Universidad de Valladolid

Doctoral Program of
Information and Telecommunication Technologies

Doctoral Thesis

Enhancing EEG-Based Brain-Computer Interfaces with Deep Learning and Explainable Artificial Intelligence

Thesis presented by **D. Sergio Pérez Velasco**
to apply for the *Ph.D. degree*
from the *University of Valladolid*

Directed by:
Dr. D. Roberto Hornero Sánchez

2025
Valladolid, Spain

A mi familia, amigos y compañeros

Defense

Título	Enhancing EEG-Based Brain-Computer Interfaces with Deep Learning and Explainable Artificial Intelligence
Autor	D. Sergio Pérez Velasco
Director	Dr. D. Roberto Hornero Sánchez
Departamento	Teoría de la Señal y Comunicaciones e Ingeniería Telemática

---

Tribunal

Presidente

Vocal

Secretario

acuerda otorgarle la calificación de

En Valladolid, a de del

Universidad de Valladolid

Escuela Técnica Superior de Ingenieros de Telecomunicación
Dpto. de Teoría de la Señal y Comunicaciones e Ingeniería Telemática

Research Stay for the International Mention

City:	Nijmegen (Netherlands)
Faculty:	Donders Institute for Brain, Cognition and Behaviour
Institution:	Radboud University
Research group:	Data-driven Neurotechnology lab
Dates:	01/04/2024 - 01/07/2024
Duration:	91 days (3 months)
Supervisor:	Dr. Michael Tangermann

DONDERS INSTITUTE

Radboud Universiteit

Agradecimientos

Me gustaría comenzar agradeciendo al Dr. Roberto Hornero Sánchez, primero, por contestar el correo de un completo desconocido con la voluntad de guiar de un educador experimentado y por, en ese mismo correo, ofrecerme la oportunidad de realizar esta investigación sobre los sistemas *brain-computer interface* (BCI), un mundo que me ha apasionado y que, sin su apoyo, seguramente no habría recorrido. Gracias también por su paciencia con mis constantes retrasos en las fechas límite de escritura.

Me gustaría ahora agradecer a todos los compañeros del Grupo de Ingeniería Biomédica (GIB). A las *altas esferas* por sus consejos y ejemplo: Carlos, Jesús, Gonzalo y María. A las *bajas esferas* (los que están, los que acaban de llegar y los que deberían ser ya solo altas esferas) por compartir los momentos más divertidos: Víctor G., Víctor R., Aarón, Adrián, Roberto, Fernando, Clara, María, Jorge, Verónica, Javier, Marcos, Marina, Máximo y Teresa. Y, por supuesto, a mis compañeros de la línea de BCI quienes más me han ayudado y apoyado en el día a día; a Eduardo y Víctor sobre todo por guiarme y ayudarme a no descarrilar del camino doctoral; a Diego por haberme acompañado todo este tiempo y ser también un apoyo emocional; y a Selene, Beatriz, Ana, Rubén y Rebeca, por haber compartido conmigo las dificultades de realizar las tareas de los proyectos. En general, a todos por haber creado un ambiente inigualable en el que no solo me he podido desarrollar como investigador sino también como persona.

I would also like to thank Dr. Michael Tangermann and all the people at the Data-Driven Neurotechnology Lab (Radboud University) for sharing with me their passion for research. My special thanks go to Pierre, for showing me around Nijmegen and including me in every plan beyond the lab; and to Jordy, Guido, Matthias, Peter and Hanneke for turning my research stay into a truly welcoming experience.

También ha sido muy importante para mí el apoyo durante este tiempo que he

recibido de mis amigos. En especial a todo el grupo de ADASAJEAR y AMYJASS: a Andrés, que a pesar de la distancia, nunca dejó de estar cerca; a Alberto, por contagiarme su alegría y pasión por cada día; a Javi, porque siempre estás ahí; a Alex, Velázquez, Daniel, Eric y Raúl porque siempre me han apoyado en las decisiones que he tomado y con vuestros consejos me animásteis a comenzar esta tesis; a Yari y María por crear en todo momento un ambiente tan familiar.

Por último, pero quizá lo más importante, a mi familia, que incondicionalmente ha estado y estará siempre conmigo: a mis padres, que me han criado e inculcado los valores hoy me definen y hacen posible alcanzar este hito; al resto de mi familia, que siempre celebran conmigo los grandes hitos; y a Sandra, que me ha apoyado durante las innumerables noches en vela por la procrastinación, por llenar cada día de alegría para poder llevar los días más difíciles del camino doctoral y por caminar a mi lado en la creación de nuestra futura familia.

A todos, gracias de corazón.

Abstract

For decades, researchers have envisioned utilizing brain signals to interact with computers and assistive devices directly through thought alone. Although the field of electroencephalography (EEG)-based brain–computer interfaces (BCIs) has seen significant advances, current systems often remain constrained by suboptimal reliability, lengthy user calibration procedures, and a limited understanding of the underlying neural dynamics. These obstacles hamper the practical deployment of BCIs in real-world scenarios, including robust assistive systems for people with severe motor impairments.

In this context, this Doctoral Thesis focuses on improving motor imagery (MI) EEG-based BCIs by leveraging state-of-the-art deep learning (DL) methods and explainable artificial intelligence (XAI) techniques. It first introduces *EEGSym*, a novel DL architecture developed to tackle the well-known variability in EEG signals that often leads to "BCI inefficiency". EEGSym incorporates hemispheric symmetry, inception modules, and residual connections to robustly decode MI commands directly from raw EEG. Furthermore, it is enhanced with data augmentation strategies and a multi-dataset transfer learning pipeline, eliminating the need for lengthy calibration across different users and recording sessions.

Alongside the proposed DL architecture, a XAI-driven methodology was introduced to reveal the spatio-temporal features most relevant for MI decoding. By adapting Shapley additive explanations (SHAP) to the EEG domain, this work uncovers the neural patterns and cortical regions the model relies upon, highlighting not only primary sensorimotor areas but also the prefrontal and parietal cortices. These findings broaden our understanding of brain activation beyond conventional sensorimotor paradigms, helping to optimize electrode setups and improve BCI usability. In particular, an eight-channel montage was selected through SHAP analysis to minimize equipment needs without compromising accuracy.

Lastly, this thesis explores motor execution (ME) to MI direct transfer learning as a means to improve MI sessions. Models exclusively trained on ME data were found to classify MI tasks with comparable performance to MI-trained models, pointing to a shared neural substrate that may help circumvent challenges, such as low user engagement or unobservable MI attempts. By unifying these insights (robust DL architectures, explainable feature attribution, and ME→MI transfer), this work provides a pathway toward more efficient, reliable, and versatile BCIs.

In the three studies that comprise this Doctoral Thesis, we evaluated the proposed solutions for novelty, robustness, and flexibility. First, in a strictly inter-subject evaluation that merged five public datasets and 280 participants, *EEGSym* achieved mean accuracies of $88.6 \pm 9.0\%$ on *Physionet*, $83.3 \pm 9.3\%$ on *OpenBMI*, $85.1 \pm 9.5\%$ on *Kaya2018*, $87.4 \pm 8.0\%$ on *Meng2019*, and $90.2 \pm 6.5\%$ on the *Carnegie Mellon* dataset while relying on only 16 electrodes; consequently, 268 of the 280 users (95.7%) surpassed the 70% BCI-control threshold, reducing the proportion of "BCI-inefficient" users to below 5%.

Second, coupling *EEGSym* with SHAP-based explainability revealed that motor imagery engages a distributed network extending from the sensorimotor cortex to prefrontal and posterior-parietal regions. Based on this insight, we selected an eight-channel montage centered on F7/F8 and bilateral centro-parietal sites that maintained high accuracy ($86.5 \pm 10.6\%$ on *Physionet* and $88.7 \pm 7.0\%$ on the *Carnegie Mellon* dataset), thus enabling reduced EEG setups for MI-based BCIs with comparable performance. Finally, we show for the first time that a model trained exclusively on ME data can decode MI with comparable accuracy, demonstrating direct ME→MI transfer that eliminates user-specific MI calibration.

Taken together, these results outline a path toward MI-based BCIs that are accurate, interpretable, and easy to deploy while remaining resilient to inter-subject and inter-session variability. Although these tests were conducted offline with healthy volunteers, the 95% control rate, portable eight-electrode configuration, and calibration-free ME→MI transfer establish a firm basis for future clinical and real-time studies, accelerating the transition of BCIs from laboratory prototypes to reliable assistive technologies for daily use.

Acronyms

AI	Artificial Intelligence
BCI	Brain-Computer Interface
CAR	Common Average Reference
CSP	Common Spatial Patterns
CNN	Convolutional Neural Network
c-VEP	Code-modulated Visual Evoked Potentials
DA	Data Augmentation
DI	Degree of Invasiveness
DL	Deep Learning
EEG	Electroencephalography
ECoG	Electrocorticography
ERP	Event-Related Potentials
ErrP	Error-Related Potentials
FBCSP	Filter Bank Common Spatial Patterns
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
ICA	Independent Component Analysis
IIR	Infinite Impulse Response
JCR	Journal Citation Reports
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-agnostic Explanations
LOSO	Leave-One-Subject-Out

LRP	Layer-wise Relevance Propagation
M1	Primary Motor Cortex
ME	Motor Execution
MEG	Magnetoencephalography
MI	Motor Imagery
ML	Machine Learning
MLP	Multilayer Perceptron (MLP)
MRCP	Movement-Related Cortical Potentials
M1	Motor Cortex
NLP	Natural Language Processing
PET	Positron Emission Tomography
PFC	Prefrontal Cortex
PPC	Posterior Parietal Cortex
RNN	Recurrent Neural Network
SCP	Slow Cortical Potentials
SHAP	Shappley additive explanations
SMR	Sensorymotor Rhythms
SNR	Signal-to-noise-ratio
SSVEP	Steady-State Visual Evoked Pontentials
SVM	Support Vector Machines
S1	Primary Somatosensory Cortex
TL	Transfer Learning
TPR	True Positive Rate
VAE	Variational Autoencoder
XAI	Explainable Artificial Intelligence

Contents

Abstract **I**

Acronyms **III**

1 Introduction **1**
1.1 Compendium of publications: thematic consistency 2
1.2 Context: biomedical engineering, biomedical signal processing and machine learning 8
1.3 Brain-computer interfaces 10
1.3.1 The brain 13
1.3.2 Measuring brain activity: Electroencephalography 15
1.3.3 Brain-computer interface control signals and paradigms . . 18
1.4 Deep learning and explainable artificial intelligence 22
1.4.1 Introduction to deep learning 23
1.4.2 Importance of explainability in deep learning models 27
1.5 State-of-the-art 29
1.5.1 Deep learning methods for motor imagery decoding 30
1.5.2 Explainable artificial intelligence in interpreting deep learning models in electroencephalography analysis 32
1.5.3 Investigating motor imagery and execution from electroencephalography 33

2 Hypothesis and objectives **37**
2.1 Hypotheses 37
2.2 Objectives 38

3 Materials and methods **41**
3.1 Electroencephalography datasets 41

3.1.1 Experimental protocol 42
3.1.2 Task description 43
3.1.3 Dataset usage across thesis studies 44
3.2 Signal preprocessing 46
3.2.1 Electrode Selection 46
3.2.2 Frequency Filtering 47
3.2.3 Spatial Filtering 48
3.2.4 Resampling 49
3.2.5 Trial Extraction 49
3.2.6 Normalization 49
3.3 Deep learning model 50
3.3.1 Baseline models 51
3.3.2 Deep learning architecture design 52
3.3.3 Data augmentation 55
3.3.4 Training procedure 56
3.3.5 Evaluation metrics 58
3.4 Explainable artificial intelligence techniques 59
3.4.1 SHAP values for motor imagery 59
3.4.2 Channel selection based on SHAP values 61
3.5 Investigation of motor imagery and motor execution relationship . 62
3.5.1 Comparative analysis of electroencephalography patterns . 62
3.5.2 Comparison metrics 63

4 Results 67
4.1 Performance of the deep learning model 68
4.1.1 Training schedule 68
4.1.2 Comparison with other deep learning networks 69
4.1.3 Ablation study 71
4.2 Explainable artificial intelligence findings 72
4.2.1 Visualization of brain patterns 73
4.2.2 SHAP based channel selection 75
4.3 Relationship between motor imagery and motor execution 76
4.3.1 Accuracy correlation 77
4.3.2 Differences and similarities in electroencephalography activation patterns 79

5 Discussion **85**
5.1 Interpretation of the developed deep learning architecture 86
5.1.1 Factors influencing model performance 86
5.1.2 Comparison with previous inter-subject deep learning approaches 87
5.2 Insights from explainable artificial intelligence 89
5.2.1 Spatial insights: a multi-regional network 90
5.2.2 Temporal insights: early vs. late electroencephalography patterns 91
5.2.3 Practical impact: channel selection and reduced montages . 91
5.2.4 Broader implications and concluding remarks 92
5.3 Analysis of motor imagery and motor execution relationship 93
5.3.1 Overlap and synergy in neural representations 93
5.3.2 Practical benefits for calibration and rehabilitation 94
5.3.3 Impact on brain-computer interface development 95
5.4 Limitations 96

6 Conclusions **99**
6.1 Contributions 100
6.2 Main conclusions 101
6.3 Future research directions 102

A Papers included in the compendium of publications **105**
A.1 Contribution 1: Perez-Velasco et al. (2022) 106
A.2 Contribution 2: Pérez-Velasco et al. (2024) 107
A.3 Contribution 3: Pérez-Velasco et al. (2025) 109

B Scientific achievements **111**
B.1 Publications 111
B.1.1 Papers indexed in the JCR 111
B.1.2 International conferences 113
B.1.3 National conferences 114
B.2 International internship 117
B.3 Awards and honors 119

C Resumen en español **121**
C.1 Introducción 121
C.2 Hipótesis y objetivos 124

C.3 Materiales y métodos 126
C.4 Resultados y discusión 128
C.5 Conclusiones 130

Bibliography **133**

List of Figures

1.1 Diagram illustrating the thematic consistency 4
1.2 Schematic representation of a EEG-based BCI control loop 11
1.3 Anatomical regions of the human cerebral cortex 14
1.4 Temporal and spatial resolution of different brain activity measurement techniques 16
1.5 10-10 international system for EEG electrode placement 18
1.6 Violin plot illustrating SHAP values for tabulated data 29

3.1 Schematic of MI protocols in the public datasets 43
3.2 Schematic of trials in the public datasets 50
3.3 Overview of the *EEGSym* architecture 54
3.4 Visualization of the cross-validation procedure applied to each dataset 57

4.1 Training and validation loss curves for pre-training on the Physionet dataset (Goldberger et al., 2000) 69
4.2 Feature attribution maps from the Physionet dataset (Goldberger et al., 2000) 74
4.3 Feature attribution maps from the Stieger2021 dataset (Stieger et al., 2021) 75
4.4 Confusion matrices illustrating inter-subject classification results . 78
4.5 Correlation chart between classification accuracies obtained in ME→ME and ME→MI scenarios across subjects 78
4.6 Correlation chart between classification accuracies obtained in ME→MI and MI→MI scenarios across subjects 79
4.7 Feature attribution maps of the left-hand class 81
4.8 Feature attribution maps of the right-hand class 82
4.9 Feature attribution maps of the resting class 83

4.10 Mean feature attribution topographic plot 84

5.1 16-electrode configuration analyzed in Perez-Velasco et al. (2022) and Pérez-Velasco et al. (2024) 92
5.2 Correlation between classification accuracy on the Physionet (Goldberger et al., 2000) and Stieger2021 (Stieger et al., 2021) datasets 93

List of Tables

3.1 Details of the datasets 42
3.2 Summary of Dataset Usage and Electrode Configurations Across Thesis Studies 45

4.1 Comparison of accuracies on target datasets for 8 and 16 electrode configurations (Perez-Velasco et al., 2022) 70
4.2 Contribution of each novelty on Physionet 72
4.3 Comparison of binary classification performance of SHAP-based selection of 8-electrode configurations 76
4.4 Accuracies of inter-subject experiments 77

5.1 Comparison with binary classification of previous literature 88

Chapter 1

Introduction

This Doctoral Thesis aims to improve current brain–computer interface (BCI) systems based on neural self-regulation, specifically employing the motor imagery (MI) paradigm. By implementing novel deep learning (DL) techniques, it is intended to enhance current electroencephalography (EEG) classification systems used in MI-based BCIs where traditional machine learning (ML) pipelines are widely applied. The ultimate goal of this Doctoral Thesis is to develop a set of DL-based methodologies that can be applied to BCI systems to enhance EEG classification and to provide physiological interpretations of how these advanced classification techniques operate through the use of explainable artificial intelligence (XAI) techniques. This research has led to the publication of 3 articles in journals indexed in the Journal Citation Reports (JCR) from the Web of Science^{TM} between January 2022 and May 2025. Each manuscript is thematically consistent in exploring the use of DL in EEG classification; therefore, this doctoral thesis is presented as a compendium of publications. The research question we seek to answer with this Doctoral Thesis is: Can the latest advances in DL classification improve current BCI platforms, deepen our understanding of underlying neural mechanisms, and overcome the existing limitations of MI decoding accuracy?

The structure of this chapter is as follows. First, the thematic consistency of the articles that make up this Doctoral Thesis is explained in Section 1.1. Then, the general context of this research is briefly described in Section 1.2, which introduces biomedical engineering, signal processing and machine learning fields. Section 1.3 is focused on BCIs, introducing the history, the brain, control signals, and applications of these systems. Following, Section 1.4 introduces the field of deep learning, and the importance of explainability methods in DL models.

Finally, Section 1.5 provides a comprehensive state-of-the-art review of previous approaches to the main topics of this Doctoral Thesis: DL for EEG decoding with a focus on MI, the application of XAI methods to these architectures, and the EEG patterns identified by DL models during both MI and ME paradigms.

1.1 Compendium of publications: thematic consistency

The quest to establish a direct connection between the brains of different individuals (the concept of telepathy) was the initial inspiration that led Hans Berger to discover EEG signals (Berger, 1929). This work laid the foundation for modern neuroscience and the non-invasive study of brain activity. Building upon Berger's pioneering research, EEG has become an indispensable tool for exploring the electrical patterns originated from brain activity (Wolpaw and Wolpaw, 2012). It has enabled scientists to investigate neural mechanisms underlying cognition, perception, and motor functions. In particular, EEG has been instrumental in the development of BCIs, which translate neural signals into commands for external devices, offering new avenues for communication and control for individuals with motor impairments, and enabling the development of novel treatments (Wolpaw and Wolpaw, 2012).

In EEG-based BCIs, electrodes are placed on the user's scalp to record the brain's electrical activity. During the processing stage, the BCI system interprets these EEG signals to predict the user's intentions and provides feedback through various modalities, such as visual displays, robotic control, or virtual environments (Ramos-Murguialday et al., 2012). EEG is the preferred neuroimaging technique for BCIs because it is non-invasive, highly portable, and offers excellent temporal resolution. However, its low spatial resolution and poor signal-to-noise ratio (SNR) make it difficult to accurately extract users' intentions (Wolpaw and Wolpaw, 2012).

Accurately discerning user's intentions from EEG signals remains a significant challenge in BCI development. While promising paradigms have been established to elicit control signals for various applications, there is considerable room for improvement in the accuracy of these systems. To tackle these challenges, this Doctoral Thesis examines enhancements in EEG-based BCIs by implementing novel DL techniques within the MI paradigm, which mimics neural patterns related to motor execution and motor planning. The Doctoral Thesis is composed of

three interrelated studies, each contributing to the overarching goal of improving MI decoding accuracy and advancing our understanding of the underlying neural mechanisms. A graphical representation of the thematic consistency and contributions of the three papers is depicted in Figure 1.1.

In the first study (Perez-Velasco et al., 2022), we introduce EEGSym, a new DL architecture designed to mitigate inter-subject variability and reduce BCI inefficiency in MI classification. EEGSym incorporates inception modules, residual connections, and introduces symmetry through the mid-sagittal plane into the network architecture. Complemented by a data augmentation (DA) technique and transfer learning across different datasets, EEGSym significantly outperforms previous state-of-the-art models. With improved inter-subject MI classification across five public datasets involving 280 subjects, EEGSym achieves BCI control for 95.7% of users, demonstrating potential to overcome typical BCI inefficiencies that impede widespread adoption.

In the second study (Pérez-Velasco et al., 2024), we apply XAI techniques to EEGSym to identify and interpret the brain patterns that the DL model focuses on MI tasks. By adapting Shapley additive explanations (SHAP) to our DL network, we uncover how EEGSym's predictions leverage not only signals from the primary motor cortex but also from broader brain regions, including the prefrontal and posterior parietal cortices. This insight challenges the traditional focus of MI-based BCIs on sensorimotor rhythms, suggesting that a broader region-based approach may optimize EEG electrode placement. Furthermore, our results demonstrate that leveraging XAI-based electrode selection can maintain high classification accuracy with significantly fewer electrodes, which is beneficial for practical BCI deployment.

In the third study (Pérez-Velasco et al., 2025), we explore the relationship between motor imagery and motor execution (ME) by evaluating the direct transferability of DL models trained on ME data to MI tasks. Our findings indicate that models trained on ME data can perform comparably in MI decoding as MI-trained models, highlighting shared neural patterns between ME and MI. This approach leverages the more straightforward and familiar nature of ME tasks to enhance MI-based BCIs. It has the potential to simplify the calibration process and improve user motivation by reducing or eliminating the need for lengthy MI calibration sessions. Additionally, using DL models trained on ME data during MI sessions could help align the training with the pathways associated with lost or damaged functions, potentially enhancing rehabilitation outcomes.

Collectively, these studies contribute to advancing BCI technology by:

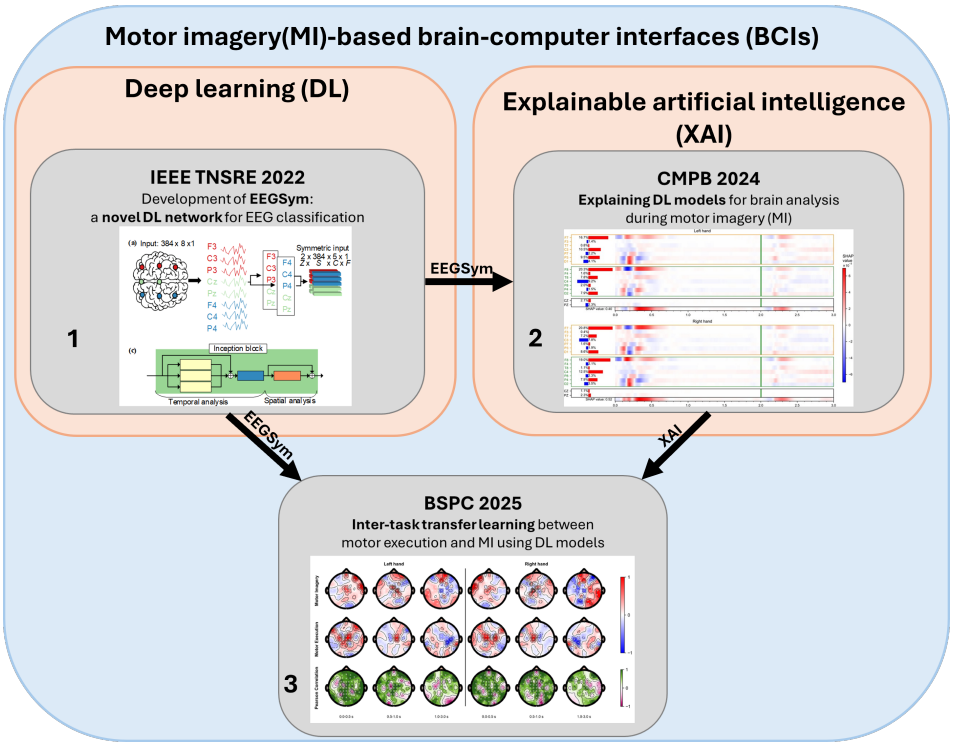


Figure 1.1: Diagram illustrating the thematic consistency among the published contributions included in this Doctoral Thesis, organized chronologically and by the main theme developed in each article. The arrows indicate how each publication builds upon previous findings. Two primary groups are identified: deep learning (DL) (Perez-Velasco et al., 2022) and explainable artificial intelligence applied to DL (Pérez-Velasco et al., 2024). Then both methodologies are used to study the transfer learning approach between motor execution and motor imagery (Pérez-Velasco et al., 2025). IEEE TNSRE: IEEE Transactions on Neural Systems and Rehabilitation Engineering; CMPB: Computer Methods and Programs in Biomedicine; BSPC: Biomedical Signal Processing and Control.

- Developing the innovative DL architecture EEGSym, which significantly enhances MI decoding accuracy while eliminating the need for calibration and achieving reliable BCI control for nearly all analyzed users.
- Providing insights into neural activation patterns through XAI techniques, which enrich our physiological understanding of MI and prove that we could optimize EEG placement on MI-based BCIs.
- Exploring direct inter-task transfer learning between ME and MI, demonstrating its potential as a viable strategy to reduce calibration burdens, improve usability, and potentially enabling more targeted

rehabilitation efforts.

By addressing both the technical challenges of EEG signal classification and the physiological insights into brain function, this Doctoral Thesis aims to enhance the effectiveness and usability of MI-based BCIs. The utilization and further enhancement of the methods developed in this Doctoral Thesis have the potential to improve the reliability, scalability, and practicality of MI-based applications in neurotechnology. Given the significance of these manuscripts in understanding the contributions of this doctoral Doctoral Thesi, they have been included in Appendix A. Additional information, metrics, and abstracts are provided for each published paper below:

EEGSym: Overcoming Inter-Subject Variability in Motor Imagery Based BCIs With Deep Learning (Perez-Velasco et al., 2022)

Sergio Pérez-Velasco, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Diego Marcos-Martínez, Roberto Hornero. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, p. 1766-1775, 2022. Impact factor in 2022: 4.909, D1 (Q1) in "REHABILITATION" (Journal Citation Reports - Clarivate Analytics).

In this study, we present a new Deep Learning (DL) architecture for Motor Imagery (MI) based Brain Computer Interfaces (BCIs) called *EEGSym*. Our implementation aims to improve previous state-of-the-art performances on MI classification by overcoming inter-subject variability and reducing BCI inefficiency, which has been estimated to affect 10-50% of the population. This convolutional neural network includes the use of inception modules, residual connections and a design that introduces the symmetry of the brain through the mid-sagittal plane into the network architecture. It is complemented with a data augmentation technique that improves the generalization of the model and with the use of transfer learning across different datasets. We compare *EEGSym*'s performance on inter-subject MI classification with ShallowConvNet, DeepConvNet, EEGNet and EEG-Inception. This comparison is performed on 5 publicly available datasets that include left or right hand motor imagery of 280 subjects. This population is the largest that has been evaluated in similar studies to date. *EEGSym* significantly outperforms the baseline models reaching accuracies of 88.6±9.0 on Physionet, 83.3±9.3 on OpenBMI, 85.1±9.5 on Kaya2018, 87.4±8.0 on Meng2019 and 90.2±6.5 on Stieger2021. At the same time, it allows 95.7% of the tested population (268 out of 280 users) to reach BCI control ($\geq$70%

accuracy). Furthermore, these results are achieved using only 16 electrodes of the more than 60 available on some datasets. Our implementation of *EEGSym*, which includes new advances for EEG processing with DL, outperforms previous state-of-the-art approaches on inter-subject MI classification.

Unraveling motor imagery brain patterns using explainable artificial intelligence based on Shapley values (Pérez-Velasco et al., 2024)

Sergio Pérez-Velasco, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Selene Moreno-Calderón, Roberto Hornero. *Computer Methods and Programs in Biomedicine*, vol. 246, p. 108048, 2024. Impact factor in 2023: 4.949, Q1 in "COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS" (Journal Citation Reports - Clarivate Analytics).

Background and objective. Motor imagery (MI) based brain-computer interfaces (BCIs) are widely used in rehabilitation due to the close relationship that exists between MI and motor execution (ME). However, the underlying brain mechanisms of MI remain not well understood. Most MI-BCIs use the sensorimotor rhythms elicited in the primary motor cortex (M1) and somatosensory cortex (S1), which consist of an event-related desynchronization followed by an event-related synchronization. Consequently, this has resulted in systems that only record signals around M1 and S1. However, MI could involve a more complex network including sensory, association, and motor areas. In this study, we hypothesize that the superior accuracies achieved by new deep learning (DL) models applied to MI decoding rely on focusing on a broader MI activation of the brain. Parallel to the success of DL, the field of explainable artificial intelligence (XAI) has seen continuous development to provide explanations for DL networks success. The goal of this study is to use XAI in combination with DL to extract information about MI brain activation patterns from non-invasive electroencephalography (EEG) signals. *Methods.* We applied an adaptation of shapley additive explanations (SHAP) to *EEGSym*, a state-of-the-art DL network with exceptional transfer learning capabilities for inter-subject MI classification. We obtained the SHAP values from two public databases comprising 171 users generating left and right hand MI instances with and without real-time feedback. *Results.* We found that *EEGSym* based most of its prediction on the signal of the frontal electrodes, i.e. F7 and F8, and on the first 1500 ms of the analyzed imagination period. We also found that MI involves a broad network not only based on M1 and S1, but also on the prefrontal cortex (PFC) and the posterior parietal cortex (PPC). We

further applied this knowledge to select a 8-electrode configuration that reached inter-subject accuracies of 86.5% ± 10.6% on the Physionet dataset and 88.7% ± 7.0% on the Carnegie Mellon University's dataset. *Conclusion.* Our results demonstrate the potential of combining DL and SHAP-based XAI to unravel the brain network involved in producing MI. Furthermore, SHAP values can optimize the requirements for out-of-laboratory BCI applications involving real users.

Bridging Motor Execution and Imagery: An Inter-Task Transfer Learning Approach (Pérez-Velasco et al., 2025)

Sergio Pérez-Velasco, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Roberto Hornero. *Biomedical Signal Processing and Control*, vol. 107, p. 107834, 2025. Impact factor in 2023: 4.907, Q1 in "BIOMEDICAL ENGINEERING" (Journal Citation Reports - Clarivate Analytics).

Motor imagery (MI)-based brain-computer interfaces (BCIs) decode movement imagination from brain activity, but improving decoding accuracy from electroencephalography (EEG) remains challenging. MI-based BCIs require calibration runs to train models; however, participant engagement cannot be externally verified. Motor execution (ME) is more straightforward and can be supervised. Deep learning (DL) leverages transfer learning (TL) to bypass calibration. This is the first work to explore wether a ME-trained DL model can reliably classify MI without finetuning to the MI task, thereby achieving direct TL between ME and MI tasks. We employed *EEGSym*, a DL network for inter-subject TL of EEG decoding, evaluating three scenarios: ME to MI, ME to ME, and MI to MI classification. We analyzed performance correlation between scenarios, and used shapley additive explanations (SHAP) to elucidate model focus patterns learned from ME or MI data. Results show that DL models trained on ME data and tested on MI perform comparably to those trained on MI data. A significant positive correlation was found between performance in ME and MI tasks for models trained on ME data. Explainable artificial intelligence (XAI) techniques revealed robust correlation between patterns in ME and MI tasks. However, between 0.5 to 1 second, the ME-trained model focused on the contralateral central region, while the MI-trained model also targeted the ipsilateral fronto-central region. Our findings demonstrate the viability of inter-task TL between ME and MI using DL models in BCI applications. This supports using ME-trained models for MI tasks to enhance targeted learning of

brain activation patterns.

1.2 Context: biomedical engineering, biomedical signal processing and machine learning

This Doctoral Thesis is situated within the interdisciplinary field of biomedical engineering, with a specific focus on biomedical signal processing and machine learning. These fields form the foundation for the development of BCIs, which enable direct communication between the brain and external devices by decoding neural activity. By addressing the challenges associated with MI-based BCIs, this research contributes to advancing assistive technologies for rehabilitation of individuals with motor impairments.

Since Hans Berger's discovery of EEG in the early 20th century, the ability to measure and analyze electrical brain activity has been instrumental in understanding neural mechanisms and developing non-invasive neurotechnological applications (Berger, 1929; Wolpaw and Wolpaw, 2012). EEG is the most widely used neuroimaging method in BCI systems due to its portability, affordability, and high temporal resolution. Despite these advantages, the EEG analysis remains challenging due to their low spatial resolution, high susceptibility to noise, and significant inter-subject and inter-session variability (Wolpaw and Wolpaw, 2012).

Biomedical engineering provides the framework to further develop BCIs by applying principles of biomedical signal processing, control theory, and machine learning to biological systems (Bronzino and Peterson, 2014). In this work, the biological system under study is the brain, analyzed through the electrical fields it produces during its activity recorded using EEG.

Biomedical signal processing is a critical component in the development of BCIs, as it facilitates the extraction and interpretation of meaningful information from EEG signals. EEG captures neural activity through electrodes placed on the scalp, recording oscillatory patterns such as sensorimotor rhythms associated with MI tasks. Signal processing techniques such as filtering, time-frequency analysis, and feature extraction play a pivotal role in isolating these patterns from noise and artifacts. By enhancing the quality and interpretability of EEG data, these methods improve the reliability of EEG decoding, which is essential for practical BCI applications.

ML, and more recently DL, has revolutionized the analysis of complex

biomedical signals. Traditional ML methods, such as support vector machines (SVM) and linear discriminant analysis (LDA), rely on handcrafted features to classify EEG signals extracted with methods like common spatial patterns (CSP) (Wolpaw and Wolpaw, 2012). In contrast, DL architectures, like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), enable end-to-end learning, automatically extracting features from raw data (Goodfellow et al., 2016). These models excel in capturing the non-linear and dynamic characteristics of EEG signals, leading to significant improvements in decoding accuracy for MI-based BCIs. Additionally, transfer learning (TL) abilities of DL could allow models to generalize across datasets, users, and sessions, reducing the need for lengthy calibration sessions and enhancing their usability (Santamaria-Vazquez et al., 2020).

Despite the advantages of DL, its adoption is often limited by its *black box* nature, which refers to the inability to understand how these end-to-end models generate their predictions. To address this challenge, XAI has emerged as a crucial tool for interpreting the decisions of complex DL models. In the context of BCIs, XAI techniques such as SHAP and layer-wise relevance propagation (LRP) provide insights into the neural features and regions that influence model predictions (Adadi and Berrada, 2018; Lundberg and Lee, 2017; Sturm et al., 2016). This enhanced transparency not only fosters trust in the technology but also deepens our understanding of the physiological mechanisms underlying MI. By revealing patterns of neural activation, XAI facilitates the optimization of electrode placement and the refinement of DL architectures, ultimately improving both model performance and practical applicability (Pérez-Velasco et al., 2024).

MI, the central paradigm of this Doctoral Thesi, activates neural pathways in the motor cortex that overlap with those used during ME (Wolpaw and Wolpaw, 2012). Understanding the physiological basis of these pathways is crucial for designing effective BCIs. Sensorimotor rhythms, originating from the primary motor cortex, provide key signals for decoding imagined movements. However, variability in the signals that record these rhythms across individuals and sessions presents a significant challenge for reliable MI classification. This Doctoral Thesis leverages advanced signal processing, DL, and XAI techniques to address these challenges, aiming to improve MI decoding accuracy while addressing session and user variability.

The interdisciplinary integration of biomedical engineering, biomedical signal processing, ML, and XAI is essential for addressing the complex challenges inherent in BCI systems. Biomedical engineering provides the platform for applying

engineering principles to medical challenges, while biomedical signal processing techniques ensure the extraction of meaningful information from noisy EEG data (Bronzino and Peterson, 2014). ML, more specifically DL, offers advanced tools for decoding the data patterns of EEG, and XAI bridges the gap between model performance and transparency, fostering trust and enhancing the interpretability of DL in BCI systems.

The contributions of this Doctoral Thesis build upon these advancements to address specific challenges in MI-based BCIs. First, novel DL architectures are developed to improve decoding accuracy and reduce inter-subject variability. Second, XAI techniques are employed to interpret these models, offering insights into neural activation patterns and guiding system optimization. Finally, the research explores TL approaches to leverage shared neural patterns across motor imagery and execution tasks. By integrating these innovations, this Doctoral Thesis aims to enhance the functionality, reliability, and accessibility of BCIs, ultimately contributing to the development of assistive technologies that improve the quality of life for individuals with motor impairments.

1.3 Brain-computer interfaces

BCIs are innovative systems that establish a direct communication pathway between the human brain and external devices, bypassing traditional neuromuscular outputs (Wolpaw and Wolpaw, 2012). By translating neural activity into actionable commands, BCIs enable individuals to control computers, prosthetics, and other assistive technologies using their thoughts alone (Ramos-Murguialday et al., 2012). This technology holds significant promise for restoring motor function, augmenting human capabilities, and providing new modes of interaction for individuals with severe motor impairments (Wolpaw and Wolpaw, 2012).

The development of BCIs involves interdisciplinary efforts, combining knowledge from neuroscience, biomedical engineering, signal processing, and ML. BCIs typically consist of several components: data acquisition, signal processing (including feature extraction, feature selection, and classification), and the application interface that provides feedback to the user (Wolpaw and Wolpaw, 2012). This general workflow scheme of BCIs is depicted in Figure 1.2. The closed-loop nature of BCIs control loop allows users to modulate their brain activity based on the feedback received, enhancing the system's efficacy over time. It is important to classify BCI systems as either synchronous or asynchronous

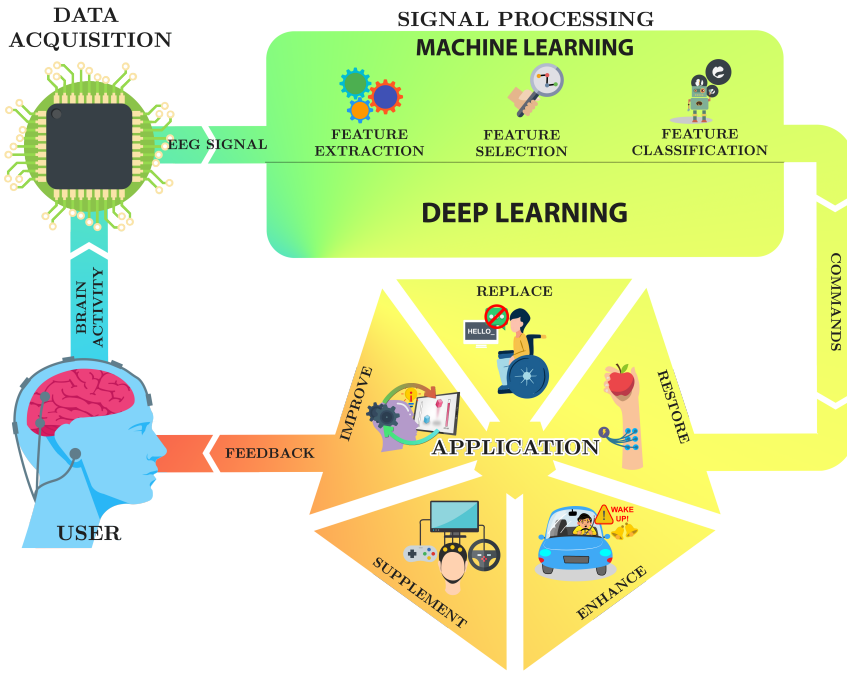


Figure 1.2: Schematic representation of a EEG-based BCI control loop. It illustrates the stages of signal acquisition, signal processing (with machine learning or deep learning), command execution, and feedback presentation. Figure adapted from (Martínez Cagigal, 2020).

(Wolpaw and Wolpaw, 2012). Synchronous, or cue-based, systems only attempt to classify the user's intent within specific time windows defined by the system, which is the approach followed in this Doctoral Thesis. In contrast, asynchronous systems operate continuously, which requires not only classifying the command but also detecting the user's intent to control, a challenge evaluated with metrics such as true positive rate (TPR) and false positives per minute (FP/min). This work focuses on maximizing decoding accuracy within the predefined intervals characteristic of synchronous MI paradigms.

The following is a concise summary of each stage:

1. **Data acquisition**: The first stage is the recording of brain activity. Various methods exist to monitor electrical, magnetic, or metabolic signals from the brain, each differing in invasiveness. However, due to its non-invasive nature, affordability, and ease of use, EEG is the most commonly employed technique in BCI studies.

2. **Signal processing**:

(a) **Machine learning**: Typically structured into the following three sequential stages.

- **Feature extraction**: With a usual first stage of pre-processing, the raw data is cleaned through filtering, artifact removal, and segmentation into epochs aligned with relevant events related to user's intentions. Key characteristics, such as time-domain patterns, spectral power in specific frequency bands, or spatial distributions, are computed from the pre-processed signals.
- **Feature selection**: The most informative features are identified, using statistical measures or algorithmic techniques, to reduce dimensionality and improve classification accuracy.
- **Classification**: ML algorithms map the selected features to specific commands, effectively translating neural activity into actionable outputs.

(b) **Deep learning**: An end-to-end paradigm that consolidates the three stages of a traditional ML pipeline into a single model.

3. **Application**: The classified commands are used to control an external device or software application, with the system providing real-time feedback that enables the user to modulate their brain activity and optimize performance over time.

The BCI performance relies on optimizing every stage, from signal acquisition and processing to command execution and feedback, to ensure consistent results across diverse users. Recognizing that each component requires continual refinement, this Doctoral Thesis focuses on enhancing the signal processing stage by integrating innovative DL techniques that boost decoding accuracy. Additionally, it emphasizes improving the interpretability of the DL models' classifications, thereby providing insights into the user's brain activity during BCI operation.

In this section, we explore the fundamental aspects of BCIs by providing an overview of the brain's structure and function (Section 1.3.1), discussing various methods used to measure brain signals (Section 1.3.2), and examining the different types of control signals employed in BCIs (Section 1.3.3).

1.3.1 The brain

The human brain is an intricate and highly specialized organ that serves as the control center of the nervous system. Weighting an average of 1.5 kg and comprising approximately 86 billion neurons interconnected through complex networks and 85 billion non-neuronal cells (Herculano-Houzel, 2009), the brain is responsible for processing sensory information, executing motor actions, and facilitating cognitive functions, such as learning, memory, or decision-making. In conjunction with the spinal cord, it constitutes the central nervous system (CNS), which processes sensory stimuli associated with external or internal events and generates responses by eliciting hormonal or neuromuscular outputs (Wolpaw and Wolpaw, 2012).

The majority of the brain's higher-order cognitive processing occurs in the cerebral cortex, the brain's outermost layer. This highly folded structure significantly increases the surface area available for synaptic connections, enhancing computational power without the need of a larger brain volume. The gyri (ridges) and sulci (grooves) that characterize the cortex are particularly abundant in humans, reflecting the complexity required for functions, such as reasoning, problem-solving, and abstract thought (Herculano-Houzel, 2009). To better understand and study the brain's functions, scientists have organized the cerebral cortex into distinct anatomical regions based on both structural and functional characteristics (Wolpaw and Wolpaw, 2012). These include the four primary lobes:

- **Frontal Lobe**: Located at the front of the brain, it is involved in executive functions, voluntary motor activity, planning, and reasoning.
- **Parietal Lobe**: Positioned behind the frontal lobe and separated from it by the central sulcus, it processes tactile sensory information such as touch, pressure, and pain.
- **Occipital Lobe**: Situated at the back of the brain, it is primarily responsible for visual processing.
- **Temporal Lobe**: Found beneath the frontal and parietal lobes, it plays a key role in auditory processing and memory formation.

An overview of these and other anatomical classifications of brain regions is illustrated in Figure 1.3.

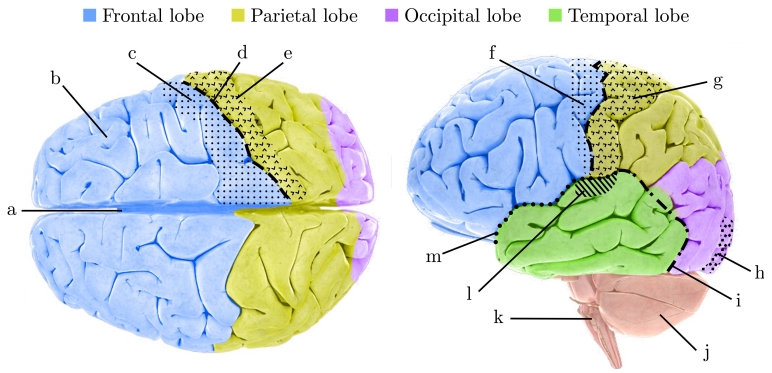


Figure 1.3: Anatomical regions of the human cerebral cortex are depicted from superior (left) and lateral (right) perspectives. The four primary lobes are distinguished by unique colors: blue for the frontal lobe, yellow for the parietal lobe, purple for the occipital lobe, and green for the temporal lobe. The labeled anatomical features include: (a) interhemispheric fissure, (b) prefrontal association cortex, (c) precentral gyrus, (d) central sulcus, (e) postcentral gyrus, (f) primary motor cortex, (g) primary somatic sensory cortex, (h) primary visual cortex, (i) preoccipital notch, (j) cerebellum, (k) brainstem, (l) primary auditory cortex, and (m) lateral sulcus. Figure reproduced with permission from (Martínez Cagigal, 2020)

In addition to the cerebral cortex, other key structures play essential roles in brain function. The brainstem, located at the base of the brain, regulates vital life functions such as breathing, heart rate, and consciousness. Meanwhile, the cerebellum, situated beneath the occipital lobe, is crucial for motor coordination, precision, and balance (Wolpaw and Wolpaw, 2012). Together, these regions support the brain's intricate interplay of functions, enabling the interaction with the body and environment.

Beyond anatomical distinctions, the brain can also be studied through functional networks, which provide insights into dynamic activity patterns. For example, the sensorimotor network connects motor and somatosensory regions, making it a focal point in MI-based BCIs. These interfaces leverage electrical signals from these specific regions, to translate neural activity into actionable commands (Wolpaw and Wolpaw, 2012).

The motor cortex, spanning the frontal and parietal lobes, is particularly significant in BCIs focused on motor functions. It includes the primary motor cortex (M1) and primary somatosensory cortex (S1), which work in tandem to initiate and refine voluntary movements (Miller and Hatsopoulos, 2012). M1 generates neural signals that control motor execution, while S1 processes sensory feedback, enabling fine motor control (Miller and Hatsopoulos, 2012).

Understanding this interplay has been crucial for optimizing electrode placement during EEG signal acquisition, thereby improving the performance of BCI systems focused on this endogenous paradigm.

The presence of distinct specialized networks highlights the intricate complexity of the brain and its profound implications for the development of advanced neurotechnological applications, such as BCIs. By integrating anatomical, functional, and network-based perspectives, researchers continue to uncover the mysteries of brain function and its potential for transformative technologies.

1.3.2 Measuring brain activity: Electroencephalography

The cerebral cortex, as discussed earlier, is the primary site of higher-order cognitive functions such as reasoning, problem-solving, and voluntary motor control. This outer layer is where critical operations related to conscious control take place, making it a key target for measuring and interpreting brain activity (Wolpaw and Wolpaw, 2012).

Researchers have developed invasive methods that offer exceptionally high spatial resolution by recording signals beneath the skull. For instance, electrocorticography (ECoG) places electrodes directly on the cortical surface, and microelectrode arrays can even capture the activity of individual neurons (Wolpaw and Wolpaw, 2012). While these approaches provide unmatched precision, they require surgical intervention and thus have limited practicality in most scenarios. Positron emission tomography (PET) despite having a non-invasive measuring device, require the user to have injected radioactive compounds (radiotracers), that bind to different chemicals present in the brain (Wolpaw and Wolpaw, 2012). It is more applicable than the other methods requiring surgical intervention mentioned above, but its application due to its low spatial resolution and the required equipment is mostly found in detecting infections or tumors (Wolpaw and Wolpaw, 2012).

In contrast, non-invasive methodologies have become the standard for most applications due to their accessibility and ease of use. Techniques such as EEG, and magnetoencephalography (MEG) measure brain signals from outside the skull, albeit through different underlying principles and with varying degrees of spatial and temporal resolution (Wolpaw and Wolpaw, 2012). Additionally, functional near-infrared spectroscopy (fNIRS) detects changes in oxygenated and deoxygenated hemoglobin in the cortex resulting from brain activity (Wolpaw and Wolpaw, 2012). Though fNIRS offers relatively good spatial insight into

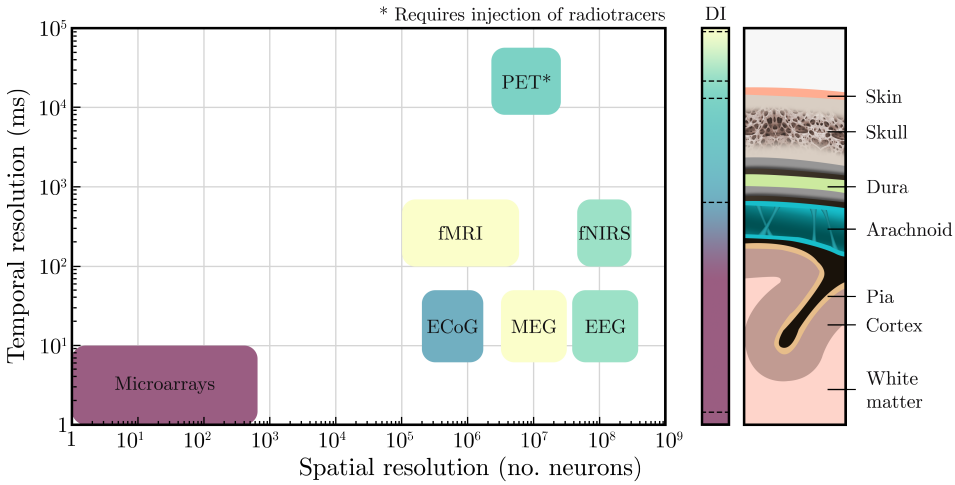


Figure 1.4: Temporal and spatial resolution of different brain activity measurement techniques. The color gradient on the right indicates the degree of invasiveness (DI) according to brain layers. ECoG: electrocorticography, MEG: magnetoencephalography, EEG: electroencephalography, fMRI: functional magnetic resonance imaging, fNIRS: functional near-infrared spectroscopy, PET: positron emission tomography (requiring injection of radiotracers), and Microarrays (intracortical electrodes) placed directly in neural tissue. Figure reproduced with permission from (Martínez Cagigal, 2020)

hemodynamic activity, it is more limited in temporal dynamics compared to EEG or MEG. Despite these trade-offs, such non-invasive methods can still capture critical brain activity needed to interpret user's intentions, making them especially suitable for BCIs, where practicality and safety are key for general adoption.

Figure 1.4 depicts presents a two-dimensional comparison of these main techniques used to monitor brain activity. The vertical axis depicts temporal resolution, where lower values indicate an ability to capture faster neural events, while the horizontal axis shows spatial resolution, in which lower values reflect greater precision in pinpointing activity within the tissue. A color gradient encodes the degree of invasiveness (DI), ranging from fully extracranial methods such as EEG and MEC to intracortical micro-electrode arrays. Metabolic or hybrid techniques, including fNIRS and PET (which requires radiotracer injection), are also positioned. This overview highlights the trade-offs each modality entails, helping to have an overview of each tool's balance of resolution and DI.

Among the non-invasive methods, EEG stands out as the most widely adopted in BCI research, owing to its comparatively low cost, ease of use, and excellent temporal resolution (Wolpaw and Wolpaw, 2012). Unlike MEG, which typically involve specialized and expensive equipment, EEG systems can be used with

relatively inexpensive hardware and minimal infrastructure. This accessibility suits a broad range of applications, from clinical rehabilitation to consumer-oriented devices.

EEG operates by placing electrodes on the scalp to record electrical potentials generated primarily by the synchronized activity of neurons in the cerebral cortex. Its millisecond-level temporal resolution enables real-time observation of dynamic processes, such as motor planning, decision-making, and sensory perception. EEG has some limitations: its spatial resolution is reduced by signal distortion as electrical activity passes through various tissue layers, and it's vulnerable to artifacts from electromyography (muscle activity), electrooculography (eye blinks), and external electrical interference. However, advanced signal processing techniques help mitigate these issues and extract meaningful patterns from the data. Consequently, EEG's combination of high temporal resolution and user-friendly implementation has made it the method of choice in most BCI applications (Wolpaw and Wolpaw, 2012).

Because EEG is generated by large, synchronized populations of neurons engaged in multiple cognitive processes simultaneously, its signal is inherently complex and difficult to interpret through simple visual inspection in the temporal domain. Frequently exhibiting oscillatory behavior with distinct spatio-temporal patterns, EEG is often transformed into the frequency domain to more effectively capture its underlying rhythmic activity. A substantial body of research indicates that certain brain functions or mental states correlate with heightened or diminished activity in specific frequency bands. Traditionally, EEG signals have been categorized into five distinct frequency bands (Wolpaw and Wolpaw, 2012). These frequency bands provide a useful framework for interpreting EEG signals, allowing researchers to investigate various mental states or cognitive processes relevant to BCI development:

- **Delta** (<4 Hz): Associated with deep sleep stages.
- **Theta** (4–8 Hz): Linked to drowsiness, meditation, and creative states.
- **Alpha** (8–13 Hz): Present during relaxed wakefulness with closed eyes or focused concentration.
- **Beta** (13–30 Hz): Related to active thinking and concentration.
- **Gamma** (30–100 Hz): Involved in higher-order cognitive functions and sensory processing.

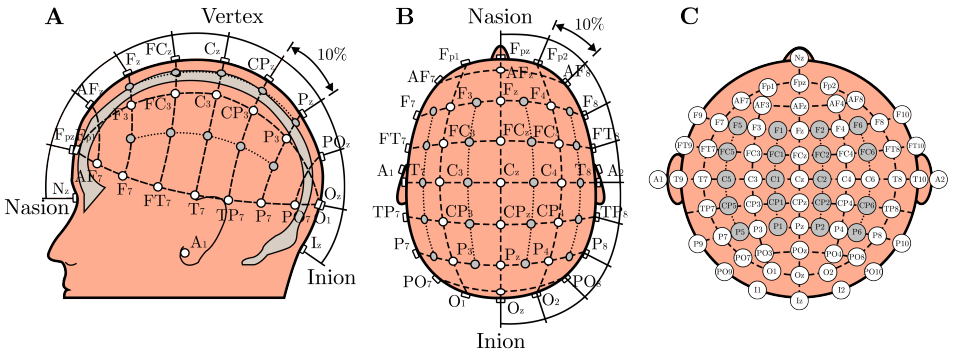


Figure 1.5: 10-10 international system for EEG electrode placement. (A) Lateral view, (B) Top view, and (C) schematic projection. The 10–10 system is an extension of the traditional 10–20 method, enhancing spatial resolution by placing electrodes at intervals corresponding to 10% of the distances between key skull landmarks (e.g.: nasion, inion). Figure reproduced with permission from (Martínez Cagigal, 2020).

EEG electrode placement is far from random, as it follows standardized protocols to allow consistent and reproducible EEG across experimental setups. Early systems, such as the international 10–20 configuration, provided a basic, low-resolution overview of brain activity, but advances in technology and cost reductions led to the application of higher-resolution systems like the 10–10 or 10–5 configurations. Figure 1.5 depicts the international 10–10 electrode-placement system. Panel A shows a lateral view of the head, Panel B a superior (top) view, and Panel C a flattened schematic projection that labels every site. The 10–10 system refines the classic 10–20 montage by positioning electrodes at 10 % intervals between the nasion and inion, thereby doubling the spatial sampling density and improving the resolution of cortical activity mapping. Despite the increased spatial detail offered by these denser arrays, they come with drawbacks such as longer setup times and reduced user comfort. Consequently, practical BCI applications tend to favor configurations with fewer electrodes, with most commercial devices using 8 or 16 channels. Depending on the task, these electrodes can be concentrated within a specific region to achieve higher spatial resolution, albeit at the expense of obtaining a comprehensive global overview of brain activity.

1.3.3 Brain-computer interface control signals and paradigms

As previously mentioned, each EEG channel captures the simultaneous activity of millions of neurons. Although this activity originates in the brain's cortex,

it is recorded at the scalp. As the electrical signals travel from the cortex to the electrodes, they pass through multiple layers (e.g., arachnoid, dura, skull, skin), each altering the signal. These alterations make it challenging to accurately pinpoint the signal's original source (Wolpaw and Wolpaw, 2012). In addition to its low spatial resolution, EEG is highly susceptible to various sources of noise, resulting in a low SNR. Combined, these issues make it extremely challenging to accurately discern a user's intent from the EEG data. EEG captures fine-grained temporal changes in brain activity but from millions of neurons at the same time, so reaching the level of discerning users thoughts is far from possible with this type of neuroimaging technique (Wolpaw and Wolpaw, 2012). However, EEG can detect changes that are common to these cluster of neurons, so BCIs rely on detecting specific patterns in brain activity that involve the synchronous activation or deactivation of a wide number of localized neurons. This evoked activity discernible on the EEG are known as control signals, and they are essential for translating neural activity into commands. These control signals can be categorized based on their origin (Kumar et al., 2020): exogenous control signals are based on natural responses of the brain elicited by external stimuli, whereas endogenous control signals are based on brain activity generated voluntarily by the user. After identifying and characterizing these control signals, researchers have developed a range of BCI paradigms that leverage them. Some of the most widely used control signals in EEG-based BCIs and their corresponding paradigms include (Wolpaw and Wolpaw, 2012):

1. **Event-Related Potentials (ERPs)**

   They are time-locked brain responses to external sensory stimuli. Being the most commonly used:

   - **Visual Evoked Potentials (VEPs)**: These are brain responses elicited by visual stimuli, without the need of the user to engage in any voluntary mental task.
     - **Steady-State Visually Evoked Potentials (SSVEPs)**
       SSVEPs are neural responses elicited in the occipital cortex when a person gazes steadily at a visual stimulus that flickers at a constant rate (Wolpaw and Wolpaw, 2012). These responses, emerging from basic sensory processing, serve as exogenous control signals that can be reliably activated without the need for any specific cognitive task or prior training. In a standard SSVEP-based BCI setup, a matrix of visual commands, each flashing at a distinct frequency,

is presented on a display. When the user directs their gaze to one particular element, the brain's visual cortical activity synchronizes to that specific frequency. By analyzing the frequency components of the EEG from that region, the system identifies the dominant frequency and translates it into the corresponding selection. It is important to note that the range of effective frequencies is limited by the display's refresh rate and the potential overlap from harmonic frequencies.

  - **Code-modulated VEPs (c-VEPs)**
    c-VEPs are elicited by flashing stimulation sequences that follow pseudo-random codes, generating evoked responses that are time-locked to the specific code the user focuses on (Martínez-Cagigal et al., 2021). To use c-VEPs for communication, a matrix of visual targets is displayed, each flashing according to its own distinct stimulation sequence. This distinctiveness is achieved either by assigning different pseudo-random codes to each target or by employing a circular shifting method that uses time-shifted versions of the same code. When a user directs their attention to a particular target, the BCI system computes the correlation between the recorded EEG activity and each stimulation sequence, thereby identifying the intended command with calibration times as brief as one minute per pseudo-random code (Martínez-Cagigal et al., 2021). However, c-VEP-based BCIs demand precise time synchronization. The most practical approach is to employ a single pseudo-random code and present each target as a time-shifted variant (Martínez-Cagigal et al., 2021). This method, however, constrains the number of distinct commands to the number of available uncorrelated codes.

- **P300**
  A positive ERP that peaks roughly 300 ms after an infrequent stimulus appears (Polich, 2007). It is elicited automatically in the parietal region of the brain by external events and therefore requires little training, making it popular for speller applications making use of the "oddball" paradigm (Santamaria-Vazquez et al., 2020). However, the P300 is unique in that its amplitude is further enhanced when users actively engage in a mental task (e.g., counting each occurrence) following the stimulus, so it falls also a bit into the endogenous category.

- **Error-related pontential (ErrPs)**

  ErrPs are distinctive brain responses that occur when a user detects an error in the system's output. Characterized by an initial negative deflection over fronto-central regions, followed by a subsequent positive component, these potentials reflect the brain's internal monitoring and evaluative processes. They can be reliably detected on a single-trial basis and have been effectively used in BCIs to either correct erroneous commands in real time or to adapt the system through reinforcement learning, ultimately enhancing overall performance (Chavarriaga et al., 2014).

2. **Slow Cortical Potentials (SCPs)**

   SCPs are slow event-related voltage shifts recorded over sensorimotor areas that occur in relation to both actual movements and MI. Typically characterized by negative potential changes preceding movement or cognitive tasks, these signals reflect the preparatory activation of the cortex. A well-known example is the Bereitschaftspotential (readiness potential), which begins about 500–1000 ms before a self-initiated movement, indicating activity in regions like the supplementary motor area and primary motor cortex (Wolpaw and Wolpaw, 2012). Another related component, the contingent negative variation, emerges 200–500 ms after a warning stimulus and persists until an imperative signal occurs, with its amplitude modulated by motivational and task-specific factors (Wolpaw and Wolpaw, 2012). These slow, time-locked phenomena underscore the role of sensorimotor cortices in both movement execution and imagery (Wolpaw and Wolpaw, 2012).

3. **Sensorimotor Rhythms (SMR)** SMRs are oscillatory patterns recorded over the sensorimotor cortices, specifically, the posterior frontal and anterior parietal areas (Pfurtscheller and Lopes da Silva, 1999). These rhythms are typically identified over three frequency bands: alpha (8–13 Hz), high beta band (18–30 Hz), and gamma (>30 Hz). While EEG predominantly captures alpha, beta, and lower gamma frequencies, higher-frequency activity is better detected with techniques like ECoG and MEG (Miller and Hatsopoulos, 2012). As a result, most SMR studies have relied on EEG, focusing primarily on alpha (also called mu) and beta rhythms. SMRs are usually elicited with motor behavior (Wolpaw and Wolpaw, 2012). A key feature of this control signal is that it can reflect actual ME or merely the mental rehearsal of movement without physical execution (i.e., MI). This property

makes attractive the use of this control signal for rehabilitation procedures (Sebastián-Romagosa et al., 2020).

Each control signal has unique characteristics that influence its suitability for different BCI applications, and these differences are not only engineering-related but also grounded in distinct neurophysiological mechanisms that suggest different translational paths to clinical use. For instance, endogenous paradigms such as MI are well aligned with neurorehabilitation settings where repeated attempted/imagined movement paired with feedback aims to promote use-dependent plasticity (Bundy et al., 2017; Moldoveanu et al., 2019; Sebastián-Romagosa et al., 2020). MI-based BCIs often require extensive user training to achieve reasonable performance in decoding user intentions. Nevertheless, beyond decoding accuracy, it is crucial to ensure that the decoded features reflect physiologically plausible motor-network recruitment and can be reliably integrated into closed-loop feedback protocols. In contrast, exogenous paradigms based on ERPs or VEPs (e.g., P300 or c-VEPs) leverage stimulus-locked responses automatically elicited by external stimuli without requiring prior training, making them ideal for communication applications (Martínez Cagigal, 2020; Martínez-Cagigal et al., 2021; Santamaria-Vazquez et al., 2020).

1.4 Deep learning and explainable artificial intelligence

In recent years, DL has emerged as the best performing approach to a wide range of machine learning problems, enabling significant advancements in fields such as computer vision, natural language processing, and biomedical signal processing (Goodfellow et al., 2016). DL models, particularly neural networks with multiple layers, have demonstrated remarkable capabilities in automatically extracting hierarchical features from raw data, outperforming traditional machine learning methods in various tasks.

The application of DL to EEG data has opened new possibilities in BCI systems, offering improved accuracy in decoding neural signals associated with cognitive and motor functions (Lawhern et al., 2018). However, the complexity and opacity of DL models often render them as *black boxes*, making it challenging to interpret their decision-making processes. This lack of transparency poses significant concerns in critical applications like healthcare, where understanding the reasoning behind model predictions is crucial for trust, accountability, and

further scientific insight.

XAI addresses this challenge by introducing techniques that illuminate the decision-making processes within DL models (Adadi and Berrada, 2018). In the context of EEG analysis and BCIs, XAI techniques can help uncover the neural patterns or features that DL models exploit. This enhanced transparency deepens our understanding of the brain functions underlying various BCI control signals and may even reveal previously undiscovered signals or interactions among them.

This section introduces the fundamental concepts of DL and underscores the importance of model interpretability.

1.4.1 Introduction to deep learning

To first tackle the definition of deep learning, we must first explore the broader concepts of artificial intelligence (AI) and its ML subfield, since DL is also a subset of ML. In the beginning of the application of AI, it consisted on solving problems that were intellectually difficult for humans but relatively easy for computers. The challenge at the time for AI was to enable computers to perform tasks traditionally requiring human intelligence. A pivotal breakthrough within AI was the advent of ML, which enables systems to automatically extract knowledge by discovering patterns directly in raw data, rather than relying solely on pre-programmed rules. This ability allows computers to tackle complex, real-world problems and make decisions that often mirror human judgment. However, traditional ML methods typically required feature engineering (handcrafted transformations based on prior task-specific knowledge) to derive useful representations (Goodfellow et al., 2016). While effective, this approach inherently limited performance by constraining the depth and richness of the features that could be extracted. DL is a subset of ML that focuses on neural networks with many layers, known as deep neural networks (Goodfellow et al., 2016). These models are inspired by the structure and function of the human brain, consisting of interconnected layers of artificial neurons that can learn complex patterns from large datasets. Each layer applies a set of simple functions to transform the input into a new representation. By composing these layers, the network gradually captures complex patterns and concepts, effectively bypassing the need for manually engineered features. The key capabilities of DL include (Goodfellow et al., 2016):

- The ability to automatically learn multiple levels of abstraction, capturing both low-level and high-level features directly from raw data.

- By applying non-linear activation functions, DL networks can represent intricate, non-linear relationships inherent in data, enabling them to tackle complex tasks effectively.
- These models are designed to handle large, high-dimensional datasets efficiently, leveraging parallel computing resources, such as Graphics Processing Units (GPUs), to accelerate training and inference processes.
- DL models are adept at learning abstractions that generalize effectively across diverse tasks and experimental conditions. They can transfer knowledge acquired from one domain to another, even with minimal additional data, through a process known as TL. By fine-tuning pre-trained networks on new datasets, researchers not only reduce training time but also enhance performance on novel tasks, making these models remarkably adaptable and efficient across various domains.

DL has undergone significant evolution since its inception (Goodfellow et al., 2016). In the 1940s, early models like the perceptron aimed to mimic biological learning processes, laying the groundwork for artificial neural networks. The 1980s saw a resurgence with the connectionism movement, introducing key concepts such as distributed representations and the backpropagation algorithm, which enhanced the training of multilayer networks (Rumelhart et al., 1986). Despite these advancements, progress stalled in the 1990s due to computational limitations and the popularity of alternative ML methods (Goodfellow et al., 2016).

2006 marked a revival of DL, driven by the availability of large datasets and advancements in hardware, particularly GPUs. Researchers demonstrated that GPUs could significantly accelerate neural network training, making it feasible to develop deeper and more complex models (Goodfellow et al., 2016). This resurgence has led to DL's prominence in various applications today. Common DL architectures include (Goodfellow et al., 2016):

1. **Multilayer perceptrons (MLPs)**

   MLPs are the foundational architecture of deep learning, consisting of multiple layers of interconnected neurons. Each neuron in a layer is connected to every neuron in the subsequent layer, facilitating the transformation of input data into a desired output. MLPs are particularly suited for tasks where data can be represented in a tabular form, such as in finance for credit scoring or in healthcare for predicting patient outcomes (Goodfellow et al., 2016).

2. **Convolutional Neural Networks (CNNs)**

   CNNs are designed to process grid-like data structures, making them highly effective for image and video analysis (LeCun and Bengio, 1998). They utilize convolutional layers to capture spatial hierarchies and patterns more efficiently than MLPs, enabling applications in object detection, facial recognition, and medical image analysis. A pivotal moment was the 2012 ImageNet Large Scale Visual Recognition Challenge, where the AlexNet CNN achieved a top-5 error rate of 15.3%, significantly outperforming the runner-up's 26.2%. This breakthrough demonstrated the potential of DL models, and in particular CNNs, leading to widespread interest beyond the research community (Krizhevsky et al., 2012).

3. **Recurrent Neural Networks (RNNs)**

   RNNs are tailored for sequential data, where the order of data points is crucial (Rumelhart et al., 1986). They maintain internal memory states that capture information about previous inputs, making them effective for time-series analysis and natural language processing (NLP). Applications include language modeling, speech recognition, and sentiment analysis. However, traditional RNNs face challenges with long-term dependencies, which led to the development of Long Short-Term Memory (LSTM) networks addressing these limitations (Hochreiter and Schmidhuber, 1997).

4. **Autoencoders**

   An autoencoder is a specialized type of artificial neural network designed to learn efficient codings of unlabeled data in an unsupervised manner. Its primary purpose is to discover an effective representation of input data, typically for dimensionality reduction or feature extraction (Bourlard and Kamp, 1988; Hinton and Zemel, 1993). The architecture of an autoencoder comprises two main components (Hinton and Zemel, 1993):

   - Encoder: This part of the network compresses the input data into a latent-space representation, effectively reducing its dimensionality.
   - Decoder: This component reconstructs the original data from the compressed representation, aiming to make the output as close to the input as possible.

   The training process involves minimizing the difference between the input and its reconstruction, thereby ensuring the model captures the most salient

features of the data. Autoencoders have been applied to various tasks, including data denoising (Vincent et al., 2010), anomaly detection (Neloy and Turgeon, 2024), and generative modeling (Block et al., 2022). Variants such as variational autoencoders (VAEs) have further expanded their applicability by enabling the generation of new data samples from the learned latent space (Kingma and Welling, 2019).

5. **Generative Adversarial Networks (GANs)**

   GANs consist of two neural networks—the generator and the discriminator—that are trained simultaneously through adversarial processes (Goodfellow et al., 2014). The generator creates synthetic data, while the discriminator evaluates its authenticity. GANs have been utilized in image generation (Trevisan de Souza et al., 2023), data augmentation (Luo and Lu, 2018), and even in creating art (Shahriar, 2021). In the medical field, GANs assist in generating synthetic medical images to augment training datasets, improving the robustness of diagnostic models (Gao et al., 2023).

6. **Transformers**

   Transformers utilize self-attention mechanisms to process input data, effectively modeling long-range dependencies without relying on recurrent structures (Vaswani et al., 2017). They have revolutionized NLP, leading to the development of models like BERT and GPT, which excel in tasks such as machine translation, text summarization, and question-answering systems (Devlin et al., 2019). Beyond NLP, transformers have been adapted for image processing tasks, demonstrating versatility across domains (Dosovitskiy et al., 2021).

These various DL architectures have each found applications in each domain. However, some have demonstrated remarkable versatility, being effectively translated to other domains. In the context of EEG analysis, for example, CNNs have emerged as the architecture of choice. This is largely because EEG data, with its inherent spatial and temporal structure, benefits greatly from the convolution operation capability to extract meaningful features (Lawhern et al., 2018; Santamaria-Vazquez et al., 2020; Schirrmeister et al., 2017).

1.4.2 Importance of explainability in deep learning models

Despite their impressive performance, DL models are often criticized as *black boxes* due to their opaque decision-making processes. Because a single network often manages all processing stages (from feature extraction to classification) without clear intermediate outputs, it can be challenging to understand how the final predictions are made. This inherent complexity introduces a trade-off between accuracy and interpretability (Adadi and Berrada, 2018). To overcome this limitation, the XAI field has emerged, with the aim of illuminating the inner workings of DL models and reveal the underlying patterns that drive their predictions. It provides insights into how predictions are generated, which in turn fosters trust and enhances the safety of AI systems. This transparency is essential for increasing confidence in model's predictions and promoting the broader adoption of AI technologies, specially so in the healthcare field (Kamath and Liu, 2021).

XAI methods applied to trained DL models can generally be divided into two broad categories: perturbation-based and backpropagation-based techniques (Shrikumar et al., 2017). Perturbation-based XAI techniques modify the input and analyze how these alterations influence the network's output, while backpropagation-based XAI techniques traces the output's significance backward through the network to highlight which input features were most influential.

Among the most well-known perturbation-based techniques are occlusion sensitivity analysis (Zeiler and Fergus, 2014) and local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016). Occlusion sensitivity analysis, originally introduced for explaining DL models in computer vision, systematically masks portions of an input image to observe how the classifier's output changes. This method helps determine whether a model is truly localizing an object or merely relying on surrounding context (Zeiler and Fergus, 2014). On the other hand, LIME generates perturbed samples around a specific input and fits a simple, interpretable model (typically a sparse linear model) to approximate the local behavior of the original network. By analyzing these approximations, LIME identifies which features contribute most to a model's prediction, providing interpretable explanations across different types of data (Ribeiro et al., 2016).

For backpropagation-based techniques, some of the most widely used approaches include layerwise relevance propagation (LRP) (Binder et al., 2016), gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2019), or integrated gradients (Sundararajan et al., 2017). LRP assigns relevance scores

to input features by redistributing the model's output backward through its layers, ensuring that highly relevant features receive stronger attribution (Binder et al., 2016). Grad-CAM, primarily used in convolutional neural networks, highlights the most important regions of an image by computing the gradient of the target class with respect to feature maps, producing heatmaps that indicate areas of high importance (Selvaraju et al., 2019). Integrated gradients, in contrast, computes the average gradient of the model's predictions along a path from a baseline (e.g., a zeroed-out input) to the actual input (Sundararajan et al., 2017).

SHAP (Lundberg and Lee, 2017) was introduced as a unifying framework for various XAI techniques, providing a theoretically grounded approach to model interpretability. SHAP values are based on three fundamental properties: local accuracy, missingness, and consistency. These properties ensure that the assigned feature importance values meaningfully reflect the model's decision-making process. The authors demonstrated how existing explanation methods adhered to one or more of these principles while also proposing small modifications to ensure they satisfied all three (Lundberg and Lee, 2017). By enforcing these criteria, SHAP guarantees a unique and consistent interpretation of model predictions, enhancing the reliability of explanations across different applications (Lundberg and Lee, 2017). Figure 1.6 displays a violin plot that summarizes SHAP values for a tabular model. Each horizontal violin corresponds to a feature; its width shows the distribution of SHAP values (i.e., the contribution of that feature to the model's prediction) along the x-axis. Points within each violin are colored from blue to pink, indicating low- to high-valued observations of the feature, so one can see at a glance whether large or small feature values tend to push the prediction upward (positive SHAP) or downward (negative SHAP). This visualization therefore conveys, in a single panel, both the importance and the direction of influence of every feature in the dataset.

As DL continues to push the boundaries of AI applications, the need for interpretability remains essential. The development of XAI methods has provided valuable tools to enhance transparency, helping researchers and practitioners understand the reasoning behind complex DL models. Perturbation-based methods, backpropagation-based techniques, and unifying frameworks like SHAP enable researchers and practitioners to gain insights into the decision-making processes of DL models. Adopting these techniques effectively bridges the gap between achieving high performance and ensuring interpretability. In the context of EEG analysis, where the *black box* nature of DL models makes it difficult to see which brain patterns they focus on, XAI techniques are essential for revealing

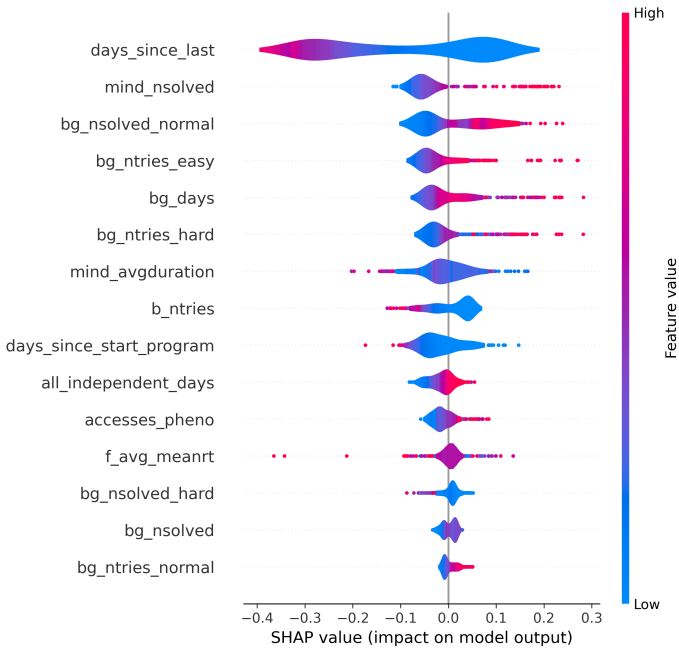


Figure 1.6: Example of a violin plot illustrating shapley additive explanations (SHAP) values for tabulated data. The x-axis indicates the impact of each feature on the model's output, while the color scale (blue to pink) reflects the feature value from low to high.

new insights into brain activity, uncovering novel control signals, and refining BCI paradigms.

1.5 State-of-the-art

The primary objective of this Doctoral Thesis is to design, develop, and assess innovative DL methodologies that improve MI decoding in EEG-based BCI systems, while offering physiological insights into the underlying neural processes. To achieve this, the research is structured around three interconnected studies, each addressing a distinct goal:

- Integrate and adapt state-of-the-art advancements in a DL architecture, called *EEGSym*, to improve MI classification accuracy (Perez-Velasco et al., 2022).
- Develop and assess the effectiveness of a SHAP-based XAI technique to identify and interpret the neural patterns that the DL model relies on,

providing a clearer understanding of the brain dynamics involved in MI (Pérez-Velasco et al., 2024).

- Further investigate the neural mechanisms underlying MI and its relationship with ME by leveraging *EEGSym* and the previously developed SHAP-based XAI framework (Pérez-Velasco et al., 2025).

Accordingly, the following subsections provide a state-of-the-art review of each of these topics, outlining the key advancements and challenges that this Doctoral Thesis seeks to address.

1.5.1 Deep learning methods for motor imagery decoding

A significant shortcoming of current BCIs is their limited performance, especially in systems based on MI. This paradigm under study in this Doctoral Thesis, is the mental rehearsal of a movement without actually performing it, producing characteristic modulations of SMR in the EEG. BCIs translate these endogenous patterns into commands that can control external devices or drive motor-rehabilitation protocols. Yet, despite its intuitive nature, the MI paradigm still struggles to deliver high performance across users and sessions. Although an accuracy of $\geq$70% is generally considered the minimum threshold for effective BCI control in binary scenarios (Lee et al., 2019; Meng and He, 2019), systems that perform only near this level may render nearly half of the potential user population unable to operate the interface reliably. Failure to reach the 70% accuracy threshold for a user has been termed BCI inefficiency, and it is estimated to affect between 10 and 50% of potential BCI users (Alkoby et al., 2018). Furthermore, one of the most extended applications of MI-based BCIs is its use in rehabilitation (Bundy et al., 2017; Moldoveanu et al., 2019; Sebastián-Romagosa et al., 2020). In this context, high accuracy is essential because it enhances the user's sense of agency, which in turn boosts motivation, a critical factor for successful rehabilitation outcomes (Jeunet et al., 2019).

The state-of-the-art in MI classification from EEG has evolved considerably over the years. Classical ML methods based on feature engineering, such as CSP and its variants like filter bank CSP (FBCSP) (Ang et al., 2008), as well as approaches based on Riemannian geometry (Congedo et al., 2017), have been widely used to extract meaningful features from the EEG. These feature extraction methods were combined with classifiers like linear discriminant analysis (LDA) or support vector machines (SVM) (Wolpaw and Wolpaw, 2012). However, this

pipeline typically requires extensive per-session calibration due to inter-subject and inter-session variability in EEG signals (Saha and Baumert, 2020). Inter-subject variability means that a model trained on one person's EEG data often fails to perform well when applied to another individual. Similarly, the inter-session variability makes it difficult to train a model on data from one session that will work effectively in future sessions for the same user. This inter-session variability arises from a confluence of physiological, psychological, and technical factors (Saha and Baumert, 2020). Within the same day, fluctuations in the user's state such as mental fatigue, shifts in attention, and motivation can alter the recorded EEG signals. When comparing across different days, these effects are often amplified by long-term neural plasticity, learning of the task, and unavoidable inconsistencies in electrode positioning and scalp impedance. Such non-stationarities in the signal pose a significant challenge for traditional models, highlighting the critical need for architectures that can generalize across these variations without requiring recalibration for every new session. These variabilities restrict the generalizability of models trained on a single subject or session, and hinders the use of large amounts of data from multiple sessions or subjects.

In contrast, DL approaches have demonstrated superior performance by automatically learning complex features from raw EEG data, thereby reducing the reliance on handcrafted features. Furthermore, these features have shown to generalize better across subjects, reaching superior performance on inter-subject scenarios (Lawhern et al., 2018). Notably, CNNs have proven particularly effective for EEG decoding (Lawhern et al., 2018; Schirrmeister et al., 2017). Schirrmeister et al. (2017) introduced two CNN architectures (*ShallowConvNet* and *DeepConvNet*) that outperformed traditional ML methods, highlighting the potential of DL in EEG analysis. Lawhern et al. (2018) developed *EEGNet*, a compact CNN tailored for EEG-based BCIs, which achieved competitive accuracy while being computationally efficient.

Subsequent studies have sought to enhance CNN architectures by incorporating advanced techniques from computer vision. Santamaria-Vazquez et al. (2020) improved EEG decoding in an ERP speller by integrating inception modules (Szegedy et al., 2015) into its CNN *EEG-Inception*, allowing the network to capture features at multiple scales. Kostas and Rudzicz (2020) adapted *DenseNet* (Huang et al., 2017) to MI decoding from EEG, demonstrating that deeper networks with dense connections can capture more informative features. Kwon et al. (2020) enhanced feature representation by transforming EEG signals into spectro-spatial feature representations from the EEG before inputting them into a CNN. Fan et al.

(2021) addressed inter-subject variability in MI by developing an enhanced CNN architecture that integrates residual connections (He et al., 2016) and an attention mechanism (Vaswani et al., 2017).

Despite these advancements, challenges persist in fully leveraging DL for MI decoding. Many CNNs rely on an initial spatial convolution that limit the complexity of spatial relationships they are able to capture in EEG data (Dose et al., 2018; Kostas and Rudzicz, 2020; Lawhern et al., 2018; Santamaria-Vazquez et al., 2020; Schirrmeister et al., 2017). Additionally, few studies have explored training models across multiple datasets to enhance generalization and exploit TL capabilities of DL and the improvement that large quantities of training data have on DL models (Goodfellow et al., 2016).

Reducing the number of EEG electrodes is another critical aspect for practical BCI applications, as it simplifies setup and increases user comfort (Acqualagna et al., 2016). Some studies have examined the impact of electrode reduction on model performance (Dose et al., 2018; Fan et al., 2021), but more research is needed to develop models that maintain high accuracy with a reduced set of channels.

1.5.2 Explainable artificial intelligence in interpreting deep learning models in electroencephalography analysis

While DL models have significantly improved EEG decoding performance, their inherent *black box* nature raises concerns about interpretability and transparency (Adadi and Berrada, 2018). Understanding how these models make decisions is crucial, especially in medical applications where trust and accountability are key. XAI aims to make AI models more interpretable by providing insights into their decision-making processes (Samek et al., 2021). In the context of EEG analysis, XAI techniques could help identify which features, time regions, frequency bands, or brain regions contribute most to model predictions, facilitating physiological interpretations and guiding improvements in BCI paradigms.

A variety of XAI techniques have been applied to EEG data, revealing crucial information about neural activity during classification tasks. Sturm et al. (2016) adapted LRP (Binder et al., 2016) for EEG data, demonstrating that their model focused on specific brain regions and time intervals when classifying MI tasks. Their findings revealed that individual trials could be classified based on the relevance of the M1 and S1 regions between 1000 and 3000 ms after MI onset, even in the absence of real-time feedback. Perturbation-based approaches have also been explored for EEG analysis. Ieracitano et al. (2021) applied occlusion sensitivity

analysis (Zeiler and Fergus, 2014) to determine the importance of different EEG regions in their DL model's predictions. Similarly, Nahmias and Kontson (2020) introduced *easyPEASI*, a perturbation-based method that systematically alters input data to identify relevant frequency bands for pathology detection.

SHAP, as discussed in Subsection 1.4.2, provides a unified framework for XAI by satisfying key properties that ensure meaningful feature attributions. SHAP assigns each input feature an importance value based on cooperative game theory (Lundberg and Lee, 2017), ensuring local accuracy and consistency. Alsuradi et al. (2020) applied SHAP to interpret tree-based classifiers built on handcrafted EEG features. Its use for explaining DL models models trained directly on raw EEG, however, remains unexplored. More recently, Miao et al. (2023) extended the application of SHAP values to temporal EEG data (filtered between 8–30 Hz) during ME and MI in a binary classification framework. Their study demonstrated the feasibility of SHAP for providing temporal insights into EEG-based DL models.

Despite these advancements, several limitations remain that need further investigation. One key challenge is the restricted focus on binary classification paradigms, which may oversimplify the underlying neural dynamics. Incorporating a resting state class into DL architectures is a promising strategy to refine the interpretability of these models. This addition is expected to provide a more comprehensive understanding of the neural dynamics at play, as it allows for the differentiation of task-related activations from baseline brain activity. Additionally, generalizability across different DL architectures remains an open question.

1.5.3 Investigating motor imagery and execution from electroencephalography

The dual paradigms of MI and ME have long been recognized as key components in EEG-based BCIs. MI involves the mental simulation of movement without actual action, whereas ME pertains to engaging in physical movement. Despite sharing underlying neural mechanisms (most notably in the M1), the EEG signatures elicited by these two tasks differ in their stability and variability. ME-related patterns are generally more robust and consistent across subjects, making them attractive for developing models with enhanced TL capabilities, while MI patterns often suffer from variability and lower SNR, which contribute to the so-called BCI inefficiency (Alkoby et al., 2018; Wolpaw and Wolpaw, 2012).

Recent research has explored the potential of leveraging the stable neural

patterns of ME to improve the classification of MI. For example, Lee et al. (2022) investigated decoding various upper limb movements, both executed and imagined, by employing a DL network augmented with a variational autoencoder (VAE). Their work highlighted that reconstructing MI examples from ME samples could improve classification accuracy. Similarly, Miao et al. (2023) demonstrated that a DL model trained on ME data could classify MI tasks effectively when supplemented with a limited amount of MI-specific fine-tuning. Shuqfa et al. (2023) further advanced this line of inquiry by training classifiers on a combined dataset of ME and MI, reporting that such hybrid training protocols outperform models trained on either task independently. Nonetheless, the feasibility of applying a DL model trained solely on ME data to classify MI tasks without finetuning remains largely unexplored, posing a compelling research question regarding the inter-task transferability of motor-related neural patterns.

In contrast to fine-tuning approaches (Lee et al., 2022; Miao et al., 2023), a promising research direction is to explore direct TL from ME to MI without any subsequent adaptation to the MI task. This entails determining whether a DL model trained exclusively on ME data can reliably classify MI events. If successful, such an approach may offer several advantages:

- **Increased reliability**

  ME data collection allows for objective verification of task performance through observable movement. In contrast, MI events may suffer from user distraction or disengagement, making compliance harder to verify.

- **New rehabilitation avenues**

  Models trained with ME data can focus on the neural activity linked to functions a rehabilitation patient has lost. Their TL capacity toward MI supports every stage of rehabilitation. Early in therapy, if the person cannot move the affected limb, MI-based feedback with these models could improve their rehabilitation. As movement returns and ME signals appear, the same model seamlessly shifts to detecting the emerging ME patterns.

- **Preparatory runs**

  ME-based preparatory runs can serve as a reliable baseline for early identification of setup issues. Additionally, by observing and optimizing the strategy during actual movement execution, users can then replicate the most effective approach during the MI task.

Current investigations are also focusing on elucidating the spatio-temporal patterns underlying ME and MI tasks through XAI techniques (Miao et al., 2023). In summary, while ME offers a more stable basis for training DL models, exploring inter-task TL from ME to MI without additional fine-tuning holds significant promise. This approach could reduce calibration requirements and mitigate user fatigue, potentially paving the way for the increased deployment of EEG-based BCIs in real-world environments.

Following this introduction of the core topics in this Doctoral Thesis, the document is structured to provide a comprehensive overview of the research. Chapter 2 outlines the hypotheses and specific objectives that drive each study. Chapter 3 details the databases and methodology, including signal pre-processing, the development of our novel DL model *EEGSym*, performance assessment, and statistical analysis. It also describes the adaptation of a SHAP-based XAI method for EEG and the transfer learning approach from ME to MI. Chapter 4 presents the main experimental results, which are subsequently discussed in Chapter 5, where current limitations are also examined. Chapter 6 summarizes the contributions and key findings and outlines future research directions. Finally, Appendix A includes a compendium of publications. To complement this work, Appendix B details the scientific achievements obtained during the Ph.D. (covering papers indexed in JCR, international and national conferences, and international internship as well as awards), and Appendix C provides a brief summary in Spanish.

Chapter 2

Hypothesis and objectives

Understanding and accurately decoding MI from EEG signals remains a significant challenge in the development of BCI systems. Despite promising paradigms that elicit control signals for various applications, there is considerable room for improvement in the accuracy of these systems. This is partly due to inherent drawbacks of EEG recordings, such as low spatial resolution, poor SNR, and inter-subject variability. Additionally, the neural mechanisms underlying MI that are being used for classification are not fully understood, limiting the effectiveness of current approaches.

To address these challenges, this Doctoral Thesis focuses on designing, developing, and evaluating novel DL methodologies for MI decoding. By implementing advanced DL techniques and applying XAI methods, we aim to enhance EEG classification systems used in BCI applications. Furthermore, we investigate the relationship between MI and ME to explore the potential of TL between these tasks, which could simplify the calibration process and improve user experience in MI-based BCIs.

The hypotheses that have motivated each study, as well as the overarching hypothesis that justifies this Doctoral Thesis, are declared in Section 2.1. The main objective and the specific objectives established to achieve it are introduced in Section 2.2.

2.1 Hypotheses

Despite advancements in BCI research, current MI-based BCIs face several limitations that hinder their widespread application outside laboratory settings.

These limitations include:

1. **Low classification accuracy** due to inter-subject variability and the complex nature of EEG signals.
2. **Lengthy calibration sessions** required for training models specific to each user, which can reduce user motivation and practicality.
3. **Limited understanding of neural mechanisms** underlying MI in EEG signals, restricting the ability to optimize BCI systems effectively.

To address these issues, the following specific hypotheses have been formulated:

H1 **Advanced DL architectures**, such as CNNs incorporating brain symmetry, inception modules and residual connections, could significantly improve MI decoding accuracy by effectively handling inter-subject variability and extracting meaningful features from raw EEG data.

H2 **Explainable artificial intelligence techniques** applied to DL models could uncover important neural patterns and brain regions involved in MI, providing physiological interpretations that deepen our understanding of the underlying neural mechanisms.

H3 **Transfer learning between ME and MI** is feasible using DL models, enabling models trained on ME data to directly perform effectively on MI tasks, without fine-tuning to the MI task, thus reducing or eliminating the need for extensive MI-specific calibration sessions.

These specific hypotheses support the overarching hypothesis of this Doctoral Thesis:

"The latest advances in DL, combined with XAI methods and leveraging the relationship between ME and MI, can improve current EEG-based BCI systems by enhancing MI decoding accuracy, deepening our understanding of brain function, and overcoming the limitations associated with MI classification."

2.2 Objectives

The general goal of this Doctoral Thesis is to design, develop, and evaluate a set of novel deep learning-based methodologies that enhance MI decoding in EEG-based BCI systems and provide physiological interpretations of the neural processes

involved. To achieve the main objective, the following specific objectives have been established:

O1: Design and implement a novel DL architecture *EEGSym* for MI decoding that incorporates brain symmetry, inception modules, and residual connections to improve classification accuracy and reduce BCI inefficiency. It will be supplemented by the use of data augmentation techniques and the use of transfer learning across different datasets. This will enhance the generalization capabilities of the DL model and address inter-subject variability.

O2: Design, develop and apply XAI methods, specifically an adaptation of SHAP, to *EEGSym*. This will reveal the neural patterns and brain areas that most influence MI classification. The insights will then serve to optimize EEG electrode placement, reducing the number of required electrodes without compromising classification accuracy.

O3: Contribute to the understanding of the neural mechanisms of MI and ME by analyzing similarities and differences in neural activation patterns, evaluating the transferability of DL models trained on ME data to MI tasks, and assessing the viability of inter-task transfer learning. The success in this objective can potentially result in the design of more effective rehabilitation and assistive technologies.

O4: Evaluate the developed methodologies on large publicly available EEG datasets involving a diverse population of subjects to demonstrate their effectiveness, generalizability, and potential impact on reducing BCI inefficiency.

By fulfilling these objectives, this Doctoral Thesis aims to advance the state of the art in EEG-based MI decoding, enhance the practicality of BCI systems, and deepen our understanding of the neural processes involved in motor imagery and execution.

Chapter 3

Materials and methods

This chapter presents the methodology adopted throughout this Doctoral Thesis to investigate DL enhancements for MI classification, DL intepretability, and the relationship between MI and ME. Section 3.1 introduces the publicly available EEG datasets used, describing their main characteristics and acquisition protocols. Section 3.2 outlines the pre-processing pipeline applied to the raw EEG data, including electrode selection, filtering, and normalization steps.

The DL model development is detailed in Section 3.3. We first compare existing CNN architectures for EEG decoding (Section 3.3.1), then introduce the proposed network *EEGSym* (Section 3.3.2), and discuss DA strategies (Section 3.3.3). Next, we describe the training procedure (Section 3.3.4) and evaluation metrics (Section 3.3.5) used in our experiments.

Section 3.4 focuses on the XAI approach followed to elucidate how the DL model arrives at its predictions. We detail the usage of SHAP values to interpret network decisions (Section 3.4.1), and outline a channel selection protocol based on these feature attributions (Section 3.4.2).

Finally, Section 3.5 examines the relationship between MI and ME through comparative analysis (Section 3.5.1) and the specific comparison metrics (Section 3.5.2).

3.1 Electroencephalography datasets

This section provides an overview of the five public EEG datasets used throughout this Doctoral Thesis and their main characteristcis are summarized in Table 3.1: Physionet (Goldberger et al., 2000), OpenBMI (Lee et al., 2019), Kaya2018 (Kaya

Table 3.1: Details of the datasets

Dataset	Subjects	#	MI duration (s)	Sessions	Trials/ session	Sampling Frequency (Hz)
Physionet (Goldberger et al., 2000)	109	64	3*	1	45	128/160
OpenBMI (Lee et al., 2019)	54	62	4	4	100	1000
Kaya2018 (Kaya et al., 2018)	13	38	1	5	900	200
Meng2019 (Meng and He, 2019)	42	64	6	3	250	1000/1024
Stieger2021 (Stieger et al., 2021)	62	62	2-8	7-11	450	1000

Subjects: number of healthy subjects. #: number of electrodes. MI: motor imagery.
*: Physionet has also motor execution trials.

et al., 2018), Meng2019 (Meng and He, 2019), and Stieger2021 (Stieger et al., 2021). These datasets were selected for their subject count, trial count, and consistency of the MI paradigm. Notably, all participants in these datasets are healthy. As can be seen in Table 3.1, the datasets differ markedly in scale and recording protocol: the number of healthy participants ranges from 13 in Kaya2018 to 109 in the PhysioNet dataset, while channel density varies between 38 and 64 electrodes. MI segments last from 1 s (Kaya2018) to 6 s (Meng2019), and each subject contributed between one and eleven recording sessions, with per-session trial counts spanning 45 to 900. Sampling frequencies also cover a broad spectrum (from 128 Hz up to 1,024 Hz) providing a diverse test bed for assessing both temporal and spatial robustness of the DL models. Of note, Physionet dataset also include ME trials with the same number and duration as the MI events.

3.1.1 Experimental protocol

All datasets employ a similar experimental protocol (see Figure 3.1). The typical trial sequence is:

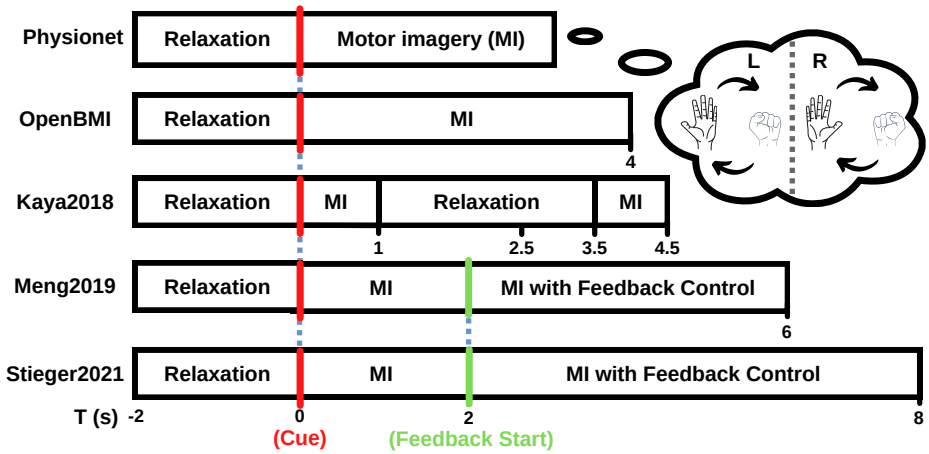


Figure 3.1: Schematic of MI protocols in the public datasets.

1. **Initial Resting Period (1–4 s)**

   A visual fixation cue (typically a cross or simple symbol) is presented to center the subject's gaze.

2. **MI Period**

   A visual cue instructs subjects to imagine performing a left or right hand movement. The specific durations of the MI period for each dataset are detailed in Table 3.1. In both Meng2019 and Stieger2021 datasets, a feedback period begins 2 seconds after the cue (Meng and He, 2019; Stieger et al., 2021). a visual feedback phase begins two seconds after the cue. During this interval, participants watch a moving marker on the same screen that displayed the instructions. The marker's position updates in real time to indicate whether the system has classified their imagery as left- or right-hand, enabling users to refine their mental strategy online.

3. **Final Resting Period (2–6 s)**

   Subjects relax until the next trial begins.

3.1.2 Task description

This subsection details the task followed to record the EEG activity across the datasets:

1. **Physionet** (Goldberger et al., 2000)

   Subjects imagine (MI) or perform (ME) continuously opening and closing their left or right hand over a 3 s trial. Notably, unlike the other datasets, this dataset also includes ME data collected using the same protocol followed for MI data.

2. **OpenBMI** (Lee et al., 2019)

   Subjects imagine left versus right hand movements over a 4 s trial; real-time feedback is provided on 2 of the 4 sessions.

3. **Kaya2018** (Kaya et al., 2018)

   Subjects are cued to perform a single instance of left or right hand MI lasting 1 s, followed by a relaxation period ranging from 2.5 to 3.5 s.

4. **Meng2019** (Meng and He, 2019)

   Subjects imagine left versus right hand movements over a 6 s trial within a cue-based protocol. The feedback start's 2 seconds after cue presentation. Subjects where instructed to move a cursor to either the left or right side of the screen. When participants imagined a left-hand movement, the onscreen cursor drifted toward the left edge, while imagining a right-hand movement made it shift toward the right.

5. **Stieger2021** (Stieger et al., 2021)

   Subjects imagine left or right hand movements with trial durations varying from 2 s to 8 s, depending on feedback-driven target achievement. The feedback start's 2 seconds after cue presentation. During the feedback phase, participants had up to six seconds to steer the cursor toward the designated target. Cursor motion was driven by variations in alpha-band power detected over both motor cortices.

3.1.3 Dataset usage across thesis studies

Throughout this Doctoral Thesis, the five publicly available MI EEG datasets described in Section 3.1 are utilized under varying research objectives and extracting diverse electrode configurations. The following details are consolidated in Table 3.2, which provides a quick reference to the datasets and electrode configurations employed in each study:

Table 3.2: Summary of Dataset Usage and Electrode Configurations Across Thesis Studies

Study	Datasets Used	Electrode Configurations
Perez-Velasco et al. (2022)	Physionet (MI) OpenBMI Kaya2018 Meng2019 Stieger2021	*8 channels*: [F3, C3, P3, Cz, Pz, F4, C4, P4] *16 channels*: +[F7, T7, P7, O1, F8, T8, P8, O2]
Pérez-Velasco et al. (2024)	Physionet (MI) Stieger2021	*16 channels* as in Perez-Velasco et al. (2022)
Pérez-Velasco et al. (2025)	Physionet (MI, ME, and rest)	All 64 channels of the dataset

- Perez-Velasco et al. (2022): This study introduced the *EEGSym* DL architecture for MI classification with an emphasis on mitigating inter-subject variability and reducing the proportion of users failing to achieve BCI control.
  - **Datasets:** All five MI datasets spanning 280 subjects, the largest population tested to date in similar research.
  - **Configurations:**
    - *8-channel configuration:* F3, C3, P3, Cz, Pz, F4, C4, P4. These channels emphasize bilateral sensorimotor and midline sites, aligning with *EEGSym*'s design, which leverages the brain's bilateral symmetry.
    - *16-channel configuration:* The 8-channel set plus F7, T7, P7, O1, F8, T8, P8, O2. Adding more frontal, temporal, parietal, and occipital sensors broadens coverage of cortical regions, thereby increasing classification accuracy and improving user coverage in BCI control.
- Pérez-Velasco et al. (2024): The central goal here was to integrate XAI methods, specifically SHAP with *EEGSym* to understand the full cortical activation patterns behind MI.
  - **Datasets:** Physionet (MI) and Stieger2021.
  - **Configuration:** The 16-channel montage introduced in Perez-Velasco et al. (2022).

It was intended to study how several cortical regions beyond the sensorimotor cortex are involved during MI, making the extended configuration essential for comprehensive analysis.

- Pérez-Velasco et al. (2025): This work examined inter-task TL between ME and MI to potentially eliminate the need for lengthy MI calibration sessions. By training *EEGSym* on ME data and subsequently evaluating MI classification performance, the study investigated whether ME-trained models transfer effectively to MI tasks.
  - **Dataset:** Physionet exclusively, including MI, ME, and resting-state data.
  - **Configuration:** All 64 available channels.

    Leveraging both ME and MI recordings, and a complete electrode array, maximized the model's ability to capture shared and distinct activation patterns. This choice was essential for analyzing various TL scenarios (ME→MI, ME→ME, MI→MI) and conducting correlation analyses of performance across tasks.

3.2 Signal preprocessing

The raw EEG data from each dataset underwent a standardized preprocessing pipeline, consistently applied across the three studies in this compendium. Preprocessing is a vital step in EEG signal analysis because raw data often contain substantial noise and artifacts (electrical interference, muscle activity, and eye movements) that can compromise subsequent analyses. By reducing or eliminating these confounding factors, the reliability, interpretability, and SNR of the EEG signals are greatly enhanced. The preprocessing workflow is outlined as follows:

3.2.1 Electrode Selection

Although advanced network architectures and data augmentation strategies can handle variable input dimensions (Guetschel et al., 2024), ensuring consistent input shape across datasets simplifies model design and training. Additionally, in our first study, ensuring training compatibility with other models made a configuration-constrained architecture the most reasonable choice.

In Perez-Velasco et al. (2022), to achieve a standard input for all DL experiments, a common subset of electrodes from the five datasets was extracted to

form two configurations. The **8-channel configuration** includes F3, C3, P3, Cz, Pz, F4, C4, and P4. The **16-channel configuration** additionally incorporates F7, T7, P7, O1, F8, T8, P8, and O2. These electrode configurations are common to all datasets and are typically available to setup in commercial EEG caps. The choice of these smaller montages helped mitigate practical constraints (e.g., real-time BCI implementation), while still offering wide coverage of the entire scalp, providing a global overview of brain activity.

In Pérez-Velasco et al. (2024), the 16-channel configuration was specifically chosen to capture the broader cortical areas implicated in MI. This setup not only builds on the analyses performed (Perez-Velasco et al., 2022), but also enables a subsequent channel selection approach informed by XAI analysis.

Conversely, Pérez-Velasco et al. (2025) used a single dataset (Physionet, Goldberger et al. (2000)) and opted for the complete 64-channel setup. Since transfer learning (ME $\rightarrow$ MI) and spatial correlation analyses were of primary interest, maintaining all available electrodes allowed for a more fine-grained exploration of brain patterns during both motor execution and imagination tasks

3.2.2 Frequency Filtering

Although a wide range of EEG frequencies (approximately 0.1–100 Hz) can carry meaningful neural information (Section 1.3.2), many practical considerations dictate a narrower bandwidth. In particular, the electrical grid at 50 or 60 Hz must also be removed to reduce power-line interference.

In this work, **a fourth-order Butterworth infinite impulse response (IIR) notch filter** was applied to remove the 50/60 Hz power-line artifact in datasets where such filtering was not already handled in hardware (Goldberger et al., 2000; Lee et al., 2019).

Subsequently, and prior to resampling, a **fourth-order Butterworth IIR low-pass filter** was applied as an anti-aliasing filter. The cutoff frequency for this filter was set precisely at 63 Hz. This choice is critical to satisfy the Nyquist theorem for the target sampling rate of 128 Hz, as it ensures that all frequencies above the Nyquist frequency (64 Hz) are adequately attenuated before the resampling step. This prevents high-frequency noise from being aliased into the frequency band of interest, thereby preserving the integrity of the crucial spectral information for MI classification while removing components with a typically low SNR.

We opted for an IIR filter on both operations due to its computational efficiency

and straightforward implementation, acknowledging that while finite impulse response (FIR) filters have a linear phase response and inherent numerical stability, the IIR filter's reduced computation time is critical for online BCI applications, ensuring that our offline analyses closely mirror real-time conditions.

3.2.3 Spatial Filtering

Spatial filters are crucial for enhancing the signal quality on individual channels by effectively mitigating artifacts across the electrode array. In this work, each selected electrode was re-referenced using the common average reference (CAR) approach (Ludwig et al., 2009). Specifically, for a given channel c, the CAR filtered signal x_c^* is obtained as:

$$x_c^* = x_c - \frac{1}{N_c} \sum_{i=1}^{N_c} x_i \tag{3.1}$$

where x_c is the original signal recorded at channel c, N_c is the total number of channels in the subset of electrodes selected, and x_i represents the signal recorded at channel i. This operation subtracts the overall mean signal of all channels from each channel individually, thereby reducing artifacts common to all electrodes. By implementing CAR, the signal of interest in each channel is emphasized relative to ubiquitous noise, thus improving the quality and interpretability of the spatially filtered EEG data (Wolpaw and Wolpaw, 2012).

While CAR may introduce a slight bias, the approximation error diminishes for higher-density electrode configurations, making it well suited for capturing global brain activity. Moreover, the method remains effective even with lower-density montages, as long as the electrode array evenly covers the entire scalp. However, if the electrode array is disproportionately dense in a specific brain region, the common activity from that region may be inadvertently reflected on distant channels and subtracted from its original location.

However, this potential limitation was mitigated in our study by two key factors. First, the chosen electrode configurations provided a **wide and symmetrical distribution across the scalp** (including frontal, central, and parietal sites), rather than being clustered in a single area, which helps ensure the calculated average is more representative of a global potential. Second, and more importantly, a key strength of an end-to-end deep learning approach is its ability to **learn the specific spatial patterns** resulting from the preprocessing pipeline. The network is not designed under the assumption of a perfect reference; instead, it learns to classify based on the consistent spatial signatures produced

after CAR has been applied, adapting to any systematic effects introduced by the re-referencing step. Therefore, CAR remains a practical and effective method for standardizing the data and reducing common-mode noise within the context of this DL framework.

3.2.4 Resampling

All datasets were resampled to 128 Hz to standardize the sampling rate across datasets.

3.2.5 Trial Extraction

Trials were extracted by segmenting a 3-second time window immediately following the onset of each event. This 3-second window (yielding 384 samples at 128 Hz) is the largest common interval available across datasets, minimizing the need to discard or pad trials.

3.2.6 Normalization

Each trial underwent channel-wise z-score standardization, ensuring that every channel's signal has zero mean and unit variance. This step removes the continuous component and conditions the data for the subsequent classification with DL.

After these steps, the preprocessed data is organized as a 4D tensor with dimensions [*trials* $\times$ *samples* $\times$ *electrodes* $\times$ 1], where *samples* equals 384 (3 s $\times$ 128 Hz), and the final dimension is required by the DL network. Figure 3.2 presents a schematic representation of the trial structure and its correspondence to the datasets, underscoring the heterogeneity of the EEG segments used for decoding. In PhysioNet, the analysis window coincides with the full duration of the imagined (or, in this dataset also executed) movement. In OpenBMI only a portion of the 4 s imagination period is retained. Kaya2018 always includes a 2 s relaxation interval preceding imagination. Whereas Meng2019 and Stieger2021 the final second captures MI performed while the participant receives real-time visual feedback on task performance.

Additionally, to incorporate a neutral class that increases interpretability in the last work of this Doctoral Thesis (Pérez-Velasco et al., 2025), the same preprocessing pipeline was applied to the baseline resting recordings of Physionet (Goldberger et al., 2000). For this neutral resting class, consecutive 3-second segments were extracted from the 1-minute eyes-open baseline recordings, yielding

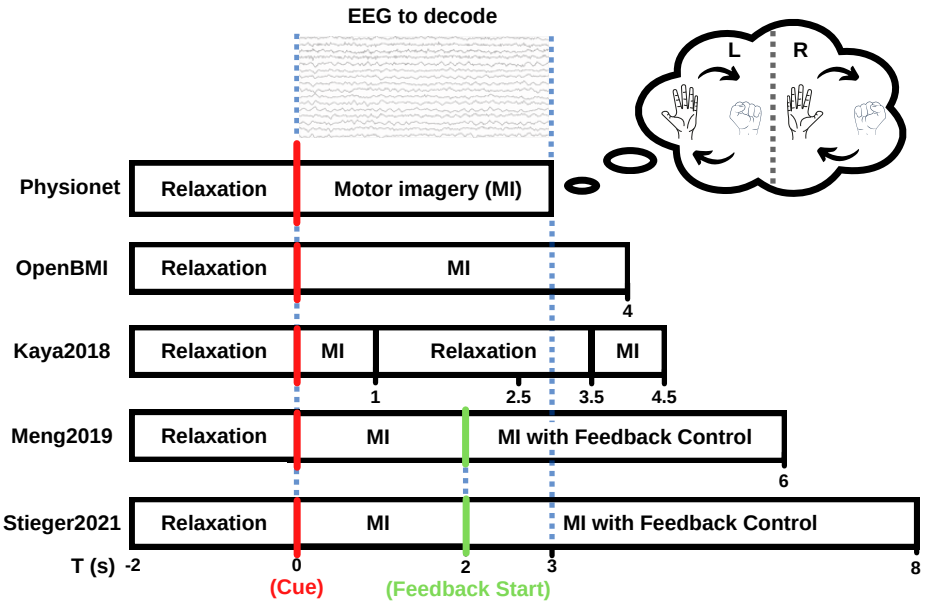


Figure 3.2: Schematic of MI protocol in the public datasets, illustrating the feature window extracted from the preprocessing stage and its correspondence to each dataset protocol.

20 trials per subject as done previously in Dose et al. (2018). It is important to note that this constitutes a passive resting state, where participants were instructed to relax with their eyes open but were not given any specific cognitive task to perform. This approach ensures a neutral baseline against which motor-related activity can be clearly distinguished by the model. This approach ensures that the resting trials are formatted consistently with the MI and ME trials.

3.3 Deep learning model

This section outlines the DL framework used for MI classification in this Doctoral Thesis. Building upon prior methodological advancements of DL in other fields, we introduce *EEGSym* (Perez-Velasco et al., 2022), a novel architecture designed to leverage the spatial and spectral characteristics of EEG signals. To contextualize its development, we compare multiple baseline models, detail the architectural design choices, and discuss the data augmentation strategies employed to enhance generalization. Additionally, we outline the training procedure and define the evaluation metrics used to assess performance. The proposed framework offers

a robust solution for MI classification in EEG-based BCIs by addressing key challenges, including removing the need for cumbersome calibration procedures for each session.

3.3.1 Baseline models

To benchmark the performance of *EEGSym*, we consider four widely cited CNN-based architectures for EEG decoding: ShallowConvNet and DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), and EEG-Inception (Santamaria-Vazquez et al., 2020). These models provide a comprehensive overview of existing DL strategies applied in EEG-based BCIs.

1. **ShallowConvNet and DeepConvNet** (Schirrmeister et al., 2017)

   These architectures were among the first CNN designs tailored for EEG decoding directly from raw signals, bypassing the need for handcrafted feature extraction. ShallowConvNet is crafted to learn frequency-specific band power features by first applying a temporal convolution with a relatively large kernel, followed by a spatial filtering step reminiscent of CSP, then using a squaring nonlinearity, average pooling, and a logarithmic activation to mimic traditional log-variance computations. In contrast, DeepConvNet utilizes a deeper, more generic structure that stacks several convolutional and pooling layers with ELU activations, thereby capturing more complex, hierarchical representations of EEG data. Both networks achieve competitive performance in decoding task-related information and support interpretability through visualization techniques that reveal the spatial and spectral features learned by the models. We adopt the TensorFlow implementations in Lawhern et al. (2018), which follow the hyperparameters and layer configurations specified in the original publication.

2. **EEGNet** (Lawhern et al., 2018)

   EEGNet is a lightweight CNN architecture specifically engineered for EEG-based BCI applications, capable of generalizing across diverse paradigms even when training data is scarce. By leveraging depthwise and separable convolutions, EEGNet dramatically reduces the number of trainable parameters while maintaining strong performance, enabling efficient training with limited data. Our implementation follows the official TensorFlow code

provided by the authors, ensuring fidelity to the design and hyperparameter choices outlined in the original publication (Lawhern et al., 2018).

3. **EEG-Inception** (Santamaria-Vazquez et al., 2020)

   EEG-Inception integrates inception modules (Szegedy et al., 2015) into an EEG-specific architecture, enabling multiple convolutional kernels of varying sizes to operate in parallel within the same layer. This multi-scale design effectively captures diverse time scales in EEG signals, thereby emphasizing different frequency components in a manner similar to FBCSP. This inception-based architecture demonstrated robust performance in ERP detection, outperforming both traditional machine learning methods and the other DL baseline architectures presented earlier. Our implementation follows the publicly available code provided by the authors.

3.3.2 Deep learning architecture design

The first core contribution of this Doctoral Thesis is *EEGSym* introduced in Perez-Velasco et al. (2022), a CNN-based architecture designed to inherently exploit the bilateral symmetry of the human brain. An overview of *EEGSym* is depicted in Figure 3.3, and detailed implementations for both 8- and 16-channel configurations could be found in the open-soruce implementation at https://github.com/Serpeve/EEGSym.

The various sections of this architecture, as illustrated in Figure 3.3.e, include:

1. **Symmetric Division**

   EEGSym begins by splitting the input data electrode dimension into left and right hemisphere groupings. This division aims to learn common spatial features across brain hemispheres in the early layers, leveraging the brain's intrinsic symmetry.

2. **Tempospatial Analysis**

   EEGSym builds upon prior successful methods in EEG decodification. In the early stages of this section, the network employs inception modules (Szegedy et al., 2015), as demonstrated in EEG-Inception (Santamaria-Vazquez et al., 2020), to capture multi-scale temporal information with convolutional kernels operating over different window sizes. Two inception blocks (Szegedy et al., 2015) capture multi-scale temporal information (e.g., 125 ms, 250 ms, 500 ms windows). An average pooling layer then reduces

the temporal dimension (i.e., S), which helps to prevent overfitting and reduce computational cost. Furthermore, grouped convolutions (Xie et al., 2017) are used throughout the architecture to emulate the benefits observed with depthwise convolutions in EEGNet (Lawhern et al., 2018) and EEG-Inception (Santamaria-Vazquez et al., 2020) (depthwise convolutions being a special case of grouped convolutions when the number of groups equals the number of filters). This grouped convolution operates across all channels of each hemisphere (i.e., C), reducing the channel dimension to 1, and the output is then added back to every channel via a residual connection. These inception blocks are subsequently followed by residual blocks (He et al., 2016) that further refine both temporal and spatial features. Importantly, every convolution operation in *EEGSym* is consistently followed by batch normalization, ELU activation, and dropout regularization to enhance training stability and mitigate overfitting.

3. **Channel Merging**

   After extracting low-level tempospatial features, the channel merging stage consolidates spatial information from both hemispheres into a unified representation. This stage employs convolutional layers with residual connections and grouped convolutions (kernel size $2\times1\times5$) to combine the data, preserving inter-hemispheric relationships while reducing the channel dimensionality.

4. **Temporal Merging**

   This stage progressively reduces the temporal dimension (e.g., from 384 samples to a single time point) through convolutional and pooling layers, capturing global temporal dependencies relevant to MI classification.

5. **Output Module**

   The final stage transforms the merged feature set into a compact global feature vector. It consists of a series of convolutional layers with residual connections, followed by a flattening operation and a softmax classifier that outputs the final two-class MI prediction (left- vs. right-hand imagery). Note that this module can be easily adapted to produce multiple outputs for multi-class classification tasks by adjusting the number of neurons in the final dense layer accordingly.

Notably, *EEGSym* integrates two design principles not jointly explored in earlier networks applied to EEG:

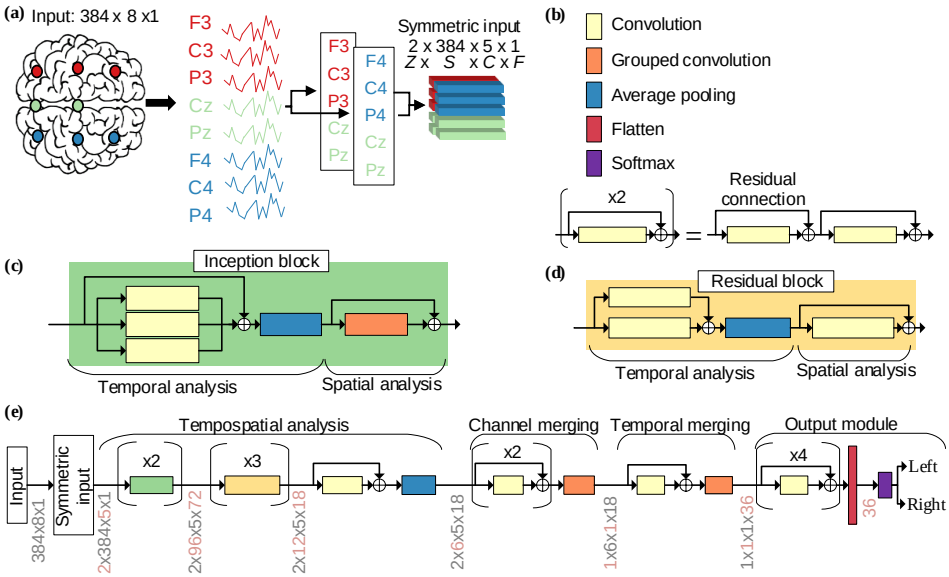


Figure 3.3: Overview of the *EEGSym* architecture. (a) Illustration of the input electrode symmetric division for the 8-electrode configuration, where Z represents hemispheres (i.e., 2), S denotes the number of samples (i.e., 384 samples), C corresponds to electrodes per hemisphere (i.e., 5), and F indicates the number of filters. (b) Legend explaining the architectural components. (c) Inception block. (d) Residual block. (e) Complete *EEGSym* architecture. Each convolutional and grouped convolutional layer is followed sequentially by batch normalization, ELU activation, and dropout regularization. Output dimensions for each operation are marked in gray, while the affected dimension at each stage is highlighted in red. Comprehensive tables detailing the 8- and 16-electrode configurations, including kernel sizes, filter counts, and other specifics, are available in the open-source implementation at https://github.com/Serpeve/EEGSym. Figure from Perez-Velasco et al. (2022).

- *Residual connections*: These connections facilitate deeper, more adaptable feature learning by introducing shortcuts that allow information to bypass certain layers. This mechanism helps counteract vanishing gradients and enables some layers to be effectively skipped, ensuring that critical spatial features and correlations across electrodes are efficiently propagated throughout the network (He et al., 2016).

- *Implicit hemispheric symmetry*: This design explicitly leverages the brain's natural symmetry by processing inputs on a hemisphere-wise basis. Drawing inspiration from gaze recognition research (Valliappan et al., 2020), the architecture first extracts common spatial characteristics from each hemisphere during the tempo-spatial analysis stage, and then models

the complex inter-hemispheric relationships during channel merging. This approach reflects the lateralized nature of sensorimotor processing.

A subsequent ablation study (Section 4.1.3) quantifies how each of these architectural innovations contributes to the improved accuracy observed on the Physionet dataset.

3.3.3 Data augmentation

DA is employed to mitigate overfitting and enhance model robustness (Roy et al., 2019). Inspired by techniques in computer vision and existing EEG literature, in Perez-Velasco et al. (2022) we apply a *uniform random selection* of one of four augmentation strategies on a trial-by-trial basis. These four strategies are:

1. **Patch Perturbation**

   Drawing inspiration from random erasing (Zhong et al., 2020), this augmentation randomly selects a contiguous time window—ranging from 0.6,s to the full 3,s trial—and a subset of channels to perturb. The selected patch is either replaced with zeros or infused with Gaussian noise (mean 0 and a standard deviation uniformly drawn between 0.01 and twice that of the original signal). Importantly, at least one channel remains unmodified, ensuring that some genuine information from that time window is preserved. This strategy encourages the model to extract meaningful features from various temporal segments and prevents over-reliance on specific time-frequency patterns.

2. **Hemisphere Perturbation**

   Recognizing that left-/right-hand motor imagery is typically reflected in the contralateral motor cortex (Wolpaw and Wolpaw, 2012), this augmentation targets one hemisphere at random. The process involves either randomly reordering the electrode data within that hemisphere or replacing it entirely with Gaussian noise (mean 0, standard deviation 1). This approach compels the model to independently decode MI-related patterns from each hemisphere, thereby increasing robustness against spatial inconsistencies or partial channel dropout.

3. **Random Shift**

   Although the precise onset cue for MI is known, individual reaction times can vary considerably across trials and subjects (Psotta, 2014). To simulate

this variability, including slower responses from distracted or fatigued participants, we shift the trial onset by a random amount, up to 0.5 s. The exact shift is drawn uniformly from a range corresponding to 1 to 64 samples (given a sampling frequency of 128 Hz). This approach helps the model generalize across a range of delayed MI onsets, accommodating fatigued or distracted participants.

4. **No Augmentation**

   To ensure that the model also learns from unaltered data distributions, a proportion of the trials are left unmodified. This control condition balances the augmented data and aids in preserving the integrity of the original signal features.

By combining these four strategies in a random manner, each training epoch presents the model with a new, stochastically altered dataset, thus improving generalization and mitigating overfitting.

3.3.4 Training procedure

The training pipeline followed in Perez-Velasco et al. (2022) and Pérez-Velasco et al. (2024) is illustrated in Figure 3.4, and it consists of three main steps:

1. **Pre-training on Multiple Datasets**

   All datasets except the target dataset (i.e., the one used for final evaluation) are merged into a single "pre-training" set. DA is applied at this stage, and training employs a relatively high learning rate (e.g., 10^{-2}). Ten trials of each class per subject are withheld for validation, while the remaining trials form the training set. Early stopping (patience = 25) halts training when validation loss fails to improve.

2. **Fine-tuning on the Target Dataset**

   Once pre-training converges, the final model weights serve as initialization for the target dataset. Trials from all target subjects except one (in a leave-one-subject-out, LOSO, manner) are used to fine-tune the model. Initially, only the final softmax layer is unfrozen and trained with a very low learning rate (e.g., 10^{-4}), to preserve the pre-trained feature extraction. Next, all layers are unfrozen and trained with the same low learning rate, allowing the model to adapt deeper layers if the target dataset differs substantially from pre-training sets. Early stopping is again employed.

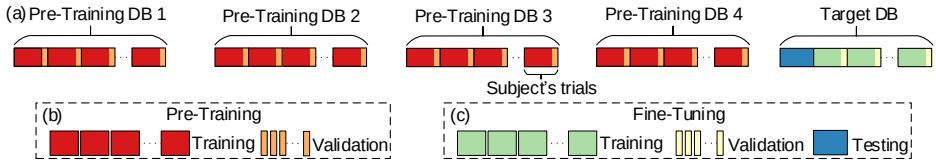


Figure 3.4: Visualization of the cross-validation procedure applied to each dataset when used as the Target DB. The **Pre-Training DB** consists of datasets utilized for pre-training the model. The **Target DB** refers to the dataset where Leave-One-Subject-Out (LOSO) cross-validation is performed, ensuring that in each iteration, one subject is used for testing while the remaining subjects are employed for fine-tuning. (a) Overview of the cross-validation framework. (b) Representation of the datasets used for pre-training. (c) Illustration of the fine-tuning process and the selected test subject. Figure from Perez-Velasco et al. (2022).

3. **Testing on the Held-Out Subject**

   The model is then evaluated on the remaining held-out subject. Accuracy from this LOSO test is recorded, and the process repeats until every subject in the target dataset has served once as the test subject.

This approach balances the benefits of pre-trained features with dataset-specific adaptation, improving the final classification accuracy and being the best-performing solution possible. However, in a real-world deployment, where the system initially lacks data from other subjects recorded under the same laboratory conditions, a slight decrease in accuracy can be expected.

For all baseline models, the hyperparameters were set to the preferred values reported by the original authors. In contrast, for *EEGSym*, the dropout rate (dr), number of filters in inception modules (N), and learning rate (lr) were determined through grid search on the validation set. In grid search, a predefined finite set of candidate values is chosen for each hyperparameter to conservatively cover a plausible range based on prior experience. The algorithm then trains a model for every combination of these candidate sets, and the configuration yielding the lowest validation error is selected. In our experiments, the search spaces were defined as: $dr = [0.2, 0.3, 0.4, 0.5]$, $N = [8, 16, 24, 32]$, and $lr = [0.01, 0.001, 0.0001]$. The optimal values found were 0.4 for the dropout rate, 24 for the number of filters, and 0.001 for the learning rate.

In Pérez-Velasco et al. (2025), we omitted the pre-training phase due to variability in channel sets across datasets and instead relied solely on data from Physionet (Goldberger et al., 2000). Consequently, the training schedule was simplified to a single training step (similar to the fine-tuning step described

above) performed without data augmentation, with all layers unfrozen, and using a learning rate of 10^{-2}. This was then followed by testing on a held-out subject.

The specifications of the hardware and software used during the training are a NVIDIA 3080Ti GPU, supported by CUDA 11.2, and cuDNN 8.1.0 within the TensorFlow 2.10.1 framework.

3.3.5 Evaluation metrics

Throughout our experiments, the primary *training objective* for all models is to minimize the **categorical cross-entropy (CCE) loss** (Mao et al., 2023), which quantifies the divergence between the true class distribution (represented as one-hot vectors) and the model's predicted probability distribution. Formally, if y_i denotes the true class label of the i-th sample (one-hot encoded) and $\hat{y}_i$ is the predicted probability distribution, the CCE loss over N training samples is defined as:

$$\mathcal{L}_{\text{CCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\,\log\big(\hat{y}_{i,c}\big), \tag{3.2}$$

where C is the total number of classes (in Perez-Velasco et al. (2022) and Pérez-Velasco et al. (2024), $C = 2$; in Pérez-Velasco et al. (2025), $C = 3$).

To evaluate classification performance, we primarily report **accuracy** (%), calculated as:

$$\text{Accuracy} = \frac{\text{Number of correctly classified trials}}{\text{Total number of trials}} \times 100. \tag{3.3}$$

Since our MI experiments typically involve two balanced classes (left vs. right hand), accuracy serves as a straightforward performance metric. However, in Pérez-Velasco et al. (2025) we also report the confusion matrix. A confusion matrix is a table that compares the true labels with the predicted labels, summarizing the performance of a classification model by displaying the counts of true positives, true negatives, false positives, and false negatives. This detailed breakdown provides insight into the types of errors made by the model and is particularly useful for understanding misclassification patterns.

To compare the performance between *EEGSym* and the baseline models, we apply statistical tests across both electrode configurations and all datasets. Specifically, we use the Wilcoxon signed-rank test (Wilcoxon, 1945) for paired comparisons, and adjust the resulting p-values using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR).

This ensures that any observed differences in accuracy and loss are statistically significant.

3.4 Explainable artificial intelligence techniques

In this section, we present the explainable artificial intelligence techniques employed to interpret the decision-making process of the DL model used for MI classification presented in Pérez-Velasco et al. (2024). Specifically, we leverage SHAP values, a game-theoretic approach for feature attribution, to analyze the contributions of different electroencephalography signal components to model predictions. This methodology allows for a deeper understanding of how electroencephalography patterns influence classification decisions. Furthermore, we introduce a channel selection strategy based on SHAP values to identify the most informative electrodes, optimizing the model for practical brain-computer interface applications.

3.4.1 SHAP values for motor imagery

Numerous approaches have been proposed to enhance the interpretability of deep neural networks using additive feature attribution methods (Ribeiro et al., 2016; Shrikumar et al., 2017). To unify these techniques under a common theoretical framework, SHAP values were introduced (Lundberg and Lee, 2017). SHAP values stem from cooperative game theory, where different "players" contribute collectively to an outcome. When applied to electroencephalography data, these "players" correspond to different signal components, and their SHAP values indicate the extent to which each component influences the model's prediction.

SHAP values satisfy three fundamental properties that ensure their reliability in explaining model decisions: local accuracy, missingness, and consistency (Lundberg and Lee, 2017). Local accuracy ensures that the sum of SHAP values equals the difference between a given prediction and the expected output. The missingness property guarantees that features not present in an input receive a SHAP value of zero, ensuring that only active features contribute to model explanations. Consistency ensures that if a feature gains importance in the presence of other features, its SHAP value does not decrease.

Since computing exact SHAP values for DL models is computationally demanding, the SHAP Python package (https://shap.readthedocs.io/en/latest/) provides approximation methods based on previous feature attribution techniques

(Lundberg and Lee, 2017). In this study, we utilize the GradientExplainer method, which combines concepts from integrated gradients (Sundararajan et al., 2017), SHAP (Lundberg and Lee, 2017), and SmoothGrad (Smilkov et al., 2017). This method defines a reference baseline input (typically an all-zero input, Sundararajan et al. (2017)) and computes gradients of the model's predictions with respect to the input at multiple points along a path from this baseline to the actual input. By integrating these gradients, SHAP values are estimated, quantifying the contribution of each EEG signal component to the model's output.

In our approach, SHAP values highlight regions of the EEG signal that are crucial for classification. Positive SHAP values indicate features that strongly support a given classification, while negative values correspond to features that hinder correct predictions. Values near zero suggest that a feature has little to no impact on the classification (Lundberg and Lee, 2017).

One key factor in approximating integrated gradients as SHAP values, while ensuring local accuracy, is to employ a background dataset rather than relying on a single reference. Although this strategy increases computational demands, it can be effectively bypassed in our application to EEG data. The preprocessing pipeline ensures that each channel's signal is centered around zero (see Section 3.2). Consequently, by the central limit theorem, the average of an infinite number of baseline EEG signals will also have a zero mean. Thus, we can substitute the background dataset with an all-zeroed baseline signal. This zero reference not only simplifies the computation of SHAP values but also provides a more accurate approximation. Additionally, integrated gradients inherently satisfy the crucial properties of missingness and consistency through its gradient-based computations.

At this stage, we have established all the necessary components for generating feature attribution maps for each trial: the approximate method GradientExplainer, the MI data for evaluation, and the CNN model from which the SHAP values will be derived. In our analysis, we focus on the averaged SHAP values across individual trials, subjects, and datasets. First, we compute the SHAP values for a single EEG input—referenced against our background. Leveraging the work of Perez-Velasco et al. (2022), in Pérez-Velasco et al. (2024) we obtain each subject's SHAP values from a model trained on data from other individuals. For a given subject, only the fine-tuned weights from the dataset that excluded their trials are used. Importantly, we focus solely on correctly classified trials, as they provide a better insight into the EEG patterns relevant the examined task.

Next, we compute the average SHAP value map for each class on a

per-user basis, and then aggregate these averages across all subjects. By averaging first within subjects and then across them, we derive a more generalizable representation of the key features involved in each task. This aggregated analysis provides a robust interpretative framework that transcends individual variability, thereby yielding insights into the shared neurophysiological mechanisms underlying MI.

In Pérez-Velasco et al. (2024), the process culminates in four feature attribution maps, one for left- and right- hand classes under two conditions (with and without feedback). In these maps, the time series is displayed along the horizontal axis, and the channels are arranged vertically, establishing a direct correlation between SHAP values and the input signal's channels or time points. By aggregating the SHAP values along each axis, we can assess the relative contribution of each EEG channel or time segment to the classification. It is important to note that a positive SHAP value in a region does not necessarily indicate the presence of a distinguishing pattern, but rather the absence of any counteracting pattern. Therefore, to fully understand the classification process, we present and examine the feature attribution maps for both classes in tandem.

An open-source implementation of this SHAP-based method for EEG data analysis is available within the MEDUSA© software ecosystem (Santamaría-Vázquez et al., 2022), a Python-based framework designed to accelerate brain-computer interface and cognitive neuroscience research.

3.4.2 Channel selection based on SHAP values

In Pérez-Velasco et al. (2024), we propose a SHAP-based channel selection method to assess the relevance of extracted feature attributions and optimize electrode configurations for MI sessions. By aggregating channel-wise SHAP values from feature attribution maps generated using the 16-electrode configuration on the Physionet dataset (Goldberger et al., 2000) and the dataset from Stieger et al. (2021), we identify electrodes with minimal contribution to the final classification in the proposed DL model (Perez-Velasco et al., 2022).

Reducing the number of electrodes is a crucial step toward making brain-computer interface systems more practical, cost-effective, and easier to set up for real-world applications. Our selection approach prioritizes electrodes that exert the strongest influence on accurate MI classification. Rather than evaluating individual electrodes, we analyze them in pairs, reflecting the brain's functional symmetry. From the selected electrode configuration, we compare seven electrode

pairs positioned on opposite hemispheres (i.e., F7–F8, F3–F4, T7–T8, C3–C4, P7–P8, P3–P4, O1–O2), alongside the central pair (Cz–Pz). The four most significant electrode pairs, based on aggregated SHAP values, are selected as an 8-electrode optimized configuration. To validate this selection, we compare the classification performance of this optimized 8-electrode setup against the original 8-electrode configuration used in the first publication (Perez-Velasco et al., 2022).

This analysis provides insights into which EEG channels contribute most to MI classification and supports the development of more efficient BCI systems that maintains high classification accuracy with a reduced number of electrodes.

3.5 Investigation of motor imagery and motor execution relationship

The aim of this section is to present the specifics of the methodology followed in Pérez-Velasco et al. (2025). We examine how patterns derived from ME relate to those observed during MI with the tools developed in our previous works (Perez-Velasco et al., 2022; Pérez-Velasco et al., 2024). Understanding this relationship is critical for developing efficient BCI systems that can potentially leverage ME data to improve MI decoding. More concisely, we investigate for the first time whether a model trained on ME data can effectively classify MI data without any fine-tuning to MI in Pérez-Velasco et al. (2025).

3.5.1 Comparative analysis of electroencephalography patterns

In order to compare how the brain activity recorded during ME relates to that from MI, we analyze the performance of three different training–testing scenarios:

1. **ME→ME:** Training on ME data and evaluating on ME data.
2. **MI→MI:** Training on MI data and evaluating on MI data.
3. **ME→MI:** Training on ME data and evaluating on MI data.

All experiments follow a LOSO scheme, ensuring the model's ability to generalize to unseen participants. In each iteration, data from all but one user are used for training; the held-out user's data serve as the test set. The training data itself is further split so that 10% of the trials from each training subject is

reserved for validation. Early stopping is implemented when no improvement in validation loss is observed for 10 consecutive epochs (Goodfellow et al., 2016).

All models are trained for three-class classification (left hand, right hand, and resting), with the resting condition included to enhance the clarity of feature attribution (Pérez-Velasco et al., 2024). After the network is trained, *binary accuracy* (left vs. right) is assessed by ignoring the predicted probability of the resting class and comparing only the probabilities of the left and right classes. This dual view of performance (three-class versus binary) provides both a broader perspective on network decisions and a more targeted measure of MI decoding accuracy.

Models trained on ME data are evaluated on both ME and MI trials for each test subject, whereas models trained on MI data are evaluated only on MI trials. This approach allows for a direct comparison of how well the network transfers knowledge from physically executed movements to imagined ones.

To gain insights into the neural correlates of both ME and MI, we generate feature attribution maps using SHAP values (see Section 3.4). Notably, in Pérez-Velasco et al. (2025) we include the resting class in the three-class setup, enhancing the interpretability of SHAP values by distinguishing true task-related features from mere absence of features (Pérez-Velasco et al., 2024).

In addition, we examine different time windows within feature maps to observe how the underlying patterns change over time for both ME and MI. Visualizing these activations provides a valuable comparison of which brain regions and temporal segments are most critical to accurate classification in each condition.

3.5.2 Comparison metrics

To ensure a rigorous and objective comparison between conditions, we employ statistical analyses that assess differences in classification performance and feature attribution distributions. These methods provide a robust framework for validating the relationships between ME and MI while accounting for variability across subjects:

1. **Comparison of Classification Accuracies**

   To evaluate whether training on ME data influences classification performance on MI, we compare the average accuracy of the three experimental conditions (ME→ME, MI→MI, ME→MI). We apply the Wilcoxon signed-rank test (Wilcoxon, 1945) for paired comparisons, ensuring that differences in performance across conditions are assessed within the

same group of subjects. To correct for multiple comparisons, we apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), controlling the FDR.

2. **Confusion Matrices**

   To gain a deeper understanding of classification performance beyond a single accuracy metric, we analyze confusion matrices for each experimental scenario. These matrices provide insights into error patterns, revealing whether misclassifications tend to favor specific classes. This analysis is particularly relevant in the three-class setting (left, right, resting), where classification errors may stem from overlapping cortical representations or systematic biases in network predictions.

3. **Correlation Analyses**

   To explore the consistency of classification performance across ME and MI, we compute Pearson's correlation coefficient between individual subject accuracies in different conditions. A high correlation between ME→ME and ME→MI suggests that subjects who are well-classified in ME tend to be similarly well-classified in MI. Additionally, we analyze the correlation between ME→MI and MI→MI to determine whether classification performance is more strongly influenced by prior ME data or the inherent discriminability of MI patterns.

4. **Feature Attribution Comparisons**

   By leveraging feature attribution methods, we examine how neural activity is weighted across experimental conditions. Using SHAP values, we quantify the spatial and temporal importance of different electroencephalography channels during classification. To measure the similarity of feature importance between conditions, we compute the Pearson correlation coefficient (r) between SHAP values derived from ME and MI. This analysis identifies channels that consistently contribute to classification, providing insights into shared neural representations across tasks.

These statistical and visualization techniques allow for a comprehensive evaluation of model performance and feature attribution, enabling a detailed investigation of the relationship between ME and MI. By integrating multiple levels of analysis (accuracy comparisons, error distributions, correlation analyses, and feature importance) we provide a robust methodology for assessing the viability of

ME-trained models in MI classification. This approach is particularly relevant for optimizing brain-computer interface protocols and reducing reliance on extensive calibration procedures.

Chapter 4

Results

This chapter presents the empirical findings of this dissertation, structured around the three research hypotheses outlined in Section 2.1. Each section corresponds to a specific aspect of the investigation, directly relating to the studies included in the compendium of publications (see Appendix A).

First, Section 4.1 examines the performance of the DL model developed for MI classification, assessing its training schedule (Section 4.1.1), benchmarking against existing DL architectures (Section 4.1.2), and evaluating the impact of key architectural innovations through an ablation study (Section 4.1.3). These results validate hypothesis **H1**, demonstrating that an advanced DL architecture incorporating brain symmetry, inception modules, and residual connections significantly improve MI decoding accuracy.

Next, Section 4.2 explores findings from XAI techniques, supporting hypothesis **H2**. We provide a detailed visualization of the neural activation patterns extracted by the DL model (Section 4.2.1) and introduce a SHAP-based channel selection strategy (Section 4.2.2), which identifies the most informative electrodes contributing to MI classification.

Finally, Section 4.3 investigates the relationship between ME and MI, addressing hypothesis **H3**. We analyze the correlation between classification accuracies in both tasks (Section 4.3.1) and examine similarities and differences between the EEG activation patterns of these two tasks (Section 4.3.2), providing evidence for the feasibility of direct TL from ME to MI.

4.1 Performance of the deep learning model

This section outlines the main results presented in Perez-Velasco et al. (2022). Details the performance of the proposed *EEGSym* architecture in inter-subject MI classification, comparing it with four CNN baselines across five publicly available EEG datasets. Additionally, an ablation study examines the contributions of key architectural choices, specifically residual connections and explicit bilateral symmetry, to overall performance.

4.1.1 Training schedule

Before conducting comparisons, we evaluated the training and validation loss behaviors across the training process. As described in Section 3.3.4, the training approach comprised two stages: a pre-training phase on multiple datasets and a subsequent fine-tuning phase performed on the target dataset using LOSO cross-validation. Early stopping criteria, based on validation loss, were employed during both stages to minimize overfitting. The exact hyperparameters used are detailed in Section 3.3.4.

Figure 4.1 illustrates the evolution of training and validation losses during both pre-training and fine-tuning on the Physionet dataset (Goldberger et al., 2000), using the 8-electrode configuration. The graph highlights a consistent decrease in loss, with fine-tuning achieving convergence within a few epochs. A dotted line marks the early stopping point in the first stage of pre-training, ensuring optimal adaptation to the target dataset. Notably, pre-training on the Physionet dataset required approximately 4 hours and 18 minutes with an 8-electrode configuration, and about 6 hours and 25 minutes with a 16-electrode setup (on the hardware and software described in Section 3.3.4). Similarly, the fine-tuning process took roughly 7 minutes for the 8-electrode configuration and 12 minutes for the 16-electrode setup. This incremental training approach takes advantage of TL across multiple datasets, thereby enhancing classification accuracy on the target dataset (Goodfellow et al., 2016).

In general, training and validation accuracies converged steadily. When learning curves flattened, validation loss stopped improving and triggered the early stopping mechanismm, preventing overfitting.

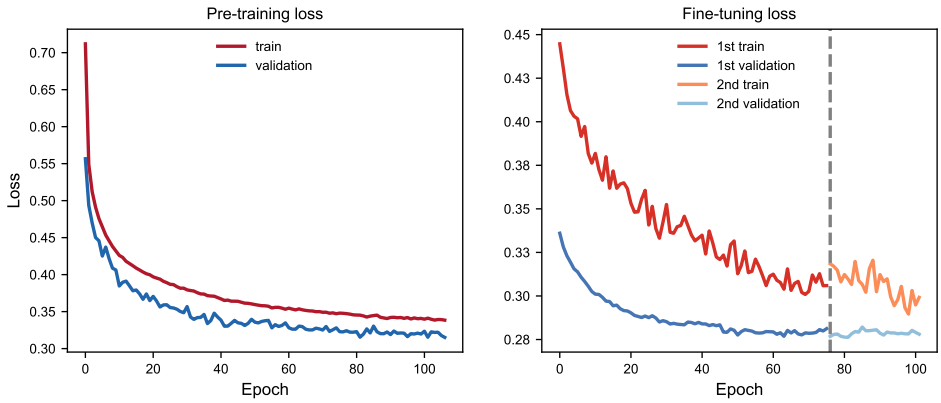


Figure 4.1: Training and validation loss curves for pre-training on the Physionet dataset (Goldberger et al., 2000) and subsequent fine-tuning across all subjects, excluding subject 2, using the 8-electrode configuration. The dotted line in the fine-tuning phase indicates the early stopping point for the initial stage of fine-tuning. Figure from Perez-Velasco et al. (2022).

4.1.2 Comparison with other deep learning networks

To assess the classification performance of *EEGSym*, we compared its results against four established CNN architectures: ShallowConvNet and DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), and EEG-Inception (Santamaria-Vazquez et al., 2020). The comparison involved five publicly available MI datasets: Physionet (Goldberger et al., 2000), OpenBMI (Lee et al., 2019), Kaya2018 (Kaya et al., 2018), Meng2019 (Meng and He, 2019), and Stieger2021 (Stieger et al., 2021). Results are summarized in Table 4.1, reporting mean accuracies with standard deviation ($\mu \pm \sigma$) and the number of users achieving BCI control (defined as accuracy $\geq$70%) for both 8-electrode and 16-electrode configurations.

The comparison results indicate that *EEGSym* consistently outperforms all baseline architectures, regardless of the electrode configuration used, achieving higher inter-subject classification accuracies. The improvements were statistically significant ($p < 0.05$) according to Wilcoxon signed-rank test (Wilcoxon, 1945), correcting the FDR with Benjamini-Hochberg approach. Notably:

1. **8-electrode setup**

   EEGSym obtained mean accuracies between 1% and 3% higher than baseline architectures across all datasets, achieving lower standard deviations and resulting in fewer subjects below the 70% accuracy threshold.

Table 4.1: Comparison of accuracies on target datasets for 8 and 16 electrode configurations (Perez-Velasco et al., 2022)

		Physionet (Goldberger et al., 2000)			**OpenBMI** (Lee et al., 2019)			**Kaya2018** (Kaya et al., 2018)			**Meng2019** (Meng and He, 2019)			**Stieger2021** (Stieger et al., 2021)		
#	**Architecture**	$\mu\pm\sigma$	**BCI Control**		$\mu\pm\sigma$	**BCI Control**		$\mu\pm\sigma$	**BCI Control**		$\mu\pm\sigma$	**BCI Control**		$\mu\pm\sigma$	**BCI Control**	
8 electrodes	ShallowConvNet	82.0±10.9	92/109	*	79.1±9.9	45/54	*	82.3±9.4	11/13	*	82.8±9.3	38/42	*	85.4±7.4	**60/62**	*
	DeepConvNet	82.8±10.7	95/109	*	80.5±9.3	**47/54**	*	82.4±9.2	**12/13**	*	83.6±9.1	39/42	*	86.2±7.2	**60/62**	*
	EEGNet	81.6±11.2	92/109	*	78.6±9.5	46/54	*	80.9±8.6	11/13	*	82.2±9.6	36/42	*	84.5±7.8	59/62	*
	EEG-Inception	82.7±10.8	92/109	*	80.3±9.4	**47/54**	*	81.7±9.1	**12/13**	*	84.4±8.4	**42/42**	*	87.3±7.0	**60/62**	*
	EEGSym	**84.5±9.7**	**99/109**		**82.0±9.6**	46/54		**84.7±9.1**	**12/13**		**85.2±8.3**	41/42		**88.4±6.5**	**60/62**	
16 electrodes	ShallowConvNet	86.2±9.6	100/109	*	80.0±9.7	46/54	*	83.1±9.6	11/13	*	85.5±8.6	40/42	*	87.5±7.3	59/62	*
	DeepConvNet	85.9±10.6	101/109	*	80.9±9.7	46/54	*	82.9±9.7	**12/13**	*	85.9±8.0	**41/42**	*	88.2±7.3	60/62	*
	EEGNet	83.1±10.7	97/109	*	79.6±9.8	45/54	*	82.1±9.1	11/13	*	82.4±9.4	39/42	*	85.7±8.2	58/62	*
	EEG-Inception	87.5±9.3	104/109	*	81.6±9.4	**48/54**	*	82.6±9.9	11/13	*	86.3±8.1	**41/42**	*	89.4±6.8	60/62	*
	EEGSym	**88.6±9.0**	**108/109**		**83.3±9.3**	46/54		**85.1±9.5**	**12/13**		**87.4±8.0**	**41/42**		**90.2±6.5**	**61/62**	

#: Number of electrodes used. $\mu\pm\sigma$: Mean classification accuracy and standard deviation, computed across all subjects using the leave-one-subject-out (LOSO) cross-validation method. BCI Control: Number of users who achieve brain-computer interface (BCI) control, defined as an accuracy of at least 70%. The highest results for each dataset and electrode configuration are highlighted in **bold**. Statistical differences between the mean accuracies of *EEGSym* and the baseline models were evaluated using the Wilcoxon signed-rank test, with false discovery rate (FDR) correction applied via the Benjamini-Hochberg method. Significant differences are indicated with * (p-value < 0.05).

2. **16-electrode setup**

   Performance further improved with additional channels, allowing *EEGSym* to achieve average accuracies between 87% and 90% across three datasets. BCI control was achieved by 268 of the 280 users (95.7%). EEG-Inception enabled control for 264 users, DeepConvNet for 260, ShallowConvNet for 258, and EEGNet for 252 users. Additionally, the proposed pre-training pipeline facilitated BCI control (≥90% of users) for all architectures using only 16 electrodes without subject-specific calibration.

Overall, these results demonstrate that balanced multi-dataset pre-training combined with *EEGSym*'s specialized design yields consistent classification performance across all evaluated datasets and supports fully inter-subject operation without the need for calibration.

4.1.3 Ablation study

To thoroughly investigate the contribution of key architectural elements within the *EEGSym* model, we performed an ablation study using the Physionet dataset (Goldberger et al., 2000). This dataset was selected for its large and diverse subject pool, enabling a reliable evaluation of the individual contributions of each architectural component on nearly half of the available subjects. The ablation analysis focused on two key features:

1. *Residual connections*, which enhance the model's ability to retain and propagate relevant information through multiple convolutional layers.

2. *Hemispheric symmetry*, which explicitly incorporates brain symmetry along the mid-sagittal plane into its architecture.

The results of applying these architectural choices separately and in combination for both 8-electrode and 16-electrode configurations are summarized in Table 4.2.

For the 16-electrode configuration, we observed statistically significant improvements ($p < 0.05$) in classification accuracy when either residual connections or hemispheric symmetry were introduced individually, relative to a baseline network that lacked both features. Combining these two architectural elements resulted in even higher accuracy, suggesting that the model benefits substantially from their joint application.

Table 4.2: Contribution of each novelty on Physionet

Res	Sym	8 electrodes Accuracy $\mu\pm\sigma$	8 electrodes BCI Control		16 electrodes Accuracy $\mu\pm\sigma$	16 electrodes BCI Control	
		82.9±10.4	90/109		87.2±9.4	104/109	
X		83.2±10.3	94/109		87.8±9.1	105/109	*
	X	84.1±10.1	96/109	*	87.7±9.8	103/109	*
X	X	84.5±9.7	99/109	*	88.6±9.0	108/109	*

Res: Indicates whether residual connections are incorporated into the architecture, facilitating spatial feature extraction across multiple processing layers. Sym: Specifies whether electrode inputs adhere to the structured hemispheric symmetry defined in Section 3.3.2. $\mu\pm\sigma$: Denotes the mean classification accuracy and standard deviation, computed across all subjects using the leave-one-subject-out (LOSO) cross-validation approach. BCI Control: Represents the number of participants who achieve brain-computer interface (BCI) control, defined as attaining an accuracy of at least 70%. Statistical significance between each tested model and the baseline was assessed using the Wilcoxon signed-rank test, with false discovery rate (FDR) correction applied via the Benjamini-Hochberg method. Significant differences over the baseline performance are marked with * (p-value < 0.05).

A similar trend was observed for the 8-electrode configuration, where combining both residual connections and hemispheric symmetry also led to the highest improvement in accuracy, though the benefit of residual connections alone was smaller and not statistically significant. Nonetheless, the inclusion of hemispheric symmetry consistently provided a significant improvement. These results underscore that residual connections and hemispheric symmetry offer complementary advantages: while symmetry enables efficient processing of common bilateral EEG patterns, residual connections facilitate deeper spatial feature extraction.

Overall, these findings reinforce that the superior performance of *EEGSym* is primarily due to its strategic architectural design choices, specifically the integration of residual connections and the explicit consideration of hemispheric symmetry. These features together significantly enhance the robustness and efficacy of *EEGSym* in EEG-based MI decoding.

4.2 Explainable artificial intelligence findings

This section outlines the main results presented in Pérez-Velasco et al. (2024). It details the application of the SHAP-based XAI methodology presented in

Section 3.4.1 to interpret the *EEGSym* architecture in MI tasks. It provides spatio-temporal feature attribution visualizations and introduces a SHAP-guided electrode selection strategy that optimizes the EEG montage for enhanced efficiency while maintaining classification performance.

4.2.1 Visualization of brain patterns

Figures 4.2 and 4.3 display the feature attribution maps generated using SHAP values for correctly classified MI trials from the Physionet dataset (Goldberger et al., 2000) (without feedback) and the Stieger2021 dataset (Stieger et al., 2021) (with feedback), respectively. Each figure depicts the temporal and spatial distribution of EEG feature relevance over a 3-second interval following the onset cue. Positive SHAP values, indicated in red, correspond to electrode signals that positively influence the classification of the given class (left- or right-hand), whereas negative contributions, shown in blue, indicate signals associated with the opposite class.

The figures 4.2 and 4.3 highlight:

- A temporal analysis shows that the highest contributions to correct classifications primarily occur within the initial 600-800 ms after cue onset, especially within the 200-400 ms interval. This timing was consistently observed across both datasets.

- Channel-wise percentage contributions, presented alongside each electrode, quantify the relative importance of each electrode for predicting the respective class. Electrodes placed over the prefrontal cortex (F7, F8) and sensorimotor cortex (C3, C4, T7, T8) generally had the highest contribution percentages, collectively accounting for a substantial proportion (around 50%) of the predictive contribution.

- During trials without feedback (Physionet dataset, Figure 4.2), electrodes from the prefrontal region (F7, F8) exhibited the greatest contributions, followed closely by central (C3, C4) and temporal electrodes (T7, T8). This pattern shows a clear lateralization based on the class, with contralateral frontal electrodes to the imagined hand providing significant positive SHAP values.

- In the dataset with real-time feedback (Stieger2021, Figure 4.3), the channel-wise distribution is similar to the Physionet dataset during the first 2 seconds

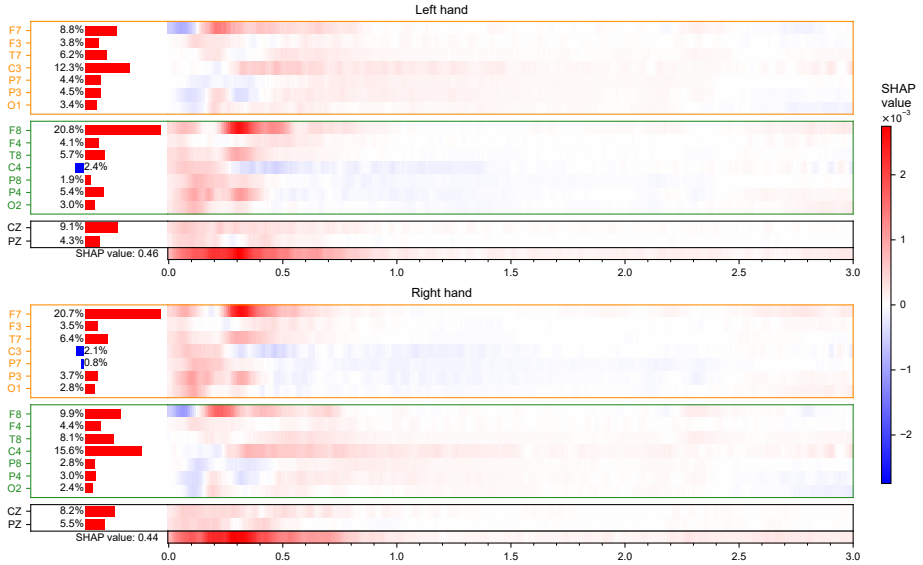


Figure 4.2: Feature attribution maps generated using shapley additive explanations (SHAP) values for motor imagery (MI) trials from the Physionet dataset (Goldberger et al., 2000), recorded without feedback. The vertical axis lists the 16 EEG channels positioned according to the 10-10 system. Electrodes located on the left hemisphere are highlighted in orange, those on the right hemisphere in green, and central electrodes in black. The color scale illustrates feature contributions: red areas indicate electrode–time regions positively contributing to the correct classification of the target class, while blue areas indicate regions negatively contributing or favoring the opposite class. Percentages next to channels represent each electrode's relative contribution. This visualization offers insight into the temporal and spatial distribution of EEG features that influence the model's predictions during the MI task without feedback. Figure from Pérez-Velasco et al. (2024).

without feedback. After the feedback onset at 2 seconds (marked by a green vertical line), there is an increased prominence of parietal (P3, P4) and occipital (O1, O2) electrode contributions. This activity remains notable throughout the feedback period, peaking shortly after feedback initiation.

- The midline electrodes (Cz and Pz) consistently showed the lowest contributions across both datasets.

Overall, these SHAP-based feature attribution visualizations clearly illustrate the EEG signal segments and specific cortical regions predominantly utilized by the model during classification, thereby confirming that the *EEGSym* architecture effectively leverages spatial and temporal EEG characteristics associated with left and right-hand motor imagery events.

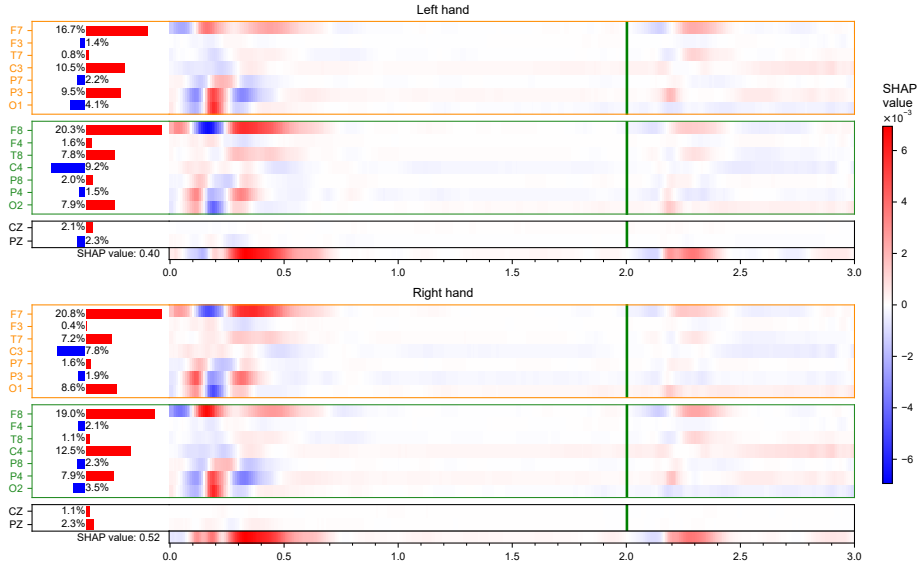


Figure 4.3: Feature attribution maps generated using SHAP values for motor imagery (MI) trials from the Stieger2021 dataset (Stieger et al., 2021), recorded with feedback. In the vertical axis, 16 channels from the 10-10 system are presented. The vertical axis lists the 16 EEG channels positioned according to the 10-10 system. Electrodes located on the left hemisphere are highlighted in orange, those on the right hemisphere in green, and central electrodes in black. The color scale illustrates feature contributions: red areas indicate electrode–time regions positively contributing to the correct classification of the target class, while blue areas indicate regions negatively contributing or favoring the opposite class. Percentages next to channels represent each electrode's relative contribution. This visualization offers insight into the temporal and spatial distribution of EEG features that influence the model's predictions during the MI task. The dashed vertical line at 2 s indicates the onset of the real-time visual feedback phase. Figure from Pérez-Velasco et al. (2024).

4.2.2 SHAP based channel selection

The feature attribution maps described in Section 3.4.1 were used to identify an optimized electrode configuration for EEG classification. The electrode pairs F7-F8, C3-C4, T7-T8, and P3-P4 emerged as having the highest cumulative SHAP contributions across both Physionet and Stieger2021 datasets. Specifically, this SHAP-based electrode montage accounted for 72% of the total absolute feature attribution, with the electrode pairs ranked as follows according to their relevance: F7-F8 (34.3%), C3-C4 (18.1%), T7-T8 (10.8%), and P3-P4 (9.4%).

Performance of the newly identified SHAP-based electrode configuration (electrodes F7, F8, C3, C4, T7, T8, P3, and P4) was compared against the original 8-electrode montage used in *EEGSym* (F3, F4, C3, C4, P3, P4, Cz, and

Table 4.3: Comparison of binary classification performance of SHAP-based selection of 8-electrode configurations

Study	Physionet (Goldberger et al., 2000) Accuracy(%)	Stieger2021 (Stieger et al., 2021) Accuracy(%)
Perez-Velasco et al. (2022)	84.5±9.7	88.4±6.5
Pérez-Velasco et al. (2024)	**86.5±10.6 ***	**88.7±7.0**

Accuracy (%): Mean accuracy and standard deviation computed across all subjects using a leave-one-subject-out (LOSO) cross-validation approach. The best performance for each dataset is highlighted in **bold**. Significant differences in mean accuracies were determined using the Wilcoxon signed-rank test, with false discovery rate (FDR) corrections applied via the Benjamini-Hochberg method. Statistically significant differences ($p < 0.01$) are indicated with an asterisk (*).

Pz), which according to this new analysis captured only 45% of the total SHAP contribution. Table 4.3 summarizes these comparisons. The newly identified 8-electrode setup demonstrated statistically significant improvements ($p < 0.01$) in the Physionet dataset, achieving a mean accuracy of 86.5% ± 10.6% compared to the original configuration's accuracy of 84.5%. For the Stieger2021 dataset, the SHAP-based configuration showed a small accuracy improvement (88.7% ± 7.0% versus 88.4% ± 6.5%), though this increase did not reach statistical significance. The statistical evaluation was performed using the Wilcoxon signed-rank test (Wilcoxon, 1945), applying false discovery rate (FDR) correction via the Benjamini-Hochberg procedure.

These results show that the electrode selection approach guided by SHAP values is effective in identifying more efficient electrode configurations that preserve classification performance.

4.3 Relationship between motor imagery and motor execution

This section outlines the main results presented in Pérez-Velasco et al. (2025) regarding the relationship between ME and MI. It details inter-subject classification experiments comparing ME, MI, and METL scenarios, and presents both accuracy correlation and spatio-temporal EEG activation pattern analyses using SHAP-based feature attribution.

4.3.1 Accuracy correlation

To investigate the direct transferability of ME learning to MI decoding, we evaluated three distinct inter-subject classification scenarios using Physionet's ME and MI datasets: ME→ME, MI→MI, and ME→MI.

The results, summarized in Table 4.4, demonstrate that ME→MI TL classification achieves comparable accuracies to conventional MI→MI for two-class scenarios (left- vs. right-hand), with accuracies of 86.21% ± 9.73% and 86.45% ± 10.33%, respectively. In the three-class scenario (left hand, right hand, rest), however, there was a minor accuracy reduction in ME→MI (78.08% ± 10.97%) relative to MI→MI (80.16% ± 10.97%).

To more precisely observe the cause for the differences in performance, Figure 4.4 presents the confusion matrices for each scenario. It shows a consistent distribution in resting class detection between ME→ME and ME→MI scenarios. This is straightforward, since the resting class is the same in ME and MI datasets. Additionally, classification of the resting class slightly improved when using ME data for training. Nevertheless, increased confusion was observed between MI (left-/right-hand) and resting classes in ME-trained models when tested on MI data (ME→MI scenario). Notably, the percentage of MI trials incorrectly classified as resting accounted for less than 5% of total trials in this ME trained model.

Correlation analysis of subject-specific inter-subject accuracies between ME→ME and ME→MI conditions (Figure 4.5) revealed a significant positive correlation (Pearson's coefficient $r = 0.6057$, $p < 0.001$). Additionally, the subject-specific accuracy comparison between MI→MI and ME→MI (Figure 4.6) indicated

Table 4.4: Accuracies of inter-subject experiments

Scenario	Accuracy (%)	
	3-class	2-class
ME to ME	83.07 ± 8.95*	88.51 ± 8.43*
MI to MI	80.16 ± 10.97*	86.45 ± 10.33
ME to MI	78.08 ± 10.97	86.21 ± 9.73

Accuracy (%): Obtained across subjects under an inter-subject (subject-independent) classification scheme. Scenario: Indicates training-to-testing data configurations, where ME denotes Motor Execution and MI denotes Motor Imagery. 3-class: Refers to classification among resting, left-hand, and right-hand conditions; 2-class: Refers to distinguishing between left-hand and right-hand classes only. Accuracies are presented as mean ± standard deviation. Statistical differences between scenarios were computed using the Wilcoxon signed-rank test, with false discovery rate (FDR) correction using the Benjamini–Hochberg method. Significant differences (compared to the other two scenarios) are marked with * (p-value < 0.01).

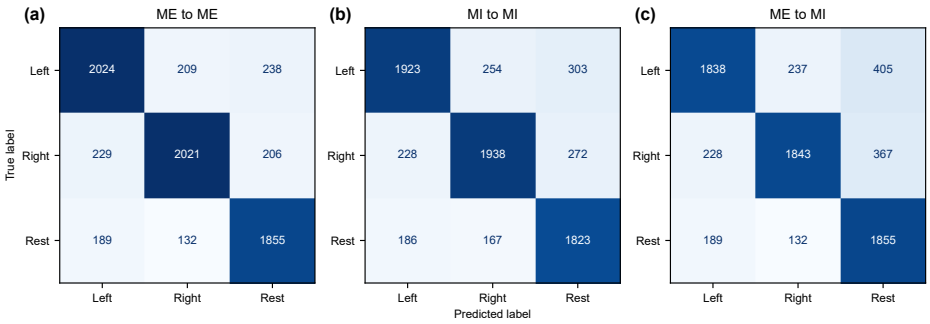


Figure 4.4: Confusion matrices illustrating inter-subject classification results for: **(a)** Motor Execution (ME) tasks classified with models trained on ME data; **(b)** Motor Imagery (MI) tasks classified using models trained on MI data; and **(c)** MI tasks classified using models trained on ME data (transfer learning scenario). Figure adapted from Pérez-Velasco et al. (2025).

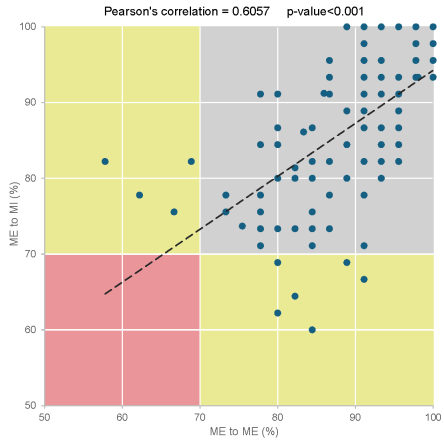


Figure 4.5: Correlation chart between classification accuracies obtained in ME→ME and ME→MI scenarios across subjects. ME: Motor Execution; MI: Motor Imagery. Figure from Pérez-Velasco et al. (2025).

that three subjects failed to achieve BCI control under both ME- and MI-trained models. Notably, training on ME data allowed three additional subjects (initially non-responsive under MI-only training) to attain successful BCI control, thus reducing the number of non-responsive subjects from nine to six.

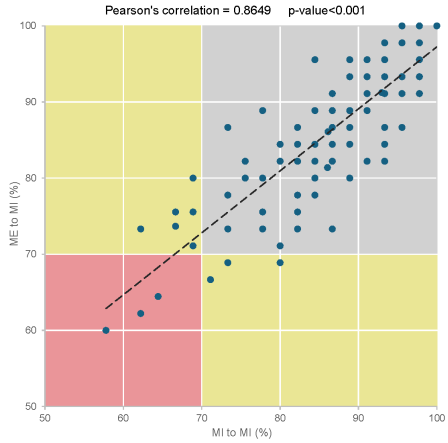


Figure 4.6: Correlation chart between classification accuracies obtained in ME→MI and MI→MI scenarios across subjects. ME: Motor Execution; MI: Motor Imagery. Figure from Pérez-Velasco et al. (2025).

4.3.2 Differences and similarities in electroencephalography activation patterns

Feature attribution maps obtained through SHAP analysis (described in Section 3.4.1) for left-hand (Figure 4.7), right-hand (symmetrical to left-hand, Figure 4.8), and resting classes (Figure 4.9) revealed both commonalities and differences in spatiotemporal patterns between the ME and MI paradigms.

Distinct spatial patterns emerged in the attribution maps for left and right-hand classes throughout the trial. Specifically, three clear temporal windows were identified based on these feature attribution maps:

1. **Early planning window (0–0.5 s)**

   For the left-hand class, both ME and MI paradigms showed significant feature attribution at electrodes AF8, FT8, and T10. Additionally, the ME paradigm emphasized contralateral central electrodes (C2, FC2, C3, FC3). The SHAP values exhibited strong positive correlations between ME and MI at contralateral fronto-temporal and parieto-occipital electrode locations, indicating similar cortical areas being engaged during initial movement planning or execution phases.

2. **Mid-interval window (0.5–1 s)**

   During this intermediate stage, both paradigms displayed prominent feature

attribution at contralateral motor cortex electrodes (C1, C2). However, notable differences emerged at frontal electrodes (FC1, FC2), where the ME paradigm focused primarily on contralateral frontal electrodes, whereas the MI paradigm relied predominantly on ipsilateral frontal electrodes. Consequently, the SHAP values at these electrode sites exhibited a strong negative correlation between paradigms. Nevertheless, a stable, high positive correlation remained present at contralateral fronto-temporal electrodes throughout this temporal window.

3. **Late sustained window (1–3 s)**

   In this late temporal window, both ME and MI paradigms exhibited sustained feature attribution at ipsilateral electrodes over the motor cortex, parietal, and temporal regions. The ME paradigm primarily highlighted electrodes C5 and C6, whereas the MI paradigm focused on electrodes C3, C4, T7, and T8. Additionally, strong positive correlations and significant SHAP attributions were consistently found at contralateral frontal electrodes (FP1/FP2, AF1/AF2, F3/F4), as well as at ipsilateral temporal electrodes (F7/F8), indicating robust consistency in neural patterns utilized by the DL model across tasks.

Additionally, topographic visualizations of mean SHAP values across temporal windows (Figure 4.10) reinforced these findings. Pearson's correlation coefficients calculated between MI→MI and ME→ME feature attribution maps across electrodes demonstrated consistently high overall correlations, with isolated regions of negative correlation primarily located at frontal electrodes (FC1/FC2) during the mid-interval temporal window.

Overall, the DL model effectively captured meaningful and consistent spatial-temporal EEG patterns across both MI and ME paradigms, with clear identification of cortical regions and temporal dynamics relevant for classification across tasks.

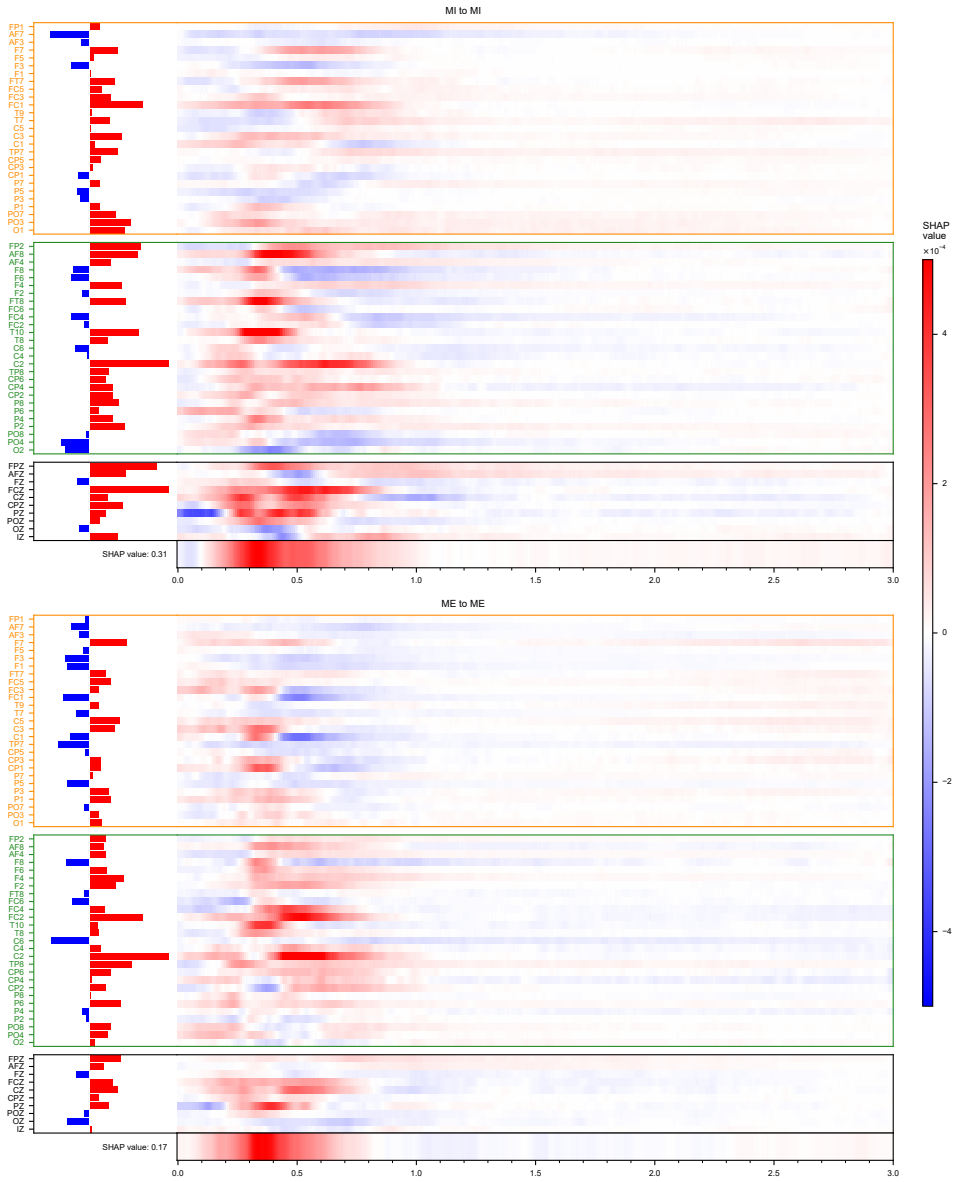


Figure 4.7: Feature attribution maps of the left-hand class showing shapley additive explanation (SHAP) values of a model trained on Motor Imagery (MI) applied to MI data and of a model trained on Motor Execution (ME) applied to ME tasks. In the vertical axis, 64 channels from the 10-10 system are presented. Channels corresponding to the left hemisphere of the scalp are marked in orange, the ones corresponding to the right hemisphere in green, and the central electrodes in black. The horizontal axis shows three seconds after the cue corresponding to the MI event. Positive SHAP values, indicated in red, highlight features that contribute to correctly predicting the target class, while negative SHAP values, shown in blue, denote features that detract from predicting the target class. The bars indicate the aggregated SHAP value of each channel. Figure from Pérez-Velasco et al. (2025).

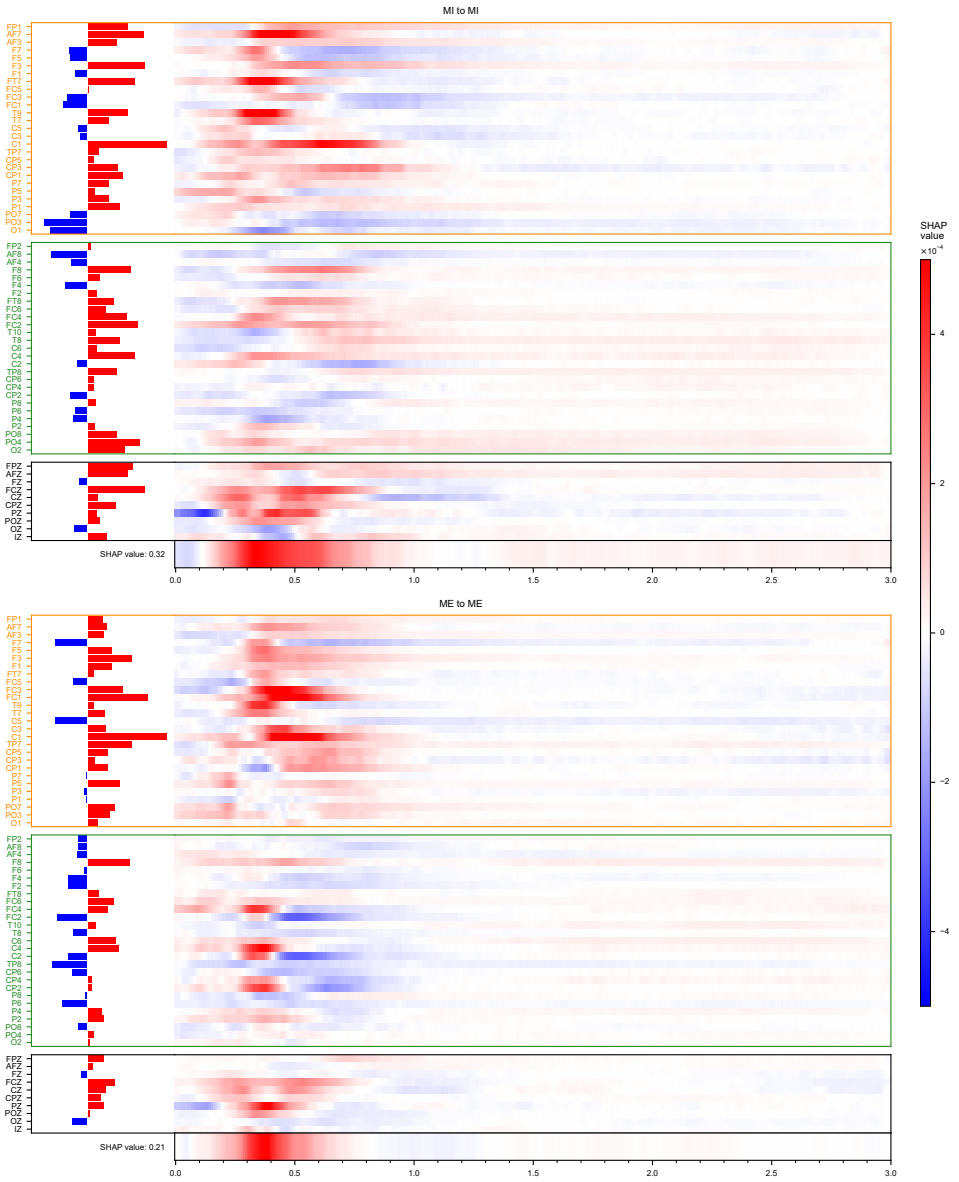


Figure 4.8: Feature attribution maps of the right-hand class showing shapley additive explanation (SHAP) values of a model trained on Motor Imagery (MI) applied to MI data and of a model trained on Motor Execution (ME) applied to ME tasks. In the vertical axis, 64 channels from the 10-10 system are presented. Channels corresponding to the left hemisphere of the scalp are marked in orange, the ones corresponding to the right hemisphere in green, and the central electrodes in black. The horizontal axis shows three seconds after the cue corresponding to the MI event. Positive SHAP values, indicated in red, highlight features that contribute to correctly predicting the target class, while negative SHAP values, shown in blue, denote features that detract from predicting the target class. The bars indicate the aggregated SHAP value of each channel. Figure from Pérez-Velasco et al. (2025).

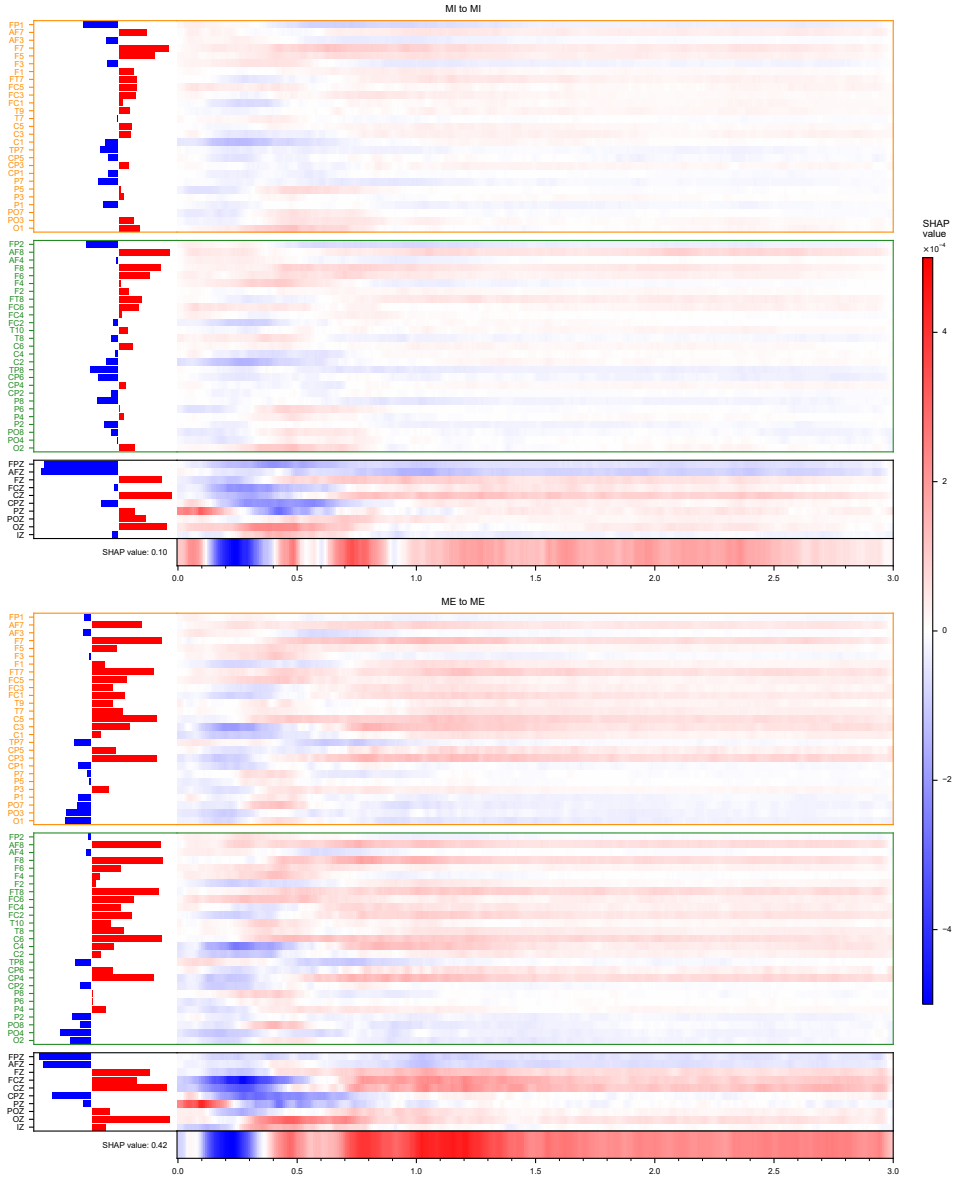


Figure 4.9: Feature attribution maps of the resting class showing shapley additive explanation (SHAP) values of a model trained on Motor Imagery (MI) applied to MI data and of a model trained on Motor Execution (ME) applied to ME tasks. In the vertical axis, 64 channels from the 10-10 system are presented. Channels corresponding to the left hemisphere of the scalp are marked in orange, the ones corresponding to the right hemisphere in green, and the central electrodes in black. The horizontal axis shows three seconds after the cue corresponding to the MI event. Positive SHAP values, indicated in red, highlight features that contribute to correctly predicting the target class, while negative SHAP values, shown in blue, denote features that detract from predicting the target class. The bars indicate the aggregated SHAP value of each channel. Figure from Pérez-Velasco et al. (2025).

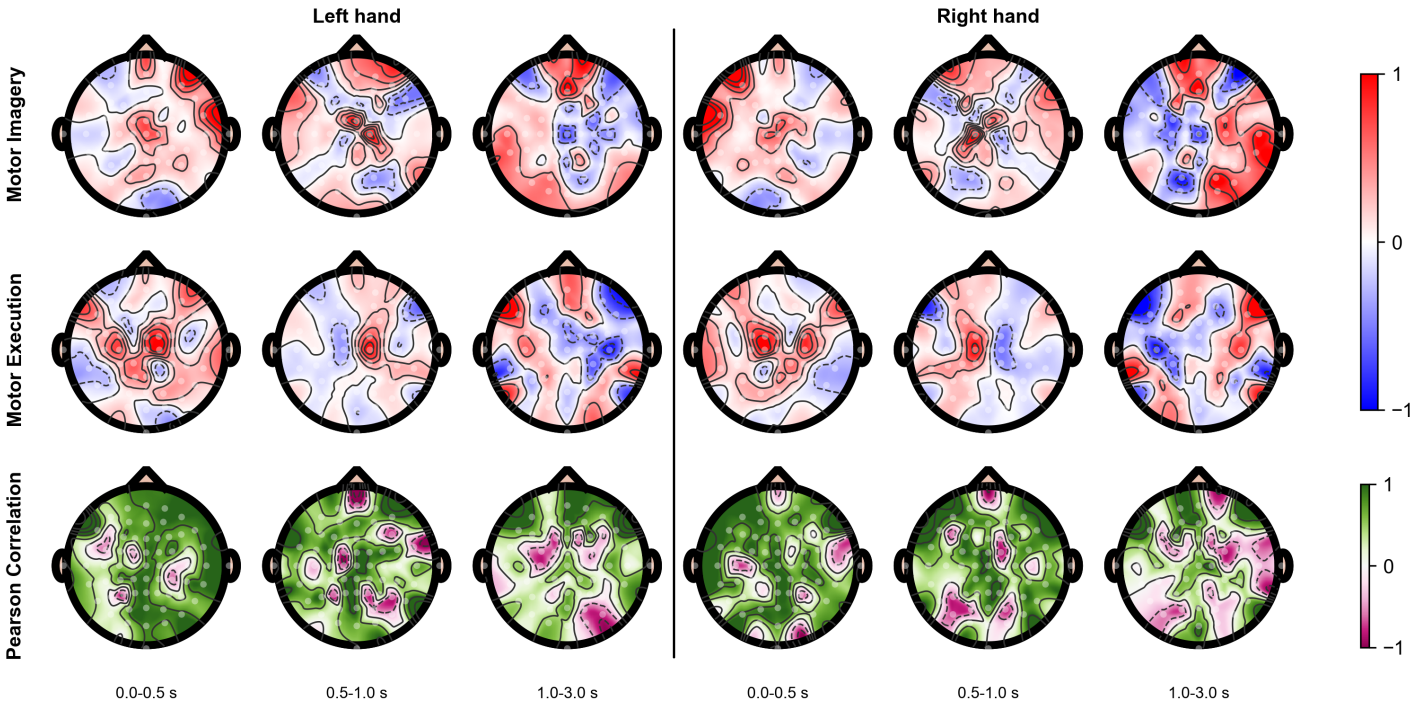


Figure 4.10: Mean feature attribution topographic plot showing shapley additive explanations (SHAP) values for Motor Imagery (MI) and Motor Execution (ME) tasks, along with the Pearson's correlation between them for the temporal windows from 0 to 0.5, 0.5 to 1, and 1 to 3 seconds. Positive SHAP values, indicated in red, highlight features that contribute to correctly predicting the target class, while negative SHAP values, shown in blue, denote features that detract from predicting the target class. Figure from Pérez-Velasco et al. (2025).

Chapter 5

Discussion

In this Doctoral Thesis, we have explored DL-based approaches for EEG decoding, the interpretability provided by XAI methods, and the possibility of direct TL between ME and MI. The findings contribute both practically and theoretically to the field of BCIs, addressing key challenges in MI classification, model interpretability, and cross-task transferability.

Firstly, we analyzed the performance of our DL model (Section 5.1), focusing on the impact of network design choices, DA strategies, and cross-dataset training. The results demonstrated that specific architectural features, such as residual connections and explicit bilateral symmetry, contributed to improved classification accuracy. Additionally, the integration of DA techniques and multi-dataset pre-training led to models that generalized effectively across different subjects and datasets.

Secondly, we enhanced model interpretability via XAI by employing a SHAP-based approach to generate feature attribution maps (Section 5.2). These maps provided a detailed visualization of the spatio-temporal EEG patterns that influence model predictions, offering reliable insights into the neural features captured during MI tasks. Furthermore, the feature attribution maps supported the selection of an optimized EEG montage, highlighting the potential of XAI to improve both model transparency and practical BCI implementation.

Lastly, we investigated the relationship between MI and ME, assessing the viability of direct TL from ME data to MI classification (Section 5.3). This finding is supported by inter-subject experiments that showed comparable performance between MI→MI and ME→MI training scenarios. Additionally, feature attribution analyses revealed shared and distinct neural activation patterns

between MI and ME, further informing the theoretical understanding of their functional overlap.

In this chapter, the key findings of the Doctoral Thesis are discussed in relation to previous work, emphasizing their implications for the development of more robust, interpretable, and adaptable BCI systems. Finally, the main limitations of this research are outlined (Section 5.4).

5.1 Interpretation of the developed deep learning architecture

In this section, we synthesize how our methodological choices culminated in a robust EEG-based classification system for MI developed in the first work of this Doctoral Thesis (Perez-Velasco et al., 2022). Our approach combines a brain-inspired DL architecture, targeted DA techniques, and extensive TL across multiple public EEG datasets. *EEGSym* leverages these facets to address BCI inefficiency on a broad cohort of participants in a fully inter-subject paradigm.

5.1.1 Factors influencing model performance

Our design is driven by three core innovations:

1. **Model Architecture and Symmetry**

   EEGSym employs a symmetric design that mimics the hemispheric organization of the brain, augmented by residual connections. This configuration facilitates the extraction of spatial features while enabling effective performance with reduced electrode configurations.

2. **Data Augmentation and Transfer Learning**

   We introduce perturbation-based DA techniques (*patch perturbation*, *hemisphere perturbation*, and *random shift*) to expand the variability in the training data. Coupled with multi-dataset pre-training, these methods mitigate the common issue of limited sample sizes, significantly reducing the prevalence of "BCI-inefficient" users and enhancing overall accuracy.

3. **Evaluation and Performance Stability**

   Our framework was validated on five datasets: Physionet (Goldberger et al., 2000), OpenBMI (Lee et al., 2019), Kaya2018 (Kaya et al., 2018),

Meng2019 (Meng and He, 2019), and Stieger2021 (Stieger et al., 2021). *EEGSym* consistently surpasses a 70% accuracy baseline, a practical indicator of BCI control, in more than 95% of users (268 out of 280), despite varying acquisition conditions.

Finally, the choice of time window for classification played a significant role. While many approaches rely on 3–4,s segments, our experiments revealed that the most discriminative signals often emerge within the first second after the cue. By focusing on these early neural signatures, the model is able to achieve high performance with shorter, more practical EEG segments.

Beyond achieving impressive accuracy gains, our results demonstrate that integrating domain knowledge, such as hemisphere symmetry, with large-scale TL can eliminate the need for individual calibration. Importantly, the ability to maintain state-of-the-art performance with only 8 or 16 electrodes further emphasizes the accessibility and practical deployment potential of our approach in rehabilitation or home-based BCI systems.

5.1.2 Comparison with previous inter-subject deep learning approaches

When evaluated against established architectures, such as *ShallowConvNet*, *DeepConvNet* (Schirrmeister et al., 2017), *EEGNet* (Lawhern et al., 2018), and *EEG-Inception* (Santamaria-Vazquez et al., 2020), the proposed approach featuring *EEGSym* **outperforms** these baselines in inter-subject classification. Notably, *EEGSym* maintains high performance even with few electrodes (8 or 16), underscoring its efficiency and transferability to realistic scenarios.

A key contributor to these gains is the **subject-independent training** paradigm, where data from multiple sources are aggregated to train a single model capable of robustly generalizing across individuals. This approach leverages diverse sessions and recording conditions, diminishing the likelihood of overfitting to dataset-specific nuances. Coupled with DA, this design attains consistent classification rates exceeding 80-90% for most users in these public large-scale databases.

Another practical merit is the reduced **inference time**, inherent to any DL network of reasonable size. The model processes individual EEG trials in around 30 ms on standard GPU hardware, making it suitable for *real-time* or *online* decoding scenarios. This short latency is especially relevant in rehabilitation tasks,

Table 5.1: Comparison with binary classification of previous literature

	Study	TW (s)	Number of electrodes	Accuracy $\mu\pm\sigma$
Physionet (Goldberger et al., 2000)	Dose et al. (2018)	3	64	80.38±12.54
			16	78.03
			9	75.85
	Kostas and Rudzicz (2020)	3	64	82.84
	Fan et al. (2021)	3	64	82.88
			14	78.98
	Varsehi and Firoozabadi (2021)	3	14	83.63
			9	81.26
	Ours (Perez-Velasco et al., 2022)	3	16	**88.56**±8.96
			8	**84.45**±9.70
OpenBMI (Lee et al., 2019)	Kwon et al. (2020)	4	19	74.15±15.83
	Zhang et al. (2021)	4	62	84.19±9.98
	Ours (Perez-Velasco et al., 2022)	3	16	**84.72**±11.73 *
			8	**82.93**±12.10 *

TW: time window duration used for classification. $\mu\pm\sigma$: mean accuracy and standard deviation obtained across all subjects in a subject-independent scheme. * Results are adjusted to mimic the reporting in Kwon et al. (2020) and Zhang et al. (2021), where only the trials from the last test run were used for evaluation.

where continuous updates foster user engagement and plasticity-oriented training, where responsiveness is crucial.

A detailed comparison with previous studies is provided in Table 5.1. Notably, the Physionet (Goldberger et al., 2000) dataset comprises data from 109 subjects; however, several studies used for comparison excluded data from a few subjects. For example, Dose et al. (2018) did not specify the excluded subjects, while Fan et al. (2021) and Varsehi and Firoozabadi (2021) removed subjects S088, S092, S100, and S104 due to data corruption, and Kostas and Rudzicz (2020) excluded S088, S090, S092, and S100. In contrast, our approach utilizes all available subjects and trials, thereby providing a more robust evaluation metric.

Table 5.1 summarizes the binary classification performance of our method against previous studies. Our approach, evaluated on both Physionet (Goldberger

et al., 2000) and OpenBMI (Lee et al., 2019) datasets, demonstrates that *EEGSym* outperforms earlier DL approaches (even when using only 16 electrodes), while some baseline studies required all 64 electrodes. Furthermore, *EEGSym* achieves competitive performance with only 8 electrodes on the Physionet dataset, and in the OpenBMI dataset our method attains comparable accuracies using only 16 out of the 62 available electrodes.

The improved performance of *EEGSym* can be largely attributed to the enhanced TL strategy. Unlike previous works that exploit data from a single dataset for inter-subject training (Dose et al., 2018; Fan et al., 2021; Kostas and Rudzicz, 2020; Kwon et al., 2020; Zhang et al., 2021), our approach pre-trains the network using multiple publicly available datasets sharing the same MI paradigm. This TL scheme results in higher inter-subject accuracies, as evidenced by our results on the Physionet dataset, where both baseline models and *EEGSym* outperform earlier deep learning approaches that operated on full 64-channel configurations.

In summary, the combined use of advanced preprocessing, data augmentation, and extensive pre-training significantly elevates deep learning performance in MI classification tasks. These findings underscore the potential of our approach to advance DL-based BCIs, achieving state-of-the-art results with a reduced electrode set.*EEGSym* builds upon and improves existing CNN models by integrating brain-inspired design principles and residual connections, thereby offering a viable pathway toward calibration-free MI-based BCIs and potentially eliminating a significant barrier to widespread adoption. The subsequent sections delve deeper into how XAI methods further illuminate these results and how MI classification intertwines with ME paradigms.

5.2 Insights from explainable artificial intelligence

DL architectures have demonstrated superior performance in MI classification tasks; however, they are often criticized as "black-box" models due to their limited interpretability. The work of Pérez-Velasco et al. (2024) underscored the necessity of illuminating these inner workings to better understand how DL algorithms discriminate MI-related patterns. In this Doctoral Thesis, we describe the employed SHAP-based XAI method to unveil the spatiotemporal features most critical for MI classification. By quantifying each electrode's and each

time window's contribution to the DL model's decisions, we aimed to provide transparent evidence of which brain regions and time intervals are essential for accurate decoding.

5.2.1 Spatial insights: a multi-regional network

Our SHAP-based analysis of feature attribution maps revealed that multiple cortical regions, not just the sensorimotor strip, exhibit significant involvement in MI tasks. Specifically:

1. **Prefrontal Cortex (PFC)**

   Frontal electrodes, particularly F7 and F8, displayed high SHAP values within the first second after the cue onset. This aligns with the notion that the PFC contributes to the planning and initiation of imaginary movements, mirroring its role in ME studies (Miller and Hatsopoulos, 2012). The stronger PFC engagement in no-feedback conditions suggests a reliance on internal goal setting for movement imagination as previously noted by Marcos-Martínez et al. (2021).

2. **Posterior Parietal Cortex (PPC)**

   Parietal and occipital electrodes (e.g., P3, P4, O1, O2) gained prominence when real-time feedback was provided, echoing the known involvement of the PPC in integrating visual or spatial information. Figure 4.3 illustrates how P3, P4, and nearby sites became increasingly important after the onset of feedback, suggesting that MI can also engage visually guided neural circuits akin to ME.

3. **Sensorimotor Regions (M1 and S1)**

   While the sensorimotor strip (C3, C4) remained integral to classification, the SHAP analysis showed that MI tasks recruit a more distributed network than often assumed. Central electrodes still contributed robustly to the discrimination of left- versus right-hand MI, but their importance was modulated by feedback conditions and by activity in frontal and parietal areas.

Together, these spatial insights indicate that MI involves a complex interplay among PFC, PPC, and sensorimotor regions, supporting the shift of focus on sensorimotor areas of previous works (Lee et al., 2019; Sebastián-Romagosa et al., 2020; Stieger et al., 2021). In clinical populations, such as stroke survivors,

these networks may be reorganized. Therefore, a BCI that accurately captures this extended neurophysiological footprint, including both the core SMR and the auxiliary frontoparietal activity, may offer benefits for clinical translation. Such a BCI that employs these pre-trained networks could capture the patient's remaining neural capacity and offer meaningful feedback.

5.2.2 Temporal insights: early vs. late electroencephalography patterns

Our analysis also highlighted a temporal hierarchy in MI decoding:

1. **Early Planning Window (0-1 s)**

   A large fraction of the network's discriminative power emerged within the initial second of each trial. SHAP values revealed that cortical signals related to the transition from resting state to motor intention are especially salient, echoing the preparatory activity documented in ME studies (Miller and Hatsopoulos, 2012).

2. **Prolonged Feedback Influence**

   In the presence of real-time visual cues, electrodes over parietal-occipital sites (e.g., P3, P4, O1, O2) contributed more strongly during the second half of the trial. This delayed effect underscores how feedback-related processes can extend the window of neural relevance for MI discrimination, as seen in Figure 4.3.

These temporal findings suggest that focusing on shorter time segments (approximately the first second after cue) can yield robust decoding, which will represent a faster feedback of actual brain activity in online BCI scenarios.

5.2.3 Practical impact: channel selection and reduced montages

One of the most notable outcomes of the SHAP analysis is the identification of a *reduced 8-electrode configuration* that can preserve the accuracy previously achieved with 16 channels. By systematically ranking electrodes according to their cumulative SHAP values, we found that positions F7, F8, T7, T8, C3, C4, P3, and P4 accounted for a significant majority of the total feature attribution.

This finding has clear **practical implications**. Reducing the number of electrodes streamlines preparation for both researchers and end users, thereby

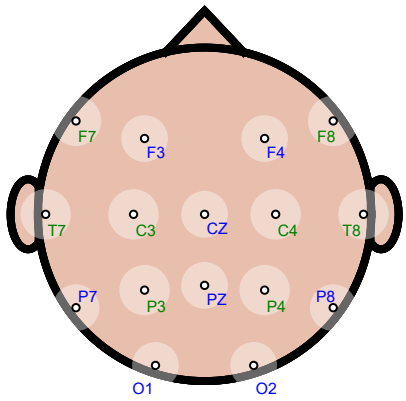


Figure 5.1: 16-electrode configuration analyzed in Perez-Velasco et al. (2022) and Pérez-Velasco et al. (2024). SHAP-based 8-electrode selected configuration are labeled in green. Figure from Pérez-Velasco et al. (2024).

lowering hardware expenses and setup time. Moreover, shorter setup durations may bolster user motivation, an important factor in rehabilitation and home-based BCI contexts (Lee et al., 2019; Stieger et al., 2021). Figure 5.1 depicts the selected electrodes in green, and Figure 5.2 illustrates how incremental additions of electrode pairs affect overall model accuracy.

5.2.4 Broader implications and concluding remarks

Overall, these XAI-driven insights enrich our understanding of how DL models interpret MI-specific brain activity, offering an evidence-based rationale for focusing on certain spatial regions (PFC, PPC, sensorimotor) and temporal windows (first second post-cue). This improved transparency tackles a major drawback of traditional CNN-based pipelines: the uncertainty about which neural signals contribute to high classification accuracy.

By bridging neurophysiological theories of motor planning with practical BCI requirements, our approach paves the way for more efficient, calibration-free (or minimally calibrated) setups. Future work could extend these findings to more complex paradigms (e.g., multi-limb or continuous control) or investigate additional feedback modalities. In essence, explainable DL frameworks not only improve end-user experience through reduced electrode montages but also advance the scientific community's grasp of the cortical networks that underlie MI and ME alike.

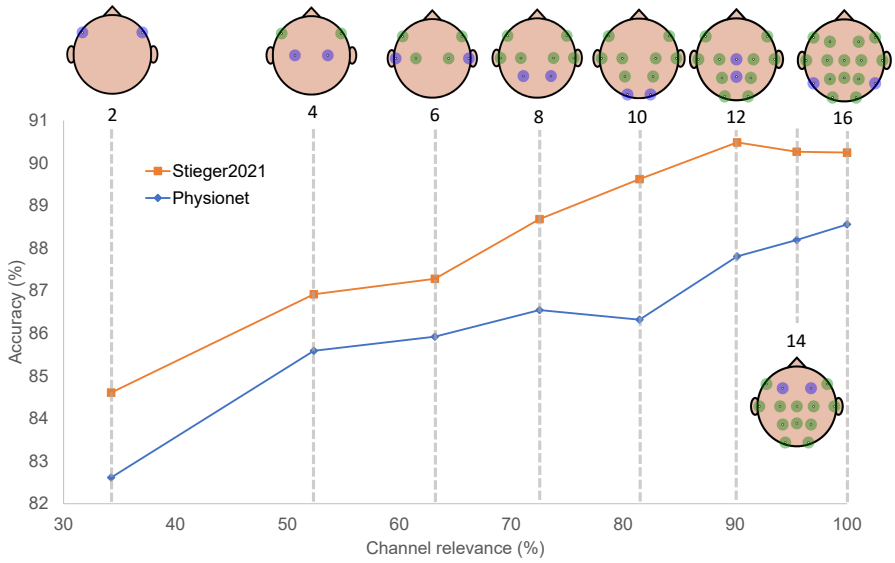


Figure 5.2: Correlation between classification accuracy on the Physionet (Goldberger et al., 2000) and Stieger2021 (Stieger et al., 2021) datasets and each channel's relevance, where relevance is defined as the percentage of absolute SHAP values contributed by that channel relative to the total absolute SHAP values across all channels and time intervals. Each point in the plot corresponds to a particular electrode configuration used for classification, with the newly added electrode pair shown in blue and previously added pairs depicted in green. Figure from Pérez-Velasco et al. (2024).

5.3 Analysis of motor imagery and motor execution relationship

This section examines how ME and MI relate within the context of DL classification, drawing on observations of the third study of this Doctoral Thesis: (Pérez-Velasco et al., 2025). Notably, we found that DL models trained *exclusively* on ME data can classify MI trials at accuracies comparable to MI-trained models. Such findings underscore an extensive overlap in the spatiotemporal patterns underpinning ME and MI, offering a compelling avenue for more efficient BCI protocols.

5.3.1 Overlap and synergy in neural representations

Physiological evidence has long suggested that ME and MI engage common sensorimotor processes. Our results support this viewpoint, showing that

models trained on ME data transferred effectively to MI tasks (e.g., achieving 86.21% ± 9.73% accuracy, nearly matching MI-trained models at 86.45% ± 10.33%). Additionally, the correlation between ME→ME and ME→MI accuracies (Figure 4.5) indicates that participants excelling in ME classification could similarly perform well in MI tasks. Conversely, users who fail to modulate their EEG for MI tasks also exhibit limited discrimination during ME runs. This pattern suggests that "BCI inefficiency" may reflect difficulties in reliably detecting or interpreting a user's *existing* neural modulations, rather than a complete lack of capacity to produce them.

The inter-task TL outcomes described above reinforce the notion that ME and MI rely on a shared neural activity. Consequently, a portion of the subject pool identified as "non-responsive" in MI paradigms might be revisited with improved EEG measurement or classification techniques, as our observations imply that fundamental sensorimotor rhythms are present but remain underexplored or inadequately captured. In this context, the inclusion of the resting class played a critical role. By forcing the model to learn a defined no-control state, it prevents the misclassification of neural noise or non-task-related activity as a motor command, a function that is essential for the reliability of any practical BCI system. The confusion matrices in Figure 4.4 confirm that the model effectively learned to separate intentional motor patterns from this baseline. This insight expands upon earlier work (Lee et al., 2022; Miao et al., 2023) that also highlighted substantial commonalities between ME and MI brain activity but had not investigated *direct* ME→MI TL in a calibration-free manner. Overall, recognizing this overlap can reorient future research toward refining feature extraction, instead of attributing low performance solely to user-specific deficits.

5.3.2 Practical benefits for calibration and rehabilitation

BCIs that can switch smoothly from decoding ME to MI offer clear practical benefits. We have demonstrated that training the DL network on verifiable ME data creates feature representations that later generalise to MI. An strategy that employs these models with capabilities on ME and MI data lowers the entry barrier for newcomers, boosts motivation during therapy, and adapts as motor ability recovers. The three main benefits are detailed below:

1. **Streamlined calibration**

   One pragmatic advantage of ME-based training is the ease of gathering *task-compliant* EEG data. Physical limb movement is overtly verifiable,

thereby reducing ambiguity about user engagement, an issue encountered in MI studies. By training on data that reliably reflects valid motor signals, the DL model extracts robust feature representations that, as our study demonstrates, transfer seamlessly to MI. This approach can simplify calibration routines and shorten the learning curve in multi-user or clinical settings.

2. **Motivation and user experience**

   Because ME tasks require less mental effort than sustained MI, they can serve as an accessible entry point for new BCI users or for patients in rehabilitation scenarios. Learners initially perform *actual* movements, becoming acquainted with the BCI's feedback and control interface before progressing to MI. The correlation seen in Figure 4.5 also suggests that performance during ME training could provide a useful indicator of subsequent MI proficiency, though further work is needed to fully quantify its predictive capacity.

3. **Targeted neurorehabilitation**

   In rehabilitation contexts, providing real-time feedback for partially executed movements can reinforce neural pathways tied to actual motor function. Over time, as users improve or transition to pure imagery, the same model can smoothly accommodate MI data with minimal or no additional calibration. This "ME-first" approach can be especially beneficial for individuals with limited or inconsistent motor control, offering them a more intuitive and confidence-building gateway into MI-based BCI therapy.

5.3.3 Impact on brain-computer interface development

From a **design perspective**, leveraging ME data to train DL models for MI classification could substantially reduce the overhead and uncertainty traditionally associated with MI calibration. Obtaining verifiably correct EEG labels during ME sessions mitigates the core challenge of ensuring user compliance, since involuntary or improper MI attempts remain more difficult to detect. This synergy between ME and MI has several tangible benefits:

1. **Accelerated Training**

   Practitioners can rely on relatively short ME sessions to acquire high-quality EEG signals, thus rapidly reaching functional performance levels when the model is repurposed for MI tasks.

2. **Enhanced User Engagement**

   By beginning with the more intuitive, physically verifiable motor actions, early success rates improve and encourage user adherence, crucial in demanding applications like stroke rehabilitation.

3. **Long-Term Adaptability**

   As users recover or gain MI proficiency, the model can be further finetuned (or left unaltered if no improvement is necessary) to reflect the evolving neural representation of imagined movements, supporting an integrative, staged therapy.

Overall, the synergy between ME and MI classification presented here carries substantial implications for real-world BCI deployment. By beginning calibration with physically visible tasks, researchers and clinicians establish a reliable baseline for each user, simplifying the subsequent transition to MI-centric paradigms. This perspective extends to applications ranging from neurorehabilitation protocols, where bridging residual movements and mental practice is vital, to interactive domains such as gaming. In both cases, quick and accurate user control is essential for an engaging experience.

5.4 Limitations

Although the DL approaches proposed in this dissertation have shown considerable promise for MI-based BCIs, several limitations must be acknowledged. Our first two evaluations (Perez-Velasco et al., 2022; Pérez-Velasco et al., 2024) primarily focused on binary classification (e.g., left- vs. right-hand MI), which may not capture the full complexity of neural activity that could emerge from multi-class paradigms including additional movements or a resting state. Expanding the analysis to a broader array of movement types provided more informative feature attribution maps in Pérez-Velasco et al. (2025).

The datasets used in this work were collected from healthy participants, introducing a potential cohort bias. Since individuals with neurological conditions or motor impairments often exhibit altered cortical dynamics, further validation with clinical populations is necessary. While this focus on healthy subjects was essential for developing and validating the core methodologies on large-scale, consistent data, the direct applicability of these findings to a clinical context, where neural signals can be weaker or more variable, remains to be demonstrated.

Additionally, While the ME-based preparatory approach may work well for healthy or mildly impaired users, its feasibility for severely motor-impaired individuals remains uncertain and warrants additional investigation.

Another limitation arises from the reliance on offline datasets and single-session evaluations. The absence of real-time feedback with the methods developed and longitudinal studies means that factors such as learning effects, performance stability, and user adaptation over time have not been fully explored. Future research should incorporate online and long-term assessments to determine whether these DL models can effectively adapt to evolving EEG signals and user-specific changes. Furthermore, this work is exclusively centered on synchronous BCI paradigms. The transition to asynchronous, or self-paced, control, which would allow users to issue commands at will, presents a distinct set of challenges, including the need for robust intent detection.

Moreover, the resting state data used in this work was collected during a passive, eyes-open condition. This may not fully represent the variety of non-control mental states a user might experience in a real-world setting, such as active thinking, listening, or distraction. The model's ability to generalize and maintain a low false positive rate during these more complex active rest periods remains an open question and a critical area for future online validation.

Additionally, although the inter-subject performance of the models was robust, fine-tuning to the target dataset to adapt to new datasets or hardware variations enhanced the accuracy reported, which diminishes the "plug-and-play" convenience of the approach. Collecting more diverse data from multiple centers and user groups could improve model generalizability to unseen datasets.

This work also concentrated on CNN-based architectures, such as *EEGSym*. While these models provided strong performance and reliable SHAP-based feature attributions, emerging Transformer-based models (e.g., BENDR (Kostas et al., 2021), MAEEG (Chien et al., 2022), s-JEPA (Guetschel et al., 2024)) may capture different features or offer additional advantages. A direct comparison between CNNs and Transformer architectures in terms of the feature attribution patterns observed remains an important direction for future research.

Finally, although fixed 8- and 16-electrode configurations maintained high accuracies, these setups may not be optimal for every application or user type. Future work could explore adaptive or individualized electrode selection strategies, particularly for more complex, multi-session, or multi-class scenarios.

In summary, the results presented here emphasize the **potential feasibility** of large-scale, inter-subject DL training for MI-based BCIs, including approaches

leveraging ME for more accessible calibration. At the same time, addressing the above limitations—particularly by testing a broader spectrum of paradigms, subject populations, and modeling techniques—will be essential for transitioning from promising laboratory results to robust, *clinically and practically deployable* BCI systems.

Chapter 6

Conclusions

The central objective of this Doctoral Thesis has been the development and integration of advanced DL and XAI techniques to enhance MI-based BCIs. Initially, a novel DL architecture termed *EEGSym* was proposed, explicitly designed to exploit hemispheric brain symmetry and incorporating inception modules and residual connections. This architecture demonstrated state-of-the-art decoding accuracy, while substantially reducing the prevalence of users facing "BCI inefficiency". Subsequently, the classification performance of *EEGSym* was further strengthened through the application of targeted DA techniques and a multi-dataset TL pipeline. These strategies improved the generalization capabilities across diverse EEG datasets and user populations, significantly enhancing robustness and consistency of MI classification. To address model interpretability, a SHAP-based XAI approach was adapted to EEG data, enabling detailed identification of critical spatio-temporal EEG features influencing the model's predictions. Leveraging these insights, a selection of a more compact yet equally effective electrode montage was proposed. Finally, building upon the developed DL architecture and XAI insights, this Doctoral Thesis demonstrated the feasibility of a model trained solely on ME data could reliably classify MI, pointing to a feasible path for increased reliability of training data for MI-based BCI systems.

This chapter consolidates the main insights, findings, and contributions derived from the research carried out in this Doctoral Thesis. Section 6.1 highlights the original contributions of the compendium of publications included herein. Section 6.2 presents the principal conclusions drawn from the collective interpretation of the results. Finally, Section 6.3 outlines potential lines of research

to extend and enhance the approaches proposed in this Doctoral Thesis.

6.1 Contributions

The core contributions provided by the studies in this doctoral work are summarized below:

1. **Design of a novel deep learning architecture (*EEGSym*) for MI decoding**

   A novel CNN-based model was introduced in Perez-Velasco et al. (2022) to address EEG-based MI classification. The architecture explicitly leverages the brain's hemispheric symmetry, incorporates inception modules to capture multi-scale temporal features, and uses residual connections for deeper and more stable feature learning. This design achieved state-of-the-art performance on multiple public datasets while reducing calibration needs across users.

2. **Development of a DA and TL pipeline**

   In tandem with the proposed architecture, DA techniques and multi-dataset pre-training were implemented to improve inter-subject accuracy in MI-based BCIs. These strategies, applied in Perez-Velasco et al. (2022), mitigated the effect of limited training samples and allowed the model to generalize across five different EEG datasets, elevating the proportion of users surpassing the 70% BCI control threshold.

3. **Adaptation of a SHAP-based XAI framework for EEG signals**

   A major contribution of Pérez-Velasco et al. (2024) was the adaptation of SHAP to raw EEG data. This method elucidated which spatio-temporal signal components were key to the model's classification decisions, thus enhancing interpretability. By highlighting relevant cortical areas, and time windows, the research demonstrated its usability to advance our neurophysiological understanding of EEG-based applications.

4. **Demonstration of direct transfer learning between motor execution and imagery**

   The study in Pérez-Velasco et al. (2025) investigated whether a model trained exclusively on ME data could directly classify MI trials, revealing comparable accuracy to a conventional MI-trained model. This finding expands the

horizon of calibration-free BCI protocols by harnessing objectively verifiable ME data to refine or replace MI calibration sessions.

6.2 Main conclusions

Drawing on the experimental results and discussions in Chapters 4 and 5, as well as the original publications (Perez-Velasco et al., 2022; Pérez-Velasco et al., 2024; Pérez-Velasco et al., 2025), the main conclusions are summarized below:

1. **Advanced deep learning architectures can improve BCI performance**

   By incorporating hemispheric symmetry, inception modules, and residual connections, *EEGSym* significantly boosted MI classification accuracies across large cohorts of users in five public datasets. With sixteen electrodes, *EEGSym* attained the highest mean accuracy in every case, averaging approximately 86.9% and exceeding the strongest competing DL architecture by 1–2 percentage points on each dataset (e.g., 88.6 % versus 87.5 % on PhysioNet; 90.2 % versus 89.4 % on Stieger 2021). With eight electrodes, *EEGSym* continued to outperform all baselines, achieving an average accuracy of about 85.0% and retaining an advantage of 0.8–2.3 percentage points. These results demonstrate that *EEGSym* offers state-of-the-art performance while remaining robust to reduced sets of electrode configuration, a property of practical value for real-world BCI usage.

2. **Data augmentation and multi-dataset transfer learning considerably enhance generalization**

   The integration of perturbation-based data augmentation strategies (e.g., patch perturbation, hemisphere perturbation, random shifts) along with large-scale transfer learning across multiple datasets notably increased the robustness of the DL models. These strategies allowed the developed models to generalize effectively to diverse EEG datasets and recording conditions, substantially improving inter-subject classification performance. More than 95% of subjects achieved above 70% accuracy in fully inter-subject conditions (268 out of 280), thereby reducing the number of users that suffer "BCI-inefficiency".

3. **Explainable AI provides crucial insights into cortical activation beyond classical sensorimotor regions**

SHAP-based feature attribution maps revealed that frontoparietal areas, notably F7, F8, P3, and P4 can be as relevant as sensorimotor sites (C3, C4) for decoding MI. Moreover, significant neural features often emerge in the first second after the cue, indicating early motor planning or motor intention processes.

4. **Optimized electrode montages maintain accuracy while reducing complexity and cost**

   Selecting electrode pairs that consistently displayed high SHAP values (e.g., F7–F8, C3–C4, T7–T8, P3–P4) yielded reliable performance comparable to the original 16-channel setup. This highlights the practicality of reduced montages, facilitating setup times and affordability of BCI systems.

5. **Direct ME→MI TL is feasible and effective**

   Results showed that models trained solely on ME data can classify MI instances with accuracies close to those of MI-trained models. Some participants unable to achieve BCI control under conventional MI-based models succeeded when ME data was used for training. This finding underscores the shared cortical overlap between actual and imagined movements.

Despite the previous conclusions, these promising offline results in healthy volunteers are only a first step. Large-scale studies with clinical populations and real-time feedback are imperative to validate the applicability of these methods in rehabilitation and assistive settings.

6.3 Future research directions

While this Doctoral Thesis has made significant strides in DL for MI-based BCIs, a number of open questions remain, offering opportunities for further exploration:

1. **Expansion to clinical populations and multi-class paradigms**

   Testing *EEGSym* and ME→MI transfer learning with users who have neurological conditions (e.g., stroke or spinal cord injury) would clarify how altered cortical functions affect decoding. For instance, the models could be tested on publicly available EEG datasets from stroke survivors (Liu and Lv, 2022) or individuals with spinal cord injury (Ofner et al., 2019) who perform motor imagery or attempted movements. Additionally, moving

beyond binary paradigms to more complex multi-class tasks can broaden the range of BCI-based applications.

2. **Longitudinal and online evaluations**

   Future work should implement and assess *EEGSym* under real-time closed-loop conditions over extended periods. Examining learning effects, session-to-session variability, and user adaptation would provide a deeper understanding of the method's robustness in realistic scenarios.

3. **Hybrid calibration protocols and advanced feedback strategies**

   Combining short ME runs with MI sessions may expedite calibration, ensuring consistent label quality while encouraging user engagement. Novel feedback modalities (e.g., augmented or virtual reality) could further refine the user's mental simulation of movement and enhance system efficacy.

4. **Integration of alternative DL architectures**

   Beyond CNNs, approaches such as Transformers or self-supervised models may capture complementary spatio-temporal patterns in EEG data. Comparing their feature attribution patterns with SHAP could unveil new insights into neural dynamics during MI.

5. **Advanced neurorehabilitation and user-centered design**

   Combining *EEGSym* and explainable DL methods into broader neurorehabilitation programs warrants further investigation. Tailoring BCI-based exercises to each patient's progress and motor recovery goals, with iterative feedback loops between clinicians and researchers, may propel these systems closer to routine clinical application.

6. **Scalable implementations and edge computing**

   Exploring lightweight model adaptations and deploying these systems on portable devices or edge computing platforms would facilitate real-time, mobile, and home-based BCI applications. Such advances could significantly reduce the need for expensive computing hardware, promoting broader adoption and accessibility.

7. **Cross-modal integration and multimodal BCIs**

   Integrating EEG-based MI decoding with other physiological signals or modalities, such as electromyography (EMG), eye-tracking, or brain

stimulation techniques, could open new avenues for enhancing accuracy and robustness in BCI applications, especially in clinical and assistive contexts.

The work presented here serves as a solid foundation for advancing BCI technologies. By refining the synergy between deep neural architectures, interpretability methods, and ME/MI paradigms, future research can continue to optimize BCI protocols, and improve clinical efficacy, ultimately propelling BCIs from laboratory prototypes to broadly accessible tools that transform the lives of individuals.

Appendix A

Papers included in the compendium of publications

A.1 Contribution 1: Perez-Velasco et al. (2022)

EEGSym: Overcoming Inter-Subject Variability in Motor Imagery Based BCIs With Deep Learning (Perez-Velasco et al., 2022)

Sergio Pérez-Velasco, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Diego Marcos-Martínez, Roberto Hornero. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, p. 1766-1775, 2022. Impact factor in 2022: 4.909, D1 (Q1) in "REHABILITATION" (Journal Citation Reports - Clarivate Analytics).
DOI: 10.1109/tnsre.2022.3186442

In this study, we present a new Deep Learning (DL) architecture for Motor Imagery (MI) based Brain Computer Interfaces (BCIs) called *EEGSym*. Our implementation aims to improve previous state-of-the-art performances on MI classification by overcoming inter-subject variability and reducing BCI inefficiency, which has been estimated to affect 10-50% of the population. This convolutional neural network includes the use of inception modules, residual connections and a design that introduces the symmetry of the brain through the mid-sagittal plane into the network architecture. It is complemented with a data augmentation technique that improves the generalization of the model and with the use of transfer learning across different datasets. We compare *EEGSym*'s performance on inter-subject MI classification with ShallowConvNet, DeepConvNet, EEGNet and EEG-Inception. This comparison is performed on 5 publicly available datasets that include left or right hand motor imagery of 280 subjects. This population is the largest that has been evaluated in similar studies to date. *EEGSym* significantly outperforms the baseline models reaching accuracies of 88.6±9.0 on Physionet, 83.3±9.3 on OpenBMI, 85.1±9.5 on Kaya2018, 87.4±8.0 on Meng2019 and 90.2±6.5 on Stieger2021. At the same time, it allows 95.7% of the tested population (268 out of 280 users) to reach BCI control (≥70% accuracy). Furthermore, these results are achieved using only 16 electrodes of the more than 60 available on some datasets. Our implementation of *EEGSym*, which includes new advances for EEG processing with DL, outperforms previous state-of-the-art approaches on inter-subject MI classification.

A.2 Contribution 2: Pérez-Velasco et al. (2024)

Unraveling motor imagery brain patterns using explainable artificial intelligence based on Shapley values (Pérez-Velasco et al., 2024)

Sergio Pérez-Velasco, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Selene Moreno-Calderón, Roberto Hornero. *Computer Methods and Programs in Biomedicine*, vol. 246, p. 108048, 2024. Impact factor in 2023: 4.949, Q1 in "COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS" (Journal Citation Reports - Clarivate Analytics).
DOI: 10.1016/j.cmpb.2024.108048

Background and objective. Motor imagery (MI) based brain-computer interfaces (BCIs) are widely used in rehabilitation due to the close relationship that exists between MI and motor execution (ME). However, the underlying brain mechanisms of MI remain not well understood. Most MI-BCIs use the sensorimotor rhythms elicited in the primary motor cortex (M1) and somatosensory cortex (S1), which consist of an event-related desynchronization followed by an event-related synchronization. Consequently, this has resulted in systems that only record signals around M1 and S1. However, MI could involve a more complex network including sensory, association, and motor areas. In this study, we hypothesize that the superior accuracies achieved by new deep learning (DL) models applied to MI decoding rely on focusing on a broader MI activation of the brain. Parallel to the success of DL, the field of explainable artificial intelligence (XAI) has seen continuous development to provide explanations for DL networks success. The goal of this study is to use XAI in combination with DL to extract information about MI brain activation patterns from non-invasive electroencephalography (EEG) signals. *Methods.* We applied an adaptation of shapley additive explanations (SHAP) to *EEGSym*, a state-of-the-art DL network with exceptional transfer learning capabilities for inter-subject MI classification. We obtained the SHAP values from two public databases comprising 171 users generating left and right hand MI instances with and without real-time feedback. *Results.* We found that *EEGSym* based most of its prediction on the signal of the frontal electrodes, i.e. F7 and F8, and on the first 1500 ms of the analyzed imagination period. We also found that MI involves a broad network not only based on M1 and S1, but also on the prefrontal cortex (PFC) and the posterior parietal cortex (PPC). We further applied this knowledge to select a 8-electrode configuration that reached inter-subject accuracies of 86.5% ± 10.6% on the Physionet dataset and 88.7%

± 7.0% on the Carnegie Mellon University's dataset. *Conclusion.* Our results demonstrate the potential of combining DL and SHAP-based XAI to unravel the brain network involved in producing MI. Furthermore, SHAP values can optimize the requirements for out-of-laboratory BCI applications involving real users.

A.3 Contribution 3: Pérez-Velasco et al. (2025)

Bridging Motor Execution and Imagery: An Inter-Task Transfer Learning Approach (Pérez-Velasco et al., 2025)

Sergio Pérez-Velasco, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Roberto Hornero. *Biomedical Signal Processing and Control*, vol. 107, p. 107834, 2025. Impact factor in 2023: 4.907, Q1 in "BIOMEDICAL ENGINEERING" (Journal Citation Reports - Clarivate Analytics).
DOI: 10.1016/j.bspc.2025.107834

Motor imagery (MI)-based brain-computer interfaces (BCIs) decode movement imagination from brain activity, but improving decoding accuracy from electroencephalography (EEG) remains challenging. MI-based BCIs require calibration runs to train models; however, participant engagement cannot be externally verified. Motor execution (ME) is more straightforward and can be supervised. Deep learning (DL) leverages transfer learning (TL) to bypass calibration. This is the first work to explore wether a ME-trained DL model can reliably classify MI without finetuning to the MI task, thereby achieving direct TL between ME and MI tasks. We employed *EEGSym*, a DL network for inter-subject TL of EEG decoding, evaluating three scenarios: ME to MI, ME to ME, and MI to MI classification. We analyzed performance correlation between scenarios, and used shapley additive explanations (SHAP) to elucidate model focus patterns learned from ME or MI data. Results show that DL models trained on ME data and tested on MI perform comparably to those trained on MI data. A significant positive correlation was found between performance in ME and MI tasks for models trained on ME data. Explainable artificial intelligence (XAI) techniques revealed robust correlation between patterns in ME and MI tasks. However, between 0.5 to 1 second, the ME-trained model focused on the contralateral central region, while the MI-trained model also targeted the ipsilateral fronto-central region. Our findings demonstrate the viability of inter-task TL between ME and MI using DL models in BCI applications. This supports using ME-trained models for MI tasks to enhance targeted learning of brain activation patterns.

Appendix B

Scientific achievements

B.1 Publications

B.1.1 Papers indexed in the JCR

1. **Sergio Pérez-Velasco**, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Diego Marcos-Martínez, Roberto Hornero, EEGSym: Overcoming Inter/subject Variability in Motor Imagery Based BCIs with Deep Learning, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1766-1775, June, 2022, DOI: 10.1109/TNSRE.2022.3186442

2. **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Selene Moreno-Calderón, Roberto Hornero, Unraveling motor imagery brain patterns using explainable artificial intelligence based on shapley values, *Computer Methods and Programs in Biomedicine*, vol. 246, pp. 108048, April, 2024, DOI: 10.1016/j.cmpb.2024.108048

3. **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Roberto Hornero, Bridging Motor Execution and Motor Imagery BCI paradigms: An Inter-Task Transfer Learning Approach, *Biomedical Signal Processing and Control*, vol. 107, pp. 107834, Septiembre, 2025, DOI: 10.1016/j.bspc.2025.107834

4. Víctor Martínez-Cagigal, Jordy Thielen, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Peter Desain, Roberto Hornero, Brain–computer interfaces based on code-modulated visual evoked potentials (c-VEP): a

literature review, *Journal of Neural Engineering*, vol. 18 (6), pp. 061002, November, 2021, DOI: 10.1088/1741-2552/ac38cf

5. Diego Marcos-Martínez, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Roberto Hornero, Neurofeedback Training Based on Motor Imagery Strategies Increases EEG Complexity in Elderly Population, *Entropy*, vol. 23 (12), pp. 1574, November, 2021, DOI: 10.3390/e23121574

6. Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Robust Asynchronous Control of ERP-Based Brain-Computer Interfaces using Deep Learning, *Computer Methods and Programs in Biomedicine*, vol. 215, March, 2022, DOI: 10.1016/j.cmpb.2022.106623

7. Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Diego Marcos-Martínez, Víctor Rodríguez-González, **Sergio Pérez-Velasco**, Selene Moreno-Calderón, Roberto Hornero, MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research, *Computer Methods and Programs in Biomedicine*, vol. 230, pp. 107357, March, 2023, DOI: 10.1016/j.cmpb.2023.107357

8. Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, **Sergio Pérez-Velasco**, Víctor Rodríguez-González, Ana Martín-Fernández, Selene Moreno-Calderón, Roberto Hornero, ITACA: An open-source framework for Neurofeedback based on Brain-Computer Interfaces, *Computers in Biology and Medicine*, vol. 160, June, 2023, DOI: 10.1016/j.compbiomed.2023.107011

9. Selene Moreno-Calderón, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Combining brain-computer interfaces and multiplayer video games: an application based on c-VEPs, *Frontiers in Human Neuroscience*, vol. 17, August, 2023, DOI: 10.3389/fnhum.2023.1227727

10. Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Selene Moreno-Calderón, Roberto Hornero, Non-binary m-sequences for more comfortable brain–computer interfaces based on c-VEPs, *Expert Systems With Applications*, vol. 232 (120815), December, 2023, DOI: 10.1016/j.eswa.2023.120815

11. Diego Marcos-Martínez, **Sergio Pérez-Velasco**, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, Roberto Hornero, Calibration-Free Ocular Artifact Reduction in EEG signals using a Montage-Independent Deep Learning Model, Biomedical Signal Processing and Control, vol. 110, pp. 108147, Diciembre, 2025, DOI: https://doi.org/10.1016/j.bspc.2025.108147

B.1.2 International conferences

1. **Sergio Pérez-Velasco**, Gonzalo C. Gutiérrez-Tobal, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, Roberto Hornero, Assessment of Residual Deep Neural Networks and AdaBoost to predict adherence to digital-based active and healthy aging interventions, *IUPESM World Congress on Medical Physics and Biomedical Engineering (IUPESM 2022)*, Singapur (Singapore), June 12 - June 17, 2022

2. **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Beatriz Pascual-Roa, Roberto Hornero, Inter-task transfer learning between upper-limb motor execution and motor imagery, *9th Graz BCI Conference (GBCIC 2024)*, pp. 409-413, Graz (Austria), September 9 - September 12, 2024, DOI: 10.3217/978-3-99161-014-4-072

3. Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Víctor Rodríguez-González, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, MEDUSA: A Novel Platform for Modern Non-invasive Brain-computer Interfaces, *IUPESM World Congress on Medical Physics and Biomedical Engineering (IUPESM 2022)*, pp. 124, Singapur (Singapore), June 12 - June 17, 2022

4. Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Selene Moreno-Calderón, Roberto Hornero, Non-binary m-sequences for reliable, high-speed Brain–Computer Interfaces based on c-VEP: a pilot study, *IUPESM World Congress on Medical Physics and Biomedical Engineering (IUPESM 2022)*, Singapur (Singapore), June 12 - June 17, 2022

5. Selene Moreno-Calderón, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Assessing the Potential of Brain-Computer Interface Multiplayer

Video Games using c-VEPs: A Pilot Study, *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)*, ISBN: 979-8-3503-2447-1, Sydney (Australia), July 24 - July 27, 2023

6. Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Selene Moreno-Calderón, Roberto Hornero, Nonparametric Early Stopping Detection for c-VEP-based Brain-Computer Interfaces: A Pilot Study, *45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2023)*, ISBN: 979-8-3503-2447-1, Sydney (Australia), July 24 - July 27, 2023

B.1.3 National conferences

1. **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Selene Moreno-Calderón, Roberto Hornero, Caracterización espacio-temporal de la clasificación de imaginación motora con herramientas de explainable artificial intelligence (XAI), *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 316-319, Valladolid (Spain), November 23 - November 25, 2022

2. **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Rubén Ruiz-Gálvez, Roberto Hornero, Evaluación del Impacto del Aprendizaje Auto-Supervisado en la Precisión de Interfaces Cerebro-Ordenador basadas en Imaginación Motora, *XLI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2023)*, ISBN: 978-84-17853-76-1, pp. 397-400, Cartagena (Spain), November 22 - November 24, 2023

3. Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Sobre la Eficacia del Principio de Vecinos Equivalentes en Sistemas BCI basados en c-VEP, *12º Simposio CEA de Bioingeniería*, pp. 32-36, Madrid (Spain), June 3 - June 4, 2021

4. Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Víctor Martínez-Cagigal, Roberto Hornero, EEG-InceptionGen: Una Red Convolucional de Próposito General para la Clasificación de señales EEG, *XXXIX Congreso*

Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2021), ISBN: 978-84-09-36054-3, pp. 163-166, Madrid (Spain), November 25 - November 26, 2021

5. Selene Moreno-Calderón, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Conecta 4: un Videojuego Multijugador para Sistemas Brain-Computer Interface basados en c-VEPs, *Jornadas de Robótica, Educación en Automática y Bioingeniería 2022 (JREB 2022)*, pp. 246-252, Málaga (Spain), May 18 - May 20, 2022

6. Diego Marcos-Martínez, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, **Sergio Pérez-Velasco**, Selene Moreno-Calderón, Roberto Hornero, ITACA: Un nuevo sistema de entrenamiento cognitivo mediante Neurofeedback, *Jornadas de Robótica, Educación en Automática y Bioingeniería 2022 (JREB 2022)*, Málaga (Spain), May 18 - May 20, 2022

7. Selene Moreno-Calderón, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Roberto Hornero, Evaluación de un videojuego multijugador basado en Brain Computer Interfaces utilizando c-VEPs, *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 324-327, Valladolid (Spain), November 23 - November 25, 2022

8. Laura Gutiérrez-de Pablo, **Sergio Pérez-Velasco**, Víctor Rodríguez-González, Víctor Gutiérrez-de Pablo, Carlos Gómez, Jesús Poza, Aplicación de Deep Learning para el procesado automático de componentes ICA de registros de magnetoencefalografía, *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 316-319, Valladolid (Spain), November 23 - November 25, 2022

9. Ana Martín-Fernández, Diego Marcos-Martínez, Víctor Martínez-Cagigal, **Sergio Pérez-Velasco**, Roberto Hornero, Validación preliminar de ITACA: Un entorno novedoso para estudios de Neurofeedback, *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 125-128, Valladolid (Spain), November 23 - November 25, 2022

10. Diego Marcos-Martínez, Ana Martín-Fernández, **Sergio Pérez-Velasco**, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Selene Moreno-

Calderón, Roberto Hornero, Análisis de los cambios en la conectividad funcional tras un entrenamiento cognitivo mediante Neurofeedback, *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 27-30, Valladolid (Spain), November 23 - November 25, 2022

11. Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, **Sergio Pérez-Velasco**, Diego Marcos-Martínez, Selene Moreno-Calderón, Roberto Hornero, Un nuevo método de parada temprana no paramétrico para sistemas Brain–Computer Interface basados en c-VEP, *XL Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2022)*, ISBN: 978-84-09-45972-8, pp. 196-199, Valladolid (Spain), November 23 - November 25, 2022

12. Amalia Gil-Correa, **Sergio Pérez-Velasco**, Víctor Rodríguez-González, Hideyuki Hoshi, Yoshihito Shigihara, Carlos Gómez, Jesús Poza, Detector automático de artefactos en señales neuronales basado en técnicas de Inteligencia Artificial, *XLI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2023)*, ISBN: 978-84-17853-76-1, pp. 654-657, Cartagena (Spain), November 22 - November 24, 2023

13. Diego Marcos-Martínez, Víctor Rodríguez-González, **Sergio Pérez-Velasco**, Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Roberto Hornero, Influencia de los sistemas Brain-Computer Interface basados en Neurofeedback en las características de la red cerebral, *XLI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2023)*, ISBN: 978-84-17853-76-1, pp. 500-503, Cartagena (Spain), November 22 - November 24, 2023

14. Rubén Ruiz-Gálvez, Eduardo Santamaría-Vázquez, Diego Marcos-Martínez, **Sergio Pérez-Velasco**, Beatriz Pascual-Roa, Roberto Hornero, Identificación de biomarcadores de rendimiento cognitivo a partir del análisis del EEG basal en personas entre 65 y 75 años, *XLII edición del Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2024)*, Sevilla (Spain), November 13 - November 15, 2024

15. Beatriz Pascual-Roa, Eduardo Santamaría-Vázquez, Diego Marcos-Martínez, **Sergio Pérez-Velasco**, Rubén Ruiz-Gálvez, Roberto Hornero, Estimación de la carga cognitiva durante la realización de tareas de memoria

de trabajo mediante la señal de EEG, *XLII edición del Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2024)*, Sevilla (Spain), November 13 - November 15, 2024

B.2 International internship

A three-month stay at the Donders Institute for Brain, Cognition and Behaviour in Radboud University, Nijmegen, Netherlands.

- **Purpose of the internship**

  The main goal of the stay was to deepen collaboration with this group and to research the application of DL algorithms to enhance the identification of user intentions in BCI systems. Three specific objectives were set:

  1. Learn how the Joint Embedding Predictive Architectures (JEPA) model for self-supervised learning was adapted to EEG signals and investigate the state-of-the-art techniques to obtain insights into the learned representations.
  2. Incorporate XAI methods to interpret the decisions made by DL models. The focus is on using SHAP to generate feature attribution maps that highlight the most influential signal regions and explain the importance of intermediate representations.
  3. Participate in the writing of joint articles, as well as in the request for projects that allow the consolidation of the bilateral relationship established between the Biomedical Engineering Group (GIB) of the University of Valladolid and CIBER BBN and Donders Institute for Brain, Cognition and Behaviour at Radboud University.

  More specifically, the tasks that were carried out are defined below:

  1. Learned the training processes of JEPA models applied to EEG, focusing on extracting various learned representations from the contextual encoder and predictor components within the architecture.
  2. Investigated current techniques applied to other JEPA architectures to gain insights into the condensed representations learned from models trained on images and videos. One common approach identified involves using diffusion models to generate EEG signal from representations of masked patches, and identify which information is retained. Although

this research shows promising prospects, it has not yet yielded satisfactory results and will be further pursued.

3. Adapted XAI techniques initially developed for TensorFlow models to be compatible with PyTorch models. This included extending the functionality to explain the significance of rpresentations generated by intermediate layers, enhancing the interpretability of DL models applied to BCI systems. This methodology can be applied to downstream models build over the JEPA architecture on classification tasks or other DL models trained from scratch.

Moreover, this stay is expected to promote a research collaboration between the Biomedical Engineering Group of the University of Valladolid and the Donders Institute for Brain, Cognition and Behaviour of Radboud University. By establishing this new bond between both institutions, the quality of the research is meant to increase, fostering a productive exchange of ideas and potential joint research initiatives.

- **Quality indicators of the institution**

  The Data-Driven NeuroTechnology Lab under the supervision of Prof. Dr. Michael Tangermann, who serves as the lab director, is part of the Donders Institute for Brain, Cognition and Behaviour of Radboud University, within the Department of Artificial Intelligence. The research focus of the lab is on advancing neurotechnology through the use of ML methods specialized in the analysis of brain signals. These methods not only facilitate the real-time decoding of brain states but also enable the active modulation of these states.

  The application of these innovative methods allows for an in-depth exploration of the brain's perceptual, motor, and cognitive functions in both healthy conditions and pathological states. Additionally, the lab is dedicated to the technological transfer of these ML models toward the development of cutting-edge BCI applications aimed at improving the quality of life for both patients and healthy users.

  The lab's team, composed of eight outstanding professionals—all holding MD or PhD degrees—is at the forefront of applying machine learning techniques to BCI systems and the interpretation of brain signals. In recent years, they have focused their efforts on exploring various BCI paradigms, such as event-related potentials, code-modulated visual evoked

potentials, and motor imagery. They also apply advanced DL techniques to enhance the analysis of these phenomena and to eliminate artifacts in EEG signals. Therefore, this research environment, which combines advanced knowledge and innovative methodological approaches, aligns perfectly with the objectives of the doctoral thesis.

Prof. Dr. Michael Tangermann, the supervisor of the stay, holds the position of director of the Data-Driven NeuroTechnology Lab and is an associate professor at Radboud University. His curriculum includes 18 scientific publications in high-impact journals in the fields of neuroscience and ML over the past five years, with an h-index of 30. Before his current position, Prof. Tangermann was a professor at the University of Freiburg, where he led the Brain State Decoding Lab in the Department of Computer Science.

B.3 Awards and honors

- **3rd place in the BR41N.IO Brain-Computer Interface Designers' Hackathon** at BCI & Neurotechnology Spring School 2021 organized by g.Tec and IEEE Brain for the project "Towards P300 calibration-less single-trial classification".

- **3rd place in the IFMBE SC Competition 2022**. For the project presented at the IFMBE SCIENTIFIC CHALLENGE COMPETITION in conjunction with the World Congress on Medical Physics and Biomedical Engineering:

  Sergio Pérez-Velasco, Gonzalo C. Gutiérrez-Tobal, Víctor Martínez-Cagigal, Eduardo Santamaría-Vázquez, Roberto Hornero, "Assessment of Residual Deep Neural Networks and AdaBoost to predict adherence to digital-based active and healthy aging interventions", *IUPESM 2022*, Singapur (Singapore), June 12 - June 17, 2022

- **Finalist of the Cajamar UniversityHack 2022** in the category "Cajamar Water Footprint" after winning the local phase of this national competition held between the 18th of February and the 3rd of May in 2022.

Apéndice C

Resumen en español

C.1 Introducción

El afán de lograr la "telepatía", conexión directa entre cerebros, llevó a Hans Berger a registrar por primera vez la actividad eléctrica cerebral, hallazgo que marcó el nacimiento de la electroencefalografía (EEG) y sentó las bases de la neurociencia moderna. Desde entonces, la EEG se ha convertido en una herramienta esencial para estudiar de forma no invasiva los patrones eléctricos subyacentes a la cognición, la percepción y el control motor. Su capacidad para desvelar mecanismos neurales ha sido decisiva en el desarrollo de los *brain–computer interfaces* (BCI). Los sistemas BCI establecen un canal de comunicación directo entre la actividad cerebral y dispositivos externos, lo cual elimina la necesidad de vías musculares y abre nuevas posibilidades de comunicación, control y rehabilitación para personas con limitaciones motoras.

En la práctica, los sistemas BCI se apoyan mayoritariamente en técnicas no invasivas porque combinan seguridad y accesibilidad. Métodos como la EEG, la magnetoencefalografía (MEG) o la espectroscopía de infrarrojo cercano funcional (fNIRS) registran la actividad cerebral desde el exterior del cráneo, cada uno con sus propios compromisos en resolución temporal y espacial. Aunque fNIRS ofrece una visión hemodinámica razonablemente detallada, su respuesta es más lenta que la de EEG o MEG; aun así, todas estas técnicas capturan información suficiente para descifrar las intenciones del usuario y resultan idóneas cuando la portabilidad y la seguridad son prioritarias.

Dentro de este conjunto, la EEG destaca como la opción más extendida en investigación y desarrollo de BCIs, gracias a su bajo coste, la sencillez del

equipamiento y su excelente resolución temporal. A diferencia de la MEG, que exige dispositivos voluminosos y costosos, la EEG puede implementarse con hardware relativamente económico y sin infraestructura especializada, lo que facilita su adopción tanto en entornos clínicos como en aplicaciones de consumo.

Un sistema BCI basado en EEG capta primero la actividad eléctirca cerebral mediante electrodos colocados según sistemas estandarizados (10–20, 10–10, 10–5); después la procesa con técnicas de filtrado, extracción de características y clasificación; y, finalmente, traduce la salida del clasificador en acciones que se retroalimentan al usuario en tiempo real para que aprenda a modular su propia actividad neuronal.

A pesar de sus numerosas ventajas, la correcta clasificación de señales EEG es muy complicada por su baja relación señal–ruido, su resolución espacial limitada y la variabilidad inter-sujeto e inter-sesión. Por ello, los BCIs recurren a **paradigmas** que inducen **señales de control** en el EEG, lo suficientemente robustas como para distinguirse del ruido y reflejar de manera fiable las intenciones del usuario. Entre las señales de control más empleadas en los BCIs basados en EEG se encuentran los potenciales relacionados con eventos (*event-related potentials*, ERPs). Dentro de esta categoría figuran los potenciales evocados visuales (*visual evoked potentials*, VEP), los potenciales evocados visuales de estado estacionario (*steady-state visual evoked potentials*, SSVEP), los potenciales evocados visuales modulados por código (*code-modulated visual evoked potentials*, c-VEP), el potencial P300 y los potenciales relacionados con los errores (*error-related potentials*, ErrPs). Completan el conjunto los potenciales corticales lentos (*slow cortical potentials*, SCPs) y los ritmos sensorimotores (*sensorimotor rhythms*, SMR).

Cada tipo de señal presenta rasgos propios que determinan su idoneidad en función del objetivo del BCI. Los sistemas basados en la la imaginación motora (*motor imagery*, MI), por ejemplo, ofrecen un control muy natural y un gran potencial rehabilitador, aunque suelen exigir un período considerable de entrenamiento al usuario. En cambio, los ERPs como el P300 o los c-VEPs se desencadenan de manera automática ante estímulos externos y prácticamente no requieren aprendizaje previo, lo que los convierte en la opción preferente para interfaces de comunicación.

Esta Tesis Doctoral se centra en el desarrollo de técnicas de *deep learning* (DL) para optimizar sistemas BCI basados en la MI registrados con EEG. Entre los distintos paradigmas, como se ha mencionado anteriormente, la imaginación de un movimiento sin ejecutarlo resulta especialmente atractiva por sus aplicaciones en la rehabilitación motora. La señal de control producida por este paradigma

que generalmente se usa para detectar las intenciones del usuarios son los ritmos sensoriomotores (*sensorimotor rhythms*, SMR) generados en la corteza motoral. No obstante, otras áreas cerebrales se ven implicadas en la imaginación, como la zona prefrontal implicada en la planificación del movimiento. Pese a que el paradigma MI resulta idóneo para promover la plasticidad cerebral asociada a la ejecución motora (*motor execution*, ME), numerosos usuarios no alcanzan niveles de clasificación lo suficientemente altos para considerar que controlan el sistema (por encima de 70 % de precisión), lo que se denomina *ineficiencia BCI*. Para superarla, esta Tesis Doctoral introduce nuevas estrategias de DL e inteligencia artificial explicable (*explainable artificial intelligence*, XAI), para adquirir nuevos conodimientos sobre este paradigma que van más allá de los métodos clásicos de aprendizaje automático basados en características seleccionadas manualmente.

El núcleo de este trabajo se materializa en tres estudios encadenados y reflejados en un compendio de artículos. En primer lugar, *EEGSym* (Perez-Velasco et al., 2022) se presenta como una arquitectura novedosa de DL para clasificar tareas de MI con una alta precisión inter-sujeto. Este modelo emplea módulos *inception*, conexiones residuales y una disposición simétrica inspirada en la anatomía cerebral, junto con un método de aumento de datos y transferencia del aprendizaje (*transfer learning*, TL) a partir de cinco bases de datos de EEG públicas. Ello resulta en una mejora de la robustez en la clasificación, logrando que la mayoría de usuarios (268 de 280, 95,7 %) alcancen el umbral necesario para controlar el sistema BCI.

En segundo lugar, para entender las bases fisiológicas que utilizan estas redes profundas, se desarrolla y aplica un método de explicabilidad basado en valores (*shapley additive explantions, SHAP*). El análisis realizado en Pérez-Velasco et al. (2024) revela que el modelo *EEGSym* no se limita a la señal de la corteza motora primaria, sino que incorpora la información de regiones frontales y parietales en la clasificación de las intenciones del usuario a partir del EEG. Estos hallazgos contribuyen a optimizar la colocación de electrodos y pueden ayudar a comprender mejor el funcionamiento del cerebro durante la MI. Confirman la relevancia de zonas cerebrales no siempre contempladas en las configuraciones de BCIs usadas normalmente para este paradigma.

El tercer estudio profundiza en la relación entre ME y MI, evaluando la posibilidad de emplear modelos entrenados con datos de ME para clasificar tareas de MI (Pérez-Velasco et al., 2025). Sorprendentemente, el entrenamiento exclusivo en ME logra resultados comparables a los de un entrenamiento específico en MI, lo que sugiere que ambas modalidades comparten patrones neuronales

relevantes. Este enfoque podría mejorar las sesiones de calibración en usuarios con movimiento, ya que el usuario puede realizar tareas de ejecución motora verificables antes de la transición a la MI, beneficiando la motivación y la adaptabilidad del sistema. También abre nuevas vías de rehabilitación, en las que el mismo modelo puede empezar clasificando MI para posteriormente clasificar el ME según se vaya recuperando la movilidad.

Desde un punto de vista más amplio, esta Tesis Doctoral aporta: (1) un modelo DL robusto para la clasificación de MI, superando la variabilidad entre sujetos (Perez-Velasco et al., 2022); (2) un marco de explicabilidad que arroja luz sobre redes cerebrales implicadas en las predicciones del modelo (Pérez-Velasco et al., 2024); y (3) una estrategia de TL entre ME y MI que simplifica la recolección de datos de entrenamiento y puede aportar una via de rehabilitación alternativa a las actuales (Pérez-Velasco et al., 2025). Estos hallazgos proveen de una base para la adopción de BCIs basados en EEG más versátiles y transparentes, con gran potencial en ámbitos clínicos y aplicaciones de neurorehabilitación.

C.2 Hipótesis y objetivos

En el contexto de los sistemas BCI, la clasificación de la MI a partir de señales de EEG constituye un reto debido a la baja resolución espacial, la escasa relación señal–ruido y la variabilidad entre sujetos. A pesar de los avances logrados en paradigmas que aprovechan los SMR para controlar dispositivos, su precisión sigue siendo insuficiente para su aplicación amplia más allá del entorno de laboratorio. Con el fin de mejorar estos sistemas, esta Tesis Doctoral propone el diseño, desarrollo y validación de nuevas metodologías basadas en DL, así como su complementación con técnicas de inteligencia artificial explicable XAI. Estas últimas permiten revelar cómo los modelos de DL toman sus decisiones, aportando información fisiológica y facilitando la interpretación de patrones neuronales relevantes. Así, se formulan tres hipótesis principales: (1) la incorporación de arquitecturas de DL que combinen simetría cerebral, módulos de tipo *inception* y conexiones residuales podría mejorar sustancialmente la precisión de la clasificación de MI, reduciendo la dependencia de calibraciones específicas para cada sujeto; (2) la aplicación de métodos de XAI, concretamente técnicas derivadas de los valores SHAP, podría ofrecer interpretaciones fisiológicas de los patrones cerebrales subyacentes, aclarando el rol de distintas regiones cerebrales en la tarea de MI; y (3) el uso de TL entre ME y MI mediante redes de DL permitiría entrenar modelos sobre datos de ME que funcionen directamente para

la clasificación de MI, minimizando las largas sesiones de calibración o la necesidad de un ajuste fino adicional. Para validar estas hipótesis, los objetivos planteados incluyen la creación de una arquitectura novedosa de DL denominada *EEGSym*, el desarrollo de técnicas de aumento de datos y transferencia entre múltiples conjuntos de EEG, la adaptación de *SHAP* para identificar las características más influyentes en la toma de decisiones del modelo, y la evaluación de la factibilidad de entrenar redes con datos de ME para su posterior aplicación en MI.

Estas hipótesis específicas respaldan la hipótesis global de esta Tesis Doctoral: *"Los avances más recientes en DL, combinados con métodos de XAI y el aprovechamiento de la relación entre la ME y la MI, pueden mejorar los sistemas BCI basados en EEG al aumentar la precisión de la decodificación de MI, profundizar en nuestra comprensión del funcionamiento cerebral y superar las limitaciones asociadas a su clasificación."*

El objetivo general de esta Tesis Doctoral es diseñar, desarrollar y evaluar nuevas metodologías basadas en DL que mejoren la decodificación de la MI en BCIs basados en EEG y, al mismo tiempo, aporten interpretaciones fisiológicas de los procesos neurales implicados. Para alcanzarlo se formulan los siguientes objetivos específicos:

O1: Crear e implementar la arquitectura *EEGSym*, que integre simetría cerebral, módulos *inception* y conexiones residuales. Aprovechando también técnicas de aumento de datos y transferencia del aprendizaje entre datos, con el fin de aumentar la precisión y reducir la ineficacia de los BCIs.

O2: Desarrollar y aplicar métodos de inteligencia artificial explicable, en particular una adaptación de los valores SHAP, para identificar los patrones y regiones cerebrales más relevantes, optimizando así la colocación de electrodos sin sacrificar rendimiento.

O3: Profundizar en la comprensión de los mecanismos neuronales de la MI y de la ME, evaluando la capacidad de TL de modelos entrenados con ME a tareas de MI para mejorar las terapias de rehabilitación y las tecnologías asistivas.

O4: Validar las metodologías propuestas en grandes bases de datos públicas y heterogéneas, demostrando su eficacia, capacidad de generalización e impacto potencial.

C.3 Materiales y métodos

Este trabajo emplea cinco bases de datos públicas de EEG registradas durante MI y, en un caso, ME, con el fin de desarrollar y evaluar metodologías de DL para la clasificación de MI (Goldberger et al., 2000; Kaya et al., 2018; Lee et al., 2019; Meng and He, 2019; Stieger et al., 2021). Cada base de datos presenta un protocolo experimental similar, en el que los sujetos imaginan movimientos de la mano izquierda o derecha siguiendo indicaciones visuales y descansan entre ensayos. Varias difieren, no obstante, en la duración de la tarea, la presencia de realimentación y el número de sesiones. Para unificar el formato y simplificar el entrenamiento de modelos, se extrajeron ventanas de 3 segundos inmediatamente posteriores a la señal de inicio de la tarea, re-muestreadas a 128 Hz y normalizadas electrodo a electrodo.

Los registros de EEG se preprocesan con el fin de atenuar interferencias y ruidos: se aplican filtros IIR de cuarto orden para eliminar la frecuencia de línea (50/60 Hz) y eliminar el aliasing por el re-muestreado a 128 Hz (filtrando por encima de 63 Hz antes del re-muestreo). Se realiza también un re-referenciado mediante referencia media común (*common average reference*, CAR), restando de cada canal la media global de todos los canales extraidos en cada configuración. Para garantizar consistencia, se eligen subconjuntos comunes de electrodos en cada base de datos; en Perez-Velasco et al. (2022) se emplean configuraciones de 8 o 16 canales comunes a las cinco bases de datos, en Pérez-Velasco et al. (2024) se emplea la misma configuración de 16 canales para dos bases de datos (Goldberger et al., 2000; Stieger et al., 2021), mientras que en Pérez-Velasco et al. (2025) se utilizan los 64 canales disponibles en la única base de datos que incluye ME y MI (Goldberger et al., 2000).

El desarrollo del modelo de DL se apoya en arquitecturas convolucionales (*convolutional neural networks*, CNNs) previamente propuestas para EEG, como ShallowConvNet, DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018) y EEG-Inception (Santamaria-Vazquez et al., 2020), las cuales sirven de referencia para medir las mejoras de clasificación que la red propuesta aporta. En Perez-Velasco et al. (2022) se presenta *EEGSym*, una red neuronal que integra: (1) divisiones hemisféricas (simetría cerebral) para capturar características relacionadas con la lateralización de MI; (2) módulos *inception*; y (3) conexiones residuales para poder incluir más capas sin presentar problemas de gradientes que desaparecen que impedirían el entrenamiento. Además, *EEGSym* utiliza *batch normalization*, activaciones ELU y regularización por *dropout* en cada

etapa convolucional, combinando estrategias previamente exitosas en visión por computadora y análisis de EEG.

La estrategia de entrenamiento consiste en dos fases principales: (1) un pre-entrenamiento en varias bases de datos, aplicando una tasa de aprendizaje alta y técnicas de aumento de datos (*data augmentation*, DA); y (2) una adaptación fina o *fine-tuning* al conjunto objetivo mediante validación cruzada *leave-one-subject-out* (LOSO). En esta última fase, cada sujeto se reserva para conjunto de test, mientras que el resto contribuye al entrenamiento y a la validación. La técnica de DA incluye perturbaciones temporales y espaciales (como *patch perturbation*, alteraciones aleatorias en un hemisferio o desplazamientos en la ventana de inicio) diseñadas para dar robustez al modelo frente a artefactos y variaciones en el tiempo de reacción de los sujetos. La métrica de evaluación es la exactitud (%), a la que se aplican pruebas estadísticas (Wilcoxon) (Wilcoxon, 1945) y corrección por tasa de descubrimientos falsos (*false discovery rate*, FDR) (Benjamini and Hochberg, 1995) para confirmar diferencias significativas con las precisiones de las redes comparadas.

En Pérez-Velasco et al. (2024), para dotar de interpretabilidad a las redes, se adopta un enfoque de XAI basado en SHAP (Lundberg and Lee, 2017). Se calcula la contribución de cada componente temporal y espacial del EEG respecto a la clasificación realizada por la red, usando una adaptación del gradientes integrados (`GradientExplainer`) (Sundararajan et al., 2017). Además, se propone una selección de canales basada en SHAP para reducir el número de electrodos manteniendo altos niveles de exactitud, lo que favorece aplicaciones prácticas de BCI.

Finalmente, en Pérez-Velasco et al. (2025) se investiga la relación entre ME y MI. Se entrena *EEGSym* exclusivamente con datos de ejecución motora (ME) y se analiza su rendimiento al clasificar señales de MI (sin *fine-tuning*), utilizando el conjunto Physionet (Goldberger et al., 2000) que incluye ambas condiciones. Se exploran tres escenarios (ME→ME, MI→MI, y ME→MI) en un esquema LOSO. Además de la exactitud global, se examinan matrices de confusión y se calcula la correlación entre desempeños para cada individuo, evaluando así la estabilidad de patrones compartidos entre ejecución e imaginación. La interpretación de los mapas de activación mediante SHAP contribuye a comparar las fuentes neuronales implicadas en ambos paradigmas. Dicho análisis sugiere que modelos entrenados con ME pueden mantener una clasificación competitiva en MI, aportando evidencias en favor de la transferencia de conocimientos entre tareas motoras y abriendo vías para reducir las sesiones de calibración o mejorar

la motivación del usuario.

En conjunto, las metodologías presentadas, incluyendo la estandarización de bases de datos heterogéneas, el preprocesamiento unificado, el diseño de la red *EEGSym*, la aplicación de DA e interpretabilidad con SHAP, y el estudio de la relación ME→MI, sientan las bases para sistemas BCI más precisos, sin la necesidad de sesiones de calibración diarias, y con mayor potencial de aplicabilidad real.

C.4 Resultados y discusión

Este trabajo ha mostrado, en diversos niveles, las ventajas de emplear enfoques de DL en la clasificación de tareas de MI y ME en sistemas BCI. Los resultados se han distribuido en tres líneas principales: (1) diseño y rendimiento de una arquitectura de DL que mejora la clasificación en MI; (2) aplicación de técnicas de XAI para entender cómo el modelo toma sus decisiones y entender mejor el funcionamiento cerebral durante este paradigma; y (3) exploración del potencial de transferencia directa de modelos entrenados con ME para su uso en MI. A continuación, se sintetizan los hallazgos y sus implicaciones.

En primer lugar, la arquitectura *EEGSym* (Perez-Velasco et al., 2022), diseñada para explotar la simetría cerebral, los módulos *inception* y las conexiones residuales, demostró un rendimiento superior al de arquitecturas convencionales como *ShallowConvNet*, *DeepConvNet* (Schirrmeister et al., 2017), *EEGNet* (Lawhern et al., 2018) o *EEG-Inception* (Santamaria-Vazquez et al., 2020). Al evaluar en cinco bases de datos públicas de EEG (Goldberger et al., 2000; Kaya et al., 2018; Lee et al., 2019; Meng and He, 2019; Stieger et al., 2021), *EEGSym* alcanzó de forma consistente tasas de exactitud entre el 80 % y 90 % (superando con significancia estadística a los modelos comparados y permitió que más del 95 % de la población analizada (268 de 280 sujetos) superase el umbral de 70 % de exactitud requerido para estimar que se tiene control BCI. Además, estos resultados se alcanzaron usando sólo 8 o 16 electrodos, mostrando la factibilidad de configuraciones simplificadas para aplicaciones reales que son más cómodas para el usuario y más rápidas de preparar.

Los componentes clave que explican este éxito incluyen: (1) la división de entradas acorde con la simetría cerebral para aumentar la eficiencia de los parámetros de *EEGSym*; (2) el uso de conexiones residuales que facilitan la propagación de características a través todas las capas de la red; y (3) un entrenamiento secuencial (pre-entrenamiento con múltiples bases de datos de EEG,

seguido de un ajuste fino en la base dde datos del sujeto a probar) que aprovecha la capacidad de TL del DL (Goodfellow et al., 2016). Asimismo, se incorporó un DA específico (p. ej., perturbaciones en un hemisferio, desplazamientos temporales) para robustecer el modelo frente a variaciones entre sujetos y artefactos típicos del EEG.

En segundo lugar, para comprender mejor en qué se fijaba *EEGSym* para realizar estas predicciones, se introdujo un análisis XAI basado en SHAP Pérez-Velasco et al. (2024). Este método permitió visualizar qué canales y qué momentos temporales son más relevantes en la decisión de la red, y evidenció que la MI implica una activación distribuida en el cerebro detectable con EEG: no sólo en las zonas sensoriomotoras, sino también en áreas prefrontales y parietales, aumentando su importancia cuando existe realimentación en tiempo real. Los resultados mostraron que la ventana inicial (0-1 s tras la señal de inicio) es la que más contribuye a la clasificación, que además se fija a la etapa de planificación motora antes de la imaginación mantenida en la zona central. Concretamente, los electrodos F7, F8, C3, C4, T7, T8, P3 y P4 fueron los que mayor relevancia presentaron, lo que condujo a una propuesta de configuración óptima de 8 canales diferentes a la configuración usada en Perez-Velasco et al. (2022). En comparación con esa configuración, esta nueva selección basada en SHAP logró mejoras significativas en la exactitud en Physionet, demostrando la validez de la metodología de XAI presentada para mejorar la interpretabilidad de *EEGSym*.

Por último, se investigó la relación entre la ME y la MI, evaluando si un modelo entrenado exclusivamente con datos de ME puede clasificar de forma directa los ensayos de MI sin ajuste fino adicional (Pérez-Velasco et al., 2025). Los experimentos mostraron que los modelos entrenados en ME alcanzan tasas de exactitud comparables (86-87 %) a las de un modelo entrenado solo en MI, evidenciando que ME y MI comparten patrones neuronales registrados por la EEG suficientes para que TL sea factible entre tareas. Además, el modelo ME→MI mejoró la clasifcación en varios sujetos que inicialmente no habían superado el umbral de control BCI con un modelo entrenado solo en MI. Desde una perspectiva práctica, este hallazgo apunta a sesiones de calibración menos tediosas, ya que verificar la ejecución motora es más sencillo y reduce la incertidumbre sobre la correcta realización de la tarea. En rehabilitación, a medida que el usuario se recupera o adquiere mayor destreza en la MI, el modelo puede ajustarse finamente (o mantenerse tal cual si ya ofrece un rendimiento óptimo) para adaptarse a los cambios en la actividad neuronal del movimiento imaginado. Según el sujeto recupera movilidad, el mismo modelo puede clasificar la actividad neuronal del

incipiente ME recuperado. De este modo, el sistema acompaña las distintas etapas de la terapia, integrándose de forma progresiva y personalizada en el proceso de rehabilitación.

En síntesis, este trabajo respalda la idea de que las técnicas de DL, adecuadamente diseñadas (*EEGSym*) y explicadas (SHAP), pueden mejorar de forma sustancial la exactitud de clasificación de la imaginación motora, reducir el número de electrodos necesarios y ofrecer interpretaciones fisiológicamente coherentes sobre la actividad cerebral implicada. Además, la transferencia directa desde ME a MI abre la puerta a calibraciones más ágiles y puede favorecer la adopción de BCIs en contextos clínicos y de rehabilitación. Futuras investigaciones podrían abordar la evaluación con poblaciones clínicas, estudios longitudinales que analicen la adaptación del usuario al sistema o el uso de arquitecturas de vanguardia (por ejemplo, basadas en *transformers*) para seguir impulsando la usabilidad y eficacia de los BCIs basados en EEG.

C.5 Conclusiones

El objetivo central de esta Tesis Doctoral ha sido el desarrollo e integración de técnicas avanzadas de DL y XAI para mejorar los BCIs basados en MI. En primer lugar, se propuso la arquitectura *EEGSym*, diseñada explícitamente para aprovechar la simetría hemisférica del cerebro y que incorpora módulos *inception* y conexiones residuales. Dicha arquitectura mostró un incremento notable en la precisión de clasificación, reduciendo la prevalencia de usuarios que no alcanzaban el control en BCI (*BCI inefficiency*). A continuación, se mejoró aún más el rendimiento de *EEGSym* mediante estrategias de DA específicas y un protocolo de TL entre varias bases de datos, mejorando la capacidad de generalización frente a distintos usuarios y condiciones experimentales.

Para dotar de interpretabilidad a estos modelos, se adaptó un método basado en SHAP (Lundberg and Lee, 2017), permitiendo identificar con detalle los componentes espacio-temporales del EEG que influyen en las decisiones de la red. A partir de estos análisis, se propuso un montaje de electrodos más reducido que mantiene la eficacia del modelo, lo que aporta ventajas prácticas (menos costes de equipos y tiempo de preparación). Por último, se demostró la viabilidad de entrenar el modelo exclusivamente con señales de ME para aplicarlo directamente en clasificación de MI, abriendo la posibilidad de mejorar la fase de calibración específica de MI, al contar con la recogida de datos de entrenamiento externamente verificables.

A continuación, se sintetizan los aportes más destacados de la Tesis Doctoral y sus conclusiones principales:

- **Las arquitecturas avanzadas de aprendizaje profundo elevan el rendimiento de los BCIs**

  Al incorporar simetría hemisférica, módulos *inception* y conexiones residuales, *EEGSym* incrementó de forma notable la precisión en la clasificación de MI en cinco bases de datos públicas con grandes cohortes de usuarios. Con dieciséis electrodos alcanzó la mayor exactitud media en todos los casos (aprox. 86,9 %), superando en 1-2 puntos porcentuales a la mejor red comparada (p. ej. 88,6 % frente a 87,5 % en PhysioNet; 90,2 % frente a 89,4 % en Stieger 2021). Con solo ocho electrodos mantuvo estas mejoras con respecto al resto de redes comparadas (media cercana al 85,0 %).

- **La ampliación de datos y el aprendizaje por transferencia multidataset refuerzan la generalización**

  La combinación de técnicas de DA basadas en perturbaciones (desplazamientos aleatorios, alteraciones por hemisferios o por parches), junto con el TL a gran escala entre múltiples bases de datos, aumentó considerablemente la robustez de los modelos. Más del 95 % de los sujetos (268 de 280) superó el 70 % de acierto en condiciones totalmente inter-sujeto, reduciendo la incidencia de la denominada ineficacia BCI.

- **La IA explicable revela activaciones corticales más allá de las regiones sensorimotores clásicas**

  Los mapas de atribución basados en SHAP mostraron que zonas frontoparietales (F7, F8, P3 y P4) pueden ser tan relevantes como los sitios sensorimotores C3 y C4 para descodificar la MI. Además, muchas características neuronales significativas emergen durante el primer segundo tras la instrucción, lo que apunta a procesos tempranos de planificación o intención motora.

- **Montajes de electrodos optimizados mantienen la precisión y reducen complejidad y coste**

  Seleccionar pares de electrodos con valores SHAP elevados (F7–F8, C3–C4, T7–T8, P3–P4) proporciona un rendimiento comparable al montaje original de 16 canales, lo que simplifica la preparación del sistema y abarata su implementación sin sacrificar exactitud.

- **La transferencia directa de ME a MI es viable y eficaz**

  Los modelos entrenados exclusivamente con datos de ME clasificaron instancias de MI con precisiones muy cercanas a las de los modelos entrenados directamente con MI. Algunos participantes que no alcanzaban control BCI con enfoques MI tradicionales lograron hacerlo cuando se empleó ME para el entrenamiento, lo que subraya el solapamiento cortical entre movimiento real e imaginado.

En conjunto, las investigaciones realizadas sientan las bases para BCIs basados en EEG más eficientes, explicables y adaptables. El aprovechamiento simultáneo de técnicas avanzadas de DL, métodos de explicabilidad y la fusión ME–MI posibilitan avanzar hacia BCIs más prácticas y versátiles, abriendo nuevas perspectivas tanto en el ámbito clínico como en aplicaciones de consumo.

Bibliography

Acqualagna, L., Botrel, L., Vidaurre, C., et al., 2016. Large-Scale Assessment of a Fully Automatic Co-Adaptive Motor Imagery-Based Brain Computer Interface. PLOS ONE 11 (2), e0148886.

Adadi, A. and Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160.

Alkoby, O., Abu-Rmileh, A., Shriki, O., and Todder, D., 2018. Can We Predict Who Will Respond to Neurofeedback? A Review of the Inefficacy Problem and Existing Predictors for Successful EEG Neurofeedback Learning. Neuroscience 378 (January), 155–164.

Alsuradi, H., Park, W., and Eid, M., 2020. Explainable Classification of EEG Data for an Active Touch Task Using Shapley Values. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 12424 LNCS. pp. 406–416.

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C., 2008. Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface. Proceedings of the International Joint Conference on Neural Networks, 2390–2397.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57 (1), 289–300.

Berger, H., 1929. Über das Elektrenkephalogramm des Menschen. Archiv für Psychiatrie und Nervenkrankheiten 87 (1), 527–570.

Binder, A., Montavon, G., Bach, S., et al., 2016. Layer-wise relevance propagation for neural networks with local renormalization layers.

Block, A., Mroueh, Y., and Rakhlin, A., 2022. Generative modeling with denoising auto-encoders and langevin sampling.

Bourlard, H. and Kamp, Y., 1988. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics 59 (4-5), 291–294.

Bronzino, J. D. and Peterson, D. R., 2014. Biomedical Engineering Fundamentals. CRC Press.

Bundy, D. T., Souders, L., Baranyai, K., et al., 2017. Contralesional Brain-Computer Interface Control of a Powered Exoskeleton for Motor Recovery in Chronic Stroke Survivors. Stroke 48 (7), 1908–1915.

Chavarriaga, R., Sobolewski, A., and MillÃ¡n, J. d. R., 2014. Errare machinale est: the use of error-related potentials in brain-machine interfaces. Frontiers in Neuroscience 8 (8 JUL), 1–13.

Chien, H.-Y. S., Goh, H., Sandino, C. M., and Cheng, J. Y., 2022. MAEEG: Masked Auto-encoder for EEG Representation Learning. In: NeurIPS 2022 Poster presentation.

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

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (Mlm), 4171–4186.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Fan, C. C., Yang, H., Hou, Z. G., et al., 2021. Bilinear neural network with 3-D attention for brain decoding of motor imagery movements from the human EEG. Cognitive Neurodynamics 15 (1), 181–189.

Gao, C., Killeen, B. D., Hu, Y., et al., 2023. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. Nature Machine Intelligence 5 (3), 294–308.

Goldberger, A. L., Amaral, L. A., Glass, L., et al., 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (23).

Goodfellow, I., Bengio, Y., and Courville, A., 2016. Deep Learning. MIT Press, `http://www.deeplearningbook.org`.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial networks.

Guetschel, P., Moreau, T., and Tangermann, M., 2024. S-JEPA: towards seamless cross-dataset transfer through dynamic spatial attention. In: Proceedings of the 9th Graz Brain-Computer Interface Conference 2024. pp. 11–16.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem, 770–778.

Herculano-Houzel, S., 2009. The human brain in numbers: a linearly scaled-up primate brain. Frontiers in Human Neuroscience 3 (NOV), 1–11.

Hinton, G. E. and Zemel, R., 1993. Autoencoders, minimum description length and helmholtz free energy. In: Cowan, J., Tesauro, G., and Alspector, J. (Eds.), Advances in Neural Information Processing Systems. 6. Morgan-Kaufmann.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., 2017. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017-Janua. IEEE, pp. 2261–2269.

Ieracitano, C., Mammone, N., Hussain, A., and Morabito, F. C., 2021. A novel explainable machine learning approach for EEG-based brain-computer interface systems. Neural Computing and Applications 0123456789 (Dl).

Jeunet, C., Glize, B., McGonigal, A., et al., 2019. Using EEG-based brain computer interface and neurofeedback targeting sensorimotor rhythms to improve motor skills: Theoretical background, applications and prospects. Neurophysiologie Clinique 49 (2), 125–136.

Kamath, U. and Liu, J., 2021. Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Springer International Publishing, Cham.

Kaya, M., Binli, M. K., Ozbay, E., et al., 2018. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. Scientific Data 5 (1), 180211.

Kingma, D. P. and Welling, M., 2019. An Introduction to Variational Autoencoders. Foundations and Trends® in Machine Learning 12 (4), 307–392.

Kostas, D. and Rudzicz, F., 2020. Thinker invariance: Enabling deep neural networks for BCI across more people. Journal of Neural Engineering 17 (5).

Kostas, D., Aroca-Ouellette, S., and Rudzicz, F., 2021. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. Frontiers in Human Neuroscience 15 (June), 1–15.

Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (Eds.), Advances in Neural Information Processing Systems. 25. Curran Associates, Inc.

Kumar, S., Sharma, S. K., Chaturvedi, J., and Sharma, A., 2020. Brain computer interface advancement in neurosciences : Applications and issues. Interdisciplinary Neurosurgery 20 (February), 100694.

Kwon, O. Y., Lee, M. H., Guan, C., and Lee, S. W., 2020. Subject-Independent Brain-Computer Interfaces Based on Deep Convolutional Neural Networks. IEEE Transactions on Neural Networks and Learning Systems 31 (10), 3839–3852.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., et al., 2018. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. Journal of Neural Engineering 15 (5), 1–30.

LeCun, Y. and Bengio, Y., 1998. Convolutional networks for images, speech, and time series. MIT Press, Cambridge, MA, USA, p. 255–258.

Lee, D.-Y., Jeong, J.-H., Lee, B.-H., and Lee, S.-W., 2022. Motor Imagery Classification Using Inter-Task Transfer Learning via a Channel-Wise Variational Autoencoder-Based Convolutional Neural Network. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 226–237.

Lee, M. H., Kwon, O. Y., Kim, Y. J., et al., 2019. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. GigaScience 8 (5), 1–16.

Liu, H. and Lv, X., 2022. EEG datasets of stroke patients. Dataset. https://doi.org/10.6084/m9.figshare.21679035.v5.

Ludwig, K. A., Miriani, R. M., Langhals, N. B., et al., 2009. Using a Common Average Reference to Improve Cortical Neuron Recordings From Microelectrode Arrays. Journal of Neurophysiology 101 (3), 1679–1689.

Lundberg, S. and Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. NIPS 2017 32 (2), 1208–1217.

Luo, Y. and Lu, B.-L., 2018. Eeg data augmentation for emotion recognition using a conditional wasserstein gan. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 2535–2538.

Mao, A., Mohri, M., and Zhong, Y., 2023. Cross-entropy loss functions: theoretical analysis and applications. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23. JMLR.org.

Marcos-Martínez, D., Martínez-Cagigal, V., Santamaría-Vázquez, E., et al., 2021. Neurofeedback training based on motor imagery strategies increases EEG complexity in elderly population. Entropy 23 (12), 1–19.

Martínez Cagigal, V., 2020. Toward Practical P300-based Brain–Computer Interfaces: Asynchrony, Channel Optimization and Assistive Applications. Ph.D. thesis, Universidad de Valladolid.

Martínez-Cagigal, V., Thielen, J., Santamaría-Vázquez, E., et al., 2021. Brain–computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review. Journal of Neural Engineering 18 (6), 061002.

Meng, J. and He, B., 2019. Exploring Training Effect in 42 Human Subjects Using a Non-invasive Sensorimotor Rhythm Based Online BCI. Frontiers in Human Neuroscience 13 (April), 1–19.

Miao, M., Yang, Z., Zeng, H., et al., 2023. Explainable cross-task adaptive transfer learning for motor imagery EEG classification. Journal of Neural Engineering 20 (6), 066021.

Miller, L. E. and Hatsopoulos, N., 2012. Neuronal Activity in Motor Cortex and Related Areas. In: Wolpaw, J. and Wolpaw, E. W. (Eds.), Brain–Computer InterfacesPrinciples and Practice. Oxford University Press, pp. 15–44.

Moldoveanu, A., Ferche, O. M., Moldoveanu, F., et al., 2019. The TRAVEE system for a multimodal neuromotor rehabilitation. IEEE Access 7, 8151–8171.

Nahmias, D. O. and Kontson, K. L., 2020. Easy Perturbation EEG Algorithm for Spectral Importance (easyPEASI). In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 2398–2406.

Neloy, A. A. and Turgeon, M., 2024. A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. Machine Learning with Applications 17, 100572.

Ofner, P., Schwarz, A., Pereira, J., et al., 2019. Attempted Arm and Hand Movements can be Decoded from Low-Frequency EEG from Persons with Spinal Cord Injury. Scientific Reports 9 (1), 7134.

Perez-Velasco, S., Santamaria-Vazquez, E., Martinez-Cagigal, V., et al., 2022. EEGSym: Overcoming Inter-Subject Variability in Motor Imagery Based BCIs With Deep Learning. IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 1766–1775.

Pérez-Velasco, S., Marcos-Martínez, D., Santamaría-Vázquez, E., et al., 2024. Unraveling motor imagery brain patterns using explainable artificial intelligence based on Shapley values. Computer Methods and Programs in Biomedicine 246 (January), 108048.

Pfurtscheller, G. and Lopes da Silva, F., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clinical Neurophysiology 110 (11), 1842–1857.

Polich, J., 2007. Updating p300: An integrative theory of p3a and p3b. Clinical Neurophysiology 118 (10), 2128–2148.

Psotta, R., 2014. The visual reaction time distribution in the tasks with different demands on information processing. Acta Gymnica 44 (1), 5–13.

Pérez-Velasco, S., Marcos-Martínez, D., Santamaría-Vázquez, E., et al., 2025. Bridging motor execution and motor imagery bci paradigms: An inter-task transfer learning approach. Biomedical Signal Processing and Control 107, 107834.

Ramos-Murguialday, A., Schürholz, M., Caggiano, V., et al., 2012. Proprioceptive Feedback and Brain Computer Interface (BCI) Based Neuroprostheses. PLoS ONE 7 (10), e47048.

Ribeiro, M. T., Singh, S., and Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier.

Roy, Y., Banville, H., Albuquerque, I., et al., 2019. Deep learning-based electroencephalography analysis: A systematic review. Journal of Neural Engineering 16 (5).

Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Saha, S. and Baumert, M., 2020. Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review. Frontiers in Computational Neuroscience 13 (January), 1–8.

Samek, W., Montavon, G., Lapuschkin, S., et al., 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE 109 (3), 247–278.

Santamaria-Vazquez, E., Martinez-Cagigal, V., Vaquerizo-Villar, F., and Hornero, R., 2020. EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-based Brain-Computer Interfaces. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28 (12), 2773–2782.

Santamaría-Vázquez, E., Martínez-Cagigal, V., Marcos-Martínez, D., et al., 2022. MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. Computer Methods and Programs in Biomedicine 230, (Under Review).

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., et al., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping 38 (11), 5391–5420.

Sebastián-Romagosa, M., Cho, W., Ortner, R., et al., 2020. Brain Computer Interface Treatment for Motor Rehabilitation of Upper Extremity of Stroke Patients—A Feasibility Study. Frontiers in Neuroscience 14 (October), 1–12.

Selvaraju, R. R., Cogswell, M., Das, A., et al., 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128 (2), 336–359.

Shahriar, S., 2021. Gan computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network.

Shrikumar, A., Greenside, P., and Kundaje, A., 2017. Learning Important Features Through Propagating Activation Differences. 34th International Conference on Machine Learning, ICML 2017 7, 4844–4866.

Shuqfa, Z., Belkacem, A. N., and Lakas, A., 2023. Decoding Multi-Class Motor Imagery and Motor Execution Tasks Using Riemannian Geometry Algorithms on Large EEG Datasets. Sensors 23 (11), 5051.

Smilkov, D., Thorat, N., Kim, B., et al., 2017. SmoothGrad: removing noise by adding noise.

Stieger, J. R., Engel, S., Jiang, H., et al., 2021. Mindfulness Improves Brain–Computer Interface Performance by Increasing Control Over Neural Activity in the Alpha Band. Cerebral Cortex 31 (1), 426–438.

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R., 2016. Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods 274, 141–145.

Sundararajan, M., Taly, A., and Yan, Q., 2017. Axiomatic Attribution for Deep Networks. 34th International Conference on Machine Learning, ICML 2017 7, 5109–5118.

Szegedy, C., Wei Liu, Yangqing Jia, et al., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 07-12-June. IEEE, pp. 1–9.

Trevisan de Souza, V. L., Marques, B. A. D., Batagelo, H. C., and Gois, J. P., 2023. A review on generative adversarial networks for image generation. Computers Graphics 114, 13–25.

Valliappan, N., Dai, N., Steinberg, E., et al., 2020. Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature Communications 11 (1), 1–12.

Varsehi, H. and Firoozabadi, S. M. P., 2021. An EEG channel selection method for motor imagery based brain–computer interface and neurofeedback using Granger causality. Neural Networks 133, 193–206.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention Is All You Need. Advances in Neural Information Processing Systems 2017-Decem (Nips), 5999–6009.

Vincent, P., Larochelle, H., Lajoie, I., et al., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11 (110), 3371–3408.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1 (6), 80.

Wolpaw, J. and Wolpaw, E. W., 2012. Brain–Computer InterfacesPrinciples and Practice. Oxford University Press.

Xie, S., Girshick, R., Dollar, P., et al., 2017. Aggregated Residual Transformations for Deep Neural Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017-Janua. IEEE, pp. 5987–5995.

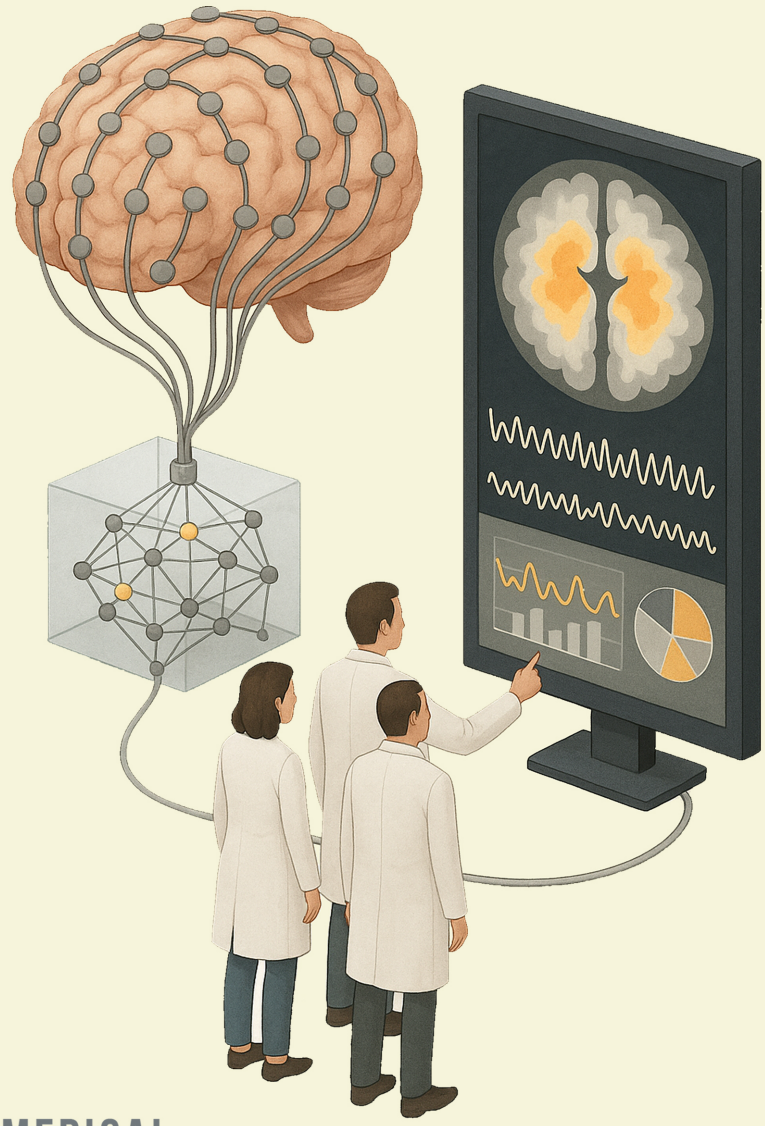
Zeiler, M. D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 818–833.

Zhang, K., Robinson, N., Lee, S. W., and Guan, C., 2021. Adaptive transfer learning for EEG motor imagery classification with deep Convolutional Neural Network. Neural Networks 136, 1–10.

Zhong, Z., Zheng, L., Kang, G., et al., 2020. Random Erasing Data Augmentation. Proceedings of the AAAI Conference on Artificial Intelligence 34 (07), 13001–13008.

Enhancing EEG-based Brain-Computer Interfaces with Deep Learning and Explainable Artificial Intelligence

Doctoral Dissertation



BIOMEDICAL ENGINEERING GROUP
UNIVERSITY OF VALLADOLID

Sergio Pérez-Velasco

Advisor: Roberto Hornero Sánchez

This Doctoral Thesis investigates how deep learning (DL) and explainable artificial intelligence (XAI) can enhance the performance and interpretability of EEG-based brain–computer interfaces (BCIs) driven by motor imagery (MI). While BCIs hold enormous promise as assistive technologies, particularly for individuals with motor impairments, their real-world deployment is often hindered by limited accuracy, inter-session variability, and the lack of physiological transparency in decoding neural signals.

To address these limitations, this work introduces EEGSym, a novel DL architecture designed to overcome inter-subject variability and achieve robust MI classification directly from raw EEG signals. EEGSym is complemented by transfer learning and data augmentation strategies, resulting in generalizable models capable of reducing BCI inefficiency. Additionally, a tailored XAI approach based on Shapley values is used to uncover the neural features most relevant to MI decoding, revealing activation patterns beyond the sensorimotor cortex and enabling reduced but effective EEG montages.

Finally, this thesis explores a direct transfer learning paradigm between motor execution (ME) and MI, showing that models trained exclusively on ME can classify MI with comparable accuracy. Together, these contributions lay the foundation for building BCIs that are not only accurate and calibration-free, but also physiologically interpretable and ready for practical application.

UVa

Doctoral Dissertation

Compendium of publications

International Mention