalón, Pau Folch, Daniel Álvarez, Debora Gil, and Aura Hernàndez i Sabaté. Dataset to Predict Mental Workload Based on Physiological Data, 2022.

[146] Zhong Yin and Jianhua Zhang. Recognition of Cognitive Task Load Levels Using Single Channel EEG and Stacked Denoising Autoencoder. In *Chinese Control Conference, CCC*, volume 2016-Augus, pages 3907–3912. IEEE Computer Society, 8 2016.

[147] Mark S. Young, Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. State of Science: Mental Workload in Ergonomics. *Ergonomics*, 58(1):1–17, 1 2015.

[148] Qi Yuan, Weidong Zhou, Liren Zhang, Fan Zhang, Fangzhou Xu, Yan Leng, Dongmei Wei, and Meina Chen. Epileptic Seizure Detection Based on Imbalanced Classification and Wavelet Packet Transform. *Seizure*, 50:99–108, 8 2017.

[149] Chen Yun and Wang Xuefeng. Association Between Seizures and Diabetes Mellitus: A Comprehensive Review of Literature. *Current diabetes reviews*, 9(4):350–354, 6 2013.

[150] L Zhang, C Zhang, Hiroshi Higashi, Jianting Cao, and Toshihisa Tanaka. Common Spatial Pattern Using Multivariate EMD for EEG Classification. In *APSIPA ASC 2011 - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2011*, pages 244–248. APSIPA, 2011.

[151] Pengbo Zhang, Xue Wang, Junfeng Chen, Wei You, and Weihang Zhang. Spectral and Temporal Feature Learning with Two-Stream Neural Networks for Mental Workload Assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1149–1159, 6 2019.

[152] Yawei Zhao, Jiabei Tang, Yong Cao, Xuejun Jiao, Minpeng Xu, Peng Zhou, Dong Ming, and Hongzhi Qi. Effects of Distracting Task with Different Mental Workload on Steady-State Visual Evoked Potential Based Brain Computer Interfaces—an Offline Study. *Frontiers in Neuroscience*, 12(FEB):79, 2018.

[153] Yu Zhonggen. A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology*, 2019, 2019.

[154] Jiazheng Zhou, Li Liu, Yan Leng, Yuying Yang, Bin Gao, Zonghong Jiang, Weiwei Nie, and Qi Yuan. Both Cross-Patient and Patient-Specific Seizure Detection Based on Self-Organizing Fuzzy Logic. *International Journal of Neural Systems*, 32(6), 2022.

[155] Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. Epileptic Seizure Detection Based on EEG Signals and CNN. *Frontiers in Neuroinformatics*, 12:95, 12 2018.