CROSS-SESSION EEG-BASED MENTAL WORKLOAD CLASSIFICATION USING GRAPH NEURAL NETWORKS FOR REPRODUCIBLE BRAIN-COMPUTER INTERFACE APPLICATIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

GÜLİZ DEMİREZEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

DECEMBER 2024

CROSS-SESSION EEG-BASED MENTAL WORKLOAD CLASSIFICATION USING GRAPH NEURAL NETWORKS FOR REPRODUCIBLE BRAIN-COMPUTER INTERFACE APPLICATIONS

submitted by **GÜLİZ DEMİREZEN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems Department, Middle East Technical University** by,

Date: 20.12.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: GÜLİZ DEMİREZEN

Signature :

ABSTRACT

CROSS-SESSION EEG-BASED MENTAL WORKLOAD CLASSIFICATION USING GRAPH NEURAL NETWORKS FOR REPRODUCIBLE BRAIN-COMPUTER INTERFACE APPLICATIONS

DEMİREZEN, GÜLİZ

Ph.D., Department of Information Systems

Supervisor: Prof. Dr. Tuğba Taşkaya Temizel

Co-Supervisor: Prof. Dr. Anne-Marie Brouwer

December 2024, 219 pages

Human mental workload refers to the portion of an operator's cognitive capacity used during a task. Performance of an operator may deteriorate if resource requirements of a task is higher than the available capability. Effective workload measurement is critical for tasks requiring sustained focus, such as in aviation, driving, and command-control environments. Identifying mental workload fluctuations can help reduce errors and improve user performance by alerting operators when cognitive load exceeds safe thresholds. Additionally, understanding mental workload informs the design of human-computer interaction systems and brain-computer interfaces.

The main objective of this thesis is to develop a reproducible methodology for classifying mental workload using EEG signals collected across multiple sessions. To achieve this, the current state of EEG mental workload classification studies is assessed and guidelines for reproducible and generalizable research in this domain are constructed. The employed methodology for mental workload classification leverages graph neural networks to model the dependencies between EEG channels. Furthermore, the use of optimal transport to address cross-session variability, a significant challenge in EEG-based classification, is investigated. Subject-specific evaluations using diverse metrics enable a comprehensive analysis of model performance. The outcome of this research will contribute to more robust and generalizable mental workload classification methods for BCI and other applications.

Keywords: Mental Workload, EEG, Physiological Measurement, Reproducibility, Generalizability, Machine Learning, Cross-Session Classification, Graph Neural Networks, Domain Adaptation, Optimal Transport, Brain-Computer Interface, Neuroergonomics, Neuroscience

ÖZ

ÇAPRAZ OTURUMLARDA EEG TABANLI ZİHİNSEL İŞ YÜKÜ SINIFLANDIRMASI İÇİN GRAFİK SİNİR AĞLARININ KULLANIMI: TEKRARLANABİLİR BEYİN-BİLGİSAYAR ARAYÜZÜ UYGULAMALARI

DEMİREZEN, GÜLİZ

Doktora, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Tuğba Taşkaya Temizel

Ortak Tez Yöneticisi: Prof. Dr. Anne-Marie Brouwer

Aralık 2024, 219 sayfa

İnsan zihinsel iş yükü, bir operatörün bir görevi yerine getirirken kullandığı bilişsel kapasite miktarını ifade eder. Bir görevin kaynak gereksinimleri mevcut kapasiteden yüksekse operatörün performansı düşebilir. Havacılık, sürüş ve komuta-kontrol ortamları gibi uzun süreli odaklanma gerektiren görevler için etkili zihinsel iş yükü ölçümü kritik öneme sahiptir. Zihinsel iş yükü dalgalanmalarını tespit etmek, operatörlerin bilişsel iş yüklerinin güvenli sınırları aştığında uyarılması yoluyla hataları azaltabilir ve kullanıcı performansını artırabilir. Ayrıca, zihinsel iş yükünü anlamak, insan-bilgisayar etkileşim sistemlerinin ve beyin-bilgisayar arayüzlerinin tasarımına da bilgi sağlar.

Bu tez çalışmasının ana amacı, farklı oturumlardan elde edilen EEG sinyalleri kullanılarak zihinsel iş yükü sınıflandırması için tekrarlanabilir bir metodoloji geliştirmektir. Bu amaçla, EEG tabanlı zihinsel yük sınıflandırma çalışmalarının mevcut durumu değerlendirilmiş ve bu alanda tekrarlanabilir ve genelleştirilebilir araştırmalar için yönergeler oluşturulmuştur. Zihinsel iş yükü sınıflandırmasında kullanılan metodoloji, EEG kanalları arasındaki bağımlılıkları modellemek için grafik sinir ağlarından faydalanmaktadır. Ayrıca, EEG sınıflandırma çalışmalarının bir zorluğu olan çapraz oturum değişkenliğini ele almak için optimal taşıma yöntemlerinin kullanımı da araştırılmaktadır. Farklı metrikler kullanılarak yapılan kişiye özgü değerlendirmeler, model performansının kapsamlı bir şekilde analiz edilmesini sağlamaktadır. Bu araştırmanın sonuçları, beyin-bilgisayar arayüzleri ve diğer uygulamalar için daha sağlam ve genelleştirilebilir zihinsel iş yükü sınıflandırma yöntemlerinin geliştirilmesine katkıda bulunacaktır.

Anahtar Kelimeler: Zihinsel İş Yükü, EEG, Fizyolojik Ölçüm, Tekrarlanabilirlik, Genelleştirilebilirlik, Makina Öğrenmesi, Çapraz Oturumlarda Sınıflandırma, Grafik Sinir Ağları, Etki Alanı Uyarlaması, Optimal Taşıma, Beyin-Bilgisayar Arayüzü, Nöroergonomi, Nörobilim



To my family

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Tuğba Taşkaya Temizel, for her unwavering support, guidance, trust and for embracing an optimistic and creative approach along this long journey. Her reliability, strong work ethic, and empathy have been a great source of inspiration. I would also like to extend my heartfelt appreciation to my co-supervisor, Prof. Dr. Anne-Marie Brouwer, for her support, guidance, and encouragement. Despite the physical distance between us, working with her has been a rewarding experience, and I feel privileged to have had such a knowledgeable and supportive mentor by my side. I feel fortunate to have conscientious advisors who have always promoted a positive and constructive atmosphere, believed in me and nurtured my potential.

I am thankful to the thesis supervising committee members Prof. Dr. Banu Günel Kılıç and Assoc. Prof. Dr. Murat Perit Çakır for their insightful guidance and critiques towards the successful completion of this thesis. I would also like to thank the thesis examining committee members Assoc. Prof. Dr. Hulusi Kafalıgönül and Assist. Prof. Dr. Hasan Onur Keleş for their constructive feedback and encouraging demeanor.

I would also like to thank Dr. Marcel F. Hinss for his help with questions related to the dataset and preprocessing and Dr. Nattapong Thammasan for insightful discussions on preprocessing. I would like to thank Arif Ozan Kızıldağ and Ümit Mert Çağlar for their support and effort with the laboratory computers.

I am thankful to Ludwig Göransson for the Mandalorian theme, which helped me focus during the busiest and most challenging times.

I would like to extend my appreciation to my close friends and family who did not complain but instead reached out to ask how I was doing, even when we couldn't meet for a long time. Now is the time we will make up for those days.

I would like to express my profound appreciation to my wonderful family for their endless love, care, support, compassion, and encouragement throughout my life. I am grateful to my mother, Gülden, and my father, İzzettin. I am also thankful to my mother-in-law, Ceylan, for her understanding, encouragement and love. I will forever be thankful for their presence in my life.

Last but not least, my deepest gratitude is for my dear husband, Umut. I am truly blessed to have a partner who understands and supports me both psychologically and intellectually. He taught me resilience, to rise again and again after recursive setbacks, while always comforting and motivating me with kindness. I would not have been able to complete this thesis without his unwavering support. I'm thankful for meeting him, for his presence in my life, his love and the joy he brought in it.

As a final note, I would like to thank myself. Despite the challenges of balancing a full-time job, its ups and downs, the COVID-19 pandemic, and the fatigue from long nights and weekends of intense studying, I did not quit the PhD program. I cultivated determination and learned to accept things as they are, and I am proud of myself.

I acknowledge the support from Higher Education Institutions (YÖK) through 100/2000 scholarship. I would also like to thank ROKETSAN Missiles Inc. for its support in this study.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

DEDICATION viii

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS xi

LIST OF TABLES xviii

LIST OF FIGURES xx

LIST OF ABBREVIATIONS xxiv

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Research Questions 2

1.3 Research Contributions 4

1.4 Organization of the Thesis 5

2 RELATED WORK AND BACKGROUND 7

2.1 Mental Workload 7

2.1.1 Definition and Models 8

2.1.2 Measures 9
2.1.2.1 Subjective measures 11
2.1.2.2 Task performance measures 11
2.1.2.3 Physiological measures 11
2.1.3 Electroencephalography and Mental Workload 12
2.1.3.1 EEG Frequency Bands and Cognitive Functions 12
2.1.3.2 EEG Markers of Mental Workload 12
Frontal Theta Activity 12
Parietal Alpha Suppression 12
2.1.3.3 Implications for Mental Workload Assessment 13
2.2 Machine Learning and Mental Workload 13
2.3 Domain Adaptation with Optimal Transport in EEG 15
3 REPRODUCIBILITY IN MACHINE LEARNING STUDIES USING EEG 19
3.1 Abstract 19
3.2 Introduction 20
3.3 Reproducibility in Machine Learning 22
3.4 Reproducibility in EEG 24
3.5 Reproducibility in Mental Workload Studies using EEG 25
3.5.1 Literature Search Strategy 25
3.5.2 Eligibility Criteria 26
3.5.3 Analysis of the Studies 27
3.6 Establishing Reproducible EEG Machine Learning Pipelines: Guidelines and Checklist Compilation 27

3.6.1 Business Understanding 28
3.6.2 Data Understanding 28
3.6.3 Data Preparation 30
3.6.4 Modeling 31
3.6.5 Evaluation 33
3.6.6 Deployment 34
3.7 Reproducibility in Machine Learning Models to Predict Mental Workload using EEG 34
3.7.1 Literature Search Strategy 34
3.7.2 Eligibility Criteria 35
3.7.3 Analysis of the Studies 36
3.7.4 Reproducibility Analysis 39
3.8 Discussion 45
4 GRAPH NEURAL NETWORKS FOR CROSS-SESSION CLASSIFICATION OF MENTAL WORKLOAD USING EEG 49
4.1 Methodology 50
4.1.1 Dataset and Experimental Task 50
4.1.2 Exploratory Data Analysis 51
4.1.3 Artifact Cleaning and Preprocessing 52
4.1.4 Feature Extraction 53
4.1.4.1 Segmentation 53
4.1.4.2 Node Features 54
4.1.4.3 Edge Features 55

4.1.4.4 Graph Representation 55
4.1.4.5 Node Selection 55
4.1.4.6 Data Split 55
4.1.5 Model Development 56
4.1.5.1 Graph Neural Networks 56
Message Passing and Neighborhood Aggregation 57
4.1.5.2 Model Architecture 58
4.1.5.3 Training 58
4.1.6 Performance Evaluation and Model Selection 58
4.1.6.1 Performance Metrics 58
4.1.6.2 Hyperparameters and Model Selection 59
4.1.6.3 Self Organizing Maps 60
4.1.6.4 Statistical Significance of Models 60
4.2 Experiments 61
4.2.1 Variations in Data Split 61
4.2.2 Hyperparameter Search Space 61
4.2.3 Epoch Length and Overlap 62
4.2.4 Graph Structures 62
4.2.5 Reduced Features and Node Selection 63
4.2.6 Batch Normalization 63
4.2.7 Binary Classification 63
4.3 Results 64

4.3.1 Exploratory Data Analysis 64
4.3.1.1 Raw Data 64
4.3.1.2 Features 64
4.3.1.3 Task Performance 64
4.3.2 Artifact Cleaning and Preprocessing 66
4.3.3 Graph Representation 66
4.3.4 Experiment Results 70
4.3.4.1 Model Performance 70
4.3.4.2 Significance of Experiment Results 76
4.3.4.3 Model Complexity 78
4.3.4.4 Computation Environment and Effort 81
4.3.4.5 Performance in Literature 84
pBCI competition of the Neuroergonomics Conference 2021: 84
COG-BCI Dataset: 85
4.3.4.6 Relation to Task Performance 86
4.4 Discussion 89
5 CROSS-SESSION GENERALIZABILITY OF MACHINE LEARNING MODELS . 91
5.1 Methodology 91
5.1.1 The Theory of Optimal Transport 91
5.1.1.1 The Monge Problem 92
5.1.1.2 The Kantorovich Relaxation 93
5.1.1.3 Dual Formulation 93

5.1.1.4 Wasserstein Distance 94
5.1.1.5 Entropic Regularization 94
5.1.1.6 Sinkhorn's Algorithm 95
5.1.2 Optimal Transport for Cross-Session Estimation of MWL using EEG . . 96
5.2 Experiments 97
5.3 Results 97
5.3.1 Optimal Transport of Features 97
5.3.2 Model Performance 98
5.3.3 Significance of Experiment Results 104
5.4 Discussion 104
6 CONCLUSION AND DISCUSSION 107
6.1 Discussion 107
6.2 Contributions and Impact of the Thesis 109
6.3 Limitations 111
6.4 Future Work 112
REFERENCES 115
A COMPILED GUIDELINES 135
B EXPLORATORY DATA ANALYSIS PLOTS 141
C MODEL RESULTS 169
C.1 Subject-Wise Results of COG_BCIg1 Variations 169
C.2 Model Results for Each Subject 169
C.3 Model Complexity for Each Subject 190

D CURRICULUM VITAE 217

LIST OF TABLES

Table 1 Checklist Items related to Data Collection - 1 29

Table 2 Checklist Items related to Data Collection - 2 30

Table 3 Inclusion and Exclusion Criteria 35

Table 4 List of Publications 36

Table 5 Reported Percentages of Checklist Items in the Data Understanding Phase-1 . 41

Table 6 Reported Percentages of Checklist Items in the Data Understanding Phase-2* 42

Table 7 Reported Percentages of Checklist Items in the Data Preparation Phase* 43

Table 8 MATB Experiments 52

Table 9 Behavioral Data 52

Table 10 Frequency Bands and Intervals 54

Table 11 Experiments with Variations in Data Split 61

Table 12 Hyperparameters - Baseline 62

Table 13 Hyperparameters - Small Architecture 62

Table 14 Experiments with Variations in Data Split 70

Table 15 Experiment Variations Compared to COG_BCIg1 71

Table 16 COG_BCIg1 Results with respect to Subjects 71

Table 17 Binary Classification Results Compared to COG_BCIg1 73

Table 18 Friedman Test Results 77

Table 19 Wilcoxon Test Results for Variations in Data Split Experiments 78

Table 20 Wilcoxon Test Results for Different Hyperparameter Search Spaces, Epoch Length and Overlap Experiments 79

Table 21 Wilcoxon Test Results for Graph Structures and Reduced Features and Nodes Experiments 80

Table 22 Python Packages and Versions 84
Table 23 Results of pBCI Hackathon Papers ([1]) 85
Table 24 OT Domain Adaptation Experiments 97
Table 25 Experiments with and without Optimal Transport 98
Table 26 Binary Classification Results Compared to COG_BCIg1 and COG_BCIg5 . . 101
Table 27 COG_BCIg5 Results with respect to Subjects 102
Table 28 Friedman Test Results 104
Table 29 Wilcoxon Test Results for Optimal Transport Experiments 105
Table 30 Compiled Checklist in Relation to CRISP-DM Phases 135
Table 31 COG_BCIg1 Variation Models and Figures 169

LIST OF FIGURES

Figure 1 Underload vs Overload [2] 8

Figure 2 Human Information Process (Modified from [3]) 10

Figure 3 Search Strategy for Literature Review 1 26

Figure 4 Search Strategy for Literature Review 2 35

Figure 5 Number of Participants 44

Figure 6 Overview of the Methodology 50

Figure 7 Experimental Procedure [4] 51

Figure 8 Preprocessing Pipeline 53

Figure 9 Overlapping and Non-overlapping Epochs 54

Figure 10 A Simple Graph Structure with Four Nodes and Four Edges 56

Figure 11 Image and Text as Graphs 56

Figure 12 Message Passing and Neighbourhood Aggregation [5] 57

Figure 13 GNN Architecture 58

Figure 14 SOM Plots of Features 65

Figure 15 TRACK Task Performance 67

Figure 16 SYSMON Task Performance 68

Figure 17 Graph Structure for Single Connected Component (p = 6%) using All Nodes 69

Figure 18 knn Graph Structures 69

Figure 19 5nn Graph Structures with Node Selection (Subject 1) 70

Figure 20 Subject-Wise Results for Models COG_BCIg1 and COG_BCIg2 74

Figure 21 Subject-Wise Results for Models COG_BCIg3 and COG_BCIg4 75

Figure 22 Best Heuristic Loss vs Model Complexity - Subject 1 82

Figure 23 Best Heuristic Loss vs Model Complexity - Subject 2 83

Figure 24 Behavioral Performance Data Comparison to COG_BCIg1 Model Results 88

Figure 25 Optimal Transport from Validation and Test Datasets to Train Dataset Domain 97

Figure 26 SOM Plots of Features (Subject 1) 99

Figure 27 SOM Plots of Features (Subject 2) 100

Figure 28 Performance Comparison of COG_BCIg1 and COG_BCIg5 and for each Subject 103

Figure 29 SOM Plots of Features (Subject 1) 142

Figure 30 SOM Plots of Features (Subject 2) 143

Figure 31 SOM Plots of Features (Subject 3) 144

Figure 32 SOM Plots of Features (Subject 4) 145

Figure 33 SOM Plots of Features (Subject 5) 146

Figure 34 SOM Plots of Features (Subject 6) 147

Figure 35 SOM Plots of Features (Subject 7) 148

Figure 36 SOM Plots of Features (Subject 8) 149

Figure 37 SOM Plots of Features (Subject 9) 150

Figure 38 SOM Plots of Features (Subject 10) 151

Figure 39 SOM Plots of Features (Subject 11) 152

Figure 40 SOM Plots of Features (Subject 12) 153

Figure 41 SOM Plots of Features (Subject 13) 154

Figure 42 SOM Plots of Features (Subject 14) 155

Figure 43 SOM Plots of Features (Subject 15) 156

Figure 44 SOM Plots of Features (Subject 16) 157

Figure 45 SOM Plots of Features (Subject 17) 158

Figure 46 SOM Plots of Features (Subject 18) 159

Figure 47 SOM Plots of Features (Subject 19) 160

Figure 48 SOM Plots of Features (Subject 20) 161

Figure 49 SOM Plots of Features (Subject 21) 162
Figure 50 SOM Plots of Features (Subject 22) 163
Figure 51 SOM Plots of Features (Subject 23) 164
Figure 52 SOM Plots of Features (Subject 24) 165
Figure 53 SOM Plots of Features (Subject 26) 166
Figure 54 SOM Plots of Features (Subject 29) 167
Figure 55 Subject-Wise Results for Models COG_BCIg1_smallArch and COG_BCIg1_batchNorm 170
Figure 56 Subject-Wise Results for Models COG_BCIg1_frontal and COG_BCIg1_parietal 171
Figure 57 Subject-Wise Results for Models COG_BCIg1_theta_alpha and COG_BCIg1_2sec 172
Figure 58 Subject-Wise Results for Models COG_BCIg1_2sec_no and COG_BCIg1_4sec_no 173
Figure 59 Subject-Wise Results for Models COG_BCIg1_3nn and COG_BCIg1_5nn 174
Figure 60 Subject-Wise Results for Models COG_BCIg1_50p_3nn and COG_BCIg1_50p_5nn 175
Figure 61 Subject-Wise Results for Models COG_BCIg1_25p_5nn 176
Figure 62 Model Results for Subjects 1 and 2 177
Figure 63 Model Results for Subjects 3 and 4 178
Figure 64 Model Results for Subjects 5 and 6 179
Figure 65 Model Results for Subjects 7 and 8 180
Figure 66 Model Results for Subjects 9 and 10 181
Figure 67 Model Results for Subjects 11 and 12 182
Figure 68 Model Results for Subjects 13 and 14 183
Figure 69 Model Results for Subjects 15 and 16 184
Figure 70 Model Results for Subjects 17 and 18 185
Figure 71 Model Results for Subjects 19 and 20 186
Figure 72 Model Results for Subjects 21 and 22 187
Figure 73 Model Results for Subjects 23 and 24 188
Figure 74 Model Results for Subjects 26 and 29 189
Figure 75 Best Heuristic Loss vs Model Complexity - Subject 1 191

Figure 76 Best Heuristic Loss vs Model Complexity - Subject 2 192
Figure 77 Best Heuristic Loss vs Model Complexity - Subject 3 193
Figure 78 Best Heuristic Loss vs Model Complexity - Subject 4 194
Figure 79 Best Heuristic Loss vs Model Complexity - Subject 5 195
Figure 80 Best Heuristic Loss vs Model Complexity - Subject 6 196
Figure 81 Best Heuristic Loss vs Model Complexity - Subject 7 197
Figure 82 Best Heuristic Loss vs Model Complexity - Subject 8 198
Figure 83 Best Heuristic Loss vs Model Complexity - Subject 9 199
Figure 84 Best Heuristic Loss vs Model Complexity - Subject 10 200
Figure 85 Best Heuristic Loss vs Model Complexity - Subject 11 201
Figure 86 Best Heuristic Loss vs Model Complexity - Subject 12 202
Figure 87 Best Heuristic Loss vs Model Complexity - Subject 13 203
Figure 88 Best Heuristic Loss vs Model Complexity - Subject 14 204
Figure 89 Best Heuristic Loss vs Model Complexity - Subject 15 205
Figure 90 Best Heuristic Loss vs Model Complexity - Subject 16 206
Figure 91 Best Heuristic Loss vs Model Complexity - Subject 17 207
Figure 92 Best Heuristic Loss vs Model Complexity - Subject 18 208
Figure 93 Best Heuristic Loss vs Model Complexity - Subject 19 209
Figure 94 Best Heuristic Loss vs Model Complexity - Subject 20 210
Figure 95 Best Heuristic Loss vs Model Complexity - Subject 21 211
Figure 96 Best Heuristic Loss vs Model Complexity - Subject 22 212
Figure 97 Best Heuristic Loss vs Model Complexity - Subject 23 213
Figure 98 Best Heuristic Loss vs Model Complexity - Subject 24 214
Figure 99 Best Heuristic Loss vs Model Complexity - Subject 26 215
Figure 100 Best Heuristic Loss vs Model Complexity - Subject 29 216

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ASR	Artifact Subspace Reconstruction
BEM	Boundary Element Model
BCI	Brain Computer Interface
BIDS	Brain Imaging Data Structure
CNN	Convolutional Neural Network
CLT	Cognitive Load Theory
CRISP-DM	CRoss Industry Standard Process for Data Mining
DA	Domain Adaptation
EDF	European Data Format
EEG	Electroencephalography
fNIRS	Functional Near-Infrared Spectroscopy
FFT	Fast Fourier Transform
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GCNN	Graph Convolutional Neural Network
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
GSR	Galvanic Skin Response
HR	Heart Rate
HRV	Heart Rate Variability

ICA	Independent Component Analysis
KDD	Knowledge Discovery in Databases
KSS	Karolinska Sleepiness Scale
LSTM	Long-Short Term Memory
ML	Machine Learning
MATB	Multi-Attribute Task Battery
MWL	Mental Workload
N/A	Not Applicable
NASA	National Aeronautics and Space Administration
NASA-TLX	NASA Task Load Index
NN	Neural Network
OT	Optimal Transport
pBCI	Passive BCI
PCA	Principle Component Analysis
PSD	Power Spectral Density
RF	Random Forest
RGNN	Regularized Graph Neural Network
RNN	Recurrent Neural Network
RSME	Rating Scale Mental Effort
SIMKAP	Simultaneous Capacity/Multi-Tasking
SDAE	Stacked Denoising Auto Encoder
SEMMA	Sample, Explore, Modify, Model, Assess
SVM	Support Vector Machine
SWAT	Subjective Workload Assessment Technique
TDSP	Team Data Science Process
tSNE	t-Distributed Stochastic Neighbour Embedding

CHAPTER 1

INTRODUCTION

1.1 Motivation

Brain-Computer Interfaces (BCIs) are systems that decode brain signals to control a connected device, such as a wheelchair or a computer ([6]). These systems evolved especially for the use of disabled people to aid them with communication, motor control or motor restoration, locomotion as well as for preventing seizures of epileptic patients or treatment of diseases such as autism, attention deficit hyperactivity disorder, depression and schizophrenia ([7]). Passive BCIs are introduced as as an extension of BCIs, as connected computer system used to assess mental states either to present this measurement to the user or to adjust the system by using the measured state as feedback ([8, 9]). A passive BCI aims to measure spontaneous user states from brain activity that occurs without voluntary control or stimulation ([10]).

The rapidly developing digital world alters the load on the human operator from physical to mental. Modeling and assessing mental workload have received increasing interest for the last few decades [2]. It can be measured to quantify the performance of operators as well as the system so that it does not impose unnecessary load on the human operator. Therefore, assessing mental workload serves to keep operators from unnecessary mental load and also helps for design purposes such that end products induce the least amount of load on users while ensuring the performance requirements are met. The primary aim would be to reduce user errors due to excessive load which is of great importance for performance-critical professions that require long-time focus such as air traffic controllers, drivers, pilots, or command control officers.

Mental workload is inherently multi-dimensional and complex, making it difficult to reach a consensus on its precise definition. It is often described as a function of task demands, individual capacities, and performance outcomes [11]. Each additional task or increased task difficulty requires more information processing from the user. If the cognitive demands exceed the individual's processing capacity, performance is likely to deteriorate. However, the experience of workload is subjective and can be influenced by various factors such as prior experience, skill level, and emotional state [12]. This variability necessitates robust and adaptable measurement techniques capable of capturing the nuances of cognitive load.

Mental workload measurement techniques can be subjective measures, task-related performance measures, or physiological measures. EEG is widely used for mental workload assessment because it provides a direct measure of neural activity with high temporal resolution.

EEG signals are capable of capturing variations in cognitive states related to changes in mental workload [13]. Unlike other physiological measures, EEG reflects real-time brain activity, making it suitable for monitoring rapid changes in cognitive load. Additionally, EEG is non-invasive, relatively low-cost, and portable, facilitating its application in real-world scenarios.

EEG data is inherently structured, as they represent signals collected from different brain regions with spatial relationships between the electrodes. Traditional machine learning approaches often treat EEG signals as a collection of independent time series, overlooking the spatial dependencies. Graph Neural Networks (GNNs) offer a powerful solution by explicitly modeling the relationships between EEG channels as a graph [14].

GNNs can capture both the local interactions between nearby electrodes and the global patterns across the entire scalp. This capability makes GNNs particularly suited for EEG analysis, as they can leverage the underlying brain connectivity to improve the classification of mental states. Additionally, GNNs can adapt to varying graph structures, allowing the incorporation of different connectivity measures (e.g., k-nearest neighbors, functional connectivity), thus offering flexibility in handling different EEG datasets [15].

However, EEG signals are highly non-stationary, and their characteristics can vary across different sessions, tasks, and individuals [1]. This variability presents significant challenges for developing robust models, particularly in real-world applications where recording conditions and individual differences cannot be controlled. Addressing this issue requires innovative modeling approaches capable of handling the inherent variability of EEG data.

Developing machine learning models that generalize well across sessions and subjects remains a significant challenge in EEG-based studies. EEG signals are subject to substantial variability due to changes in electrode positioning, environmental conditions, and individual differences. Cross-session modeling aims to address the non-stationarity of EEG signals, ensuring that models trained on data from one session can still perform well when applied to data from a different session [16].

Subject-based modeling is considered in this thesis because of the pronounced individual differences in EEG signals. Each subject's brain activity patterns are influenced by their unique anatomical and psychological characteristics. Consequently, models trained on data from multiple subjects often struggle to generalize well. By training separate models for each subject, this thesis aims to develop a cross-session modeling methodology to quantify mental workload using EEG measurements. Moreover, research towards reproducible machine learning using EEG data is investigated.

1.2 Research Questions

This thesis aims to answer the following research questions:

- RQ1: What is the status of reproducibility in machine learning studies to estimate mental workload using EEG?

- RQ2: How can we achieve reproducible and generalizable machine learning models that use EEG?
- RQ3: How can mental workload be predicted using EEG data and graph neural networks?
  - RQ3.1: What is the effect of different graph structures on model performance?
  - RQ3.2: What are the possible solutions to prevent overfitting?
  - RQ3.3: Which metrics and evaluation methods can offer a more comprehensive assessment of the models?
  - RQ3.4: What is the effect of model complexity on model performance?
  - RQ3.5: What insights can we collect from subject-specific analysis?
  - RQ3.6: Can we identify the least number of channels that would minimize EEG device requirements?
- RQ4: How can we mitigate effects of variability between sessions to achieve models that are generalizable over time using domain adaptation techniques?

Within the scope of this thesis, three main studies are conducted to answer the above research questions:

The first study entails establishing guidelines to support reproducibility and generalizability in EEG modeling studies using machine learning and assessing status of reproducibility in mental workload classification using EEG data. For this purpose, current efforts in establishing reproducibility in machine learning and in EEG are investigated followed by a systematic literature review to find studies about reproducibility in mental workload prediction using EEG. Using these, guidelines are formulated as structured along the widely recognized Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. By using these guidelines, researchers can ensure transparency and comprehensiveness of their methodologies, therewith enhancing collaboration and knowledge-sharing within the scientific community, and enhancing the reliability, usability and significance of EEG and machine learning techniques in general. A second systematic literature review extracts machine learning studies that used EEG to estimate mental workload. The reproducibility status of these studies are evaluated using the constructed guidelines to extract areas studied and overlooked as well as to identify current challenges against reproducibility.

In the second part of the thesis, a reproducible pipeline is developed for estimating mental workload from EEG signals using GNNs. The proposed approach leverages GNNs' ability to model the relationships in EEG data, incorporating spectral features from multiple frequency bands and graph-based representations of the signals constructed using spatial relations. Comprehensive experiments explore the effects of variations in model parameters, data segmentation, graph representations, data splits, and node selection on model performance, providing insights into the robustness and limitations of GNNs in EEG-based mental workload classification. Reliable model assessment is emphasized by evaluating generalizability through test performance on unseen datasets, ensuring that results reflect real-world applicability. In addition to accuracy, confusion matrices are utilized to provide deeper insights into model performance across different classes. Furthermore, subject-specific Self-Organizing

Map (SOM) visualizations of feature distributions between sessions are analyzed, highlighting the impact of session variability on feature separability and its relationship to model performance. This framework emphasizes reproducibility, offering a transparent and extensible pipeline for future research in this domain.

Finally, the third part of the thesis addresses the challenge of cross-session variability, which is shown in the second part as a critical obstacle to achieving generalizable EEG-based models. Reaching a generalizable model using EEG is challenging and necessitates innovative computational solutions. Domain adaptation is employed to mitigate these effects. Optimal Transport (OT) aligns EEG data distributions across sessions, enabling the transfer of features without requiring retraining. The application of OT demonstrates improvements in cross-session performance, advancing the field toward the development of EEG-based models that are not only accurate but also robust to session-to-session variations.

1.3 Research Contributions

The contributions of this thesis are summarized as follows:

1. Guidelines for reproducibility and generalizability: Presenting a framework to ensure the reproducibility of EEG machine learning models, which is essential for building trust and collaboration in the scientific community.

2. Reproducibility status assessment: Evaluating the reproducibility of studies that estimate mental workload using EEG, highlighting challenges and providing recommendations for future research.

3. Graph neural networks for mental workload estimation: Developing graph neural network (GNN) models to predict mental workload from EEG data and presenting methods to reliably assess the performance of these models.

4. Cross-session modeling of mental workload: Cross-session modeling using EEG data and identifying the challenges posed by session-to-session variability in EEG data

5. Domain adaptation techniques: Applying OT domain adaptation to mitigate the effects of cross-session variability, improving model generalizability.

6. Identification of subject-specific features: Comprehensive exploration of subject-specific variability in EEG signals to identify challenges in generalization to unseen sessions, using techniques like SOM visualizations and model performance metrics.

The findings of this thesis aim to improve the performance and real-world usability of EEG-based BCI systems, contributing to a deeper understanding of mental workload and cognitive states.

1.4 Organization of the Thesis

The organization of this thesis is as follows: Chapter 2 provides background information on mental workload and reproducibility in Machine Learning (ML) using EEG Data. The chapter continues with the literature review of GNNs for EEG and Domain Adaptation with Optimal Transport.

Chapter 3 introduces guidelines for achieving reproducibility and generalizability in EEG-based classification models. This chapter includes a systematic literature review assessing the current status of reproducibility in mental workload estimation using EEG.

Chapter 4 presents GNN-based modeling studies for mental workload estimation and methods to improve model performance, including the analysis of different graph structures and techniques to prevent overfitting.

Chapter 5 explores domain adaptation techniques, specifically focusing on the use of OT to mitigate session variability and enhance model generalizability.

Chapter 6 concludes the thesis by stating contributions and limitations and suggesting possible future work to improve upon the findings presented.

CHAPTER 2

RELATED WORK AND BACKGROUND

In this chapter, background information on mental workload is presented, focusing on its definition, measurement methods, and associated neural correlates to establish a comprehensive understanding of this key concept. The next section explores machine learning approaches that utilize EEG, emphasizing the role of Graph Neural Networks (GNNs). It discusses the fundamental components of GNNs and recent advancements, particularly in the context of EEG analysis. GNNs have gained significant traction in neuroscience and brain-computer interface research due to their ability to model complex brain connectivity patterns. Finally, this chapter reviews related work on domain adaptation techniques, with a specific focus on optimal transport methods, which have shown promise in addressing the challenges of variability in EEG-based mental workload classification.

2.1 Mental Workload

With advancements in technology shifting the burden on human operators from physical to cognitive tasks, mental workload (MWL) has become a central topic of research, particularly in psychology, ergonomics, and human factors domains. The importance of MWL studies spans various existing and emerging fields. In professions where human performance is critical, continuous monitoring of MWL can help minimize user errors by alerting the operator or facilitating adaptive system interventions. Additionally, MWL assessments are utilized as a design consideration to optimize materials and user interfaces, ensuring minimal cognitive load. This is particularly beneficial in adaptive learning systems, where the interface can be dynamically adjusted based on the user's cognitive state. Furthermore, monitoring MWL is crucial in brain-computer interface applications to assess whether a user can effectively engage with the system at a given moment.

This section provides background information by outlining key research on the definition of mental workload, its measurement methods, and its neural correlates as observed in EEG studies.

2.1.1 Definition and Models

Historically, MWL has lacked a universally accepted definition due to its complex and multidimensional nature as it involves various factors such as task demands, cognitive resources, operator performance, and contextual influences. Nonetheless, common components are often found across different definitions and models of MWL.

In one of the early attempts to define MWL for man-machine interface research, [3] reviewed related literature and proposed a model with three main components: input load, operator effort and performance. Input load refers to external factors affecting the operator, categorized into environmental, design-induced, and procedural sources. Operator effort is an internal component influenced by input load, shaped by factors such as psychophysical characteristics, background, and personality as well as changing factors of experience, motivation, and attentiveness. Additionally, operator effort is affected by the level of performance required by the task. Performance represents the outputs generated by the operator, measured in relation to the specific task requirements.

The relationship between task difficulty and performance, MWL and activation levels is shown in Figure 1.

Task difficulty increases as task demands increase which results in an increase of mental workload by definition, leading to changes in performance. When task demands are very low, the operator is in an "underload" state, with resources underutilized. As task demands increase, performance improves up to an optimal level. However, beyond a certain threshold, performance declines sharply due to "overload", where task demands exceed the operator's resources. This highlights the need for optimizing task demands to maintain high performance. Notably, mental activation levels do not have a direct relationship with task demands.

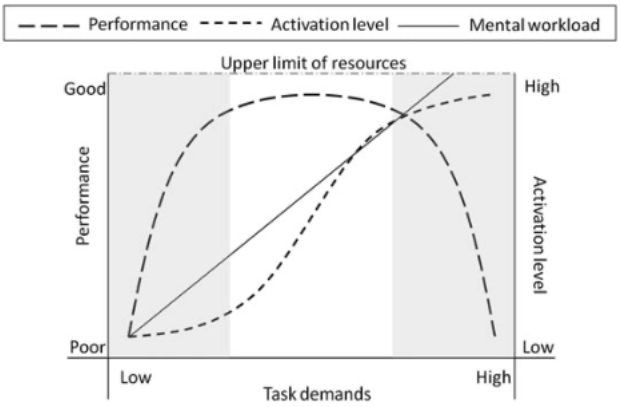


Figure 1: Underload vs Overload [2]

Another framework is Cognitive Load Theory (CLT), which focuses on cognitive load during learning within instructional design and cognitive processes. Cognitive load is categorized into three types: "intrinsic", "extraneous", and "germane" [17].

Intrinsic load is related to the complexity and interactivity of elements within the task. Extraneous load arises from poor instructional design, causing unnecessary cognitive burden. Germane load occurs when learners are appropriately challenged, facilitating deeper cognitive processing [18]. [19] emphasized the importance of task strategies in cognitive load, highlighting the role of domain-specific knowledge in distinguishing between novices and experts during problem-solving. Attention to problem states and cognitive processing capacity are factors of learning during problem solving. Measures of cognitive load must consider task difficulty, subject knowledge, and task strategy, all of which are influenced by prior experience.

Using an analogy with physical load, mental workload can also be described in terms of "stress" and "strain" [2]. "Stress" occurs as a result of task-related requirements, such as time constraints or task difficulty, while "strain" is related to the operator's response, influenced by available cognitive resources.

Many informal definitions of MWL from the literature are listed in [20] all of which include some or all of the aforementioned task, operator, and performance components.

More recently, [21] presented a comprehensive review of MWL theories and definitions and introduced a novel inclusive definition of mental workload, aiming to synthesize the different operational definitions and provide a framework that accounts for the interaction between task characteristics, operator attributes, cognitive resources, and the context of task execution. They define MWL as follows:

"Mental workload (MWL) represents the degree of activation of a finite pool of resources, limited in capacity, while cognitively processing a primary task over time, mediated by external stochastic environmental and situational factors, as well as affected by definite internal characteristics of a human operator, for coping with static task demands, by devoted effort and attention." ([21])

This comprehensive approach helps unify the existing fragmented views on MWL, while acknowledging it as a dynamic and evolving construct rather than a static measure.

2.1.2 Measures

Different approaches have been developed to measure MWL, focusing on either input load or operator effort. In [3], input load measures are predefined based on task characteristics, including environmental, design, and procedural variables. Emphasis is placed on assessing operator effort, which is classified into four main categories: time-line analysis, information processing studies, operator activation-level studies, and subjective effort ratings. Time-line analysis is a load-oriented method that examines the duration of task elements relative to performance requirements, such as time constraints and expected time pressures. Information processing studies conceptualize the operator as a limited-capacity system and aims to measure excess mental capacity with a focus on task performance. Common approaches include secondary tasks (or loading tasks) and control/information theory models to assess mental workload. Operator activation-level studies hypothesize that effort influences physiological

responses, such as changes in heart rate or pupillary dilation. Finally, subjective effort ratings provide an assessment of the operator's conscious perception of overall effort.

[3] highlights the importance of distinguishing between different types of loading factors and effort when evaluating workload measures. A simple sketch is presented for human information process. Combining the descriptions, a modified version is presented in Figure 2. Different types of effort are elicited based on input load characteristics. The model identifies key stages of effort: scanning and perception, which depend on attention levels; information processing, where the input is transformed to generate an output; and output manifestation, through actions such as movement or speech. Performance is evaluated based on the generated output, while all stages rely on memory recall and are influenced by the operator's emotional state. Different measures may be suitable for assessing specific types of effort, emphasizing the need for task-specific modeling of information processing.

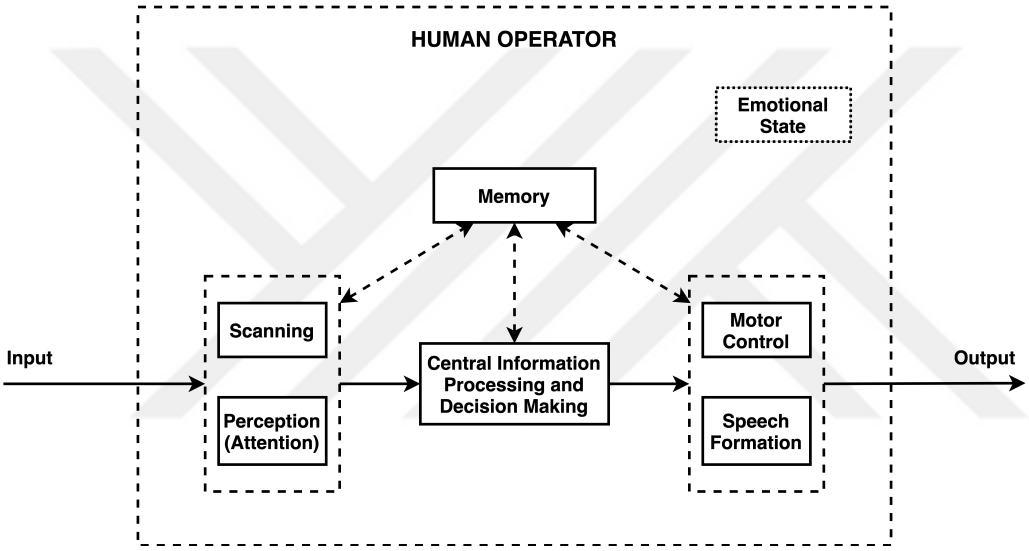


Figure 2: Human Information Process (Modified from [3])

In a broader classification, [2] categorizes mental workload measurement techniques into three main types:

1. Subjective measures
2. Task-related measures
3. Physiological measures

This classification aligns with previously mentioned methods: time-line analysis and information processing studies correspond to task-related measures, operator activation levels relate to physiological measures, and subjective effort ratings match subjective measures. These categories are described briefly along with examples in the following subsections. [21] describes these measures in detail and in relation with a comprehensive review to define mental workload.

2.1.2.1 Subjective measures

Subjective self-reports capture the operator's consciously perceived mental workload, typically administered at the end of a task. Most well-known scales include NASA Task Load Index [22], the Subjective Workload Assessment Technique (SWAT) [23], and the Rating Scale Mental Effort (RSME) [24]. These instruments assess various dimensions of workload (e.g., mental demand, effort, and frustration), providing a comprehensive evaluation of the operator's perceived effort.

Subjective measures are advantageous because they are easy to administer and provide direct insights into the operator's experiences. However, they are also limited by their reliance on self-reports, which can be influenced by individual biases and may not accurately reflect real-time workload changes.

2.1.2.2 Task performance measures

Mental workload is typically studied using either single-task loading or secondary tasks. In single-task paradigms, different levels of task difficulty are employed to observe variations in performance, allowing inferences about mental workload. When secondary tasks are used to induce mental workload, they are introduced as an additional load, requiring the operator to maintain performance on the primary task while also performing a secondary activity. The performance on the primary and/or secondary task can be used as an indicator of the operator's remaining cognitive resources.

Examples of tasks used to measure mental workload include n-back task [25], mental arithmetic task [26], and various cognitive tasks such as mazes, information sampling, and multitasking [27, 28]. Additionally, specialized tasks, such as flight, driving or military platform simulations are often employed in specific domains to measure MWL under realistic conditions.

2.1.2.3 Physiological measures

Physiological measures capture changes in the operator's bodily state in response to task demands. Commonly used physiological metrics include electroencephalography (EEG), eye movement, heart rate, and respiration [20]. Emerging techniques such as functional Near-Infrared Spectroscopy (fNIRS), functional Magnetic Resonance Imaging (fMRI), and oximetry are also gaining popularity. An extensive list is provided by [29], which includes measures like galvanic skin response (GSR), skin temperature, posture, and speech.

Physiological features are often extracted to quantify workload changes, as these indicators typically vary with increasing task demands until reaching a saturation point, where the operator's cognitive resources are exhausted, and performance begins to decline. For instance, EEG signals have been widely used to study MWL, leveraging changes in brain activity patterns across different frequency bands (e.g., alpha, theta, beta) to assess cognitive load.

Physiological measures offer the advantage of continuous, objective monitoring without relying on the operator's self-assessment. However, they often require specialized equipment and may be influenced by factors unrelated to mental workload, such as physical activity or emotional state.

2.1.3 Electroencephalography and Mental Workload

Electroencephalography (EEG) is a non-invasive method for recording electrical brain activity through electrodes placed on the scalp. The signals captured by EEG represent the summed electrical activity of neurons near the electrodes, making it a time-series data modality with high temporal but relatively low spatial resolution. Each electrode corresponds to a specific channel in the EEG recording, providing localized information about brain activity in the area it monitors. This localized measurement allows researchers to capture neural oscillations, which are rhythmic fluctuations in brain activity across various frequency bands.

2.1.3.1 EEG Frequency Bands and Cognitive Functions

Brain oscillations within specific frequency ranges are associated with distinct neural functions and cognitive states. These frequency bands typically include delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (>30 Hz) rhythms. The analysis of these oscillatory patterns has proven useful in understanding various mental processes, including attention, memory, and mental workload (MWL) [13, 30].

2.1.3.2 EEG Markers of Mental Workload

Event Related Potentials (ERPs) are indicative of mental workload, specifically the P300 decreases with increasing workload ([31]). Mental workload has also been linked to characteristic changes in specific EEG frequency bands. Numerous studies have shown that increases in task difficulty and cognitive demand are accompanied by distinct changes in frontal and parietal brain regions. In particular:

Frontal Theta Activity : An increase in theta power over the frontal cortex (e.g., Fz electrode) is consistently reported as an indicator of heightened cognitive control and working memory load [32]. This response is often attributed to increased activation of the anterior cingulate cortex, which is engaged during tasks requiring sustained attention and cognitive effort.

Parietal Alpha Suppression : A decrease in alpha power in parietal regions (e.g., Pz electrode) is associated with increased task engagement and information processing. Alpha oscillations are typically linked to a relaxed state or inactivity, and their suppression reflects greater attentional demands and cognitive workload [33, 34].

One widely used metric for estimating mental workload is the ratio of frontal theta power to parietal alpha power (θ Fz / α Pz). This ratio has been shown to increase with rising cognitive demands, providing a reliable indicator of mental workload. [35] demonstrated that this ratio outperforms other derived metrics, such as multiplicative or logarithmic combinations of theta and alpha power. Consequently, a well-established metric is the ratio between theta power over frontal sites and alpha power over parietal sites.

In addition to theta and alpha bands, studies have explored changes across a broader range of frequencies. [36] conducted a detailed analysis during multitasking scenarios and observed increased power in frontal delta, theta, and beta bands, while occipital alpha power significantly decreased under higher mental workload conditions. These findings align with the hypothesis that increased frontal activation reflects higher cognitive control, whereas reduced occipital alpha indicates suppressed visual processing during cognitively demanding tasks. Additionally, [37] found that theta long-range coherence increases and anterior upper alpha short-range connectivity decreases when task demands increase in working memory.

2.1.3.3 Implications for Mental Workload Assessment

The ability of EEG to capture real-time changes in brain activity makes it an invaluable tool for assessing mental workload. The well-established markers, such as increased frontal theta and decreased parietal alpha, provide insights into cognitive states and task demands. These markers are particularly useful in applications such as brain-computer interfaces (BCIs), adaptive human-computer interaction, and cognitive workload monitoring in real-world environments.

In this thesis, EEG data will be analyzed to identify neural signatures of mental workload using advanced machine learning methods, including Graph Neural Networks (GNNs), which are capable of modeling the complex connectivity patterns inherent in brain activity.

2.2 Machine Learning and Mental Workload

Automated evaluation of mental workload using data-driven machine learning models has been applied intensively in the last couple of years. A variety of classifiers have been explored, including K-nearest neighbour (KNN) [38], logistic regression [39], Support Vector Machines (SVMs) [25, 39], Artificial Neural Networks (ANNs) [40, 41]. More complex architectures such as adaptive Stacked Denoising Auto Encoder (SDAE) [42], Convolutional Neural Networks (CNNs) [39, 43], Long Short-Term Memory (LSTM) networks [44, 45] have also been applied, demonstrating improved classification performance in many cases.

A promising direction for modeling mental workload from EEG signals involves graph-based methods, where the spatial or temporal information can be explicitly integrated into the models. Graph theory-based machine learning algorithms, such as graph neural networks (GNNs) and transformer networks utilizing attention mechanisms, have shown potential for capturing complex connectivity patterns in EEG data ([46]). However, research employing these algorithms for mental workload assessment remains limited. Recent studies have applied dynamic graph convolutional neural networks (DGCNNs) [47] and regularized GNNs [48] for

emotion recognition tasks using EEG signals, both reporting superior performance compared to state-of-the-art methods in related applications.

A recent survey showed that no studies have yet employed GNNs or transformer-based models specifically for mental workload classification using EEG data ([46]). This gap presents an opportunity to leverage graph theory-based machine learning models, which are well-suited for mapping the complex, dynamic nature of brain connectivity. In this study, these models are explored for their potential to detect characteristic features and patterns associated with mental workload.

Graph Neural Networks (GNNs) were first introduced in 2009 as a neural network-based method designed to extend traditional neural network models to graph-structured data domains [49]. The initial GNN framework defined a mapping from nodes and their features in a graph to a multi-dimensional Euclidean space, enabling the model to capture complex relationships in structured data. This foundational model relied on relaxation methodologies and an information diffusion approach, which paved the way for further developments in GNN architectures. Following this pioneering work, various GNN models have been proposed, each introducing novel message-passing and aggregation mechanisms to handle diverse graph types, such as Graph Convolutional Networks (GCN) [50], Graph Attention Networks (GAT) [51], and Graph Isomorphism Networks (GIN) [52].

Graphs are ubiquitous in real-life applications, including computer networks, social networks, citation networks, brain connectivity, the Internet, knowledge graphs, scene graphs, molecules and traffic networks. Recent advancements have demonstrated the efficacy of GNNs in solving complex problems that were previously challenging for traditional machine learning models. In a prominent application, GNNs were used for drug discovery by predicting molecular properties. Here, atoms serve as nodes, and chemical bonds form the edges. This approach led to the discovery of Halicin, an antibiotic structurally distinct from conventional antibiotics, which effectively targets a wide range of bacteria in mice [53]. In another example, AlphaFold by DeepMind managed to solve a 50-year-old problem of protein folding in biology. AlphaFold predicted the 3D structure of a protein from its 1D amino acid sequence by representing the folded protein as a spatial graph and applying neural networks with attention mechanisms to predict node-level features [54].

The suitability of GNNs for analyzing brain signals, particularly EEG data, lies in their capacity to model relationships between different channels (nodes) and the connectivity patterns (edges) that represent neural interactions. The brain involves extensive communication between distant, distinct regions ([55]). EEG signals can be naturally represented as graphs, where nodes correspond to EEG electrodes, and edges can reflect spatial proximity or functional connectivity metrics like correlation, coherence, or mutual information. GNNs excel at capturing both local and global brain connectivity, making them particularly advantageous for analyzing the complex, non-Euclidean structure of EEG data. Modern GNN approaches have introduced adaptive edge learning, where the graph structure is refined dynamically based on the data, allowing the model to uncover latent neural connectivity patterns.

Recent research has demonstrated the potential of GNNs in various EEG-related tasks. For instance, [56] proposed a graph signal-based deep learning framework to classify videos using EEG data, employing Graph Convolutional Neural Networks (GCNNs) and constructing

intra-band and inter-band edges based on correlation and distance metrics. Their approach showed improved performance over baseline classifiers like k-NN and random forest.

Another significant contribution is the Dynamic Graph Convolutional Neural Network proposed by [47] for EEG-based emotion recognition. This model was notable for its ability to adaptively learn the relationships between EEG channels, thereby capturing dynamic connectivity patterns that enhance feature extraction. Similarly, [57] applied a spatio-temporal GCNN for EEG motor imagery classification, utilizing a learnable mask to uncover meaningful latent graph structures that align with neuroscientific insights.

[48] introduced a Regularized Graph Neural Network (RGNN) that integrates biological topology information to capture both local and global associations of EEG channels. Their work incorporated domain adversarial training and emotion-aware distribution learning to address cross-subject variability and noisy labels, demonstrating superior performance compared to traditional methods.

In the context of clinical applications, [58] proposed a GNN model for classifying abnormal scalp EEGs in patients with neurological disorders, achieving better diagnostic accuracy than human experts and conventional machine learning methods. Their approach highlighted the advantage of GNNs in capturing both spatial and functional connectivity, essential for distinguishing subtle pathological patterns in EEG data.

For brain-computer interface (BCI) applications, [59] developed a GNN with a multilevel feature fusion structure, combining outputs from convolutional layers across different stages to leverage both local and global neural information. Their model significantly improved decoding performance on the Physionet motor imagery dataset, surpassing standard algorithms.

Hybrid GNN models that combine graph convolution with other deep learning components, such as Long Short-Term Memory (LSTM) networks, have also emerged. [60] proposed a novel fusion model that integrates GCNNs and LSTM units for emotion recognition from EEG signals. The GCNNs capture spatial graph information, while LSTMs extract temporal dependencies, resulting in improved classification accuracy on the DEAP dataset.

The application of Graph Neural Networks to EEG data has grown rapidly in recent years, driven by the ability of GNNs to effectively capture complex brain connectivity patterns. This thesis aims to leverage GNNs for modeling mental workload, exploring their potential to uncover characteristic features in EEG signals and providing insights into the underlying neural dynamics.

2.3 Domain Adaptation with Optimal Transport in EEG

Domain adaptation methods for EEG signals include diverse approaches such as transfer learning, meta-learning, feature extraction, deep learning, hybrid models, few-shot learning, and cross-modal transfer learning. These techniques have greatly improved EEG signal processing, enhancing classification accuracy and robustness in applications like clinical diagnostics and brain-computer interfaces. As the field progresses, developing innovative domain

adaptation strategies is essential to manage the variability of EEG data across subjects and conditions.

Optimal transport (OT) [61] has emerged as a powerful technique for domain adaptation, addressing the challenges associated with transferring knowledge from a source domain to a target domain. This method leverages the mathematical framework of optimal transport to align probability distributions of different domains, facilitating effective learning even when the domains exhibit significant discrepancies.

One of the primary advantages of optimal transport in domain adaptation is its ability to handle distributions that do not share common support. Traditional domain adaptation methods often rely on the assumption that the source and target distributions have overlapping support, leading to reduced performance when this assumption is violated. In contrast, optimal transport aligns distributions by minimizing the cost of transferring mass between them, making it suitable for aligning disjoint domains [62]. This flexibility is particularly beneficial in real-world applications where data from different domains may not be directly comparable. Additionally, optimal transport uses the Wasserstein distance to measure the difference between probability distributions. This metric accounts for the cost of transforming one distribution into another and is sensitive to the geometry of the data, making it particularly effective in scenarios where the structure of the data is important [63, 64]. This characteristic enables optimal transport to preserve the underlying relationships within the data, which is often lost in simpler distance metrics used in traditional domain adaptation methods.

Another significant advantage of optimal transport is its ability to incorporate regularization techniques, which can enhance the stability and robustness of the mappings. Regularization in optimal transport can help mitigate issues such as model instability, particularly in high-dimensional spaces, by imposing constraints on the transport plan ([65]). Additionally, optimal transport excels in scenarios with limited labeled data in the target domain. By leveraging the structure of the source domain and utilizing unlabeled data from the target domain, it facilitates effective knowledge transfer [66]. This capability is critical in fields like medical imaging, where labeled data is often scarce [67].

In addition, optimal transport methods can be computationally efficient, especially with advancements in algorithms and optimization techniques. Recent improvements, such as entropic regularization and approximate solvers, have significantly enhanced the scalability of OT, enabling its application to large datasets and even real-time scenarios [65, 68]. These methods address challenges that traditional domain adaptation techniques often struggle with, such as label shift, where the label distribution in the target domain differs from the source domain. By aligning these distributions, OT improves classification performance in the target domain [69]. Furthermore, the flexibility to align distributions makes it suitable for domain generalization, where models must perform well across multiple unseen domains [70]. Additionally, OT enhances robustness in domain adaptation by managing noise and outliers, a common issue in real-world datasets. Its incorporation of robust statistical measures ensures reliable performance even when data quality can vary significantly [62, 71].

In conclusion, optimal transport represents a significant advancement in the field of domain adaptation, offering a flexible, principled, and computationally efficient approach to aligning disparate data distributions. Its advantages, including the ability to handle non-overlapping

supports, incorporate regularization, and manage label shifts, make it a superior choice compared to traditional domain adaptation methods.

Recent research on domain adaptation in electroencephalography (EEG) has shown promising results. One study by [66] proposed a method that combines optimal transport (OT) with frequency mix-up for EEG-based motor imagery recognition. This approach effectively aligns EEG signal distributions across subjects, improving model generalization. The integration of OT into the transfer learning framework significantly enhanced classification performance, especially in scenarios with limited target domain data ([66]). [72] proposed a method that utilized OT to align the distributions of motor imagery EEG signals, thereby enhancing the model's ability to generalize across different subjects. Their findings demonstrated that optimal transport played a crucial role in mitigating the effects of domain shift, leading to improved classification outcomes. [73] proposed backward optimal transport for domain adaptation, where new session data is transported to a calibration session to avoid model retraining. When the labels of the first 20 trials (out of a total of 140 trials for Dataset 1 and 180 trials for Dataset 2) of the new session are used for computing the transport, classification performance rises. The proposed method is able to mitigate the cross-session variability in motor imagery BCIs. These studies collectively demonstrate the potential of optimal transport in addressing domain shift challenges in EEG-based applications, particularly for motor imagery classification. As optimal transport techniques continue to evolve, they offer substantial potential for improving the generalization and robustness of BCI systems across different subjects and sessions in real-world scenarios.

CHAPTER 3

REPRODUCIBILITY IN MACHINE LEARNING STUDIES USING EEG

Reproducibility is a fundamental pillar of scientific research, serving as the basis for verifying results and building upon previous work. In the field of machine learning, where complex models are often used to analyze large datasets, reproducibility has become increasingly challenging. The situation is prominent in domains involving physiological data, such as electroencephalography (EEG), due to the variability inherent in both the data and experimental protocols.

In the context of brain-computer interfaces (BCIs) and neuroergonomics, machine learning techniques are frequently employed to decode mental states from EEG signals, with mental workload estimation being one of the most common applications ([74]). However, the reproducibility of these machine learning models remains a significant barrier to their broader adoption and reliable use.

The journal paper published in April 2024 in "Open Science to Support Replicability in Neuroergonomic Research" Research Topic of Frontiers in Neuroergonomics journal ([75]), addresses the gap in reproducibility by proposing methods to enhance reproducibility in machine learning studies using EEG and assesses the reproducibility status of publications that classify mental workload using EEG. The following research questions are addressed with this publication:

- RQ1: What is the status of reproducibility in machine learning studies to estimate mental workload using EEG?
- RQ2: How can we achieve reproducible and generalizable machine learning models that use EEG?

3.1 Abstract

This study addresses concerns about reproducibility in scientific research, focusing on the use of electroencephalography (EEG) and machine learning to estimate mental workload. We established guidelines for reproducible machine learning research using EEG and used these to assess the current state of reproducibility in mental workload modeling. We first started by

summarizing the current state of reproducibility efforts in machine learning and in EEG. Next, we performed a systematic literature review on Scopus, Web of Science, ACM Digital Library, and Pubmed databases to find studies about reproducibility in mental workload prediction using EEG. All of this previous work was used to formulate guidelines, which we structured along the widely recognized Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. By using these guidelines, researchers can ensure transparency and comprehensiveness of their methodologies, therewith enhancing collaboration and knowledge-sharing within the scientific community, and enhancing the reliability, usability and significance of EEG and machine learning techniques in general. A second systematic literature review extracted machine learning studies that used EEG to estimate mental workload. We evaluated the reproducibility status of these studies using our guidelines. We highlight areas studied and overlooked and identify current challenges for reproducibility. Our main findings include limitations on reporting performance on unseen test data, open sharing of data and code, and reporting of resources essential for training and inference processes.

3.2 Introduction

Reproducibility is fundamental for research advancement. Reproducing results, not only by the owners of the original study but also by other researchers, enables establishing a solid foundation that can be built upon for global research progress. The ability to repeat a study of others using the exact same methodology and produce the same results facilitates the verification and validation of study findings, identification or reduction of errors, and accurate comparison of newly developed methodologies. This process not only increases the trustworthiness of findings but also bolsters the credibility of the researchers involved and science in general. Moreover, reproducibility ensures the seamless deployment and long-term usability of applications.

However, findings suggest that most research is not reproducible. A Nature survey, for instance, revealed that 70% of researchers could not reproduce another researcher's experiments, while over 50% could not reproduce their own research ([76]). [77] investigated the reproducibility status of 400 papers from the IJCAI and AAAI conference series and concluded that only approximately 25% of the variables required for reproducibility were adequately documented. A systematic and transparent reporting is essential to support reproducibility.

In the rapidly evolving field of neuroergonomics and Brain-Computer Interface (BCI) applications, the need to ensure the reproducibility of research findings is high. Most neuroergonomic and BCI applications require artificial intelligence (AI) and machine learning (ML) technologies to identify patterns in brain signals with the aim of decoding user's intentions or distinguishing mental states. For example, these techniques are used in passive BCIs to evaluate mental workload in real time and adapt the tasks using the estimated workload, which could be beneficial for monitoring and supporting professionals whose work requires high focus. It can also aid in selecting alternatives for human-computer interaction (HCI) systems that induce the least amount of load on the users. Although the evaluation of mental workload from EEG signals is extensively studied in the literature ([78]), many challenges remain to be addressed towards developing real-life applications. These challenges include ability to generalize across subjects ([79]), across sessions or over time ([80], [1]), across tasks and across

contexts ([4], [81], [82]). A systematic and reproducible approach can foster a more collaborative research environment, enabling more effective and rapid solutions to these challenges.

Despite increased attention for reproducibility in the literature, definitions of reproducibility remain unclear and even conflicting ([83]). In this paper, we adopt the definition from (p.1645) [77] as expressed for reproducibility in AI, which was set forth as *"Reproducibility in empirical AI research is the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team."*. Achieving reproducibility necessitates documenting research at a certain level of detail. [77] grouped documentation into method, data, and experiment categories. They also defined three levels of reproducibility, namely R1: Experiment Reproducible, R2: Data Reproducible, and R3: Method Reproducible. R1 reproducibility necessitates sharing all three documentation categories. In R1, the results are expected to be the same, except for minor differences due to hardware changes, as the same implementation is executed on the same data. R1 corresponds to fully reproducible research ([84]) and technical reproducibility ([85]). As the reproducibility level increases from R1 to R2 and R3, the generalizability of the models increases, and documentation requirements decrease at the cost of reduced transparency. The generalizability of a model is the degree to which the outcomes of a study are applicable to diverse contexts or populations. R2 reproducibility is attributed when an alternative implementation of the method is executed on the same data, hence requiring the openness of method description and data but not the scripts, and it is generalizable to alternative implementations of the method. Finally, R3 reproducibility is expected to yield the same results with alternative implementations on different data, thus necessitating only the method documentation. Obtaining similar performance in this case is a step in concluding that the improvement of research was made possible by the proposed method, and the method is generalizable. It should be noted that reproducibility does not necessarily guarantee accuracy. Even if the results are not favorable, the study can be reproducible. Moreover, there is not a single best solution for a given problem, which is another reason for detailed reporting ([86]).

In this study, we propose guidelines considering full (R1) reproducibility with the aim of maximum transparency, enabling the generation of the same results with the same implementation and on the same data. This level of sharing can be tailored for R2 reproducibility level by leaving out the reporting of the experiment and for R3 reproducibility level by leaving out the reporting of both the experiment and data. These approaches reduce the degree of reproducibility but are steps toward generalizable solutions. In cases where the scripts or data are not made available, authors need to be willing to assist other researchers in constructing the baseline ([87]).

While guidelines for the reproducibility of machine learning and EEG studies exist independently in the literature, there is a lack of integrated guidance covering both. EEG guidelines primarily emphasize standardized procedures for data collection, preprocessing, sharing, and statistical analysis. Recommendations for machine learning stress best practices in feature engineering, modeling, and evaluation and highlight code transparency and dataset availability. The necessity to connect these guidelines becomes apparent with the rising number of publications employing machine learning on EEG data, combined with the already mentioned challenges in the generalizability of EEG-based mental state estimations across subjects and contexts or over time. In the current manuscript, we aim to close this gap in the literature and combine a reproducible and standardized ML pipeline with EEG guidelines with a focus on

the classification of mental workload. Based on previous work, we establish guidelines and a checklist for reproducible EEG machine learning. Using this checklist, we systematically assess to what extent studies currently adhere to this checklist.

To scope our research, we chose to focus on workload recognition since it represents a substantial and relatively well-defined sub-area of mental state monitoring and passive BCI. In fact, workload and multitasking emerged as the most common mental state or process, according to the survey conducted by [74].

The proposed guidelines and checklist have the potential to be applicable to most other types of EEG ML mental state assessment studies. However, the specific nuances of each domain must be considered during implementation.

Our manuscript is outlined as follows. In Section 3.3, we first introduce the current reproducibility status in machine learning and explain the CRISP-DM methodology, which is a commonly used standard for data mining machine learning projects. Then, we present the reproducibility efforts in EEG studies (Section 3.4) and systematically review papers that studied reproducibility in mental workload prediction using EEG (Section 3.5). In Section 3.6, we combine the findings in the literature with our contributions and propose guidelines for a reproducible EEG machine learning pipeline that is incorporated into the CRISP-DM phases. Following from these guidelines, we then create a compiled checklist of the requirements for reproducibility. In Section 3.7, adhering to the proposed checklist, we assess the current reproducibility status of machine learning models that utilize EEG to measure mental workload based on a comprehensive systematic literature review. We performed both systematic literature reviews following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram ([88]). Finally, Section 3.8 is allocated for the discussion.

3.3 Reproducibility in Machine Learning

In research fields where machine learning solutions are applied, the challenge of reproducibility is prominent. Independent researchers often struggle to replicate the same results solely based on information provided in publications ([89], [76], [77]). In light of recent discussions on the reproducibility crisis, efforts to examine reproducibility in publications and introduce guidelines or checklists have expanded. Organizations and academic publishers have developed reproducibility checklists to ensure that research incorporates a minimum set of essential information and statistical checks, promoting openness in order to transparently report reproducible work ([90]). Leading journal editors, funding agencies, and scientific leaders collaboratively established a comprehensive set of Principles and Guidelines in Reporting Preclinical Research in June 2014 ([91]), and a considerable number of journals have agreed to support it. These principles include rigorous statistical analysis and transparency in reporting together with a proposed set of key information and data and material sharing. Academic organizations have also introduced checklists to promote reproducibility in machine learning studies. For example, [92] generated "The Machine Learning Reproducibility Checklist" which was used in NeurIPS 2019. This checklist includes items for models and algorithms, theoretical claims, and figures and tables. Authors emphasize significant cultural and organizational changes

besides code submission policy or a checklist to achieve reproducibility. As discussed in the previous section, [77] curated a checklist to investigate the status of reproducibility. Following specific guidelines facilitates a systematic process for conducting reproducible research.

The most widely used methodology for structuring data mining machine learning projects is the CRISP-DM. Introduced in 2000, CRISP-DM is a baseline process model to define and standardize data science life cycle in industry ([93]). This iterative process comprises six phases, each of which is briefly explained below. We use these phases to structure our checklist in Section 3.6.

1. Business Understanding: The initial phase aims to identify business objectives, metrics, and success criteria for subsequent model evaluation. Additionally, it involves defining and planning available resources, as well as establishing strategies to mitigate potential project risks throughout the project lifecycle. In addition to these fundamental project management activities, data mining objectives and corresponding technical success criteria are determined during the business understanding phase. Finally, a project plan is devised for each subsequent phase of the project, ensuring a cohesive and strategic approach.

2. Data Understanding: This phase consists of the tasks of data collection, data description, data exploration, and data quality verification. Data collection adheres to established best practices within the relevant domain, with a clear presentation of data definitions, types, and additional requirements. This phase also entails the examination of data for cleanliness, addressing issues like missing values, noise, outliers, and data imbalance. In this phase, data is understood, and subject matter knowledge is acquired so that each member of the project has a common ground on terminology and domain knowledge. Moreover, future decisions on data preparation, modeling, evaluation and deployment can be made informed only if the context specific to the domain is well understood.

3. Data Preparation: This phase involves organizing data for modeling purposes. Tasks encompass data selection based on goals and limitations, which may include technical or quality considerations. Additionally, this phase includes data cleaning, filtering, and the creation of new attributes or samples through data transformation, augmentation, and integration from multiple sources. Feature engineering and selection are also integral components of this phase.

4. Modeling: The modeling phase starts with the selection of an appropriate method tailored to the specific problem. The rationale behind this selection and any underlying modeling assumptions need to be documented. Reasons to select an algorithm may be related to data and problem characteristics or may arise from some constraints such as development time or hardware limitations. Certain methods incorporate feature selection, which can be another factor to take into account. After the selection of the modeling technique, test design is performed, and model building is initiated. As models are developed, they are assessed and ranked based on predefined evaluation criteria, also taking into account the business success criteria when possible. Model parameters are adjusted iteratively based on these evaluations until a satisfactory model is achieved.

5. Evaluation: The model's compliance to predefined business objectives is assessed in this phase rather than the model performance that was considered in the previous modeling phase. Testing the models in deployment environments can also be anticipated. If the results prove to be insufficient, it may be necessary to revisit earlier phases. This could entail fine-tuning the hyperparameters, exploring alternative algorithms, or reevaluating data preparation and conducting more comprehensive data exploration. Upon achieving satisfactory results in the evaluation phase, a final review of the process is necessary to address any potential oversights. Favorable outcomes from this review pave the way for the subsequent deployment stage.

6. Deployment: At the beginning of this phase, a strategy for deployment is developed. This phase highlights the significance of generalization, as the system or solution is implemented in real-world settings. Here, inference is done on novel data which was never encountered by the model previously. The model's ability to adapt to various scenarios and different users is put to test. Constructing and thoroughly testing the deployment environment is a crucial component of this phase. Furthermore, this phase involves meticulous planning for the continuous monitoring and maintenance of the system. Considering the evolving nature of businesses, model drift may occur, necessitating the retraining of the model with recent data to capture updated business aspects.

Various other data process models besides CRISP-DM are available, such as "Knowledge Discovery in Databases (KDD)" ([94]), "Sample, Explore, Modify, Model, Assess (SEMMA) [1]" and "Team Data Science Process (TDSP) [2]" are available. We chose to use CRISP-DM not only because it is widely adopted ([95]), but also because its phases align well with EEG processing and the machine learning pipeline and because it covers "Business Understanding" and "Deployment" phases, which are necessary to build applications. "Business Understanding" and "Deployment" phases are not included in KDD or SEMMA models. While phases of CRISP-DM and TDSP are similar, CRISP-DM incorporates more detailed phases related to data processing, modeling and evaluation which are fundamental steps for conducting machine learning studies using EEG.

3.4 Reproducibility in EEG

Reproduction of EEG studies comes with challenges, some of which are inherent to scientific research, while others arise from the nature of EEG data. Variations in data collection settings, such as the environment, electrode placement, or online filters, can lead to differences in results. Individuals differ in terms of anatomical and neurophysiological characteristics. Order of preprocessing steps and a large number of parameters that are used within different preprocessing methodologies can cause large differences ([96]). These problems can be mitigated through systematic and transparent reporting. In EEG research, there is a considerable number of publications that aim to standardize data formats, data collection methodologies, data analysis (particularly statistical analysis and preprocessing), and data sharing.

[1] https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

[2] https://learn.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle

The Brain Imaging Data Structure (BIDS) standard was developed to standardize MRI datasets by defining file structure, format, and naming conventions as well as guidelines for presenting metadata ([97]). [98] established EEG-BIDS to introduce this standard to the EEG domain. Specific to EEG data, they recommended the European Data Format (EDF) and the BrainVision Core Data Format, alongside allowing two unofficial data formats due to their common usage and to ease adoption of EEG-BIDS: EEGLAB's format (".set" and ".fdt" files) and the Biosemi format (".bdf").

In 2014, a committee appointed by the Society for Psychophysiological Research reported comprehensive guidelines for studies using EEG and MEG with a detailed checklist for reporting ([99]). The covered topics are hypotheses, participants, recording characteristics and instruments, stimulus and timing parameters, data preprocessing, measurement procedures, figures, statistical analysis, spectral analysis, source-estimation procedures, Principle Component Analysis (PCA) and Independent Component Analysis (ICA), multimodal imaging, current source density and Laplacian transformations, and single-trial analyses. The Organization for Human Brain Mapping (OHBM) neuroimaging community (Committee on Best Practices in Data Analysis and Sharing (COBIDAS) MEEG - where MEEG refers to MEG and EEG) compiled best practices for data gathering, analysis, and sharing ([86]). Recommendations encompassed MEEG data acquisition and data analysis terminologies, definitions, and basic experimental attributes to include in an article. They also listed MEEG preprocessing and MEEG connectivity modeling parameters to be reported and their impact on reproducibility. The authors also state the importance of a dynamic guideline, which is to be adapted as new technology and methods arise. Similarly, [100] included the most commonly used clinical EEG terms and proposed a standardized and structured EEG report form.

[74] created an overarching experiment model that provides a formal structure for presenting HCI research using brain signal data to enhance reproducibility and reusability. They further conducted statistical analysis to understand reporting structures and identified reporting gaps for 110 papers from dedicated HCI conferences or journals. The recommendations and discussions on future challenges offer valuable insights for the advancement of HCI practices. While the focus of this publication was on HCI, the aspects they list are mostly applicable to EEG ML studies in general.

We refer to these established EEG guidelines to develop a compiled checklist in Section 3.6.

3.5 Reproducibility in Mental Workload Studies using EEG

In this section, we present the outcomes of our systematic literature review on reproducibility studies related to EEG and mental workload. Our aim is to determine to what extent studies that use EEG to measure mental workload have focused reproducibility.

3.5.1 Literature Search Strategy

Figure 3 shows the search strategy. In phase I, we conducted a search using specific terms across Scopus, Web of Science, ACM Digital Library (ACM DL), and Pubmed databases to

assess the current state of reproducibility in this field. We searched in titles, abstracts, and keywords with the following search term: *("Reproducibility" OR "Replicability" OR "Generalizability") AND "EEG" AND ("Workload" OR "Cognitive Load" OR "Mental Effort" OR "Mental Load")* in February 2024. The search was constrained to a publication year up to and including 2023 at the latest.

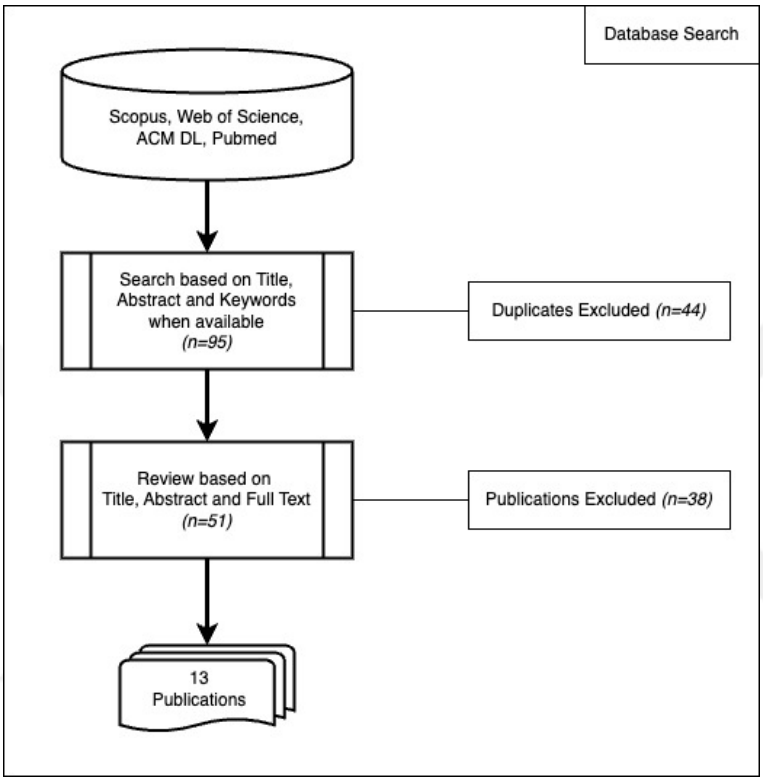


Figure 3: Search Strategy for Literature Review 1

3.5.2 Eligibility Criteria

In phase II, we considered a publication relevant according to the following inclusion criteria:

a) has a focus on mental workload

b) has a focus on or uses EEG data

c) has a direct focus on reproducibility or evaluates a method across different settings, for example, in different tasks or at different times

3.5.3 Analysis of the Studies

Search in the databases produced 95 publications in total. The Scopus search produced 45 articles. Web of Science and ACM DL search yielded seven and one indifferent results, respectively. PubMed search produced 42 publications, six of which were distinct from the ones already found. As a result, we had 51 unique articles from these four databases. Only 13 of these 51 were considered relevant according to the eligibility criteria. This is an indication that few studies focus on reproducibility in the domain of mental workload classification from EEG signals.

Among the relevant thirteen studies, reproducibility was demonstrated for different electrode configurations and preprocessing pipelines ([101]), for different settings of 2D and 3D environments ([102]), for a larger number of participants ([103]), for different tasks ([104], [105], [106]), and over time ([107], [108], [109] and [34], [110], [111], [1]). [34] also tested their findings on separate tasks to check cross-task performance and finally on data from a new participant to observe cross-subject performance.

3.6 Establishing Reproducible EEG Machine Learning Pipelines: Guidelines and Checklist Compilation

The objective of this section is to establish guidelines for constructing reproducible EEG machine learning pipelines. We compiled and tailored established guidelines from the reproducibility literature in both the EEG and machine learning research domains, supplemented these by our own contributions and structured the guidelines following CRISP-DM phases. Below, we discuss the guidelines per CRISP-DM phase. Table 30 outlines the finally resulting, complete list of checklist items and related research steps. Items on the checklist marked with a "†" rather reflect best practices or are related to generalizability, while those that are not marked directly affect reproducibility.

Researchers intending to use these guidelines should adapt them according to their specific methods if they are not covered. Moreover, applied methods should adhere to best practices and guidelines outlined in the relevant literature. Considering how an independent researcher can replicate the analysis or develop the same models by using only the content provided in the publications and supplementary materials is important. This requires a systematic approach during both the research process and the publication phase. As the field evolves and new methods emerge, any new essential information should be incorporated to align with the aforementioned focus on reproducibility. To achieve a comprehensive understanding necessary for replicating the results, a thorough comprehension of the methodologies employed is needed rather than relying solely on the direct execution of open-source code.

As a general approach for all phases, sharing of scripts and properties of the computation environment is required for complete reproducibility ([112]). Text in code is expected to be human-readable, with necessary explanations provided in the comments. Reproducible code practices such as PEP-8 ([113]) are suggested to use. Additionally, open-sourcing raw or at least preprocessed data with the definition of data and data structure should be ensured for full

reproducibility. Shared resources ought to be easily accessible, and permanent access should be preferred.

3.6.1 Business Understanding

In the context of machine learning research utilizing EEG data, stating the problem, including specific research questions or hypotheses and corresponding predictions ([99]), along with the related assumptions and literature provides a clear foundation and facilitates choosing the proper methodologies. A full grasp of the research problem, as well as the associated terminologies, is required to accomplish this phase. For this aspect, the following items are included in the checklist: "Problem/Scope statement" and "Related literature".

3.6.2 Data Understanding

This phase encompasses data collection, data description, data exploration, and data quality verification. Data collection and experiment design constitute a huge component of machine learning research with EEG data. Therefore, to better capture this important and multifaceted process, we divide the "Data Collection" task of this phase into multiple research steps, namely, "General", "Participant Selection", "Experimental Setup", "Experimental Task Information", "Task-free Recordings", "Behavioral Measures", "Subjective Measures" and "Labeling". Table 1 and Table 2 show the items of the checklist grouped by data collection research step together with their main reference. We mostly used items from [86], [99] and [74].

For recordings during real-life applications, such as driving or flying an airplane, marking recordings with respect to events would be more appropriate than using stimuli. In these cases, intertrial intervals or stimulus properties would not be applicable, and the checklist needs to be tailored to reflect such nuances.

EEG data is heavily dependent on the experimental settings and also the user's state of mind. Collecting additional data, such as subjective assessments and behavioral data, would be beneficial to mitigate the effects of these dependencies. These additional data can be instrumental during the evaluation of results, can be directly integrated into the models to normalize the data, or serve as separate input.

Labels should be clearly defined - e.g., whether workload labels are derived from task difficulty, subjective measures, or judgments by subject matter experts.

We consider data description, data exploration, and data quality verification tasks of CRISP-DM "Data Understanding" phase under the "Analysis" research step. For this step, we include in the checklist "Recording length", "Statistical analyses to justify the number of trials and the number of participants" ([86]), and "Statistical Analysis for descriptives of the collected measurements" ([99]). Exploratory data analysis to check for quality and descriptive analysis to better understand the data is advised to make informed decisions in the upcoming phases.

Table 1: Checklist Items related to Data Collection - 1

Research Step	**Compiled Checklist**	**Reference**
General	Dataset (name if public or private)	
Participant selection	Number of participants	
	Participant recruitment method (e.g., direct mailing, advertisements)	[86] [74]
	Participant sampling strategy (that constrain inclusion to a particular group/including population from which the participants were sampled)	[86] [74]
	Age of participants	[86] [99] [74]
	Gender of participants	[86] [99] [74]
	Education level of participants	[99]
	Medications taken by the participants	[86]
	Prior/Current illness of participants	
	Information on sleep deprivation	[100]
	Handedness of participants	
	Consent of participants	[86]
Experimental setup	Type of EEG sensor/device (including make and model)	[99] [74]
	Number of sensors	[99]
	Sensor locations	[99] [74]
	Sampling rate	[99] [74]
	Online filters (Type of filter and parameters)	[99]
	Electrode impedance	[99] [74]
	Amplifier characteristics	[99]
	Measurement procedures	[99]
	Recording environment	[86] [74]
	Participant seated or lying down status	[86]
Experimental task information	Task description	
	Characteristics of stimuli	[86] [99]
	Instructions for the task	[86] [74]
	Number of runs and sessions	[86]
	Clear timeline including - Timing of all stimuli/events - Intertrial intervals	[99] [74]
	Software and hardware used for stimulus presentation	[86] [74]

Table 2: Checklist Items related to Data Collection - 2

Research Step	Compiled Checklist	Reference
Task-free recordings	Definition	
	Timing	
	Eyes open vs closed status	[86]
	If eyes open, fixation point usage	[86]
Behavioral measures	Nature of the response	[86]
	Acquisition device and parameters	[86]
	Interface with EEG data and calibration procedures	[86] [74]
	Errors and outliers handling	[86]
Subjective measures	Subjective assessments recorded - Timing - Method	
Labeling	Definition	[74]

Furthermore, we also encourage the sharing of data to facilitate reproduction studies, along with the disclosure of source code/software used for data collection and task execution ([74]) under "Open-sourcing" step. An experiment cannot be reproduced to gather new data if the details of execution and data collection are left out. To prevent having to report every detail, standardized data collection methodologies and experiment software are required. This transparency enables independent labs to conduct the same experiment and replicate the results using their own data. Storing data in a standardized structure, such as EEG-BIDS ([98]), is essential. Sharing of physiological data raises ethical considerations and informed consent of participants for the study, and usage or sharing of their data is obligatory during data collection ([114]).

3.6.3 Data Preparation

This phase entails the tasks of selecting, cleaning, constructing, integrating, and formatting data in accordance with CRISP-DM. These tasks correspond to a large portion of the overall research process, from data preprocessing and feature generation to feature selection and feature transformation. The flow of steps used for preprocessing, feature generation, selection, and transformation should be well-defined ([99]) as well as the methods used and their related parameters. Seeds for random number generators need to be used and reported to prevent randomness in results ([115]).

We put special emphasis on data preprocessing steps, keeping in mind that there is not a common single pipeline and applications vary as well as the implementations and tools ([116], [96], [117], [118], [119]). Therefore, we aim to list the most commonly used techniques in our checklist (Table 30), in no particular order, leaving it to interested parties to tailor the same detailed approach for their own research. We took into account [86], [99], and [74] to list the most used preprocessing methods. More than one preprocessing pipeline can be used, yet consistency in feature generation, selection, and transformation steps is important throughout the study. All algorithms and corresponding parameters should be explicitly defined, and best practices for the applied methods should be followed. For example, [120] reviewed filtering

and explained how to choose the right filter. Keeping track of input and output data, data types, and data size at each research step supports coherence throughout the project.

Feature generation and feature selection are the next steps after preprocessing to prepare data for machine learning ([74]). Features are expected to be defined together with the method and parameters used to construct and select them. The total number of features, as well as the selected features and their number, should be stated. If descriptive or inferential statistics were analyzed, their method and parameters need to be reported ([99]).

Cross-validation is a widely employed technique to enhance model performance and generalizability. While cross-validation is typically performed during the modeling or evaluation phases, the initial step of splitting the data and setting aside a test set to prevent data leakage falls under the data preparation phase. In cases involving models that necessitate hyperparameter tuning, such as deep learning models or other parametric models, the hyperparameters are fine-tuned based on the performance metrics of validation sets. Consequently, an independent, unseen dataset for reporting the model's performance is required since the utilization of the validation set for hyperparameter optimization inherently introduces bias to the outcomes from the validation set. For these cases, it is common practice to divide the dataset into train, validation, and test sets. To ensure an unbiased estimate of the model's performance, the unseen test data should be set aside, excluding it from both model development and assessment until the final reporting stage to demonstrate the model's generalizability and avoid wrongly optimistic performance. Data split needs to return independent sets according to the task at hand to prevent leakage. For example, cross-subject estimation requires a subject-wise split, while cross-session models necessitate a session-specific split. Finding that models' performance cannot be reproduced from one individual to another, or from one session to another, will lead the research community to use other features or develop other types of models that can be generalized or to the conclusion that individual models are required. After the data is split into train, validation, and test sets, any data augmentation, transformation, or normalization should be executed using only the training set parameters to avoid potential data leakage. [121] emphasizes the importance of selecting such parameters separately from the test set and using independent training and test sets as good classification practice. This approach establishes an unbiased common ground for the comparison of different algorithms. As a result, we add "Data split (Method, Parameters)" ([92]), "Separate test set", "Feature transformation (Method, Parameters)", "Feature transformations applied using training data?", "Data Augmentation (Method, Parameters)", and "Data Augmentation applied using training data?" items to our checklist.

Overall, the process to obtain the feature sets should be provided in detail to prevent any gaps when generating them from scratch. Additionally, open-sourcing the scripts or providing the processed data would also mitigate these concerns ([77], [92], [74]).

3.6.4 Modeling

This phase consists of selecting the model, generating the test design, and building and assessing the model. Best practices in the literature need to be followed for these tasks. For example, [122] provides practical recommendations for training deep neural networks.

When selecting a model, one should provide a detailed explanation of the rationale behind the choice and its intended behavior ([77]). In the case of opting for an existing validated method, the report should reference relevant packages, functions, or repositories. Additionally, one should explicitly state model parameters, including the loss function, regularization, other internal settings, and model structure, if applicable. Similar to data preparation, the use and reporting of random number generator seeds for reproducibility and obtaining deterministic results should also be ensured.

Generating the test design is inherent to the training strategy. For parametric methods, hyperparameters of the model and the method for tuning them together with their ranges and number of trials should be reported, including the selected hyperparameters ([92], [74]). After the test design, the upcoming step in the project is model training. The optimization method and its parameters as well as the number of training epochs or iterations ([92]) should be defined at this stage. Additional techniques utilized during training, such as early stopping, need to be stated with the relevant parameters ([122]).

Once test design and model building are completed, generated models need to be assessed technically and compared to choose the best model or models. Evaluation metrics should be defined with their reasoning ([92], [74]) and their chance-level values need to be included. Naive baseline models and naive predictions are important to build, in particular when dealing with class imbalanced datasets. A naive model, which generates the majority class label at all times could imply whether the developed model is useful. Chance-level values can be extracted by using random estimators.

Additionally, attention is required when the model performance is to be compared with baseline parametric models proposed in the literature. Using model parameters as they are would lead to wrong conclusions since they would be tuned specifically to the dataset of the original study. Parameters of baseline models should also be adjusted, if possible, with the methodology provided in the original paper to perform a fair comparison ([123]). This would be possible if the reference study is also reproducible.

When reporting the results, the data split that the metrics are calculated on (train, validation, or test) must be explicitly stated. When the dataset is imbalanced, metrics other than accuracy should be used. Using confusion matrices is encouraged to identify regions where the model does not fit completely. This is specifically relevant for machine learning with EEG since these datasets are usually small. Using a combination of complementary metrics rather than relying on a single metric helps a more extensive understanding of machine learning performance ([124]). Moreover, categories of data can be used to break down performance measures to understand the results in different regions ([123]).

Machine learning models are prone to computational environment changes; therefore, a description of the computing hardware and software infrastructure needs to be presented together with the dependencies, including external libraries and their versions or virtual environment with all dependencies ([77], [92]). It would be beneficial to test whether the same set of packages works on other related environments, such as on a different device or operating system, before moving on to deployment.

In conclusion, similar to the approach in the Data Preparation phase, the process for modeling should be described in detail so that an independent researcher can reproduce the results. Open-sourcing the modeling scripts and providing the trained models are also encouraged for details that may have been left out or to mitigate misunderstandings from written text.

Data Preparation and Modeling phases are managed iteratively since the two phases affect each other closely.

3.6.5 Evaluation

In this phase, results are discussed in line with the research questions or hypotheses stated in the Business Understanding phase.

Due to high efforts required in EEG data collection, the number and variety of participants are usually low. Therefore, the distribution of the whole population cannot be normally captured equally within data splits. Model selection, assessment, and comparison need to be performed on validation sets since training sets are used for model training. After finding the best model, to generate an unbiased estimate of the performance, an independent test set should be used for reporting. This test set should not be included in model development or selection. Results on the test set need to be presented to check for generalizability and prevent misleading optimistic findings. For the most reliable results, nested cross-validation is recommended ([86]). Finally, statistical analysis should be performed to ascertain the significance of results. For comparison of classifiers, appropriate statistical tests need to be used ([125]), such as (non-parametric) Wilcoxon signed ranks and the Friedman test ([126]).

Recently, [127] emphasized the increase in computational resources of machine learning research as larger models are trained with larger amounts of data for performance improvement. They advise researchers to report training time and sensitivity to hyperparameters. Moreover, they are expected to prioritize computationally efficient hardware and algorithms and be mindful of energy sources powering their computing. [128] proposed Green AI, where the focus of research would be efficiency rather than accuracy. This approach aims to reduce the environmental impact of model training and the entry barriers to the field, both caused by increased computational resource requirements. Computational resources for training, such as model size, run time or power consumption, and carbon emissions, need to be reported to promote responsible AI that is energy-efficient. Releasing code and data or models also helps reduce carbon emissions as it will reduce the energy spent on replicating the results by other researchers [129]). Open-sourcing the scripts for evaluation is also encouraged to perform appropriate comparisons.

After the evaluation of results in this phase, the process is reviewed, and the next steps are determined as to reiterate from the Business Understanding phase or move on with Deployment.

3.6.6 Deployment

While deploying models is essential for real-world applications, current EEG (workload) research tends to focus more on developing new methods for classification or data processing approaches rather than on deployment specifics. When models are deployed, it is important to provide details about the deployment hardware, software infrastructure, and dependencies. Additionally, reporting required computational resources for inference, such as inference time, power consumption, and carbon emissions, is necessary. Challenges arise with the growing sizes of recent models, like those in natural language or image processing, as they may pose difficulties in deployment due to constraints on size or cost in practical applications. Deployment techniques, including low-rank factorization or model quantization, along with computational optimization methods, can be employed to address these challenges ([130]). If such techniques are used, it is important to report the methods and parameters involved.

Interfaces and schematic or sample views need to be presented for a good understanding of the application. The performance of the model is required to be verified to yield sufficiently similar results in both development and production environments given identical input.

In an optimal scenario, for results to be deemed appropriate for a real-world application, developed models should exhibit consistent and acceptable performance across diverse subjects and various time frames. If the conditions permit, it would be best to model and evaluate these aspects to demonstrate the generalizability before deployment. Between the development and deployment environments, data flow must be consistent end-to-end, from preprocessing the data to generating the features and inferring the results. Moreover, after deployment, performance needs to be monitored and maintained continuously to prevent any problems and model drift. The method to achieve this monitoring can be reported for transparency.

3.7 Reproducibility in Machine Learning Models to Predict Mental Workload using EEG

We performed a comprehensive literature review to assess the extent to which the aspects in the checklist (Table 30) have been implemented within the domain of mental workload classification studies utilizing EEG data. Although this section is dedicated to assessing the reproducibility status of mental workload classification using EEG, the guidelines and checklist have the potential to be applicable to most other EEG machine learning studies.

3.7.1 Literature Search Strategy

Figure 4 shows our search strategy. During phase I, we searched in titles, abstracts, and keywords in Scopus, Web of Science, ACM Digital Library (ACM DL), and Pubmed databases with the following search term: *"Machine Learning" AND "EEG" AND ("Workload" OR "Cognitive Load" OR "Mental Effort" OR "Mental Load")* in September 2023. We did not include limits on the language at this stage. We also searched in the "Frontiers in Neuroergonomics" journal from the webpage as at the time of the search, this journal was not yet

indexed in the aforementioned databases and its scope directly entails our topic. We searched in full-text for this journal because a search based on only titles, abstracts, and keywords was not possible.

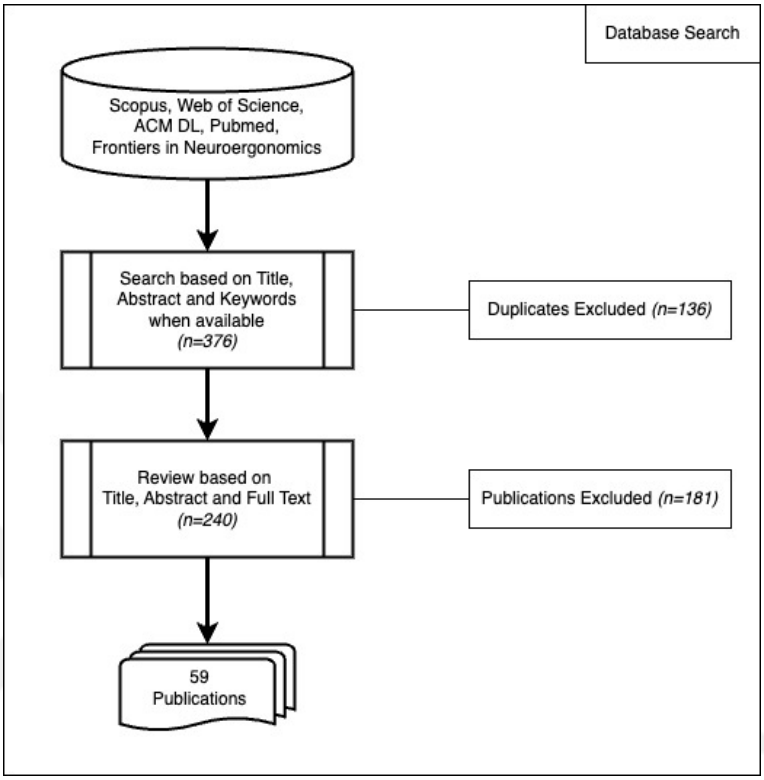


Figure 4: Search Strategy for Literature Review 2

3.7.2 Eligibility Criteria

In phase II, we selected publications according to the criteria given in Table 3. All inclusion criteria are domain-specific, and all exclusion criteria are generic.

Table 3: Inclusion and Exclusion Criteria

Inclusion criteria		**Exclusion criteria**	
I1	Models mental workload/cognitive load	E1	Original language is not English
I2	Applies machine learning	E2	Full text is not available
I3	Uses only EEG data features	E3	Review paper
I4	Uses at least one classifier		
I5	Uses EEG data at the sensor level		

3.7.3 Analysis of the Studies

The search in the databases produced 376 publications in total. The Scopus search produced 210 articles. Web of Science yielded 73 papers, but none of them were different. Only one out of ten results from ACM DL was distinct. Pubmed search produced 61 results, 7 of which were new. The search in Frontiers in Neuroergonomics generated 22 results. As a result, we had 240 unique articles from these four databases when duplicates were removed. From these, 59 publications met the eligibility criteria (4). In phase III, we inspected these publications in detail . We showcase the status of reproducibility among the selected papers by following the checklist given in Table 30. Results of phase III are presented in Section 3.7.4.

Table 4: List of Publications

Authors	Year	Title	Journal/Conference Name
[131]	2023	Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight	Scientific Reports
[132]	2023	Using EEG signals to assess workload during memory retrieval in a real-world scenario	Journal of Neural Engineering
[133]	2023	Recognition of Pilot Mental workload in the Simulation Operation of Carrier-based Aircraft Using the Portable EEG	ACM International Conference Proceeding Series
[134]	2023	Modeling of Brain Cortical Activity during Relaxation and Mental Workload Tasks Based on EEG Signal Collection	Applied Sciences (Switzerland)
[135]	2023	A Mental Workload Classification Method Based on GCN Modified by Squeeze-and-Excitation Residual	Mathematics
[136]	2023	Cross-Task Mental Workload Recognition Based on EEG Tensor Representation and Transfer Learning	IEEE Transactions on Neural Systems and Rehabilitation Engineering
[137]	2023	Decoding EEG Signals with Visibility Graphs to Predict Varying Levels of Mental Workload	2023 57th Annual Conference on Information Sciences and Systems, CISS 2023
[138]	2023	On Time Series Cross-Validation for Deep Learning Classification Model of Mental Workload Levels Based on EEG Signals	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[139]	2023	Inter-subject cognitive workload estimation based on a cascade ensemble of multilayer autoencoders	Expert Systems with Applications
[140]	2023	Cognitive load detection using circulant singular spectrum analysis and Binary Harris Hawks Optimization based feature selection	Biomedical Signal Processing and Control
[141]	2022	Optimal classification of N-back task EEG data by performing effective feature reduction	Sadhana - Academy Proceedings in Engineering Sciences
[142]	2022	Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment	Frontiers in Artificial Intelligence
[143]	2022	Evaluation of a New Lightweight EEG Technology for Translational Applications of Passive Brain-Computer Interfaces	Frontiers in Human Neuroscience
[144]	2022	Self-Paced Dynamic Infinite Mixture Model for Fatigue Evaluation of Pilots' Brains	IEEE Transactions on Cybernetics

Table 4: cont.

[145]	2022	An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload	Frontiers in Neuroinformatics
[146]	2022	Assessing Distinct Cognitive Workload Levels Associated with Unambiguous and Ambiguous Pronoun Resolutions in Human–Machine Interactions	Brain Sciences
[147]	2022	Cognitive load detection using Binary salp swarm algorithm for feature selection	2022 IEEE 6th Conference on Information and Communication Technology, CICT 2022
[148]	2022	EEG based Mental Workload Assessment by Power Spectral Density Feature	2022 IEEE International Conference on Mechatronics and Automation, ICMA 2022
[149]	2022	Analysis of Mental Task Ability in Students based on Electroencephalography Signals	SPICES 2022 - IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems
[150]	2022	Real-Time EEG-Based Cognitive Workload Monitoring on Wearable Devices	IEEE Transactions on Biomedical Engineering
[151]	2021	Driving-induced neurological biomarkers in an advanced driver-assistance system	Sensors
[152]	2021	Cognitive performance detection using entropy-based features and lead-specific approach	Signal, Image and Video Processing
[153]	2021	EEG Fingerprints of Task-Independent Mental Workload Discrimination	IEEE Journal of Biomedical and Health Informatics
[154]	2021	Prediction and Detection in Change of Cognitive Load for VIP's by A Machine Learning Approach	3rd IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2021
[155]	2021	Tracking of mental workload with a mobile eeg sensor	Sensors
[156]	2021	FINE-GRAINED and MULTI-SCALE MOTIF FEATURES for CROSS-SUBJECT MENTAL WORKLOAD ASSESSMENT USING BI-LSTM	Journal of Mechanics in Medicine and Biology
[157]	2021	Brain–computer interface for assessment of mental efforts in e-learning using the nonmarkovian queueing model	Computer Applications in Engineering Education
[158]	2021	Detecting Mental Workload in Virtual Reality Using EEG Spectral Data: A Deep Learning Approach	Proceedings - 2021 4th IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2021
[159]	2021	The Cognitive Load Evaluation Based on EEG with K-Nearest Neighbor Algorithm	ISPACS 2021 - International Symposium on Intelligent Signal Processing and Communication Systems: 5G Dream to Reality, Proceeding
[160]	2021	Mental Effort Estimation by Passive BCI: A Cross-Subject Analysis	Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS
[161]	2020	Deep Learning Identifies Brain Cognitive Load Via EEG Signals	2020 IEEE 17th India Council International Conference, INDICON 2020

Table 4: cont.

[162]	2020	Estimating the cognitive load in physical spatial navigation	2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020
[163]	2020	Mental Workload Estimation Using EEG	Proceedings - 2020 5th International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2020
[164]	2020	Feature selection model based on eeg signals for assessing the cognitive workload in drivers	Sensors
[165]	2020	Ternary-task convolutional bidirectional neural turing machine for assessment of EEG-based cognitive workload	Biomedical Signal Processing and Control
[166]	2019	A three-class classification of cognitiveworkload based on EEG spectral data	Applied Sciences (Switzerland)
[167]	2019	Individual-specific classification of mental workload levels via an ensemble heterogeneous extreme learning machine for EEG modeling	Symmetry
[168]	2019	EEG based mental workload assessment via a hybrid classifier of extreme learning machine and support vector machine	Chinese Control Conference, CCC
[169]	2019	Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework	Neurocomputing
[170]	2019	Spectral and Temporal Feature Learning with Two-Stream Neural Networks for Mental Workload Assessment	IEEE Transactions on Neural Systems and Rehabilitation Engineering
[171]	2019	EEG-Based Workload Index as a Taxonomic Tool to Evaluate the Similarity of Different Robot-Assisted Surgery Systems	Communications in Computer and Information Science
[172]	2019	On the Use of Machine Learning for EEG-Based Workload Assessment: Algorithms Comparison in a Realistic Task	Communications in Computer and Information Science
[173]	2019	Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment	IEEE Transactions on Neural Systems and Rehabilitation Engineering
[174]	2018	Investigating the generalizability of EEG-based cognitive load estimation across visualizations	Proceedings of the 20th International Conference on Multimodal Interaction, ICMI 2018
[175]	2018	Quantifying cognitive workload in simulated flight using passive, dry EEG measurements	IEEE Transactions on Cognitive and Developmental Systems
[176]	2018	Towards robust neuroadaptive HCI: Exploring modern machine learning methods to estimate mental workload from EEG signals	Conference on Human Factors in Computing Systems - Proceedings
[177]	2018	Deep Convolutional Neural Networks for mental load classification based on EEG data	Pattern Recognition
[178]	2018	Classification of EEG signals for cognitive load estimation using deep learning architectures	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Table 4: cont.

[179]	2018	Mental workload estimation from EEG signals using machine learning algorithms	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[180]	2017	Mental workload classification in n-back tasks based on single-trial EEG	Yi Qi Yi Biao Xue Bao/Chinese Journal of Scientific Instrument
[181]	2017	Cross-session classification of mental workload levels using EEG and an adaptive deep learning model	Biomedical Signal Processing and Control
[182]	2017	Monitoring cognitive workload in online videos learning through an EEG-based brain-computer interface	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
[183]	2017	Classification of EEG features for prediction of working memory load	Advances in Intelligent Systems and Computing
[184]	2016	Adaptive automation triggered by EEG-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment	Frontiers in Human Neuroscience
[185]	2015	Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks	Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS
[186]	2015	Towards an effective cross-task mental workload recognition model using electroencephalography based on feature selection and support vector machine regression	International Journal of Psychophysiology
[187]	2015	Cognitive Workload Assessment Based on the Tensorial Treatment of EEG Estimates of Cross-Frequency Phase Interactions	Annals of Biomedical Engineering
[188]	2012	Temporal factors of EEG and artificial neural network classifiers of Mental Workload	Proceedings of the Human Factors and Ergonomics Society
[189]	2008	Feasibility and pragmatics of classifying working memory load with an Electroencephalograph	Conference on Human Factors in Computing Systems - Proceedings

3.7.4 Reproducibility Analysis

By inspecting the selected 59 publications based on the full text according to the guidelines presented in Table 30, we aimed to establish which elements of the guidelines in our list are commonly adhered to, and which elements of the guidelines in our list are commonly ignored in machine learning research that models mental workload using EEG.

1. Business Understanding: Related to the Business Understanding phase, we considered "Problem/Scope statement" present when the objective of the paper was stated in the Abstract or Introduction sections. Additionally, if the problem was described in the Introduction with references or a separate "Literature Review" section, "Related literature" was marked as present. According to our analysis, all publications defined the problem and presented relevant literature, although the extent of their coverage differed.

2. Data Understanding: Checklist items regarding Participant Selection, Experimental Setup, Experimental Task Information, Task-free recordings, Behavioral Measures, Subjective Measures, Labeling, and Analysis research steps are evaluated for the selected papers. Table 5 shows the reported percentages of the checklist items related to the participants, experiment, labeling, and statistical analysis. Table 6 shows the reported percentages of additionally collected data, namely, task-free recordings, behavioral and subjective measures.

   Sixteen of the publications used an open dataset. When a publication referenced an open dataset, we checked the relevant publication to analyze if the checklist items were reported. Additionally, when more than one dataset was used, we marked an item present if it was included for at least one of the datasets. We considered the "Education level of participants" provided if the "Participant sampling strategy" stated information about education level, for example, graduate students or pilots. "Prior/Current illness of participants" was marked as reported if it was explicitly stated or the participants were stated to be healthy. Participants identified as healthy were presumed to be free from medication use.

   "Amplifier characteristics" were considered present when an amplifier model or amplifier properties such as channel number or time constant were specified. "Participant seated or lying down status" was marked as present if it was explicitly stated or it could be inferred from the recording environment or the task.

   "Recording length" was considered given if it was explicitly stated or it could be calculated from given information.

   Characteristics of stimuli were marked as given when it was explicitly stated or it could be inferred from the task description, for example, visual or auditory stimuli. Detailed instructions for the experimental task are required for reproduction. We marked a study to have reported instructions, whether the instructions were related to the experiment execution or physical restraints, such as refraining from movement. Even if most of the studies (80%) reported instructions, capturing all information in reports to enable the execution of tasks by other researchers is hard. Although 73% had their own recording and dataset, only three of them had the raw data available upon request, and two of them had the preprocessed data available. Open datasets and standardized data collection methodologies and experiment settings should be established to overcome most of these challenges. Similarly, open-source codes help to reproduce the methodologies and provide a common baseline for comparisons, yet only two of the publications shared their data processing repositories. In addition to open-sourcing, authors need to be willing to help other researchers perform experiments. Expanding the knowledge base towards generalizable models for real-life applications is possible by achieving a collaborative research environment.

3. Data Preparation: In our analysis, we considered the "flow of algorithms" included even if it was listed in one sentence. Flow of algorithms used to preprocess data, to generate and select features and to develop models were reported for most of the studies (93%). 'Seeds for random number generators' were marked as given when they were stated explicitly or the code was open. Two publications, which shared their codes, were consequently marked as reporting them and one publication provided the seed number.

Table 5: Reported Percentages of Checklist Items in the Data Understanding Phase-1

Research Step	**Checklist Item**	**Percentage Reported (%)**
Participant Selection	Participant recruitment method	8
	Participant sampling strategy	75
	Age of participants	71
	Gender of participants	76
	Education level of participants	59
	Medications taken by the participants	34
	Prior/Current illness of participants	64
	Information on sleep deprivation	7
	Handedness of participants	42
	Consent of participants	66
Experimental Setup	Type of EEG sensor/device Number of Sensors	86 98
	Sensor Locations	85
	Sampling rate	88
	Online filters	15
	Electrode impedance	34
	Amplifier characteristics	31
	Measurement procedures	37
	Recording environment	73
	Participant seated or lying down status	75
Experimental Task Information	Task Description	100
	Characteristics of stimuli	54
	Instructions for the task	81
	Number of runs and sessions	93
	Timing of all stimuli/events	68
	Intertrial intervals	61
	Software and hardware for stimulus presentation	56
Labeling	Definition	98
Analysis	Recording Length	86
	Statistical Analysis to justify the number of trials and number of participants	2
	Statistical Analysis for descriptives of the collected measurements	39

Table 6: Reported Percentages of Checklist Items in the Data Understanding Phase-2*

Research Step	**Usage Percentage (%)**	**Parameter Percentages (%)**
Task-free Recordings	54	Timing: 84 Eyes open or closed status: 69 If eyes open, fixation point usage: 13
Behavioral Measures	51	Acquisition device: 43 Interface with EEG data and calibration procedures: 23 Method for errors and outlier handling: 3
Subjective Measures	32	Timing: 100 Method: 100

*The third column refers to percentages of the subset in the second column

Table 7 shows the status of preprocessing items. The percentage of application of the research steps and the percentage that parameters were reported among them are presented. Similar to performing the experiments, data preparation, and modeling would be best understood by independent researchers when code and data are shared to prevent having to state all parameters in detail.

Feature generation was unclear for three of the publications, and the number of features was not explicitly stated for 29% of the publications. Feature generation method and parameters were not explicitly stated for 32% and 41% of the publications, respectively. Thirty four% of the publications performed descriptive statistics and the method was specified for 85% of them.

53% of the publications performed feature selection, and 60% of those that performed feature selection indicated the number of selected features. 93% of the publications stated method for data splits, and 25% among them listed their parameters in the form of percentages, fold numbers, or session-based splits.

Fifteen% and 42% of the publications provided information about the computing infrastructure and dependencies, respectively. "Dependencies" were marked given even if only one software package or software was stated (e.g., Python, scikit-learn, Tensorflow, EEGLAB (version 14.2.0), MATLAB2019b).

4. Modeling: For the modeling phase, 64% of the studies explain the algorithm used and the motivation to apply it. Sixty-nine% of the publications state the hyperparameters, 56%, 44% and 41% of them report the method for hyperparameter tuning, state ranges of the hyperparameters, and present selected hyperparameters, respectively. Only one of the publications that use methods other than grid search reported the number of trials for hyperparameter tuning. Here, we exclude grid search as the number of trials for it can be deduced from parameter ranges. Detailed information on models or model training, such as loss function, regularization, model structure, optimizer, or number of training epochs/iterations, are not applicable to all models. Therefore, they could only be investigated where applicable. To present the general situation, we extracted their reporting percentages without considering the related models. Optimization method,

Table 7: Reported Percentages of Checklist Items in the Data Preparation Phase*

Research Step	**Usage Percentage (%)**	**Parameter Percentages (%)**
Sensor/Segment Removal	19	Interpolation: 18 Removed sensors: 55
Artifact removal/correction	39	Range of parameters: 13 Types of artifacts identified: 69 Criteria to identify: 30 Number/proportion of removed artifacts: 4 Position of removed artifacts: 0
Signal–noise separation methods	39	Parameters: 17 Number of ICs: 4 How non-brain ICs were identified: 30 How back-projection was performed: 17
Downsampling	17	Method: 40 Parameters: 80
Detrending	5	Method: 100 Parameters: 0
Filtering	76	Filter type: 100 Parameters: 96
Segmentation	80	Method: 100 Parameters: 96
Baseline correction	3	Method: 100 Parameters: 100
Re-referencing	19	Method: 100
Dimensionality Reduction	2	Method: 100 Parameters: 0
Feature Transformation	22	Method: 100 Parameters: 77 Applied using training data?: 31
Data Augmentation	12	Method: 100 Parameters: 43 Applied using training data?: 14

*The third column refers to percentages of the subset in the second column

number of training epochs/iterations, and additional methods used during training were reported for 22%, 24%, and 17 % of the publications, respectively.

All publications except one report the metrics used, and five of them state the chance-level value. Data split that the metrics are calculated on is not clearly explained for 17% of the publications. Additionally, 20% report confusion matrices.

When we consider computational environment and open sourcing, 29% and 44% report computing infrastructure and dependencies, respectively. None of them open-source their trained models and only three of them open-source their code for modeling.

5. Evaluation: Statistical analysis for significance of results was carried out by 42% of the publications, 68% of which also included parameters such as alpha parameter, confidence interval, or p-value. During the Evaluation phase, an unseen test set to report the performance of the model is mandatory for unbiased estimates and to present the generalizability of results. However, 39% of the studies hold out a test set and only 34% report the results on the test set. For EEG modality, setting aside an unseen test set can be difficult considering the limited amount of data and low number of participants. EEG data collection is time-consuming, and it may be difficult to find participants who satisfy the inclusion criteria and are willing to participate in the experiment. To illustrate this, Figure 5 shows a histogram of the number of participants where most studies include 15 or less participants. One publication did not state the number of subjects.

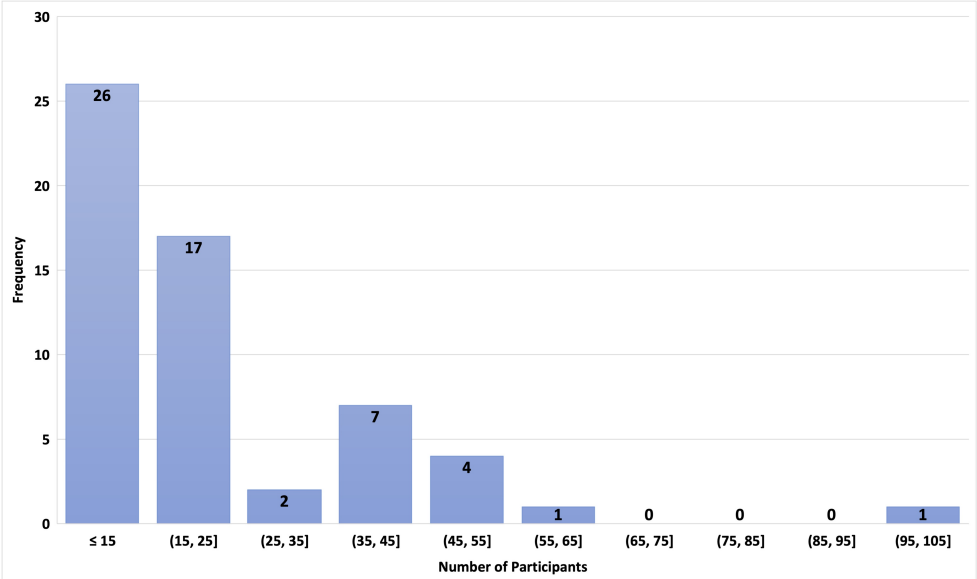


Figure 5: Number of Participants

Few studies (20%) report the computational resources for training such as model size, training or inference times, power consumption, and carbon emissions.We consider these resources reported even when one of these types of data is presented. These aspects are closely related to both the limitations of the deployment environment and sustainable AI.

All publications related their results to the problem statement. Three of them open-sourced their code for evaluation.

6. Deployment: Only one of the publications considered deployment. Deployment techniques, computational resources required for inference, deployment environment, and deployment tests need to be considered after finalizing the model in the development environment. Deployed systems must retain their performance and be reliable, scalable, maintainable, and adaptable ([130]).

3.8 Discussion

This study introduced guidelines, compiled in a checklist aligned with the CRISP-DM framework, for improving the reproducibility of machine learning research utilizing EEG data. A systematic evaluation of EEG mental workload studies shed light on commonly employed strategies, frequently overlooked aspects, and the existing gaps that impede progress toward achieving reproducible science for practical applications.

The key revelation from our analysis is the prevalent limitation in reproducibility across the examined studies. Notably, a significant number of publications fall short in reporting performance on unseen test data, an important aspect that is informative of the model's generalizability. This omission poses a potential problem to the applicability of these models in diverse settings and under varying conditions.

Furthermore, our investigation reveals that only a minority of studies share essential resources, such as data or scripts, crucial for achieving full reproducibility. Given the inherent complexity of capturing every detail in the machine learning pipeline, the open sharing of data and code emerges as a key factor in increasing the credibility of models. This not only builds trust but also helps speed up progress by making it easier to understand new research, saving time on reproducing results, and creating starting points for future work.

A third noteworthy finding is the inadequate reporting of resources essential for training and inference processes. Now that the detrimental environmental effects of AI are becoming increasingly clear, reporting the training and inference times, power consumption, and carbon emissions has become a recommended practice. The inclusion of such information is important for fostering environmentally conscious practices in machine learning research. Deployment techniques to compress models or optimize inference are being developed. With only one study found in our survey specifically addressing deployment considerations, there is an apparent need to study and discuss deployment strategies for EEG classification using machine learning.

Our study has several implications.

Firstly, the introduction of a guideline and checklist, aligned with the CRISP-DM framework, provides a foundational framework for researchers in the field. Adhering to these guidelines will result in a clearer understanding and validation of the methodologies employed, enable the reduction of errors, and improve the credibility and reliability of machine learning studies utilizing EEG data, their authors, and the scientific field as a whole, promoting better scientific practices and accelerated progress.

Secondly, by using the introduced checklist, models can be more fairly compared, ensuring a comprehensive evaluation. With models being compared more fairly, the results and experiments become more transparent and interpretable.

Thirdly, key findings from the reproducibility assessment highlight areas for improvement and future work.

While the present study has contributed valuable insights, there are limitations and promising paths for future research.

Search terms for the systematic literature reviews could be added to enhance coverage and inclusivity. Terms could be expanded to include similar words, such as “Electroencephalography” in addition to “EEG” and “Classification” in addition to “Machine Learning”. Additionally, in the first literature review, not all studies examining reproducibility will have emerged using our terms "reproducibility", "replicability", and "generalizability". Moreover, we focus on the reproducibility status of mental workload estimation studies using EEG. This work could still be extended to include the reproducibility status of EEG studies in general.

Our study focused on mental workload estimation studies. The proposed checklist has the potential to be applied to EEG machine learning studies in general, in particular mental state monitoring in a broader sense. Future work could explore reproducibility of machine learning studies using EEG across various domains, e.g., mental states besides workload, therewith broadening the scope of the reproducibility results and checking in detail for applicability of the proposed checklist across domains. In addition, it would be of interest to examine how reproducibility of different aspects depends on the working domain or expertise of the authors. Mental workload estimation is an interdisciplinary topic. Authors’ background and main expertise likely affect the degree of reproducibility of different aspects, and interdisciplinary teams will likely increase the overall quality of reproducibility.

Transparency and explainability are now integral components of Responsible AI, and are as such requested in various standards, recommendations, and regulations, including the EU AI Act, OECD AI principles, and ISO/IEC 42001:2023. These principles are also catalyzing the acceleration of reproducible studies in the field of machine learning. In the future, the proposed guidelines could incorporate Responsible AI aspects, such as the growing significance of explainability features in model development. These features are increasingly becoming essential, even mandated, in the regulations of certain countries. Further research is needed to explore and address deployment strategies, especially considering the environmental impact and practical applications.

The current study did not account for the time frame of the considered papers. A crucial aspect for future exploration involves investigating whether reproducibility and other good practices have undergone changes over time. Given the increasing topic-related standards and publication requirements in recent years, it is pertinent to examine if these shifts have influenced reproducibility in more recent papers.

In conclusion, the proposed guidelines for reproducible machine learning research using EEG, as well as the overview of the current state of the literature regarding reproducibility, have the potential to support and motivate the community to further improve the current state of affairs. Our findings highlight the necessity for a change in research methods, putting a focus

on transparency, sharing data openly, and reporting resources in detail. Tackling these issues is crucial for moving the field forward, building trust in models, improving the quality of studies, and lessening the environmental impact of machine learning applications.

CHAPTER 4

GRAPH NEURAL NETWORKS FOR CROSS-SESSION CLASSIFICATION OF MENTAL WORKLOAD USING EEG

This chapter presents the methodology used to develop graph neural network (GNN) models for classifying mental workload across sessions, along with the conducted experiments and their results to answer research questions. GNNs are capable of modeling complex relationships between graph nodes through graph edges. EEG signals can naturally be represented as graphs due to the spatial correlation between electrode channels. Surface EEG electrodes capture signals influenced by neighboring regions, and GNNs facilitate a comprehensive representation of this data using message-passing mechanisms.

In the Methodology section (Section 4.1), the entire pipeline - from the raw dataset to model selection - is detailed. The studies and reporting follow reproducibility and generalizability guidelines established in Chapter 3. The COG-BCI Dataset ([190]) is used for the thesis studies. Multi-Attribute Task Battery (MATB) task is selected to classify three levels of mental workload, corresponding to easy, medium and difficult conditions (Section 4.1.1). Experiments described in Section 4.2 aim to improve model performance and address the following research questions:

- RQ3: How can mental workload be predicted using EEG data and graph neural networks?
  - RQ3.1: What is the effect of different graph structures on model performance?
  - RQ3.2: What are the possible solutions to prevent overfitting?
  - RQ3.3: Which metrics and evaluation methods can offer a more comprehensive assessment of the models?
  - RQ3.4: What is the effect of model complexity on model performance?
  - RQ3.5: What insights can we collect from subject-specific analysis?
  - RQ3.6: Can we identify the least number of channels that would minimize EEG device requirements?

Section 4.3 presents the results and Section 4.4 concludes the chapter with a discussion.

4.1 Methodology

The methodology employed in this study is illustrated in the flowchart presented in Figure 6. The subsequent subsections provide a detailed explanation of each step in the process. Data exploration, artifact cleaning and preprocessing, and feature extraction from the EEG signals are performed using the open-source EEGLAB toolbox (v2024.2) [191] that runs on MATLAB (R2023a) [192]. The graph neural network models used for classifying mental workload are developed and assessed using Python (v3.8) within notebook environments.

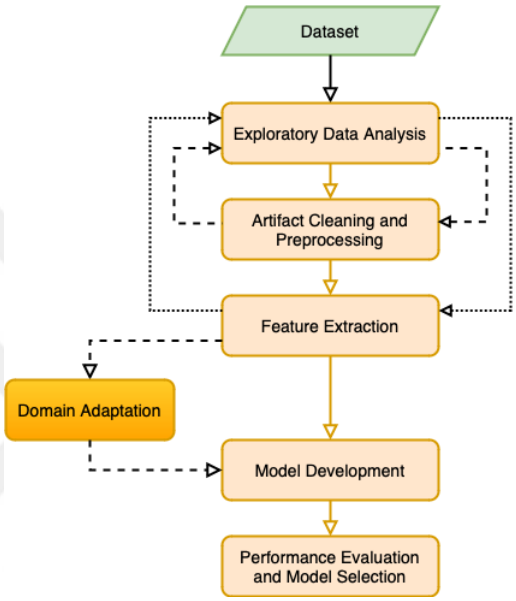


Figure 6: Overview of the Methodology

4.1.1 Dataset and Experimental Task

This thesis utilized the Cognitive BCI dataset (COG-BCI) ([190]), which is the extension of the passive BCI (pBCI) competition dataset. pBCI dataset, first used in the Neuroergonomics Conference 2021 [193], included preprocessed data from 15 subjects.

The COG-BCI dataset ([190]) contains raw data from 29 subjects (11 Female, 18 Male; mean age: 23.9 (std: 3.2) years), along with behavioral and subjective data. The raw recordings include data from the Psychomotor Vigilance Task (PVT), Flanker task, n-back task (n = 0, 1, 2), and the Multi-Attribute Task Battery (MATB) task with three difficulty levels. EEG data was acquired using a 64-electrode system with the reference electrode placed at Fpz. For participants 1-9, the Cz electrode was not recorded, and one electrode was used for electrocardiographic (ECG) data. No filtering was applied during data acquisition.

The experimental procedure, as detailed in [4], is illustrated in Figure 7. Three sessions were recorded at one-week intervals, with each session lasting approximately two hours. Behavioral data, indicating task performance, was collected for each task in terms of accuracies and

reaction times. Resting-state recordings included one minute of eyes-open and one minute of eyes-closed conditions at the start and end of each session. Tasks were performed in a pseudo-randomized order during each session. Subjective data was gathered using the Rating Scale Mental Effort (RSME) after each task, and the Karolinska Sleepiness Scale (KSS) was administered after the PVT task. Detailed information about the data collection process is provided in [4], ensuring transparency and reproducibility.

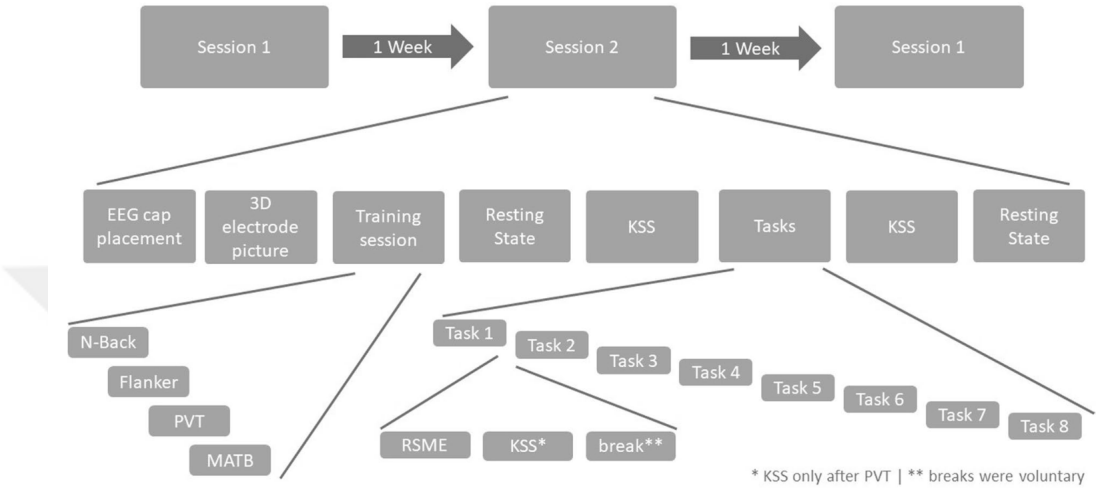


Figure 7: Experimental Procedure [4]

This recent dataset enables cross-session, subject-based estimation studies, although it requires substantial preprocessing. The availability of raw recordings and additional data offers flexibility for various research objectives, despite the increased computational and time demands associated with preprocessing, feature extraction, and model development due to the large number of participants and the unprocessed nature of the data.

The MATB task is selected for performing three-class classification among difficult, medium and easy conditions, as it better reflects real-life scenarios compared to the n-back task. The MATB-II task, developed by NASA, is designed to evaluate task-switching and mental workload in realistic piloting environments [194]. For the MATB task in the COG-BCI dataset, three five-minute runs with varying difficulty levels are designed, utilizing a combination of the four tasks: tracking (TRACK), system monitoring (SYSMON), communication (COMM), and resource management (RESMAN) (Table 8). As the difficulty level increases, the number of tasks or the complexity of tasks is adjusted accordingly. Details of these tasks are provided in [4]. Behavioral data, as summarized in Table 9, is collected for each task.

4.1.2 Exploratory Data Analysis

To explore the data, statistics of the raw data are first examined for inconsistencies or outliers across each session, difficulty level, and channel. In addition, behavioral data are inspected to assess performance variability between subjects, sessions, and difficulty levels. Finally, the

Table 8: MATB Experiments

MATB difficulty level	Tasks
Easy	SYSMON and TRACK
Medium	SYSMON, TRACK and RESMAN
Difficult	SYSMON, TRACK, RESMAN, COMM (Also, TRACK task was made more difficult)

Table 9: Behavioral Data

MATB Task	Behavioral Data
TRACK	- X and Y coordinates of the tracking tasks
SYSMON	- onset of a specific alarm and reaction time
RESMAN	- the amount of fuel in the relevant reservoirs
COMM	For each radio message, - was the message a target or not - radio to be changed - target frequency - did the participant change some radio frequency - was the change in frequency correct

distribution of features is analyzed using SOM visualizations to facilitate the interpretation of the model results.

4.1.3 Artifact Cleaning and Preprocessing

Electroencephalography (EEG) is a non-invasive technique used to measure and record electrical activity in the brain by placing multiple electrodes on the scalp. However, EEG signals are often contaminated by physiological artifacts such as eye and muscle movements, heart rate, as well as non-physiological artifacts such as line noise and measurement noise. Preprocessing is performed to clean and prepare raw data for further computations ([99]).

An automated preprocessing pipeline is constructed, as shown in Figure 8. The ECG channel is removed from the data after loading. The channel locations are imported using spherical coordinates from the Boundary Element Model (BEM) Dipfit model available in the EEGLAB toolbox. The data is then re-referenced to right mastoid (TP10) channel and high-pass filtered at 1 Hz to prepare it for Independent Component Analysis (ICA) computation [195]. Following this filter, the channels are rejected if their standard deviation exceeds two standard deviations of the average standard deviation of all channels.

Applying ICA to reject artifacts is a commonly used methodology applied successfully for artifact rejection of EEG signals and EEGLAB has a built-in ICA calculation capability ([196]). This algorithm separates the signal into statistically independent components which correspond to brain related independent components or artifacts. ICA components are extracted using the Picard algorithm ([197]) and automatically labeled using the ICLabel toolbox [198],

which labels components as muscle, eye, heart, brain, or other. The Swartz Center for Computational Neuroscience developed ICLabel to create an informative and reliable automated classifier for independent components. The vast amount of data which was collected over the years was crowd-labeled and machine learning models were trained using them ([199]). Components identified as muscle, eye, or heart with a probability greater than 95% are rejected. Since the interpolation of channels introduces dependency and results in a rank-deficient data matrix, ICA may yield inaccurate results if performed post-interpolation ([200]). Thus, spherical interpolation is applied only after ICA and component rejection. Finally, the data is low-pass filtered at 40 Hz and downsampled to 250 Hz. Both high- and low-pass filters were FIR filters applied using EEGLAB's *eegfiltnew()* function with default parameters. The automated preprocessing pipeline is executed for all subjects, sessions, and the MATB task.

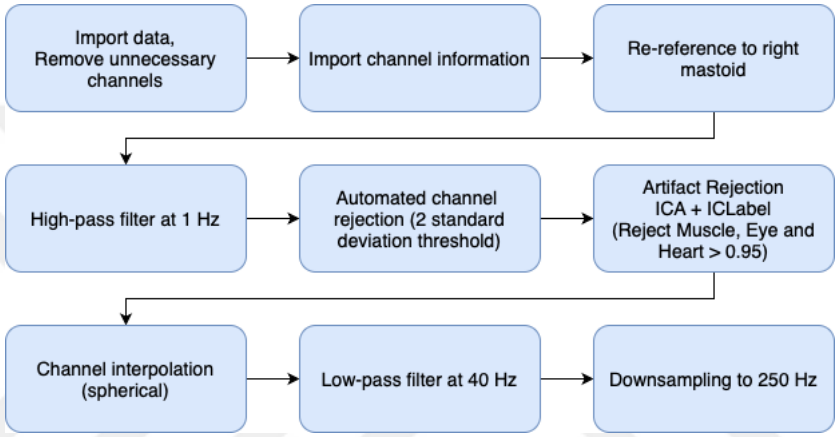


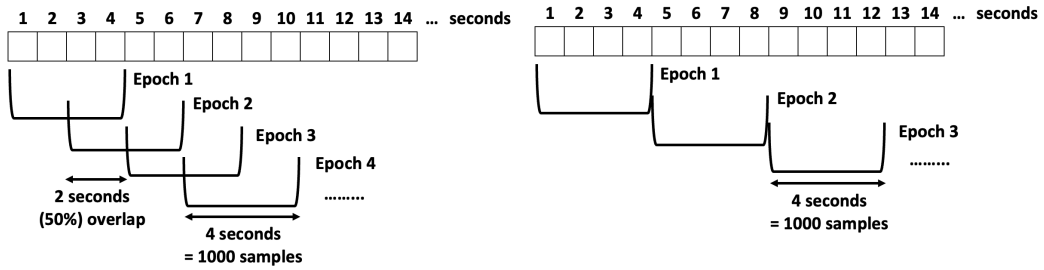
Figure 8: Preprocessing Pipeline

4.1.4 Feature Extraction

To utilize graph neural networks, the EEG data is transformed into a graph representation by extracting node and edge features, which are then integrated with an adjacency matrix. The following subsections detail this process. Before feature extraction, *Cz* channel, which is not recorded for the first nine subjects, is removed from the data for all participants to ensure consistency.

4.1.4.1 Segmentation

The systematic literature in 3 showed a varying range of epoch sizes used for EEG classification models. There are studies in the literature that investigate different epoch lengths in EEG ([201, 202]). Effects of epoch length can change with respect to task or data. In the preprocessed pBCI competition dataset, described in Section 4.1.1, 2-second non-overlapping epochs were used. In this thesis, EEG data is segmented into 4-second 50% overlapping epochs as a baseline. Additionally, 2-second epochs with either 50% overlapping or no overlapping as well as 4-second non-overlapping epochs were used in experiments to analyze



(a) 50% Overlapping 4-second Epochs (b) Non-overlapping 4-second Epochs

Figure 9: Overlapping and Non-overlapping Epochs

effects of different epoch lengths and overlap (Section 4.2.3). Figure 9 illustrates the differences between overlapping and non-overlapping 4-second epochs. Overlapping is employed to increase the information content, especially to account for the tapered edges of convolution signals during spectral feature calculation.

Each epoch is assigned a label based on the MATB task difficulty levels as Easy, Medium and Difficult as described in 4.1.1 to perform 3-class classification. Epoching into 4-second overlapping segments resulted in 444 samples for each of the training, validation and test sets, where classes were balanced and there were 148 samples for each class.

4.1.4.2 Node Features

EEG Power Spectral Density (PSD) is frequently used as a node feature for mental workload estimation due to its established relevance ([33], [34], [32]). In this study, PSD mean power values are computed for each epoch using Welch's method, implemented with EEGLAB's *spectopo()* function. A 1-second Welch window with 50% overlap is selected to enhance the information content. The node features consist of the mean power values for the frequency bands defined in Table 10. The node features are MinMax scaled for normalization using only train dataset statistics.

Table 10: Frequency Bands and Intervals

Frequency Band	Frequency Interval (Hz)
theta	4 - 8
alpha	8 - 13
beta1	13 - 20
beta2	20 - 30

4.1.4.3 Edge Features

Edge features represent the weights applied during message passing between connected nodes. In this study, reciprocal of the three-dimensional Euclidean distances between electrode positions are used as edge features to capture spatial relationships since closer edges need to have larger weights for message passing. The weights are normalized using the maximum value, as described by Equation 1, where n is the total number of nodes, d_{ij} and w_{ij} denote the Euclidean distance and the edge weight between nodes i and j, respectively.

$$w_{ij} = \frac{1}{d_{ij}.max\{\frac{1}{d_{ij}} : i = 1..n, j = 1..n\}} \tag{1}$$

4.1.4.4 Graph Representation

The graph representation is completed by defining the adjacency matrix along with the extracted node and edge features. Two approaches are employed for constructing the graph:

1. Single Connected Component: The graph is formed by selecting a percentage, p, of the edges with the highest weights, ensuring a single connected component.
2. k-Nearest Neighbor (k-NN) Graph Generation: kNN graphs are generated using PyTorch Geometric's *"knn_graph()"* function such that the in-degree of each node equals *k*.

4.1.4.5 Node Selection

In some experiments, a subset of nodes with the highest channel-wise mutual information between node features and the ground truth labels is selected to potentially enhance classification performance by decreasing model complexity. This selection is conducted only on the training set to prevent data leakage into the validation or test sets. After the nodes with the highest mutual information are identified, the graph structures are constructed using the k-nearest neighbor (k-NN) method.

4.1.4.6 Data Split

To investigate cross-session estimation, the data is split into training, validation, and test sets based on sessions. Two strategies were employed for creating the training and validation sets. In the first strategy, sessions are kept separate between the training and validation sets to evaluate cross-session training performance. In the second strategy, the data from two sessions is shuffled and split 50-50 to assess performance when cross-session variation was minimized between the training and validation sets. The test set always comprises a distinct

third session to ensure that generalizability was evaluated without cross-session data leakage into the test set. Detailed descriptions of the data splits for each experiment are provided in Section 4.2.1.

The training set is used for model training, while the validation set is employed for hyper-parameter optimization and model selection. The test set, containing only unseen data, is reserved for evaluating the model's performance and assessing its generalizability.

4.1.5 Model Development

4.1.5.1 Graph Neural Networks

Graph neural networks are first proposed in [49]. These neural network models operate on graph structures, represented as $G(N, E)$, where N denotes the set of nodes and E denotes the set of edges. By representing data graph form, a variety of tasks can be performed, such as node classification, link prediction, graph classification, clustering, and graph generation. To perform graph classification, the graph G is mapped to a vector of integers in $\mathbb{N}^m$ by a function $\tau(G)$, where m is the number of classes.

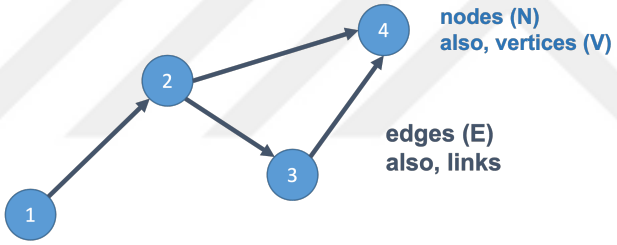


Figure 10: A Simple Graph Structure with Four Nodes and Four Edges

GNNs offer several advantages. They effectively leverage relational information inherent in graph structures, preserving the topological dependencies between nodes. Unlike images (fixed size grid graphs) or text (sequential - line graphs) (Figure 11), graphs do not require fixed node ordering or a reference point. Additionally, GNNs can incorporate both node features and edge features (weights), enriching the representation of complex data.

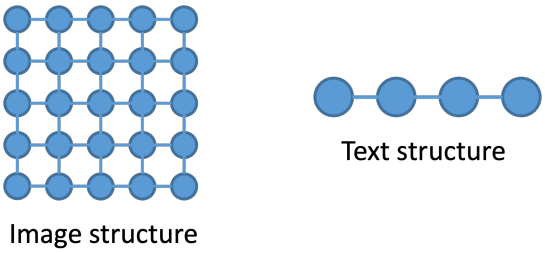


Figure 11: Image and Text as Graphs

Message Passing and Neighborhood Aggregation During the training process of GNNs, node features are propagated throughout the graph based on the directions and weights of edges, a process known as Message Passing.

Each node then aggregates information passed from its neighbors to construct node embeddings, a process referred to as Neighborhood Aggregation. The aggregation operator (gray boxes in Figure 12) must be permutation invariant (i.e., order-independent) to ensure consistent learning. Common permutation-invariant functions include averaging, summing, or applying a neural network layer.

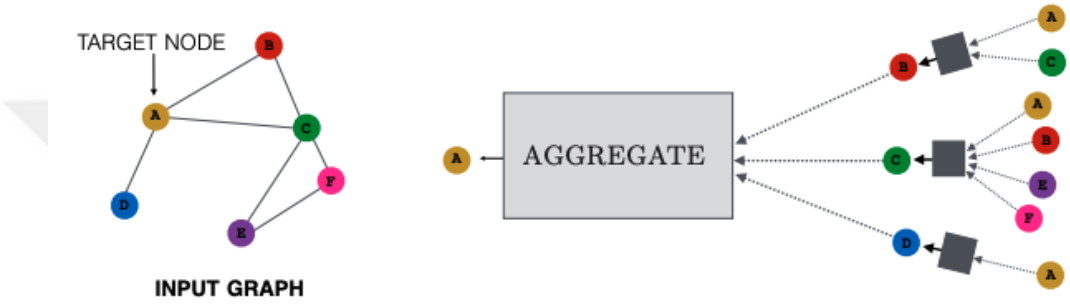


Figure 12: Message Passing and Neighbourhood Aggregation [5]

Through message passing and neighborhood aggregation, nodes gain information about their neighboring nodes in addition to their own features. This enhanced representation improves the overall expressive power of the graph. Since graphs do not have inherent local receptive fields (as in convolutional neural networks), graph convolutions are defined by message passing followed by local neighborhood aggregation, which constructs the graph embedding. Each layer of graph convolution corresponds to information passed from a 1-hop neighborhood of nodes. Consequently, the number of graph convolution layers indicates the depth of neighborhood information (i.e., the number of hops) that the model captures.

After obtaining node embeddings, a graph embedding is computed to enable graph-level classification. This involves applying an order-invariant pooling function (e.g., global mean pooling, global max pooling) to the node embeddings. The resulting graph embedding is then passed to a classifier network to generate the final class predictions.

Different GNN architectures primarily differ in their message passing and aggregation mechanisms. One of the foundational GNN architectures is the Graph Convolutional Network (GCN), which employs local convolutions and incorporates self-loops in the adjacency matrix to prevent loss of node-specific features ([50]). The GCN model is implemented in the PyTorch Geometric package as the *GCNConv()* class. Another commonly used architecture integrates higher-order WL concepts into GNN architectures, which enhances the expressiveness of GNNs ([203]). This architecture is available as the *GraphConv()* class in PyTorch Geometric.

4.1.5.2 Model Architecture

The employed GNN architecture is illustrated in Figure 13. The input graph data is processed through multiple graph convolution layers with nonlinear activation functions to generate node embeddings. At the readout layer, global pooling is applied as a permutation-invariant function to produce graph-level embeddings, which serve as input to the final classifier. The classifier consists of fully connected linear layers with activation functions, with dropout applied at the beginning to mitigate overfitting. In the experiments, global max pooling was used at the readout layer.

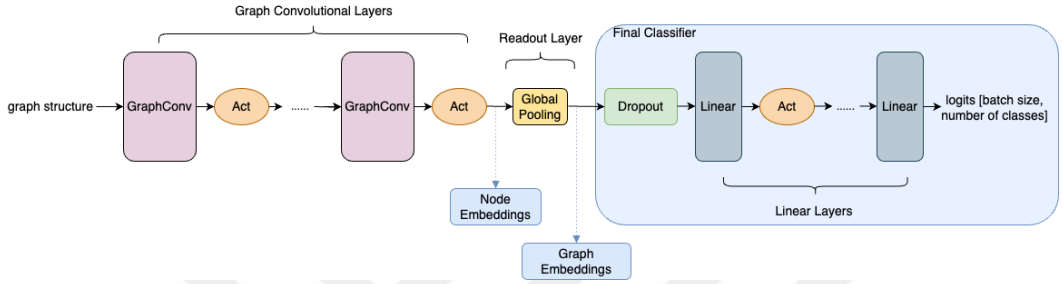


Figure 13: GNN Architecture

The models are implemented in Python using the open-source PyTorch Geometric package [204], which is widely adopted for the development of graph neural networks and provides numerous graph operators.

4.1.5.3 Training

The cross-entropy loss function in PyTorch was optimized using the Adam optimizer with weight decay to mitigate overfitting. This loss function inherently applies a softmax operation to the output of the final layer to generate class probabilities. An additional measure against overfitting was the use of learning rate scheduling, which reduced the learning rate when the loss did not improve for a specified number of epochs. Such measures were necessary due to the limited sample size of the dataset.

The Weights and Biases framework was utilized to track the experiments and visualize metrics [205]. Each trial, corresponding to a distinct set of hyperparameters, was trained for 500 epochs, with the best epoch being logged based on the heuristic loss defined in Section 4.1.6.2.

4.1.6 Performance Evaluation and Model Selection

4.1.6.1 Performance Metrics

The COG-BCI dataset is balanced such that each subject has an equal number of recordings for each mental workload level, with uniform length. Therefore, accuracy is used as the

primary metric for evaluating model performance. Additionally, confusion matrices are utilized as complementary metrics to gain a more comprehensive understanding of the model's classification performance ([124]). These supplementary metrics help in analyzing and understanding the model's learning behavior beyond what is captured by accuracy alone.

4.1.6.2 Hyperparameters and Model Selection

Hyperparameter optimization is conducted using the Optuna framework [206], which facilitates efficient search through Bayesian optimization. Optuna seeding is employed to ensure the reproducibility of experiments across different runs. The hyperparameters tuned during the training of the graph neural network are detailed in Section 4.2.2. They include activation functions, batch size, learning rate, dropout ratio, weight decay, number of layers, and number of neurons in each graph or linear layer.

Model selection is particularly challenging in this study due to the difficulties in achieving generalization across sessions and the variability observed between subjects. The high inter-subject variability often results in inconsistent learning behavior, where some models show high performance for certain subjects while performing poorly for others. Despite regularization techniques applied to mitigate overfitting, some models exhibit a tendency to overfit the training data. In these cases, training accuracy remains high, but the validation accuracy drops significantly, indicating poor generalization. On the other hand, underfitting models fail to capture the underlying patterns in the data, leading to poor performance on both the training and validation sets.

Relying solely on the highest validation accuracy to select the best model can be problematic. The model with the highest validation accuracy might correspond to an underfitting model that coincidentally performs well on the validation set without learning meaningful patterns from the training data. Conversely, it could also represent an overfitting model if the difference between the training and validation losses is not considered. In such cases, the model effectively memorizes the training data but fails to generalize, giving a poor validation performance.

To address these issues, a weighted heuristic loss function is defined (Equation 2) to aid in model selection. This heuristic loss function accounts for both the validation loss and the difference between training and validation losses, penalizing models that exhibit signs of overfitting or underfitting. The defined loss function automates the process of selecting the best trial among Optuna runs and the best epoch within each trial, thus improving time efficiency while maintaining robust model evaluation.

$$loss_{heuristic} = 0.75 * loss_{validation} + 0.25 * \begin{cases} loss_{diff}, & \text{if } loss_{diff} >= 0. \\ 1e6, & \text{otherwise.} \end{cases} \quad (2)$$

Here, $loss_{diff}$ represents the difference between training and validation losses. The penalty term $1e6$ is introduced for negative values of $loss_{diff}$ to strongly penalize models where the training loss is higher than the validation loss, a situation indicative of unstable or incorrect learning behavior. By combining validation loss with the difference between training and

validation losses, the heuristic loss function discourages both overfitting and underfitting, leading to the selection of models that generalize better across sessions.

This approach allows for an automated and objective selection of the optimal model, reducing the risk of selecting a model based solely on high validation accuracy, which might be a result of chance or show poor generalization capability.

4.1.6.3 Self Organizing Maps

The Self-Organizing Map (SOM) is an unsupervised neural network model introduced by Kohonen ([207], [208]) used for dimensionality reduction, data visualization and clustering. It projects high-dimensional data onto a 2-dimensional grid of neurons, preserving the topological relationships of the input. This grid of neurons attempts to capture the structure of the input data while preserving its intrinsic spatial relationships. Each neuron is represented by a weight vector, which adjusts during training to approximate the input data distribution.

During SOM training, neuron weights are initialized and for each input sample, the neuron with the closest weight vector (smallest distance) is identified as the Best Matching Unit (BMU). The BMU and its neighboring neurons update their weights to better match the input, guided by a decreasing learning rate and neighborhood function. Similar data points are mapped to nearby neurons, forming clusters on the SOM grid. Clusters can be identified using a Unified Distance Matrix (U-Matrix), where distances between neurons indicate cluster boundaries.

In this study, SOM is used to visualize baseline feature alignment to compare train, validation, and test features before training the model. For both cases, the SOM is trained only on the train set and used to map validation and test sets. Labels, while not used during SOM training, are visualized to interpret class separability. This technique provides intuitive 2-dimensional visualization of data distributions with high dimensions.

4.1.6.4 Statistical Significance of Models

The Friedman test is a method used to assess the joint statistical significance of multiple models, serving as a non-parametric alternative to repeated-measures ANOVA. This test ranks the performance of algorithms separately for each dataset and returns a p-value under the null hypothesis that all models perform equivalently ([126]). Unlike parametric methods, the Friedman test can be used for cases where the dependent variable is ordinal or not normally distributed.

The Wilcoxon Signed-Ranks test, a non-parametric alternative to the paired t-test, evaluates the performance differences between two models. It compares the locations of two populations to determine whether one is consistently shifted relative to the other by analyzing the sum of ranks of the differences. This test ignores the signs of the differences, instead ranking them and comparing positive and negative differences ([126]). The Wilcoxon test is particularly valuable when the dependent variable is not normally distributed, as it does not rely on

assumptions of normality. It is commonly used for detecting pairwise statistically significant differences between models.

Both tests provide robust methods for statistical analysis in scenarios where parametric assumptions do not hold, ensuring the validity of results in diverse experimental conditions.

4.2 Experiments

The GNN architecture explained in 4.1.5.2 is utilized for the experiments. The details of the experiments performed to analyze and improve the classifier performance on the COG-BCI dataset, as well as to address the research questions, are reported in the following subsections.

4.2.1 Variations in Data Split

For exploring cross-session classification of EEG data, the data split for training, validation, and test sets is performed with respect to sessions. For the baseline model, COG_BCIg1, session 1, 2, and 3 are assigned as the train, validation, and test sets, respectively. To analyze differences between sessions, an additional experiment (COG_BCIg3) is conducted where validation and test sets were swapped, keeping session 1 as the training set. Further experiments are designed to investigate the effects of distribution differences between sessions, where the training and validation sessions are shuffled and split 50-50 to generate train and validation sets, as shown in Table 11. A 50-50 split is preferred to make the experiment comparable to the baseline and to ensure balanced data across the training, validation, and test sets. The test session is kept independent from train and validation sets to assess the generalizability of the models to unseen sessions.

Table 11: Experiments with Variations in Data Split

Model Name	Data Split
COG_BCIg1	Train Session:1, Validation Session:2 Test Session:3
COG_BCIg2	Train Session, Validation Session: 1+2, Split 50-50 Test Session:3
COG_BCIg3	Train Session:1, Validation Session:3 Test Session:2
COG_BCIg4	Train Session, Validation Session: 1+3, Split 50-50 Test Session:2

4.2.2 Hyperparameter Search Space

Hyperparameter search space is refined based on the results of initial iterations where the parameter ranges were checked for convergence. The final limits of the hyperparameters are given in Table 12, which were used for the baseline experiment (COG_BCIg1). Due to the number of hyperparameters, 150 Optuna trials were conducted. To better analyze the

effect of reduced model complexity and its potential against overfitting, a smaller architecture (COG_BCIg1_smallArch) was also tested, with changes in hyperparameters listed in Table 13.

Table 12: Hyperparameters - Baseline

activation function	[gelu, relu, tanh]
batch size	[16, 64]
learning rate	[1e-4, 1e-2]
dropout	[1e-2, 0.5]
weight decay	[1e-4, 1e-1]
number of hidden graph layers	[1, 4]
number of nodes in graph layers	[4, 32]
number of linear layers	[2, 5]
number of nodes in layers	[8, 64]

Table 13: Hyperparameters - Small Architecture

number of hidden graph layers	[1, 2]
number of nodes in graph layers	[1, 5]
number of linear layers	[2, 3]
number of nodes in layers	[1, 5]

4.2.3 Epoch Length and Overlap

The baseline configuration uses 4-second epochs with 50% overlap (COG_BCIg1). The variations are analyzed to investigate the effect of epoch length and overlap on model performance as described in Section 4.1.4.1. Corresponding model names are given below.

1. COG_BCIg1: 4-second epochs, with 50% overlap
2. COG_BCIg1_2sec: 2-second epochs, with 50% overlap
3. COG_BCIg1_4sec_no: 4-second epochs, non-overlapping
4. COG_BCIg1_2sec_no: 2-second epochs, non-overlapping

4.2.4 Graph Structures

Different graph structures are analyzed to assess the effects of graph representation on model performance. As described in 4.1.4.4, two methods are employed to construct the graph structures. For the single connected component graph representation, 6% threshold is applied when using all channels as nodes in the graph. This representation serves as the baseline model (COG_BCIg1). Additionally, k-NN graphs are generated with k=3 and k=5. The following variations are explored as graph structures:

1. COG_BCIg1: Single Connected Component, threshold = 6%
2. COG_BCIg1_3nn: knn graph with k=3
3. COG_BCIg1_5nn: knn graph with k=5

4.2.5 Reduced Features and Node Selection

Large feature dimension can be another reason for the models to overfit or failure to generalize. To analyze this behavior, the variations below with reduced features or node selections are tested. In the first three experiments, regions and frequency bands identified in the literature as being correlated with mental workload (Section 2.1.3.2) are selected to evaluate whether this results in a reduction of device requirements, such as the number of channels. Additionally, in experiments that involve selecting a percentage of nodes, the goal is to include the most salient channels based on mutual information with the true labels, thereby decreasing model complexity. The effects of these modifications on model performance are then observed.

1. COG_BCIg1_theta_alpha Used only theta and alpha band features
2. COG_BCIg1_parietal: Only parietal channels are used
3. COG_BCIg1_frontal: Only frontal channels are used
4. COG_BCIg1_50p_3nn: 50% of the nodes are selected, knn graph with *k*=3
5. COG_BCIg1_50p_5nn: 50% of the nodes are selected, knn graph with *k*=5
6. COG_BCIg1_25p_5nn: 25% of the nodes are selected, knn graph with *k*=5

4.2.6 Batch Normalization

Another measure to introduce regularization against overfitting is batch normalization [209]. Another experiment (COG_BCIg1_batchNorm) is executed with batch normalization at both the graph layers and linear layers.

4.2.7 Binary Classification

In addition to the three class classification, binary classification between the Easy and Difficult conditions is performed to assess the performance of the model to classify dual-task mental workload levels (COG_BCIg1_binary). Only data for the highest and lowest workload level are included for this classification, leaving out the medium level for which the actual workload level is more ambiguous and feature distributions tend to overlap with either the Easy or the Difficult conditions according to preliminary exploratory analysis.

4.3 Results

4.3.1 Exploratory Data Analysis

4.3.1.1 Raw Data

Data exploration of the raw dataset showed that subjects 17 and 27 had identical data. Therefore, data from subject 27 was excluded to avoid redundancy. Additionally, subjects 25 and 28 had the same data for session 1. Since it was not possible to know the association of subsequent sessions to their subjects, both subjects were excluded from the study. As a result, data from 26 subjects were used in the studies. Moreover, large anomalous values were identified for Subject 14, session 2, FT9 channel. These were automatically removed during preprocessing.

4.3.1.2 Features

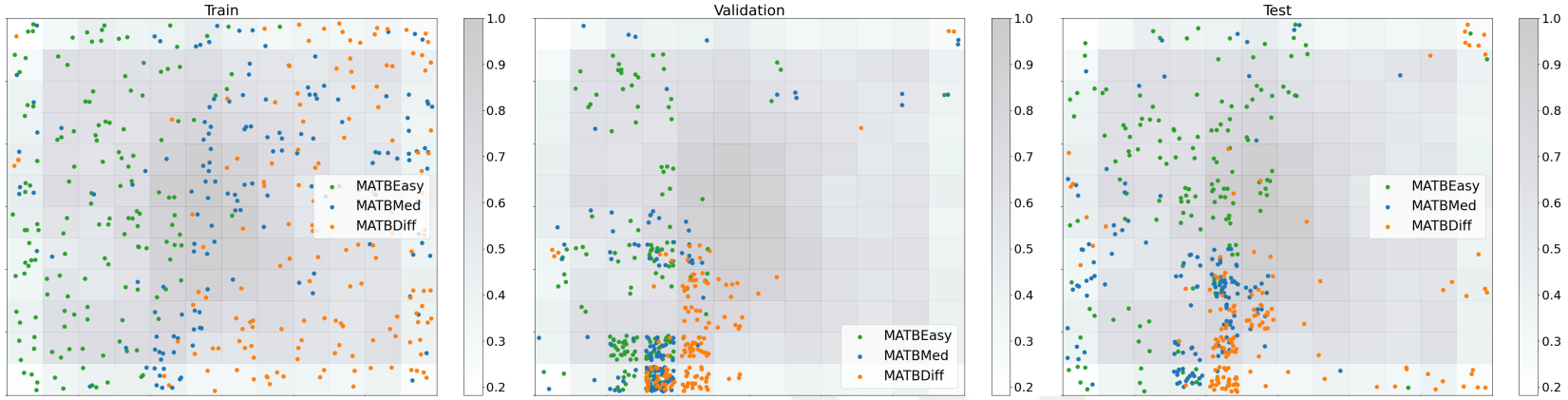
SOM plots of features for subjects 1 and 2 are shown in Figure 14, with columns corresponding to data from sessions 1, 2, and 3, respectively. The SOM is trained in an unsupervised manner using only session 1 data, reducing its dimensionality to two for visualization. This trained model is then applied to transform data from sessions 2 and 3. While labels are excluded during training, they are included in the plots to evaluate class separability visually. For subjects 1 and 2, the distributions of the validation and test sets differ noticeably from that of the training set, highlighting session-specific variability. Plots for all subjects are provided in Appendix B.

4.3.1.3 Task Performance

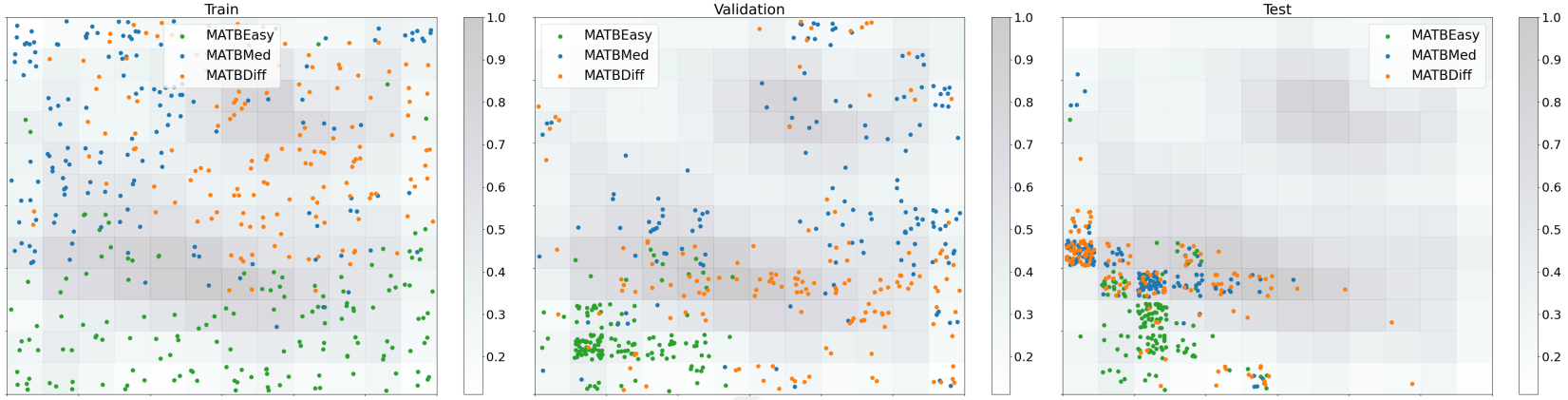
In the COG-BCI dataset, behavioral data indicative of task performance was collected alongside EEG recordings. Exploratory analysis was conducted for the TRACK and SYSMON tasks, as these tasks were consistently included across all difficulty levels.

For the TRACK task, the behavioral data comprised X and Y coordinates representing the participant's tracking performance. The root mean squared distance was computed using these coordinates to quantify how closely the participant maintained the target within the central area, with smaller distances indicating better performance. For the SYSMON task, the behavioral metric analyzed was the reaction time of participants following the onset of an alarm, serving as an indicator of attentiveness and responsiveness.

Figures 15 and 16 illustrate the distributions of track distance and reaction times, respectively, for each subject, session, and difficulty level. These boxplots reveal notable patterns in task performance. While individual performance varies, a general deterioration in performance is observed under more difficult conditions, aligning with the expectation that increased task difficulty leads to higher mental workload. Additionally, performance tends to improve across sessions, which can be attributed to task learning and adaptation over time. These findings



(a) Subject 1



(b) Subject 2

Figure 14: SOM Plots of Features

are consistent with prior research linking behavioral measures to cognitive workload. Statistical analyses presented in [4] further support these observations, providing evidence of the relationship between task difficulty, workload, and performance.

Note that the relation between task performance and task difficulty (in this case, cognitive workload condition) is not straightforward ([210]). Increasing task difficulty may lead to higher mental workload in individuals who manage to keep the task performance at the same level as when the task is easy. Individuals who do not spend any extra effort with higher task difficulty (e.g. due to lack of motivation) should be considered to have unvarying workload across workload conditions, even though performance is affected. However, participants are assumed to be motivated to perform well across all conditions, and that the effect of workload condition on performance indicates that the levels differ strongly enough to be able to model cognitive workload.

4.3.2 Artifact Cleaning and Preprocessing

The rejected channels and components are saved during automated preprocessing described in 4.1.3. On average, 3.37 $\pm$ 0.88 channels and 2.32 $\pm$ 2.17 components are rejected for MATB tasks, calculated over each subject and session.

4.3.3 Graph Representation

Figures 17 - 18 show the graph representations for Single Connected Component (p = 6%) and 3nn and 5nn graphs, respectively, when all nodes are used. Higher density of edges of 5nn graph is seen. Figure 19 show the resulting 5nn graph representations for subject 1 as an example when 50% and 25% of the nodes are selected. After node selection, each subject has different graph representations.

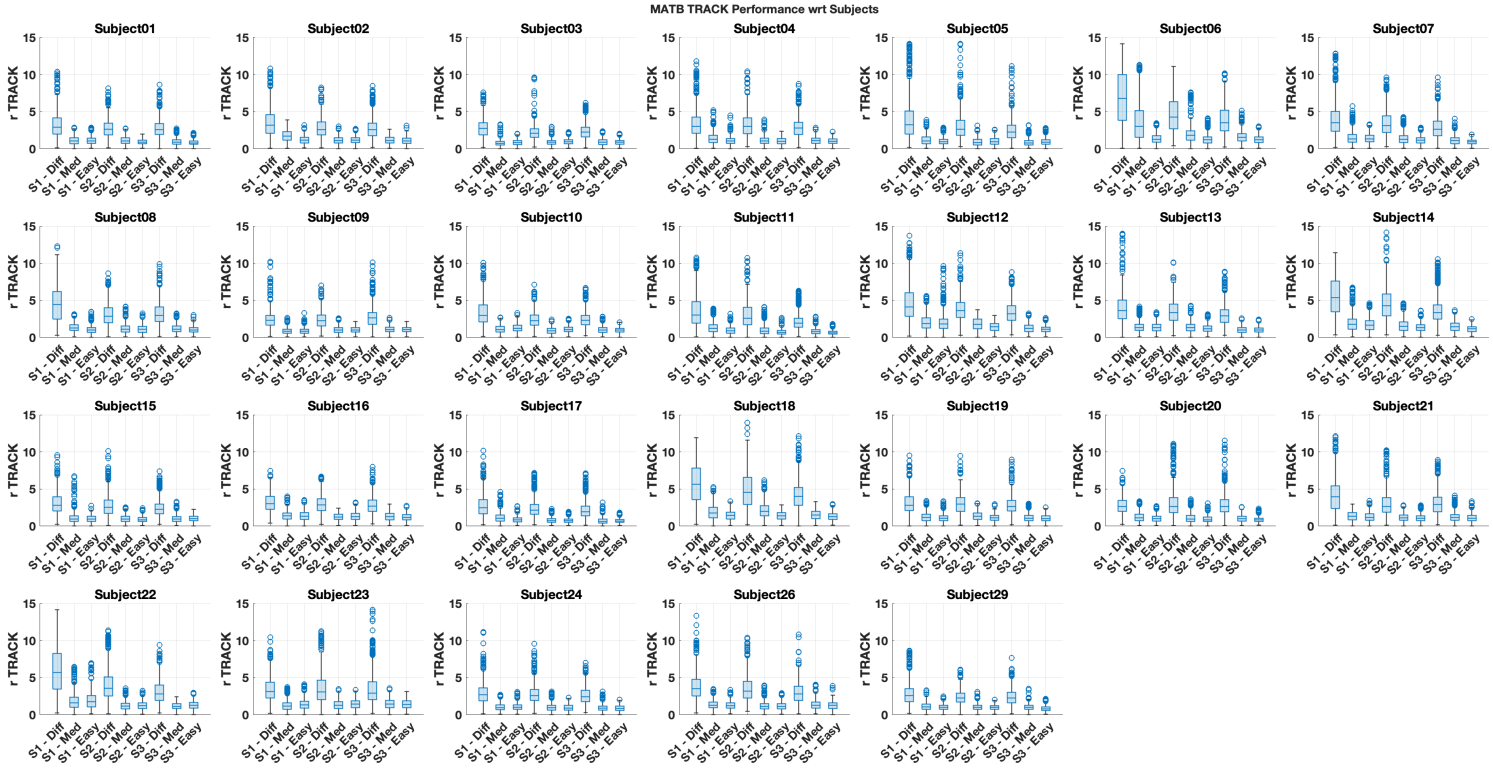


Figure 15: TRACK Task Performance

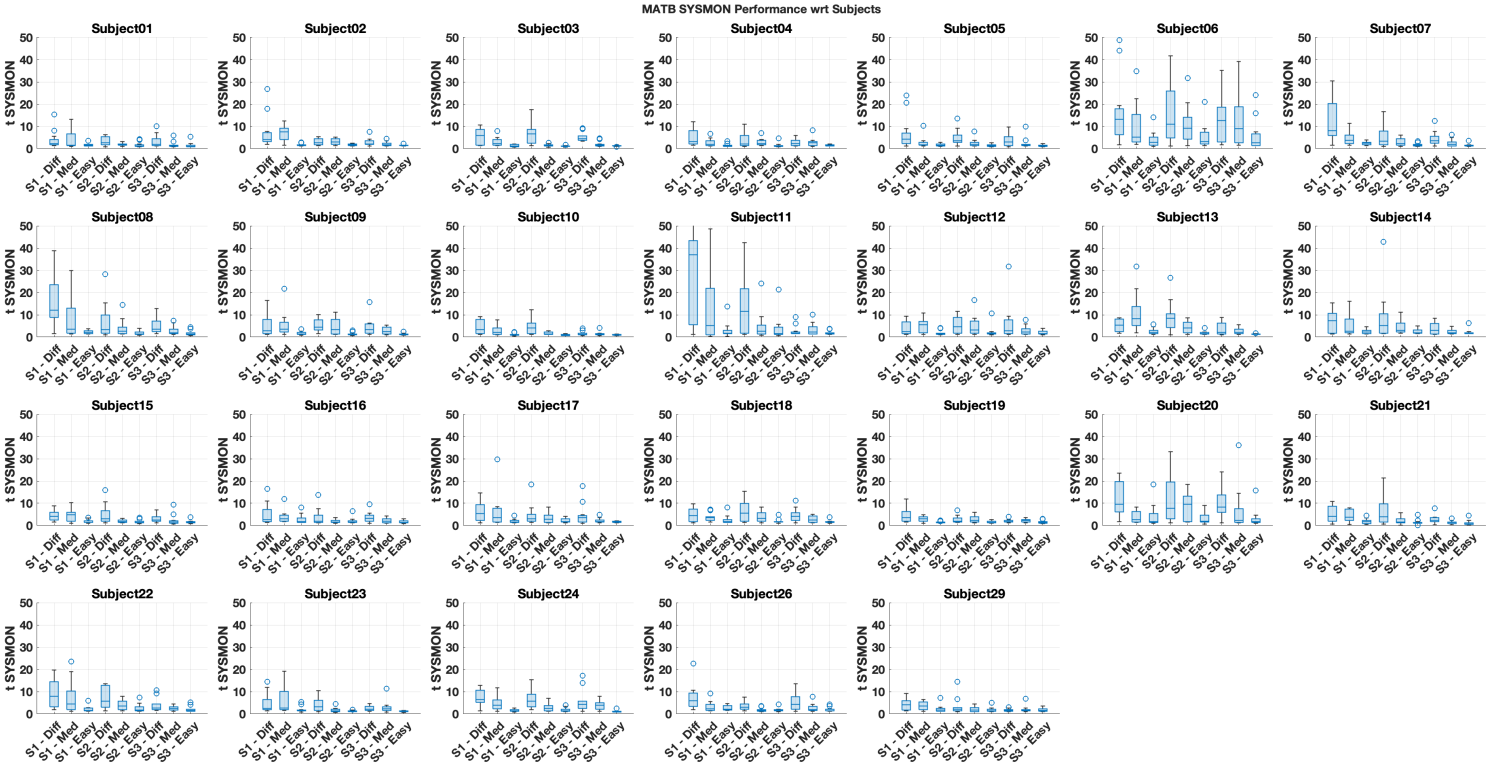


Figure 16: SYSMON Task Performance

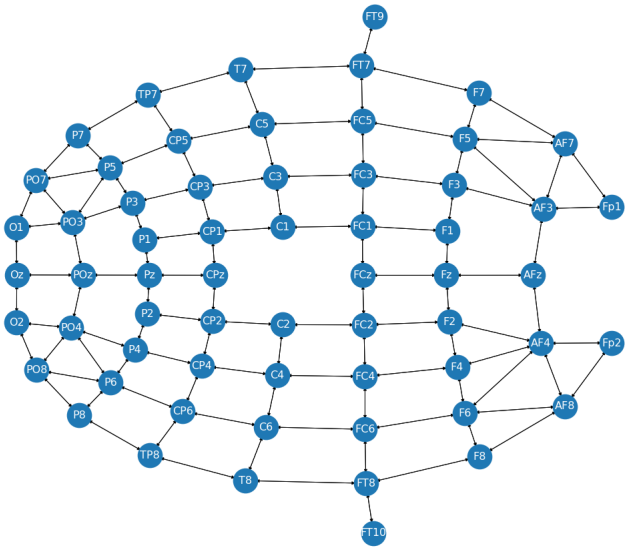
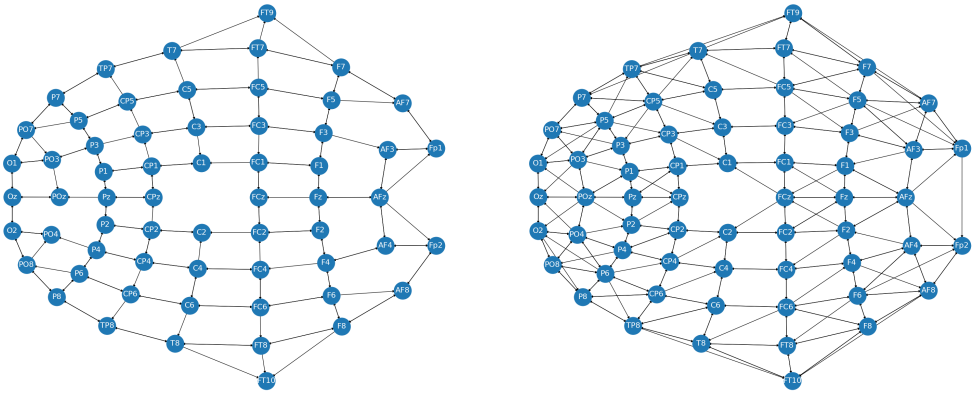


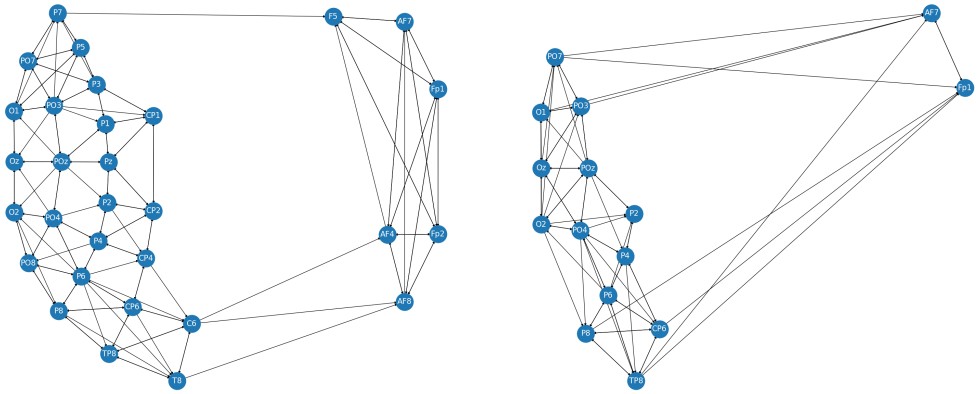
Figure 17: Graph Structure for Single Connected Component ($p = 6\%$) using All Nodes



(a) 3nn Graph Structure

(b) 5nn Graph Structure

Figure 18: knn Graph Structures



(a) 50% Node Selection (b) 25% Node Selection

Figure 19: 5nn Graph Structures with Node Selection (Subject 1)

4.3.4 Experiment Results

4.3.4.1 Model Performance

Train, validation and test set accuracies for the best models averaged over subjects are presented in Table 14 for the experiments with variations in data split and in Table 15 for the remaining experiments. When averaged with respect to subjects, all models exceeded adjusted chance level accuracy for the test set, which is 38% for 3-class classification and 148 samples per class and 95% confidence interval ([1, 125]). Subject based results including confusion matrices are presented in Table 16 for the COG_BCIg1 experiment.

Table 14: Experiments with Variations in Data Split

Model Name	**Train Accuracy (mean ± std)**	**Validation Accuracy (mean ± std)**	**Test Accuracy (mean ± std)**
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg2	0.9815 ± 0.0295	0.9535 ± 0.0374	0.4238 ± 0.1762
COG_BCIg3	0.5796 ± 0.1493	0.5500 ± 0.1329	0.4724 ± 0.1235
COG_BCIg4	0.9866 ± 0.0198	0.9551 ± 0.0342	0.4902 ± 0.1557

Table 15: Experiment Variations Compared to COG_BCIg1

Model Name	**Train Accuracy (mean ± std)**	**Validation Accuracy (mean ± std)**	**Test Accuracy (mean ± std)**
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg1 smallArch	0.6596 ± 0.1139	0.6097 ± 0.1328	0.4320 ± 0.1232
COG_BCIg1_batchNorm	0.7974 ± 0.1525	0.7427 ± 0.1375	0.4481 ± 0.1208
COG_BCIg1_frontal	0.6154 ± 0.1532	0.5754 ± 0.1552	0.4526 ± 0.0924
COG_BCIg1_parietal	0.6272 ± 0.1601	0.5823 ± 0.1621	0.4311 ± 0.1086
COG_BCIg1_theta_alpha	0.5963 ± 0.1458	0.5752 ± 0.1243	0.4462 ± 0.1053
COG_BCIg1_2sec	0.5920 ± 0.1784	0.5707 ± 0.1785	0.4235 ± 0.1018
COG_BCIg1_2sec_no	0.6015 ± 0.1810	0.5733 ± 0.1774	0.4313 ± 0.1030
COG_BCIg1_4sec_no	0.6460 ± 0.1694	0.5994 ± 0.1668	0.4305 ± 0.1076
COG_BCIg1_3nn	0.6500 ± 0.1628	0.5898 ± 0.1722	0.4435 ± 0.1182
COG_BCIg1_5nn	0.6284 ± 0.1753	0.5868 ± 0.1610	0.4546 ± 0.1142
COG_BCIg1_50p_3nn	0.6712 ± 0.1745	0.6039 ± 0.1658	0.4343 ± 0.1152
COG_BCIg1_50p_5nn	0.6565 ± 0.1741	0.6028 ± 0.1667	0.4541 ± 0.1235
COG_BCIg1_25p_5nn	0.6610 ± 0.1899	0.6134 ± 0.1902	0.4406 ± 0.0949

Table 16: COG_BCIg1 Results with respect to Subjects

Subject	Accuracy			Confusion Matrix		
	Train	Validation	Test	Train	Validation	Test
1	0.6779	0.5811	0.4752	$\begin{bmatrix}117 & 13 & 18\\ 27 & 57 & 64\\ 3 & 18 & 127\end{bmatrix}$	$\begin{bmatrix}136 & 7 & 5\\ 62 & 35 & 51\\ 37 & 24 & 87\end{bmatrix}$	$\begin{bmatrix}102 & 25 & 21\\ 90 & 35 & 23\\ 33 & 41 & 74\end{bmatrix}$
2	0.7703	0.4910	0.4572	$\begin{bmatrix}128 & 8 & 12\\ 9 & 92 & 47\\ 4 & 22 & 122\end{bmatrix}$	$\begin{bmatrix}148 & 0 & 0\\ 15 & 20 & 113\\ 66 & 32 & 50\end{bmatrix}$	$\begin{bmatrix}148 & 0 & 0\\ 93 & 55 & 0\\ 84 & 64 & 0\end{bmatrix}$
3	0.5315	0.5901	0.3491	$\begin{bmatrix}74 & 32 & 42\\ 25 & 14 & 109\\ 0 & 0 & 148\end{bmatrix}$	$\begin{bmatrix}132 & 8 & 8\\ 76 & 20 & 52\\ 16 & 22 & 110\end{bmatrix}$	$\begin{bmatrix}1 & 2 & 145\\ 20 & 15 & 113\\ 4 & 5 & 139\end{bmatrix}$
4	0.8131	0.7275	0.5338	$\begin{bmatrix}128 & 18 & 2\\ 0 & 110 & 38\\ 0 & 25 & 123\end{bmatrix}$	$\begin{bmatrix}137 & 10 & 1\\ 16 & 79 & 53\\ 0 & 41 & 107\end{bmatrix}$	$\begin{bmatrix}63 & 59 & 26\\ 20 & 114 & 14\\ 2 & 86 & 60\end{bmatrix}$
5	0.7703	0.6126	0.2928	$\begin{bmatrix}104 & 29 & 15\\ 24 & 102 & 22\\ 1 & 11 & 136\end{bmatrix}$	$\begin{bmatrix}13 & 134 & 1\\ 18 & 129 & 1\\ 3 & 15 & 130\end{bmatrix}$	$\begin{bmatrix}27 & 121 & 0\\ 46 & 102 & 0\\ 14 & 133 & 1\end{bmatrix}$
6	0.8941	0.8784	0.3739	$\begin{bmatrix}143 & 0 & 5\\ 0 & 121 & 27\\ 9 & 6 & 133\end{bmatrix}$	$\begin{bmatrix}139 & 4 & 5\\ 4 & 109 & 35\\ 0 & 6 & 142\end{bmatrix}$	$\begin{bmatrix}34 & 34 & 80\\ 10 & 14 & 124\\ 5 & 25 & 118\end{bmatrix}$
7	0.6284	0.5270	0.4617	$\begin{bmatrix}140 & 8 & 0\\ 35 & 100 & 13\\ 28 & 81 & 39\end{bmatrix}$	$\begin{bmatrix}79 & 50 & 19\\ 8 & 66 & 74\\ 10 & 49 & 89\end{bmatrix}$	$\begin{bmatrix}70 & 76 & 2\\ 30 & 74 & 44\\ 14 & 73 & 61\end{bmatrix}$
8	0.8063	0.7568	0.4144	$\begin{bmatrix}141 & 7 & 0\\ 4 & 133 & 11\\ 1 & 63 & 84\end{bmatrix}$	$\begin{bmatrix}145 & 3 & 0\\ 1 & 90 & 57\\ 0 & 47 & 101\end{bmatrix}$	$\begin{bmatrix}142 & 6 & 0\\ 136 & 12 & 0\\ 44 & 74 & 30\end{bmatrix}$
9	0.8198	0.8176	0.4257	$\begin{bmatrix}111 & 32 & 5\\ 28 & 113 & 7\\ 0 & 8 & 140\end{bmatrix}$	$\begin{bmatrix}132 & 14 & 2\\ 10 & 102 & 36\\ 0 & 19 & 129\end{bmatrix}$	$\begin{bmatrix}107 & 40 & 1\\ 68 & 79 & 1\\ 32 & 113 & 3\end{bmatrix}$
10	0.9977	0.8761	0.3333	$\begin{bmatrix}148 & 0 & 0\\ 0 & 147 & 1\\ 0 & 0 & 148\end{bmatrix}$	$\begin{bmatrix}144 & 1 & 3\\ 2 & 111 & 35\\ 0 & 14 & 134\end{bmatrix}$	$\begin{bmatrix}0 & 0 & 148\\ 0 & 0 & 148\\ 0 & 0 & 148\end{bmatrix}$

Table 16: cont.

11	0.7185	0.6982	0.5901	$\begin{bmatrix}121 & 18 & 9\\ 29 & 90 & 29\\ 9 & 31 & 108\end{bmatrix}$	$\begin{bmatrix}124 & 20 & 4\\ 27 & 62 & 59\\ 11 & 13 & 124\end{bmatrix}$	$\begin{bmatrix}123 & 23 & 2\\ 36 & 46 & 66\\ 9 & 46 & 93\end{bmatrix}$
12	0.8243	0.6937	0.4032	$\begin{bmatrix}122 & 19 & 7\\ 25 & 113 & 10\\ 11 & 6 & 131\end{bmatrix}$	$\begin{bmatrix}129 & 15 & 4\\ 25 & 68 & 55\\ 3 & 34 & 111\end{bmatrix}$	$\begin{bmatrix}59 & 60 & 29\\ 0 & 12 & 136\\ 0 & 40 & 108\end{bmatrix}$
13	0.6171	0.5923	0.4527	$\begin{bmatrix}118 & 11 & 19\\ 23 & 74 & 51\\ 27 & 39 & 82\end{bmatrix}$	$\begin{bmatrix}135 & 8 & 5\\ 30 & 102 & 16\\ 50 & 72 & 26\end{bmatrix}$	$\begin{bmatrix}123 & 24 & 1\\ 66 & 74 & 8\\ 33 & 111 & 4\end{bmatrix}$
14	0.3941	0.5225	0.3311	$\begin{bmatrix}27 & 0 & 121\\ 9 & 0 & 139\\ 0 & 0 & 148\end{bmatrix}$	$\begin{bmatrix}148 & 0 & 0\\ 134 & 0 & 14\\ 64 & 0 & 84\end{bmatrix}$	$\begin{bmatrix}0 & 0 & 148\\ 3 & 0 & 145\\ 1 & 0 & 147\end{bmatrix}$
15	0.5901	0.5495	0.5991	$\begin{bmatrix}126 & 0 & 22\\ 10 & 4 & 134\\ 16 & 0 & 132\end{bmatrix}$	$\begin{bmatrix}101 & 4 & 43\\ 54 & 5 & 89\\ 9 & 1 & 138\end{bmatrix}$	$\begin{bmatrix}116 & 1 & 31\\ 41 & 3 & 104\\ 1 & 0 & 147\end{bmatrix}$
16	0.8378	0.6351	0.7095	$\begin{bmatrix}126 & 4 & 18\\ 20 & 125 & 3\\ 25 & 2 & 121\end{bmatrix}$	$\begin{bmatrix}34 & 46 & 68\\ 3 & 114 & 31\\ 5 & 9 & 134\end{bmatrix}$	$\begin{bmatrix}139 & 7 & 2\\ 17 & 80 & 51\\ 28 & 24 & 96\end{bmatrix}$
17	0.5721	0.5833	0.6036	$\begin{bmatrix}137 & 9 & 2\\ 88 & 50 & 10\\ 21 & 60 & 67\end{bmatrix}$	$\begin{bmatrix}110 & 26 & 12\\ 15 & 58 & 75\\ 13 & 44 & 91\end{bmatrix}$	$\begin{bmatrix}123 & 21 & 4\\ 58 & 59 & 31\\ 23 & 39 & 86\end{bmatrix}$
18	0.4572	0.4212	0.4640	$\begin{bmatrix}57 & 64 & 27\\ 35 & 93 & 20\\ 24 & 71 & 53\end{bmatrix}$	$\begin{bmatrix}75 & 19 & 54\\ 22 & 13 & 113\\ 41 & 8 & 99\end{bmatrix}$	$\begin{bmatrix}106 & 2 & 40\\ 67 & 1 & 80\\ 49 & 0 & 99\end{bmatrix}$
19	0.4640	0.3221	0.4009	$\begin{bmatrix}82 & 22 & 44\\ 18 & 67 & 63\\ 53 & 38 & 57\end{bmatrix}$	$\begin{bmatrix}117 & 15 & 16\\ 92 & 24 & 32\\ 141 & 5 & 2\end{bmatrix}$	$\begin{bmatrix}43 & 43 & 62\\ 15 & 86 & 47\\ 55 & 44 & 49\end{bmatrix}$
20	0.7568	0.7432	0.3266	$\begin{bmatrix}136 & 7 & 5\\ 40 & 101 & 7\\ 3 & 46 & 99\end{bmatrix}$	$\begin{bmatrix}135 & 10 & 3\\ 12 & 88 & 48\\ 4 & 37 & 107\end{bmatrix}$	$\begin{bmatrix}137 & 10 & 1\\ 141 & 7 & 0\\ 131 & 16 & 1\end{bmatrix}$
21	0.7860	0.6824	0.5135	$\begin{bmatrix}148 & 0 & 0\\ 3 & 98 & 47\\ 12 & 33 & 103\end{bmatrix}$	$\begin{bmatrix}144 & 0 & 4\\ 3 & 88 & 57\\ 7 & 70 & 71\end{bmatrix}$	$\begin{bmatrix}89 & 6 & 53\\ 21 & 81 & 46\\ 14 & 76 & 58\end{bmatrix}$
22	0.3333	0.3333	0.3333	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$
23	0.5856	0.5495	0.4910	$\begin{bmatrix}101 & 4 & 43\\ 7 & 73 & 68\\ 21 & 41 & 86\end{bmatrix}$	$\begin{bmatrix}121 & 9 & 18\\ 29 & 42 & 77\\ 32 & 35 & 81\end{bmatrix}$	$\begin{bmatrix}102 & 16 & 30\\ 29 & 40 & 79\\ 18 & 54 & 76\end{bmatrix}$
24	0.3333	0.3333	0.3333	$\begin{bmatrix}148 & 0 & 0\\ 148 & 0 & 0\\ 148 & 0 & 0\end{bmatrix}$	$\begin{bmatrix}148 & 0 & 0\\ 148 & 0 & 0\\ 148 & 0 & 0\end{bmatrix}$	$\begin{bmatrix}148 & 0 & 0\\ 148 & 0 & 0\\ 148 & 0 & 0\end{bmatrix}$
26	0.9572	0.9550	0.3356	$\begin{bmatrix}148 & 0 & 0\\ 18 & 130 & 0\\ 1 & 0 & 147\end{bmatrix}$	$\begin{bmatrix}138 & 8 & 2\\ 1 & 146 & 1\\ 2 & 6 & 140\end{bmatrix}$	$\begin{bmatrix}123 & 3 & 22\\ 139 & 7 & 2\\ 111 & 18 & 19\end{bmatrix}$
29	0.7072	0.5788	0.4572	$\begin{bmatrix}127 & 21 & 0\\ 30 & 51 & 67\\ 1 & 11 & 136\end{bmatrix}$	$\begin{bmatrix}134 & 12 & 2\\ 55 & 55 & 38\\ 31 & 49 & 68\end{bmatrix}$	$\begin{bmatrix}134 & 7 & 7\\ 55 & 55 & 38\\ 96 & 38 & 14\end{bmatrix}$
Mean	0.6786	0.6174	0.4408			
Std	0.1826	0.1659	0.1049			

Train and validation set mean accuracies are 68 and 62% for COG_BCIg1 and 58 and 55% for COG_BCIg3. The models can not generalize to the test set such that test set performances are clearly lower than that of train and validation sets (44 and 47%, respectively for COG_BCIg1 and COG_BCIg3). When the train and validation sets are shuffled and split, mean train and validation performance increase to over 98% and 95%, respectively, while standard deviation over subjects decreases from 18 and 15% to 3 and 2%. This suggests poor performance of the experiments COG_BCIg1 and COG_BCIg3 were due to cross-session variability between train and validation sessions. This is supported by SOM plots of features given in 4.3.1.2. The difference in test performances of unshuffled and shuffled train and validation experiments shows the importance of reporting the results on an unseen test set to assess generalizability.

Subject-wise model performance is given in Figures 20 - 21 for COG_BCIg1, COG_BCIg2, COG_BCIg3 and COG_BCIg4 models, respectively. When train and validation sessions are shuffled and split to generate train and validation sets in experiments COG_BCIg2 and COG_BCIg4, increased train and validation performance is observed for all subjects. Nonetheless, test accuracies still decrease and cannot increase significantly when compared to COG_BCIg1 and COG_BCIg3 experiments where train and validation sessions are kept independent.

Failure to generalize persist for other COG_BCIg1 variation experiments although there are small improvements in mean test accuracy when using only frontal nodes and 5nn graphs with or without node selection. Among all the COG_BCIg1 variations, only using batch normalization could improve train and validation accuracies although the improvement was not reciprocated in the test set. Subject-wise results of these experiments are given in Appendix C.

The results of the binary classification model are presented in Table 17 compared to the baseline model, COG_BCIg1, averaged across subjects. An improvement in accuracies is observed, also after correcting for chance level, indicating that the Easy and Difficult conditions are more separable, as also evidenced by the exploratory analysis of features in Section 4.3.1.2.

Table 17: Binary Classification Results Compared to COG_BCIg1

Model Name	**Train Accuracy (mean $\pm$ std)**	**Validation Accuracy (mean $\pm$ std)**	**Test Accuracy (mean $\pm$ std)**
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg1_binary	0.8744 ± 0.1567	0.8576 ± 0.1568	0.6631 ± 0.1647

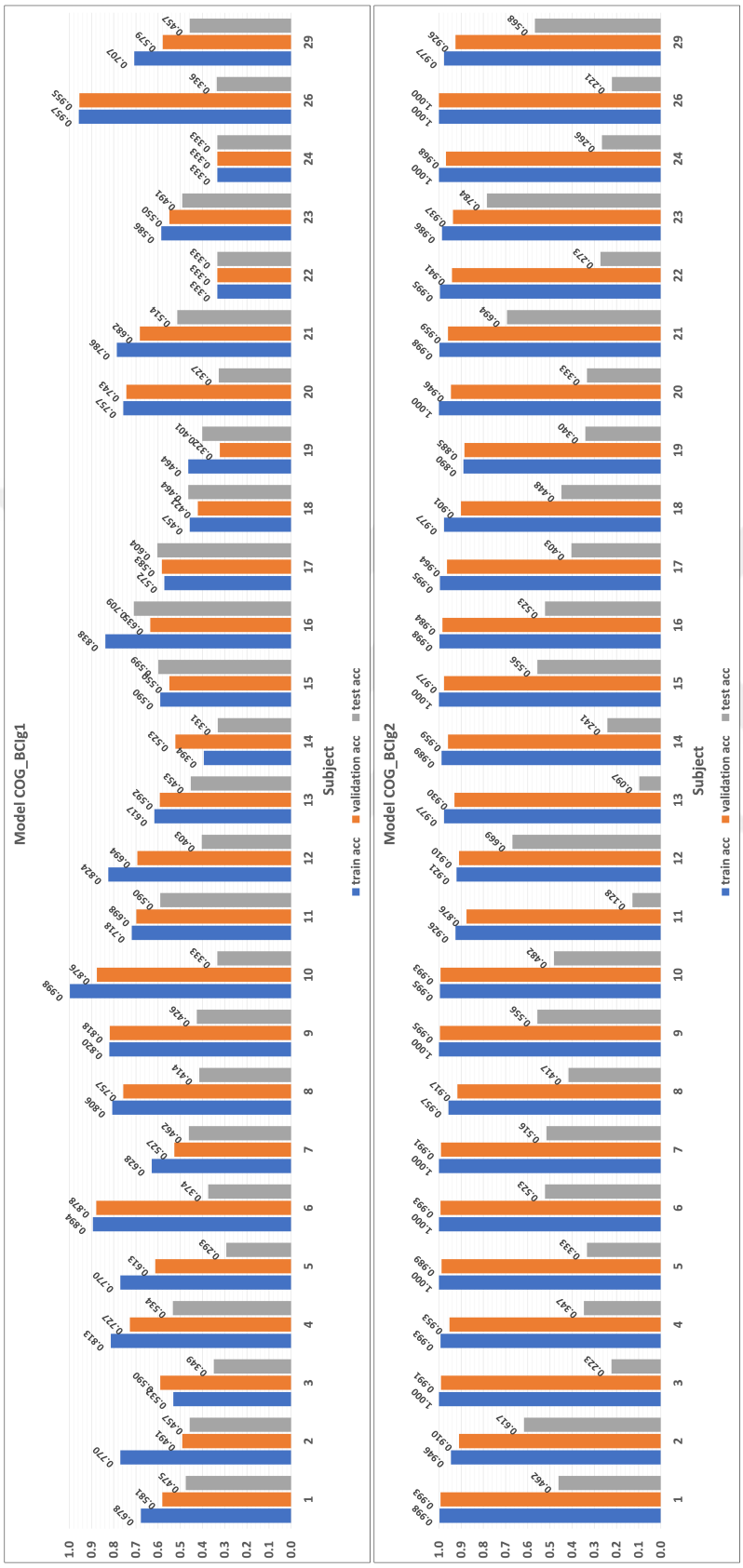


Figure 20: Subject-Wise Results for Models COG_BCIg1 and COG_BCIg2

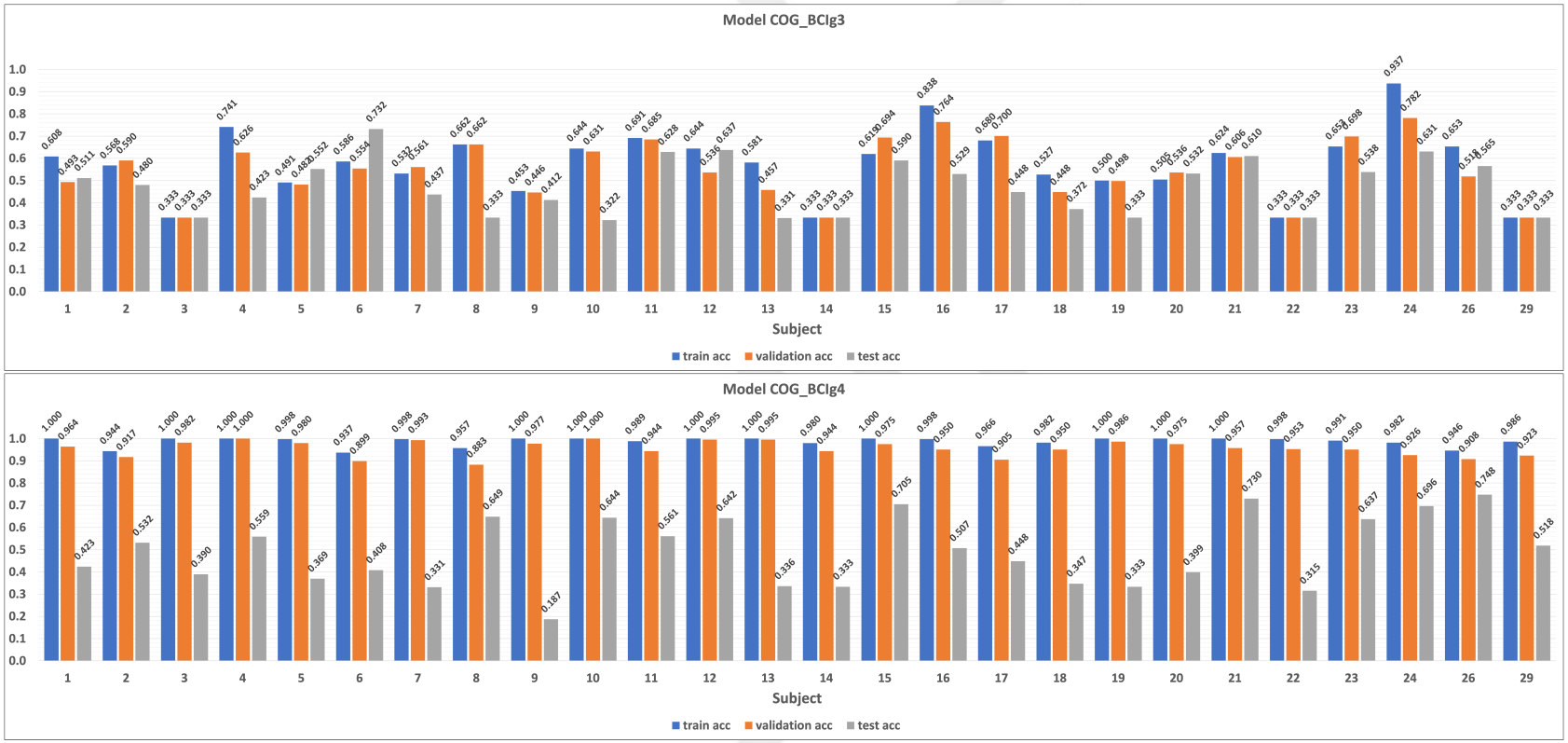


Figure 21: Subject-Wise Results for Models COG_BCIg3 and COG_BCIg4

4.3.4.2 Significance of Experiment Results

To validate the observations discussed in the previous section, train, validation and test accuracy values of the selected models for each subject were used for statistical comparisons, as described in 4.1.6.4. This comparison involved 26 data points per experiment, corresponding to the number of subjects. All computations were performed using MATLAB.

The null hypothesis for the Friedman test states that the performance metric is the same for all models. A p-value smaller than $\alpha = 0.05$ for a 95% confidence interval indicates that this hypothesis can be rejected, signifying statistically significant differences between the models. The Friedman test results for the experiment groups are given in Table 18.

The results show statistically significant differences in train and validation accuracies between the baseline COG_BCIg1 experiment and other experiment groups, including data split variation experiments, the batch normalization experiment, and the experiments on epoch length and overlap. However, for experiments involving reduced features and node selection, significant differences were observed only in train accuracy. Importantly, no statistically significant differences were found in test accuracy across any of the experiments.

Pairwise comparisons were conducted to identify significantly better-performing models using Wilcoxon signed-rank tests. A p-value smaller than $\alpha = 0.05$ was considered indicative of statistically significant differences between model pairs. The results of these tests are presented in Table 19 for variations in data splits, Table 20 for experiments involving small architectures, batch normalization, and variations in epoch size and overlap, and Table 21 for experiments with different graph structures, reduced features, and node selection.

In the experiments involving data split variations, statistically significant differences in train accuracies are observed for all model pairs except COG_BCIg2 - COG_BCIg4. For validation accuracies, all pairs except COG_BCIg1 - COG_BCIg3 and COG_BCIg2 - COG_BCIg4 showed significant differences. Models COG_BCIg2 and COG_BCIg4 outperform COG_BCIg1 and COG_BCIg3 in both train and validation accuracies. These results suggest that shuffling sessions to create training and validation sets enhanced training performance by eliminating cross-session variability in these cases.

In experiments involving batch normalization, epoch size, and overlap, significant differences in train accuracies were found for COG_BCIg1_batchNorm, COG_BCIg1_2sec, COG_BCI g1_2sec_no and COG_BCIg1_4sec_no compared to the baseline COG_BCIg1. For validation accuracies, only COG_BCIg1_batchNorm and COG_BCIg1_2sec_no showed significant differences. Among these, COG_BCIg1_batchNorm achieved the highest train and validation accuracies, outperforming the baseline. However, these improvements did not translate to test set performance.

For experiments involving graph structures, reduced features and nodes, significant differences in train accuracies were observed between COG_BCIg1 and g1_frontal, g1_parietal, g1_theta_alpha and g1_5nn, with COG_BCIg1 emerging as the best-performing model. No significant differences were observed in validation accuracies among these experiments.

Additionally, test accuracies did not exhibit statistically significant differences across any of the models, as determined by the Wilcoxon signed-rank tests. Findings from both statistical tests on test dataset accuracies highlight the challenge of achieving consistent generalization across different sessions. While variations in experimental setups can impact training and validation performance, their influence does not necessarily translate to improved generalization on unseen test data. This highlights the necessity of exploring different approaches to enhance cross-session generalizability.

Table 18: Friedman Test Results

Models	Train Accuracy	Validation Accuracy	Test Accuracy
COG_BCIg1 COG_BCIg2 COG_BCIg3 COG_BCIg4	**2.146e-13**	**3.421e-13**	0.653
COG_BCIg1 COG_BCIg1_smallArch	0.221	0.683	0.683
COG_BCIg1 COG_BCIg1_batchNorm	**0.0186**	**0.000415**	0.433
COG_BCIg1 COG_BCIg1_2sec COG_BCIg1_2sec_no COG_BCIg1_4sec_no	**1.731e-05**	**0.0307**	0.58043
COG_BCIg1 COG_BCIg1_3nn COG_BCIg1_5nn	0.326	0.526	0.616
COG_BCIg1 COG_BCIg1_frontal COG_BCIg1_parietal COG_BCIg1_theta_alpha COG_BCIg1_50p_3nn COG_BCIg1_50p_5nn COG_BCIg1_25p_5nn	**0.0228**	0.555	0.650

Table 19: Wilcoxon Test Results for Variations in Data Split Experiments

	g1	g2	g3	g4	Model Name
Train	1.0000	**0.0000**	**0.0066**	**0.0000**	g1
		1.0000	**0.0000**	0.4656	g2
			1.0000	**0.0000**	g3
				1.0000	g4
Validation	1.0000	**0.0000**	0.0875	**0.0000**	g1
		1.0000	**0.0000**	0.9899	g2
			1.0000	**0.0000**	g3
				1.0000	g4
Test	1.0000	0.6751	0.5812	0.2180	g1
		1.0000	0.3473	0.1307	g2
			1.0000	0.5531	g3
				1.0000	g4

4.3.4.3 Model Complexity

Model complexity is often associated with the number of learnable parameters in a neural network. The Vapnik–Chervonenkis (VC) dimension provides a theoretical framework to relate model complexity to the required number of samples for effective learning ([211]). Specifically, the VC dimension quantifies the capacity of a model by determining the largest set of data points it can perfectly classify. A higher VC dimension implies a more complex model, requiring more data to avoid overfitting.

In graph neural networks (GNNs), the VC dimension is not yet well established, primarily due to their unique structures and relational dependencies. However, foundational work such as [212] provides necessary formulations for specific GNN architectures, which can serve as a basis for exploring their theoretical capacity. While the VC dimension sets theoretical boundaries on model complexity, practical performance improvements can often be achieved through methods such as dropout or regularization, which help mitigate overfitting and improve generalization.

For the experiments conducted in this thesis, model complexity was evaluated by calculating the total number of learnable parameters, determined by the number of graph neural network layers, linear layers, and the nodes in these layers that are in the hyperparameter search space outlined in Section 4.2.2. To analyze the relationship between model complexity and performance, joint plots were generated for the trials of COG_BCI baseline experiment and the COG_BCIg1_smallArch experiment, where a smaller search space was used. Underfitting models, identified by penalized heuristic loss values or low-performing models that exceeded a certain threshold, were excluded from these plots to focus on models with reasonable performance. Figures 22 and 23 illustrate these relationships for subject 1 and subject 2, respectively. Figures for all subjects are included in Appendix C.

Table 20: Wilcoxon Test Results for Different Hyperparameter Search Spaces, Epoch Length and Overlap Experiments

	g1	g1_smallArch	g1_batchNorm	g1_2sec	g1_2sec_no	g1_4sec_no	Model Name
Train	1.0000	0.3173	**0.0006**	**0.0012**	**0.0006**	**0.0383**	g1
		1.0000	**0.0001**	**0.0031**	**0.0046**	0.5998	g1_smallArch
			1.0000	**0.0000**	**0.0000**	**0.0001**	g1_batchNorm
				1.0000	0.8433	**0.0074**	g1_2sec
					1.0000	**0.0039**	g1_2sec_no
						1.0000	g1_4sec_no
Validation	1.0000	0.8638	**0.0002**	0.0633	**0.0164**	0.3860	g1
		1.0000	**0.0001**	**0.0050**	**0.0059**	0.1547	g1_smallArch
			1.0000	**0.0000**	**0.0000**	**0.0000**	g1_batchNorm
				1.0000	1.0000	0.1361	g1_2sec
					1.0000	**0.0392**	g1_2sec_no
						1.0000	g1_4sec_no
Test	1.0000	0.6891	0.5677	0.1808	0.6272	0.4114	g1
		1.0000	0.6751	0.7127	0.9678	0.8401	g1_smallArch
			1.0000	0.1513	0.2699	0.3539	g1_batchNorm
				1.0000	0.3615	0.4291	g1_2sec
					1.0000	0.9757	g1_2sec_no
						1.0000	g1_4sec_no

Table 21: Wilcoxon Test Results for Graph Structures and Reduced Features and Nodes Experiments

	g1	g1_frontal	g1_parietal	g1_theta_alpha	g1_3nn	g1_5nn	g1_50p_3nn	g1_50p_5nn	g1_25p_5nn	Model Name
Train	1.0000	**0.0106**	**0.0358**	**0.0068**	0.2375	**0.0258**	0.9544	0.2418	0.2060	g1
		1.0000	0.5720	0.5097	0.1007	0.6476	**0.0450**	0.2035	0.2871	g1_frontal
			1.0000	0.1500	0.3395	0.9658	0.0789	0.6570	0.7570	g1_parietal
				1.0000	**0.0384**	0.1411	**0.0143**	0.0614	**0.0214**	g1_theta_alpha
					1.0000	0.4154	0.2158	0.8191	0.7982	g1_3nn
						1.0000	**0.0162**	0.7642	0.3913	g1_5nn
							1.0000	0.0553	0.1250	g1_50p_3nn
								1.0000	0.8583	g1_50p_5nn
									1.0000	g1_25p_5nn
Validation	1.0000	0.1793	0.2530	0.0926	0.3095	0.0857	0.9095	0.8303	0.7164	g1
		1.0000	0.4928	1.0000	0.4593	0.4839	0.2939	0.3835	0.2472	g1_frontal
			1.0000	0.5539	0.6570	0.7970	0.2713	0.4273	0.5271	g1_parietal
				1.0000	0.3673	0.4273	0.2208	0.4320	0.2086	g1_theta_alpha
					1.0000	0.7959	0.6170	0.6069	0.5449	g1_3nn
						1.0000	0.1208	0.7088	0.2839	g1_5nn
							1.0000	0.8977	0.6668	g1_50p_3nn
								1.0000	0.6265	g1_50p_5nn
									1.0000	g1_25p_5nn
Test	1.0000	0.3818	0.7425	0.3229	0.7317	0.2245	0.5034	0.2208	0.9893	g1
		1.0000	0.3002	0.6570	0.5538	0.7210	0.3532	0.9317	0.6070	g1_frontal
			1.0000	0.2357	0.3808	0.3229	0.8552	0.3064	0.9431	g1_parietal
				1.0000	0.9785	0.8077	0.6070	0.9893	0.8639	g1_theta_alpha
					1.0000	0.8191	0.7265	0.6482	0.7750	g1_3nn
						1.0000	0.0974	0.7960	0.4928	g1_5nn
							1.0000	0.1658	0.9772	g1_50p_3nn
								1.0000	0.4937	g1_50p_5nn
									1.0000	g1_25p_5nn

In these joint plots, the first and second rows compare the number of learnable parameters in the graph layers and linear layers, respectively, while the columns separate experiments for granularity. The first column aggregates all trials from both experiments, while the second and third columns isolate trials from COG_BCIg1 and COG_BCIg1_smallArch, respectively. This comparison allows for a more detailed inspection of how varying model complexity affects model performance. The joint plots illustrate the relationship between the best heuristic loss and the number of learnable parameters. Alongside these scatter plots, histograms at the top and right display the distributions of the number of parameters and the best heuristic loss values, respectively.

Due to the use of Optuna Bayesian search for hyperparameter optimization, the distribution of parameter counts is skewed, with a higher concentration of trials featuring smaller numbers of layers and learnable parameters. This indicates that the optimization process favored models with lower complexity. For subject 1, slightly lower heuristic loss values were observed in experiments with smaller architectures (COG_BCIg1_smallArch in the third column), suggesting marginally improved performance for simpler models. However, this trend was not evident for subject 2. When the differences between the baseline COG_BCIg1 and COG_BCIg1_smallArch experiments were analyzed statistically across all subjects, no significant difference in performance was detected (Section 4.3.4.2). This outcome suggests that reducing model complexity within the explored range does not adversely impact model performance, highlighting the potential for smaller, more efficient architectures in this context.

4.3.4.4 Computation Environment and Effort

Experiments are conducted on four workstations. COG_BCIg1 experiments are performed on a workstation equipped with Intel® Xeon® CPU, 32 GB RAM, and an NVIDIA GeForce RTX 4070 Ti SUPER GPU.

A conda environment with the packages shown in Table 22 with the corresponding versions when applicable.

All experiments are conducted for all subjects, with 150 Optuna trials per subject. The runtime of each trial is recorded. On average, COG_BCIg1 experiments required 98.6 ± 34.0 seconds per trial, calculated over 150 trials and 26 subjects. The total runtime of the COG_BCIg1 experiment for all subjects and trials was approximately 384,710.6 seconds, equivalent to 106.9 hours. Runtime values exclude logging and file transfers outside of the training process.

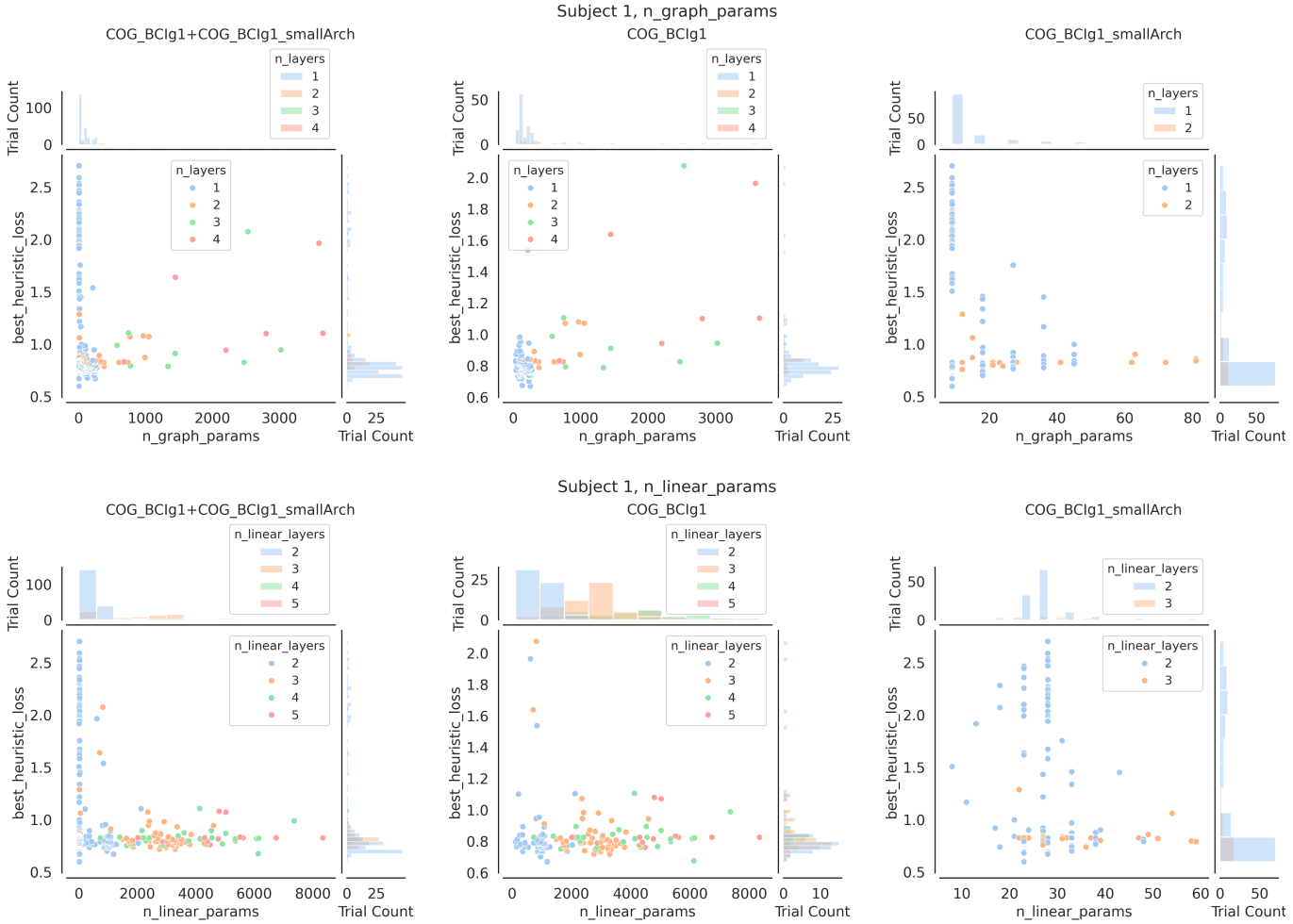


Figure 22: Best Heuristic Loss vs Model Complexity - Subject 1

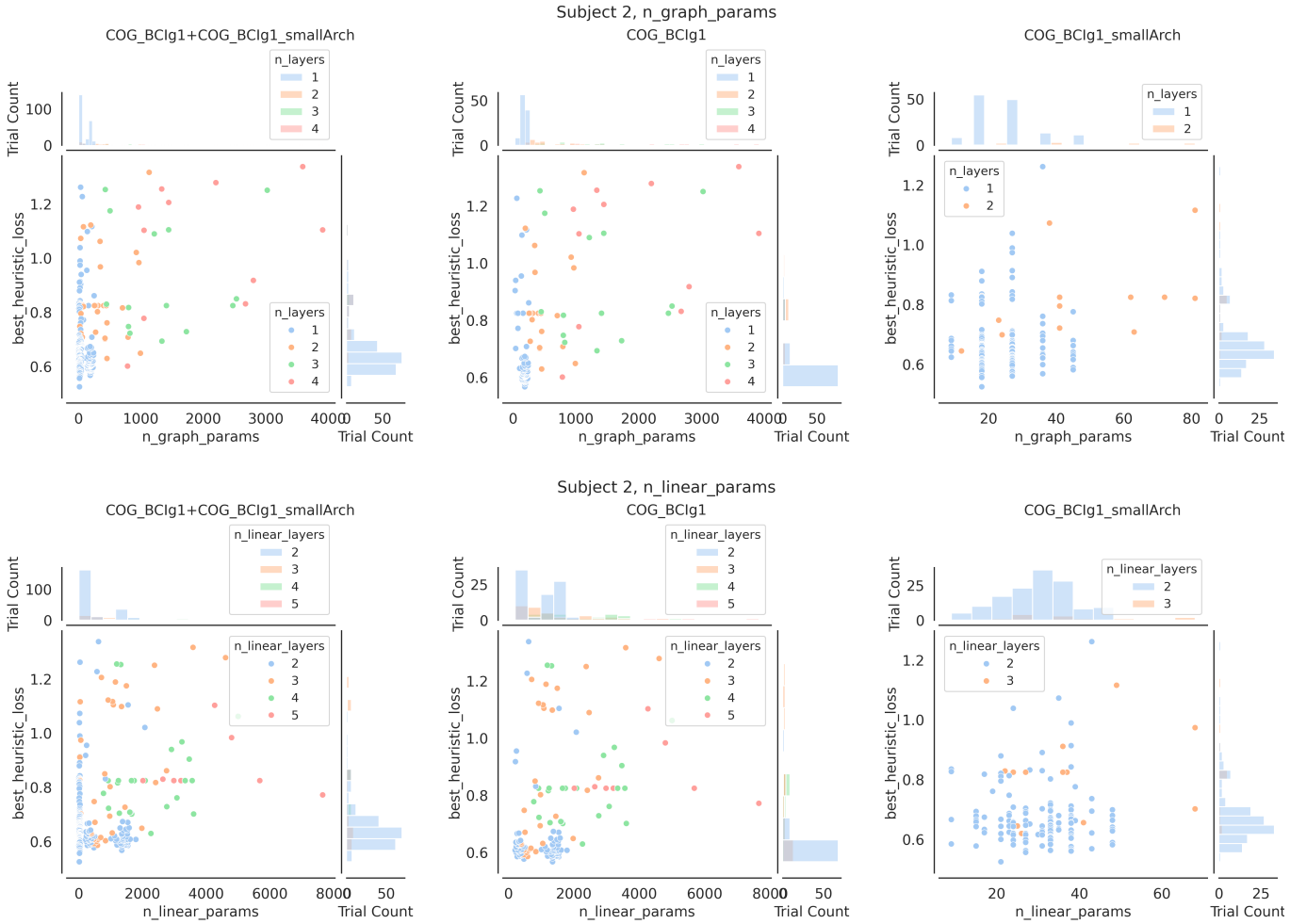


Figure 23: Best Heuristic Loss vs Model Complexity - Subject 2

Table 22: Python Packages and Versions

Package	Version
torch	2.1.2
pyg	-
pytorch-cluster	-
pytorch-scatter	-
matplotlib	-
wandb	-
notebook	-
optuna	-
plotly	-
torchinfo	-
openpyxl	-
pandas	1.5.3

4.3.4.5 Performance in Literature

pBCI competition of the Neuroergonomics Conference 2021: The rankings of the pBCI competition of the Neuroergonomics Conference 2021 are published in [1]. During the competition, data from the third session was not released and the models were tested by the organizers to assess performance on unseen dataset. The adjusted chance level accuracy is 38% with a 95% confidence interval for a 3-class classification problem. Although most of the test results were higher than chance level, the highest test accuracy achieved for the unseen third session was only 54%. These results verify the need to improve and achieve generalizable models for EEG classification in order to use them for long periods. The performance of the winning papers and the methods they used are summarized in Table 23. Riemannian geometry proved successful as three of the first four rankings used it. Another interesting observation is that three worst rankings which are below chance-level used deep learning models with the exception of second paper that used CNN. When validation and test accuracies are compared, it is seen that deep learning models tend to overfit. Using too many parameters and too deep layers can explain the overfitting behaviour.

The experiment results presented in this thesis are not directly comparable to the competition results due to differences in the number of subjects and preprocessing techniques. Nonetheless, findings of both this study and the competition emphasize that cross-session generalization remains a challenge in EEG-based mental workload classification. Despite variations in datasets and methodologies, the consistent decline in model performance when applied to unseen sessions underscores the need for more robust strategies to address session variability in practical BCI applications.

Table 23: Results of pBCI Hackathon Papers ([1])

Ranking	Test accuracy (for the third session)	Validation accuracy	Method
1	54.26%	59.4%	Riemannian Geometry + automatic electrode selection + SVM classifier
2	48.20%	46.46%	Spherical CNN with Rank-1 constraint
3	48.13%	45.30%	Riemannian Geometry + functional connectivity features
4	46.30%	51.25%	Riemannian Geometry
5	44.67%	50.28%	Random forest on classical features (300)
6	43.79%	89.51%	RNN on bandpower + approximate entropy features
7	38.49%	NA	Random forest on CSP features
8	33.18%	58.36%	1 dimension CNN
9	31.89%	94.02%	RNN on frequency, statistical, morphological, time-frequency, linear, and non-linear features
10	31.32%	92.00%	RNN and CNN

COG-BCI Dataset: Four recent publications utilized MATB task of the COG-BCI dataset as done in this thesis.

Wavelet images were used as features in ResNet and EfficientNet models in [213]. Subjects 5, 9, 14 and 17 were excluded due to excessive noise and data issues. Three-class classification (High, Medium, Low MWL) was conducted with cross-session evaluation, yielding test accuracies of 63.94%, 52.96% and 58.50% when sessions 1, 2, and 3 were used as the test set, respectively.

[214] investigated the day-to-day stability of raw PSD features and their periodic and aperiodic components, decomposed using the Fitting Oscillations and One-Over-F algorithm. PSDs were computed using the Welch method, applying Hann windows of 2 seconds with 50% overlap. The best results were obtained using an SVM with an rbf kernel and grid optimization. Data from two days was used, with one day serving as the training set and the other as the test set. Binary classification to distinguish rest from task condition yielded accuracies averaged over all subjects as 93.8% ± 4.6% for the periodic features and 73.4% ± 25.4% for the raw PSD features. Moreover, 3-class (rest vs. low MWL vs. high MWL) and binary (rest vs. task) classification were applied to a self-collected dataset. A significant drop in accuracy was observed for 3-class classification, declining from 84.2% ± 11.0% to 58.9% ± 11.5% when periodic features were used. Raw PSD features resulted in accuracies of 69.9% ± 18.5% in binary classification and 52.7% ± 14.0% in 3-class classification.

Mean PSD, relative PSD, differential entropy, and coherence were utilized as features in [215] with an adaptive classifier for 3-class classification (High, Medium, Low MWL). The approach achieved 74.89% mean accuracy. For this classification, training data included other sessions from the same subject as well as data from other subjects.

Finally, [216] employed semi-supervised cross-task domain adaptation (SCDA) using PSD features. Two prototype classifiers were trained using samples from the source task and calibration samples from the target task. Cross-task classification between n-back and MATB tasks, as well as cross-task, cross-subject classification, was performed. Twenty seconds of target data was used for calibration. Binary classification (Easy vs. Difficult) results showed 96.61% ± 4.35% accuracy for cross-task classification and 90.98% ± 9.36% accuracy for cross-task, cross-subject classification.

Findings from the literature highlight the challenges associated with cross-session classification, primarily due to variability in the data. Additionally, it is evident that 3-class classification typically yields lower performance compared to binary classification. However, the introduction of novel features or the application of domain adaptation techniques has been shown to mitigate these performance limitations. Furthermore, incorporating calibration data from the target task or sessions has demonstrated potential in enhancing model performance by addressing domain shifts.

4.3.4.6 Relation to Task Performance

Figure 24 illustrates the comparison between the mean TRACK and SYSMON performances for session 3 and the test accuracies achieved by the COG_BCIg1 model. Each column in the figure represents one difficulty level (Easy, Medium, or Difficult), and each data point corresponds to a single subject, with labels identifying the subjects. The visualization aims to explore potential relationships between task performance and model accuracy within session 3.

Upon analysis, no relationship was observed between task performance and model results. This lack of correlation suggests that the features extracted from EEG signals, which form the basis of the model's predictions, may capture aspects of mental workload that are not directly reflected in behavioral metrics. Alternatively, this discrepancy could indicate that the behavioral metrics and EEG features respond differently to the cognitive demands of the task, underscoring the complexity of linking behavioral and neural measures of mental workload.

It is also possible that inter-subject variability plays a significant role in obscuring any potential relationships. For instance, some subjects may exhibit strong EEG-based workload patterns that do not align with their behavioral performance, either due to compensatory strategies or individual differences in task execution.

These findings emphasize the need for a broader investigation into the relations between behavioral and neural measures. Future work could benefit from analyzing data across multiple sessions, incorporating additional behavioral metrics, and exploring methods to align neural and behavioral features more effectively. This would provide a more comprehensive under-

standing of how EEG-based workload classification aligns with task performance, potentially leading to improved model designs and enhanced interpretability.

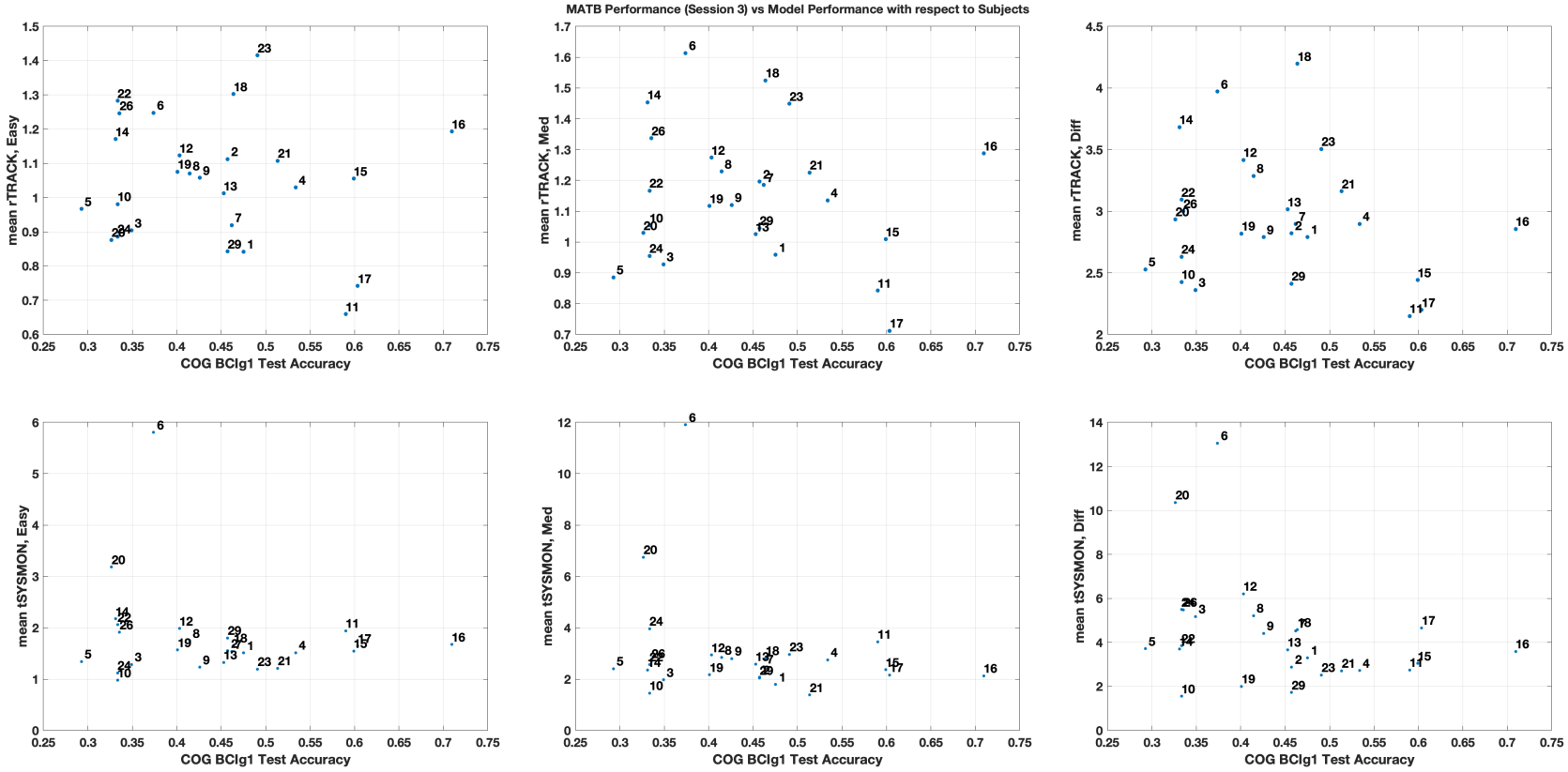


Figure 24: Behavioral Performance Data Comparison to COG_BCIg1 Model Results

4.4 Discussion

A fundamental necessity of machine learning is for the train, validation and test datasets to be sampled from the same distribution ([217]). In the context of EEG-based mental workload classification, this requirement is critical for ensuring that models generalize well to unseen data. However, SOM visualizations of the features show that the data varies considerably between sessions, requiring the models to estimate in regions where they were not trained on. This cross-session variability complicates the ability of machine learning models to generalize, as they are exposed to test data that may differ substantially from the training data distribution.

In the experiments conducted, the best mean test accuracy of 49.02% was achieved for best mean validation accuracy of 95.5%. This significant drop in test performance underscores the difficulties associated with cross-session generalization. For this experiment, session 2 was held out as the test set, with data from sessions 1 and 3 shuffled and split to generate train and validation datasets. The drop in performance suggests that while the model fits well to the training and validation sets, it struggles to generalize to the test set, highlighting the model's sensitivity to session-specific features. This finding is consistent with previous research on cross-session variability and generalization in EEG classification tasks.

The second best test accuracy of 47% was observed when still having session 2 as the test set, but keeping train and validation sets separate as session 1 and 3, respectively. In this case both train and validation accuracies were significantly lower, 58 and 55% respectively. This further emphasizes the impact of cross-session variability on performance. The observed decline in accuracy between the training and validation phases suggests that the model was not able to capture common patterns across sessions, leading to a performance gap.

Similar behavior is observed in [1] when complex models are used. The best three models in pBCI hackathon had test accuracies of 54%, 48% and 48%, respectively. As a side note, these results were obtained using preprocessed data from 15 subjects and the subject level correspondence to the COG-BCI dataset is unknown. Moreover, the data was already preprocessed in the competition and a different preprocessing pipeline can alter the results.

None of the competition entrants used graph neural networks. The findings above from SOM mappings support that even though the graph neural network models fit well to the training dataset, because of the cross-session variations in validation and test datasets, it is very difficult for them to generalize to these sets.

The experiment results, when assessed in conjunction with SOM plots of the features, provide a clear illustration of the cross-session variability problem. The use of SOM visualizations prior to model development serves as a valuable tool for identifying and assessing session-specific differences in the data. By visualizing the feature distributions across different sessions, it becomes possible to detect potential areas of variability that may affect model performance. This preemptive analysis can inform data preprocessing and feature selection strategies, allowing for more effective model development.

Furthermore, subject-specific analysis is crucial for understanding the variability in model performance and highlights the importance of data quality checks before modeling. Given that

EEG signals are highly individualistic, the quality of the data for each subject can significantly impact the overall model performance. Using confusion matrices, in addition to accuracy metrics, provides a more granular understanding of the model's performance. Confusion matrices offer insights into the distribution of classification errors, which helps in identifying specific areas where the model is failing, such as misclassifications between certain classes or sessions.

Despite implementing measures to prevent overfitting and increase the generalizability of the models, such as regularization techniques and reducing model complexity, the performance on the unseen test set still deviated significantly from the training and validation results. This is a well-known challenge in real-world EEG applications, where variability across subjects, sessions, and even data acquisition conditions can severely impact model robustness. The difficulty in achieving consistent performance across different sessions is particularly pronounced in EEG-based BCI systems, where session-specific factors such as electrode placement, skin impedance, and environmental noise can introduce substantial variations.

In the next chapter, techniques to address and accommodate these challenges will be explored. Specifically, domain adaptation using optimal transport will be discussed. This technique aims to reduce the performance gap between train, validation, and test sets, ultimately enabling more reliable and reproducible EEG-based mental workload classification systems for real-world applications.

CHAPTER 5

CROSS-SESSION GENERALIZABILITY OF MACHINE LEARNING MODELS

A major challenge in EEG-based machine learning is the issue of cross-session variability. EEG signals can be highly inconsistent across different sessions due to a variety of factors, including changes in electrode placement, variations in mental state, environmental noise, and physiological differences. Such variability severely impacts the performance of machine learning models trained on data from a single session when they are applied to unseen sessions. Modeling results in Chapter 4 show this challenge, as the test set performance degraded considerably for all subjects (Section 4.3). Ensuring generalizability across sessions is crucial for developing robust and reliable Brain-Computer Interface (BCI) applications, especially in real-world settings where collecting new training data is often impractical. With this motivation, this chapter aims answering the following research question:

- RQ4: How can we mitigate effects of variability between sessions to achieve models that are generalizable over time using domain adaptation techniques?

Several approaches have been explored to address this challenge, including transfer learning, domain adaptation, and normalization techniques. However, these methods often require retraining or fine-tuning, which limits their practical utility. Optimal Transport (OT) theory offers an alternative by enabling the alignment of data distributions without necessitating retraining, making it a powerful tool for cross-session generalizability.

The following sections will first introduce the theoretical aspects of Optimal Transport (OT), providing a background on key concepts such as the Monge problem, the Kantorovich relaxation, and the computation of Wasserstein distances using the Sinkhorn algorithm. Building on this foundation, the application of OT to align EEG data distributions across sessions is described, enhancing the model's ability to generalize without additional training.

5.1 Methodology

5.1.1 The Theory of Optimal Transport

The concept of Optimal Transport was initially introduced by Gaspard Monge in 1781, motivated by the problem of efficiently moving a pile of sand to form a target pile of a desired

shape, minimizing the total cost of transport. The formal problem can be described as finding the minimal effort to morph one probability distribution into another, where effort is quantified by an underlying cost function.

Consider two probability distributions, represented by measures α and β over spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. The goal is to transport the mass from α to β in an optimal way, minimizing a cost function. Optimal Transport has seen a surge in interest in recent years due to its applications in data science, including generative models, domain adaptation, and clustering. The Wasserstein distance, derived from OT, can better capture geometric properties between distributions compared to other divergence metrics, such as Kullback-Leibler divergence [218].

5.1.1.1 The Monge Problem

The original problem posed by Monge can be described as follows: given two probability measures α and β, supported on two spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ that defines the cost of transporting mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$, find a map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes the total transportation cost:

$$\inf_T \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{3}$$

where T is the transport map. The integral $\int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$ represents the total cost of transporting mass from distribution α to β, and $c(x, T(x))$ is the cost of moving a unit mass from point x to point $T(x)$. The function T must satisfy the condition that T pushes forward the measure α onto β, which is written as $T_{\#}\alpha = \beta$. This push-forward condition implies that for any measurable set $B \subset \mathcal{Y}$, the mass transported to B must match the mass of β over B:

$$\int_{T^{-1}(B)} d\alpha(x) = \int_B d\beta(y). \tag{4}$$

One major drawback of Monge's formulation is that it may not always admit a solution, particularly in cases where no feasible map exists that satisfies the push-forward condition [219, 220]. In practice, the Monge problem can be highly non-linear and non-convex, which complicates its solution. Moreover, in certain cases, it is impossible to guarantee the existence of a continuous transport map, especially when dealing with measures that are not absolutely continuous with respect to each other. To overcome this, alternative formulations, such as the Kantorovich relaxation, have been developed.

The Monge formulation also has strong connections to other fields of mathematics. In the context of partial differential equations, the Monge-Ampère equation is associated with the problem of finding optimal transport maps. This connection has enabled advances in both fields by utilizing tools from OT to solve PDEs, and vice versa.

5.1.1.2 The Kantorovich Relaxation

To address the challenges associated with Monge's formulation, Leonid Kantorovich introduced a relaxation by allowing mass splitting. Instead of restricting to a deterministic map T, Kantorovich proposed finding a joint distribution π on $\mathcal{X} \times \mathcal{Y}$, which represents the amount of mass moved from x to y. This relaxed problem is written as:

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y), \tag{5}$$

where π is called a *coupling* of α and β. The set $\Pi(\alpha,\beta)$ is the set of all couplings of α and β, defined as:

$$\Pi(\alpha,\beta) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : (P_{\mathcal{X}})_{\#}\pi = \alpha, (P_{\mathcal{Y}})_{\#}\pi = \beta\}, \tag{6}$$

where $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ represents the set of probability measures on $\mathcal{X}\times\mathcal{Y}$. The operators $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are the projections onto $\mathcal{X}$ and $\mathcal{Y}$, respectively, ensuring that the marginals of π are indeed α and β. Unlike the Monge problem, the Kantorovich formulation is always feasible, and it admits a solution under mild conditions [220, 221].

The Kantorovich problem is particularly attractive because it can be formulated as a linear programming problem, which guarantees the existence of efficient solution methods. Modern approaches use computational methods, such as network simplex algorithms, interior point methods, and entropic regularization, which make the Kantorovich formulation highly suitable for large-scale data problems.

The Kantorovich relaxation has numerous advantages beyond computational feasibility. It allows for probabilistic transport in cases where no deterministic map can be found, thus providing a generalized solution to optimal mass transfer.

5.1.1.3 Dual Formulation

The Kantorovich problem is a linear program, and it admits a dual formulation that can provide more insight into its properties. The dual problem is given by:

$$\sup_{(f,g)\in \mathrm{ADM}(c)} \left[\int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y) \right], \tag{7}$$

where the set of admissible functions $\mathrm{ADM}(c)$ consists of pairs of functions (f,g) such that:

$$f(x) + g(y) \leq c(x,y), \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}. \tag{8}$$

In these equations, f and g are called *Kantorovich potentials*, and they can be interpreted as dual variables that measure the benefit of having mass at specific points in $\mathcal{X}$ and $\mathcal{Y}$. The

dual formulation plays a crucial role in understanding the properties of OT distances, as it allows us to interpret the OT problem in terms of potential functions that capture the balance between mass movement costs [222]. The duality between the primal and dual formulations also provides important geometric insights, as it links OT to the geometry of functional spaces, allowing for the development of numerical techniques based on gradient flows and convex optimization.

The dual formulation also provides an efficient way to approximate OT distances in practice. Leveraging the computational properties of the dual allows one to avoid constructing explicit transport plans and instead focus on optimizing potential functions, which often involves fewer variables.

5.1.1.4 Wasserstein Distance

The Kantorovich formulation allows us to define a distance between probability measures, known as the Wasserstein distance. For a cost function $c(x, y) = d(x, y)^p$, where $d(\cdot, \cdot)$ is a metric on $\mathcal{X}$, the p-Wasserstein distance between α and β is defined as:

$$W_p(\alpha, \beta) = \left(\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} d(x, y)^p d\pi(x, y) \right)^{1/p}. \tag{9}$$

In this equation, $W_p(\alpha, \beta)$ represents the Wasserstein distance of order p between the probability measures α and β. The Wasserstein distance W_p is a true metric on the space of probability measures, and it has many desirable properties, including the ability to quantify differences between distributions with non-overlapping supports, unlike metrics such as the Kullback-Leibler divergence [219, 223].

The Wasserstein distance, often referred to as the Earth Mover's Distance (EMD), captures the idea of moving "mass" in a geometrically meaningful way, making it suitable for tasks in data science where spatial relationships matter. The Wasserstein distance is especially useful in comparing empirical distributions, which frequently arise in machine learning as samples from complex distributions. In particular, the Wasserstein distance has been successfully used in generative adversarial networks (GANs) to provide a stable and meaningful loss function for training, leading to the development of Wasserstein GANs (WGANs) [218].

The Wasserstein metric also has a unique differentiable structure, which makes it suitable for optimization problems involving probability measures. This differentiable structure, often referred to as the Riemannian structure of the Wasserstein space, allows for the development of optimization algorithms based on gradient flows, making OT a powerful tool in continuous optimization problems over probability distributions [224].

5.1.1.5 Entropic Regularization

One of the limitations of the original Kantorovich formulation is its computational complexity, especially when dealing with large datasets. To address this issue, entropic regularization

has been proposed to make the problem more tractable. By adding an entropy term to the objective, we obtain the regularized OT problem:

$$L_\epsilon(\alpha, \beta) = \inf_{\pi \in \Pi(\alpha,\beta)} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon H(\pi) \right], \tag{10}$$

where $H(\pi) = -\int_{\mathcal{X}\times\mathcal{Y}} \pi(x, y) \log \pi(x, y)\, dx\, dy$ is the entropy of the transport plan π, and $\epsilon > 0$ is a regularization parameter. The entropy term encourages smoother transport plans, effectively spreading the mass more evenly and avoiding overly concentrated solutions.

This entropic regularization allows for the use of the Sinkhorn algorithm, an iterative method that is computationally efficient and scales well with large problems [225]. The Sinkhorn algorithm leverages the fact that the entropic regularization makes the objective strictly convex, thereby ensuring a unique solution and significantly improving computational efficiency. The regularized OT problem can be solved using matrix scaling methods, which are faster than traditional linear programming techniques for OT.

Entropic regularization has important applications in machine learning and statistics. It allows for faster computation of distances between empirical distributions, which is especially useful in high-dimensional settings. Additionally, the parameter ϵ can be used to control the trade-off between the exactness of the transport plan and computational efficiency, making it a practical approach for large-scale problems [226]. The regularized OT framework also enables differentiability, which is advantageous when OT is integrated into more complex learning models, such as deep neural networks.

5.1.1.6 Sinkhorn's Algorithm

The Sinkhorn algorithm solves the entropic regularization problem by iteratively scaling the rows and columns of a kernel matrix $K = e^{-C/\epsilon}$, where C is the cost matrix. Let u and v be scaling vectors initialized to $\mathbf{1}$ (the all-ones vector). The update steps are given by:

$$u \leftarrow \frac{\alpha}{Kv}, \quad v \leftarrow \frac{\beta}{K^T u}, \tag{11}$$

where the divisions are elementwise. The vectors u and v are used to iteratively normalize the rows and columns of K so that the resulting matrix represents an approximate coupling between α and β. The process is repeated until convergence, yielding an approximate solution to the optimal transport problem that is faster and more memory-efficient compared to solving the original linear program [226].

The Sinkhorn algorithm has several advantages: It provides a natural way to approximate OT distances while maintaining differentiability, crucial for incorporating OT into end-to-end learning systems; it can be easily parallelized; and its convergence properties have been well-studied. However, the Sinkhorn algorithm does not scale well when dealing with large discrete distributions or continuous measures, requiring the use of other approaches [227].

Another advantage of the Sinkhorn algorithm is its robustness to noise. By introducing entropy into the transport plan, the algorithm becomes less sensitive to small perturbations in the data, which can be beneficial when working with empirical distributions derived from noisy measurements, such as EEG signals in brain-computer interface applications.

5.1.2 Optimal Transport for Cross-Session Estimation of MWL using EEG

Given the inherent variability in EEG data across sessions, Optimal Transport (OT) provides a promising framework for domain adaptation in the context of mental workload (MWL) estimation. The core idea is to align the feature distributions across different sessions using OT, thereby reducing discrepancies and improving the model's generalizability.

In this study, a backward optimal transport domain adaptation strategy, inspired by [73], is employed. This method involves aligning the feature distributions of the validation and test sets to match the training set distribution. By transforming the features into the domain of the training dataset, the method ensures compatibility with the pre-trained model, avoiding the need for retraining. During inference, predictions are made exclusively using the features transported to the training dataset domain.

This approach offers several key advantages:

- No Retraining Required: Traditional domain adaptation methods often require fine-tuning the model on new data. In contrast, OT-based alignment directly operates on features, allowing the pre-trained model to remain unchanged.

- Robustness to Variability and Improved Generalization: By minimizing the Wasserstein distance between the feature distributions from different sessions, OT effectively mitigates session-to-session variability. This alignment addresses the non-stationarity of EEG signals, improving the model's performance on unseen data.

For the estimation of mental workload, the same graph representation samples used in the development of the GNN models (discussed in Chapter 4 are employed. OT is utilized to compute the optimal mapping of feature distributions from the validation and test sets to the training set distribution (Figure 25). This mapping is computed using the Sinkhorn algorithm, which provides an efficient entropic regularization of the OT problem.

The implementation leverages the *pot* Python package ([228]). The regularization parameter for Sinkhorn transport can be adjusted in this package. In this study, it is initialized to a small value (1×10^{-3}) and gradually increased until convergence is achieved. This iterative approach ensures that the parameter is fine-tuned to approximate the ideal transport solution.

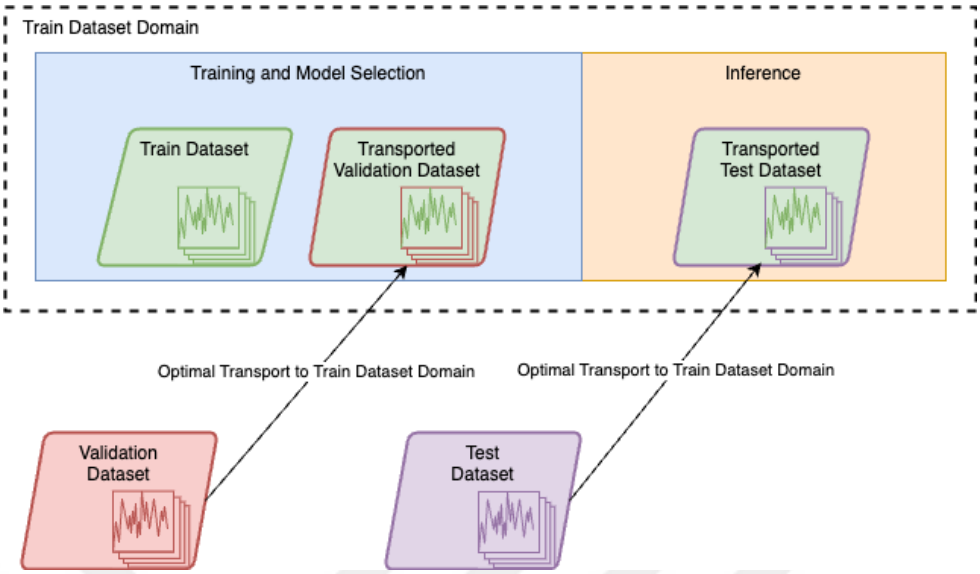


Figure 25: Optimal Transport from Validation and Test Datasets to Train Dataset Domain

5.2 Experiments

Domain adaptation using OT is applied on the experiments defined in Chapter 4, Table 11 and for the binary classification model. The traceability of these experiments with their OT variants is given in Table 24.

Table 24: OT Domain Adaptation Experiments

Baseline Model Name	OT Domain Adaptation Model Name
COG_BCIg1	COG_BCIg5
COG_BCIg2	COG_BCIg6
COG_BCIg3	COG_BCIg7
COG_BCIg4	COG_BCIg8
COG_BCIg1_binary	COG_BCIg5_binary

5.3 Results

5.3.1 Optimal Transport of Features

Figures 26 and 27 present SOM plots of features across sessions before and after applying optimal transport (OT) for Subjects 1 and 2. The experiments involve using Sessions 1, 2, and 3 as the training, validation, and test sets, respectively, corresponding to experiments COG_BCIg1 and COG_BCIg5. The SOM plots are two-dimensional, with the feature size reduced to two during the SOM process for visualization. The first row in each figure rep-

resents features prior to transport, while the second row illustrates the transported validation and test set features in columns 2 and 3. OT enhances class separation; however, label shift persists, as true labels are not utilized during the transport process.

5.3.2 Model Performance

The train, validation and test set accuracies averaged over subjects are presented in Table 25 for the experiments with variations in data split considering sessions with and without optimal transport. All models showed an increase in the accuracy of the mean test set, hence the generalization performance, when OT was applied.

Table 25: Experiments with and without Optimal Transport

Model Name	**Data Split**	**Train Accuracy (mean $\pm$ std)**	**Val Accuracy (mean $\pm$ std)**	**Test Accuracy (mean $\pm$ std)**
COG_BCIg1	Train:1, Val:2 Test:3	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg2	Train, Val: 1+2, 50-50 Test:3	0.9815 ± 0.0295	0.9535 ± 0.0374	0.4238 ± 0.1762
COG_BCIg3	Train:1, Val:3 Test:2	0.5796 ± 0.1493	0.5500 ± 0.1329	0.4724 ± 0.1235
COG_BCIg4	Train, Val: 1+3, 50-50 Test:2	0.9866 ± 0.0198	0.9551 ± 0.0342	0.4902 ± 0.1557
COG_BCIg5 (COG_BCIg1+OT)	Train:1, Val:2 Test:3	0.6483 ± 0.1428	0.6114 ± 0.1336	$\mathbf{0.5054 \pm 0.1213}$
COG_BCIg6 (COG_BCIg2+OT)	Train, Val: 1+2, 50-50 Test:3	0.9111 ± 0.0731	0.8604 ± 0.0755	$\mathbf{0.5236 \pm 0.1116}$
COG_BCIg7 (COG_BCIg3+OT)	Train:1, Val:3 Test:2	0.5742 ± 0.1385	0.5274 ± 0.1149	$\mathbf{0.5289 \pm 0.1171}$
COG_BCIg8 (COG_BCIg4+OT)	Train, Val: 1+3, 50-50 Test:2	0.9096 ± 0.0739	0.8539 ± 0.0695	$\mathbf{0.5612 \pm 0.0953}$

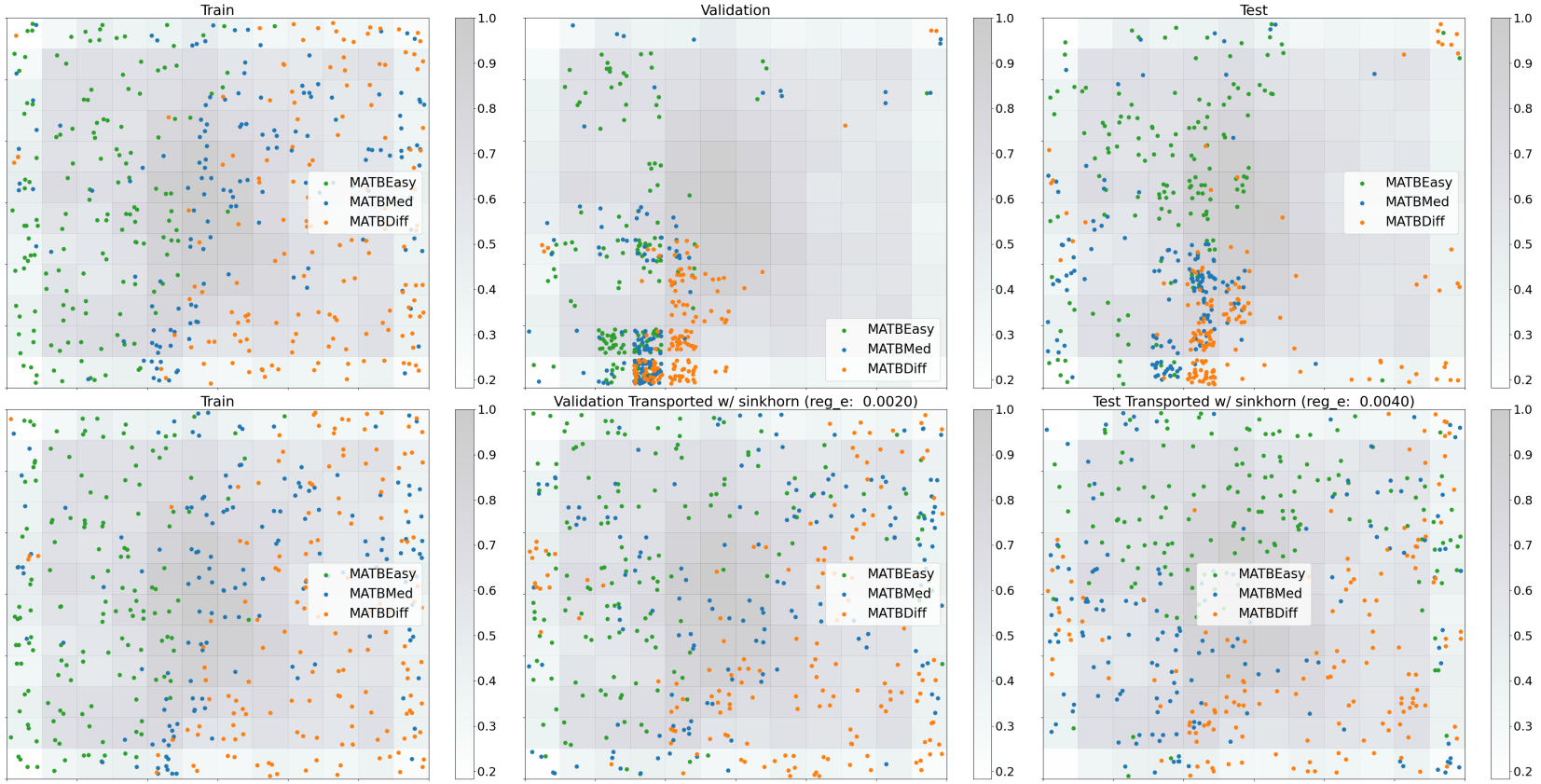


Figure 26: SOM Plots of Features (Subject 1)

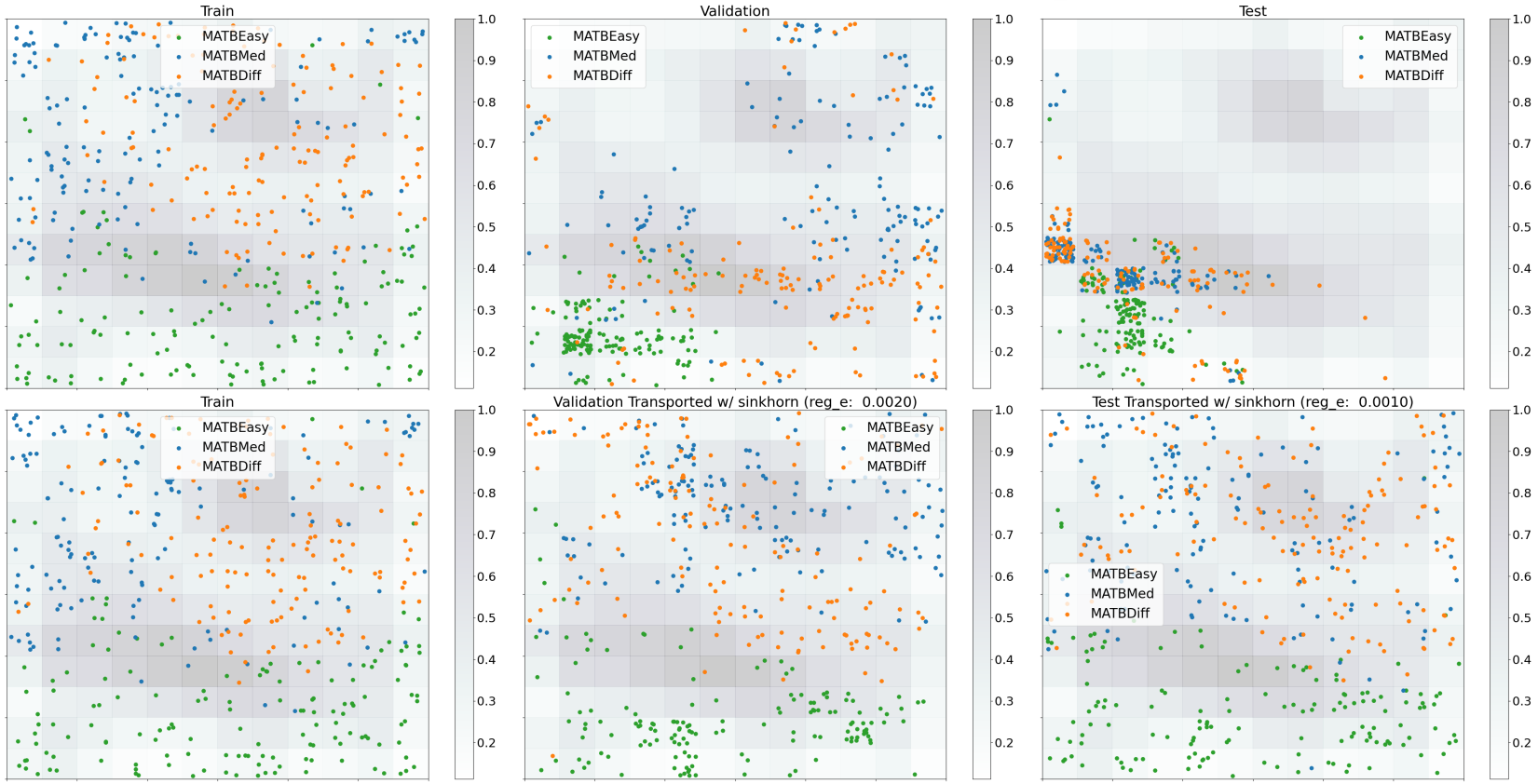


Figure 27: SOM Plots of Features (Subject 2)

The results of the binary classification model are presented in Table 26 compared to the baseline models, COG_BCIg1 and COG_BCIg5, averaged across subjects. An improvement in accuracies is observed, indicating that the Easy and Difficult conditions are more separable, as also evidenced by the exploratory analysis of features in Section 4.3.1.2. Moreover, optimal transport achieved higher performance for the binary classification as well.

Table 26: Binary Classification Results Compared to COG_BCIg1 and COG_BCIg5

Model Name	**Train Accuracy (mean ± std)**	**Validation Accuracy (mean ± std)**	**Test Accuracy (mean ± std)**
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg5	0.6483 ± 0.1428	0.6114 ± 0.1336	**0.5054 ± 0.1213**
COG_BCIg1_binary	0.8744 ± 0.1567	0.8576 ± 0.1568	0.6631 ± 0.1647
COG_BCIg5_binary	0.8526 ± 0.1560	0.8316 ± 0.1505	**0.7216 ± 0.1840**

Subject based results including confusion matrices are presented in Table 27 for COG_BCIg5 experiment. Subject based performance comparison of experiments COG_BCIg1 and COG_BCIg5 is shown in Figure 28.

SOM visualization for Subject 1 (Figure 26) reveals distinct patterns in the separation of conditions across feature sets and data splits. For the features before transport, samples from the Easy condition were more distinctly separated compared to other conditions. Following optimal transport, the separation improved for both Easy and Difficult condition samples in the train and validation sets. However, in the test set, a different pattern emerged: Difficult condition samples were well-distinguished, whereas Easy condition samples overlapped with those from the Medium condition. These observations are consistent with the confusion matrices for the respective models. For the COG_BCIg1 model, the Easy condition exhibited a higher number of true positives, reflecting the improved separability of these samples in the SOM plot. Similarly, for the COG_BCIg5 model, both Easy and Difficult conditions showed a higher number of true positives in the train and validation sets. In contrast, for the test set, the overlap between Easy and Medium conditions in the transported feature space likely contributed to fewer true positives for the Easy condition. The model performed better for the Difficult condition only, aligning with the improved separability of these samples in the SOM visualization.

This analysis demonstrates how SOM visualizations can complement confusion matrix evaluations by providing deeper insights into feature separability and model performance. The findings, along with the consistency between the model results and the observations from the SOM plots, highlight the value and importance of detailed and innovative data exploration prior to model development. SOM visualizations provided critical insights into the separability of features across conditions and data splits, offering a deeper understanding of the challenges and opportunities in the feature space. Such exploratory analyses enable informed adjustments to preprocessing and modeling pipelines, ultimately improving the interpretability and robustness of the results.

For certain conditions, however, optimal transport was unable to fully align the features. This limitation stems from the unsupervised nature of the approach, which did not utilize source

or target labels during transport. Incorporating a portion of these labels in future work could enhance the feature mapping process, potentially leading to improved alignment across conditions. This semi-supervised refinement could be particularly beneficial for conditions where separability remains challenging.

These insights emphasize the necessity of combining innovative data exploration techniques, such as SOM visualizations, with iterative refinements in preprocessing and modeling strategies. By integrating domain knowledge and leveraging a portion of labeled data, future approaches could achieve better generalization across data splits, leading to more robust EEG-based workload classification models.

Table 27: COG_BCIg5 Results with respect to Subjects

Subject	Accuracy			Confusion Matrix		
	Train	Validation	Test	Train	Validation	Test
1	0.7320	0.6847	0.4820	$\begin{bmatrix}134 & 14 & 0\\ 25 & 92 & 31\\ 2 & 47 & 99\end{bmatrix}$	$\begin{bmatrix}105 & 38 & 5\\ 28 & 78 & 42\\ 1 & 26 & 121\end{bmatrix}$	$\begin{bmatrix}48 & 85 & 15\\ 56 & 53 & 39\\ 12 & 23 & 113\end{bmatrix}$
2	0.6914	0.6554	0.6847	$\begin{bmatrix}144 & 0 & 4\\ 14 & 97 & 37\\ 9 & 73 & 66\end{bmatrix}$	$\begin{bmatrix}137 & 1 & 10\\ 0 & 107 & 41\\ 3 & 98 & 47\end{bmatrix}$	$\begin{bmatrix}138 & 0 & 10\\ 4 & 111 & 33\\ 7 & 86 & 55\end{bmatrix}$
3	0.7342	0.6554	0.3671	$\begin{bmatrix}135 & 11 & 2\\ 53 & 57 & 38\\ 3 & 11 & 134\end{bmatrix}$	$\begin{bmatrix}120 & 9 & 19\\ 71 & 45 & 32\\ 9 & 13 & 126\end{bmatrix}$	$\begin{bmatrix}60 & 10 & 78\\ 82 & 32 & 34\\ 54 & 23 & 71\end{bmatrix}$
4	0.7680	0.6982	0.5743	$\begin{bmatrix}134 & 11 & 3\\ 9 & 85 & 54\\ 0 & 26 & 122\end{bmatrix}$	$\begin{bmatrix}125 & 14 & 9\\ 30 & 67 & 51\\ 0 & 30 & 118\end{bmatrix}$	$\begin{bmatrix}99 & 24 & 25\\ 58 & 51 & 39\\ 5 & 38 & 105\end{bmatrix}$
5	0.6239	0.5450	0.6081	$\begin{bmatrix}18 & 111 & 19\\ 3 & 124 & 21\\ 0 & 13 & 135\end{bmatrix}$	$\begin{bmatrix}5 & 113 & 30\\ 9 & 126 & 13\\ 0 & 37 & 111\end{bmatrix}$	$\begin{bmatrix}11 & 104 & 33\\ 0 & 140 & 8\\ 0 & 29 & 119\end{bmatrix}$
6	0.8266	0.8131	0.5405	$\begin{bmatrix}136 & 12 & 0\\ 1 & 111 & 36\\ 3 & 25 & 120\end{bmatrix}$	$\begin{bmatrix}138 & 10 & 0\\ 0 & 129 & 19\\ 0 & 54 & 94\end{bmatrix}$	$\begin{bmatrix}89 & 44 & 15\\ 37 & 76 & 35\\ 9 & 64 & 75\end{bmatrix}$
7	0.6194	0.6126	0.5856	$\begin{bmatrix}118 & 25 & 5\\ 26 & 73 & 49\\ 19 & 45 & 84\end{bmatrix}$	$\begin{bmatrix}92 & 41 & 15\\ 31 & 70 & 47\\ 12 & 26 & 110\end{bmatrix}$	$\begin{bmatrix}80 & 46 & 22\\ 44 & 71 & 33\\ 4 & 35 & 109\end{bmatrix}$
8	0.7635	0.7275	0.6419	$\begin{bmatrix}135 & 13 & 0\\ 6 & 102 & 40\\ 1 & 45 & 102\end{bmatrix}$	$\begin{bmatrix}138 & 10 & 0\\ 3 & 105 & 40\\ 0 & 68 & 80\end{bmatrix}$	$\begin{bmatrix}94 & 45 & 9\\ 47 & 92 & 9\\ 0 & 49 & 99\end{bmatrix}$
9	0.7568	0.7230	0.6104	$\begin{bmatrix}115 & 21 & 12\\ 17 & 95 & 36\\ 0 & 22 & 126\end{bmatrix}$	$\begin{bmatrix}120 & 22 & 6\\ 8 & 86 & 54\\ 1 & 32 & 115\end{bmatrix}$	$\begin{bmatrix}78 & 47 & 23\\ 22 & 85 & 41\\ 7 & 33 & 108\end{bmatrix}$
10	0.5631	0.5586	0.5045	$\begin{bmatrix}135 & 12 & 1\\ 44 & 97 & 7\\ 14 & 116 & 18\end{bmatrix}$	$\begin{bmatrix}123 & 21 & 4\\ 15 & 120 & 13\\ 27 & 116 & 5\end{bmatrix}$	$\begin{bmatrix}104 & 43 & 1\\ 44 & 102 & 2\\ 15 & 115 & 18\end{bmatrix}$
11	0.6824	0.6892	0.5676	$\begin{bmatrix}117 & 18 & 13\\ 31 & 77 & 40\\ 16 & 23 & 109\end{bmatrix}$	$\begin{bmatrix}128 & 17 & 3\\ 41 & 76 & 31\\ 19 & 27 & 102\end{bmatrix}$	$\begin{bmatrix}122 & 24 & 2\\ 48 & 44 & 56\\ 13 & 49 & 86\end{bmatrix}$
12	0.6937	0.6779	0.5338	$\begin{bmatrix}116 & 25 & 7\\ 39 & 66 & 43\\ 6 & 16 & 126\end{bmatrix}$	$\begin{bmatrix}133 & 13 & 2\\ 25 & 67 & 56\\ 10 & 37 & 101\end{bmatrix}$	$\begin{bmatrix}120 & 16 & 12\\ 15 & 48 & 85\\ 32 & 47 & 69\end{bmatrix}$
13	0.6104	0.6014	0.4955	$\begin{bmatrix}135 & 9 & 4\\ 28 & 24 & 96\\ 17 & 19 & 112\end{bmatrix}$	$\begin{bmatrix}132 & 9 & 7\\ 16 & 21 & 111\\ 9 & 25 & 114\end{bmatrix}$	$\begin{bmatrix}96 & 15 & 37\\ 52 & 23 & 73\\ 25 & 22 & 101\end{bmatrix}$
14	0.7770	0.6419	0.1396	$\begin{bmatrix}146 & 2 & 0\\ 30 & 55 & 63\\ 0 & 4 & 144\end{bmatrix}$	$\begin{bmatrix}125 & 18 & 5\\ 29 & 34 & 85\\ 7 & 15 & 126\end{bmatrix}$	$\begin{bmatrix}0 & 0 & 148\\ 70 & 37 & 41\\ 102 & 21 & 25\end{bmatrix}$
15	0.6554	0.5698	0.6261	$\begin{bmatrix}128 & 17 & 3\\ 25 & 55 & 68\\ 12 & 28 & 108\end{bmatrix}$	$\begin{bmatrix}116 & 25 & 7\\ 58 & 39 & 51\\ 13 & 37 & 98\end{bmatrix}$	$\begin{bmatrix}128 & 16 & 4\\ 69 & 60 & 19\\ 6 & 52 & 90\end{bmatrix}$
16	0.3333	0.3333	0.3333	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$

Table 27: cont.

17	0.6689	0.6261	0.5968	$\begin{bmatrix}125 & 21 & 2\\ 26 & 42 & 80\\ 4 & 14 & 130\end{bmatrix}$	$\begin{bmatrix}127 & 17 & 4\\ 16 & 36 & 96\\ 13 & 20 & 115\end{bmatrix}$	$\begin{bmatrix}108 & 24 & 16\\ 47 & 39 & 62\\ 10 & 20 & 118\end{bmatrix}$
18	0.4932	0.4257	0.4707	$\begin{bmatrix}81 & 32 & 35\\ 34 & 59 & 55\\ 41 & 28 & 79\end{bmatrix}$	$\begin{bmatrix}74 & 55 & 19\\ 25 & 60 & 63\\ 68 & 25 & 55\end{bmatrix}$	$\begin{bmatrix}80 & 58 & 10\\ 52 & 47 & 49\\ 39 & 27 & 82\end{bmatrix}$
19	0.6171	0.5743	0.4820	$\begin{bmatrix}88 & 21 & 39\\ 12 & 100 & 36\\ 40 & 22 & 86\end{bmatrix}$	$\begin{bmatrix}64 & 40 & 44\\ 24 & 72 & 52\\ 20 & 9 & 119\end{bmatrix}$	$\begin{bmatrix}43 & 50 & 55\\ 21 & 72 & 55\\ 48 & 1 & 99\end{bmatrix}$
20	0.7320	0.7455	0.4797	$\begin{bmatrix}110 & 27 & 11\\ 29 & 93 & 26\\ 0 & 26 & 122\end{bmatrix}$	$\begin{bmatrix}130 & 15 & 3\\ 13 & 92 & 43\\ 1 & 38 & 109\end{bmatrix}$	$\begin{bmatrix}81 & 39 & 28\\ 43 & 56 & 49\\ 13 & 59 & 76\end{bmatrix}$
21	0.6779	0.6216	0.5698	$\begin{bmatrix}131 & 3 & 14\\ 7 & 98 & 43\\ 19 & 57 & 72\end{bmatrix}$	$\begin{bmatrix}122 & 0 & 26\\ 11 & 100 & 37\\ 16 & 78 & 54\end{bmatrix}$	$\begin{bmatrix}97 & 15 & 36\\ 6 & 101 & 41\\ 37 & 56 & 55\end{bmatrix}$
22	0.3333	0.3333	0.3333	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$
23	0.5833	0.5405	0.5698	$\begin{bmatrix}125 & 10 & 13\\ 28 & 112 & 8\\ 41 & 85 & 22\end{bmatrix}$	$\begin{bmatrix}128 & 19 & 1\\ 37 & 102 & 9\\ 20 & 118 & 10\end{bmatrix}$	$\begin{bmatrix}128 & 17 & 3\\ 30 & 111 & 7\\ 26 & 108 & 14\end{bmatrix}$
24	0.3333	0.3333	0.3333	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 148 & 0\\ 0 & 148 & 0\\ 0 & 148 & 0\end{bmatrix}$
26	0.8806	0.8266	0.5405	$\begin{bmatrix}126 & 9 & 13\\ 0 & 143 & 5\\ 5 & 21 & 122\end{bmatrix}$	$\begin{bmatrix}136 & 6 & 6\\ 2 & 119 & 27\\ 16 & 20 & 112\end{bmatrix}$	$\begin{bmatrix}110 & 25 & 13\\ 32 & 62 & 54\\ 30 & 50 & 68\end{bmatrix}$
29	0.7050	0.6824	0.4685	$\begin{bmatrix}135 & 13 & 0\\ 28 & 72 & 48\\ 0 & 42 & 106\end{bmatrix}$	$\begin{bmatrix}135 & 11 & 2\\ 11 & 64 & 73\\ 4 & 40 & 104\end{bmatrix}$	$\begin{bmatrix}118 & 18 & 12\\ 3 & 29 & 116\\ 18 & 69 & 61\end{bmatrix}$
Mean	0.6483	0.6114	0.5054			
Std	0.1428	0.1336	0.1213			

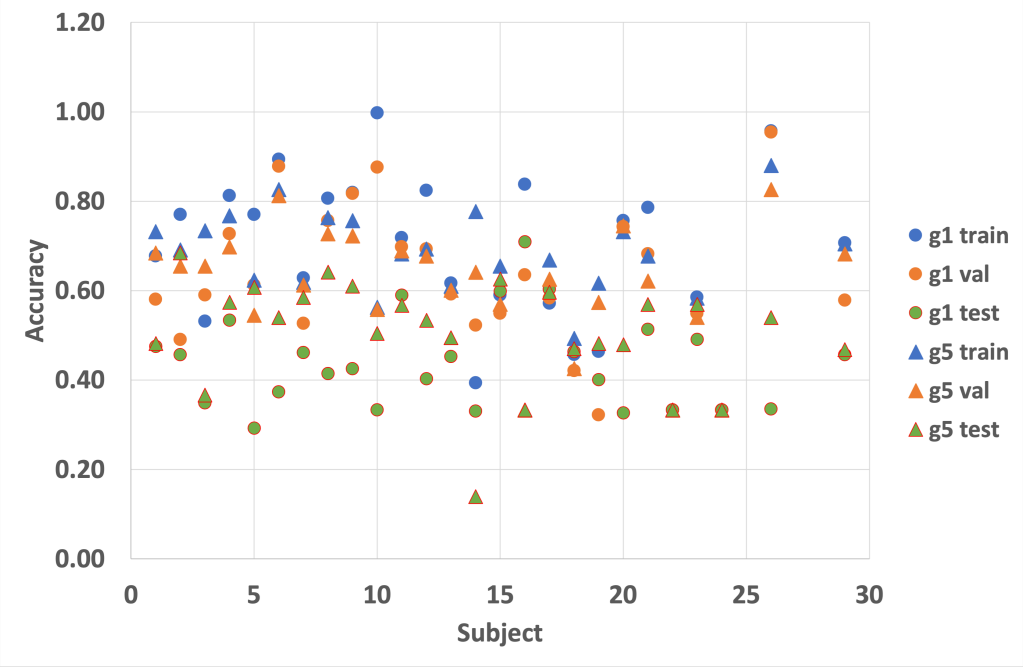


Figure 28: Performance Comparison of COG_BCIg1 and COG_BCIg5 and for each Subject

5.3.3 Significance of Experiment Results

Statistical analyses were performed to confirm the observations from the previous section, utilizing the train, validation, and test accuracy metrics of the selected models for each subject, as detailed in Chapter 4. The analysis considered 26 data points per experiment, reflecting the number of subjects, and the computations were executed using MATLAB.

Table 28: Friedman Test Results

Models	Train Accuracy	Validation Accuracy	Test Accuracy
COG_BCIg1 COG_BCIg2 COG_BCIg3 COG_BCIg4 COG_BCIg5 COG_BCIg6 COG_BCIg7 COG_BCIg8	**7.515e-29**	**8.004e-30**	**0.000639**

The Friedman test results for the experiments COG_BCIg1 - COG_BCIg8 are presented in Table 28. These results reveal statistically significant differences in train, validation, and test accuracies, differing from the findings in Section 4.3.4.2. To further analyze this difference pairwise, Wilcoxon signed-rank tests are conducted, with the results given in Table 29. The comparisons were made between experiments COG_BCIg1, COG_BCIg2, COG_BCIg3, and COG_BCIg4 experiments and their respective counterparts where optimal transport domain adaptation was applied, namely COG_BCIg5, COG_BCIg6, COG_BCIg7, and COG_BCIg8.

The Wilcoxon test results indicate statistically significant differences in test accuracies for all pairs except COG_BCIg3 - COG_BCIg7. As shown in Table 25 and Figure 28, experiments employing optimal transport domain adaptation achieved statistically significantly higher test accuracies, demonstrating improved generalization capability.

For train and validation accuracies, no significant differences were observed for the COG_BCI g1 - COG_BCIg5 and COG_BCIg3 - COG_BCIg7 pairs. However, significant changes were detected for the COG_BCIg2 - COG_BCIg6 and COG_BCIg4 - COG_BCIg8 pairs. This behavior aligns with expectations, as the training and validation sets for COG_BCIg2 and COG_BCIg4 were generated by shuffling two sessions, thereby minimizing variability between these sets and resulting in already well-aligned train and validation data.

5.4 Discussion

The incorporation of domain adaptation using optimal transport in this chapter led to improvements in the model's ability to generalize across sessions. By aligning the feature distributions between sessions, OT mitigates the mismatch between the training and validation or test data

Table 29: Wilcoxon Test Results for Optimal Transport Experiments

Models	Train Accuracy	Validation Accuracy	Test Accuracy
COG_BCIg1 COG_BCIg5	0.2246	0.9658	**0.0047**
COG_BCIg2 COG_BCIg6	**0**	**0**	**0.0086**
COG_BCIg3 COG_BCIg7	0.8911	0.3010	0.0727
COG_BCIg4 COG_BCIg8	**0**	**0**	**0.0278**

distributions. This adjustment addresses the challenges associated with cross-session variability, where the model is trained on data from one session and tested on data from another, often leading to a drop in performance due to the session-specific differences in the data. This adaptation allowed the model to perform better on previously unseen test data by reducing the cross-session variability that had hindered performance in earlier experiments.

The improvement in generalization is evident in the experimental results. Specifically, the average test accuracy improved from 49% to 56% for COG_BCI8, where the test session was Session 2, and data from Sessions 1 and 3 were shuffled and split to create the training and validation sets. Additionally, in experiments where Sessions 1, 2, and 3 were used for training, validation, and testing respectively (COG_BCIg5), the mean test accuracy increased from 44% to 51%.

The SOM plots provide further insight into how the optimal transport (OT) process affects feature distributions across sessions. These visualizations show how OT aligns the feature distributions, helping to improve generalization by reducing the differences between sessions. However, when label information is not used in the OT process, the shifts in feature distributions become more noticeable, revealing that some session-specific characteristics persist despite the adaptation. While OT reduced cross-session variability for some subjects, for others, it increased variability, and residual shifts remain. This suggests that OT can help adapt the model to session-specific data, but it does not fully eliminate cross-session differences. These results highlight that the effectiveness of OT may vary by subject, and further improvements are needed to address these remaining shifts, possibly by incorporating label information or more advanced adaptation techniques.

Moreover, the subject-specific analysis using SOM plots emphasizes the importance of detailed data quality checks prior to model training. Given the individualistic nature of EEG signals, the variability across subjects can be substantial, and without proper data preprocessing and exploration, the model's performance can be significantly compromised. The use of SOM visualizations early in the model development process allows for a more informed selection of features and better identification of potential issues in the data. This approach underscores the value of robust dataset design and comprehensive exploratory data analysis

in improving model performance and ensuring that the model is well-calibrated for real-world applications.

The MATB task was selected for this study because it more accurately reflects real-world scenarios compared to the n-back task, which is often used in controlled experimental settings. The MATB task is designed to simulate dynamic cognitive workload by incorporating both visual and auditory stimuli that vary with increasing task difficulty. This variability in stimuli alters the information processing demands and mental workload of participants, making it a more realistic approximation of real-world conditions. While automated preprocessing techniques aim to remove common artifacts such as eye and muscle movements, some residual artifacts may remain in the data. In such cases, the model may not exclusively measure mental workload, but the MATB task still offers a reasonable compromise between the controlled conditions of lab-based tasks and the unpredictability of real-world environments. This makes the MATB task a valuable tool for studying EEG-based mental workload classification.

Despite the improvements observed with the incorporation of domain adaptation, the models' test performance, although enhanced, still shows room for further enhancement. The results highlight the persistent challenge of addressing cross-session variability in EEG-based classification tasks. EEG data is complex as it is influenced by numerous factors such as electrode placement, individual differences in brain activity, and session-specific dynamics. While domain adaptation offers a promising solution for mitigating some of these challenges, it is clear that further advancements are necessary to achieve more robust performance in real-world applications.

Future research should prioritize strategies such as transfer learning, further domain adaptation techniques, and the integration of session-invariant features into the modeling pipeline. By developing approaches that can better accommodate session-specific variations, models could become more reliable and applicable to a wider range of real-world EEG applications. The findings from this chapter lay a solid foundation for exploring these adaptive strategies, which could lead to improved generalizability and more reliable performance in real-world EEG applications.

CHAPTER 6

CONCLUSION AND DISCUSSION

This thesis has explored the challenges and opportunities in cross-session EEG-based mental workload classification, with a focus on reproducibility, generalizability, and the potential of Graph Neural Networks (GNNs) and domain adaptation techniques. The findings across chapters highlight insights into the complexity of EEG data and the advancements possible through innovative modeling and adaptation strategies. Central to this work are the development of guidelines aimed at improving reproducibility for real-life applications and ensuring model robustness across different experimental sessions.

6.1 Discussion

This thesis emphasizes the broader implications of reproducibility and responsible AI practices in EEG-based machine learning research. By generating guidelines for reproducibility and generalizability, and systematically assessing mental workload classification literature and presenting findings according to these guidelines, this work provides a structured framework for advancing EEG-based machine learning studies. Transparent reporting of experimental setups, data preprocessing and model development together with resource utilization is essential for ensuring the credibility and sustainability of EEG-based machine learning research.

A major theme throughout this work is the inherent variability in EEG data, particularly across sessions. SOM visualizations consistently revealed substantial shifts in feature distributions between sessions, underscoring the difficulty of achieving cross-session generalization. This variability is a fundamental challenge for EEG-based machine learning systems, as models are often exposed to test distributions that deviate significantly from the training data. This issue is further compounded by session-specific factors, such as electrode placement or environmental noise, which are difficult to control in real-world applications.

The experiments demonstrated that GNNs, while effective at learning structured relationships within EEG data, are highly sensitive to session-specific variability. When the training, validation, and test sets were drawn from entirely different sessions, the model struggled to achieve high accuracies across all phases. This setup exposed the model to distinct session-specific distributions at each stage, increasing the difficulty of learning consistent patterns that generalized effectively. The significant variability between sessions likely introduced noise and confounding factors that the model could not overcome, resulting in lower performance

metrics. While accuracy provides an overall indication of performance, metrics like the confusion matrix offer deeper insights into the model's classification tendencies, such as biases toward specific classes or misclassification patterns. Including multiple evaluation metrics ensures a more thorough assessment of the model's strengths and weaknesses. Experiments further revealed that the models trained and validated on a mix of data from two sessions achieved high performance metrics for both training and validation, suggesting that the models effectively learned patterns within the combined data. However, test performance dropped significantly when evaluated on an unseen session, revealing that the models struggled to generalize beyond the training and validation distributions. This highlights the limitations of relying solely on training and validation metrics as indicators of model robustness, as they may not adequately reflect the challenges posed by session-specific differences. The use of SOM visualizations as a diagnostic tool proved valuable in identifying these inconsistencies, offering insights into the underlying data characteristics and guiding more informed model development strategies. These findings underscore the critical role of session-specific variability in influencing model performance and highlight the need for strategies that can better address these inter-session discrepancies to achieve robust and reliable EEG-based classification. Mental workload classification using GNNs remains underexplored, with only a few studies ([229, 230, 231]) attempting to apply these models to such tasks. However, none of these studies have tackled the challenge of cross-session modeling, which is critical for real-world applications. This study's focus on session-specific variability and cross-session generalization using GNNs represents a pioneering effort in the field.

To address these challenges, domain adaptation using optimal transport (OT) was introduced in this study. The application of OT successfully mitigated some of the cross-session variability by aligning feature distributions across sessions, leading to notable improvements in test accuracy. However, the adaptation was not uniformly effective across all subjects, with residual variability persisting in certain cases. This underscores the need for more advanced strategies, such as integrating label information into the adaptation process or developing session-invariant features. While incorporating label information from the test session during adaptation has shown potential to improve generalization, it is not always feasible in practical scenarios where test data labels are unavailable or where retraining the model for each new session introduces significant computational overhead. Therefore, approaches that rely on session-invariant features or unsupervised adaptation are particularly valuable for ensuring model applicability in real-world settings. To date, OT has not been utilized for mental workload classification using EEG data, making this study the first to explore its potential in addressing cross-session variability. This innovative application demonstrates the adaptability of OT for real-world EEG-based systems and sets the stage for further research into unsupervised domain adaptation techniques.

Cross-session variability can be assessed prior to model development to ensure data quality or to exclude highly variant signals. Using tools such as SOM visualizations, sessions with substantial deviations in feature distributions can be identified. Such analyses can guide decisions to exclude problematic data, apply individualized preprocessing pipelines, or design models that account for these differences, thereby improving overall model robustness. Such measures not only enhance performance but also reduce the likelihood of deploying models trained on data of inconsistent quality, which is crucial for building reliable and reproducible EEG-based classification systems.

Additionally, the results emphasize the importance of subject-specific analyses, as individual differences in EEG signals significantly impact model performance. EEG signals exhibit variability across sessions for the same subject due to factors such as changes in electrode placement, skin impedance, cognitive state, and environmental conditions during data acquisition. These session-specific differences can lead to shifts in feature distributions that challenge the model's ability to generalize effectively. Subject-specific analyses allow for a deeper understanding of these variations, enabling the identification of session-specific patterns or anomalies that may hinder performance. For example, SOM visualizations can highlight sessions with distinct feature distributions. This approach helps in improving data quality and aligning feature distributions, thereby enhancing model performance for individual subjects. Such analyses are critical for developing robust and reliable EEG-based systems that accommodate the inherent variability in session-specific data.

The MATB task further demonstrated the challenges of real-world EEG classification. While it provides a more realistic approximation of dynamic cognitive workload, the inherent variability in the task conditions adds complexity to the data. This highlights the trade-off between real-world applicability and controlled experimental settings in EEG-based research.

In summary, this thesis demonstrates that while advancements such as GNNs and OT offer promising avenues for improving EEG-based mental workload classification, significant challenges remain. Cross-session variability, subject-specific differences, and real-world complexities must be addressed through more sophisticated modeling techniques and preprocessing strategies. The findings provide a strong foundation for future research, emphasizing the need for innovation, interdisciplinary collaboration, and a commitment to reproducibility in advancing the field of brain-computer interface applications.

6.2 Contributions and Impact of the Thesis

The work presented in this thesis has contributed to the advancement of EEG-based mental workload estimation, with a focus on reproducibility and cross-session generalizability in brain-computer interface (BCI) applications, in several key areas:

- Reproducibility and Generalizability: Guidelines for achieving reproducibility and generalizability in EEG machine learning models were developed and assessed in Chapter 3. These guidelines emphasize the importance of transparent methodologies and comprehensive reporting, enabling greater collaboration within the scientific community. The systematic literature review conducted as part of this study identified gaps in reproducibility, highlighting challenges faced by researchers and providing a roadmap for improving the robustness and reliability of EEG-based studies.

- Graph Neural Networks for MWL Estimation using EEG data: The adoption of GNNs for modeling EEG signals was explored, demonstrating their effectiveness in representing the spatial relationships between brain regions, as detailed in Chapter 4. By leveraging the inherent structure of EEG data, GNNs provide a more sophisticated approach to modeling mental workload. The proposed framework incorporates spectral

features from multiple frequency bands and constructs graph representations of EEG signals, capturing spatial and spectral relationships between electrodes.

A comprehensive analysis was conducted to evaluate the impact of variations in model parameters, data segmentation, graph representations, data splits, and node selection on model performance. While certain configurations were found to enhance performance, the findings highlight that the primary challenge lies in ensuring generalization over time, particularly to different sessions.

- Mitigating Cross-Session Variability: Cross-session variability is a significant barrier to the deployment of EEG-based models in real-world applications. The variability arises from changes in electrode placement, mental state fluctuations, and environmental factors, among other sources. This thesis employed Optimal Transport (OT) as a domain adaptation technique to address these challenges, as detailed in Chapter 5. By aligning EEG data distributions across sessions, OT facilitates the transfer of features from training to validation and testing phases without the need for re-training, thus improving model generalizability. The effectiveness of OT was demonstrated in the context of mental workload classification, with improvements in cross-session model performance.

- Real-World Applicability: The selected COG-BCI dataset and MATB task closely mimic real-world settings, in particular the high-stakes application of air traffic control, where session-to-session variability is more pronounced. This enhances the practical relevance of the work, demonstrating its potential for real-life applicability in contexts where maintaining consistent performance across sessions is critical.

  An essential factor influencing real-world applicability is the accurate labeling of mental workload (MWL). In long multitasking experiments, factors such as task fatigue or subject disengagement can lead to misleading MWL labels. The MATB task, introduces both visual and auditory stimuli that vary with increasing task difficulty, thereby altering the participants' information processing demands in addition to mental workload. Although automated preprocessing is employed to remove eye and muscle artifacts, some residual artifacts may remain in the data. Consequently, the model may not exclusively measure mental workload. Nevertheless, the MATB task strikes a reasonable balance between the controlled conditions of laboratory tasks and the complexity of real-world scenarios. Accurate labeling necessitates thoughtful strategies to ensure that the data reliably reflects the cognitive load experienced by the subjects, as mislabeling can impact model performance. Beyond task difficulty, labeling may rely on self-report or task performance, depending on the context. However, behavioral data can be misleading; for instance, low performance may result from insufficient cognitive resources associated with high MWL or from a lack of motivation and effort indicative of low MWL. While self-reports may offer a better ground truth, they are inherently dependent on the subject's perception and willingness to report their state accurately. To address these complexities, analyzing self-reports, behavioral data, and task sequences is essential to understanding how these factors influence MWL levels. Thus, the collection of additional data on these aspects, such as those provided in the COG-BCI dataset, is strongly recommended.

  Another consideration is the requirements on the number of workload levels to classify in real-life systems. Research has shown that classifying data into more than two

workload levels can significantly reduce performance. This issue highlights the importance of considering practical constraints and ensuring that models designed for real-world applications can efficiently address realistic workload levels while maintaining accuracy. Therefore, while multi-class classification offers greater granularity, it is important to balance this with the need for robust and reliable performance in real-world scenarios.

- Subject-Specific Analysis: This study also explored the variability in EEG signals specific to each subject, as detailed in Chapters 4 and 5, aiming to identify the underlying factors contributing to challenges in generalizing to unseen sessions. Bu evaluating the distributional differences in features across sessions using SOM visualizations, a deeper understanding of the issues posed by session variability is achieved. Relating these insights to model performance metrics, such as accuracy and confusion matrices, provided a comprehensive perspective on the data characteristics.

  These insights are crucial for developing personalized EEG-based BCIs, where subject-specific factors can be leveraged to enhance the model accuracy and applicability. Moreover, the subject-specific analysis highlights the importance of conducting data quality checks prior to model development. This preliminary step can inform model adaptation strategies, ensuring greater robustness and reliability across varying session conditions.

The contributions of this thesis extend beyond the technical advancements in modeling EEG signals. By addressing the critical issue of reproducibility in EEG-based mental workload classification, this work provides a framework that can guide future research in the field. The developed guidelines are expected to have a lasting impact on ensuring the transparency and reproducibility of EEG and machine learning studies, which are essential for building trust and advancing the adoption of these technologies in real-world applications. Moreover, the application of GNNs and OT in this thesis paves the way for more sophisticated and generalizable models, addressing the practical challenges associated with deploying EEG-based systems in dynamic and unpredictable environments. These advancements are expected to have a lasting impact on the development of robust, transparent, and reliable BCI applications.

6.3 Limitations

While this thesis advances cross-session generalizability and reproducibility in EEG-based mental workload estimation, several limitations should be acknowledged as they highlight areas for improvement and future research.

First, while the dataset comprises a balanced sample of 26 subjects, ensuring diversity and reliability, the subject-specific variability remains a significant factor to be addressed. For some subjects, achieving high classification performance proves more challenging due to inherent differences in signal quality or underlying brain activity patterns.

Another limitation arises from the reliance on a specific preprocessing pipeline and the use of Power Spectral Density (PSD) as node features and Euclidean distance as edge features, which may constrain the applicability of the findings to datasets processed differently or employing

alternative feature representations. Similarly, the adjacency matrix construction methods, such as k-nearest neighbors and fixed-edge thresholds, impose assumptions about the relationships between nodes, which may not fully align with the underlying brain connectivity patterns.

Additionally, the study explores a specific range of graph neural network architectures and hyperparameters. While these configurations were optimized using Bayesian search, there may be unexplored model structures or parameter ranges that could yield further improvements. The chosen normalization, pooling, and graph construction methods also introduce potential biases, as they may not suit all subjects or sessions equally well.

The application of explainability techniques was not prioritized in this study due to the specific focus on enhancing cross-session generalizability and classification performance of mental workload using EEG data. Explainability methods, such as saliency maps or feature importance analyses, often introduce additional computational overhead and interpretability challenges when applied to complex graph-based models, such as Graph Neural Networks. Furthermore, the primary goal of the research was to establish and validate a robust pipeline for reproducible brain-computer interface applications, emphasizing classification accuracy and domain adaptation techniques over interpretive insights. Additionally, while the dataset included 26 subjects, which is a sufficient sample size for classification studies, it was not specifically designed to support causal inference or advanced explainability. Causal analyses typically require data collected under diverse scenarios, repeated measures, or controlled interventions to uncover causal relationships. Current approaches to explainability in EEG research often rely on correlation-based statistical analyses, which are insufficient for establishing causation.

Lastly, while significant effort has been devoted to ensuring reproducibility, the generalizability of the results to datasets or tasks beyond the current experimental setup is limited. Broader validation across diverse datasets, task paradigms, and recording protocols is necessary to confirm the applicability of the proposed methods to other contexts. These limitations highlight the need for continued research to refine and extend the methods presented in this thesis.

6.4 Future Work

While the results of this thesis represent a step toward improving cross-session generalizability and reproducibility in EEG-based mental workload estimation, several areas remain open for future exploration:

- Exploration of Other Domain Adaptation Techniques: Although Optimal Transport showed promise in mitigating cross-session variability, other domain adaptation techniques, such as adversarial training, domain-invariant feature learning, and multi-task learning, could be explored. These methods may offer additional insights and improvements in the transferability of models across different sessions and subjects.

- Further Refinement of Features and GNN Architectures: While GNNs were found to yield promising results, the architecture of GNNs can be further refined. Experimenting with different types of graph convolutions and attention mechanisms could provide insights into how the graph structure impact performance. Additionally, exploring the integration of temporal dependencies within GNNs could further improve the representation of dynamic changes in brain activity. Another area for improvement is the exploration of additional features beyond the Power Spectral Density (PSD) values as node features and Euclidean distances as edge features. Non-linear measures, such as connectivity or coherence, may be evaluated to determine whether they provide richer information for classifying MWL.

- Explainability: Future work could explore explainability as a complementary dimension, particularly through the collection of more extensive and diverse datasets and the development of methods tailored to causal exploration in EEG-based systems in contexts where model transparency is critical for deployment in real-world neuroergonomic applications.

- Multimodality: Expanding the feature set beyond EEG to include other physiological signals, such as eye tracking or heart rate variability (HRV), could provide a more comprehensive view of mental workload. Incorporating multimodal data could lead to more reliable systems capable of estimating MWL in complex, real-world environments, where cognitive load may be influenced by a broader range of factors beyond what is captured by EEG alone.

- Temporal Relations: Considering the temporal relations between EEG signals is a promising area for future research. Mental workload is not static and can fluctuate over time as cognitive load varies during task execution. Temporal modeling could enhance the ability of models to track and predict changes in workload more effectively. These models could capture dynamic workload states, allowing for more accurate and context-aware estimations of mental workload over time.

- Personalized Mental Workload Prediction: The identification of subject-specific features for mental workload estimation offers exciting opportunities for personalized BCI applications. Future research could focus on extracting subject-specific states or features to address subject-level variability or developing adaptive models that learn from a subject's individual EEG patterns over time, leading to more accurate mental workload predictions generalizable across subjects.

- Interdisciplinary Applications: The methodologies developed in this thesis have the potential to be applied to a wide range of interdisciplinary fields, including human-computer interaction, cognitive state monitoring, and neuroergonomics. Exploring the use of these models in non-EEG domains (e.g., fNIRS, eye-tracking) could further enhance their versatility and impact.

- Focus on Generalizability: Finally, the generalizability of EEG classifiers remains a critical focus for future work. Achieving cross-session and cross-subject generalization is essential for deploying EEG-based systems in real-world applications. Mental workload classification is inherently challenging due to its multi-faceted nature and the variability between individuals, as well as the non-stationary nature of EEG data.

Workload levels can fluctuate based on numerous factors, such as task difficulty, user experience, emotional state, and environmental conditions. Given this complexity, it is crucial that future work continues to refine methods that ensure model generalizability, allowing for the development of robust and usable passive Brain-Computer Interface (BCI) systems.

In conclusion, this thesis has made contributions in addressing the challenges of cross-session variability and reproducibility in EEG-based mental workload estimation. By leveraging advanced machine learning techniques, such as GNNs and Optimal Transport, this work has advanced the field of EEG-based brain-computer interfaces, with promising applications for both research and real-world deployment. The methodologies developed here provide a foundation for future work aimed at refining these models and making them more adaptable, personalized, and applicable to dynamic environments. Through continued innovation, EEG-based BCIs have the potential to transform human-computer interaction, improving performance in a variety of critical domains where mental workload plays a decisive role.

REFERENCES

[1] R. Roy, M. Hinss, L. Darmet, S. Ladouce, E. Jahanpour, B. Somon, X. Xu, N. Drougard, F. Dehais, and F. Lotte, "Retrospective on the first passive brain-computer interface competition on cross-session workload estimation," *Frontiers in Neuroergonomics*, vol. 3, 2022.

[2] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, 2015.

[3] G. Johannsen, "Workload and workload measurement," in *Mental Workload: Its Theory and Measurement* (N. Moray, ed.), pp. 3–11, Boston, MA: Springer US, 1979.

[4] M. F. Hinss, E. S. Jahanpour, B. Somon, L. Pluchon, F. Dehais, and R. N. Roy, "Open multi-session and multi-task eeg cognitive dataset for passive brain-computer interface applications," *Scientific Data*, vol. 10, p. 85, Feb 2023.

[5] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Society Bulletin of the Technical Committee on Data Engineering*, vol. 40, no. 3, pp. 52–74, 2017.

[6] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[7] M. Zabcikova, Z. Koudelkova, R. Jasek, and J. J. Lorenzo Navarro, "Recent advances and current trends in brain-computer interface research and their applications," *International Journal of Developmental Neuroscience*, vol. 82, no. 2, pp. 107–123, 2022.

[8] T. O. Zander and C. Kothe, "Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general," *Journal of Neural Engineering*, vol. 8, p. 025005, mar 2011.

[9] P. Arico, G. Borghini, G. Di Flumeri, N. Sciaraffa, A. Colosimo, and F. Babiloni, "Passive bci in operational environments: insights, recent advances, and future trends," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1431–1436, 2017.

[10] T. O. Zander, C. Kothe, S. Jatzev, and M. Gaertner, "Enhancing human-computer interaction with input from active and passive brain-computer interfaces," in *Brain-computer interfaces: Applying our minds to human-computer interaction* (D. S. Tan and A. Nijholt, eds.), pp. 181–199, London: Springer London, 2010.

[11] C. D. Wickens, "Multiple resources and mental workload," *Human factors*, vol. 50, no. 3, pp. 449–455, 2008.

[12] E. Galy, M. Cariou, and C. Mélan, "What is the relationship between mental workload factors and cognitive load types?," *International journal of psychophysiology*, vol. 83, no. 3, pp. 269–275, 2012.

[13] W. Klimesch, "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[14] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2021.

[15] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 249–270, 2022.

[16] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 1, pp. 4–19, 2022.

[17] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[18] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425–438, 2010.

[19] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, no. 2, pp. 257–285, 1988.

[20] B. Cain, "A review of the mental workload literature," *Technical report, Defence Research and Development Toronto (Canada)*, 2007.

[21] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human mental workload: A survey and a novel inclusive definition," *Frontiers in psychology*, vol. 13, p. 883321, 2022.

[22] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183, Elsevier, 1988.

[23] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," in *Advances in psychology*, vol. 52, pp. 185–218, Elsevier, 1988.

[24] F. R. H. Zijlstra, *Efficiency in work behaviour: A design approach for modern tools*. PhD thesis, Technische Universiteit Delft (The Netherlands), 1993. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-02-24.

[25] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring mental workload with eeg+ fnirs," *Frontiers in human neuroscience*, vol. 11, p. 359, 2017.

[26] B. Rebsamen, K. Kwok, and T. B. Penney, "Evaluation of cognitive workload from eeg during a mental arithmetic task," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1, pp. 1342–1345, 2011.

[27] B. Batun, "Gender and personality differences in cognitive tasks: a performance and human mental workload study," Master's thesis, Middle East Technical University, 2015.

[28] W. L. Lim, O. Sourina, Y. Liu, and L. Wang, "Eeg-based mental workload recognition related to multitasking," in *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1–4, IEEE, 2015.

[29] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

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

[31] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using eeg spectral power and erps in the n-back task," *Journal of neural engineering*, vol. 9, p. 045008, jul 2012.

[32] M. E. Smith, A. Gevins, H. Brown, A. Karnik, and R. Du, "Monitoring task loading with multivariate eeg measures during complex forms of human-computer interaction," *Human Factors*, vol. 43, no. 3, pp. 366–380, 2001.

[33] J. B. Brookings, G. F. Wilson, and C. R. Swain, "Psychophysiological responses to changes in workload during simulated air traffic control," *Biological psychology*, vol. 42, no. 3, pp. 361–377, 1996.

[34] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with eeg pattern recognition methods," *Human Factors*, vol. 40, no. 1, pp. 79–91, 1998. PMID: 9579105.

[35] A. Holm, K. Lukander, J. Korpela, M. Sallinen, and K. M. Müller, "Estimating brain load from the eeg," *TheScientificWorldJOURNAL*, vol. 9, pp. 639–651, 2009.

[36] W. Lim, O. Sourina, and L. P. Wang, "Stew: simultaneous task eeg workload data set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2106–2114, 2018.

[37] P. Sauseng, W. Klimesch, M. Schabus, and M. Doppelmayr, "Fronto-parietal eeg coherence in theta and upper alpha reflect central executive functions of working memory," *International Journal of Psychophysiology*, vol. 57, no. 2, pp. 97–103, 2005.

[38] M. Plechawska-Wójcik, M. Tokovarov, M. Kaczorowska, and D. Zapała, "A three-class classification of cognitive workload based on eeg spectral data," *Applied Sciences*, vol. 9, no. 24, p. 5340, 2019.

[39] J. Benerradi, H. A. Maior, A. Marinescu, J. Clos, and M. L. Wilson, "Exploring machine learning approaches for classifying mental workload using fnirs data from hci tasks," in *Proceedings of the Halfway to the Future Symposium 2019*, pp. 1–11, 2019.

[40] C. L. Baldwin and B. Penaranda, "Adaptive training using an artificial neural network and eeg metrics for within-and cross-task workload classification," *NeuroImage*, vol. 59, no. 1, pp. 48–56, 2012.

[41] S. Chandra, G. Sharma, K. Verma, A. Mittal, and D. Jha, "Eeg based cognitive workload classification during nasa matb-ii multitasking," *International Journal of Cognitive Research in Science, Engineering and Education*, vol. 3, no. 1, 2015.

[42] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using eeg and an adaptive deep learning model," *Biomedical Signal Processing and Control*, vol. 33, pp. 30–47, 2017.

[43] M. Saadati, J. Nelson, and H. Ayaz, "Convolutional neural network for hybrid fnirs-eeg mental workload classification," in *International Conference on Applied Human Factors and Ergonomics*, pp. 221–232, Springer, 2019.

[44] D. D. Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm," *Biomedical Signal Processing and Control*, vol. 60, p. 101989, 2020.

[45] U. Asgher, K. Khalil, M. J. Khan, R. Ahmad, S. I. Butt, Y. Ayaz, N. Naseer, and S. Nazir, "Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface," *Frontiers in neuroscience*, vol. 14, p. 584, 2020.

[46] D. Klepl, M. Wu, and F. He, "Graph neural network-based eeg classification: A survey," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 493–503, 2024.

[47] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[48] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, 2020.

[49] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[50] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[51] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[52] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *International Conference on Learning Representations*, 2019.

[53] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[54] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[55] F. Lotte, "A tutorial on eeg signal-processing techniques for mental-state recognition in brain–computer interfaces," in *Guide to Brain-Computer Music Interfacing* (E. R. Miranda and J. Castet, eds.), pp. 133–161, London: Springer London, 2014.

[56] S. Jang, S.-E. Moon, and J.-S. Lee, "Eeg-based video identification using graph signal modeling and graph convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3066–3070, IEEE, 2018.

[57] X. Li, B. Qian, J. Wei, A. Li, X. Liu, and Q. Zheng, "Classify eeg and reveal latent graph structure with spatio-temporal graph convolutional neural network," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 389–398, IEEE, 2019.

[58] N. Wagh and Y. Varatharajah, "Eeg-gcnn: Augmenting electroencephalogram-based neurological disease diagnosis using a domain-guided graph convolutional neural network," in *Machine Learning for Health*, pp. 367–378, PMLR, 2020.

[59] Y. Kwak, W.-J. Song, and S.-E. Kim, "Graph neural network with multilevel feature fusion for eeg based brain-computer interface," in *2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1–3, IEEE, 2020.

[60] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "Eeg emotion recognition using fusion model of graph convolutional neural networks and lstm," *Applied Soft Computing*, vol. 100, p. 106954, 2021.

[61] E. F. Montesuma, F. M. N. Mboula, and A. Souloumiac, "Recent advances in optimal transport for machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024.

[62] Q. Chen, Y. Liu, Z. Wang, I. Wassell, and K. Chetty, "Re-weighted adversarial adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7976–7985, June 2018.

[63] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, Apr. 2018.

[64] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, "Overrelaxed sinkhorn–knopp algorithm for regularized optimal transport," *Algorithms*, vol. 14, no. 5, p. 143, 2021.

[65] M. Bonafini and B. Schmitzer, "Domain decomposition for entropy regularized optimal transport," *Numerische Mathematik*, vol. 149, no. 4, pp. 819–870, 2021.

[66] P. Chen, H. Wang, X. Sun, H. Li, C. Grebogi, and Z. Gao, "Transfer learning with optimal transportation and frequency mixup for eeg-based motor imagery recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2866–2875, 2022.

[67] A. Ackaouy, N. Courty, E. Vallée, O. Commowick, C. Barillot, and F. Galassi, "Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data," *Frontiers in computational neuroscience*, vol. 14, p. 19, 2020.

[68] O. Gilo, J. Mathew, S. Mondal, and R. K. Sanodiya, "Rdaot: Robust unsupervised deep sub-domain adaptation through optimal transport for image classification," *IEEE Access*, vol. 11, pp. 102243–102260, 2023.

[69] A. Rakotomamonjy, R. Flamary, G. Gasso, M. E. Alaya, M. Berar, and N. Courty, "Optimal transport for conditional domain matching and label shift," *Machine Learning*, vol. 111, pp. 1651–1670, May 2022.

[70] Q. H. Tran, H. Janati, N. Courty, R. Flamary, I. Redko, P. Demetci, and R. Singh, "Unbalanced co-optimal transport," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 10006–10016, 2023.

[71] L. Gautheron, I. Redko, and C. Lartizien, "Feature selection for unsupervised domain adaptation using optimal transport," in *Machine Learning and Knowledge Discovery in Databases*, pp. 759–776, Springer International Publishing, 2019.

[72] C. Chu, L. Zhu, A. Huang, P. Xu, N. Ying, and J. Zhang, "Transfer learning with data alignment and optimal transport for eeg based motor imagery classification," *Journal of Neural Engineering*, vol. 21, no. 1, p. 016015, 2024.

[73] V. Peterson, N. Nieto, D. Wyser, O. Lambercy, R. Gassert, D. H. Milone, and R. D. Spies, "Transfer learning based on optimal transport for motor imagery brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 807–817, 2022.

[74] F. Putze, S. Putze, M. Sagehorn, C. Micek, and E. T. Solovey, "Understanding hci practices and challenges of experiment reporting with brain signals: Towards reproducibility and reuse," *ACM Trans. Comput.-Hum. Interact.*, vol. 29, mar 2022.

[75] G. Demirezen, T. Taşkaya Temizel, and A.-M. Brouwer, "Reproducible machine learning research in mental workload classification using eeg," *Frontiers in Neuroergonomics*, vol. 5, 2024.

[76] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, pp. 452–454, May 2016.

[77] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018.

[78] M. Saeidi, W. Karwowski, F. V. Farahani, K. Fiok, R. Taiar, P. A. Hancock, and A. Al-Juaid, "Neural decoding of eeg signals with machine learning: A systematic review," *Brain Sciences*, vol. 11, no. 11, 2021.

[79] R. N. Roy, S. Bonnet, S. Charbonnier, and A. Campagne, "Mental fatigue and working memory load estimation: Interaction and implications for eeg-based passive bci," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6607–6610, 2013.

[80] J. Millan, "On the need for on-line learning in brain-computer interfaces," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, vol. 4, pp. 2877–2882 vol.4, 2004.

[81] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, p. 031005, apr 2018.

[82] C. Mühl, C. Jeunet, and F. Lotte, "Eeg-based workload estimation across affective contexts," *Frontiers in Neuroscience*, vol. 8, 2014.

[83] National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press, 2019.

[84] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[85] M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: Still a ways to go," *Science Translational Medicine*, vol. 13, no. 586, p. eabb1655, 2021.

[86] C. Pernet, M. I. Garrido, A. Gramfort, N. Maurits, C. M. Michel, E. Pang, R. Salmelin, J. M. Schoffelen, P. A. Valdes-Sosa, and A. Puce, "Issues and recommendations from the ohbm cobidas meeg committee for reproducible eeg and meg research," *Nature Neuroscience*, vol. 23, pp. 1473–1483, Dec 2020.

[87] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research," *Commun. ACM*, vol. 59, p. 62–69, feb 2016.

[88] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *International Journal of Surgery*, vol. 8, no. 5, pp. 336–341, 2010.

[89] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, 2018.

[90] A. Kenall, S. Edmunds, L. Goodman, L. Bal, L. Flintoft, D. R. Shanahan, and T. Shipley, "Better reporting for better research: a checklist for reproducibility," *GigaScience*, vol. 4, pp. s13742–015–0071–8, 07 2015.

[91] M. McNutt, "Journals unite for reproducibility," *Science*, vol. 346, no. 6210, pp. 679–679, 2014.

[92] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)," *J. Mach. Learn. Res.*, vol. 22, jan 2021.

[93] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, *et al.*, "Crisp-dm 1.0: Step-by-step data mining guide," *SPSS inc*, vol. 9, no. 13, pp. 1–73, 2000.

[94] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37–54, Mar. 1996.

[95] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[96] K. A. Robbins, J. Touryan, T. Mullen, C. Kothe, and N. Bigdely-Shamlo, "How sensitive are eeg results to preprocessing methods: A benchmarking study," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 5, pp. 1081–1090, 2020.

[97] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, and R. A. Poldrack, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Scientific Data*, vol. 3, p. 160044, Jun 2016.

[98] C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld, "Eeg-bids, an extension to the brain imaging data structure for electroencephalography," *Scientific Data*, vol. 6, p. 103, Jun 2019.

[99] A. Keil, S. Debener, G. Gratton, M. Junghöfer, E. S. Kappenman, S. J. Luck, P. Luu, G. A. Miller, and C. M. Yee, "Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography," *Psychophysiology*, vol. 51, no. 1, pp. 1–21, 2014.

[100] N. Kane, J. Acharya, S. Beniczky, L. Caboclo, S. Finnigan, P. W. Kaplan, H. Shibasaki, R. Pressler, and M. J. van Putten, "A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the eeg findings. revision 2017," *Clinical Neurophysiology Practice*, vol. 2, pp. 170–185, 2017.

[101] A. Mastropietro, I. Pirovano, A. Marciano, S. Porcelli, and G. Rizzo, "Reliability of mental workload index assessed by eeg with different electrode configurations and signal pre-processing pipelines," *Sensors*, vol. 23, no. 3, 2023.

[102] I. Kakkos, G. N. Dimitrakopoulos, L. Gao, Y. Zhang, P. Qi, G. K. Matsopoulos, N. Thakor, A. Bezerianos, and Y. Sun, "Mental workload drives different reorganizations of functional cortical connectivity between 2d and 3d simulated flight experiments," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 9, pp. 1704–1713, 2019.

[103] T. Radüntz, N. Fürstenau, T. Mühlhausen, and B. Meffert, "Indexing mental workload during simulated air traffic control tasks by means of dual frequency head maps," *Frontiers in Physiology*, vol. 11, 2020.

[104] N. Sciaraffa, G. Di Flumeri, D. Germano, A. Giorgi, A. Di Florio, G. Borghini, A. Vozzi, V. Ronca, F. Babiloni, and P. Aricò, "Evaluation of a new lightweight eeg technology for translational applications of passive brain-computer interfaces," *Frontiers in Human Neuroscience*, vol. 16, 2022.

[105] M. J. Boring, K. Ridgeway, M. Shvartsman, and T. R. Jonker, "Continuous decoding of cognitive load from electroencephalography reveals task-general and task-specific correlates," *Journal of Neural Engineering*, vol. 17, p. 056016, oct 2020.

[106] V. Parekh, M. Bilalpur, C. V. Jawahar, S. Kumar, S. Winkler, and R. Subramanian, "Investigating the generalizability of eeg-based cognitive load estimation across visualizations," in *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*, ICMI '18, (New York, NY, USA), pp. 1–5, Association for Computing Machinery, 2018.

[107] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Pozzi, and F. Babiloni, "Chapter 10 - a passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks," in *Brain-Computer Interfaces: Lab Experiments to Real-World Applications* (D. Coyle, ed.), vol. 228, pp. 295–328, Elsevier, 2016.

[108] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, I. Graziani, J.-P. Imbert, G. Granger, R. Benhacene, M. Terenzi, S. Pozzi, and F. Babiloni, "Reliability over time of eeg-based mental workload evaluation during air traffic management (atm) tasks," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7242–7245, 2015.

[109] O. Ortiz, D. Blustein, and U. Kuruganti, "Test-retest reliability of time-domain eeg features to assess cognitive load using a wireless dry-electrode system," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2885–2888, 2020.

[110] F. Putze, M. Mülller, D. Heger, and T. Schultz, "Session-independent eeg-based workload recognition," in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing - Volume 1: BIOSIGNALS, (BIOSTEC 2013)*, pp. 360–363, INSTICC, SciTePress, 2013.

[111] E. L. Fox, M. Ugolini, and J. W. Houpt, "Predictions of task using neural modeling," *Frontiers in Neuroergonomics*, vol. 3, 2022.

[112] S. J. Eglen, B. Marwick, Y. O. Halchenko, M. Hanke, S. Sufi, P. Gleeson, R. A. Silver, A. P. Davison, L. Lanyon, M. Abrams, T. Wachtler, D. J. Willshaw, C. Pouzat, and J.-B. Poline, "Toward standard practices for sharing computer code and programs in neuroscience," *Nature Neuroscience*, vol. 20, pp. 770–773, Jun 2017.

[113] G. Van Rossum, B. Warsaw, and N. Coghlan, "Pep 8–style guide for python code," *Python. org*, vol. 1565, p. 28, 2001.

[114] S. Hendriks, C. Grady, K. M. Ramos, W. Chiong, J. J. Fins, P. Ford, S. Goering, H. T. Greely, K. Hutchison, M. L. Kelly, S. Y. H. Kim, E. Klein, S. H. Lisanby, H. Mayberg, H. Maslen, F. G. Miller, K. Rommelfanger, S. A. Sheth, and A. Wexler, "Ethical challenges of risk, informed consent, and posttrial responsibilities in human research with neural devices: A review," *JAMA Neurol*, oct 2019.

[115] T. D. Azad, J. Ehresman, A. K. Ahmed, V. E. Staartjes, D. Lubelski, M. N. Stienen, A. Veeravagu, and J. K. Ratliff, "Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery," *The Spine Journal*, vol. 21, no. 10, pp. 1610–1616, 2021.

[116] C. R. Pernet, R. Martinez-Cancino, D. Truong, S. Makeig, and A. Delorme, "From bids-formatted eeg data to sensor-space group results: A fully reproducible workflow with eeglab and limo eeg," *Frontiers in Neuroscience*, vol. 14, 2021.

[117] A. Delorme, T. Mullen, C. Kothe, Z. Akalin Acar, N. Bigdely-Shamlo, A. Vankov, and S. Makeig, "Eeglab, sift, nft, bcilab, and erica: New tools for advanced eeg processing," *Computational Intelligence and Neuroscience*, vol. 2011, p. 130714, May 2011.

[118] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The prep pipeline: standardized preprocessing for large-scale eeg analysis," *Frontiers in Neuroinformatics*, vol. 9, 2015.

[119] A. Delorme, "Eeg is better left alone," *Scientific Reports*, vol. 13, p. 2372, Feb 2023.

[120] A. de Cheveigné and I. Nelken, "Filters: When, why, and how (not) to use them," *Neuron*, vol. 102, no. 2, pp. 280–293, 2019.

[121] A.-M. Brouwer, T. O. Zander, J. B. F. van Erp, J. E. Korteling, and A. W. Bronkhorst, "Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls," *Frontiers in Neuroscience*, vol. 9, 2015.

[122] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade: Second Edition* (G. Montavon, G. B. Orr, and K.-R. Müller, eds.), pp. 437–478, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[123] D. Sculley, J. Snoek, A. B. Wiltschko, and A. Rahimi, "Winner's curse? on pace, progress, and empirical rigor," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, pp. 1–4, OpenReview.net, 2018.

[124] G. Canbek, T. Taskaya Temizel, and S. Sagiroglu, "Benchmetrics: a systematic benchmarking method for binary classification performance metrics," *Neural Computing and Applications*, vol. 33, pp. 14623–14650, Nov 2021.

[125] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random: a closer look on bci results," *International journal of bioelectromagnetism*, vol. 10, no. ARTICLE, pp. 52–55, 2008.

[126] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[127] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13693–13696, Apr. 2020.

[128] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Commun. ACM*, vol. 63, p. 54–63, nov 2020.

[129] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *J. Mach. Learn. Res.*, vol. 21, jan 2020.

[130] C. Huyen, *Designing machine learning systems*. " O'Reilly Media, Inc.", 2022.

[131] H. Taheri Gorji, N. Wilson, J. VanBree, B. Hoffmann, T. Petros, and K. Tavakolian, "Using machine learning methods and eeg to discriminate aircraft pilot cognitive workload during flight," *Scientific Reports*, vol. 13, p. 2507, Feb 2023.

[132] K.-J. Chiang, S. Dong, C.-K. Cheng, and T.-P. Jung, "Using eeg signals to assess workload during memory retrieval in a real-world scenario," *Journal of Neural Engineering*, vol. 20, p. 036010, may 2023.

[133] W. Zhu, C. Zhang, C. Liu, J. Yuan, X. Li, Y. Wang, and C. Jiang, "Recognition of pilot mental workload in the simulation operation of carrier-based aircraft using the portable eeg," in *Proceedings of the 2023 3rd International Conference on Human Machine Interaction*, ICHMI '23, (New York, NY, USA), p. 43–49, Association for Computing Machinery, 2023.

[134] K. Zemla, G. M. Wojcik, F. Postepski, K. Wróbel, A. Kawiak, and G. Sedek, "Modeling of brain cortical activity during relaxation and mental workload tasks based on eeg signal collection," *Applied Sciences*, vol. 13, no. 7, 2023.

[135] Z. Zhang, Z. Zhao, H. Qu, C. Liu, and L. Pang, "A mental workload classification method based on gcn modified by squeeze-and-excitation residual," *Mathematics*, vol. 11, no. 5, 2023.

[136] K. Guan, Z. Zhang, T. Liu, and H. Niu, "Cross-task mental workload recognition based on eeg tensor representation and transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2632–2639, 2023.

[137] A. Teymourlouei, R. J. Gentili, and J. Reggia, "Decoding eeg signals with visibility graphs to predict varying levels of mental workload," in *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2023.

[138] K. Kingphai and Y. Moshfeghi, "On time series cross-validation for deep learning classification model of mental workload levels based on eeg signals," in *Machine Learning, Optimization, and Data Science* (G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, G. Di Fatta, G. Giuffrida, and R. Umeton, eds.), (Cham), pp. 402–416, Springer Nature Switzerland, 2023.

[139] Z. Zheng, Z. Yin, Y. Wang, and J. Zhang, "Inter-subject cognitive workload estimation based on a cascade ensemble of multilayer autoencoders," *Expert Systems with Applications*, vol. 211, p. 118694, 2023.

[140] J. Yedukondalu and L. D. Sharma, "Cognitive load detection using circulant singular spectrum analysis and binary harris hawks optimization based feature selection," *Biomedical Signal Processing and Control*, vol. 79, p. 104006, 2023.

[141] R. Patel, K. Gireesan, R. Baskaran, and N. V. C. Shekar, "Optimal classification of n-back task eeg data by performing effective feature reduction," *Sādhanā*, vol. 47, p. 281, Dec 2022.

[142] I. Albuquerque, J. Monteiro, O. Rosanne, and T. H. Falk, "Estimating distribution shifts for predicting cross-subject generalization in electroencephalography-based mental workload assessment," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[143] N. Sciaraffa, G. Di Flumeri, D. Germano, A. Giorgi, A. Di Florio, G. Borghini, A. Vozzi, V. Ronca, F. Babiloni, and P. Aricò, "Evaluation of a new lightweight eeg technology for translational applications of passive brain-computer interfaces," *Frontiers in Human Neuroscience*, vol. 16, 2022.

[144] E. Q. Wu, M. Zhou, D. Hu, L. Zhu, Z. Tang, X.-Y. Qiu, P.-Y. Deng, L.-M. Zhu, and H. Ren, "Self-paced dynamic infinite mixture model for fatigue evaluation of pilots' brains," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5623–5638, 2022.

[145] B. Raufi and L. Longo, "An evaluation of the eeg alpha-to-theta and theta-to-alpha band ratios as indexes of mental workload," *Frontiers in Neuroinformatics*, vol. 16, 2022.

[146] M. Zhao, Z. Ji, J. Zhang, Y. Zhu, C. Ye, G. Wang, and Z. Yin, "Assessing distinct cognitive workload levels associated with unambiguous and ambiguous pronoun resolutions in human-machine interactions," *Brain Sciences*, vol. 12, no. 3, 2022.

[147] J. Yedukondalu and L. D. Sharma, "Cognitive load detection using binary salp swarm algorithm for feature selection," in *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, pp. 1–5, 2022.

[148] Y. Liu, S. Shi, Y. Song, Q. Gao, Z. Li, H. Song, S. Pang, and D. Li, "Eeg based mental workload assessment by power spectral density feature," in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 450–454, 2022.

[149] T. A. Babu, S. Gadde, S. Ravi, K. V. V. Rao, Y. Mamillu, and D. Krishna, "Analysis of mental task ability in students based on electroencephalography signals," in *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, vol. 1, pp. 274–278, 2022.

[150] R. Zanetti, A. Arza, A. Aminifar, and D. Atienza, "Real-time eeg-based cognitive workload monitoring on wearable devices," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 1, pp. 265–277, 2022.

[151] I. Hussain, S. Young, and S.-J. Park, "Driving-induced neurological biomarkers in an advanced driver-assistance system," *Sensors*, vol. 21, no. 21, 2021.

[152] L. D. Sharma, R. K. Saraswat, and R. K. Sunkaria, "Cognitive performance detection using entropy-based features and lead-specific approach," *Signal, Image and Video Processing*, vol. 15, pp. 1821–1828, Nov 2021.

[153] I. Kakkos, G. N. Dimitrakopoulos, Y. Sun, J. Yuan, G. K. Matsopoulos, A. Bezerianos, and Y. Sun, "Eeg fingerprints of task-independent mental workload discrimination," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3824–3833, 2021.

[154] F. Rahman, M. I. Ahmed, S. S. Saad, M. Ashrafuzzaman, S. S. Mogno, R. Rahman, and M. Z. Parvez, "Prediction and detection in change of cognitive load for vip's by

a machine learning approach," in *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–6, 2021.

[155] E. Kutafina, A. Heiligers, R. Popovic, A. Brenner, B. Hankammer, S. M. Jonas, K. Mathiak, and J. Zweerings, "Tracking of mental workload with a mobile eeg sensor," *Sensors*, vol. 21, no. 15, 2021.

[156] S. Shao, T. Wang, C. Song, Y. Su, Y. Wang, and C. Yao, "Fine-grained and multi-scale motif features for cross-subject mental workload assessment using bi-lstm," *Journal of Mechanics in Medicine and Biology*, vol. 21, no. 05, p. 2140020, 2021.

[157] B. Balamurugan, M. Mullai, S. Soundararajan, S. Selvakanmani, and D. Arun, "Brain–computer interface for assessment of mental efforts in e-learning using the nonmarkovian queueing model," *Computer Applications in Engineering Education*, vol. 29, no. 2, pp. 394–410, 2021.

[158] H. Ved and C. Yildirim, "Detecting mental workload in virtual reality using eeg spectral data: A deep learning approach," in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 173–178, 2021.

[159] J.-C. Cheng, C.-P. Hsiao, C.-W. Wei, and C.-E. Weng, "The cognitive load evaluation based on eeg with k-nearest neighbor algorithm," in *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2, 2021.

[160] N. Sciaraffa, D. Germano, A. Giorgi, V. Ronca, A. Vozzi, G. Borghini, G. Di Flumeri, F. Babiloni, and P. Aricò, "Mental effort estimation by passive bci: A cross-subject analysis," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 906–909, 2021.

[161] A. Diwakar, T. Kaur, C. Ralekar, and T. K. Gandhi, "Deep learning identifies brain cognitive load via eeg signals," in *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1–5, 2020.

[162] T.-T. N. Do, A. K. Singh, C. A. T. Cortes, and C.-T. Lin, "Estimating the cognitive load in physical spatial navigation," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 568–575, 2020.

[163] V. Pandey, D. K. Choudhary, V. Verma, G. Sharma, R. Singh, and S. Chandra, "Mental workload estimation using eeg," in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 83–86, 2020.

[164] P. Becerra-Sánchez, A. Reyes-Munoz, and A. Guerrero-Ibañez, "Feature selection model based on eeg signals for assessing the cognitive workload in drivers," *Sensors*, vol. 20, no. 20, 2020.

[165] W. Qiao and X. Bi, "Ternary-task convolutional bidirectional neural turing machine for assessment of eeg-based cognitive workload," *Biomedical Signal Processing and Control*, vol. 57, p. 101745, 2020.

[166] M. Plechawska-Wójcik, M. Tokovarov, M. Kaczorowska, and D. Zapała, "A three-class classification of cognitive workload based on eeg spectral data," *Applied Sciences*, vol. 9, no. 24, 2019.

[167] J. Tao, Z. Yin, L. Liu, Y. Tian, Z. Sun, and J. Zhang, "Individual-specific classification of mental workload levels via an ensemble heterogeneous extreme learning machine for eeg modeling," *Symmetry*, vol. 11, no. 7, 2019.

[168] H. Gu, Z. Yin, and J. Zhang, "Eeg based mental workload assessment via a hybrid classifier of extreme learning machine and support vector machine," in *2019 Chinese Control Conference (CCC)*, pp. 8398–8403, 2019.

[169] Z. Yin, M. Zhao, W. Zhang, Y. Wang, Y. Wang, and J. Zhang, "Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework," *Neurocomputing*, vol. 347, pp. 212–229, 2019.

[170] P. Zhang, X. Wang, J. Chen, W. You, and W. Zhang, "Spectral and temporal feature learning with two-stream neural networks for mental workload assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1149–1159, 2019.

[171] G. Di Flumeri, P. Aricò, G. Borghini, N. Sciaraffa, V. Ronca, A. Vozzi, S. F. Storti, G. Menegaz, P. Fiorini, and F. Babiloni, "Eeg-based workload index as a taxonomic tool to evaluate the similarity of different robot-assisted surgery systems," in *Human Mental Workload: Models and Applications* (L. Longo and M. C. Leva, eds.), (Cham), pp. 105–117, Springer International Publishing, 2019.

[172] N. Sciaraffa, P. Aricò, G. Borghini, G. D. Flumeri, A. D. Florio, and F. Babiloni, "On the use of machine learning for eeg-based workload assessment: Algorithms comparison in a realistic task," in *Human Mental Workload: Models and Applications* (L. Longo and M. C. Leva, eds.), (Cham), pp. 170–185, Springer International Publishing, 2019.

[173] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 1, pp. 31–42, 2019.

[174] V. Parekh, M. Bilalpur, C. V. Jawahar, S. Kumar, S. Winkler, and R. Subramanian, "Investigating the generalizability of eeg-based cognitive load estimation across visualizations," in *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*, ICMI '18, (New York, NY, USA), Association for Computing Machinery, 2018.

[175] J. A. Blanco, M. K. Johnson, K. J. Jaquess, H. Oh, L.-C. Lo, R. J. Gentili, and B. D. Hatfield, "Quantifying cognitive workload in simulated flight using passive, dry eeg measurements," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 373–383, 2018.

[176] A. Appriou, A. Cichocki, and F. Lotte, "Towards robust neuroadaptive hci: Exploring modern machine learning methods to estimate mental workload from eeg signals,"

in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, (New York, NY, USA), p. 1–6, Association for Computing Machinery, 2018.

[177] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on eeg data," *Pattern Recognition*, vol. 76, pp. 582–595, 2018.

[178] A. Saha, V. Minz, S. Bonela, S. R. Sreeja, R. Chowdhury, and D. Samanta, "Classification of eeg signals for cognitive load estimation using deep learning architectures," in *Intelligent Human Computer Interaction* (U. S. Tiwary, ed.), (Cham), pp. 59–68, Springer International Publishing, 2018.

[179] B. S. Cheema, S. Samima, M. Sarma, and D. Samanta, "Mental workload estimation from eeg signals using machine learning algorithms," in *Engineering Psychology and Cognitive Ergonomics* (D. Harris, ed.), (Cham), pp. 265–284, Springer International Publishing, 2018.

[180] Z. Dai, A. Bezerianos, A. S.-H. Chen, and Y. Sun, "Mental workload classification in n-back tasks based on single trial eeg," 2017.

[181] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using eeg and an adaptive deep learning model," *Biomedical Signal Processing and Control*, vol. 33, pp. 30–47, 2017.

[182] Y. Zhou, T. Xu, Y. Cai, X. Wu, and B. Dong, "Monitoring cognitive workload in online videos learning through an eeg-based brain-computer interface," in *Learning and Collaboration Technologies. Novel Learning Ecosystems* (P. Zaphiris and A. Ioannou, eds.), (Cham), pp. 64–73, Springer International Publishing, 2017.

[183] A. Abrantes, E. Comitz, P. Mosaly, and L. Mazur, "Classification of eeg features for prediction of working memory load," in *Advances in The Human Side of Service Engineering* (T. Z. Ahram and W. Karwowski, eds.), (Cham), pp. 115–126, Springer International Publishing, 2017.

[184] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Bonelli, A. Golfetti, S. Pozzi, J.-P. Imbert, G. Granger, R. Benhacene, and F. Babiloni, "Adaptive automation triggered by eeg-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[185] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, I. Graziani, J.-P. Imbert, G. Granger, R. Benhacene, M. Terenzi, S. Pozzi, and F. Babiloni, "Reliability over time of eeg-based mental workload evaluation during air traffic management (atm) tasks," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7242–7245, 2015.

[186] Y. Ke, H. Qi, L. Zhang, S. Chen, X. Jiao, P. Zhou, X. Zhao, B. Wan, and D. Ming, "Towards an effective cross-task mental workload recognition model using electroencephalography based on feature selection and support vector machine regression," *International Journal of Psychophysiology*, vol. 98, no. 2, Part 1, pp. 157–166, 2015.

[187] S. I. Dimitriadis, Y. Sun, K. Kwok, N. A. Laskaris, N. Thakor, and A. Bezerianos, "Cognitive workload assessment based on the tensorial treatment of eeg estimates of cross-frequency phase interactions," *Annals of Biomedical Engineering*, vol. 43, pp. 977–989, Apr 2015.

[188] B. N. Penaranda and C. L. Baldwin, "Temporal factors of eeg and artificial neural network classifiers of mental workload," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, pp. 188–192, 2012.

[189] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, (New York, NY, USA), p. 835–844, Association for Computing Machinery, 2008.

[190] M. F. Hinss, E. S. Jahanpour, B. Somon, L. Pluchon, F. Dehais, and R. N. Roy, "COG-BCI database: A multi-session and multi-task EEG cognitive dataset for passive brain-computer interfaces," Dec. 2022.

[191] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[192] MATLAB, *Version: 9.14.0.2206163 (R2023a)*. Natick, Massachusetts, United States: The MathWorks Inc., 2023.

[193] M. F. Hinss, L. Darmet, B. Somon, E. Jahanpour, F. Lotte, S. Ladouce, and R. N. Roy, "An EEG dataset for cross-session mental workload estimation: Passive BCI competition of the Neuroergonomics Conference 2021," July 2021. The project was validated by the local ethical committee of the University of Toulouse (CER number 2021-342).

[194] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide (nasa/tm-2011-217164)," tech. rep., Hampton, VA: National Aeronautics and Space Administration, Langley Research Center, 2011.

[195] I. Winkler, S. Debener, K.-R. Müller, and M. Tangermann, "On the influence of high-pass filtering on ica-based artifact reduction in eeg-erp," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4101–4105, 2015.

[196] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis," *NeuroImage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[197] P. Ablin, J.-F. Cardoso, and A. Gramfort, "Faster independent component analysis by preconditioning with hessian approximations," *IEEE Transactions on Signal Processing*, vol. 66, no. 15, pp. 4040–4049, 2018.

[198] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181–197, 2019.

[199] L. Pion-Tonachini, S. Makeig, and K. Kreutz-Delgado, "Crowd labeling latent dirichlet allocation," *Knowledge and information systems*, vol. 53, pp. 749–765, Dec 2017.

[200] A. Delorme and S. Makeig, "This is no "ica bug": response to the article,"ica's bug: how ghost ics emerge from effective rank deficiency caused by eeg electrode interpolation and incorrect re-referencing"," *Frontiers in Neuroimaging*, vol. 2, 2023.

[201] M. Fraschini, M. Demuru, A. Crobe, F. Marrosu, C. J. Stam, and A. Hillebrand, "The effect of epoch length on estimated eeg functional connectivity and brain network organisation," *Journal of neural engineering*, vol. 13, p. 036015, may 2016.

[202] X. J. Li, P. T. Dao, and A. Griffin, "Effect of epoch length on compressed sensing of eeg," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, IEEE, 2018.

[203] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: higher-order graph neural networks," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pp. 4602–4609, AAAI Press, 2019.

[204] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[205] L. Biewald, "Experiment tracking with weights and biases," 2020. Software available from wandb.com.

[206] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, KDD '19, (New York, NY, USA), pp. 2623–2631, Association for Computing Machinery, 2019.

[207] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[208] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013. Twenty-fifth Anniversay Commemorative Issue.

[209] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, pmlr, PMLR, 07–09 Jul 2015.

[210] A.-M. Brouwer, M. A. Hogervorst, M. Holewijn, and J. B. van Erp, "Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort," *International Journal of Psychophysiology*, vol. 93, no. 2, pp. 242–252, 2014.

[211] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[212] F. Scarselli, A. C. Tsoi, and M. Hagenbuchner, "The vapnik–chervonenkis dimension of graph and recursive neural networks," *Neural Networks*, vol. 108, pp. 248–259, 2018.

[213] V. Doğan, "Deep learning classification of cognitive workload levels from eeg wavelet transform images," Master's thesis, Middle East Technical University, 2023.

[214] Y. Ke, T. Wang, F. He, S. Liu, and D. Ming, "Enhancing eeg-based cross-day mental workload classification using periodic component of power spectrum," *Journal of Neural Engineering*, vol. 20, p. 066028, dec 2023.

[215] S. Wang, Y. Liu, and D. Hu, "An early warning approach for pilots' cognitive tipping points based multi-modal signals," *IEEE Transactions on Automation Science and Engineering*, pp. 1–12, 2024.

[216] T. Wang, Y. Ke, Y. Huang, F. He, W. Zhong, S. Liu, and D. Ming, "Using semi-supervised domain adaptation to enhance eeg-based cross-task mental workload classification performance," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 12, pp. 7032–7039, 2024.

[217] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, vol. 2, pp. 1122–1128, 2006.

[218] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[219] C. Villani *et al.*, *Optimal transport: old and new*, vol. 338. Springer, 2009.

[220] J. Bertrand and M. Puel, "The optimal mass transport problem for relativistic costs," *Calculus of Variations and Partial Differential Equations*, vol. 46, no. 1, pp. 353–374, 2013.

[221] L. V. Kantorovich, "On the translocation of masses.," *Journal of mathematical sciences*, vol. 133, no. 4, 2006.

[222] C. Villani, *Topics in optimal transportation*, vol. 58. American Mathematical Soc., 2021.

[223] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[224] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the monge-kantorovich mass transfer problem," *Numerische Mathematik*, vol. 84, no. 3, pp. 375–393, 2000.

[225] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[226] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[227] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, "Stochastic optimization for large-scale optimal transport," *Advances in neural information processing systems*, vol. 29, 2016.

[228] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, *et al.*, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.

[229] Z. Zhang, Z. Zhao, H. Qu, C. Liu, and L. Pang, "A mental workload classification method based on gcn modified by squeeze-and-excitation residual," *Mathematics*, vol. 11, no. 5, p. 1189, 2023.

[230] M.-K. Jung, H. Kim, S. Lee, J. B. Kim, and D.-J. Kim, "Interpretability of hybrid feature using graph neural networks from mental arithmetic based eeg," in *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–5, IEEE, 2023.

[231] N. Panwar, V. Pandey, and P. P. Roy, "Eeg-cognet: A deep learning framework for cognitive state assessment using eeg brain connectivity," *Biomedical Signal Processing and Control*, vol. 98, p. 106770, 2024.

Appendix A

COMPILED GUIDELINES

Table 30: Compiled Checklist in Relation to CRISP-DM Phases

CRISP-DM Phase	**Research Step**	**Compiled Checklist**
Business Understanding	Problem Definition	Problem/Scope statement
		Related literature †
Data Understanding	General	Dataset (name if public or private)
	Participant Selection	Number of participants
		Participant recruitment method † (e.g., direct mailing, advertisements)
		Participant sampling strategy (that constrain inclusion to a particular group/including population from which the participants were sampled)
		Age of participants
		Gender of participants
		Education level of participants
		Medications taken by the participants
		Prior/Current illness of participants
		Information on sleep deprivation
		Handedness of participants
		Consent of participants †
	Experimental setup	Type of EEG sensor/device (including make and model)
		Number of sensors
		Sensor locations
		Sampling rate
		Online filters (Type of filter and parameters)
		Electrode impedance
		Amplifier characteristics
		Measurement procedures
		Recording environment
		Participant seated or lying down status
	Experimental task information	Task description
		Characteristics of stimuli
		Instructions for the task

Table 30: cont.

		Number of runs and sessions
		Clear timeline including - Timing of all stimuli/events - Intertrial intervals
		Software and hardware used for stimulus presentation
	Task-free recordings	Definition
		Timing
		Eyes open vs closed status
		If eyes open, fixation point usage
	Behavioral measures	Nature of the response
		Acquisition device and parameters
		Interface with EEG data and calibration procedures
		Errors and outliers handling
	Subjective measures	Subjective assessments recorded - Timing - Method
	Labeling	Definition
	Analysis	Recording length
		Statistical analysis to justify the number of trials and number of participants †
		Statistical analysis for descriptives of the collected measurements †
	Open sourcing	Open-sourced raw data with version control
		Open-sourced code for data collection with version control
		Open-sourced code/software for task execution with version control
Data Preparation	General	Flow of algorithms for preprocessing and transforming data
		Seeds for random number generators
	Sensor/Segment removal	For sensor/segment removal - Detection method and criteria - Interpolation parameters - Removed sensors/segments
	Artifact removal	For artifact removal/correction - Method - Range of parameters - Types of artifacts identified - Criteria to identify - Number/proportion of removed artifacts - Position of removed artifacts
		For signal–noise separation methods - Method - Parameters - Number of ICs - How non-brain ICs were identified - How back-projection was performed

Table 30: cont.

	Downsampling	For downsampling - Method - Parameters
	Detrending	For detrending - Method - Parameters
	Filtering	For filtering - Filter type - Parameters
	Segmentation	For segmentation - Method - Parameters
	Baseline correction	For baseline correction - Method - Parameters
	Re-referencing	For re-referencing - Method
	Dimensionality Reduction	For dimensionality reduction - Method - Parameters
	Feature generation	For feature generation - Definition of features - Number of features - Method - Parameters
		Descriptive / Inferential statistics † - The statistical method - Parameters
	Feature selection	Feature selection - Method - Number of selected features - Parameters - Selected features
	Data split	For data split (as training, validation and test) - Method - Parameters
		Separate test set †
	Feature transformation	Feature transformation (Normalization, Standardization,. . .) - Method - Parameters
		Feature transformations applied using training data? †
	Data augmentation	Data augmentation - Method - Parameters
		Data augmentation applied using training data? †
	Environment	Computing infrastructure
		Dependencies
	Open sourcing	Open-sourced preprocessed data with version control

Table 30: cont.

		Open-sourced code for data preparation with version control
Modeling	General	ML problem (n-class MWL classification)
		Seeds for random number generators to prevent randomness in results
	Model	Algorithm name
		Explanation in detail, motivation and intended behaviour †
		Loss function and parameters
		Regularization method and parameters
		Model structure
	Training strategy	Hyperparameters of the model
		Method for hyperparameter tuning
		Hyperparameter ranges for tuning
		Number of trials for hyperparameter tuning
		Selected hyperparameters
	Model training	Optimization method and parameters
		Number of training epochs/iterations
		Additional methods used during training and their parameters if any (e.g. early stopping, ..)
	Metrics	Metrics for model evaluation
		Chance-level values of metrics †
		Data split that the metrics are calculated on
		When the dataset is imbalanced, metrics other than accuracy †
		Confusion matrix †
	Environment	Computing infrastructure
		Dependencies
	Open sourcing	Open-sourced trained models with version control
		Open-sourced code for model training with version control
Evaluation	Statistical analysis	Statistical analysis for significance of results † - Method - Parameters
	Metrics	Performance on independent test set †
		Computational resources for training †(e.g. Model size, Training time, Power consumption, Carbon emissions)
	Conclusion	Relation of results to the problem statement
	Open sourcing	Open-sourced code for evaluation with version control
Deployment	Deployment Technique	Deployment techniques for limited resources (e.g. quantization of the model, . . .) - Method - Parameters

Table 30: cont.

	Metrics	Computational resources for inference †(e.g. Model size, Inference time, Power consumption, Carbon emissions)
	Environment	Deployment infrastructure
		Dependencies
		Interfaces
		Schematic or sample view
	Deployment Test	Performance check †(to see both development and production environments yield sufficiently similar results given identical input data)
		Performance monitoring method after deployment †
[1]		

[1] Items on the checklist marked with a "†" reflect best practices or are related to generalizability while those that are not marked directly affect reproducibility.

Appendix B

EXPLORATORY DATA ANALYSIS PLOTS

In figures 29 - 54, SOM plots of features are illustrated across sessions before and after optimal transport when session 1, 2 and 3 are train, validation and test sets, respectively.

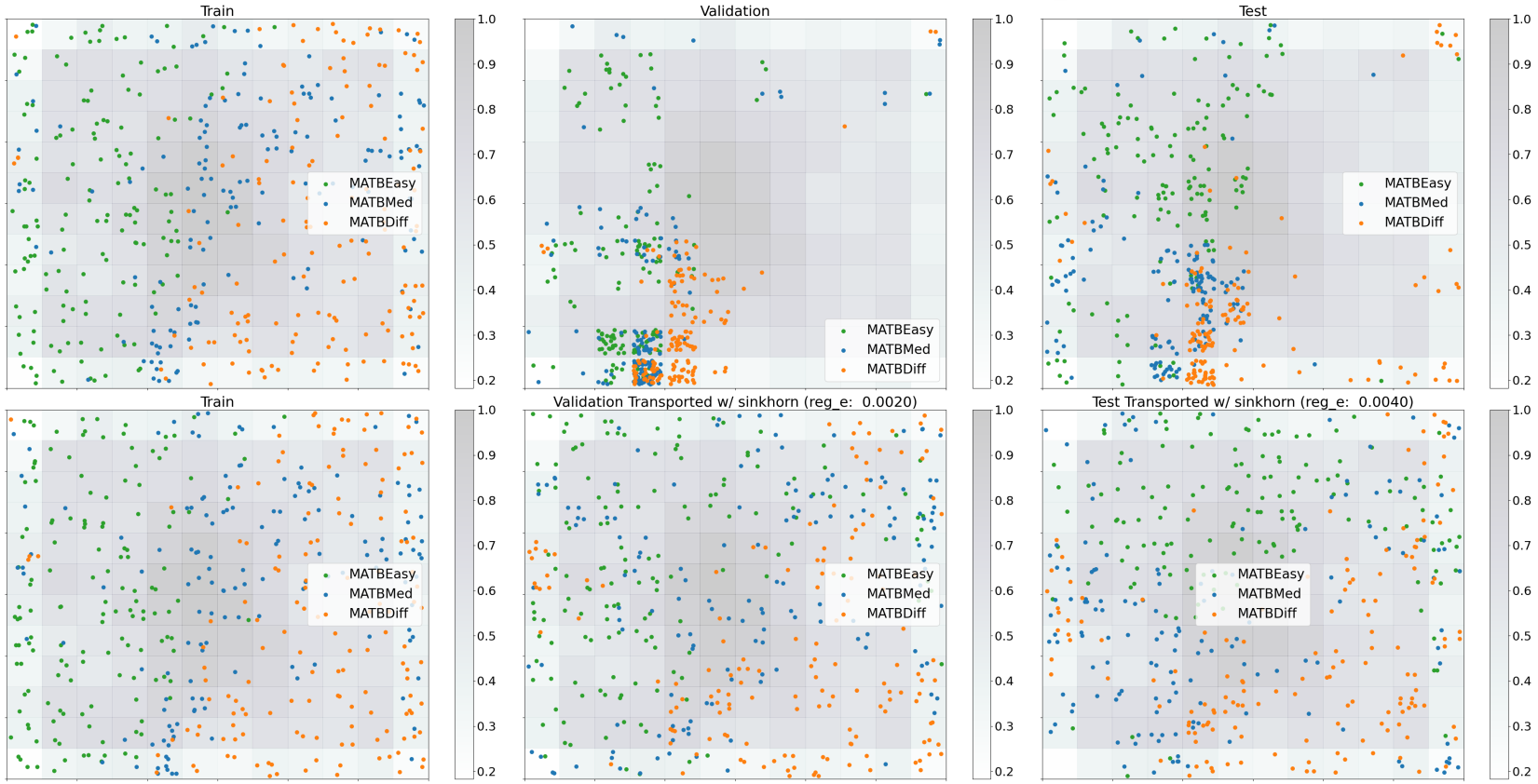


Figure 29: SOM Plots of Features (Subject 1)

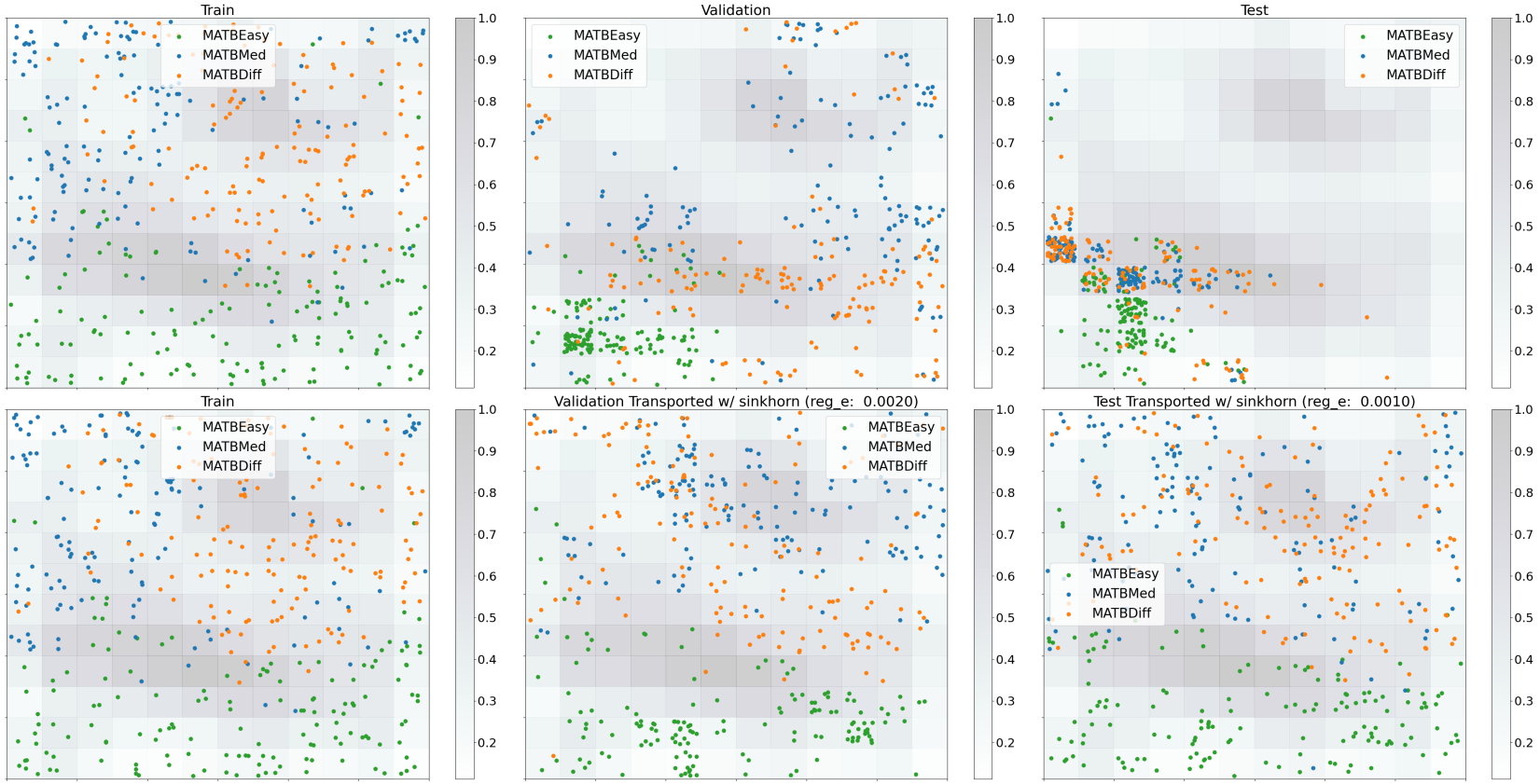


Figure 30: SOM Plots of Features (Subject 2)

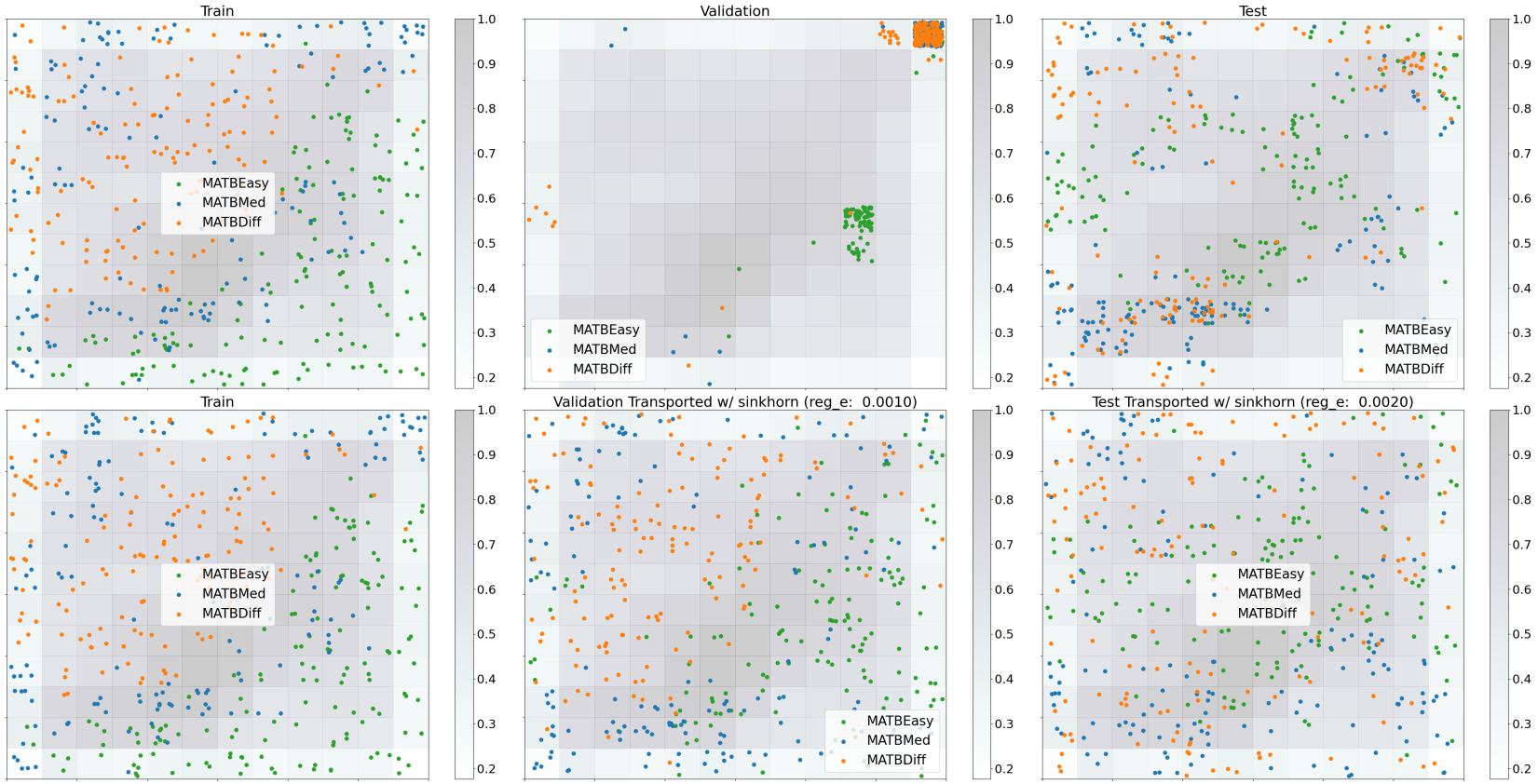


Figure 31: SOM Plots of Features (Subject 3)

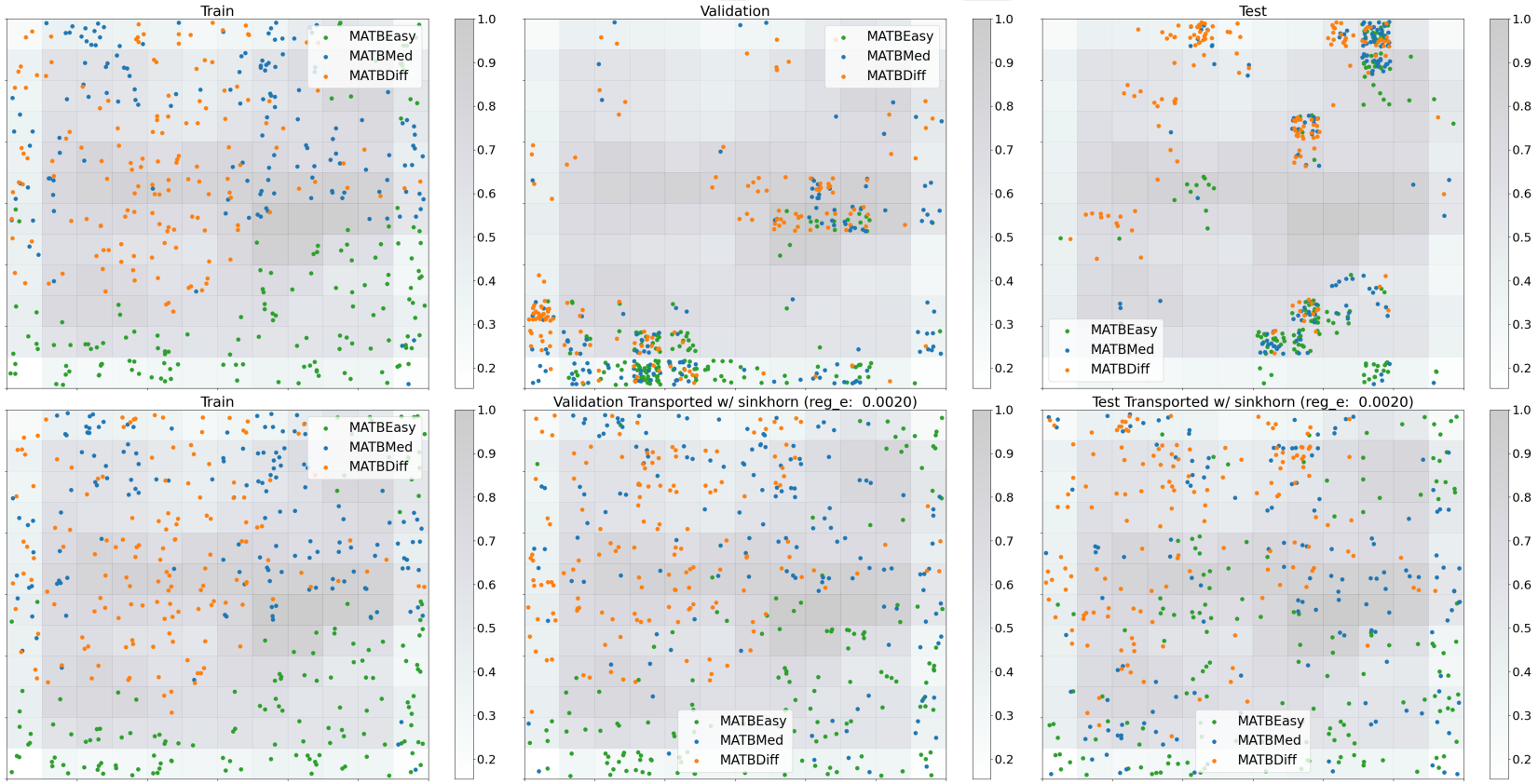


Figure 32: SOM Plots of Features (Subject 4)

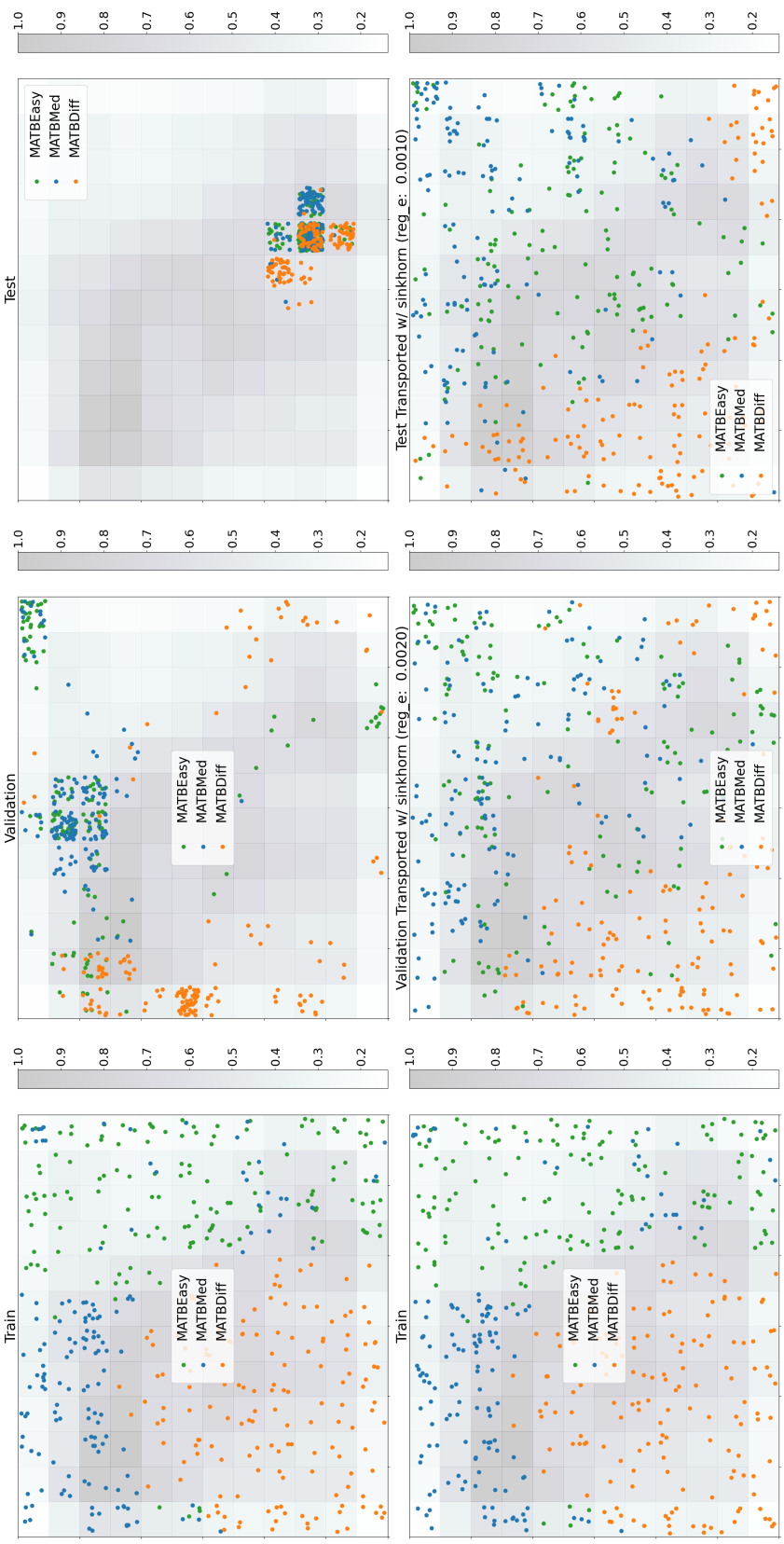


Figure 33: SOM Plots of Features (Subject 5)

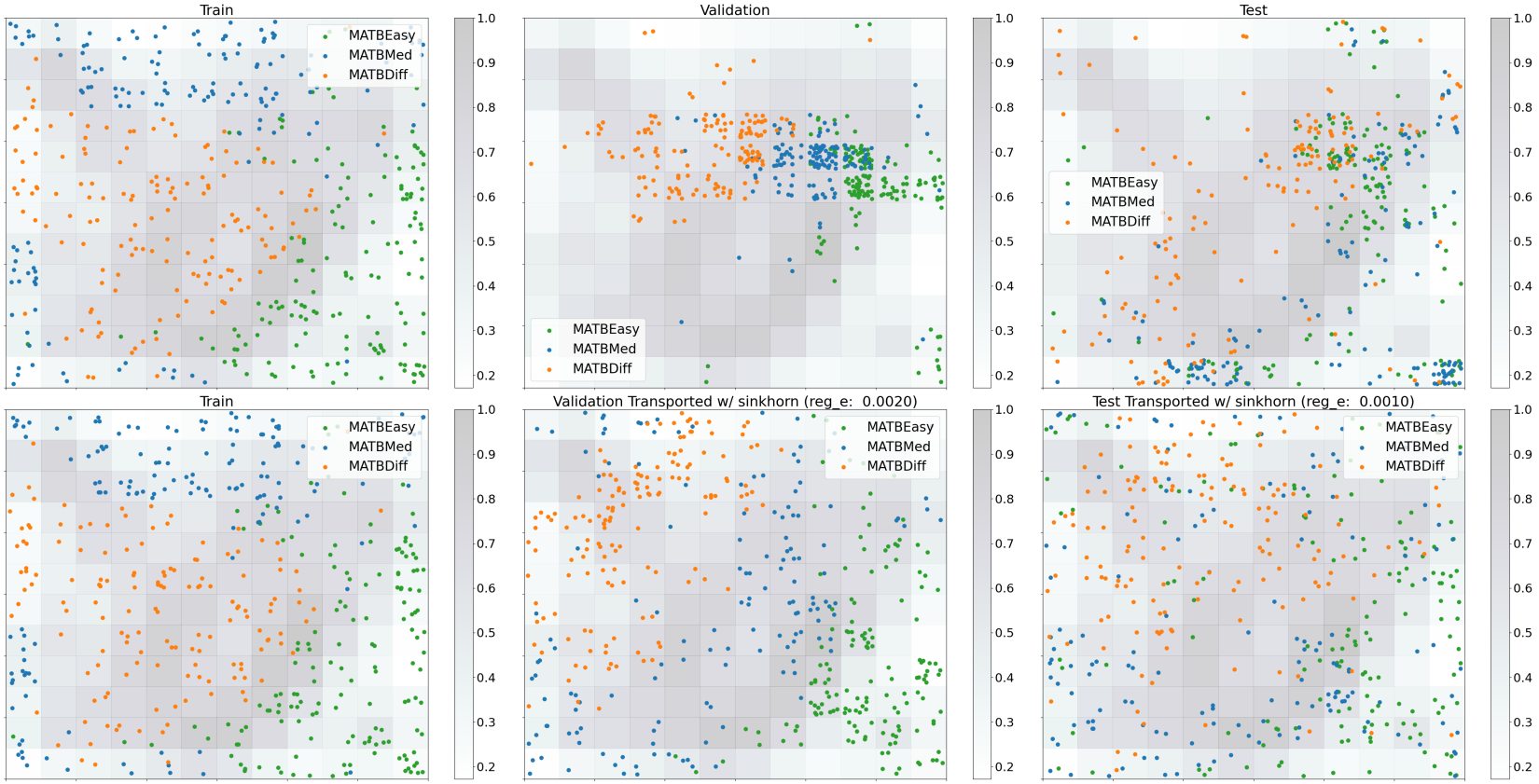


Figure 34: SOM Plots of Features (Subject 6)

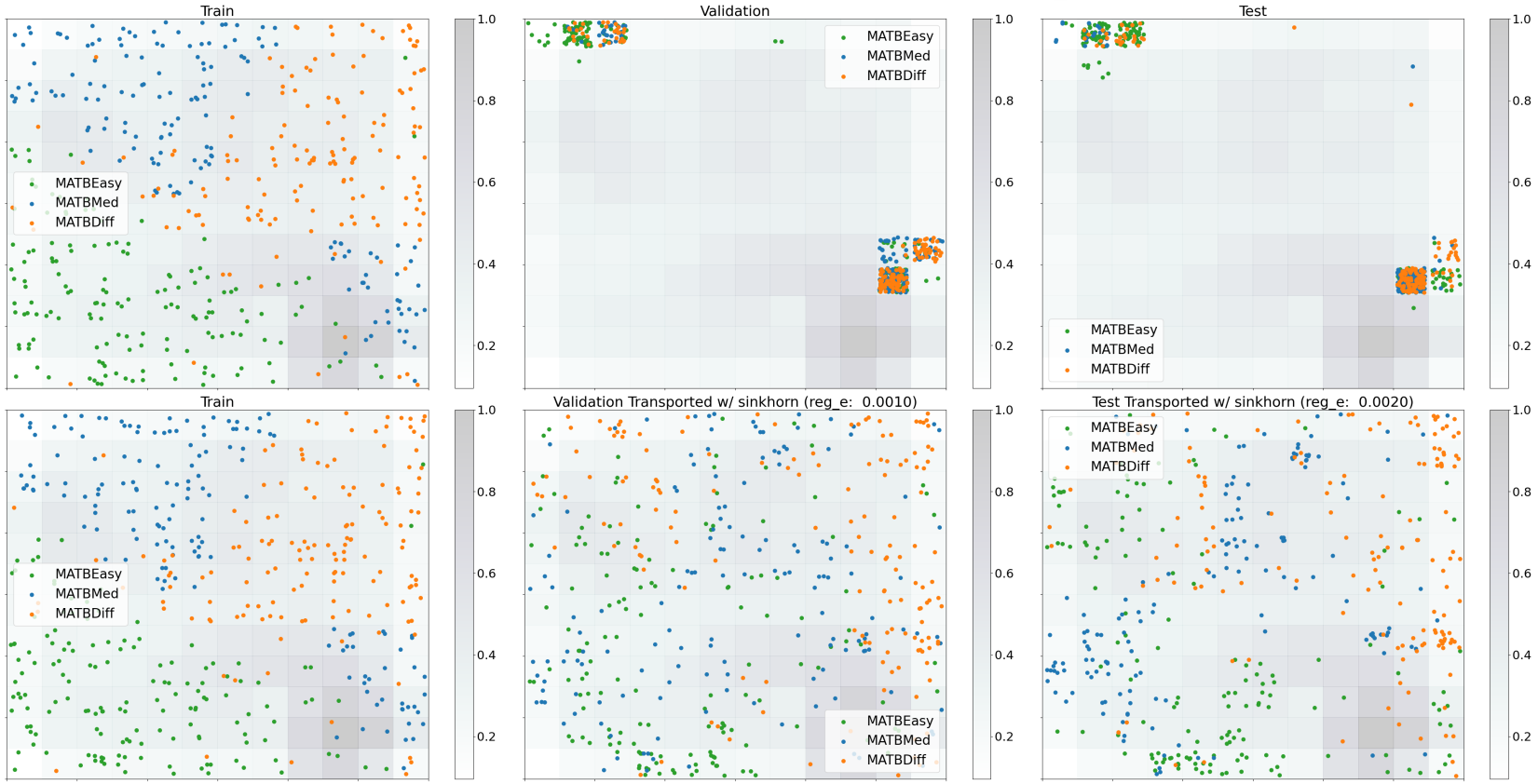


Figure 35: SOM Plots of Features (Subject 7)

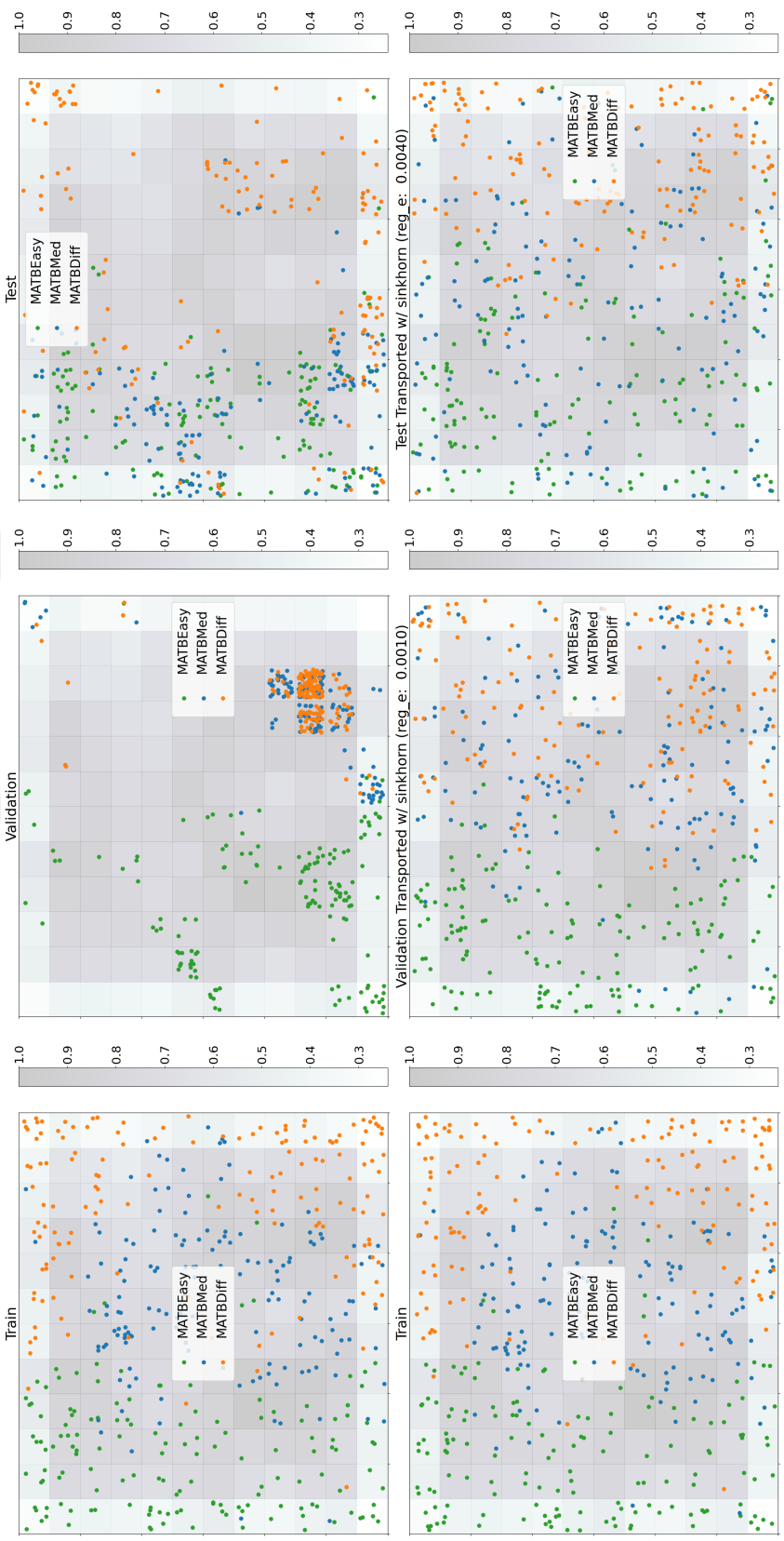


Figure 36: SOM Plots of Features (Subject 8)

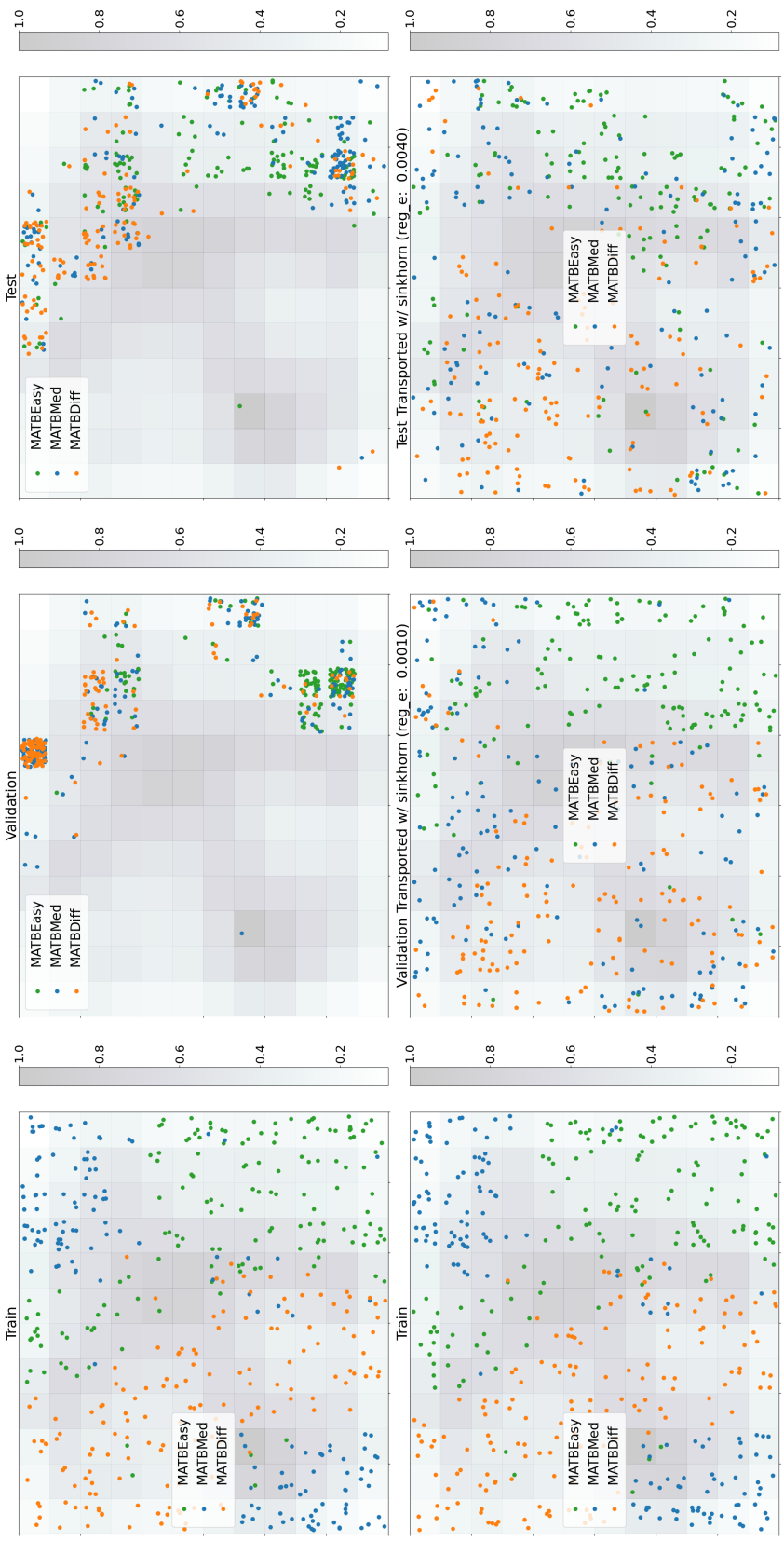


Figure 37: SOM Plots of Features (Subject 9)

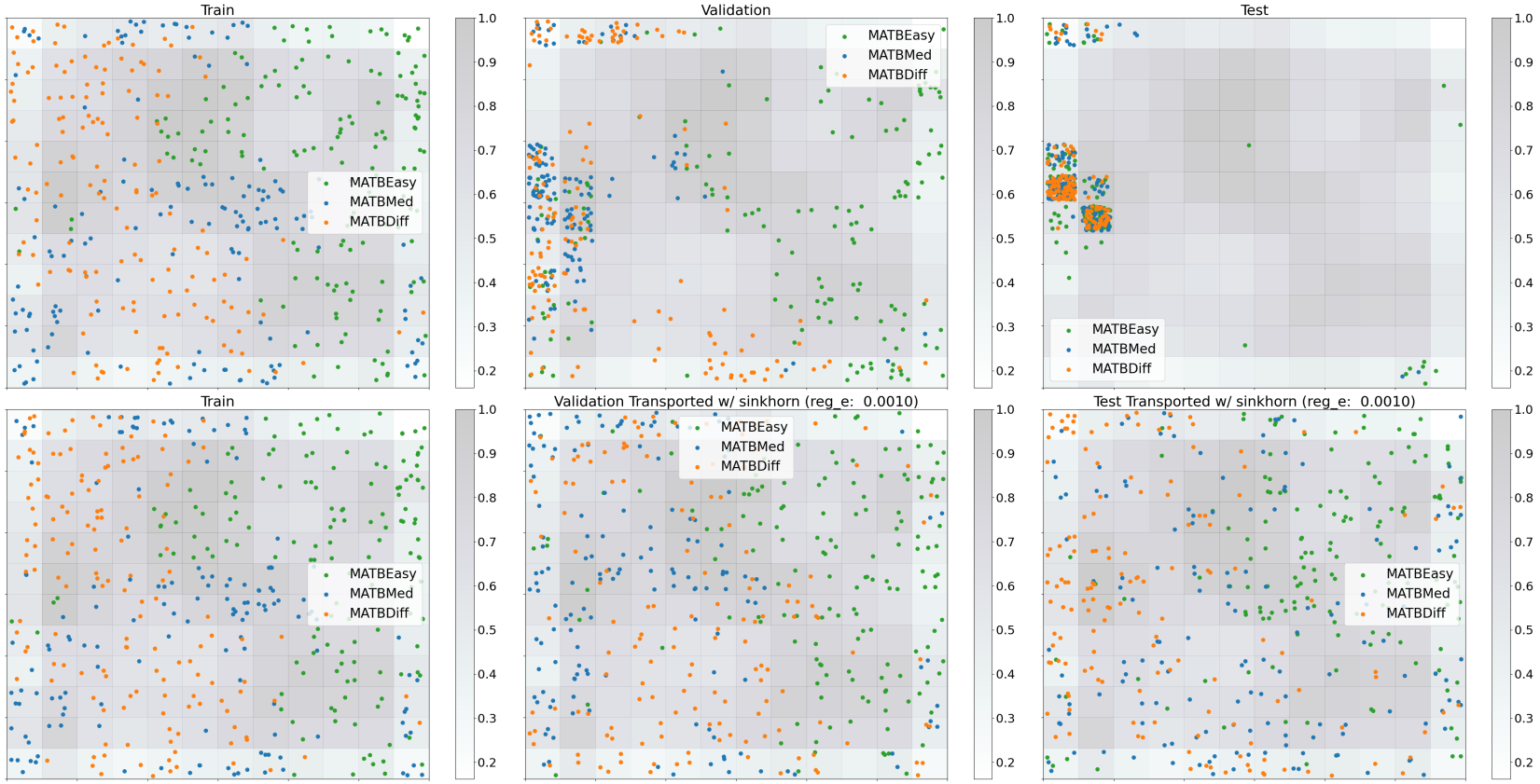


Figure 38: SOM Plots of Features (Subject 10)

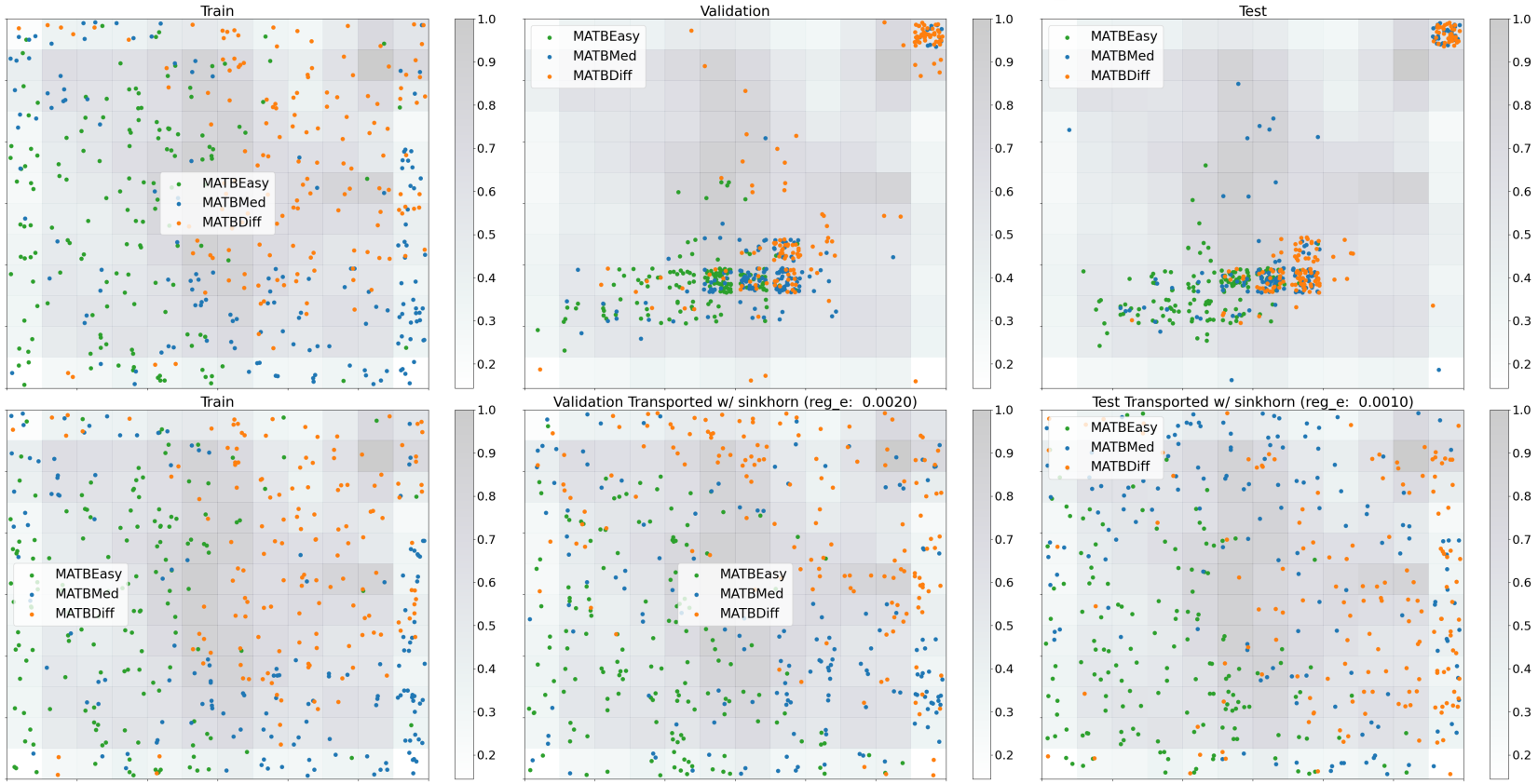


Figure 39: SOM Plots of Features (Subject 11)

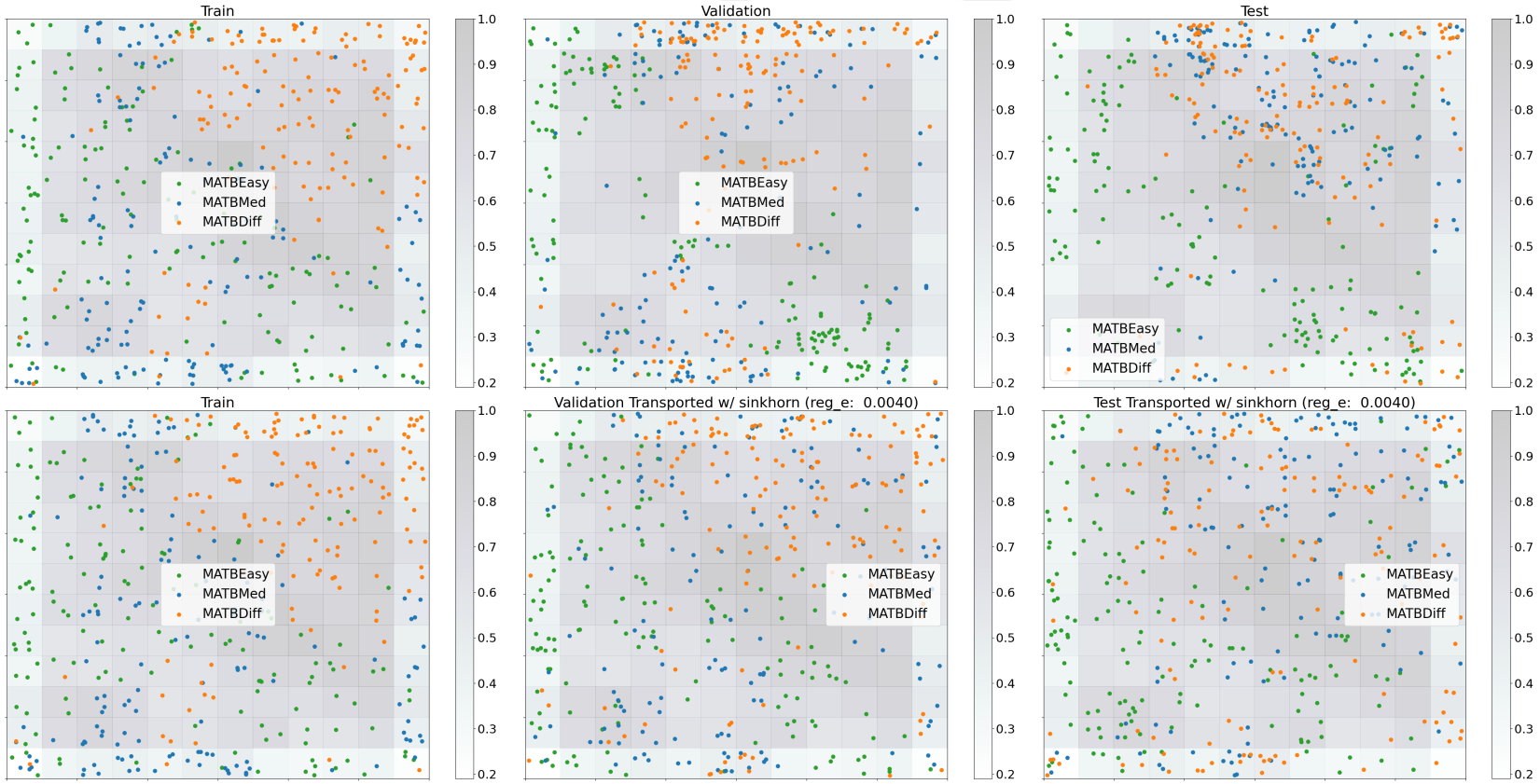


Figure 40: SOM Plots of Features (Subject 12)



Figure 41: SOM Plots of Features (Subject 13)

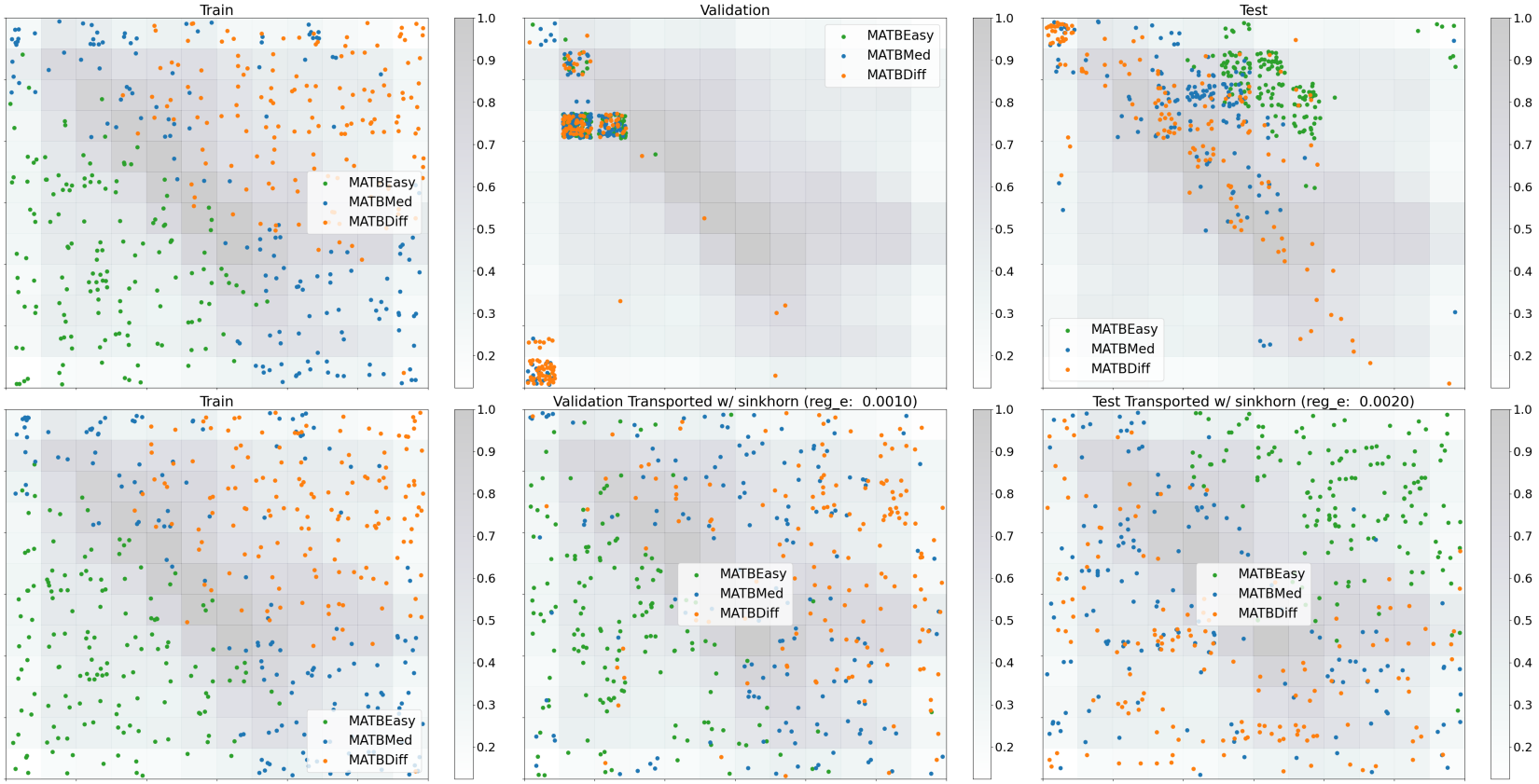


Figure 42: SOM Plots of Features (Subject 14)

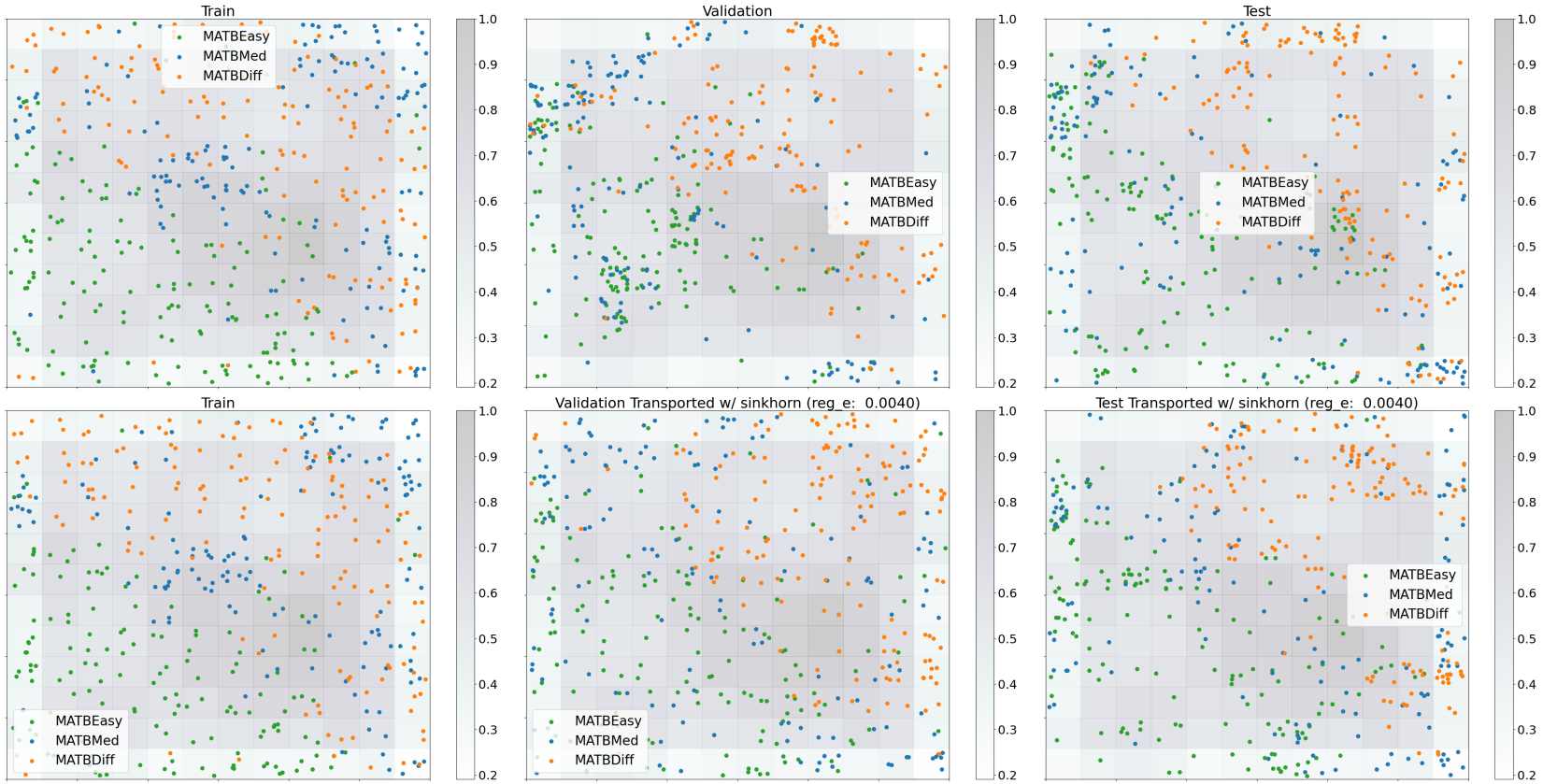


Figure 43: SOM Plots of Features (Subject 15)

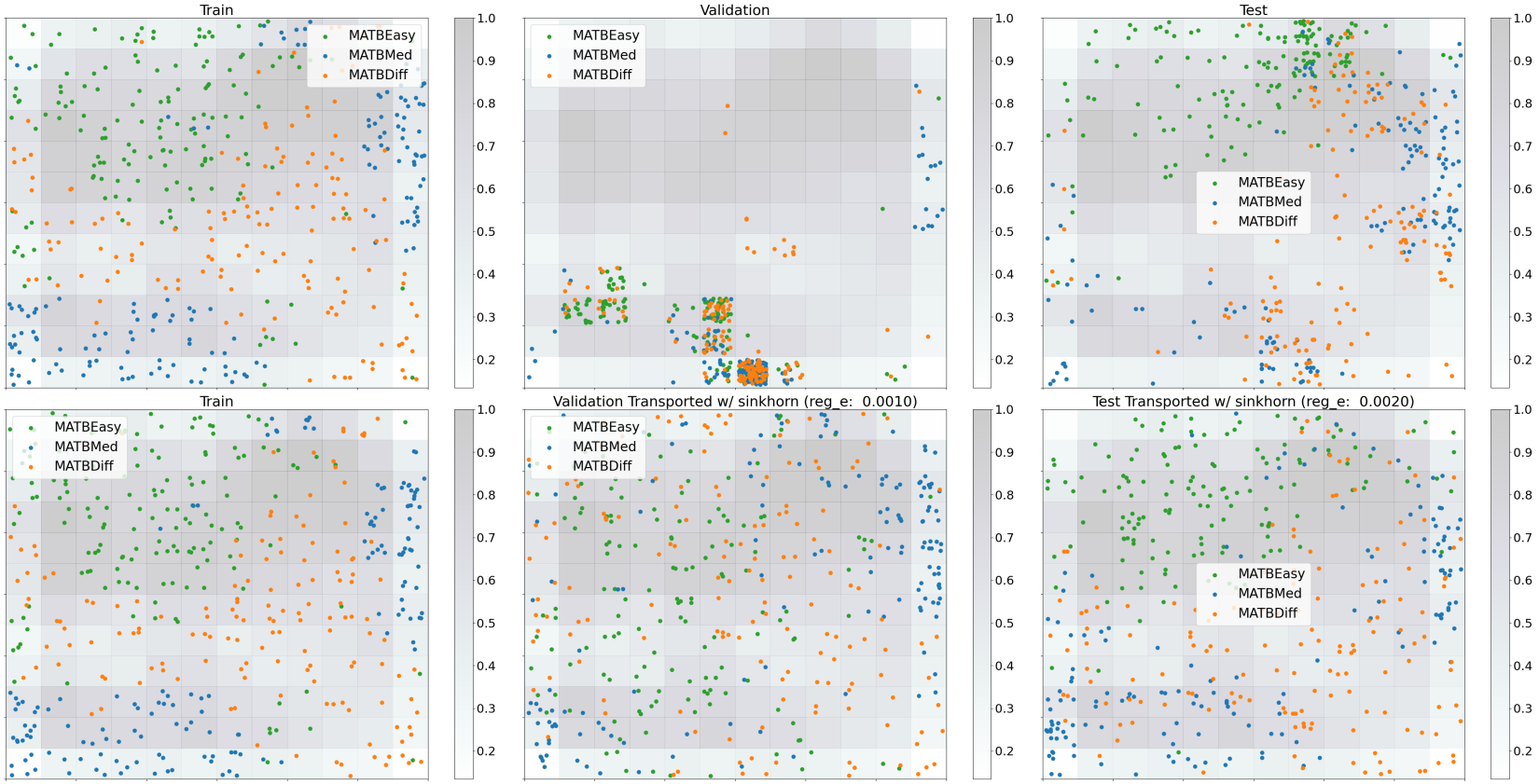


Figure 44: SOM Plots of Features (Subject 16)

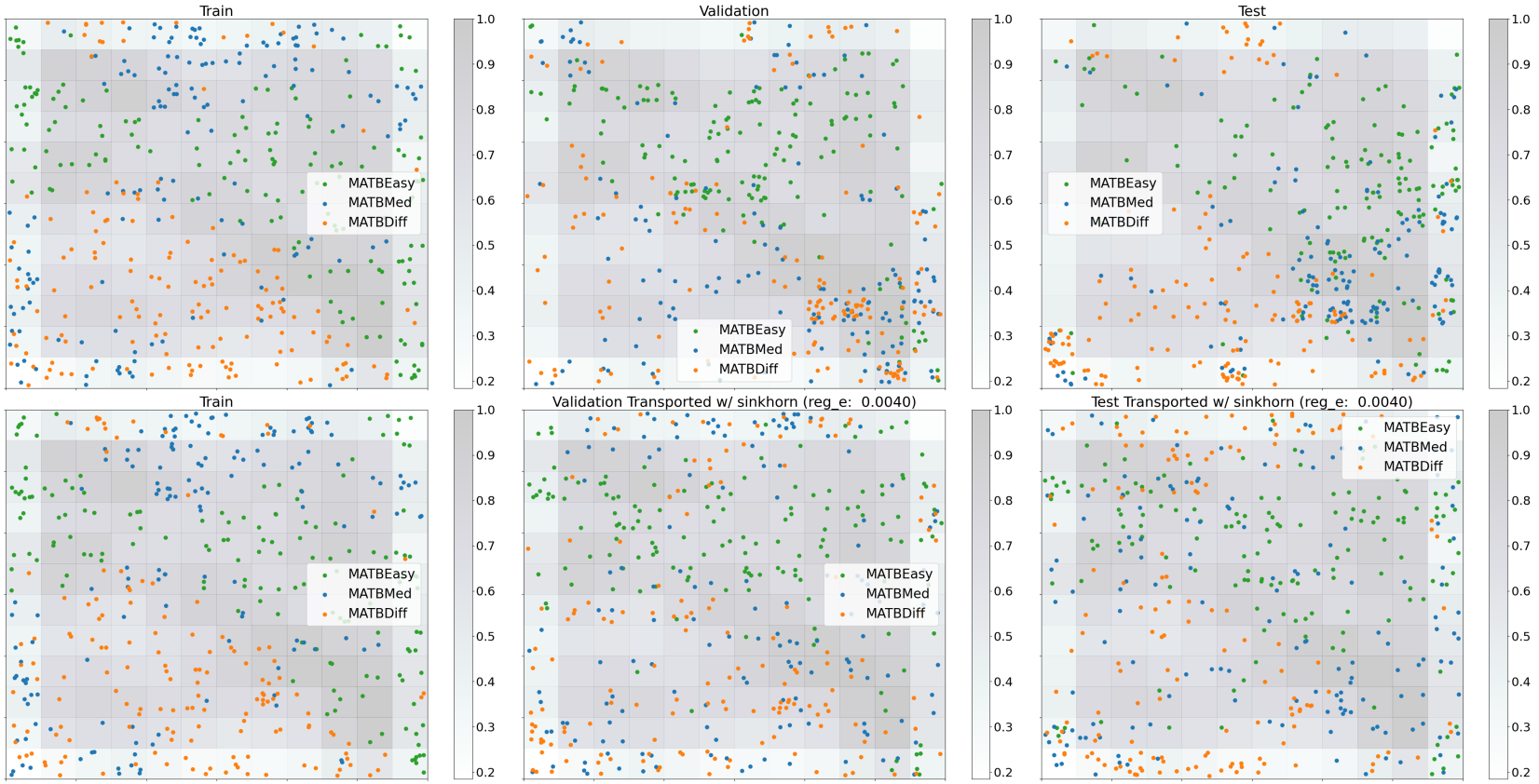


Figure 45: SOM Plots of Features (Subject 17)

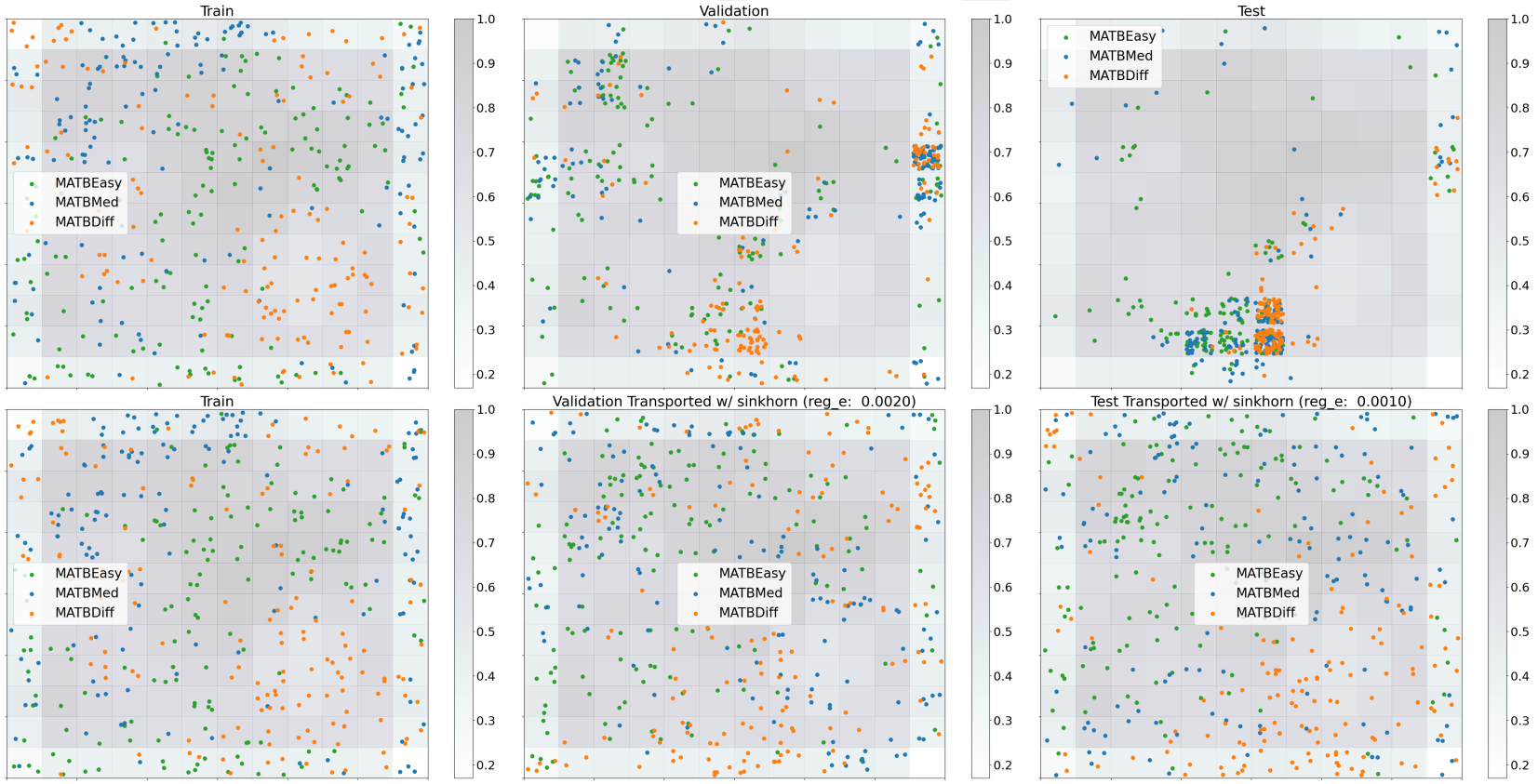


Figure 46: SOM Plots of Features (Subject 18)

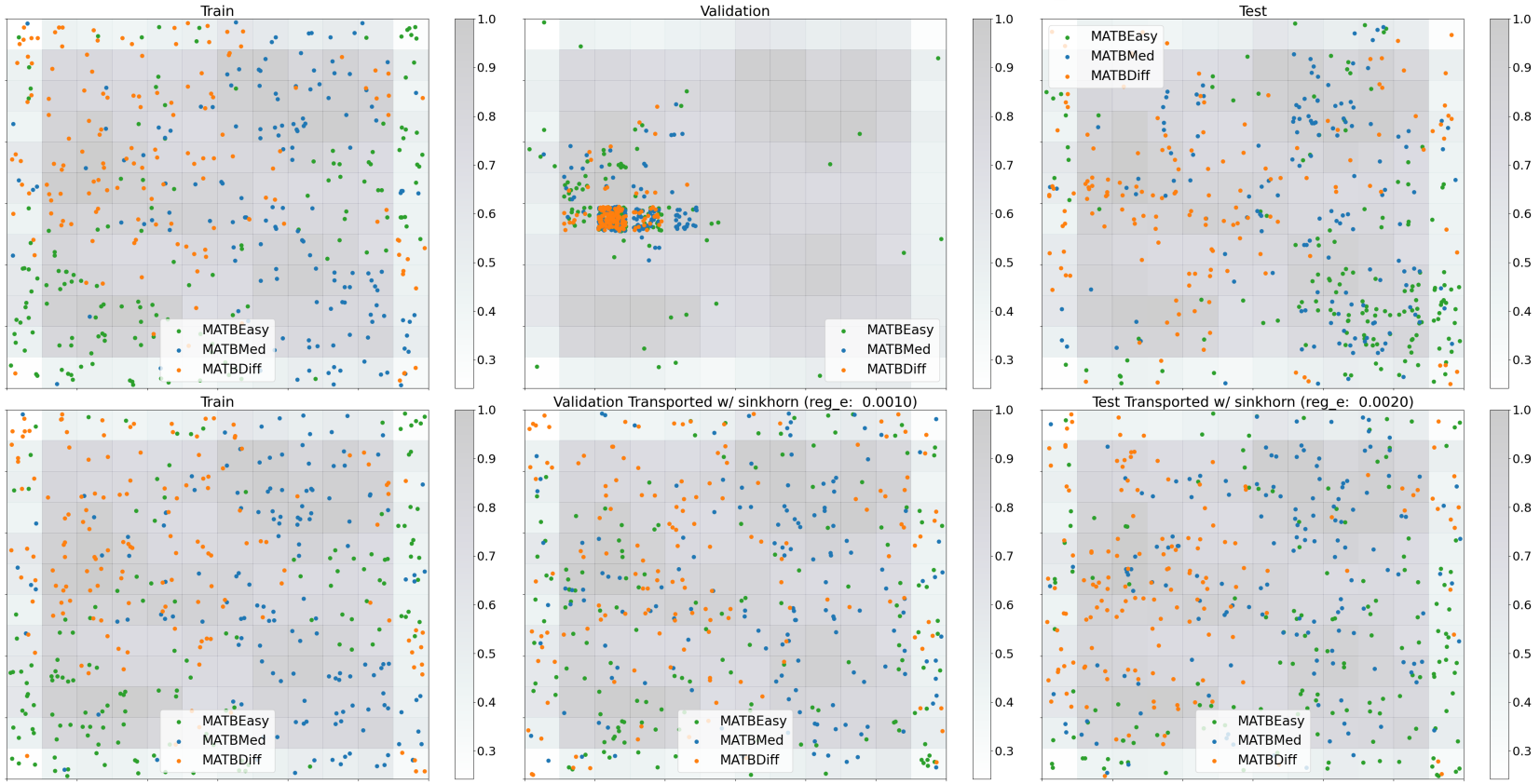


Figure 47: SOM Plots of Features (Subject 19)

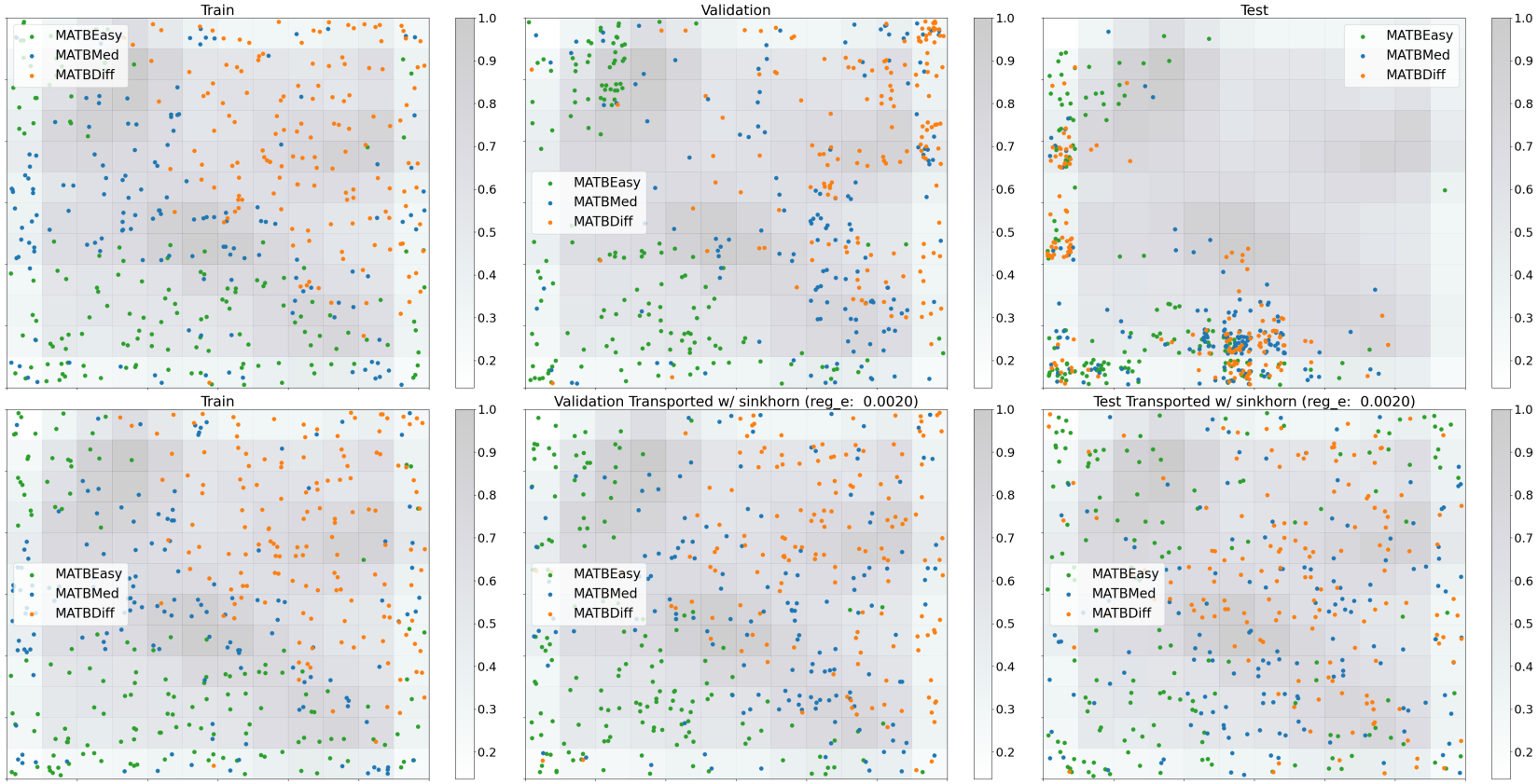


Figure 48: SOM Plots of Features (Subject 20)

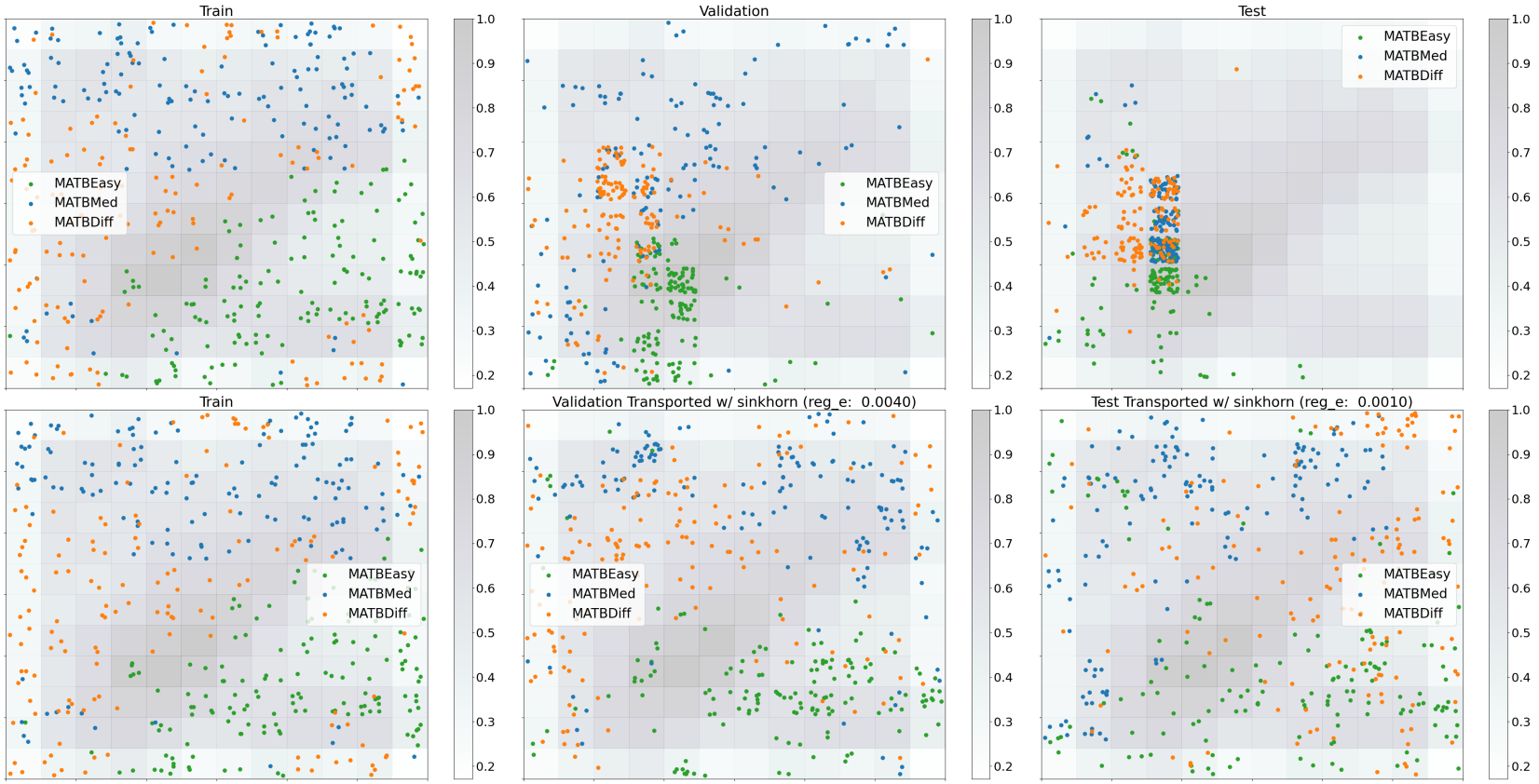


Figure 49: SOM Plots of Features (Subject 21)



Figure 50: SOM Plots of Features (Subject 22)

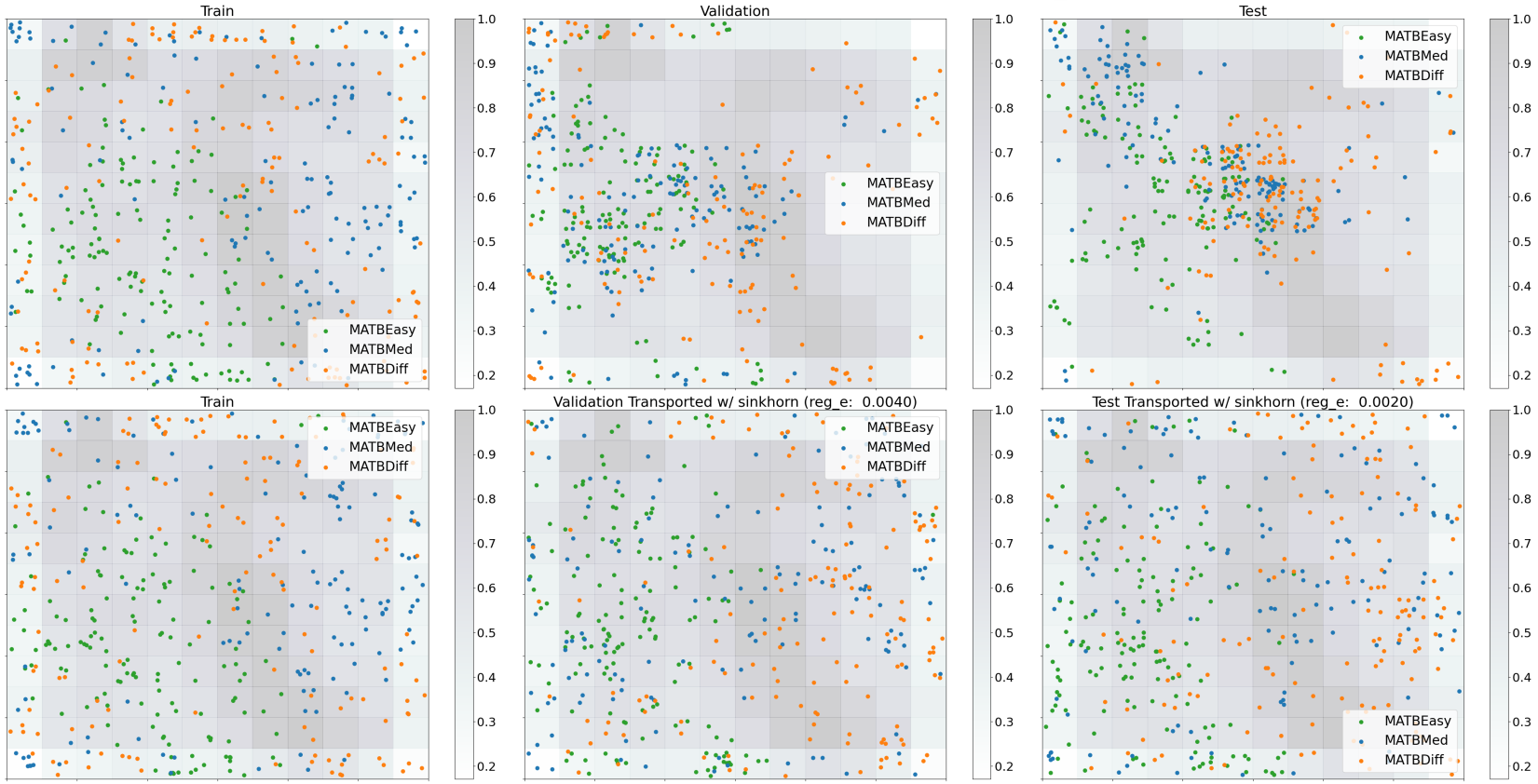


Figure 51: SOM Plots of Features (Subject 23)

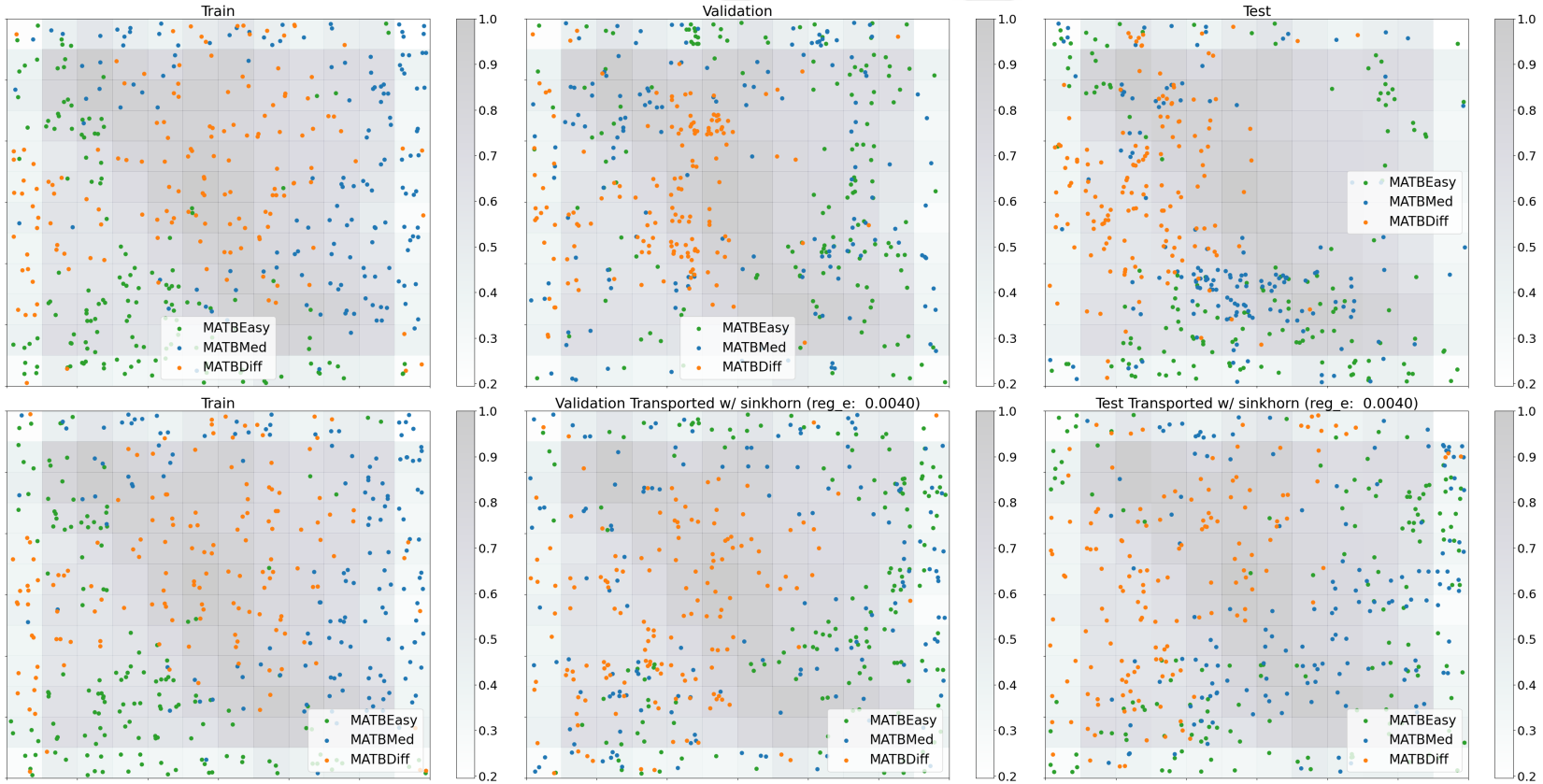


Figure 52: SOM Plots of Features (Subject 24)

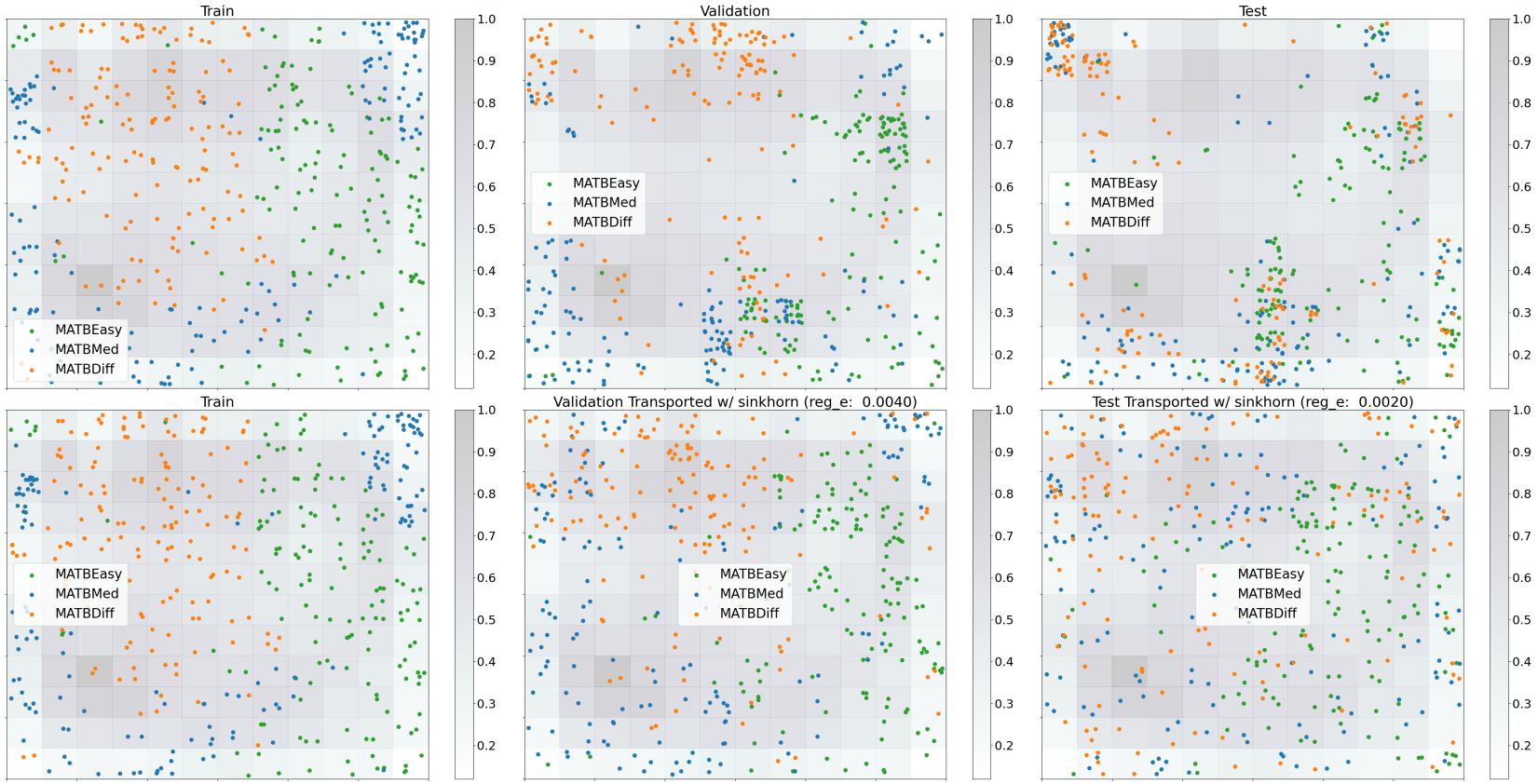


Figure 53: SOM Plots of Features (Subject 26)

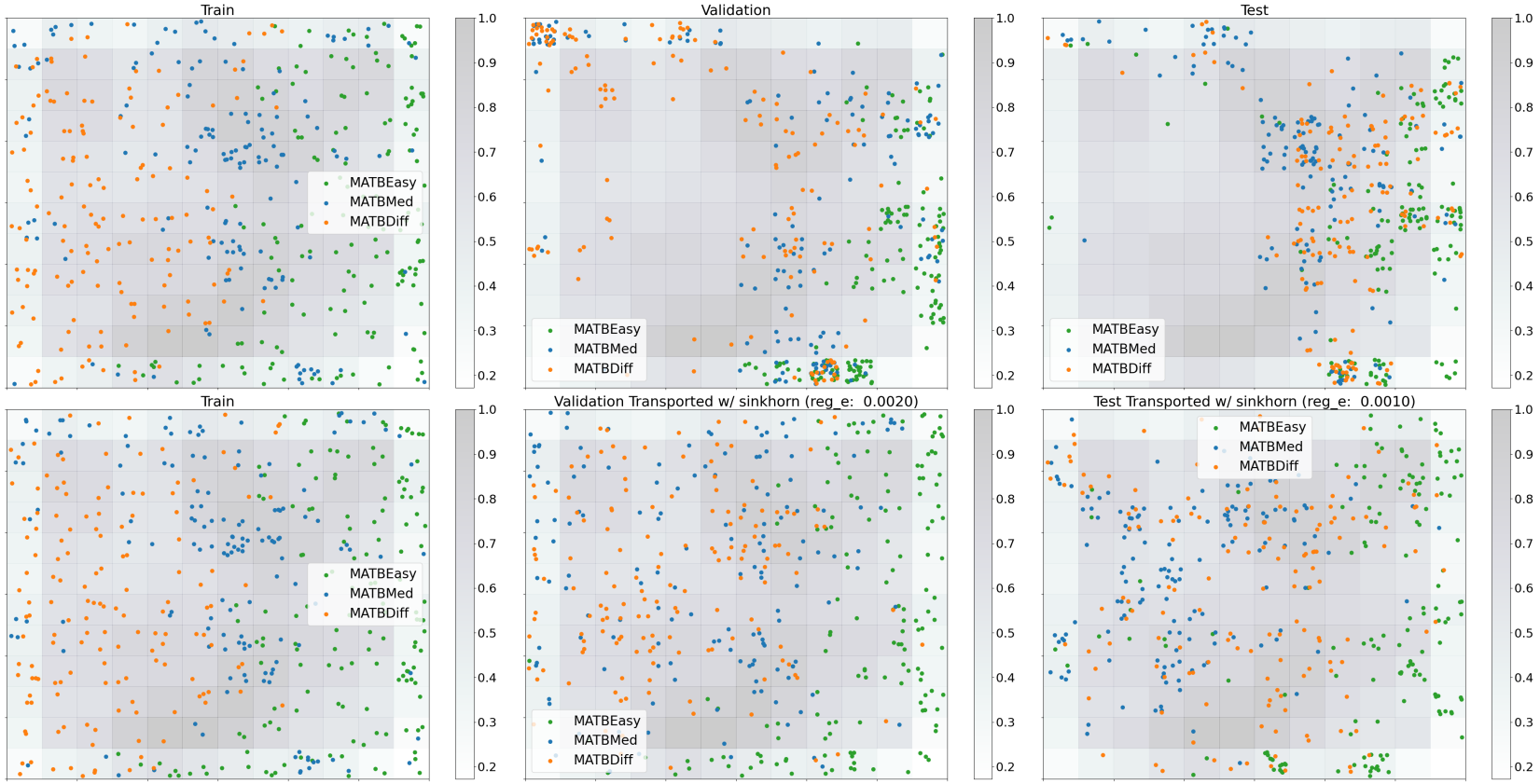


Figure 54: SOM Plots of Features (Subject 29)

Appendix C

MODEL RESULTS

This appendix presents detailed and subject-wise results of experiments.

C.1 Subject-Wise Results of COG_BCIg1 Variations

In this section, subject-wise results of COG_BCIg1 Variations are presented. Table 31 relates the models with the figures.

Table 31: COG_BCIg1 Variation Models and Figures

Model Name	Related Figure
COG_BCIg1_smallArch	Figure 55
COG_BCIg1_batchNorm	Figure 55
COG_BCIg1_frontal	Figure 56
COG_BCIg1_parietal	Figure 56
COG_BCIg1_theta_alpha	Figure 57
COG_BCIg1_2sec	Figure 57
COG_BCIg1_2sec_no	Figure 58
COG_BCIg1_4sec_no	Figure 58
COG_BCIg1_3nn	Figure 59
COG_BCIg1_5nn	Figure 59
COG_BCIg1_50p_3nn	Figure 60
COG_BCIg1_50p_5nn	Figure 60
COG_BCIg1_25p_5nn	Figure 61

C.2 Model Results for Each Subject

In this section, results of all models are presented for each subject in Figures 62 - 74.

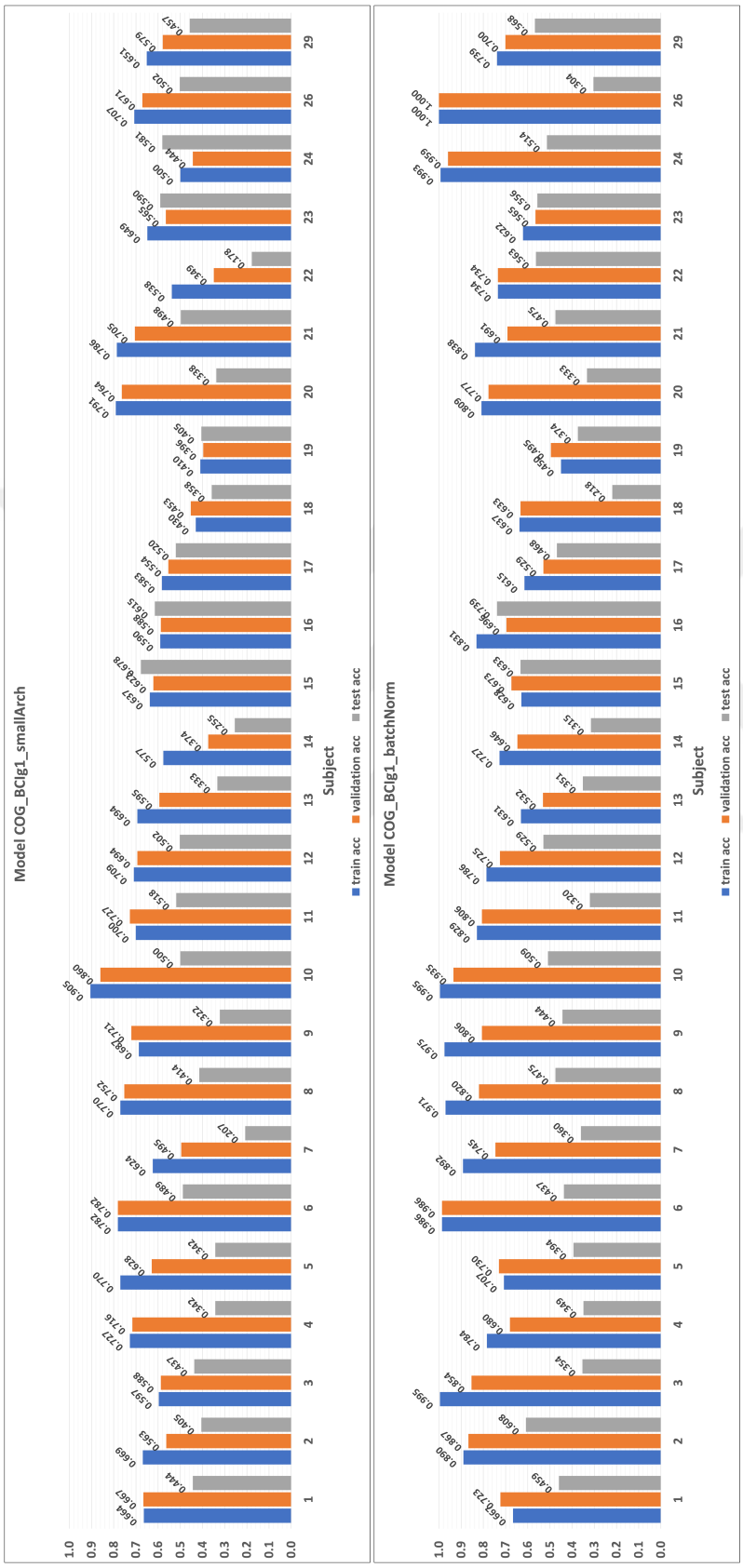


Figure 55: Subject-Wise Results for Models COG_BCIg1_smallArch and COG_BCIg1_batchNorm

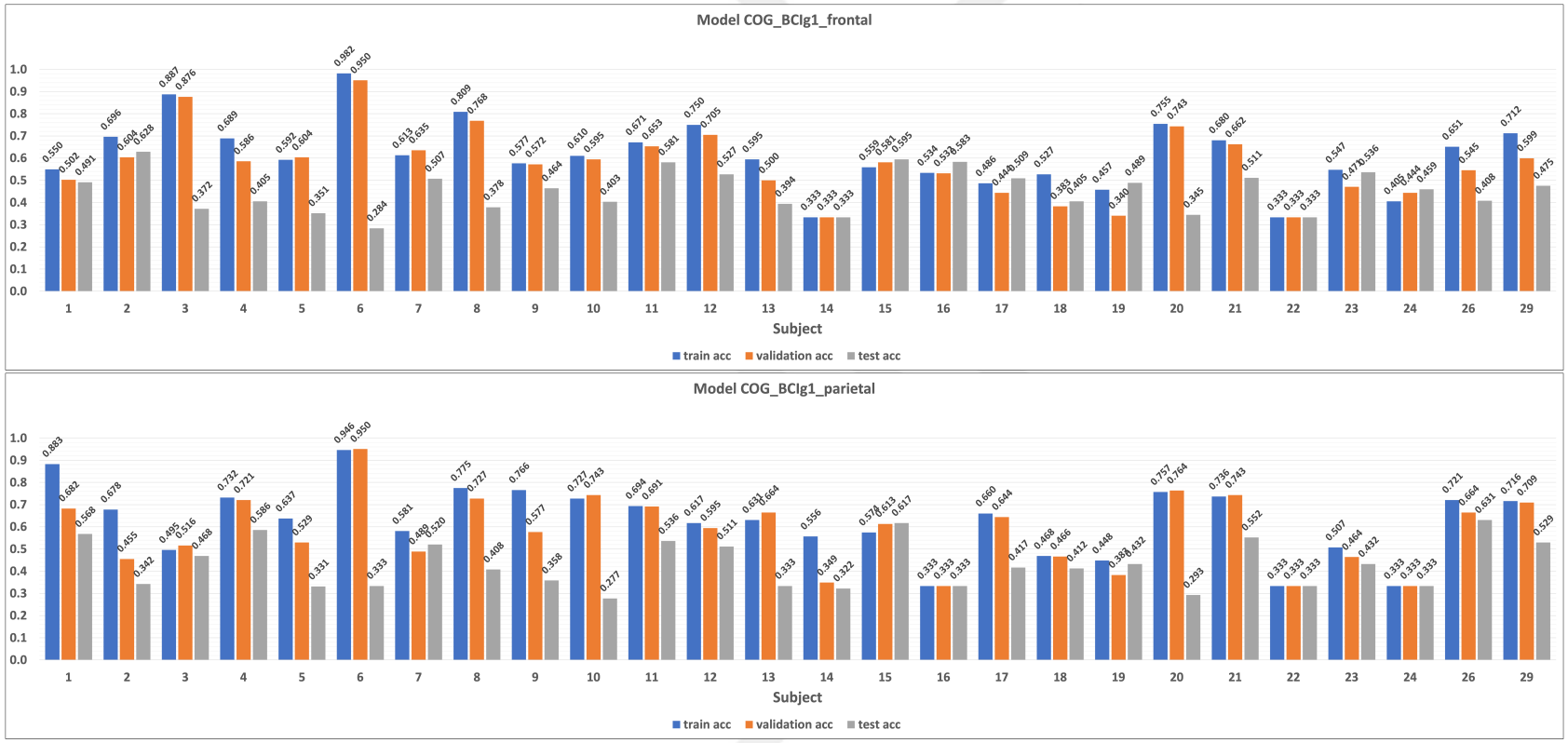


Figure 56: Subject-Wise Results for Models COG_BCIg1_frontal and COG_BCIg1_parietal

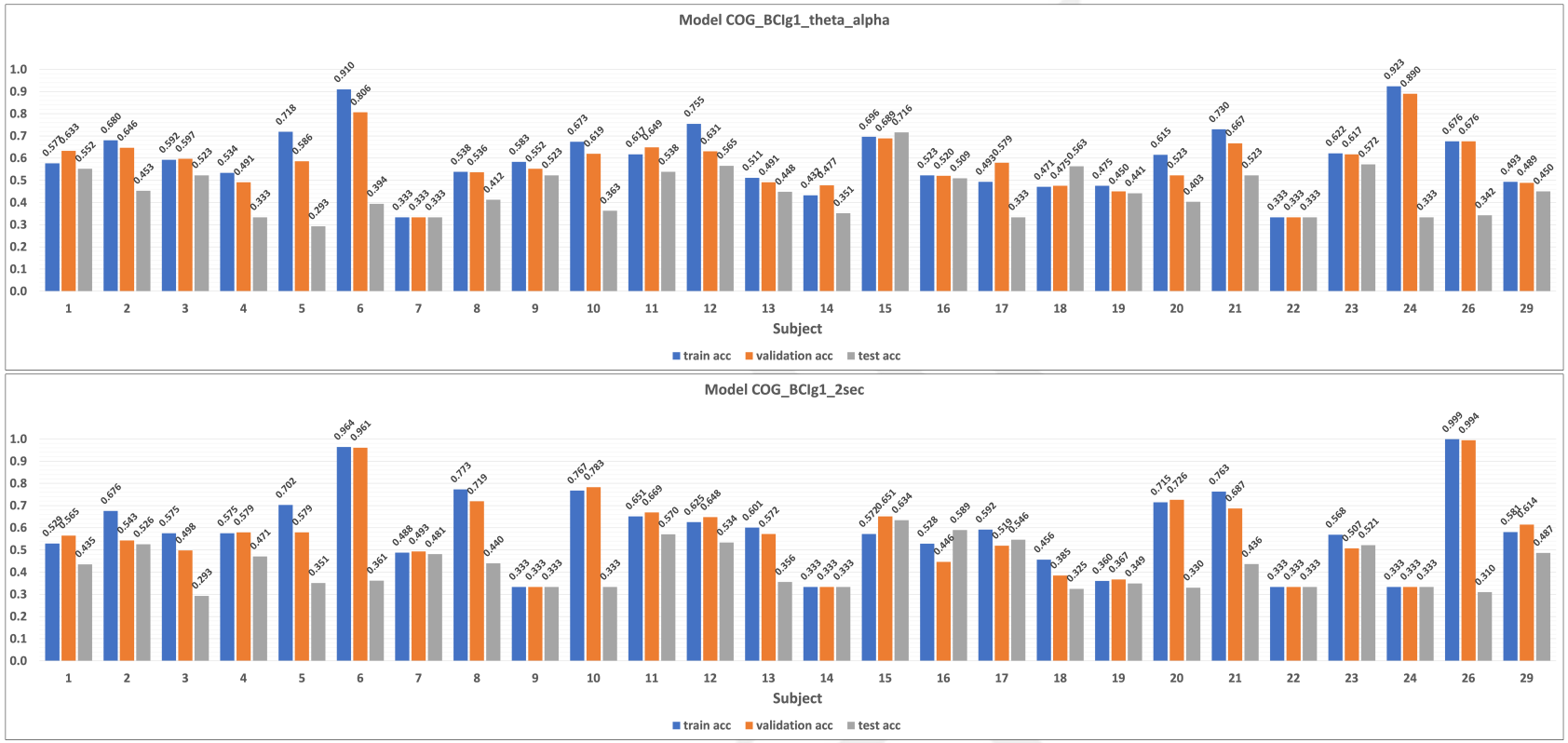


Figure 57: Subject-Wise Results for Models COG_BCIg1_theta_alpha and COG_BCIg1_2sec

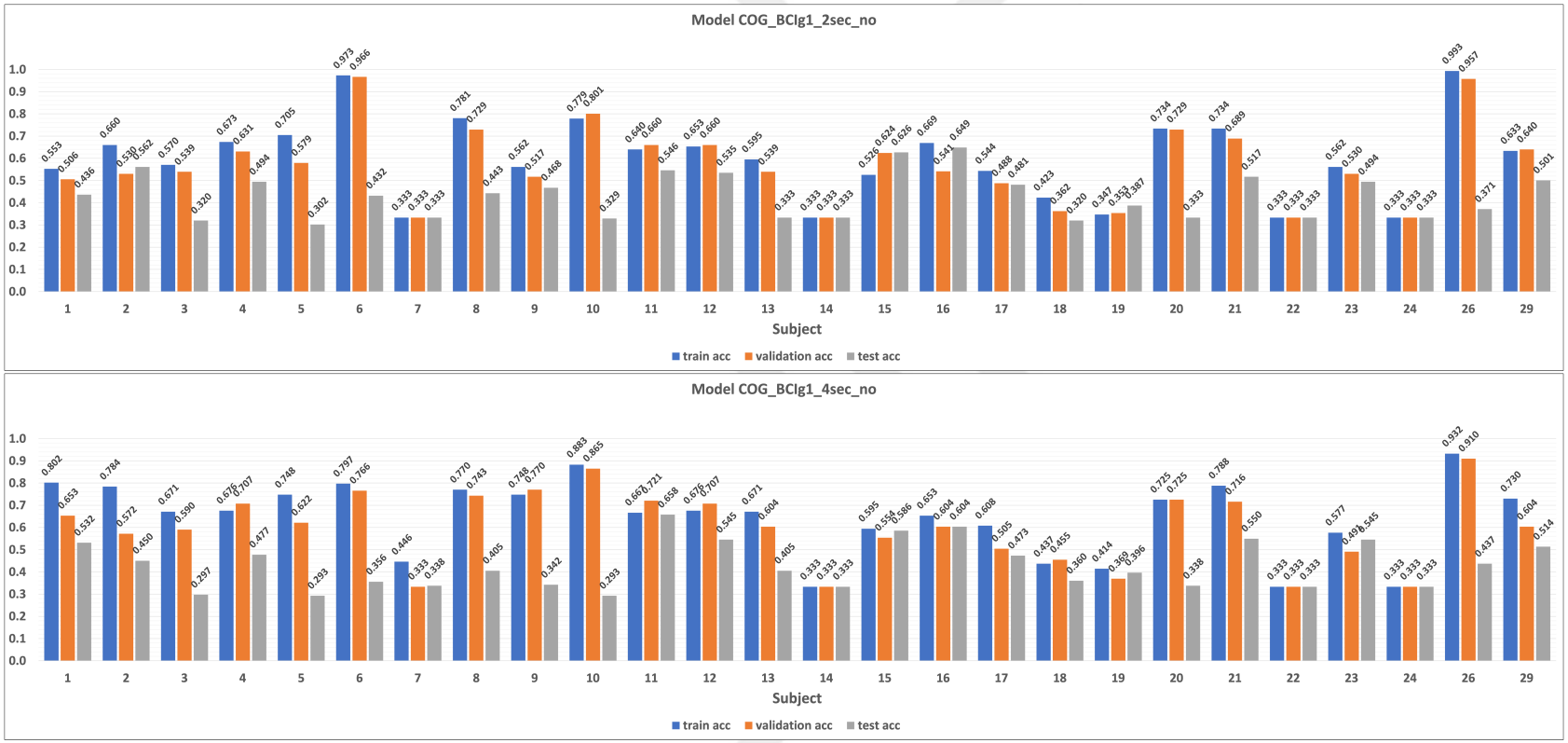


Figure 58: Subject-Wise Results for Models COG_BCIg1_2sec_no and COG_BCIg1_4sec_no

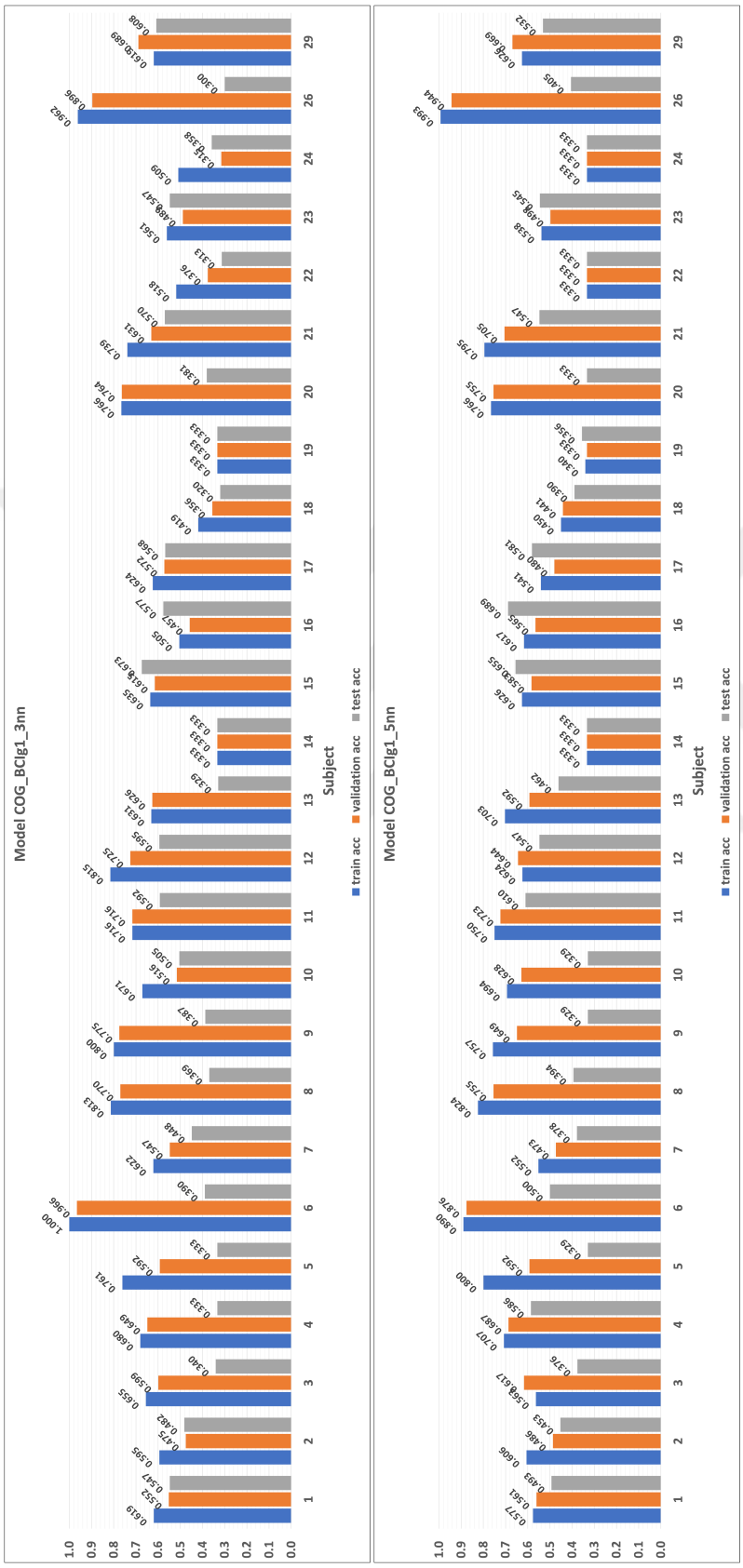


Figure 59: Subject-Wise Results for Models COG_BCIg1_3nn and COG_BCIg1_5nn

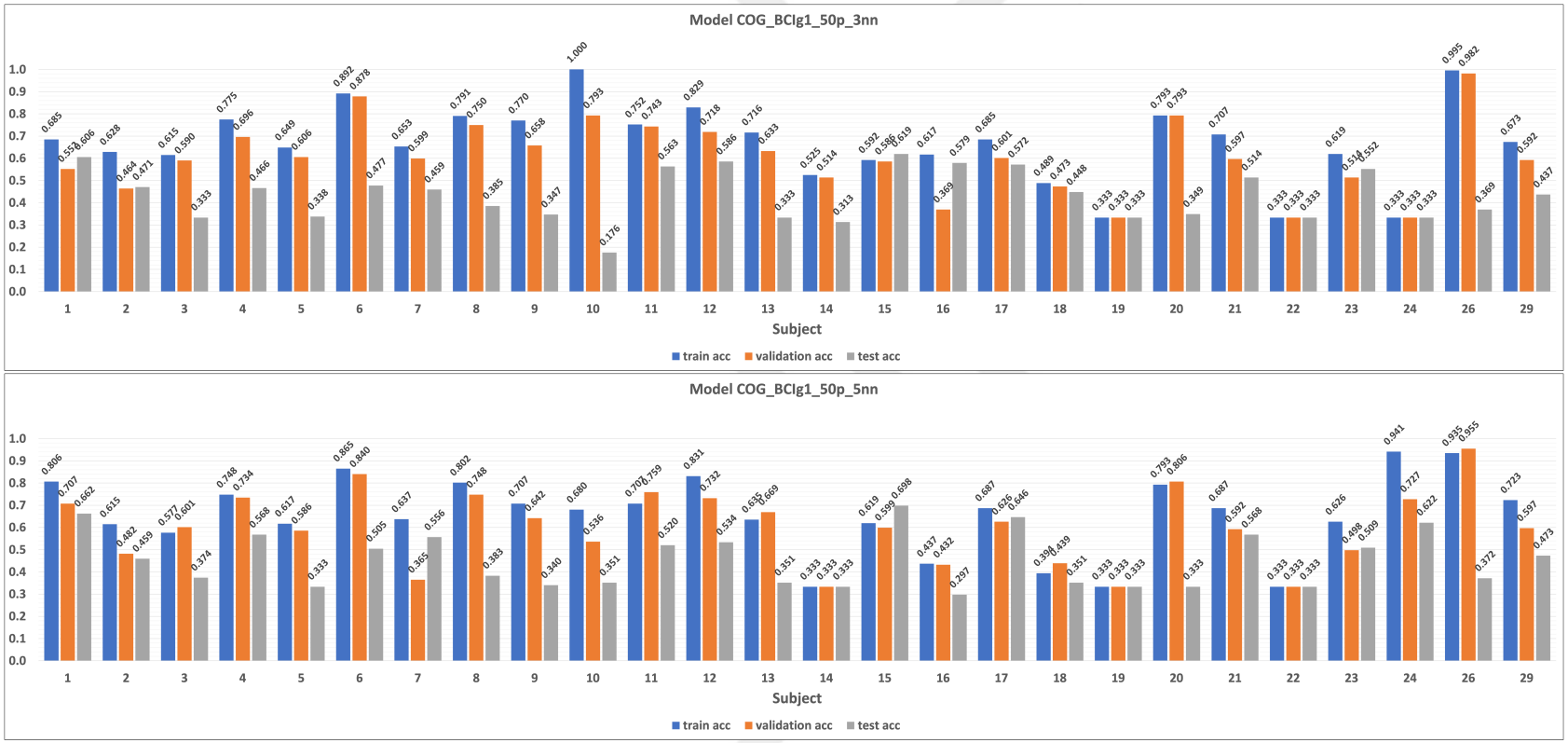


Figure 60: Subject-Wise Results for Models COG_BCIg1_50p_3nn and COG_BCIg1_50p_5nn

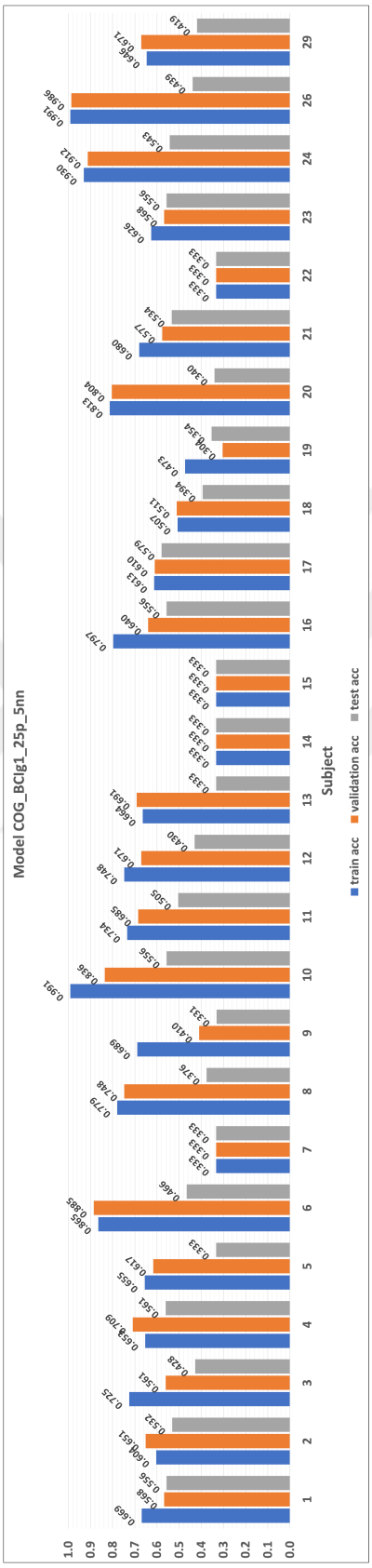


Figure 61: Subject-Wise Results for Models COG_BCIg1_25p_5nn

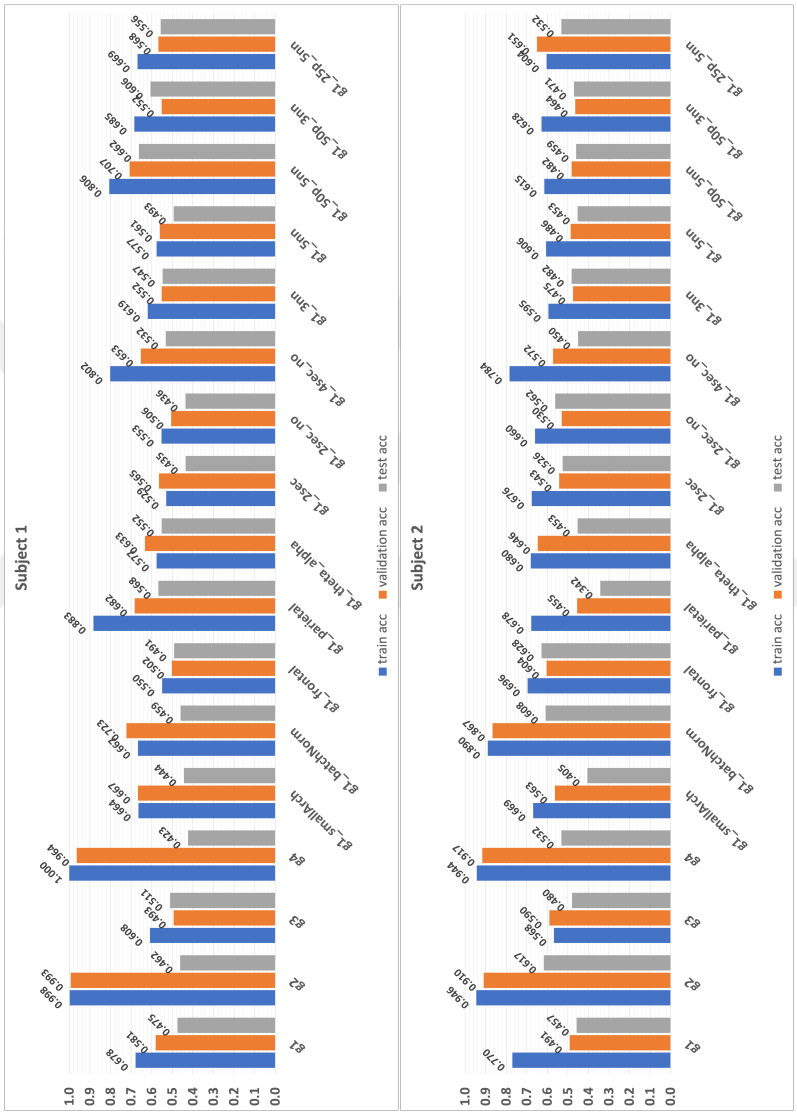


Figure 62: Model Results for Subjects 1 and 2

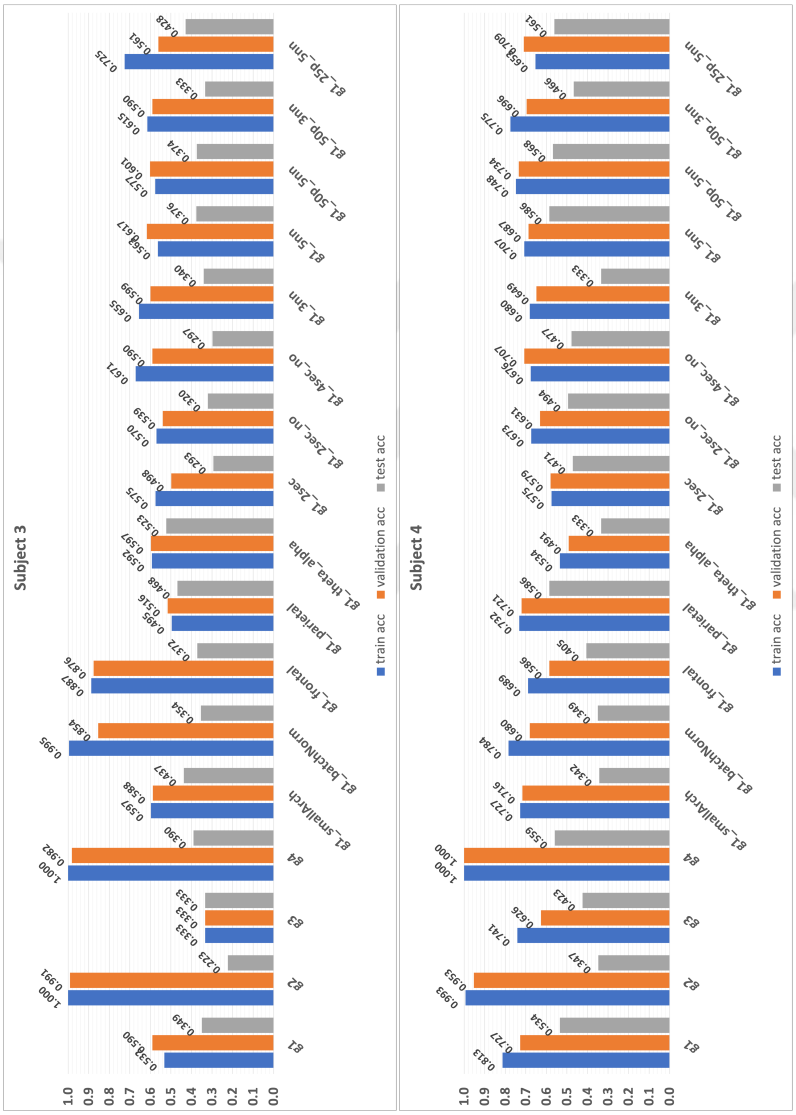


Figure 63: Model Results for Subjects 3 and 4

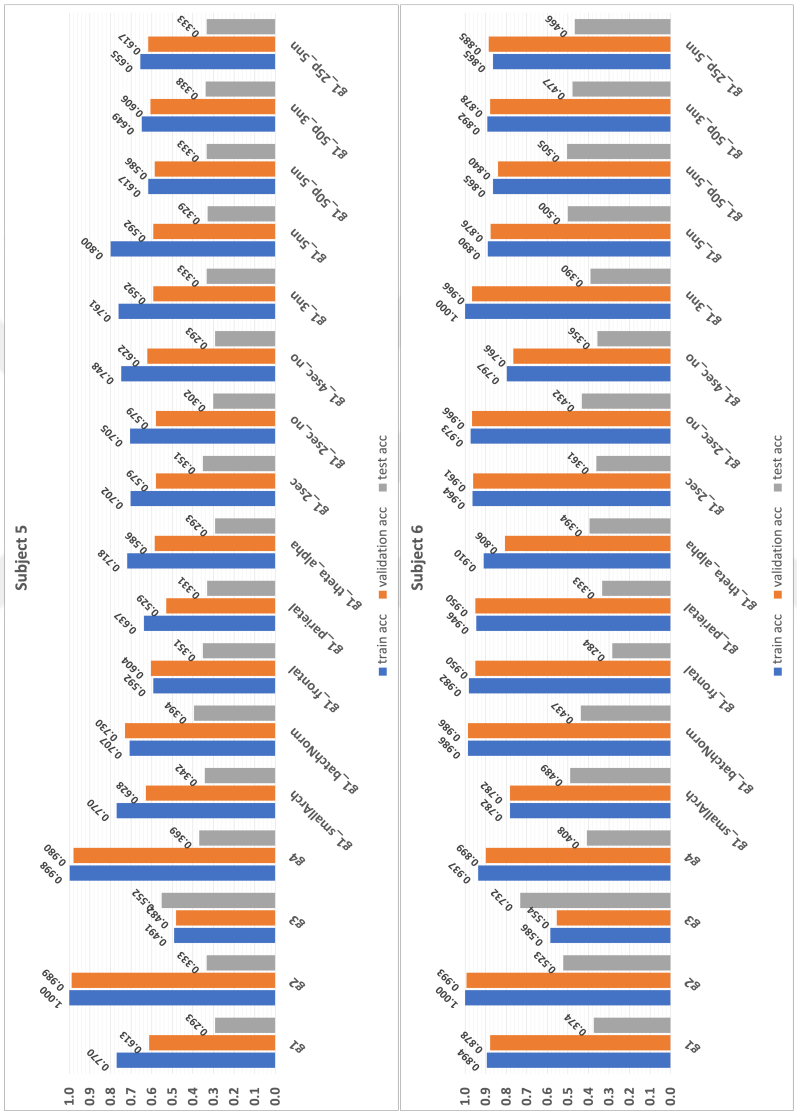


Figure 64: Model Results for Subjects 5 and 6

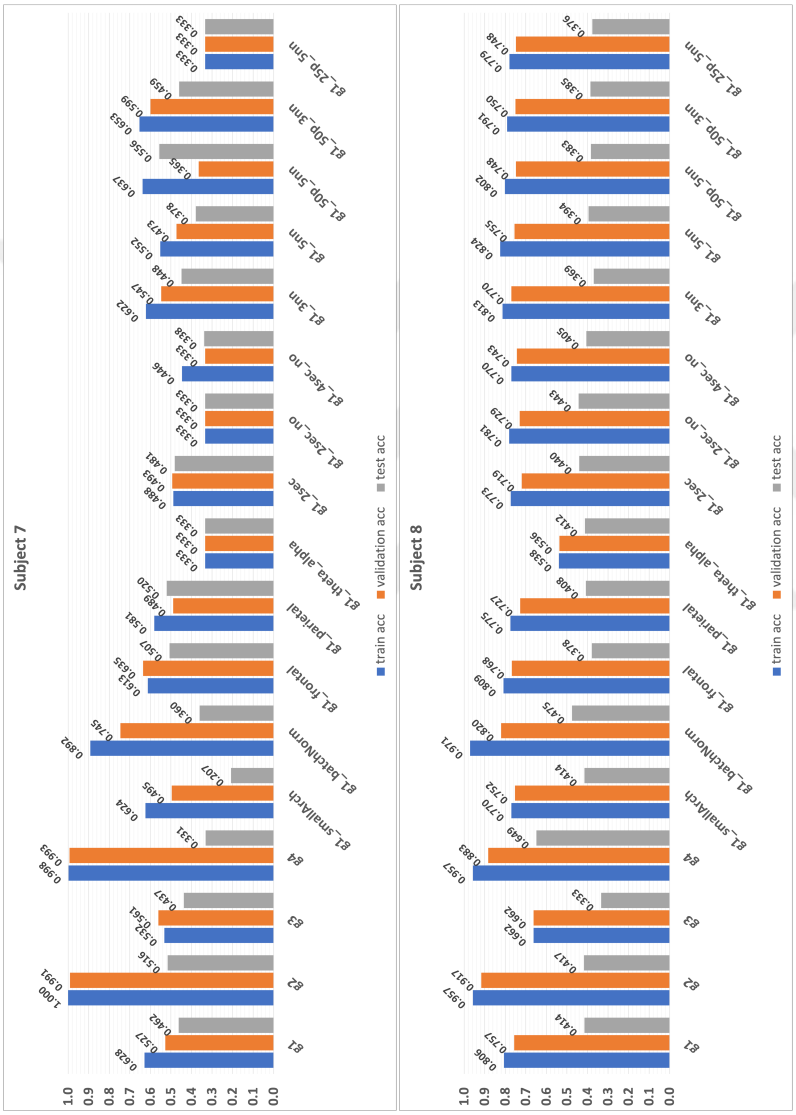


Figure 65: Model Results for Subjects 7 and 8

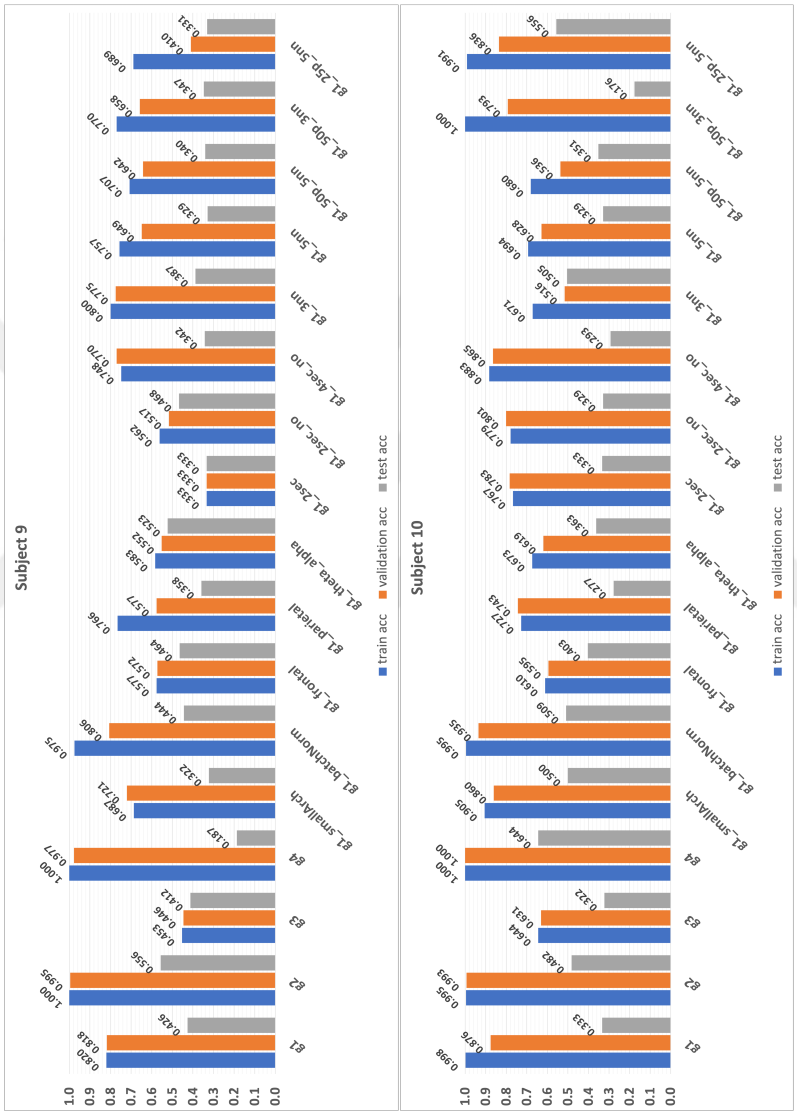


Figure 66: Model Results for Subjects 9 and 10

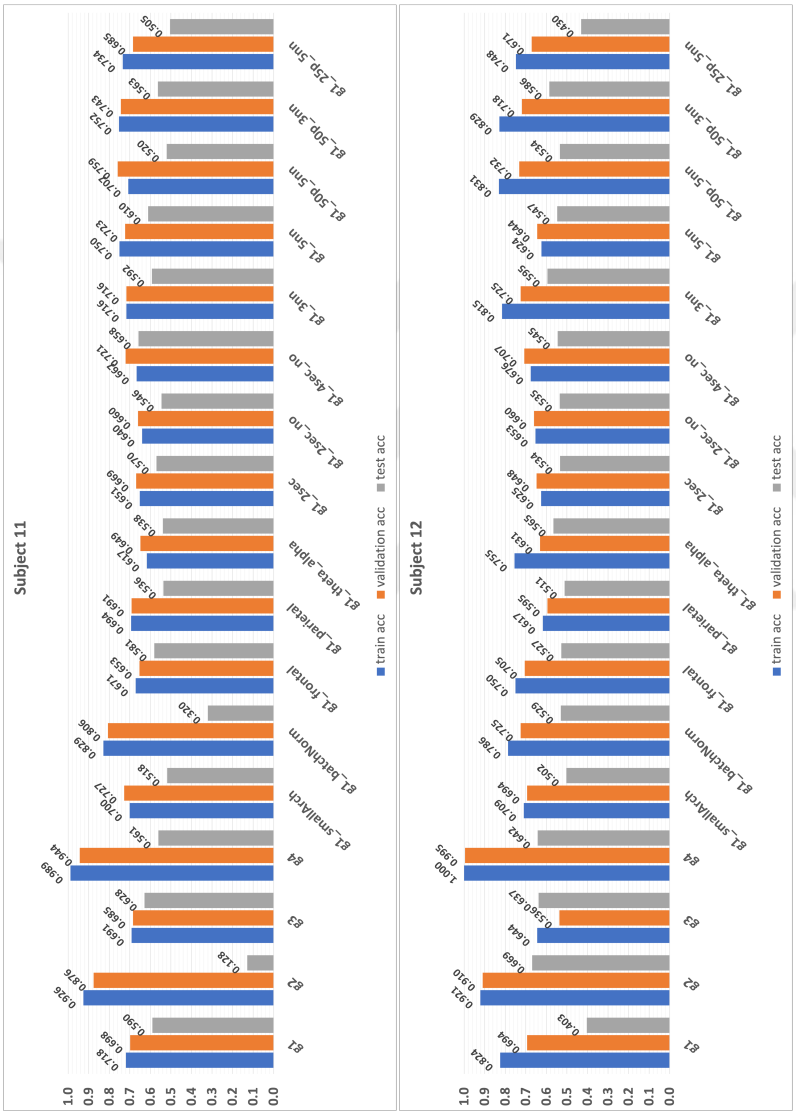


Figure 67: Model Results for Subjects 11 and 12

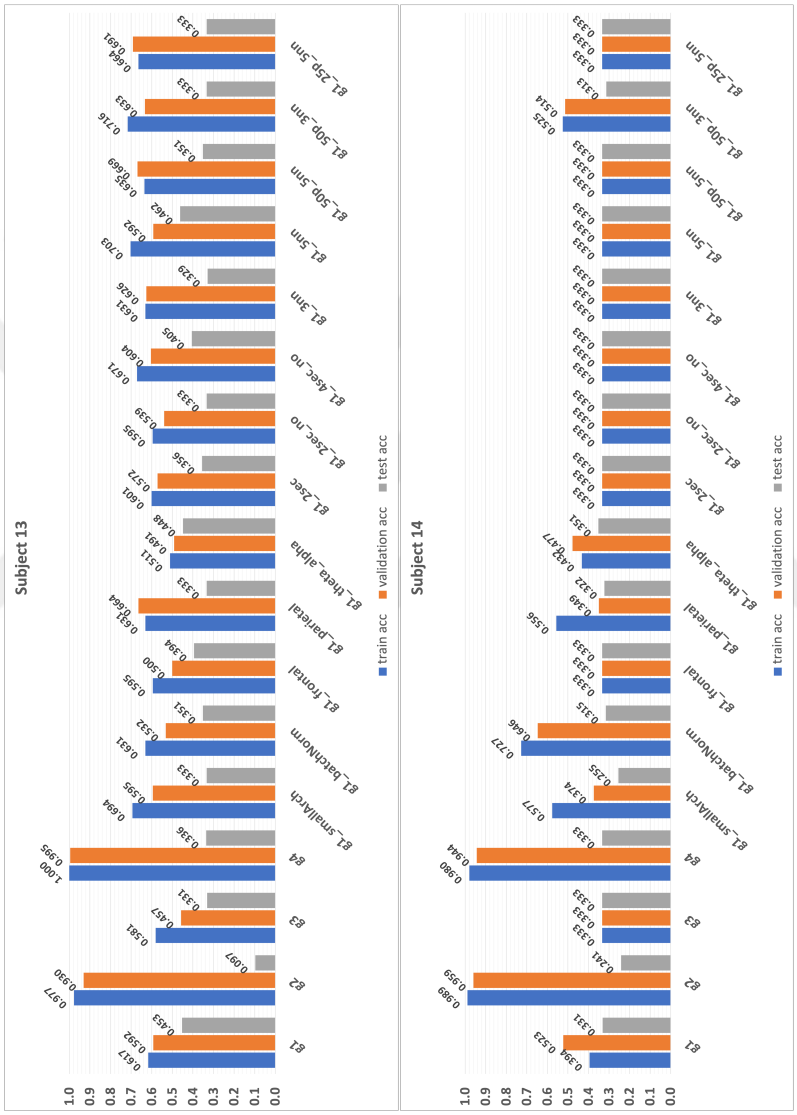


Figure 68: Model Results for Subjects 13 and 14

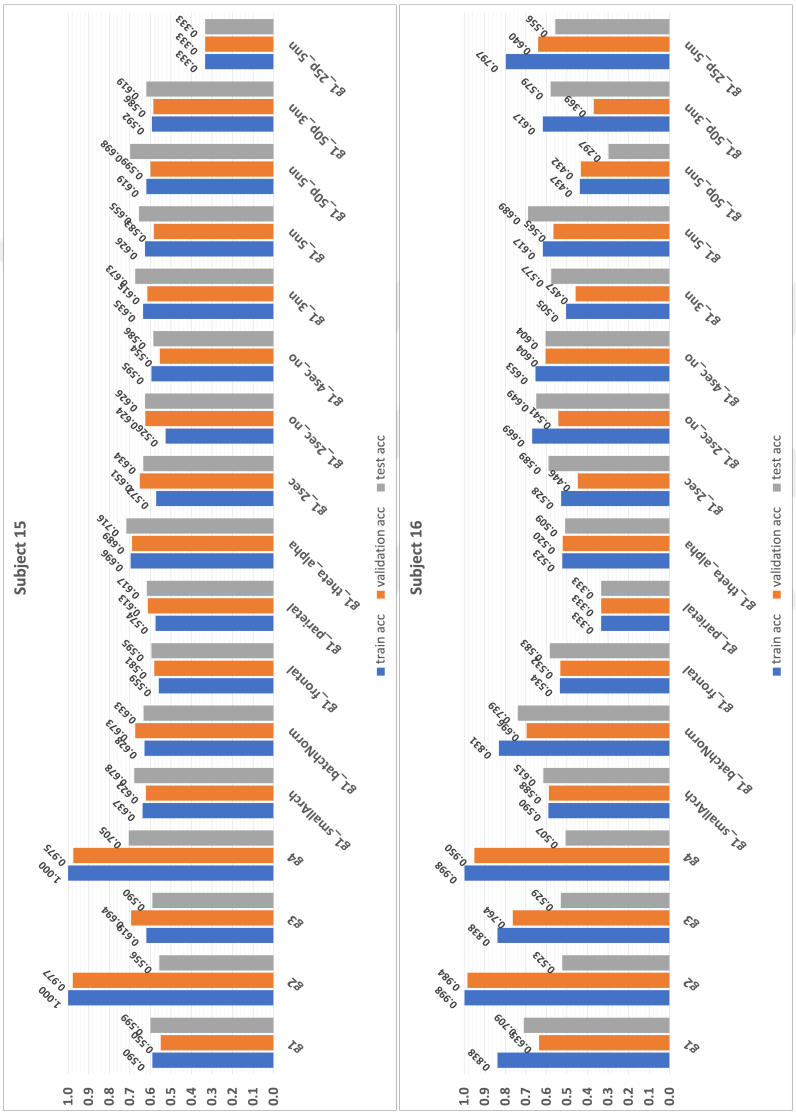


Figure 69: Model Results for Subjects 15 and 16

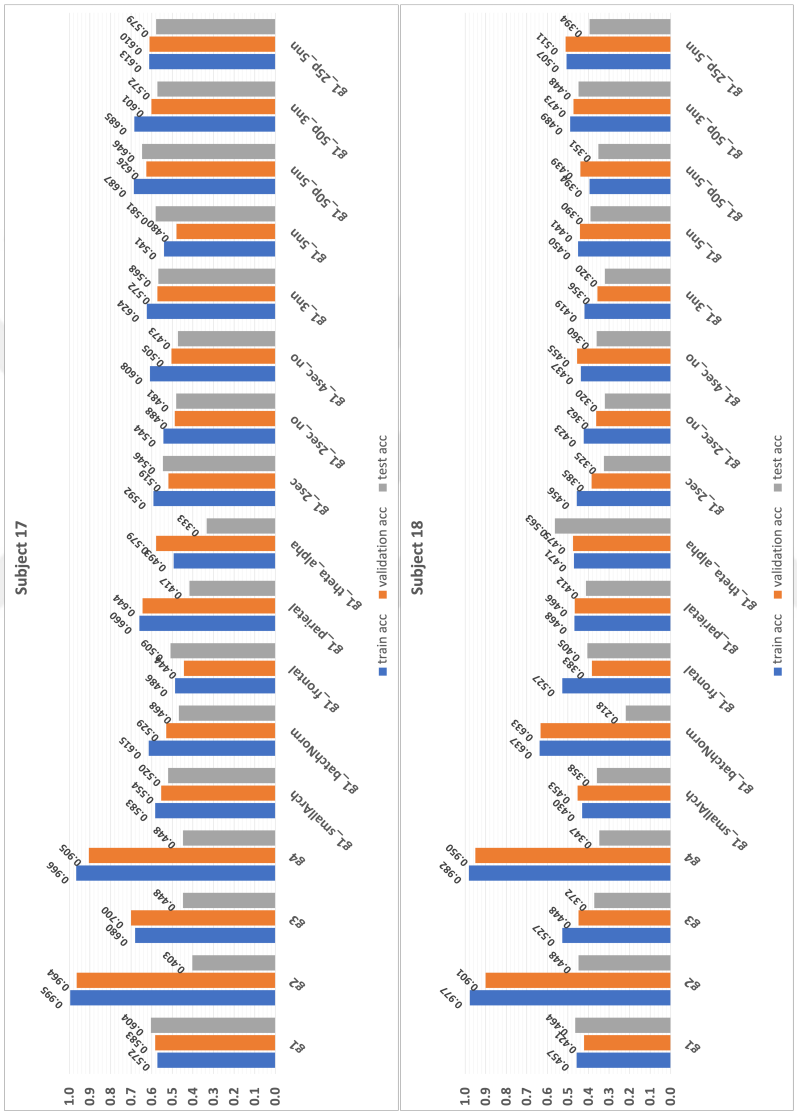


Figure 70: Model Results for Subjects 17 and 18

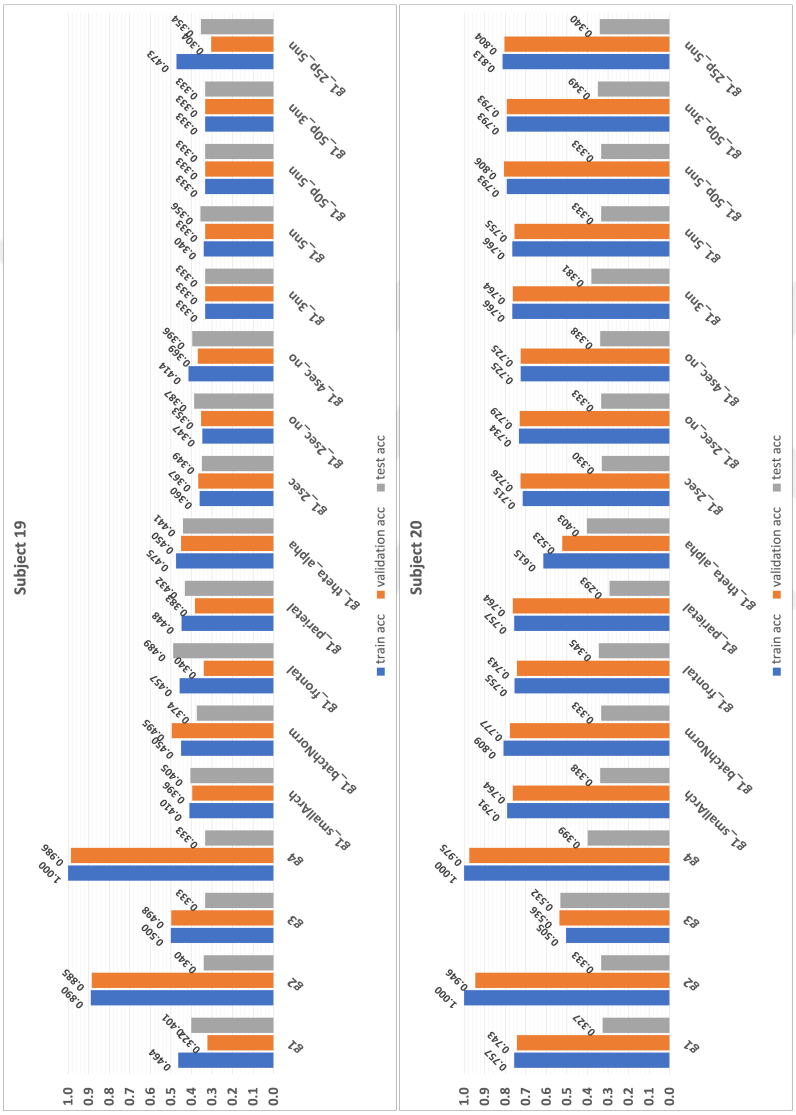


Figure 71: Model Results for Subjects 19 and 20

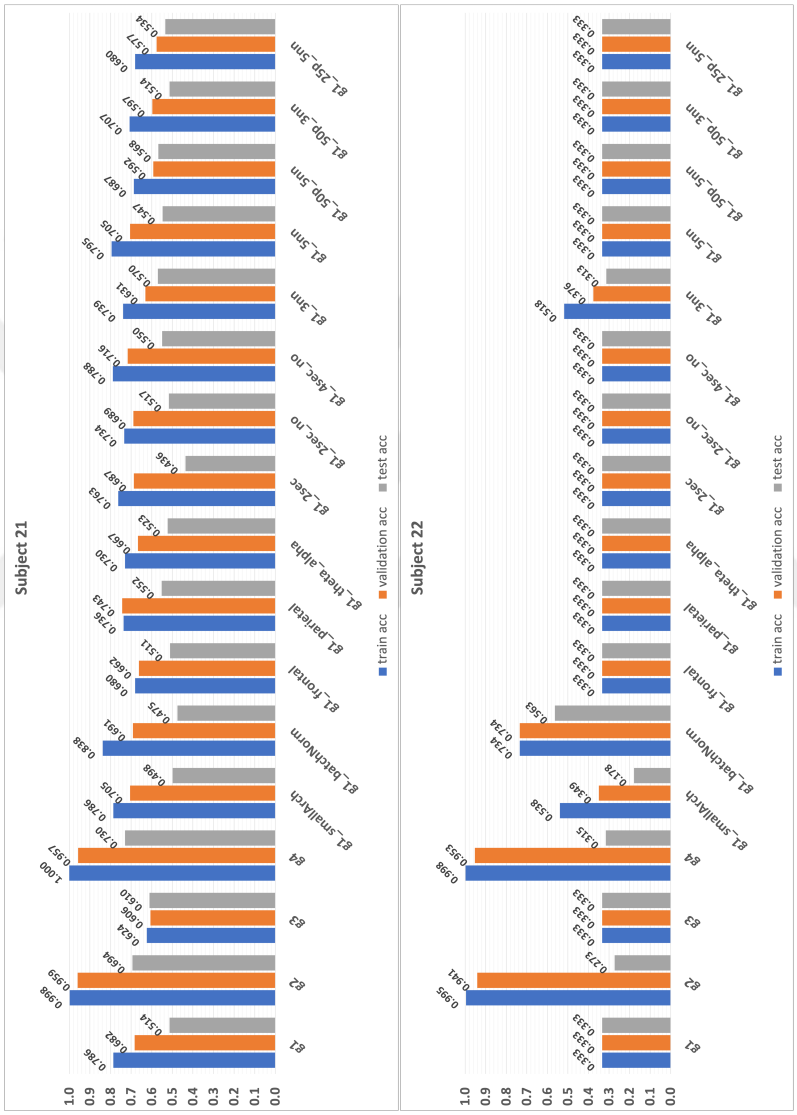


Figure 72: Model Results for Subjects 21 and 22

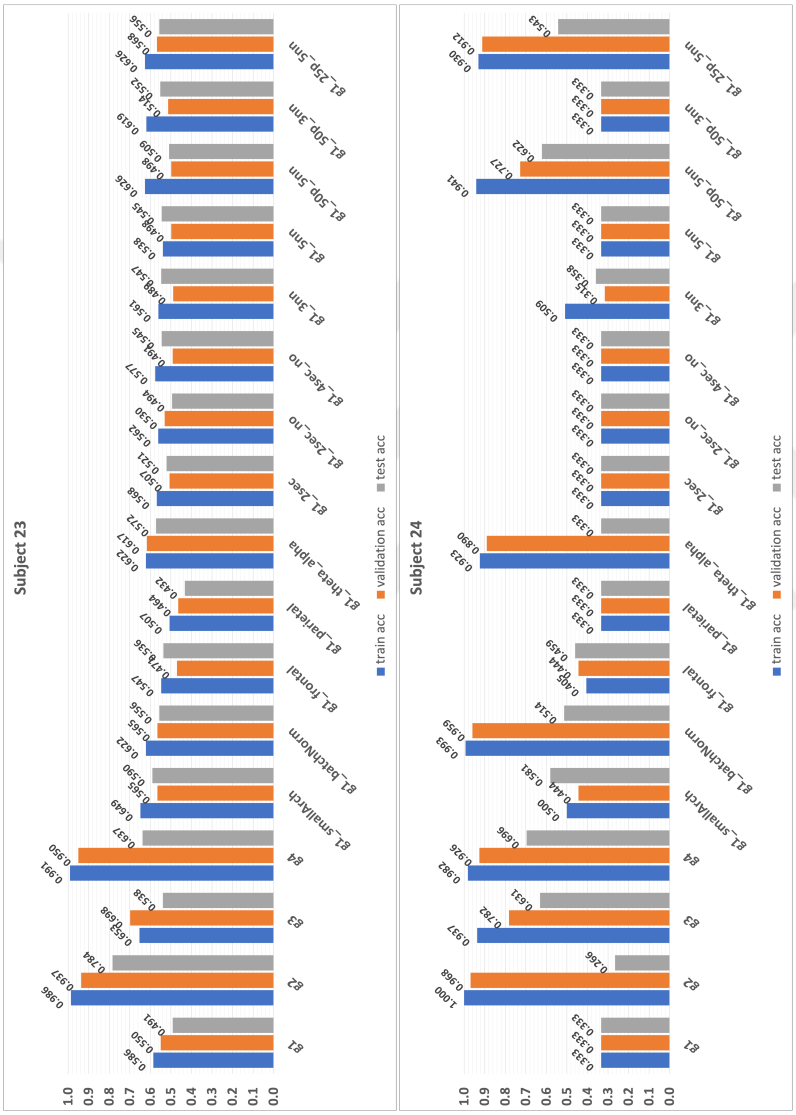


Figure 73: Model Results for Subjects 23 and 24

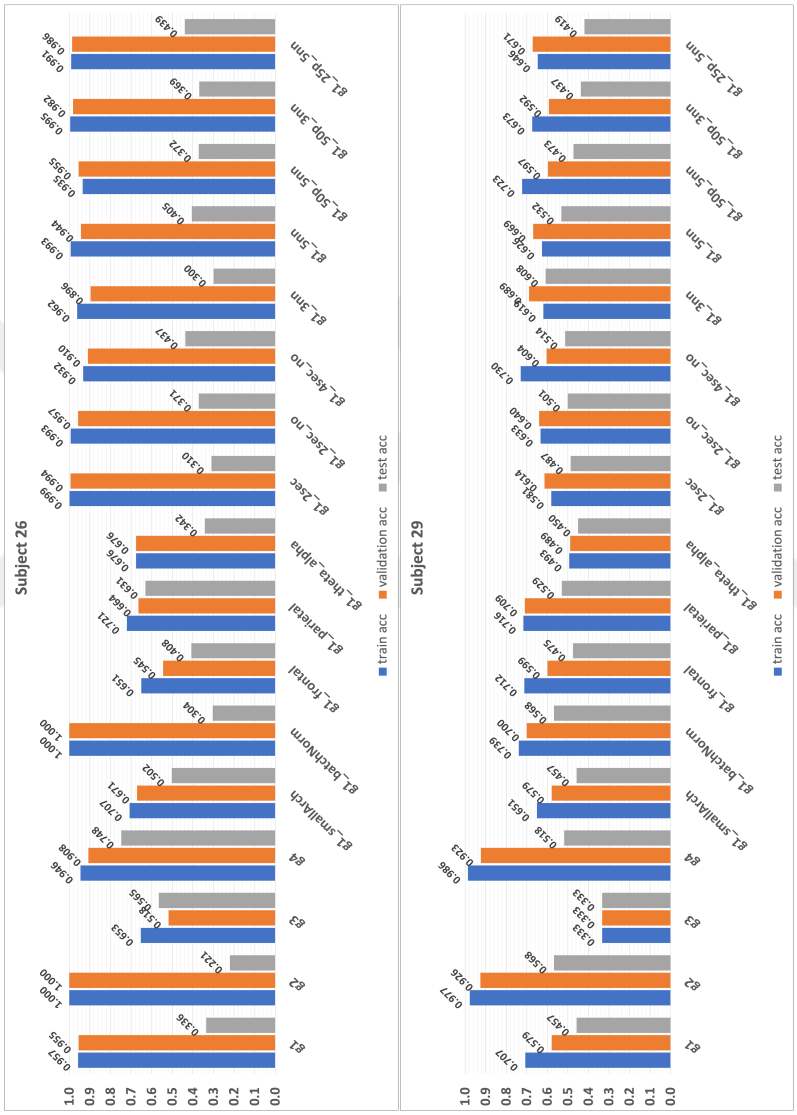


Figure 74: Model Results for Subjects 26 and 29

C.3 Model Complexity for Each Subject

In this section, joint plots of COG_BCIg1 and COG_BCIg1_smallArch experiments are illustrated for all subjects. Details on these plots are explained in 4.3.4.3.

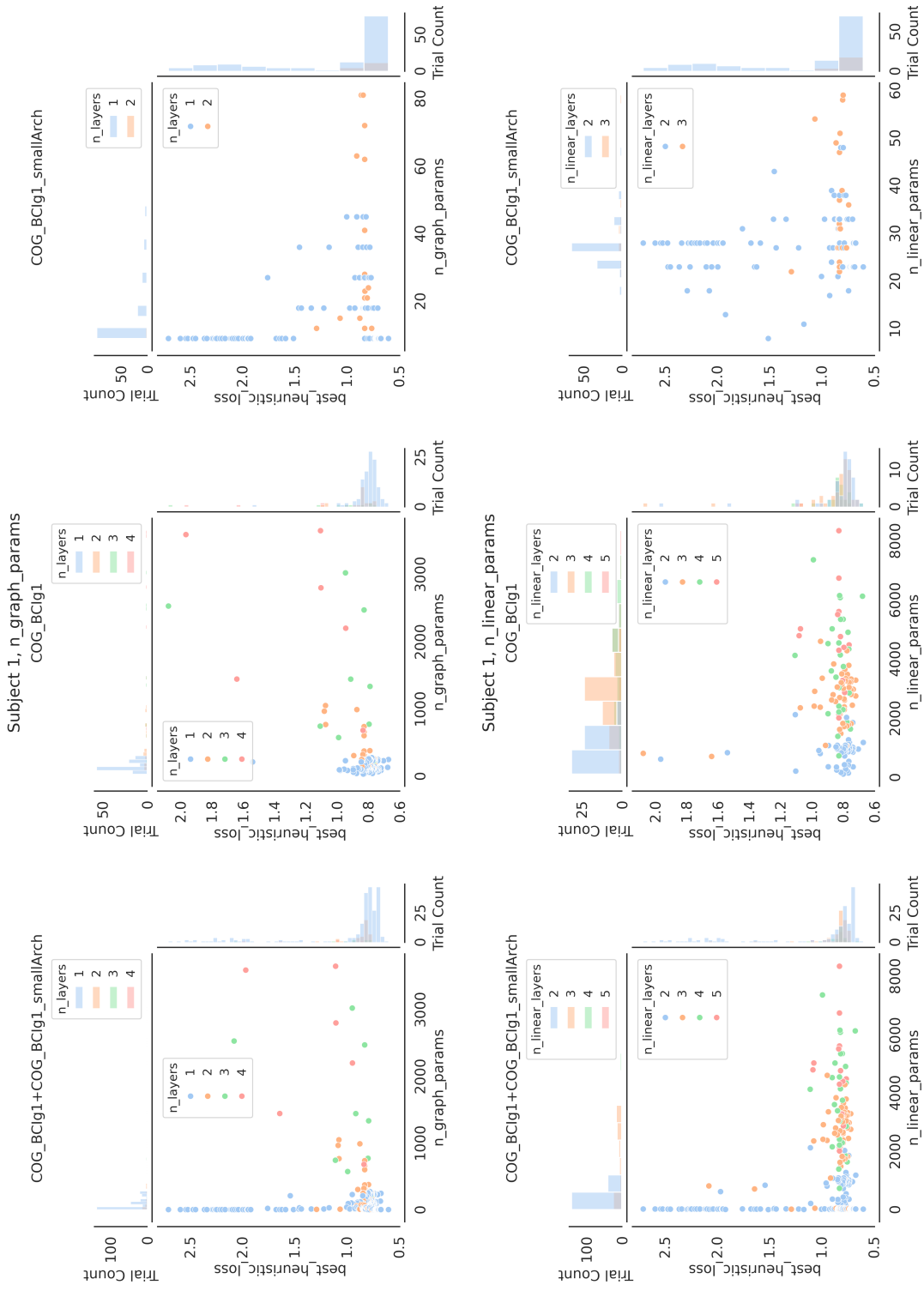


Figure 75: Best Heuristic Loss vs Model Complexity - Subject 1

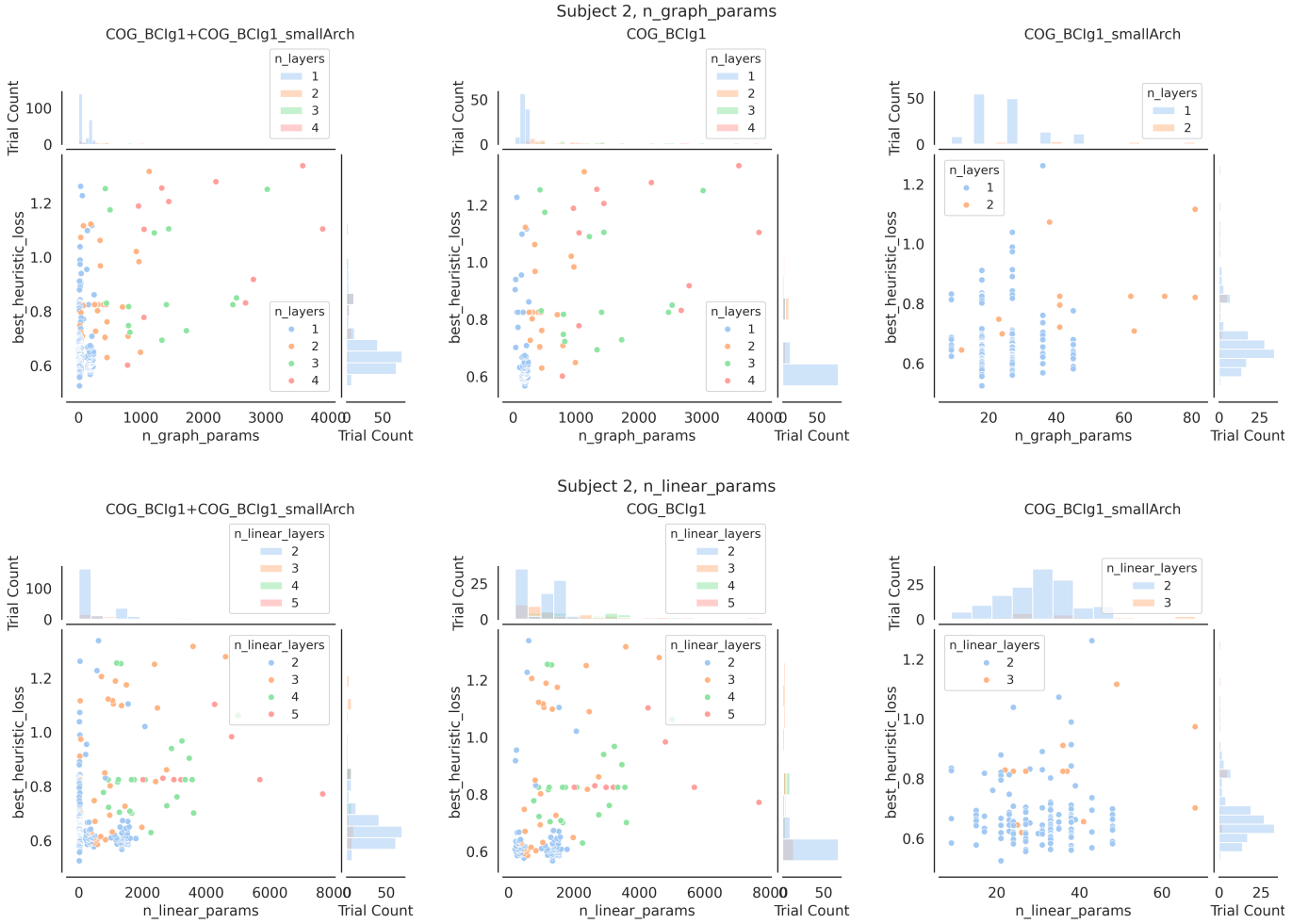


Figure 76: Best Heuristic Loss vs Model Complexity - Subject 2

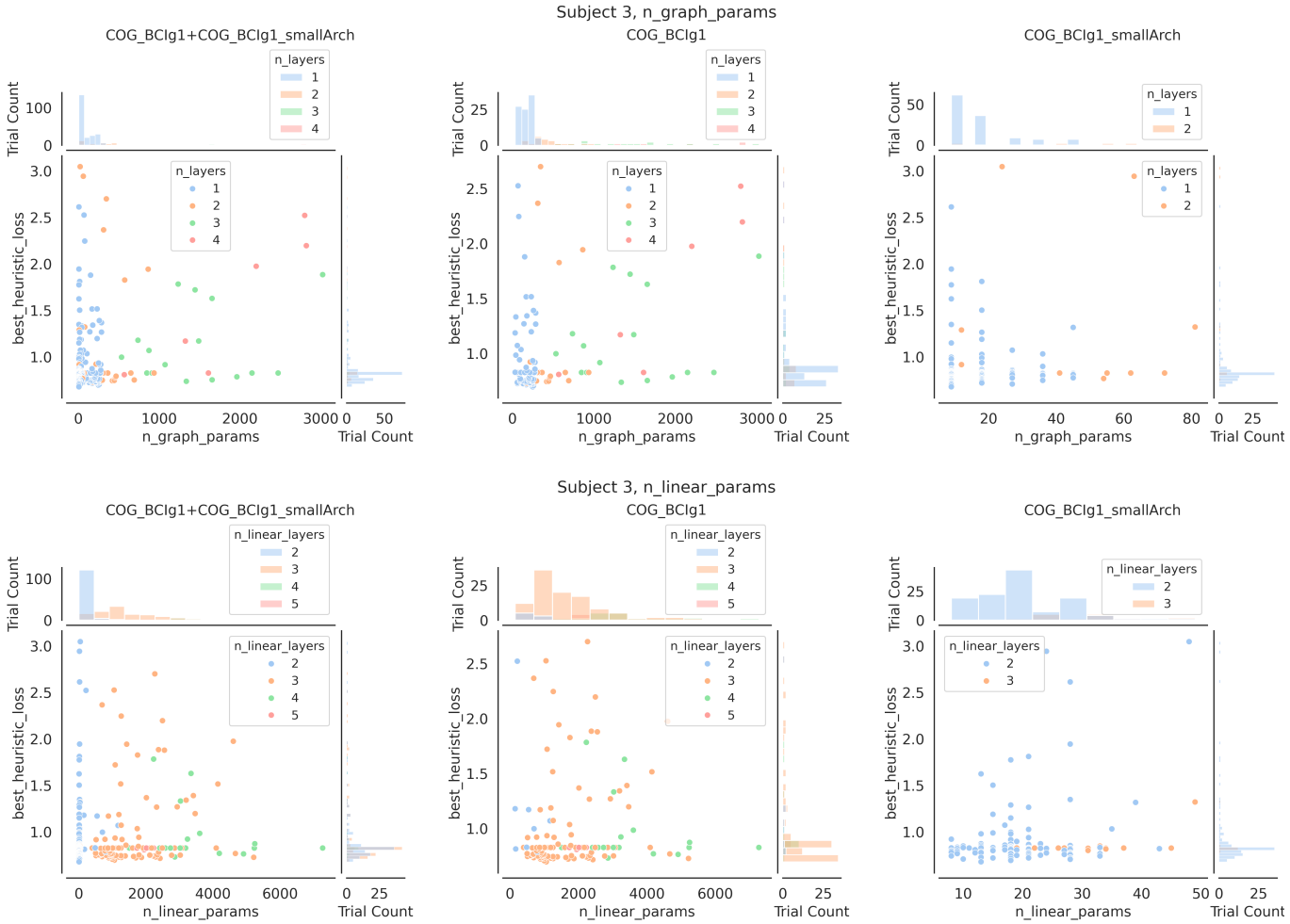


Figure 77: Best Heuristic Loss vs Model Complexity - Subject 3

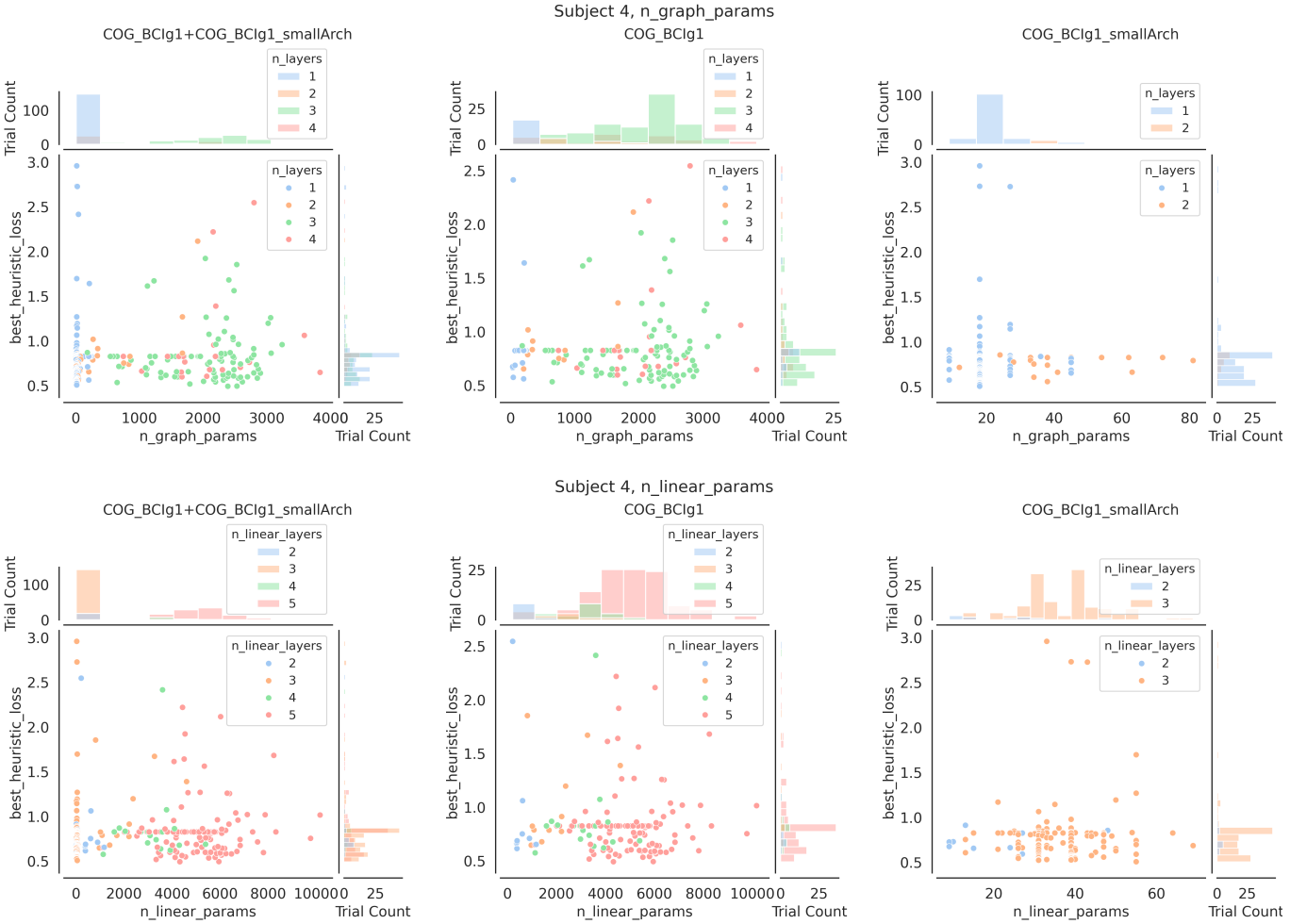


Figure 78: Best Heuristic Loss vs Model Complexity - Subject 4

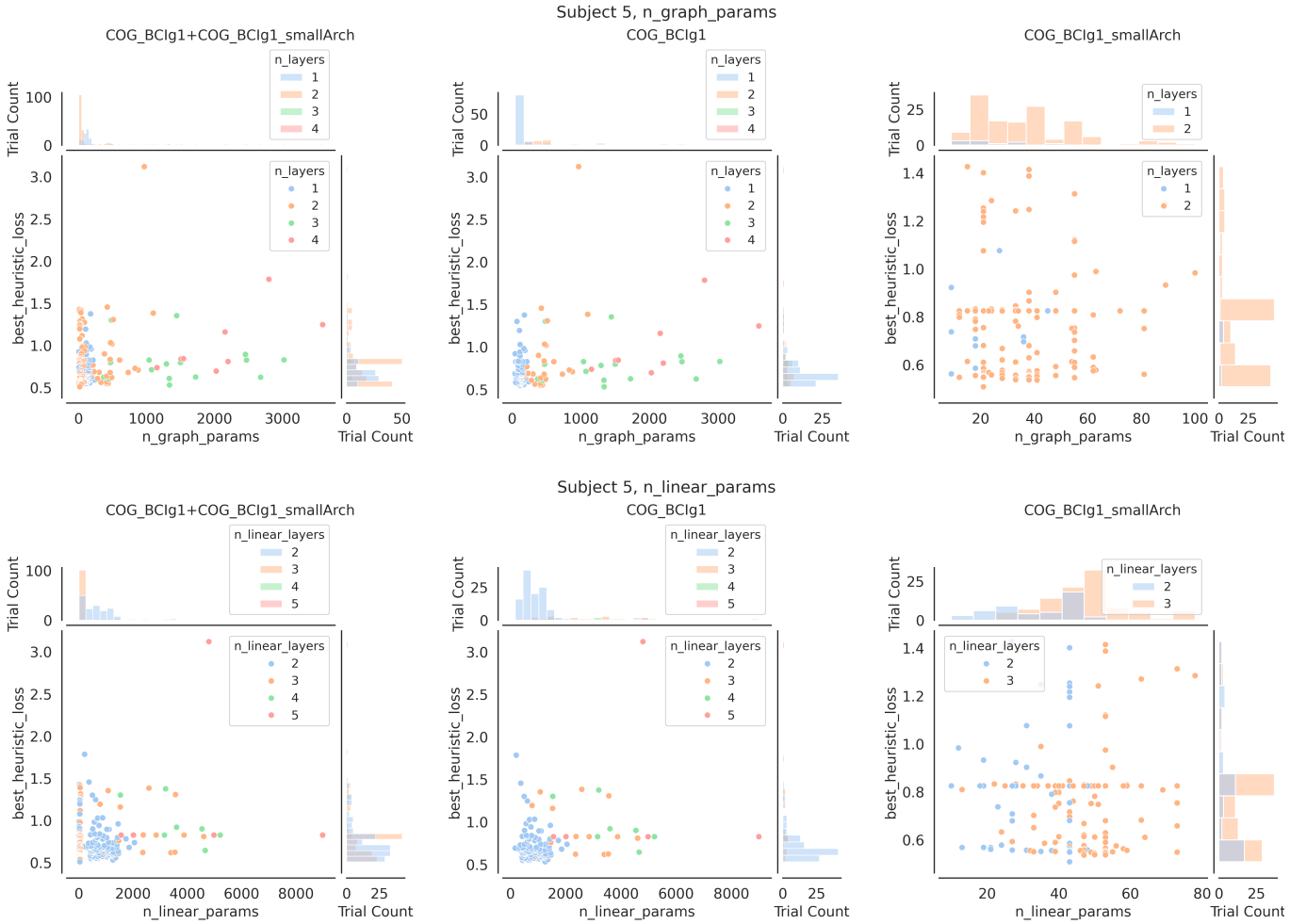


Figure 79: Best Heuristic Loss vs Model Complexity - Subject 5

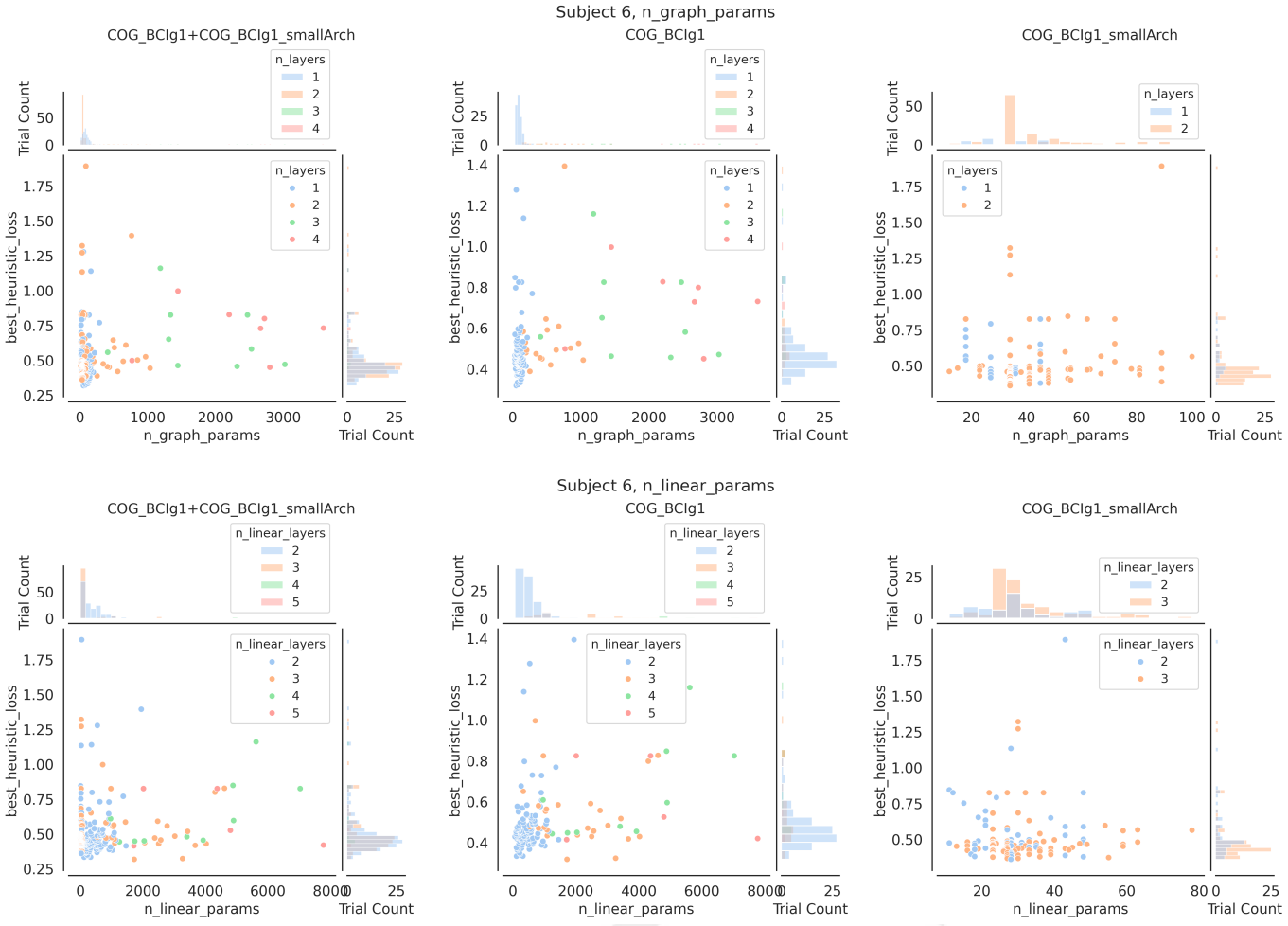


Figure 80: Best Heuristic Loss vs Model Complexity - Subject 6

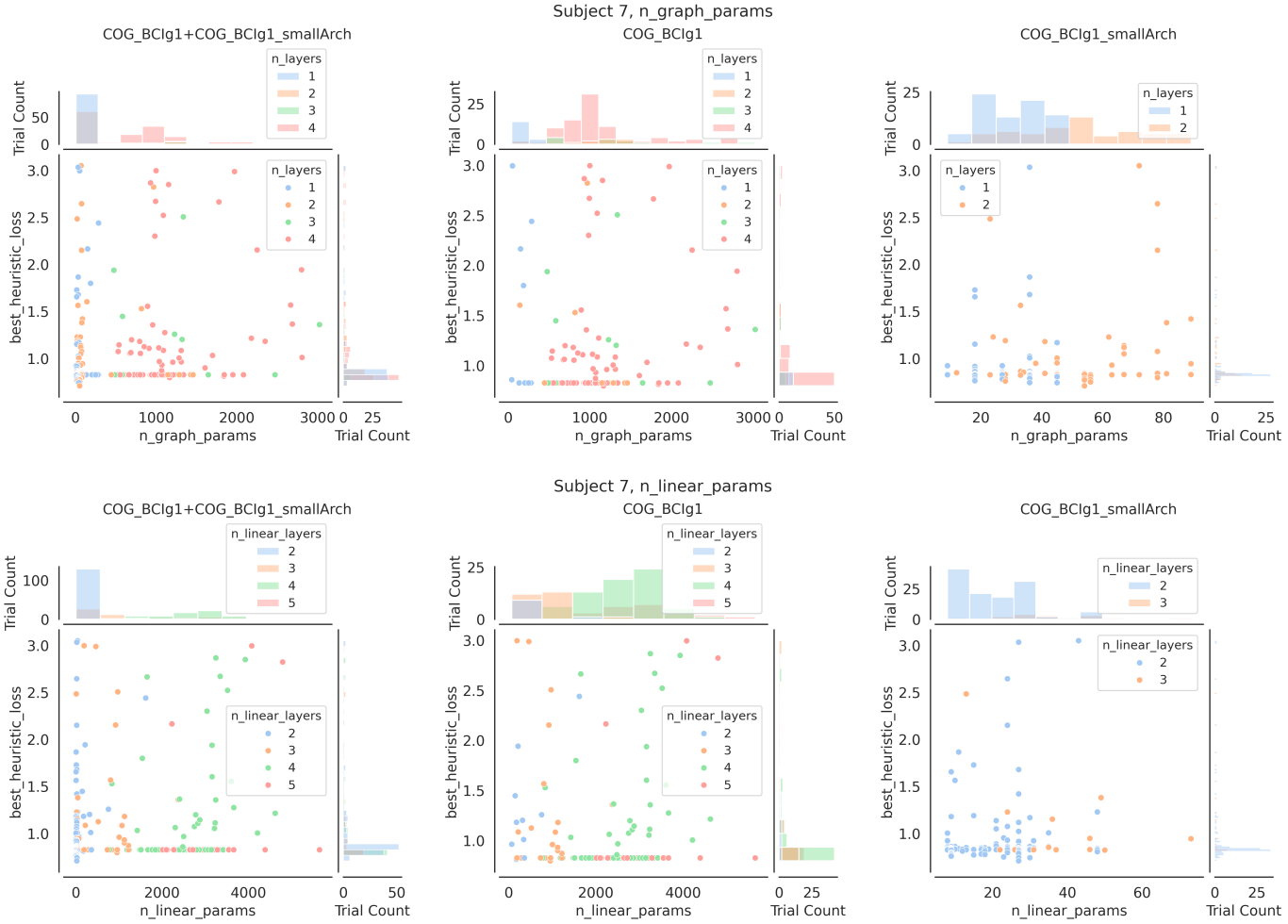


Figure 81: Best Heuristic Loss vs Model Complexity - Subject 7

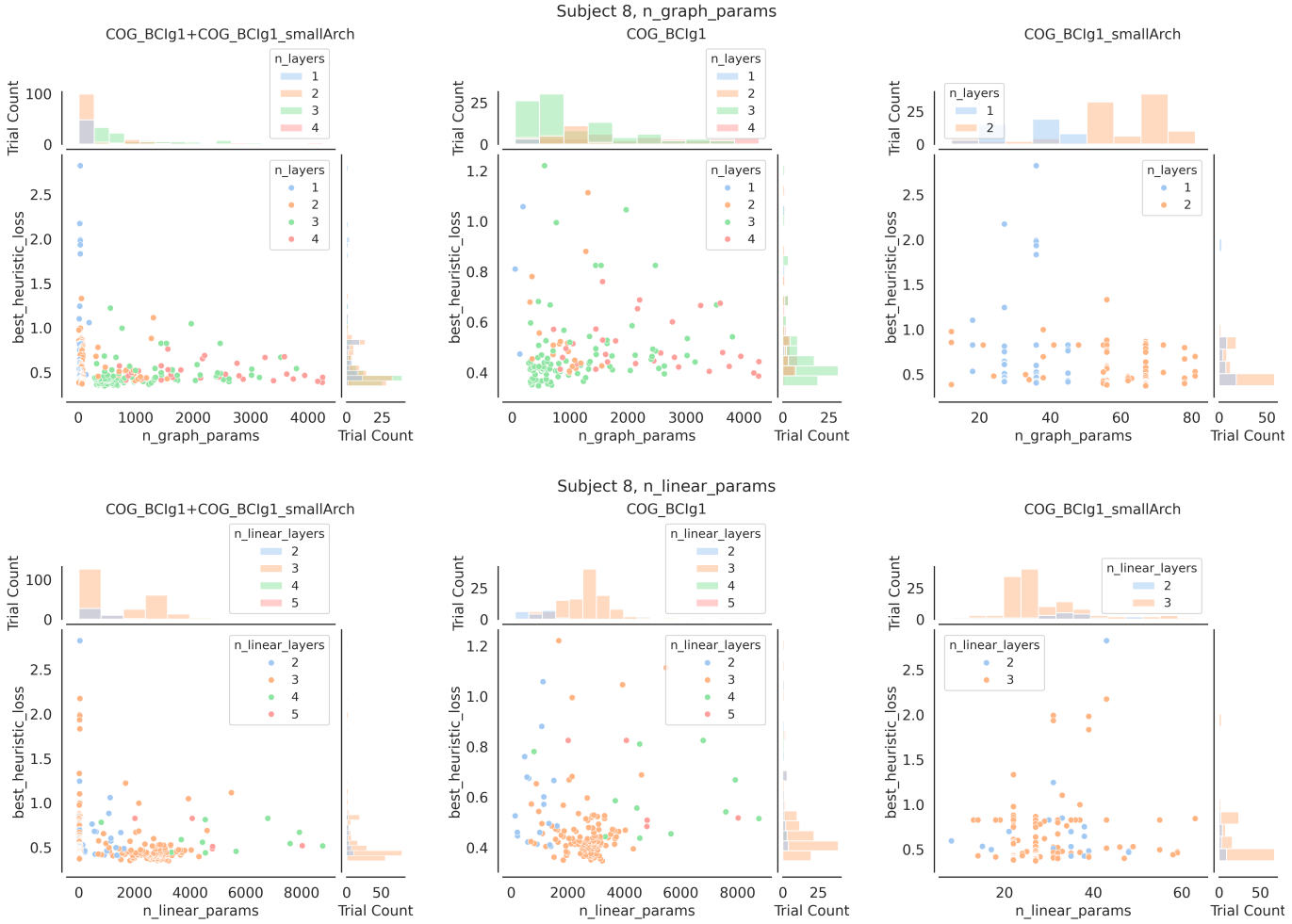


Figure 82: Best Heuristic Loss vs Model Complexity - Subject 8

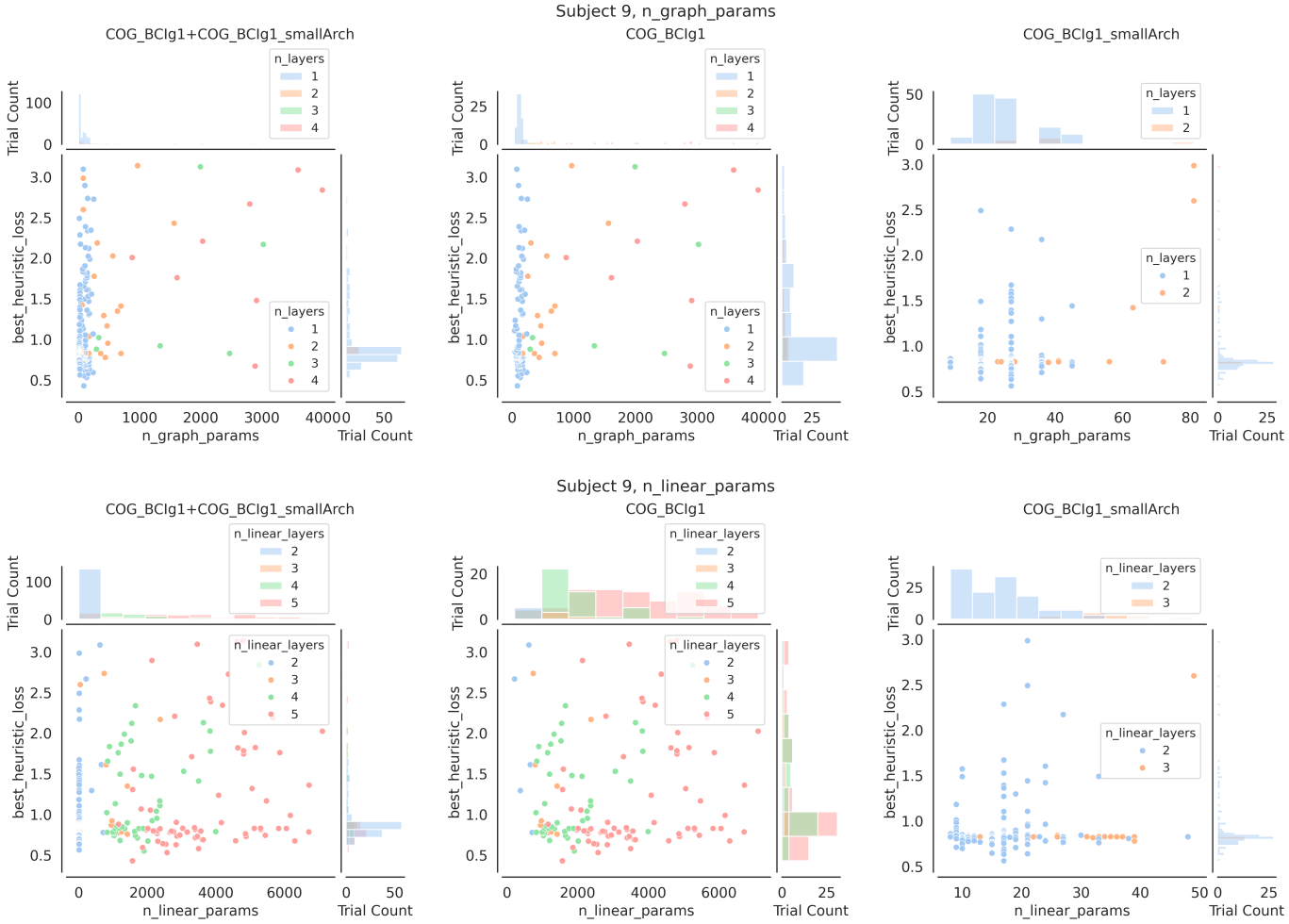


Figure 83: Best Heuristic Loss vs Model Complexity - Subject 9

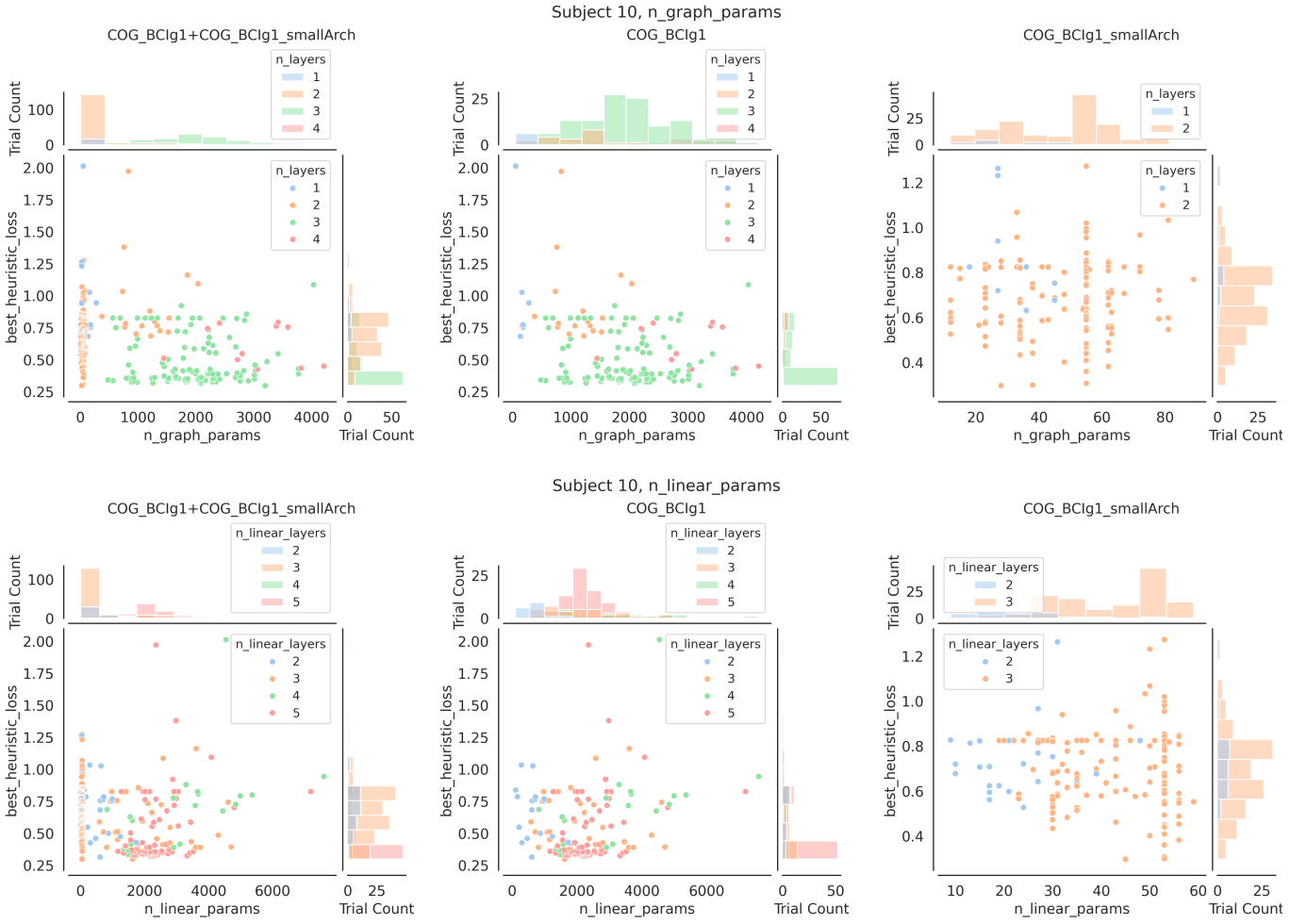


Figure 84: Best Heuristic Loss vs Model Complexity - Subject 10

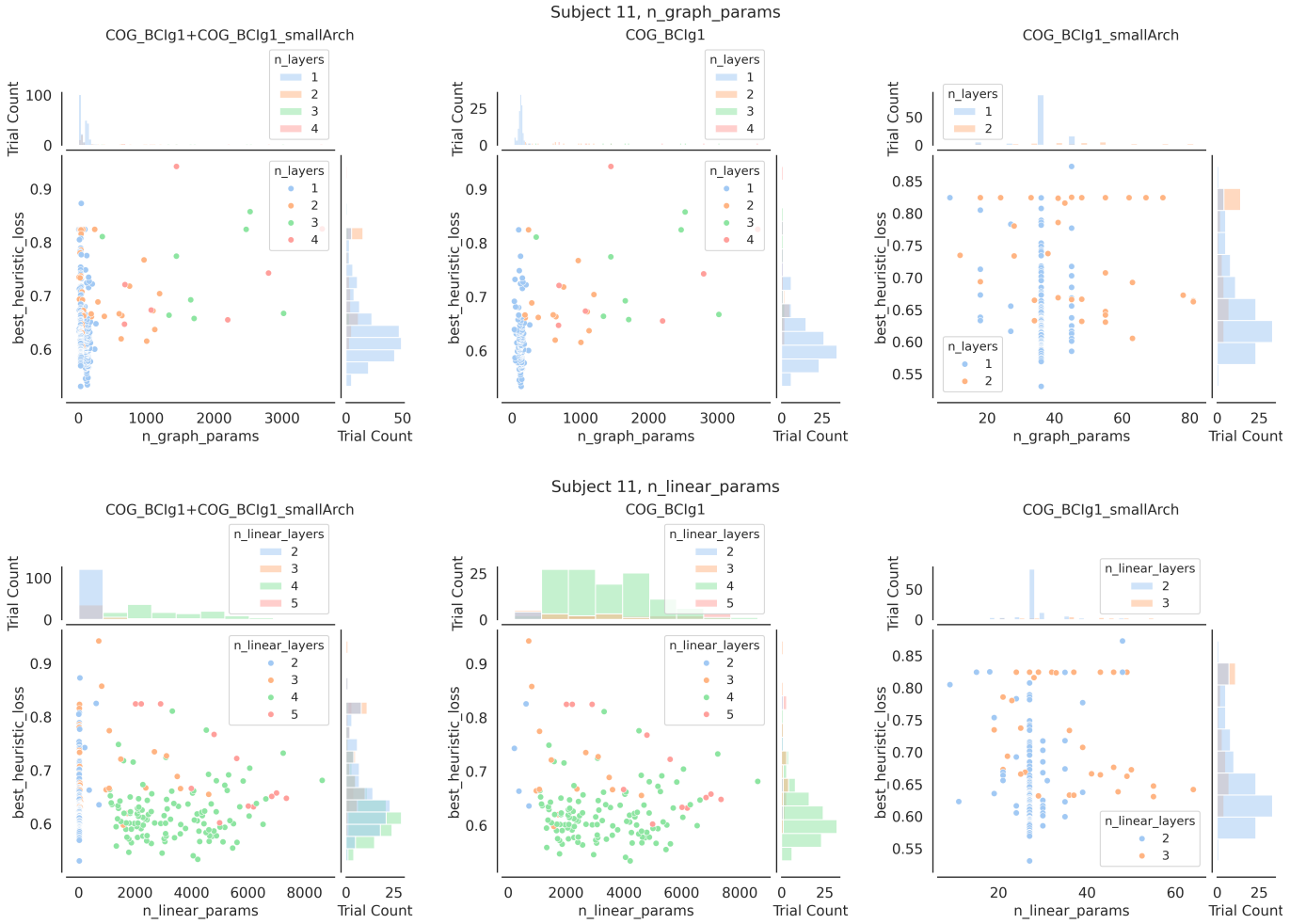


Figure 85: Best Heuristic Loss vs Model Complexity - Subject 11

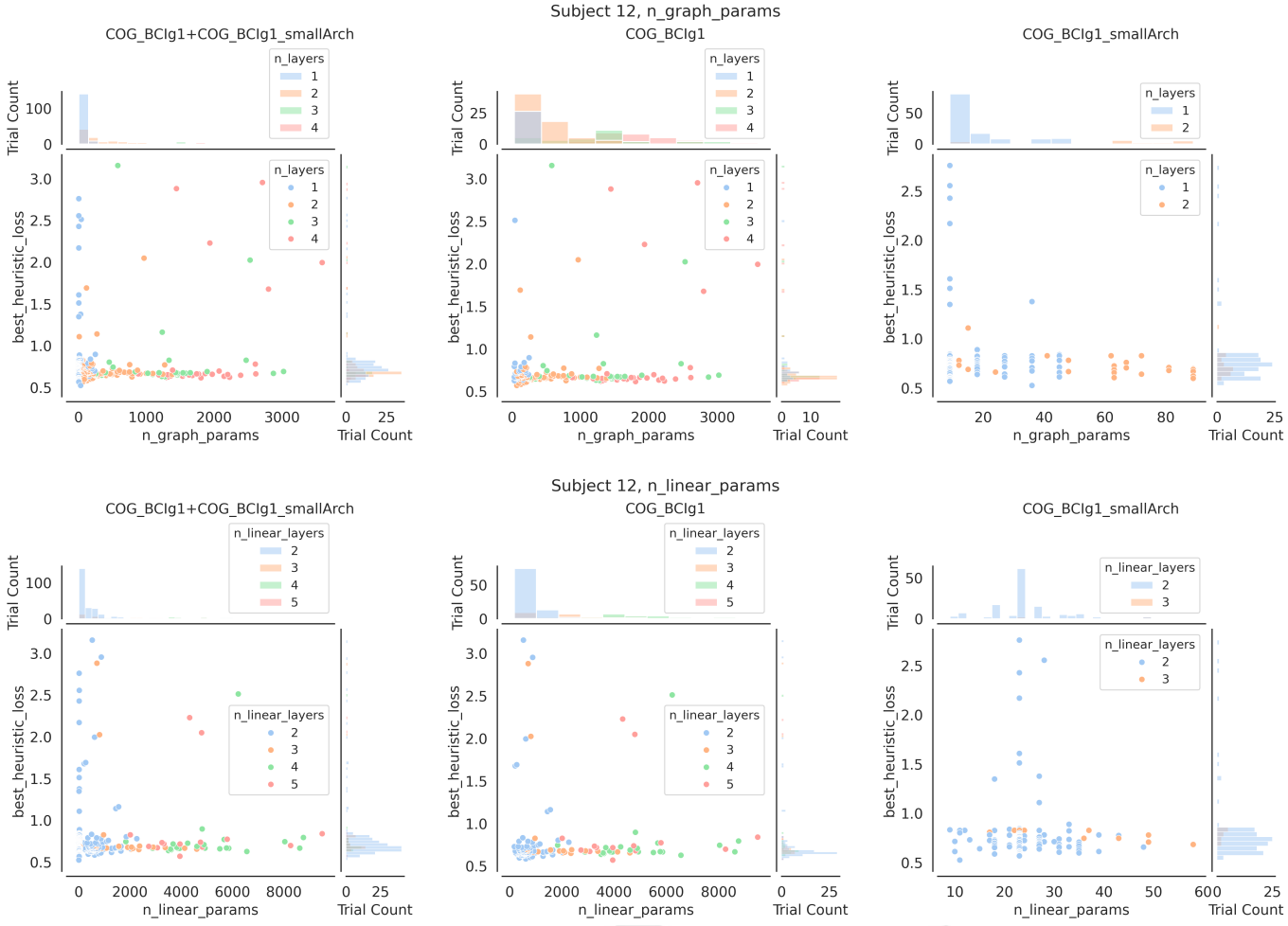


Figure 86: Best Heuristic Loss vs Model Complexity - Subject 12

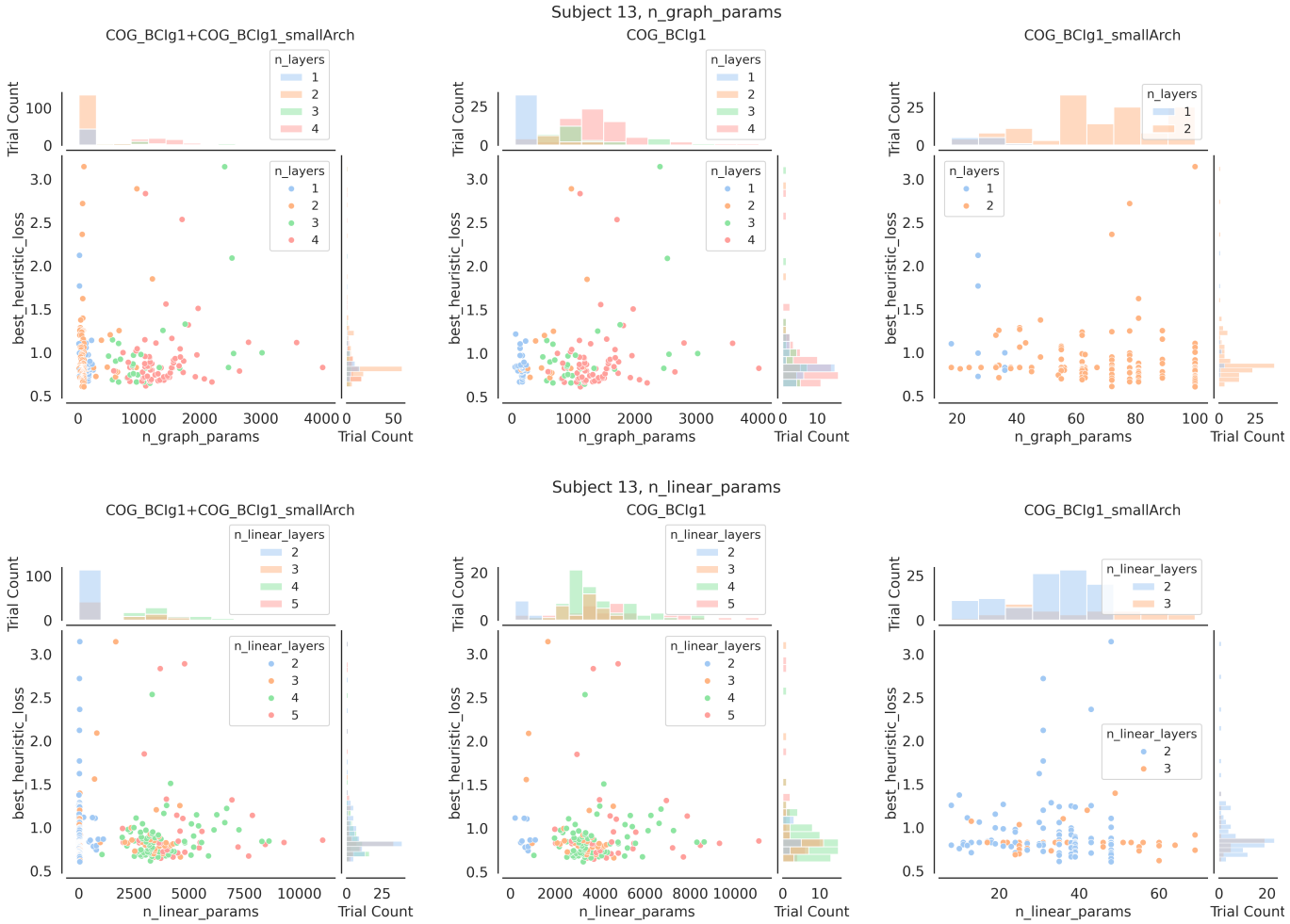


Figure 87: Best Heuristic Loss vs Model Complexity - Subject 13

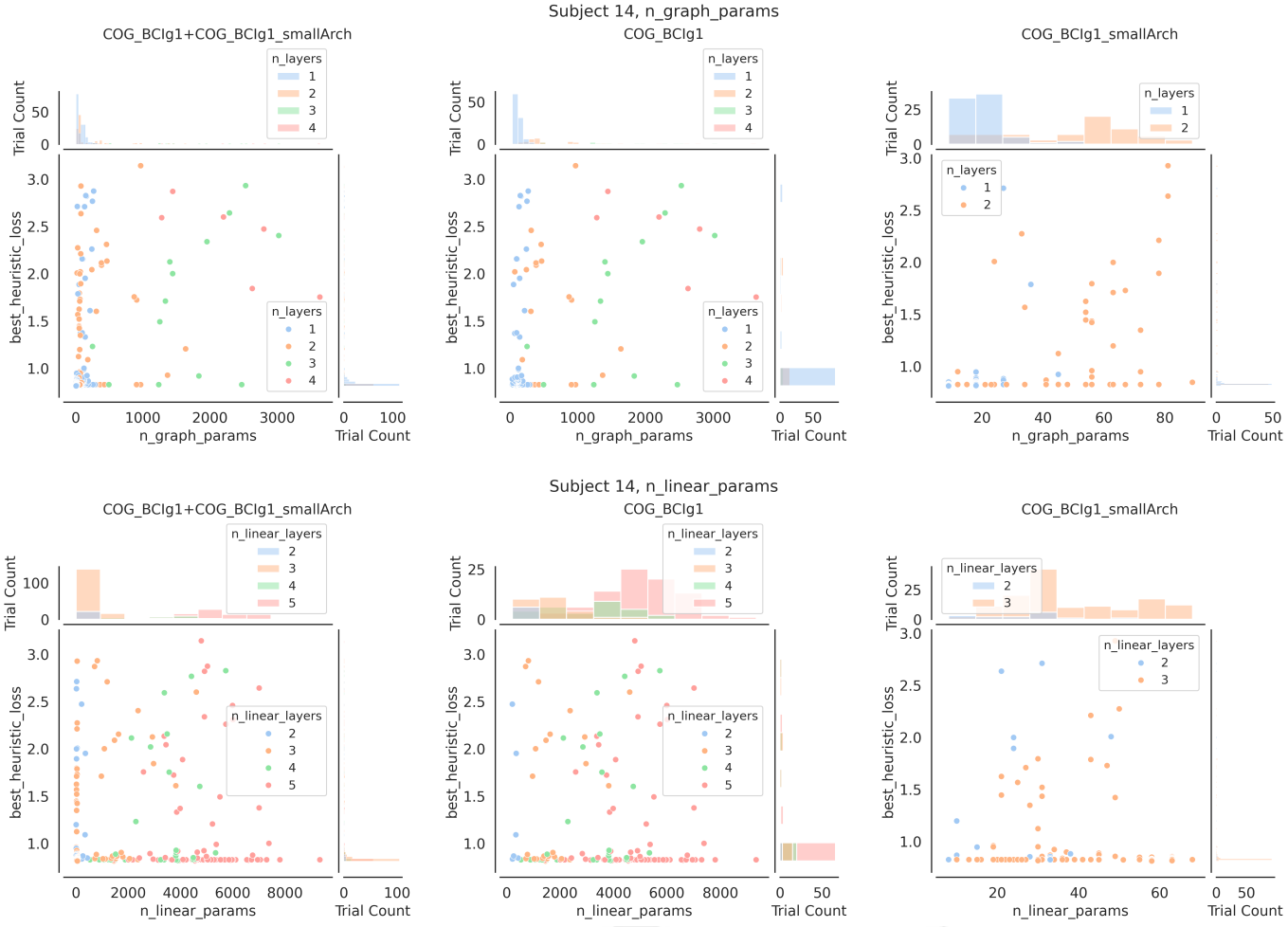


Figure 88: Best Heuristic Loss vs Model Complexity - Subject 14

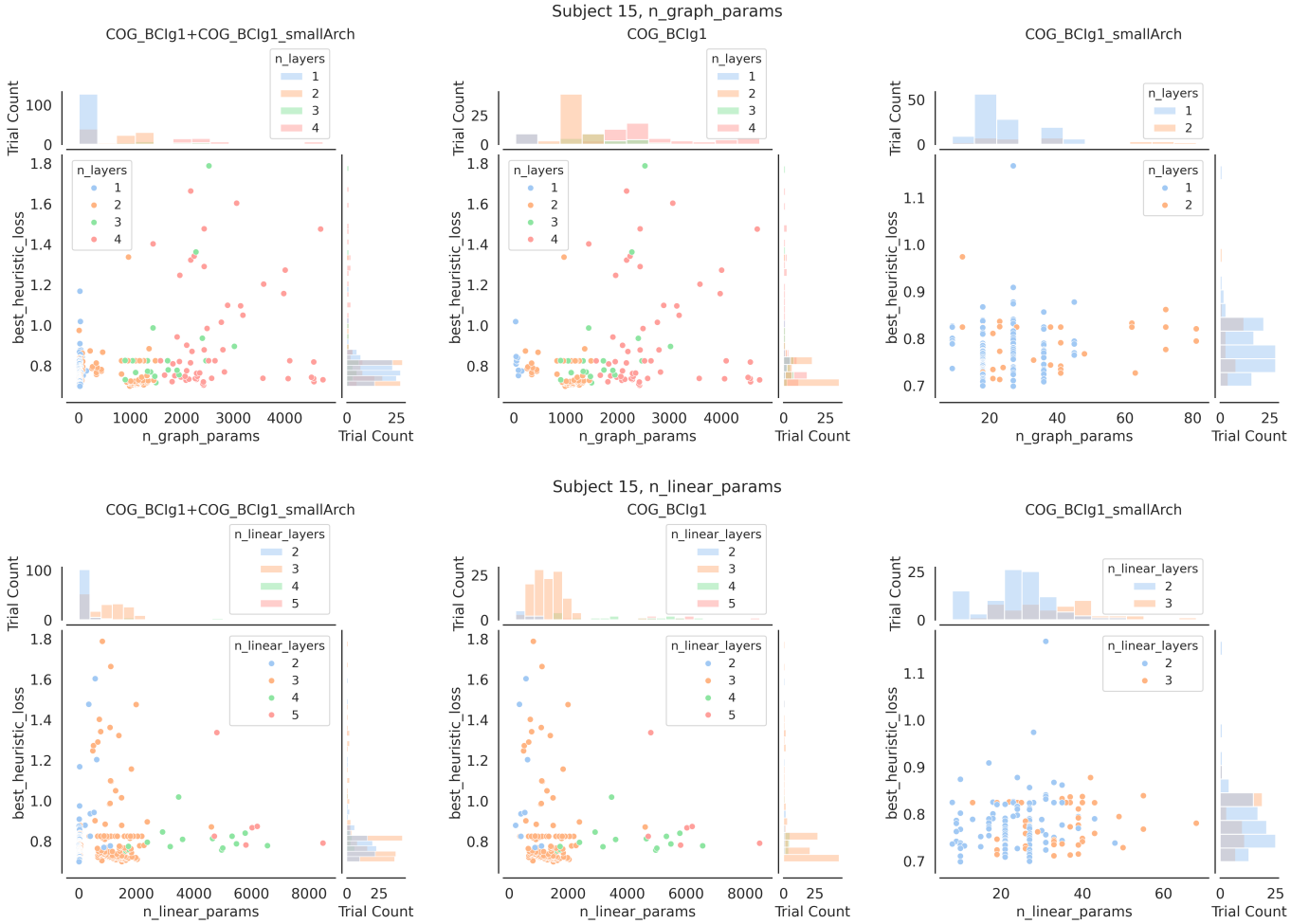


Figure 89: Best Heuristic Loss vs Model Complexity - Subject 15

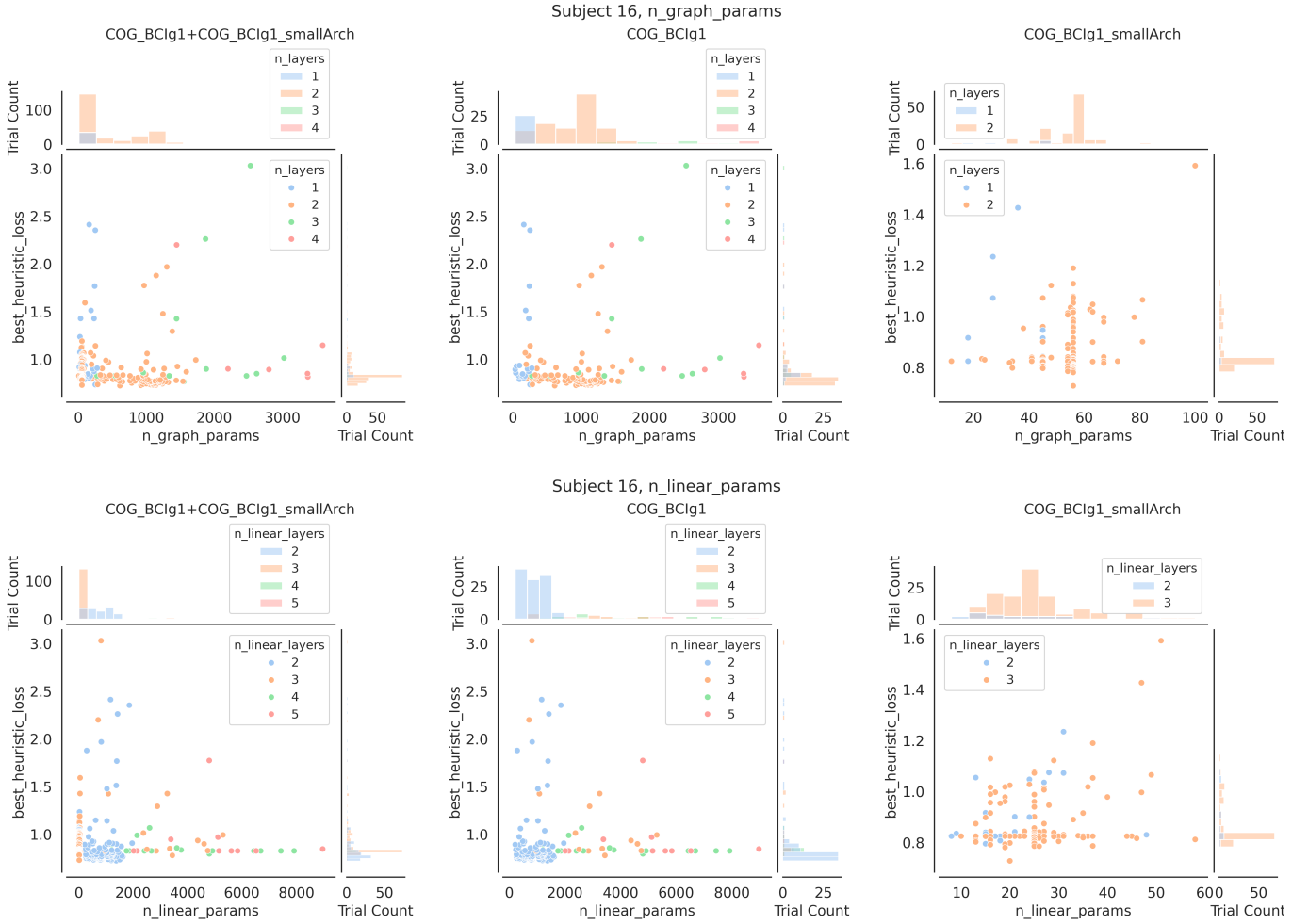


Figure 90: Best Heuristic Loss vs Model Complexity - Subject 16

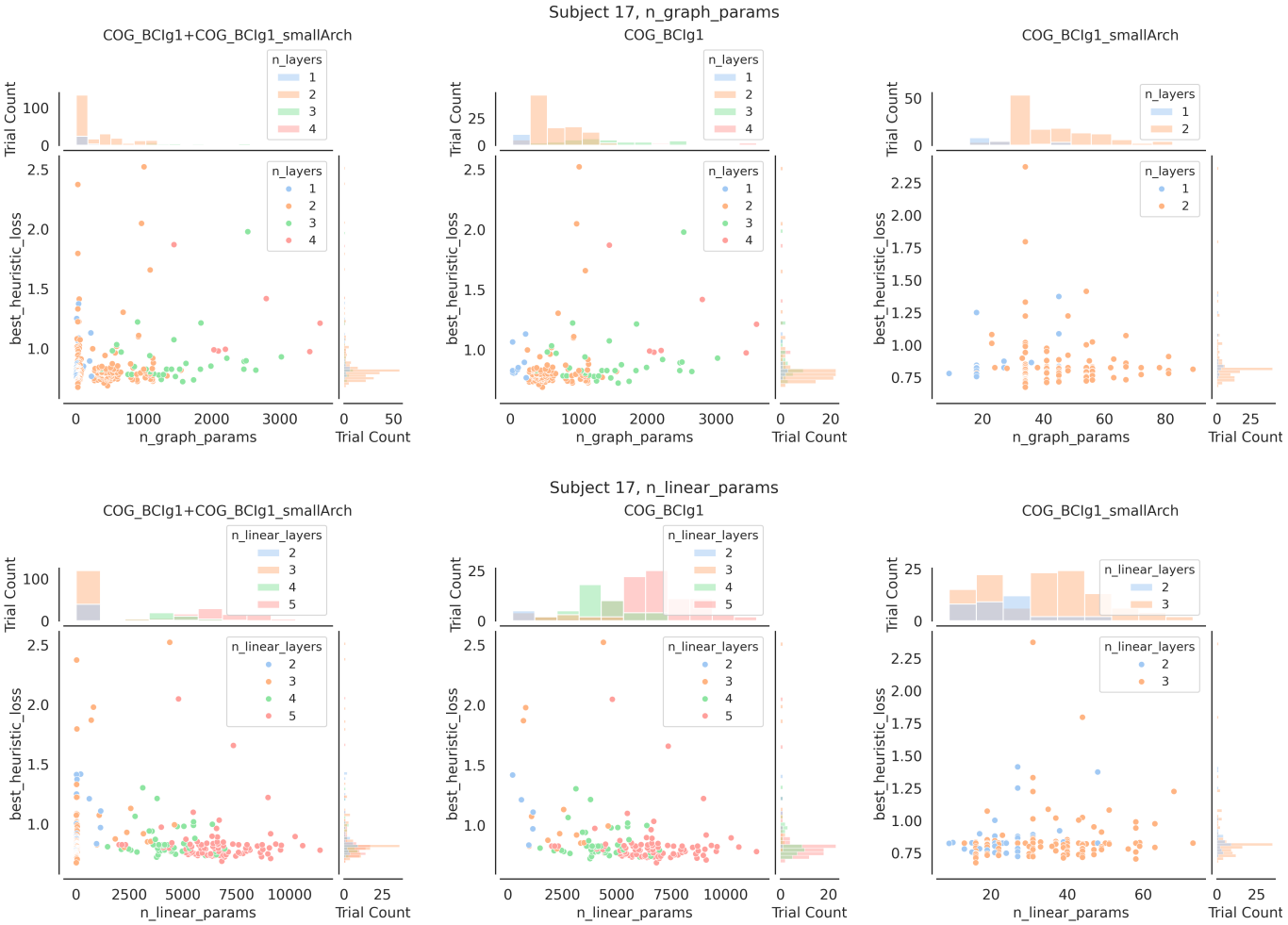


Figure 91: Best Heuristic Loss vs Model Complexity - Subject 17

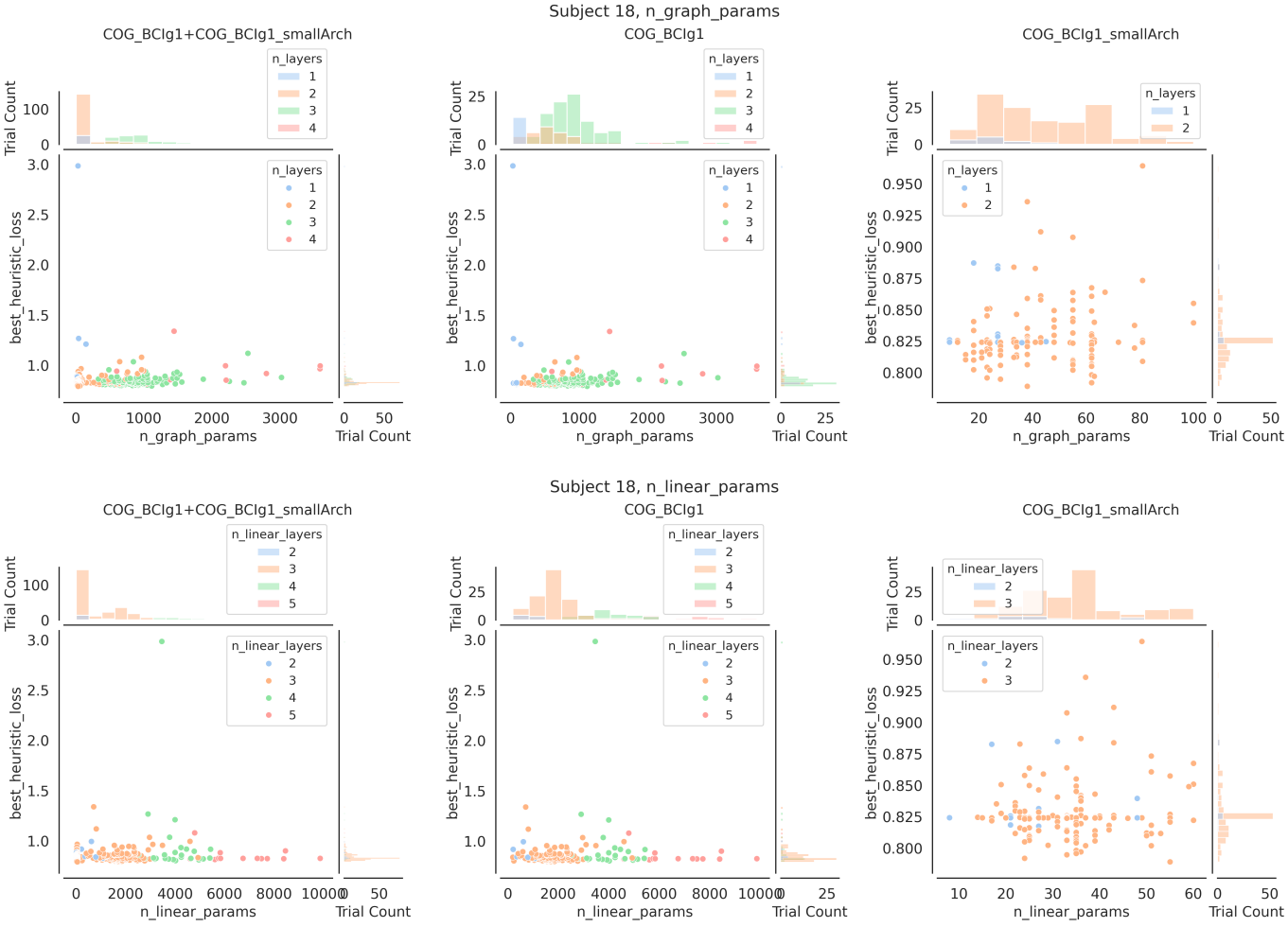


Figure 92: Best Heuristic Loss vs Model Complexity - Subject 18

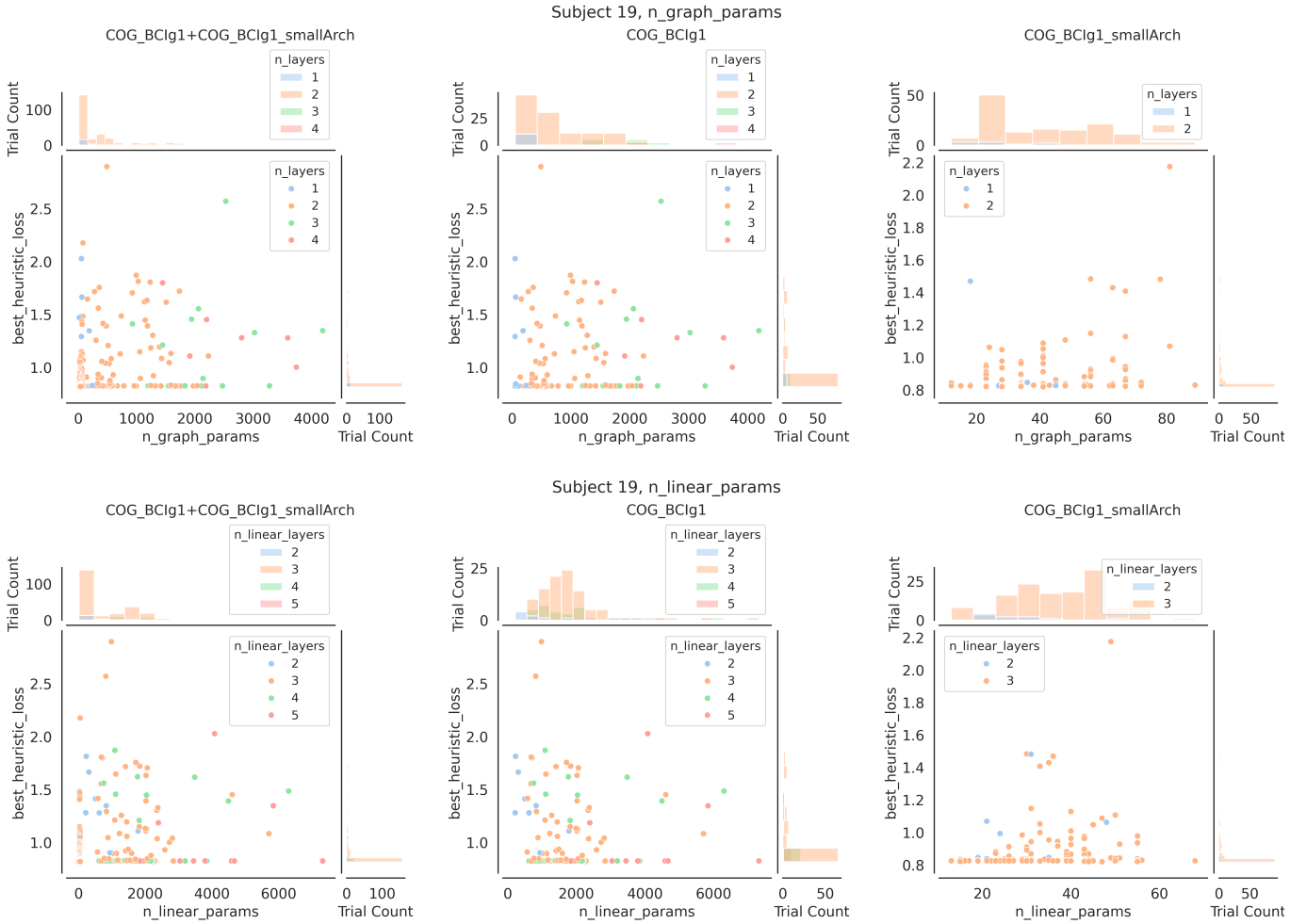


Figure 93: Best Heuristic Loss vs Model Complexity - Subject 19

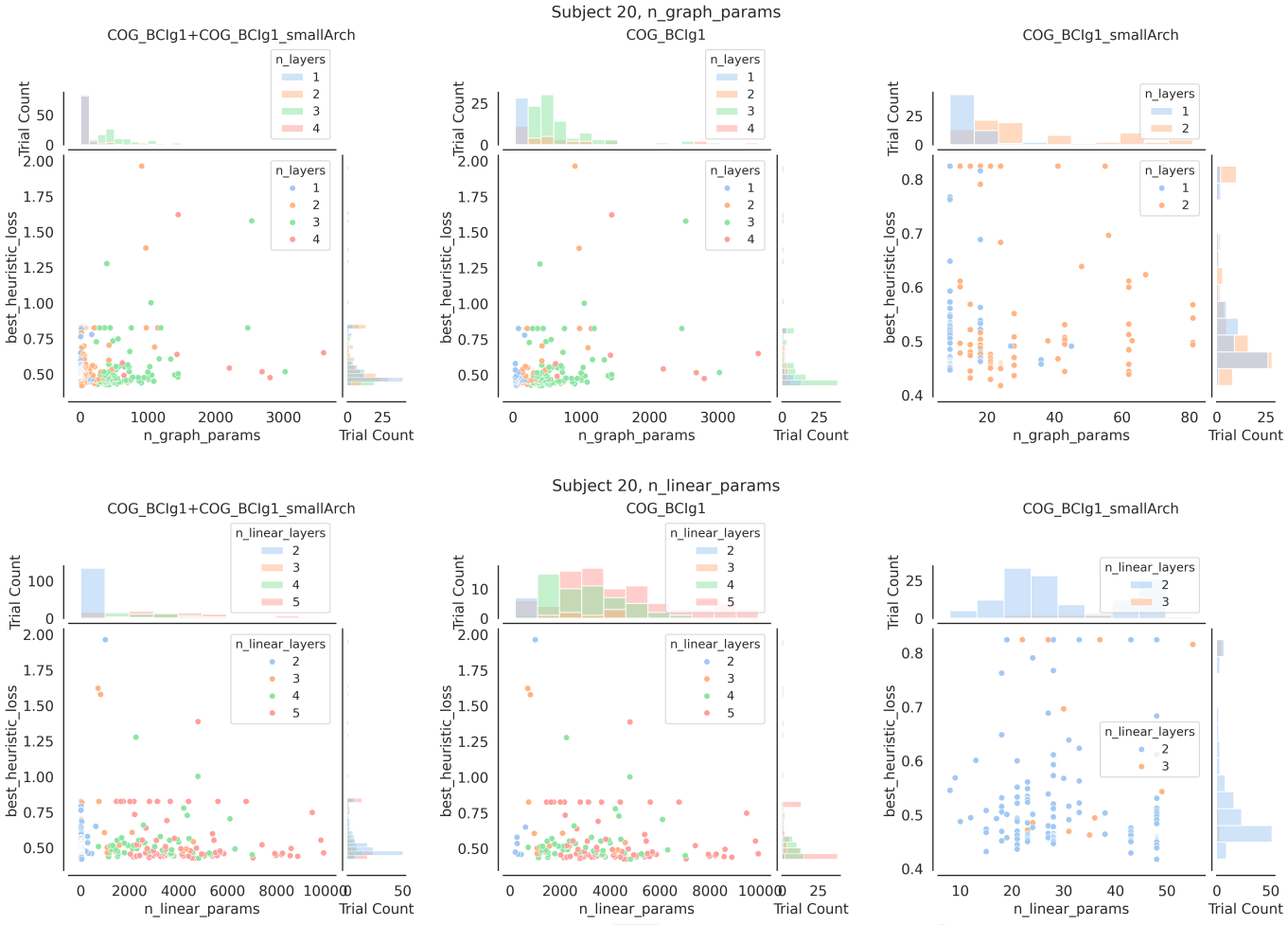


Figure 94: Best Heuristic Loss vs Model Complexity - Subject 20

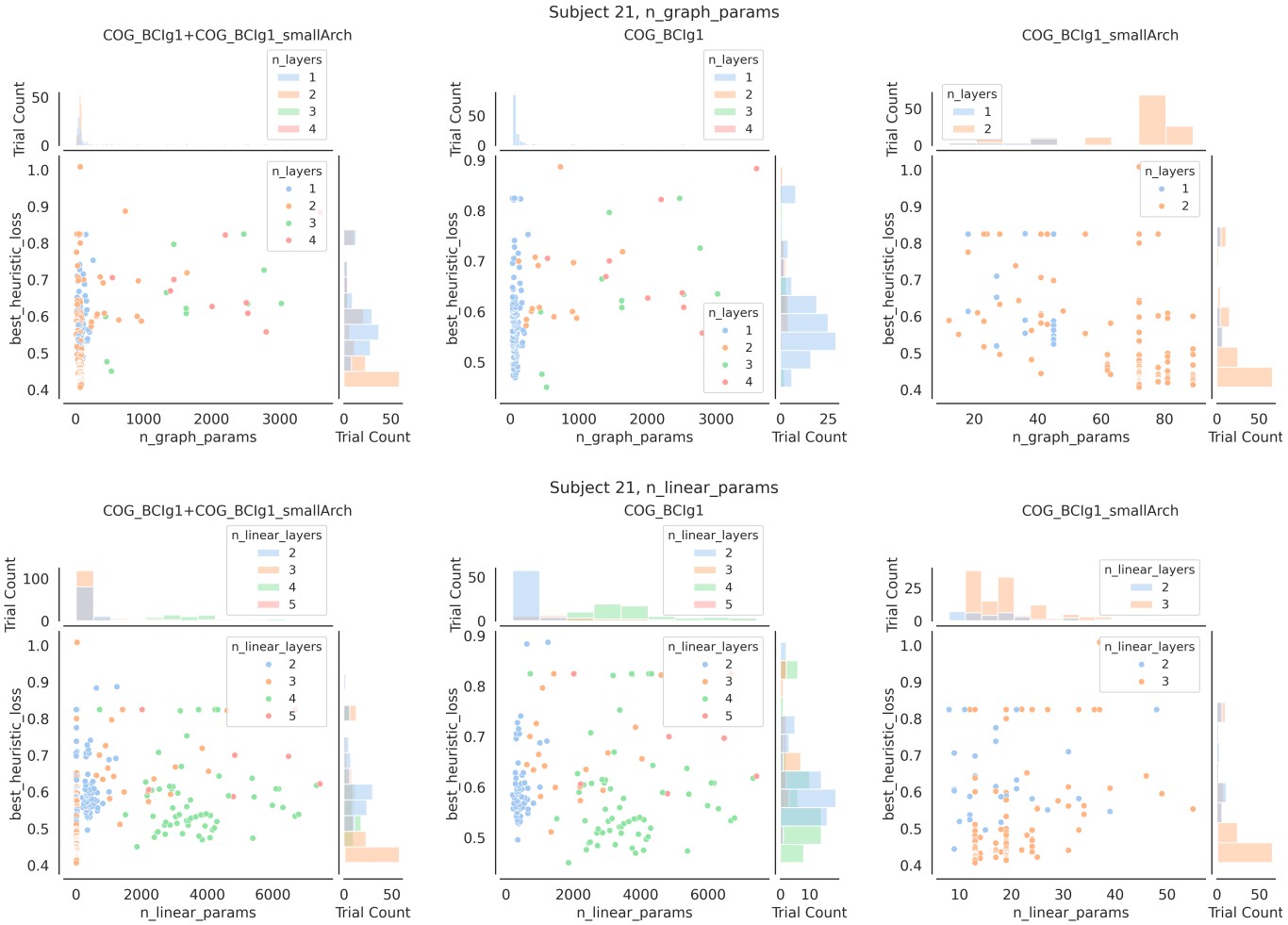


Figure 95: Best Heuristic Loss vs Model Complexity - Subject 21

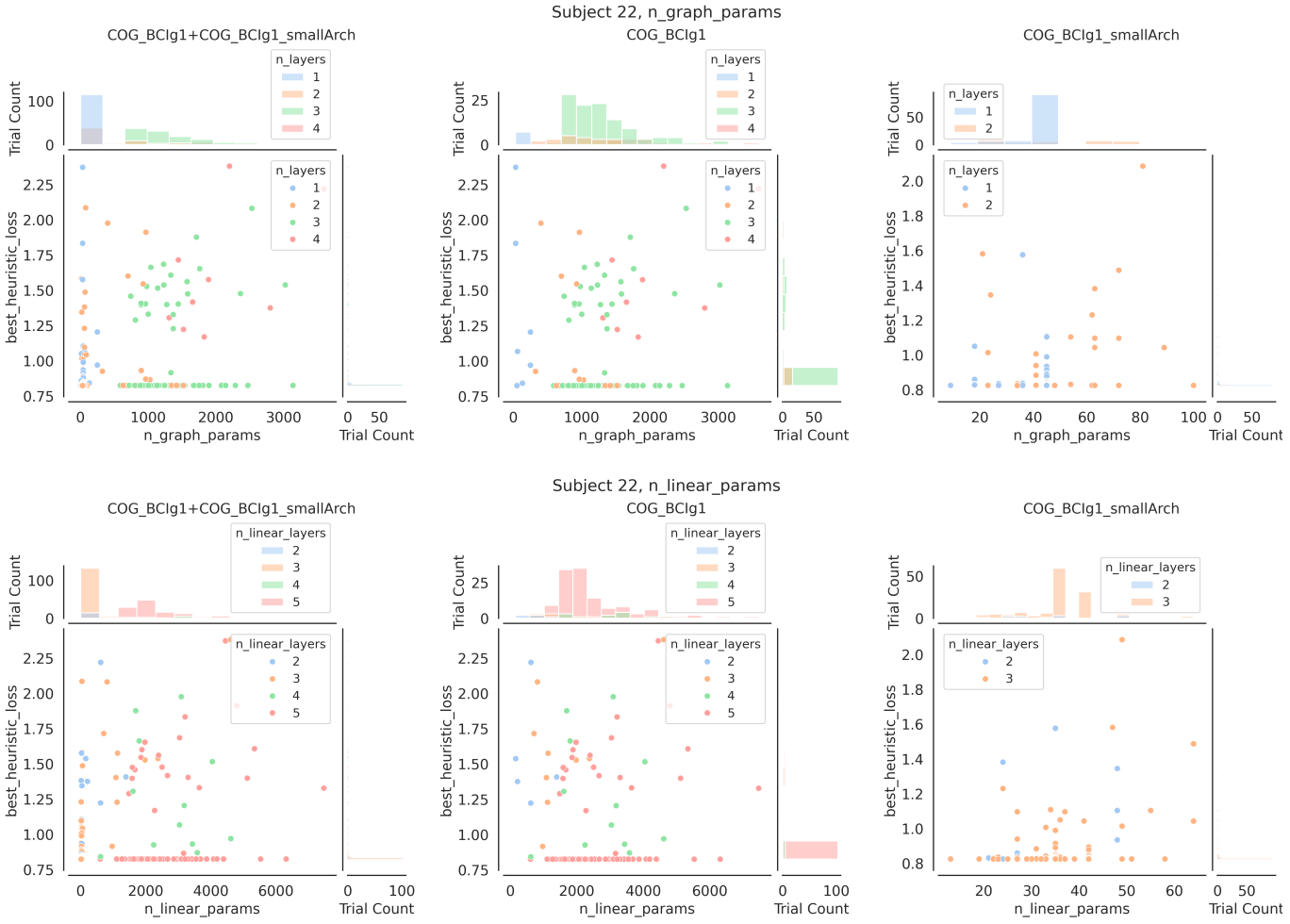


Figure 96: Best Heuristic Loss vs Model Complexity - Subject 22

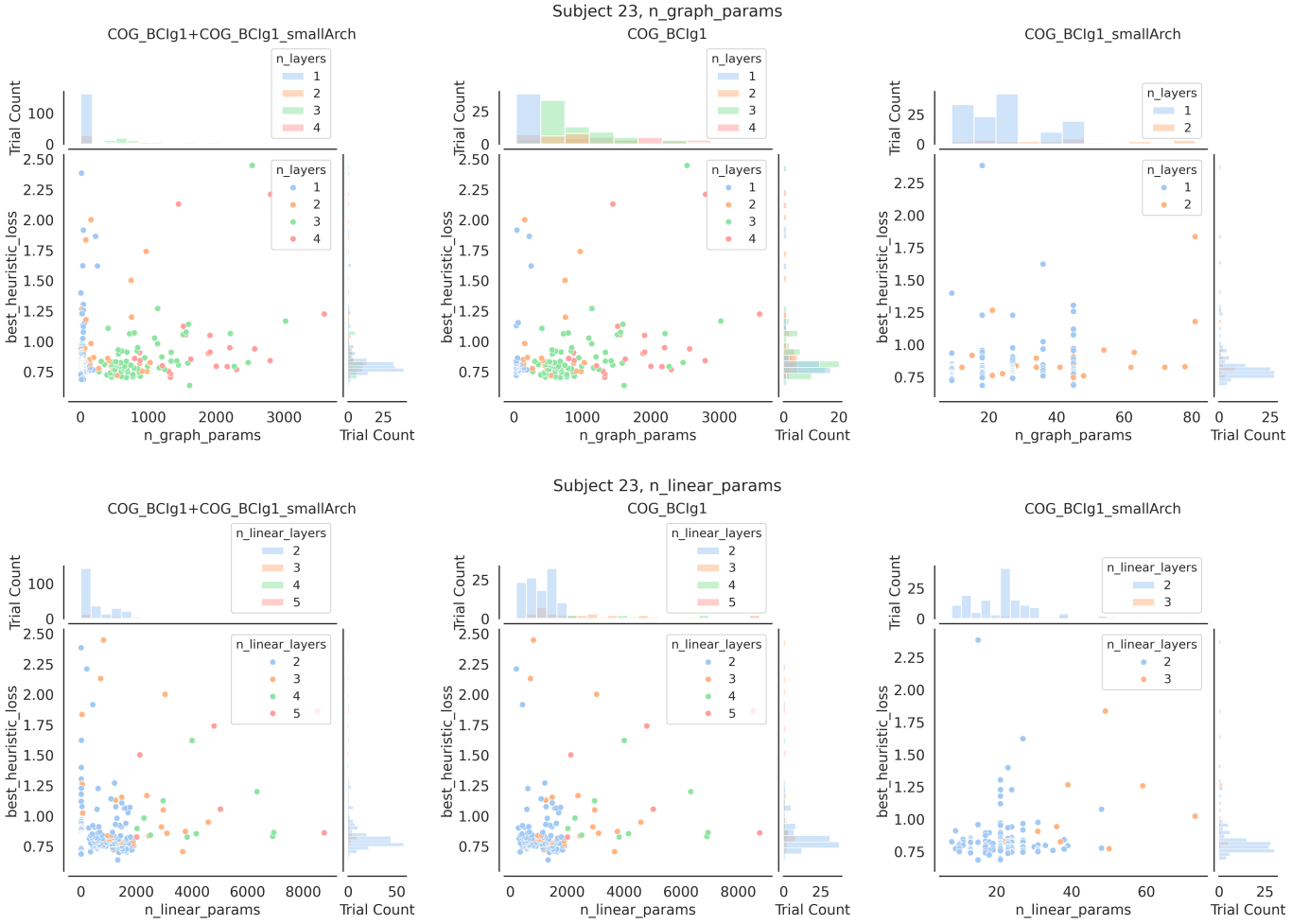


Figure 97: Best Heuristic Loss vs Model Complexity - Subject 23

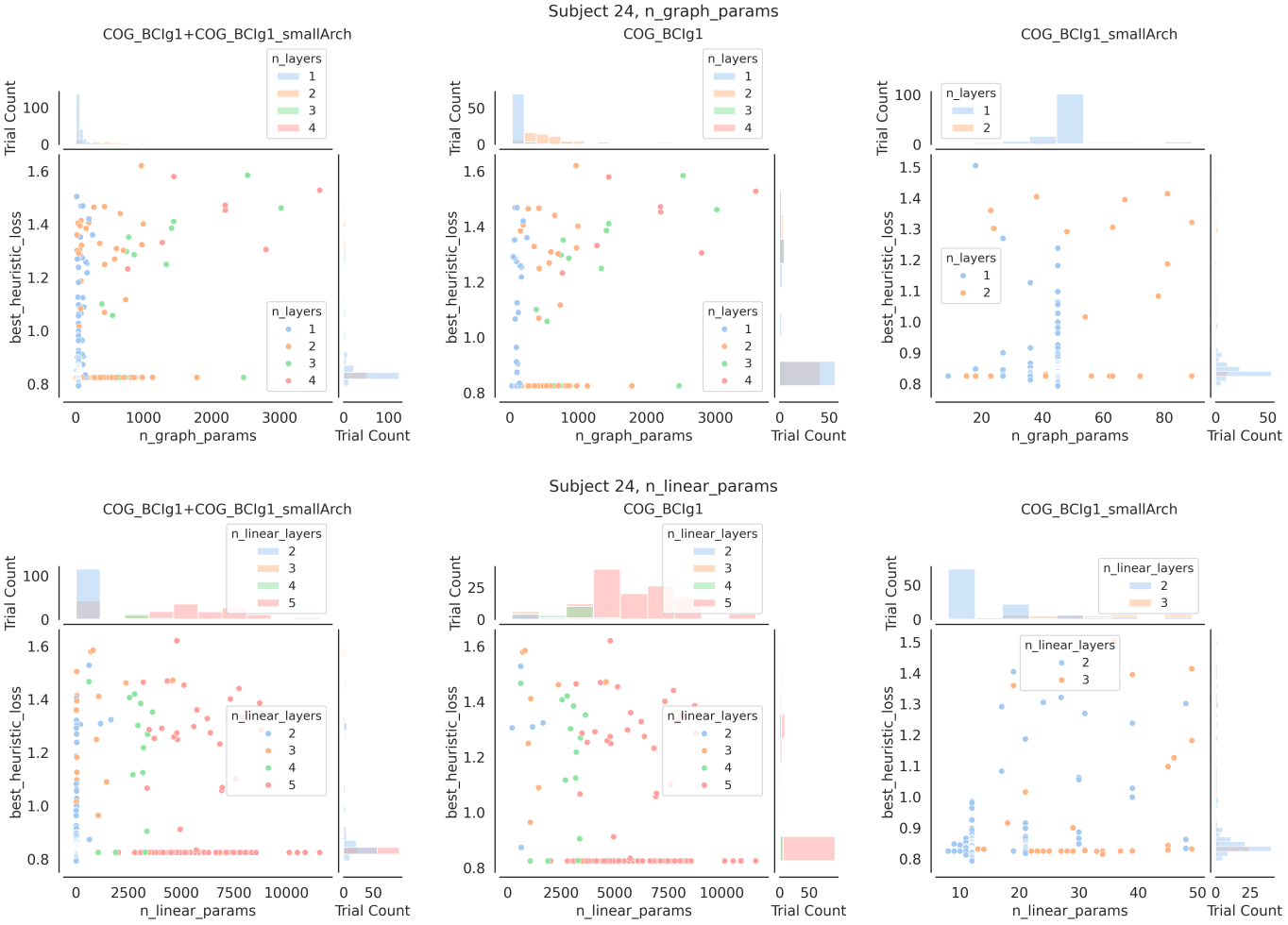


Figure 98: Best Heuristic Loss vs Model Complexity - Subject 24

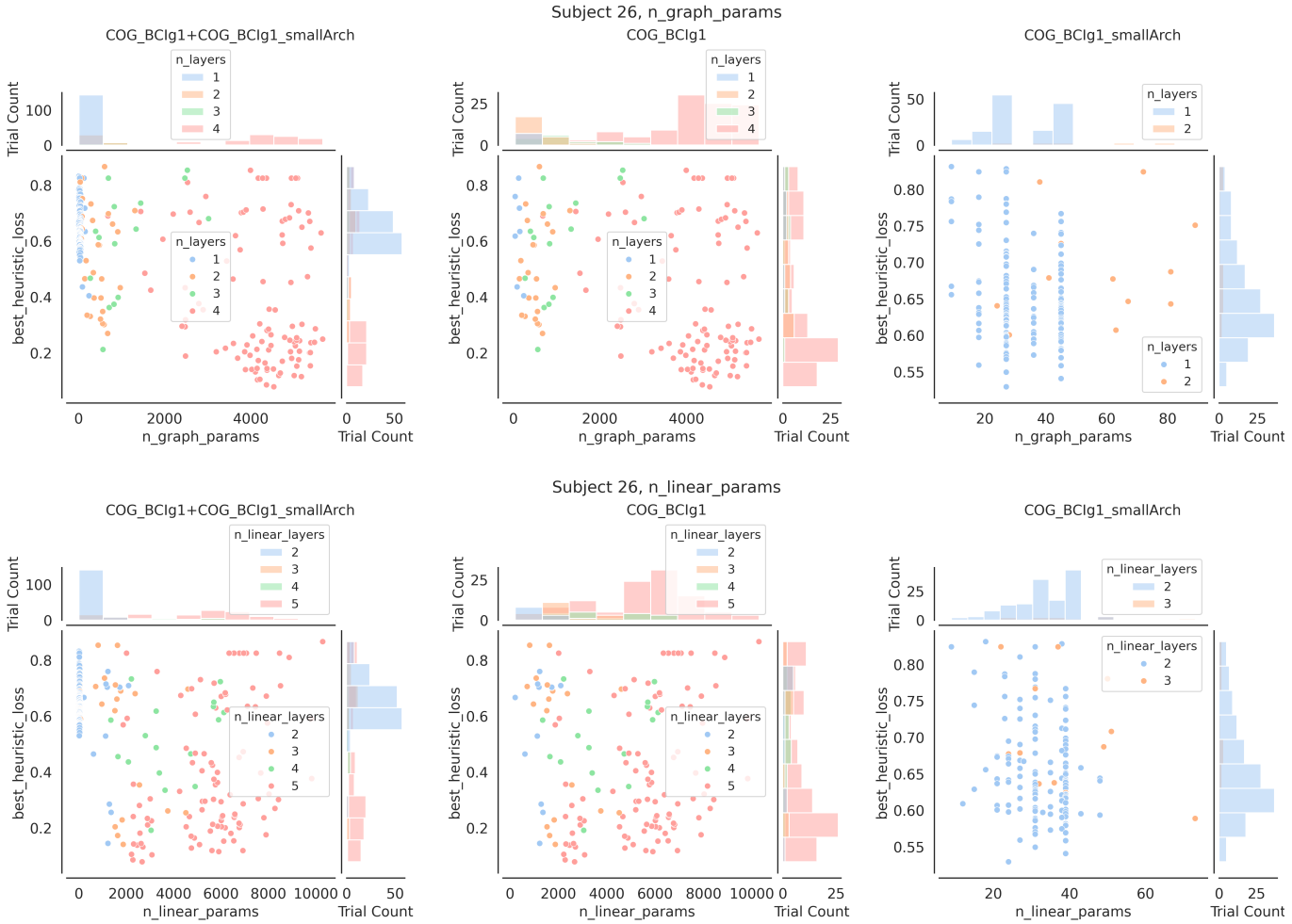


Figure 99: Best Heuristic Loss vs Model Complexity - Subject 26

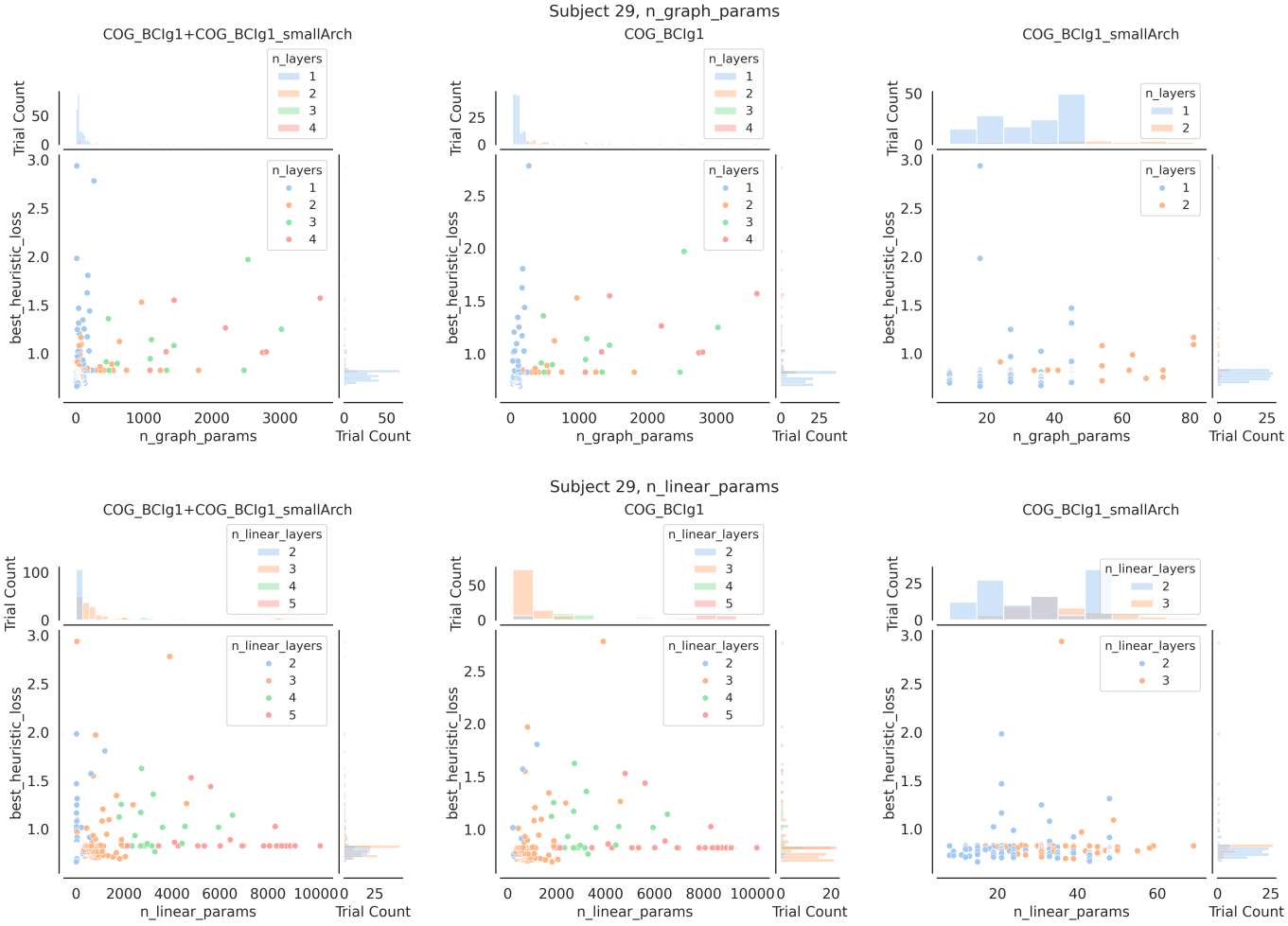


Figure 100: Best Heuristic Loss vs Model Complexity - Subject 29

Appendix D

CURRICULUM VITAE

Güliz Demirezen

EDUCATION

Middle East Technical University, Ankara, Türkiye 2018 - 2024
PhD, Informatics Institute, Information Systems CGPA: 3.71/4.00
Thesis Title: Cross-Session EEG-Based Mental Workload Classification Using Graph Neural Networks for Reproducible Brain-Computer Interface Applications

Middle East Technical University, Ankara, Türkiye 2014 - 2018
MSc, Informatics Institute, Information Systems CGPA: 3.94/4.00
Thesis Title: Automatic information coverage assessment of diabetes websites

Middle East Technical University, Ankara, Türkiye 2002 - 2006
Minor, Computer Engineering

Middle East Technical University, Ankara, Türkiye 2001 - 2006
BS, Aerospace Engineering, rank: 2

RESEARCH INTERESTS

Machine Learning, Deep Learning, Applied Artificial Intelligence, SciML, Neuroscience, Neuroergonomics, Modeling and Simulation, AI for Good

EXPERIENCE

Roketsan Missiles Inc. Ankara, Türkiye

Senior Lead Engineer 2024 - Present
System Support Equipments Development Department

Senior Lead Engineer 2022 - 2024
Software Engineering and Test Department

Senior Lead Engineer 2021 - 2022
System Technologies Development Directorate

Lead Engineer 2020 - 2021
System Technologies Development Directorate

Lead Engineer 2019 - 2020
Artificial Intelligence and Technology Management Department

Lead Engineer 2016 - 2019
Tactical Missile Systems Modeling and Simulation Department

Senior Engineer 2012 - 2015
Tactical Missile Systems Modeling and Simulation Department

Expert Engineer 2010 - 2012
Tactical Missile Systems Modeling and Simulation Department

Engineer 2006 - 2010
System Design and Integration Department, Modeling and Simulation

PROJECTS

Researcher, METU Scientific Research Project (BAP) Ankara, Türkiye
Automatic Assessment of Information Quality on Web Sites 05/2018 - 05/2019

PUBLICATIONS

Journal paper

- **Demirezen, G.**, Taskaya Temizel, T., & Brouwer, A. M. (2024). Reproducible machine learning research in mental workload classification using EEG. Frontiers in Neuroergonomics, 5, 1346794.

Proceedings

- Filgöz, A., **Demirezen, G.,** & Demirezen, M. U. (2021, October). Applying novel adaptive activation function theory for launch acceptability region estimation with neural networks in constrained hardware environments: Performance comparison. In 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) (pp. 1-10). IEEE.
  Received the *Best of Session Award*

Thesis

- **Bulut, G.** (2018). Automatic information coverage assessment of diabetes websites (Master's thesis, Middle East Technical University).

SCHOLARSHIPS

- Higher Education Institutions 100/2000 Doctorate Scholarship

VOLUNTEER EXPERIENCE

- Artificial Intelligence Association Policy (AIPA) Climate Change Commission, **Vice Chair**, 07/2024 - present
- Omdena Local Chapter Project, *Developing a Web App to Detect Hepatocellular Carcinoma Histopathology Using Deep Learning and XAI*, **Data Team Co-lead**, 04/2023 - 05/2023
- Omdena Local Chapter Project, *Predicting Autism in Toddlers Through Machine Learning*, **Modeling Team Co-lead**, 01/2023 - 03/2023