DEEP LEARNING CLASSIFICATION OF COGNITIVE WORKLOAD LEVELS FROM EEG WAVELET TRANSFORM IMAGES

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

VOLKAN DOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2023

DEEP LEARNING CLASSIFICATION OF COGNITIVE WORKLOAD LEVELS FROM EEG WAVELET TRANSFORM IMAGES

Submitted by VOLKAN DOĞAN in partial fulfillment of the requirements for the degree of **Doctor of Master of Science in Cognitive Science Department, Middle East Technical University** by,

Date: _ 06.09.2023

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, Last name : Volkan Doğan

Signature : _____

ABSTRACT

DEEP LEARNING CLASSIFICATION OF COGNITIVE WORKLOAD LEVELS FROM EEG WAVELET TRANSFORM IMAGES

Doğan, Volkan
MSc., Department of Cognitive Science
Supervisor: Assoc. Prof. Dr. Murat Perit Çakır

September 2023, 63 pages

Electroencephalogram (EEG) signals provide a non-invasive method to study cognitive processes. This study aimed to classify Multi-Attribute Task Battery (MATB) task difficulties based on wavelet transform images of EEG signals using deep learning models. An EEG dataset collected from 29 subjects while performing the MATB tasks of varying difficulties by Hinss et al. (2023) were transformed into wavelet images that can accommodate time-frequency information at the same time for further analysis. Three deep learning models, EfficientNet-B0, ResNet18, and ResNet50, were trained and tested on these images under different conditions, including pretrained and non-pretrained models, and using different optimizers. The models' performance was evaluated based on overall accuracy and accuracy by subject, session, and task difficulty. The pretrained EfficientNet-B0 model achieved the highest overall accuracy (67.52%). However, the performance varied significantly across subjects and task difficulties, indicating limited generalizability. The model's accuracy was lower for medium tasks, suggesting difficulty in distinguishing between medium and other levels of difficulty. While deep learning models can achieve high accuracy in classifying MATB task difficulty based on EEG signals, their performance varies across individuals and task difficulties. Further research is needed to improve model generalizability, optimize performance across all task difficulties, and validate the models on larger and more diverse datasets.

Keywords: EEG, Deep Learning, Cognitive Workload, Wavelet Transform, Task Difficulty Classification

ÖZ

EEG DALGACIK DÖNÜŞÜM GÖRÜNTÜLERİNDEN BİLİŞSEL YÜK SEVİYELERİNİN DERİN ÖĞRENME İLE SINIFLANDIRILMASI

Doğan, Volkan

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Doç. Dr. Murat Perit Çakır

Eylül 2023, 63 sayfa

Electroensefalografi (EEG) sinyalleri, bilişsel süreçleri incelemek için invaziv olmayan bir yöntem sunmaktadır. Bu çalışma, derin öğrenme modellerini kullanarak EEG sinyallerinin dalgacık (wavelet) dönüşüm görüntülerine dayalı olarak Çoklu Özellik Görev Bataryası (MATB) görev zorluklarını sınıflandırmayı amaçlamıştır. Tez çalışması kapsamında MATB görevlerini değişen zorluk seviyelerinde gerçekleştiren 29 denekten toplanan EEG verilerini içeren ve Hinss ve ark. (2023) tarafından toplanmış olan veri seti ilk aşamada zaman-frekans bilgilerini bir arada temsil edebilen dalgacık görüntülerine dönüştürülmüştür. EfficientNet-B0, ResNet18 ve ResNet50 olmak üzere üç derin öğrenme modeli, bu görüntüler üzerinde, önceden eğitilmiş ve eğitilmemiş modeller ile farklı optimizasyon algoritmaları kullanılarak eğitilmiş ve test edilmiştir. Modellerin performansı, genel doğruluk skoru, denek başına doğruluk, seansa göre doğruluk ve görev zorluğuna göre doğruluk temel alınarak değerlendirilmiştir. Önceden eğitilmiş EfficientNet-B0 modelinin en yüksek genel doğruluk skorunu (%67.52) sağladığı gözlenmiştir. Ancak performansın denekler ve görev zorlukları arasında önemli ölçüde değişkenlik göstermesi genellenebilirlik bakımından bazı sınırlılıkların olduğuna işaret etmektedir. Modelin doğruluk skorunun orta seviye görevler için daha düşük olması, orta seviye ve diğer zorluk seviyeleri arasında ayrım yapmada zorluk yaşandığını göstermiştir. Model genellenebilirliğini iyileştirmek, tüm görev zorluklarına göre performansı optimize etmek ve modelleri daha büyük ve daha çeşitli veri setlerinde doğrulamak için daha fazla araştırma yapılması gerekmektedir.

Anahtar Sözcükler: EEG, Derin Öğrenme, Bilişsel Yük, Dalga Dönüşümü, Görev Zorluğu Sınıflandırması

To My Father and Mother,

For your endless love, support, and encouragement throughout my journey. Your belief in me has been my guiding light, and this achievement is a testament to your unwavering faith. Thank you for everything.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Murat Perit Çakır, for his invaluable knowledge, guidance, and support throughout this research journey. His dedication and efforts have been instrumental in making this thesis possible, and I am immensely thankful for his mentorship.

I am profoundly grateful to İzzet Türkalp Akbaşlı for his unwavering support throughout this journey. His readiness to assist at any time has been a source of strength and encouragement for me.

Additionally, I am grateful to my friends, Özge Şencoşkun, Dorukhan Tüfekçi, Erencem Akça, İrem Karapolat, and Sultan Nilay Can. Their support, companionship, and efforts to keep me grounded in the real world have been invaluable to me during this time.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xiv
CHAPTERS

1. INTRODUCTION 1
2 LITERATURE REVIEW 5
2.1 Introduction to Mental Workload (MWL) 5
2.1.1 Definition of MWL 5
2.1.2 Importance of Managing MWL in Human-Computer Interaction 5
2.1.3 Motivations for Managing MWL: optimizing user performance, enhancing user engagement, minimizing errors 5
2.2 Theoretical Background of MWL 5
2.2.1 Key Theories and Models Related to MWL 5
2.2.2 Factors Influencing MWL 6
2.3 Measurement of MWL 6
2.3.1 Importance of Measuring MWL 6
2.3.2 Overview of Measurement Methods 6
2.4 EEG as a Tool for Measuring MWL 7
2.4.1 Introduction to EEG 7
2.4.2 EEG frequency bands and their significance 7
2.4.3 Advantages and limitations of using EEG to measure MWL 8
2.5 Wavelet Transform for EEG Signal Processing 9
2.5.1 Introduction to wavelet transform 9

2.5.2 Application of wavelet transform to EEG data 9
2.5.3 Advantages of using wavelet transform for EEG data analysis 9
2.6 Previous Studies Using EEG and Wavelet Transform for MWL Assessment .. 11
2.6.1 Wavelet Coefficients in EEG Analysis for Cognitive Load Classification 11
2.6.2 Emotion Recognition through EEG Signals using Wavelet Transform 11
2.7 Cross-Session Variability in pBCI: Insights from the First pBCI Competition on Workload Estimation 12
2.8 COG-BCI Database Usability Validation for Mental Workload Estimation 13
3 METHOD 15
3.1 Dataset 15
3.1.1 Data Collection 15
3.1.2 Experimental Tasks 16
3.2 Dataset Preprocessing 18
3.2.1 EEG Data Preprocessing Techniques 18
3.2.2 Libraries and Software Tools 18
3.2 Training 29
3.2.1 Preparing the Dataset for Training 29
3.2.2 Dataloader and Dataset Structuring 31
3.2.3 Model Architectures 33
3.2.4 Transfer Learning Purpose and Advantages 34
3.2.5 Model Training Details 34
3.2.6 Evaluating Models during Training 35
4 RESULTS 37
4.1 Preprocessing Results from Python Libraries 37
4.2 Model Experiments 37
4.3 Analysis by Session 38
4.4 Analysis by Subject with 5-Fold Cross Validation 42
4.5 Analysis by Task Difficulty 45
5 DISCUSSION 49
5.1 Comparison with Previous Studies 49
5.2 Model Performance 50

5.3 Potential Improvements 51
5.4 Applications and Implications 52
5.4.1 Operationalizing Cognitive Constructs: Implications and Comparisons ... 52
5.5 Limitations 53
5.6 Conclusion 54
REFERENCES 55
APPENDIX 59

LIST OF TABLES

Table 2.1: Spatial and Temporal Sensitivity Comparison of BCI Techniques (Coyle, 2005, p.35) 9

LIST OF FIGURES

Figure 2.1: Tiling of the Time-Frequency Graph for the Wavelet Transform (Issartel et al., 2015).....10
Figure 3.1: MATB-II Task Interface https://matb.larc.nasa.gov.....17
Figure 3.2: 2D Locations of 62 Electrodes. Visualized with MNE library.....19
Figure 3.3: 3D Electrode Locations. Visualized with MNE library.....20
Figure 3.4: Example EEG channels' Voltage vs Time Graph from Subject 01, Session 03, MATBeasy Task, 44 to 54 second interval from 300 seconds of data Extracted with Python MNE library. 20
Figure 3.5: Wavelet Output using Squuezepy Subject 10, Session 3, MATB Easy Task, F4 Channel 22
Figure 3.6: Wavelet Output using Squuezepy for Subject 10, Session 3, MATB Medium Task, F4 Channel.....23
Figure 3.7: Wavelet Output using Squuezepy for Subject 10, Session 3, MATB Difficult Task, F4 Channel.....24
Figure 3.8: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Easy Task, F4 Channel 26
Figure 3.9: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Medium Task, F4 Channel 27
Figure 3.10: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Difficult Task, F4 Channel 28
Figure 3.11: Wavelet Output using EEGLAB and MATLAB for Subject 9, Session 2, MATB Difficult Task, PO7 Channel 31
Figure 3.12: Example Wavelet Transform Output from Subject 26, Session 1, MATB Difficult Task AF3 Channel in approx. 1-31 Hz Frequency Range.....32
Figure 3.13: Resized Format of Figure 3.6 to 224x224 for Training Input 33
Figure 4.1: Overall Model Accuracy Comparison, Trained on Session 1-2, Tested on Session 3.....38
Figure 4.2: Overall Accuracy by Sessions for all 3 experiments where the x-axis denotes the session used as the test data.....39
Figure 4.3: Accuracy vs Density for Session 1 Test Data, Session 2 and 3 used for training 40
Figure 4.4: Accuracy vs Density for Session 2 Test Data, Session 1 and 3 used for training 41
Figure 4.5: Accuracy vs Density for Session 3 Test Data, Session 1 and 2 used for training 41
Figure 4.6: Overall Accuracy of 5-Fold Cross Validation Training, every fold includes 20 training and 5 test subjects 42

Figure 4.7: Box Plot of Overall Accuracies with 5-Fold Cross Validation Method.....43
Figure 4.8: Model Accuracy vs Density by Subject from 5-Fold Cross Validation Method 44
Figure 4.9: Boxplots of Accuracies for Each Session from 5-Fold Cross Validation Method 45
Figure 4.10: Model Accuracy Comparison by Task Difficulties for Each Session Training 46
Figure 4.11: Model Accuracy Comparison by Task Difficulties for 5-Fold Cross Validation Method 47
Figure 7.1: Precision, Recall, F1-Ratio for 5-Fold Cross Validation, Session 1, Session 2, and Session 3 Tests 59
Figure 7.2: Confusion Matrix for 5-Fold Cross Validation 60
Figure 7.3: Confusion Matrix for Session 1 Test (Session 2, and 3 is used for training) 61
Figure 7.4: Confusion Matrix for Session 2 Test (Session 1, and 3 is used for training) 62
Figure 7.5: Confusion Matrix for Session 3 Test (Session 1, and 2 is used for training) 63

LIST OF ABBREVIATIONS

EEG	Electroencephalography
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
MEG	Magnetoencephalography
MATB	Multi-Attitude Task Battery
CLT	Cognitive Load Theory
MWL	Mental Workload
SWAT	Subjective Workload Assessment Technique
NASATLX	NASA Task Load Index
RMANOVA	Repeated Measure Analysis of Variance
MDM	Minimum Distance to Mean
CNN	Convolutional Neural Network
PTWT	PyTorch Wavelet Toolbox
GPU	Graphics Processing Unit
CUDA	Compute Unified Device Architecture
BGR	Blue-Green-Red
RGB	Red-Green-Blue
BCE	Binary Cross Entropy
RMS	Root Mean Square

CHAPTER 1

1. INTRODUCTION

The complexity of tasks and the amount of information that individuals need to process in modern society has increased exponentially. This has made the understanding and management of cognitive workload crucial for optimizing performance and preventing errors across various fields. Cognitive workload refers to the amount of mental effort required to perform a task and is influenced by several factors, including task complexity, individual differences, and environmental conditions (Longo et al., 2022). Despite its importance, a rich, universally accepted method for gauging this "load" remains elusive. Drawing a parallel, one might liken the cognitive workload to a library that constantly receives new books, thereby challenging the capacity of the shelves over time. This metaphorical pathway, seeing the mental space as a dynamic entity with fluctuating capacity, beckons a deeper exploration to foster a more abstract and nuanced comprehension of cognitive workload. It is hoped that venturing into this philosophical territory will not only enrich understanding but also contribute to devising objective and reliable measures, filling the void in current academia and industry research.

Electroencephalogram (EEG) signals provide a non-invasive method to study cognitive processes and have been utilized in this study to classify task difficulties based on wavelet transform images of EEG signals using deep learning models. The dataset used includes EEG data from subjects performing tasks of varying difficulties. This dataset was collected and publicly released firstly as part of a competition organized in the 2021 Neuroergonomics conference (Roy et al., 2022), and later published in the Nature Scientific Data repository (Hinns et al., 2023) to cater to the need for benchmark datasets in EEG based passive brain-computer interface (pBCI) studies. This dataset features high-density EEG recordings obtained from 29 participants while they were engaged in 4 different tasks with varying difficulties over 3 sessions that are 1 week apart from each other. A unique feature of this dataset is that it involves recordings from the same participants at multiple time points, which allows researchers to test the generalizability of their pBCI models across participants and sessions. This is a particularly challenging classification problem due to non-stationary nature of EEG signals whose properties may vary within and across participants. Therefore, the development of robust classification models for EEG-based pBCI applications that can accommodate differences within sessions and across participants is an active research area in Cognitive Neuroergonomics (Roy et al., 2022).

Existing model development efforts on this dataset included Riemannian geometry, Random Forest, Convolutional Neural Net (CNN), and Recurrent Neural Net (RNN) based classifiers (Roy et al., 2022). These models were trained with the data from the first two sessions, and tested over the third session, so as to test the robustness of the classifiers for variability across sessions. The highest accuracy on the unseen third data set was achieved by the Riemannian classifier (54.26%) which is followed by spherical CNN (48.20%) and Random Forest (44.67%) approaches. Deep net approaches reached up to 90% validation accuracy, but their performance was dropped to chance level when they were tested on the third session's data, indicating overfitting issues. The success of the Riemannian and spherical CNN approaches is thought to be related to their inclusion of spatial information as well as temporal characteristics during training (Roy et al., 2022). Therefore, there is a need for better and explainable models to aid the development of robust pBCI systems.

In this thesis study, we employed wavelet transform images of EEG signals to develop and test deep learning models for mental workload classification. The EEG data were transformed into wavelet images, a crucial step for the analysis. The wavelet transform was chosen as it allows for the analysis of non-stationary signals, like EEG, by decomposing the signal into components associated with different frequency bands (Grossman & Morlet, 1985). This approach enables the capture of more information from the EEG signals but also increases the risk of including noise or irrelevant features in the model.

The study involved training and testing several deep learning models on these wavelet images under different conditions. The models' performance was evaluated based on overall accuracy and other criteria such as accuracy by subject and task difficulty. Although the highest overall accuracy achieved was promising, the performance varied significantly across subjects and task difficulties, indicating limited generalizability. This suggests difficulty in distinguishing between different levels of task difficulty. While deep learning models can achieve high accuracy in classifying task difficulty based on EEG signals, their performance varies across individuals and task difficulties. Further research is needed to improve model generalizability, optimize performance across all task difficulties, and validate the models on larger and more diverse datasets.

The rest of the thesis is organized as follows: The literature review chapter provides various definitions of cognitive workload from the literature, illustrates measurement methods of cognitive workload, explains the cognitive processes underlying performance, summarizes the findings of studies carried out with neurophysiological measurement techniques in various environments, and thoroughly investigates applications and studies in the domain of EEG and deep learning.

The methodology chapter explains the experimental protocols used in this study, describes the EEG data collection process, the process of transforming the EEG data into wavelet images, and introduces the deep learning models used in the study, including the scientific principles behind them and how they were configured and tested.

The results chapter reports graphs and tables derived from the analyses explained in the method chapter and details the performance of each model, including overall accuracy and accuracy by subject and task difficulty. The effects of different model configurations and optimizers are also discussed.

The discussion and conclusion chapter discusses the results derived from the methods chapter, explains the selection of used inputs, compares the algorithms in terms of their classification accuracy, evaluates the advantages and disadvantages of each approach, handles parameter tunings, and details the tools, hardware, and software environment for analyses and running of algorithms. Finally, this chapter emphasizes crucial findings, concludes with the limitations of the thesis, and outlines possible future works.

Supplementary materials complementing the analyses and the data collection process are presented in the appendices. This includes detailed results for each model and configuration, model input combinations versus accuracy scores for each algorithm used in all analyses, and graphs presented in the results chapter derived from these tables.

CHAPTER 2

2 LITERATURE REVIEW

2.1 Introduction to Mental Workload (MWL)

2.1.1 Definition of MWL

Mental workload (MWL) is a multifaceted concept that encompasses the mental effort required to perform a task. It is influenced by various cognitive, perceptual, and psychomotor processes. Theoretical and operational definitions of MWL have evolved over time, and several perspectives on MWL exist, reflecting its complexity (Longo et al., 2022)

2.1.2 Importance of Managing MWL in Human-Computer Interaction

In the context of human-computer interaction, managing MWL is a primary goal from a human factors' perspective. The optimization of user performance, enhancement of user engagement, and minimization of errors are key motivations for managing MWL. As all human activities involve some degree of mental processing, even basic physical or cognitive tasks result in a certain level of MWL (Mitchell, 2000; Longo, 2011; Longo et al., 2012).

2.1.3 Motivations for Managing MWL: optimizing user performance, enhancing user engagement, minimizing errors

Over the past two decades, technological advancements have transformed human-computer interaction, reducing the physical load on human operators while altering the nature and quantity of cognitive processing required. The ultimate goal of these advancements has been to reduce and/or regulate the human operator's MWL. The focus has been on regulating the associated cognitive, visual, auditory, perceptual, psychomotor, and communication contributors to workload (Hancock & Chignell, 1988; Myers, 1998; Miller, 2001; Longo, 2015).

2.2 Theoretical Background of MWL

2.2.1 Key Theories and Models Related to MWL

Various theories and models have been developed to understand and conceptualize mental workload (MWL). Among the most prominent ones are:

- *Cognitive Load Theory (CLT):* CLT suggests that human cognitive capacity is limited in its working memory. To optimize learning and task performance, it's crucial that cognitive load aligns with these limits. Therefore, tasks and

instructional designs should be structured to match our cognitive architecture, promoting effective learning and task execution (Sweller, 1988).

- *Multiple Resources Theory:* This theory posits that the human cognitive system operates using multiple, distinct resources. These resources span various sensory modalities, processing stages, and types of codes. For example, tasks in different sensory modalities, like visual and auditory, might not compete for the same cognitive resources, allowing for simultaneous execution with minimal interference. In contrast, tasks within the same modality might compete for resources, leading to interference (Wickens, 1984; Wickens, 2002).

2.2.2 Factors Influencing MWL

MWL experience can be shaped by multiple factors. These include the inherent challenge of a task, the expertise and skills of the individual, circadian rhythm variations, the task environment, and available tools or aids. External factors, such as interruptions, distractions, and time pressure, also play a pivotal role in determining MWL (Longo et al., 2022).

2.3 Measurement of MWL

2.3.1 Importance of Measuring MWL

Measuring mental workload (MWL) is crucial for several reasons. It not only aids in understanding the cognitive demands placed on an individual during task performance but also offers insights into optimizing human-machine interactions. By gauging MWL, designers and researchers can refine tools, systems, and environments to maximize user performance, engagement, and safety (Wickens, 2008).

2.3.2 Overview of Measurement Methods

Various methodologies have been devised to assess MWL, each offering unique insights and catering to different research and practical needs:

- *Physiological Measures:* These measures gauge the physiological responses of individuals as they perform tasks. Common physiological indicators include heart rate variability, skin conductance, pupillary dilation, and brain activity patterns measured using EEG. These measures offer objective insights into the cognitive demands of tasks, with tools like EEG providing granular data on brain activity and cognitive load (Anderson et al., 2011).

- *Performance Measures:* Performance metrics evaluate how well individuals carry out specific tasks. Measures might include task completion time, accuracy rates, error rates, and other task-specific metrics. These measures directly reflect the

impact of MWL on task execution, with increased MWL often leading to decreased performance (Splawn & Miller, 2013).

- *Subjective Measures:* Subjective methodologies, such as the Subjective Workload Assessment Technique (SWAT) and NASA Task Load Index (NASA-TLX) designed by Hart and Staveland (1988), are employed to gauge an individual's self-reported perception of MWL. SWAT, a widely utilized rating scale measurement method, operates in two phases: initially, scales with predefined attributes are established through experimenter training. Subsequently, these experimenters are tasked with evaluating both task difficulty and their performance. On the other hand, tools like NASA-TLX provide a structured approach for participants to rate their perceived workload across multiple dimensions. Although these measures stem from personal perceptions and can be affected by a variety of factors, they offer invaluable insights into the user's experience and perception of workload (Hoonakker et al., 2011; Vural, 2018).

2.4 EEG as a Tool for Measuring MWL

2.4.1 Introduction to EEG

Electroencephalography (EEG) is a non-invasive method used to record electrical activity in the brain. Electrodes placed on the scalp detect fluctuations in voltage, representing brain activity. Due to its high temporal resolution, EEG is an invaluable tool for studying cognitive processes in real-time, making it suitable for assessing mental workload (MWL) during tasks (Makeig et al., 2004).

2.4.2 EEG frequency bands and their significance

EEG data is typically categorized into different frequency bands, each linked to various cognitive and physiological states:

- Delta (1-4 Hz): Often associated with deep sleep. (Klimesch, 2012)
- Theta (4-8 Hz): Related to drowsiness, relaxation, and introspection. (Jann et al., 2010)
- Alpha (8-14 Hz): Indicates relaxation and calmness, often observed when eyes are closed. (Jann et al., 2010)
- Beta (14-30 Hz): Associated with active, analytical thought and alertness. (Uhlhaas & Singer 2010)
- Gamma (30-100 Hz): Connected to higher cognitive functions and information processing. (Engel, A. et al., 2001, Uhlhaas & Singer 2010)

These bands provide insights into the cognitive state and workload of an individual. For instance, increased beta activity might indicate heightened alertness or cognitive processing (Wilson & Russell, 2003).

2.4.3 Advantages and limitations of using EEG to measure MWL

Delorme & Makeig (2004) stated the following advantages and disadvantages of EEG-based MWL assessment:

- *Advantages:*
    - High temporal resolution: Enables the study of dynamic changes in MWL in real-time.
    - Non-invasiveness: Subjects don't undergo any invasive procedures, ensuring comfort and ease.
    - Direct measurement of brain activity: Provides a direct window into cognitive processes.
- *Limitations:*
    - Limited spatial resolution: Determining the exact origin of brain activity can be challenging.
    - Susceptibility to artifacts: External factors like muscle movements can interfere with readings.
    - Requires careful setup and calibration: Ensuring accurate readings demands meticulous electrode placement and calibration.

Table 2.1 below compares EEG to other well-known neuroimaging modalities like NIR, fMRI and PET in terms of their temporal and spatial sensitivity (Coyle, 2005). Although fMRI and PET can provide superior spatial resolution, the need for confining the participant in a laid-down position, the need for injecting radioactive tracer chemicals in PET, and the cost of the equipment limit their applicability in BCI settings. NIR is the only other viable option given its portability and affordability. Another key distinction between EEG and the abovementioned neuroimaging modalities is that, NIR/PET/fMRI all monitor brain activity indirectly through a phenomenon called hemodynamic response, which involves the response of the vascular system for supplying oxygenated blood to spiking neuron populations. This response is delayed 4-6 seconds following the initial spike burst, which limits the temporal sensitivity of these approaches.

In contrast, because EEG can detect electric potential changes due to spiking activity, it can instantly pick up such changes in neural activity. However, the complexities involved in the propagation of electrical activity inside the brain tissue and the weak nature of these signals limit the spatial sensitivity of the EEG approach. The strongest contributors of

EEG signals are considered to be the vertically aligned pyramidal neurons that are present in the cortex close to the scalp, whose electrical discharges get combined due to their alignment, which can be picked up by the electrodes placed over the scalp (Luck, 2004).

Table 2.1: Spatial and Temporal Sensitivity Comparison of BCI Techniques (Coyle, 2005, p.35)

Method	Spatial Sensitivity	Temporal Sensitivity
EEG	Around 1cm	milliseconds
PET	4-6mm	more than 10 seconds
fMRI	2mm	more than 1 second
NIR	Around 1cm	more than 1 second

2.5 Wavelet Transform for EEG Signal Processing

2.5.1 Introduction to wavelet transform

Wavelet transform is a mathematical methodology that breaks down signals into various frequency elements, examining each with a resolution tailored to its particular scale. Distinct from the Fourier transform, which supplies solely frequency data, the wavelet transform delivers insights into both time and frequency domains. This dual capability is especially advantageous for evaluating non-stationary signals, in which frequency components may evolve over time. By employing the wavelet transform, one can identify specific regions in time-frequency space where two time series exhibit significant shared power and maintain a consistent phase correlation. (Torrence & Compo, 1998, Küskü 2022).

2.5.2 Application of wavelet transform to EEG data

When applied to EEG data, the wavelet transform allows researchers to analyze changes in the EEG signal over time across different frequency bands. By decomposing the EEG signal using wavelets, researchers can pinpoint specific events or states in the EEG data and analyze their frequency content. This decomposition is especially useful for identifying transient events or non-stationary patterns in the EEG, which might correspond to specific cognitive processes or responses (Zhiwen et al., 2019).

2.5.3 Advantages of using wavelet transform for EEG data analysis

- *Time-Frequency Analysis:* Wavelet transform provides a time-frequency representation of the EEG signal, enabling the detailed study of how the frequency content of the signal changes over time (Mallat, 1989; Murugappan et al., 2010).

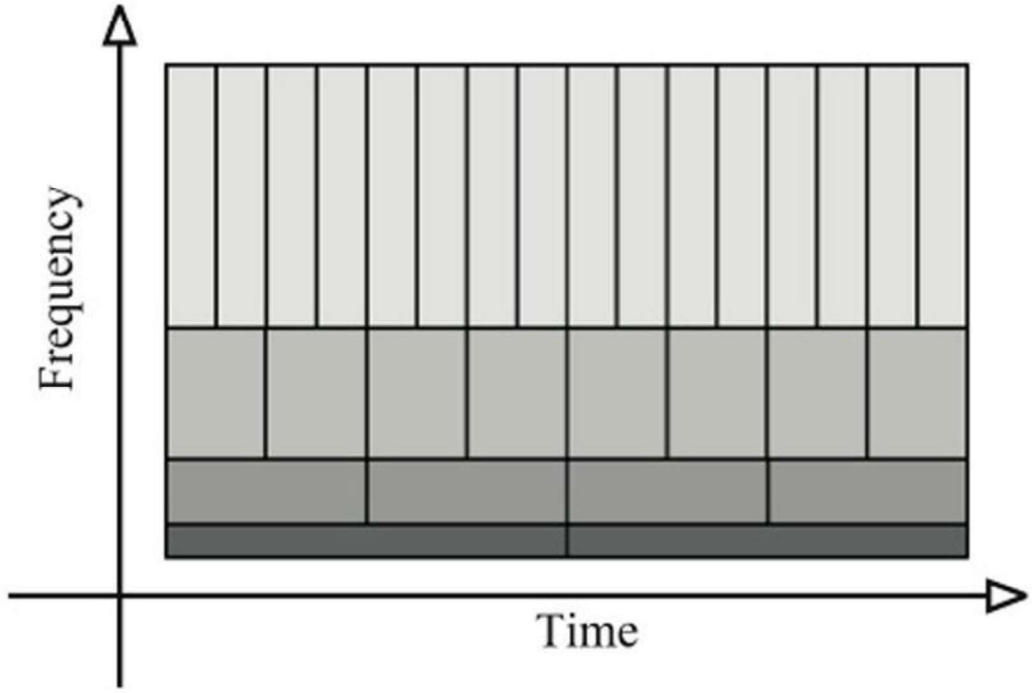

Figure 2.1: Tiling of the Time-Frequency Graph for the Wavelet Transform (Issartel et al., 2015)

- *Enhanced Signal Detection:* The complexity of EEG data often involves transient or non-stationary events, which can easily be masked or overlooked when using conventional frequency-based analyses. Wavelet transform, as elucidated by Mallat (1989), offers a refined approach that delves deeper into these complexities. By employing wavelet transform, researchers and analysts can effectively tease out these transient events, ensuring that even subtle, short-lived changes in the EEG signal are captured and analyzed. This enhanced detection capability of the wavelet transform provides a more holistic view of EEG data, ensuring that no critical information is missed during analysis.
- *Flexibility:* The wavelet transform offers a selection of wavelet functions tailored to EEG data characteristics. As highlighted by L. Senhadji et al. (1995), the choice of wavelet function can markedly influence analysis outcomes. Different wavelet functions accentuate various aspects of the EEG signal. This flexibility allows researchers to strategically choose functions that best align with specific EEG data traits, optimizing analysis and enhancing result accuracy.
- *Noise Reduction:* Wavelet transform, as highlighted by Rosanne, O., et al. (2021), offers a powerful solution for reducing noise in EEG recordings. Given the various interferences in EEG data, the wavelet transform's capabilities enable effective filtering, preserving core EEG components. This ensures clearer data quality, facilitating more accurate subsequent analyses and interpretations.

2.6 Previous Studies Using EEG and Wavelet Transform for MWL Assessment

2.6.1 Wavelet Coefficients in EEG Analysis for Cognitive Load Classification

In a study by Zarjam et al. (2015), the researchers utilized wavelet coefficients as a central feature in EEG signal analysis. The data they worked with was collected from 32 EEG channels, recorded at a 256Hz sampling rate, focusing on the 0.5 - 30Hz frequency interval. Their experimental task spanned seven levels of difficulty: Level 1 involved the summation of two 1-digit numbers, while Level 7 challenged participants with the addition of two 3-digit numbers.

Traditionally, the technique of using wavelet coefficients has been associated with pathological diagnoses. However, in this particular study, the authors ventured into the domain of cognitive load and mental task classification, marking a new trajectory in the discipline. Though there have been previous endeavors into the use of wavelet coefficients for niche EEG tasks, such as artifact removal and distinguishing between variable load levels, Zarjam et al. (2015) demonstrated the capabilities of wavelet features in EEG-based load classification.

Moreover, they introduced an innovative approach to cognitive load classification, leveraging the entropy, energy, and standard deviation of the wavelet coefficients. This methodology surpassed conventional methods like self-ratings and certain spectral-based features, registering an outstanding 98% detection accuracy in segregating the seven cognitive load levels across a spectrum of subjects. This, compared to the 31% classification accuracy of traditional self-rating methods, underscores the transformative capabilities of wavelet transform techniques in refining EEG data analysis.

2.6.2 Emotion Recognition through EEG Signals using Wavelet Transform

In a study by Murugappan et al. (2010) on human emotion recognition, EEG signals were captured from 20 subjects using 62 channels, sampled at a 256Hz rate and spanning a frequency interval of 0.5 - 70Hz. The signals were processed to induce distinct emotions such as disgust, happiness, surprise, fear, and neutrality. The "db4" wavelet function was employed for feature extraction from the EEG data via the Discrete Wavelet Transform (DWT), subsequently segregating the signals into alpha, beta, and gamma frequency bands. These extracted features were inputted into two classifiers: K Nearest Neighbor (KNN) and Linear Discriminant Analysis (LDA). When comparing the classifiers, it was evident that the proposed feature, ALREE, achieved an average classification accuracy of 83.26% with KNN and 75.21% with LDA. This highlights the effectiveness of wavelet transform features in the realm of emotion classification.

2.7 Cross-Session Variability in pBCI: Insights from the First pBCI Competition on Workload Estimation

The paper by Roy et al. (2022) provides a comprehensive overview of the organization, results, and insights gained from the first-ever pBCI competition focusing on cross-session workload estimation. This competition was a part of the 3rd International Neuroergonomics conference and aimed to address a significant challenge in the field of Brain-Computer Interfaces (BCI), specifically the variability of brain signals across different sessions. The dataset used for the competition, provided by Hinss et al. (2021), included electroencephalographic (EEG) recordings from 15 volunteers who performed the Multi-Attribute Task Battery-II (MATB-II) across three sessions, each separated by a week. The MATB-II is a well-known task developed by NASA to assess task-switching and mental workload capacities.

The competition attracted eleven teams from three continents, who submitted their work for evaluation. The results revealed a range of classification accuracies, with the best-performing algorithm achieving an accuracy of just under 60%, well above the chance level of 38% (adjusted for a 3-class classification problem). Session 1 and 2 are used as the training and validation, and session 3 is used as the test data. The winning solution utilized Riemannian geometry principles and an automatic per-subject electrode selection process to maximize the Riemannian distance between class-conditional covariance matrices. This approach highlighted the effectiveness of Riemannian classifiers for BCI applications, as three out of the four best scores were obtained using Riemannian geometry.

On the other hand, deep learning methods, despite showing promise in the validation phase, encountered generalization issues and did not outperform traditional methods. Specifically, three out of four deep learning methods performed at chance level on the test set, indicating a significant overfitting problem. This outcome underscores the need for careful design and training procedures for deep learning models applied to small datasets, such as those typically encountered in BCI research.

Overall, the competition marked a crucial step towards addressing BCI variability and promoting good research practices, including reproducibility. However, the results also highlighted the need for further research, algorithm development, benchmarks, and database collections to tackle the various sources of variability affecting BCIs, such as cross-subject, cross-context, and cross-task variabilities. The authors hope that the competition will stimulate more studies and efforts to develop practical pBCI applications that can be effectively and efficiently used across multiple sessions without the need for recalibration.

2.8 COG-BCI Database Usability Validation for Mental Workload Estimation

Hinss and colleagues (2023) undertook an extensive investigation centered on mental workload assessment using the MATB and N-Back tasks, each having three distinct levels of mental workload (MATB Easy, MATB Medium, MATB Difficult; and 0-back, 1-back, 2-back). A uniform preprocessing, feature extraction, and classification pipeline were applied across both tasks. The EEG data were filtered in the theta or alpha band, channels with a standard deviation more than two standard deviations larger than the other channels were interpolated, and an independent component analysis (ICA) with IClabel component rejection was conducted. The covariance matrix for a subset of 10 channels was computed for each epoch, and the data were partitioned into training and testing sets using a 5-fold cross-validation approach. The Riemannian Minimum Distance to Mean (MDM) classifier was then used for training and testing the data. The classifier's performance was assessed using a RMANOVA with session, difficulty, and bandwidth (theta vs. alpha) as factors.

The findings indicated that the Riemannian MDM classifier successfully identified different mental workload levels for both the MATB and N-Back tasks, achieving an accuracy of 69.40% (±12.50%) for the MATB task and 64.97% (±12.99%) for the N-Back task. Notably, the alpha band yielded significantly higher accuracies than the theta band. Accuracies also differed across sessions, with the second session recording the highest accuracy (Hinss et al. 2023).

Moreover, the researchers conducted a three-tiered validation of the gathered dataset, encompassing subjective mental workload and vigilance decrement, behavioral performance, and physiological alterations (Hinss et al. 2023). The outcomes affirmed the vigilance decrement, disparities in mental workload at the subjective, behavioral, and physiological levels, and the cerebral activity's sensitivity to errors, sessions, and congruency. The findings also verified that the COG-BCI database could be instrumental in devising and testing classification pipelines, despite the analysis not encompassing a comprehensive examination of all potential effects within the data or all possible pipelines for estimating users' cognitive states.

This research enriches the existing body of knowledge by confirming the collected dataset's suitability for passive BCI research and promoting open science practices in pBCI research and development. The employment of the Riemannian MDM classifier and the multi-level dataset validation are innovative methods that enhance the mental workload assessment and passive BCI research domains. The current thesis study aims to contribute to this line of work by exploring the utility of CNN-based mental workload classifiers trained over wavelet transformed representations of EEG signals that aim to capture time-frequency dynamics induced by changes in mental workload.

CHAPTER 3

3 METHOD

3.1 Dataset

The COG-BCI database comprises EEG and ECG recordings designed to investigate various cognitive states. Data from 29 participants have been collected, with each participant undergoing 3 sessions across 4 distinct tasks. These tasks have been crafted to elicit diverse cognitive responses, thereby enhancing the range of cognitive states represented in the dataset.

A sampling rate of 500Hz was utilized, ensuring the precision and granularity of the data. The EEG data is characterized by recordings from 62 electrode channels, offering comprehensive coverage of cerebral activity. An ECG channel has also been included, providing a perspective on cardiac activity, which is considered valuable for interpreting certain cognitive states. Depending on the tasks administered, the duration of individual tests was set at either 150 or 300 second. This rigorous data collection methodology has resulted in over 100 hours of publicly available data.

It should be noted that comprehensive validation procedures have been applied to the COG-BCI database. Validation at subjective, behavioral, and physiological levels was conducted, encompassing both cardiac and cerebral activities, affirming the dataset's relevance to the pBCI community.

Approval for this project was granted by the University of Toulouse's ethical committee (CER number 2021-342). By providing public access to this dataset, a significant contribution to the principles of open science has been made, supporting the broader adoption, and understanding of pBCI research.

3.1.1 Data Collection

In the experiment conducted by Hinss et al. (2023), an EEG system with 64 active electrodes and an amplifier, configured according to the extended 10-20 system, was used. Data was recorded at a frequency of 500 Hz, however, the electrode Cz was not available for the initial nine participants. The tenth electrode, denoted as ECG in the dataset, recorded the electrocardiographic activity and was positioned on the left fifth intercostal space. The synchronization of the stimulus display, physiological data, and participants' responses was managed by specific software. Furthermore, a 3D camera and a specialized plug-in were used to accurately determine the electrode positions on the scalp. The channel locations for each participant were meticulously documented and stored in a designated folder.

3.1.2 Experimental Tasks

- *Psychomotor Vigilance Task (PVT)*

The Psychomotor Vigilance Task is a 10-minute vigilance test where participants press the spacebar when a timer shows on the screen. Each trial begins with a variable interstimulus interval (ISI) ranging from 2-10 seconds, followed by the timer. The timer halts upon pressing the spacebar, displaying the reaction time for 500 ms. This task mirrors the PC-PVT 2.0 computer version and involves 90 trials per session.

- *N-Back*

The N-Back Task is a computer-based test that measures working memory and mental workload. Participants view numbers on a screen and press a button if the current number matches the one presented N steps back, with increasing N signifying greater difficulty. Each number displays 500 ms, followed by a 1,500 ms blank screen. In the 0-back version, participants press when a specific target number (e.g., "3") appears. All conditions have a 1/3 hit rate (16 hits in 48 trials per block). In the 2-back version, five immediate repeat numbers are added as conflict trials, meant to challenge participants without affecting performance. Participants tackle three blocks each of 0-back, 1-back, and 2-back tasks, lasting roughly 2 minutes each. Thus, each condition takes about 6 minutes for three blocks. Participants are informed about the upcoming block's condition and given brief instructions.

- *MATB-II*

NASA's MATB-II task assesses task-switching and mental workload, presenting participants with up to 6 simultaneous tasks, replicating realistic operational systems with adjustable difficulty. An adapted version, coded in MATLAB, was employed in this study. Participants completed combinations of four subtasks (Fig. 3.1):

- **Tracking (TRACK)**: Participants use a joystick to keep a moving target within a window, with adjustable difficulty based on target speed and movement.
- **System Monitoring (SYSMON)**: Participants monitor gauges and lights, responding to specific signals using keyboard commands. The number of events determines difficulty.
- **Communications (COMM)**: Participants listen to radio messages, acting only on relevant ones by changing the radio frequency.
- **Resource Management (RESMAN)**: Participants maintain fluid levels in tanks by managing interconnected pumps, with added challenges like pump failures.

In this study, participants had three 5-minute runs at different difficulty levels:

- **Easy**: SYSMON and TRACK tasks.

- **Medium**: Adds the RESMAN task.
- **Difficult**: Introduces the COMM task and increases the TRACK task's challenge. Each run began with brief instructions.

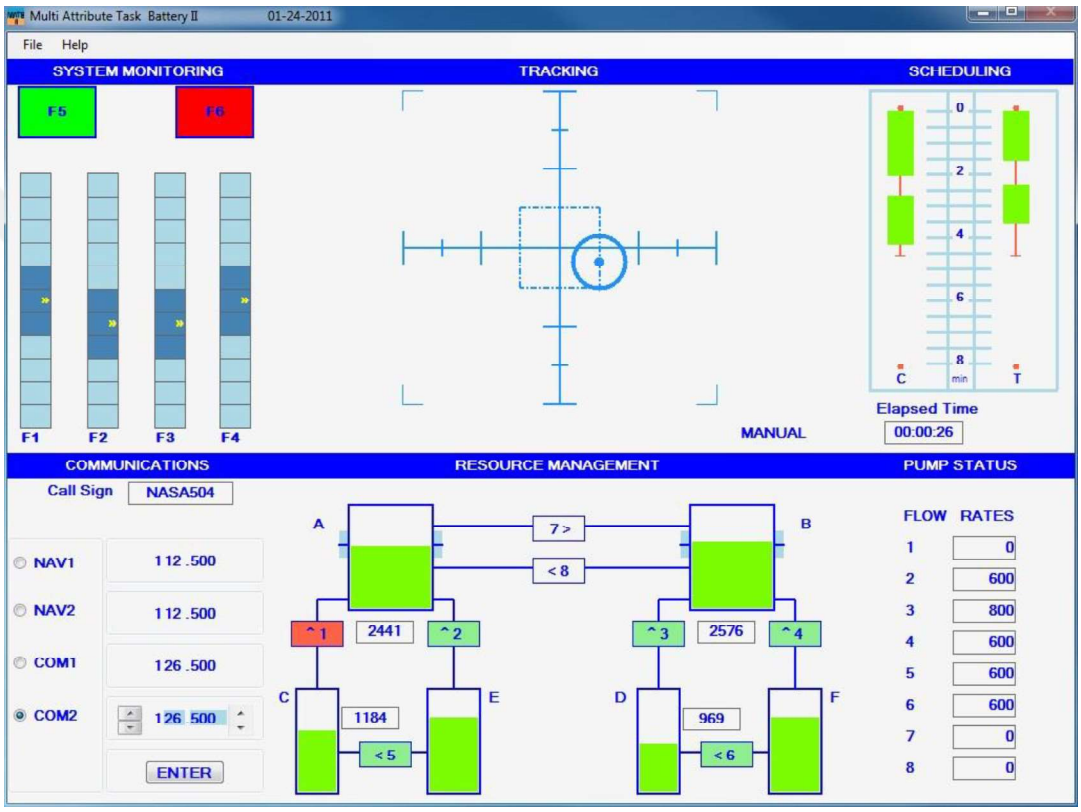

Figure 3.1: MATB-II Task Interface https://matb.larc.nasa.gov

- *Flanker*

The Flanker task is a reaction test that introduces conflict in a binary decision. Participants view 5 arrows on a screen, responding to the central arrow while ignoring the flanking arrows. The flankers can be congruent (same direction as the center) or incongruent (opposite direction). Examples of stimuli are '< < > < <' and '< < < < <', with respective responses '>' and '<'. Trials start with a 2000 ms interstimulus interval (ISI), followed by the stimulus lasting 16 ms (determined through a pilot study). Post-stimulus, a blank screen appears for 2250-2750 ms for the participant's response. Feedback is provided for 500 ms, indicating the trial's outcome.

3.2 Dataset Preprocessing

In cognitive science research, particularly when working with intricate datasets like those obtained from Electroencephalography (EEG), the preprocessing of data stands as a quintessential step. The overarching importance of preprocessing lies in its capacity to refine raw data, transforming it into a format that is both analytically amenable and representative of the underlying cognitive phenomena. This step ensures the removal of potential confounders that might obscure genuine patterns or induce spurious results. The primary goals encompass the mitigation of noise and external interferences, standardization of the dataset for uniformity across sessions and subjects, and the diligent identification and rectification of artifacts or aberrant signals. These objectives aim to provide a clean slate, setting the stage for accurate and meaningful subsequent analyses.

As we delve deeper into the specifics of EEG data preprocessing techniques in section 3.2.1, the rationales behind each chosen method, from bandpass filtering to the removal of noise and artifacts, will be elucidated. Furthermore, the tools and libraries, both from Python and MATLAB environments, which facilitated these preprocessing endeavors, will be detailed in section 3.2.2, underscoring their functionalities and applications.

3.2.1 EEG Data Preprocessing Techniques

Bandpass Filtering: EEG data were subjected to a bandpass filter spanning a frequency range of 1-31 Hz. This range was chosen based on established literature to focus on significant frequency bands, specifically the delta (1-4 Hz), alpha (8-13 Hz), beta (13-30 Hz), and lower gamma (up to approximately 31 Hz) bands, each associated with various cognitive processes. This filtering aimed to eliminate extraneous frequencies, thereby improving the signal-to-noise ratio and ensuring a consistent dataset for subsequent analyses.

Removal of Artifacts and Noise: The raw EEG recordings provided in the dataset had already undergone initial cleaning by the dataset providers. Subsequently, when a wavelet transformation was applied to this cleaned data, certain artifacts were observed, particularly in regions where wavelet values exceeded an upper threshold, resulting in pronounced black areas. To address this, a systematic approach was implemented: images containing more than 15% black area were removed from the dataset. This decision was predicated on the understanding that such prominent artifacts could significantly alter the features of the overall dataset and potentially influence subsequent analyses. This removal strategy was primarily based on visual inspection, and a more in-depth metric evaluation was not conducted at this stage.

3.2.2 Libraries and Software Tools

3.2.2.1 Python MNE (Magnetoencephalography (MEG) and Electroencephalography (EEG))

In the present research, the MNE software package was employed for the specialized processing and analysis of EEG and MEG data. MNE facilitated the extraction of visual

representations, such as: 2D visualizations of EEG electrode locations (as depicted in Fig 3.2), 3D spatial configurations of the electrodes (illustrated in Fig. 3.3), and individual EEG signal traces for each respective channel (presented in Fig. 3.4).

Furthermore, the MNE library was instrumental in parsing raw EEG data, specifically extracting data from the ".set" files associated with each experimental session.

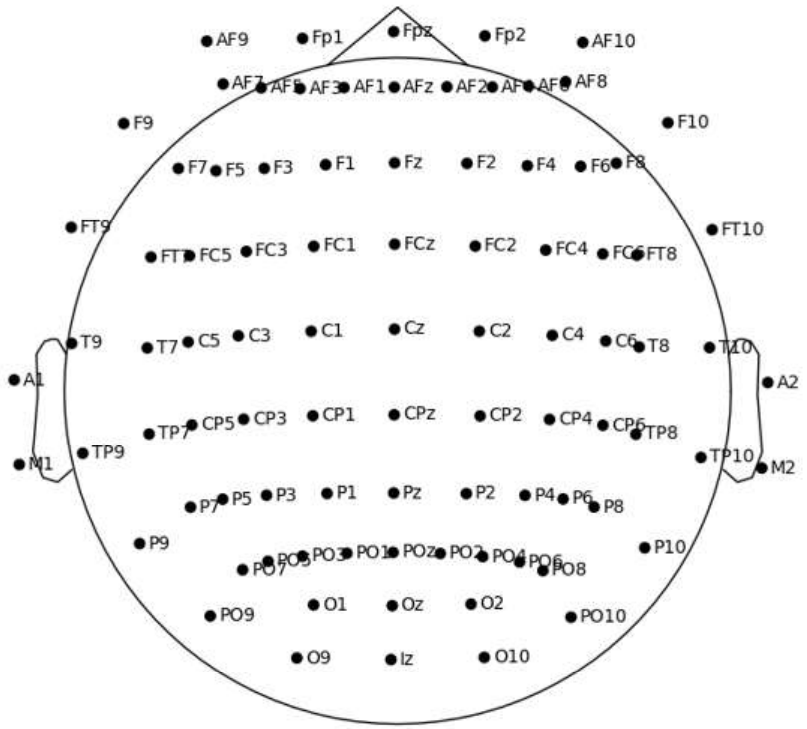

Figure 3.2: 2D Locations of 62 Electrodes. Visualized with MNE library.

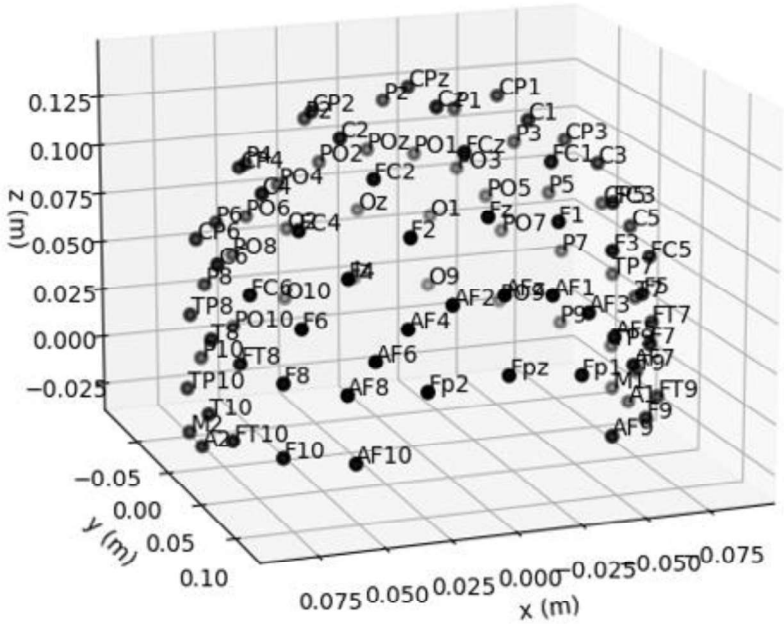

Figure 3.3: 3D Electrode Locations. Visualized with MNE library

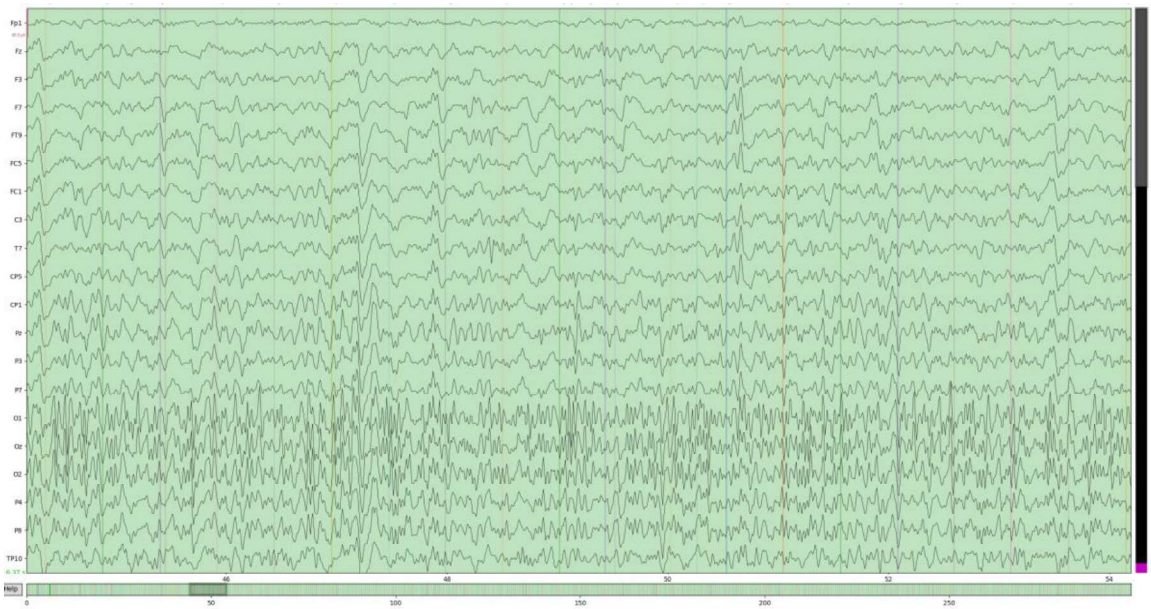

Figure 3.4: Example EEG channels' Voltage vs Time Graph from Subject 01, Session 03, MATBeasy Task, 44 to 54 second interval from 300 seconds of data Extracted with Python MNE library.

3.2.2.2 Python PyWavelets(pywt)

For the wavelet-based analyses in this study, the PyWavelets library was utilized. This Python library is specifically dedicated to facilitating wavelet transformations, supporting both one-dimensional and multi-dimensional discrete wavelet transforms. PyWavelets offers an extensive collection of built-in wavelet filters and also provides the flexibility for users to define custom wavelet filters tailored to specific requirements. It is important to note that PyWavelets, in its design, does not natively offer GPU acceleration capabilities.

3.2.2.2 Python PTWT(Parallel Tensor Wavelet Transform)

In the domain of multi-resolution analysis for tensor data, the PTWT library was employed. This tool is predominantly tailored for tasks related to image and video processing. With its specialization in tensor wavelet transforms, PTWT is optimally designed to handle multi-dimensional data, notably images and videos. While PTWT has been architected with a focus on parallel processing, it also harnesses the capabilities of CUDA to accelerate processes, thereby leveraging GPU resources for enhanced computational efficiency.

3.2.2.4 Python Ssquuezepy

In this study, the ssqueezepy library was utilized for its capabilities in signal processing, particularly its expertise in the synchrosqueezed wavelet transform. Synchrosqueezing, as a reassignment method, offers enhanced clarity in time-frequency representations. The library proved advantageous as it yielded sharper time-frequency results compared to some traditional wavelet transform approaches. Importantly, ssqueezepy is equipped with CUDA support, allowing for GPU acceleration. This feature, combined with its extensive wavelet parameter customization options and superior visualization tools, made it a valuable asset in our data analysis process.

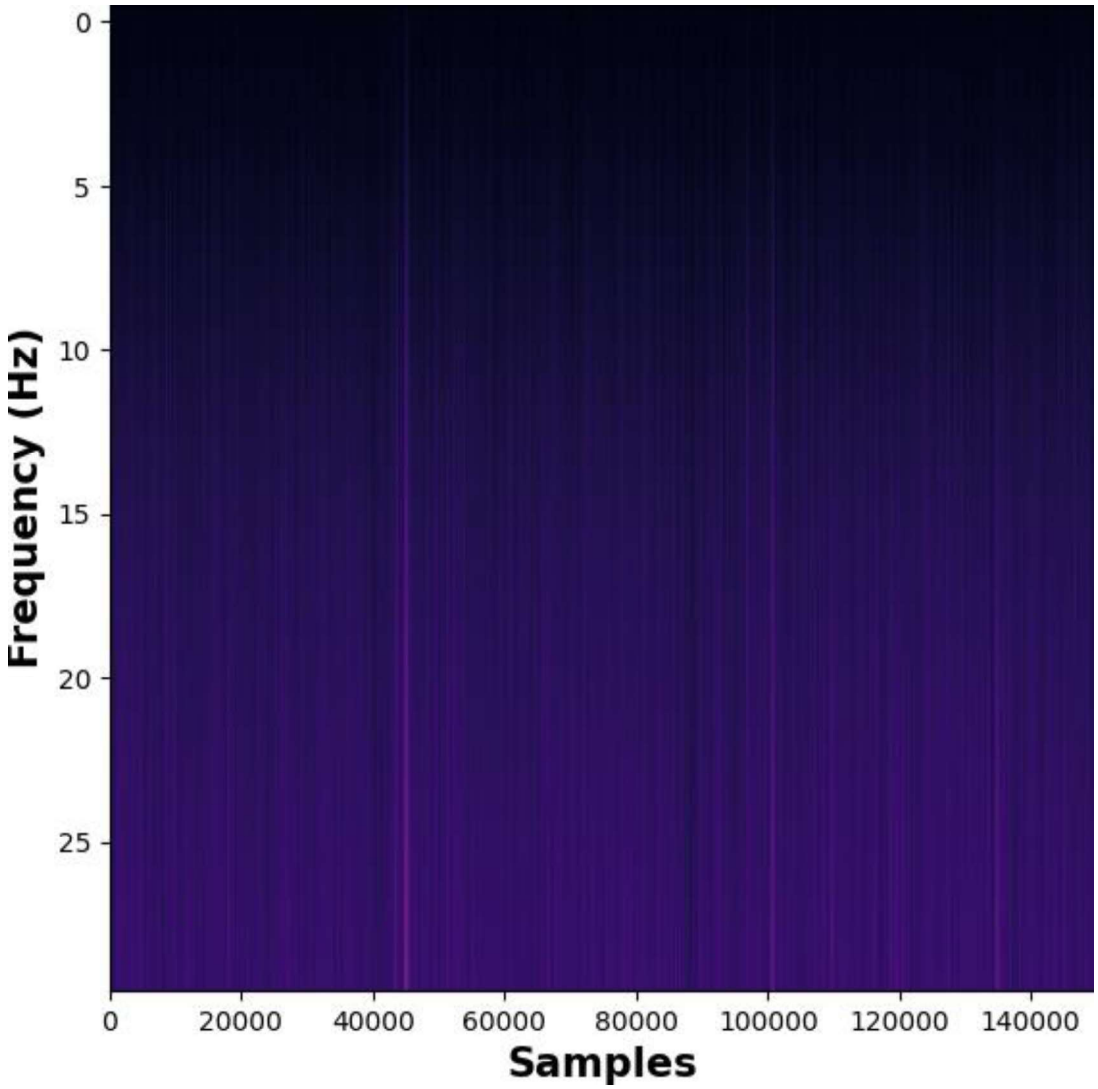

Figure 3.5: Wavelet Output using Squuezepy Subject 10, Session 3, MATB Easy Task, F4 Channel

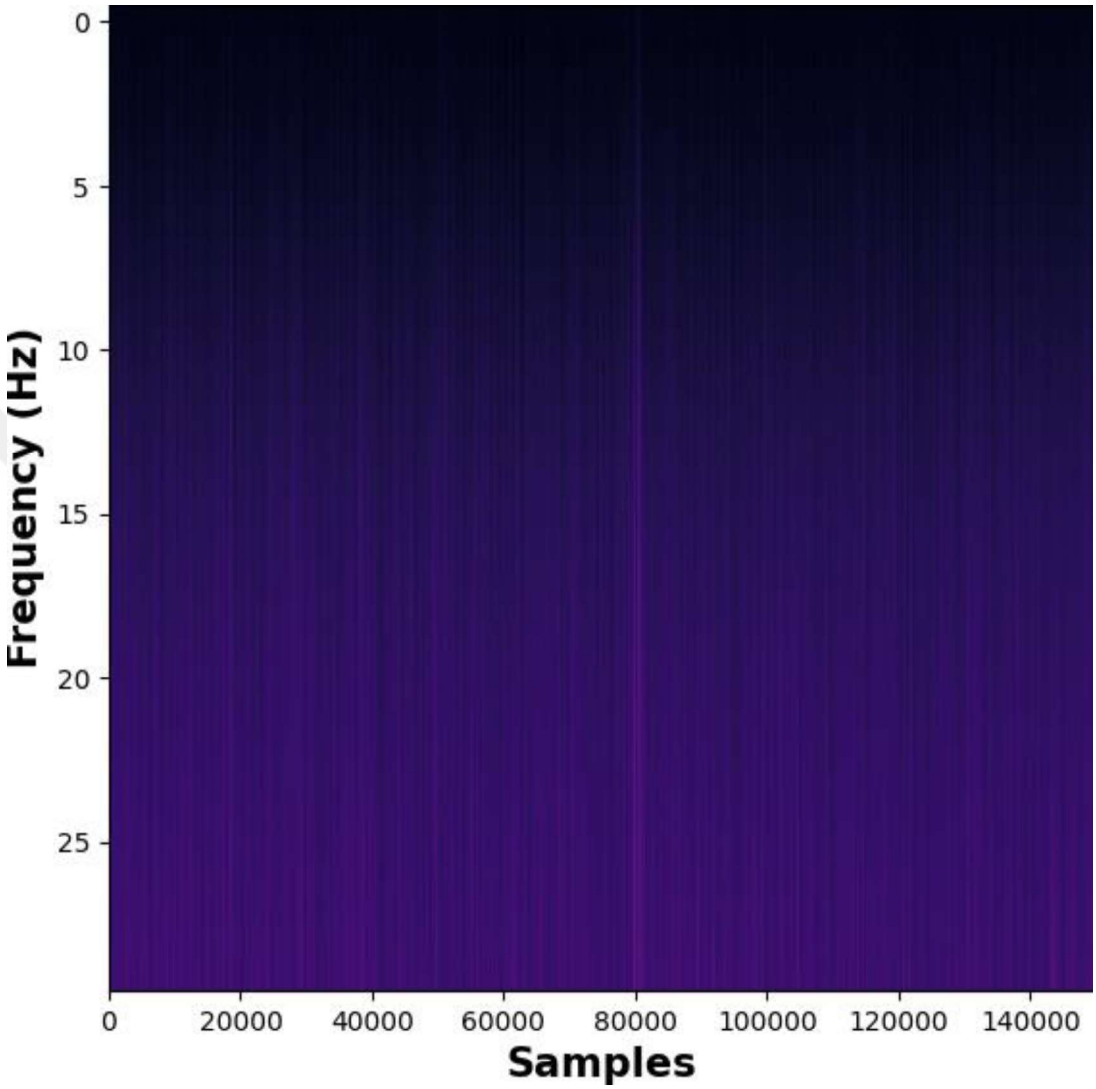

Figure 3.6: Wavelet Output using Squuezepy for Subject 10, Session 3, MATB Medium Task, F4 Channel

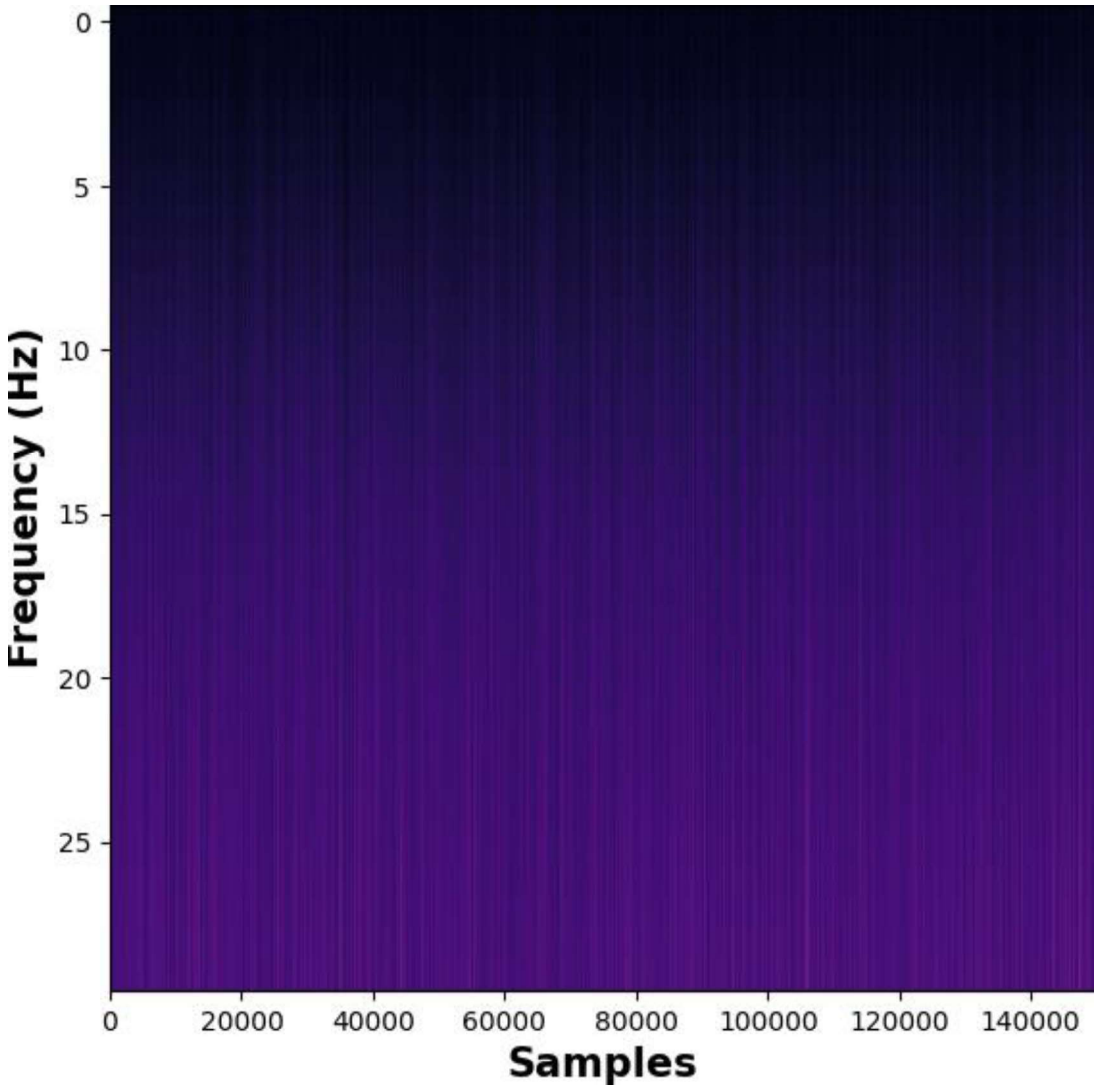

Figure 3.7: Wavelet Output using Squuezepy for Subject 10, Session 3, MATB Difficult Task, F4 Channel

In Figure 3.5, Figure 3.6, and Figure 3.7, it can be observed that the wavelet transform outputs for different task difficulties - easy, medium, and difficult - are quite similar in appearance and lack sharp distinctions. This indicates that the wavelet transform does not yield distinct or clearly separable features for different levels of task difficulty, which might pose challenges in accurately classifying the mental workload levels solely based on these outputs.

Additionally, the results obtained from PyWavelets (PyWT) and PyTorch Wavelet Toolbox (PTWT) did not exhibit any significant improvement over the SqueezePy implementation, leading to a similar lack of distinct features for different levels of task difficulty. Consequently, this prompted a transition to the MATLAB EEGLAB toolkit,

which is a more comprehensive and specialized tool for EEG data analysis, with the expectation that it might provide more discriminative features and ultimately yield better classification performance.

3.2.2.5 MATLAB EEGLAB Toolkit

In the preprocessing of the EEG data, EEGLAB, a popular MATLAB toolbox for electrophysiological data analysis, was utilized for extracting EEG data from the .set files, a format commonly used in EEGLAB. Each .set file was read into MATLAB, and the wavelet transform was subsequently performed using a custom MATLAB function. This `wt` function, part of Grindset et al.'s (2002) MATLAB library, was employed to perform the continuous wavelet transform on the time series data of each channel. Default parameters for the transform, such as padding the time series with zeros, were used, with the number of sub-octaves per octave set to 1/12 and the minimum scale set to twice the sampling interval. The mother wavelet was chosen as 'Morlet,' as it is one of the most used function types in EEG data (NK Al-Qazzaz et al. 2015), and the AR1 coefficient of the time series was automatically estimated.

The wavelet transform was executed on the input time series, and the wavelet power spectrum was calculated by taking the square of the absolute value of the wavelet transform. Significance levels for the power spectrum were determined, and the results were displayed in a figure, with contour lines indicating the significance levels and a color bar representing the power values.

While other platforms such as MNE, PyWavelets, PTWT, and ssqueezepy were considered, the combination of EEGLAB for data extraction and the `wt` function for wavelet transform delivered visually sharper wavelet outputs, providing a more streamlined and effective analysis process. This analysis approach significantly contributed to the research objectives and provided a robust foundation for the subsequent stages of the study.

As seen in Figure 3.8, 3.9, 3.10, the wavelet images have a period range spanning from 0 to 32,768 units and a time range extending up to 150,000 samples. Using the device's sampling rate of 500 Hz, our primary interest was the frequency range of 1 to 31 Hz. When translated into period units, the 1 Hz frequency corresponds to a period of approximately 500 units (or 1 second), while the 31 Hz frequency translates to a period close to 16 units (or 0.032 seconds). This specific period range on the wavelet image encapsulates the critical EEG signal features pivotal to our research.

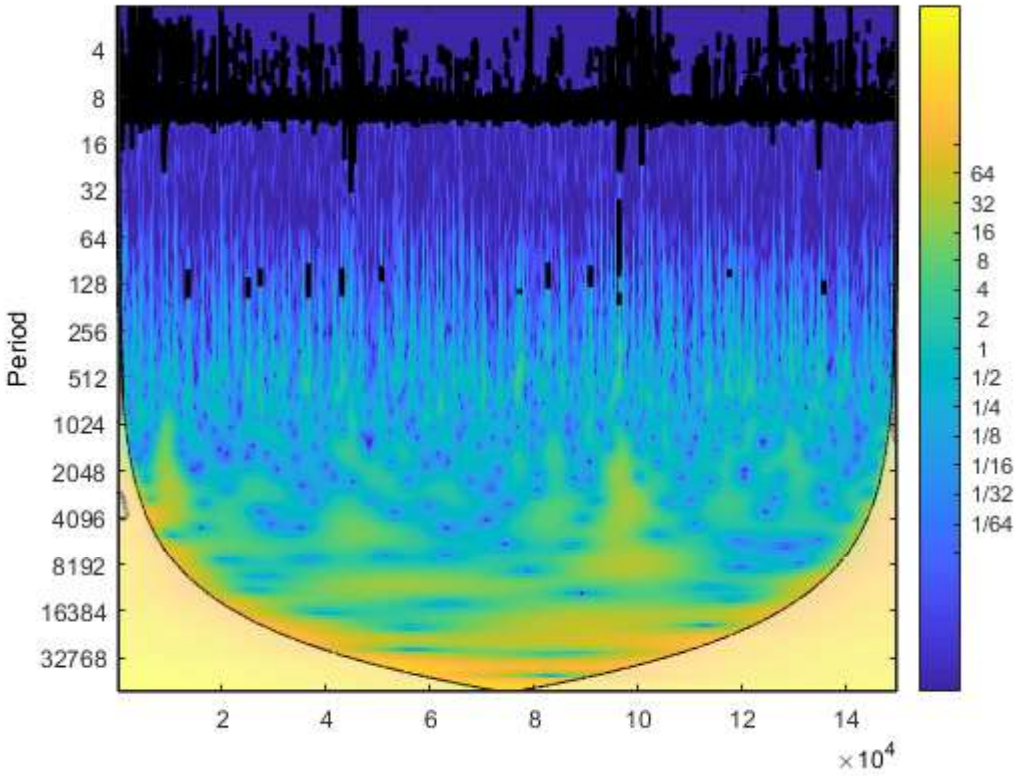

Figure 3.8: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Easy Task, F4 Channel

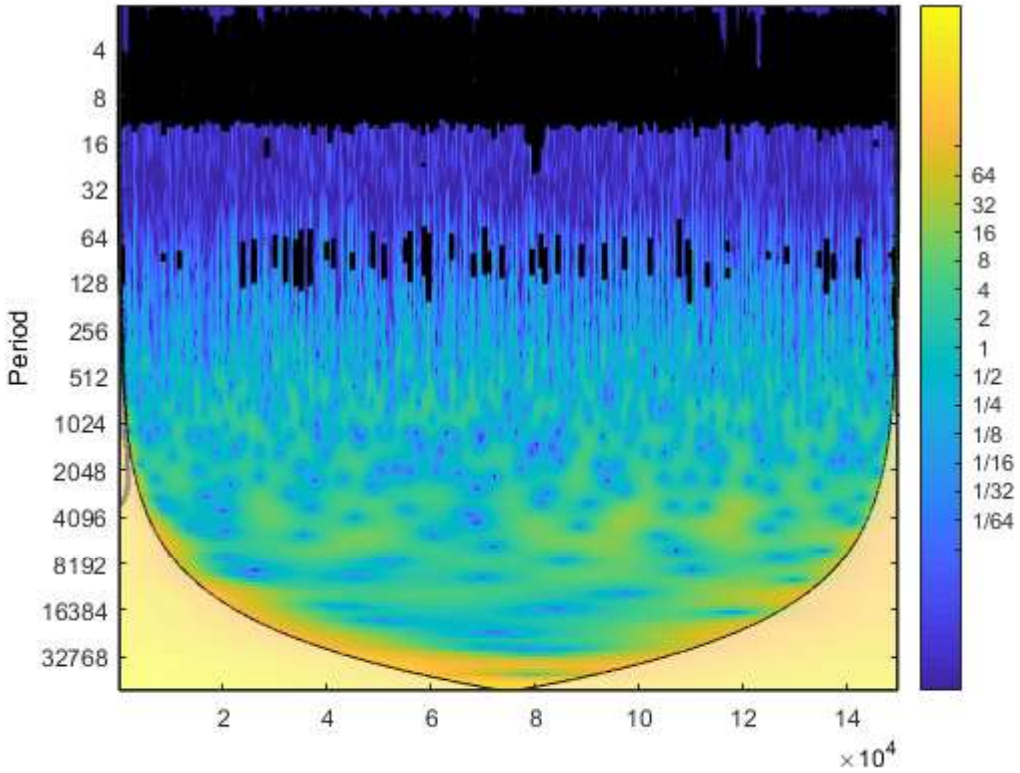

Figure 3.9: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Medium Task, F4 Channel

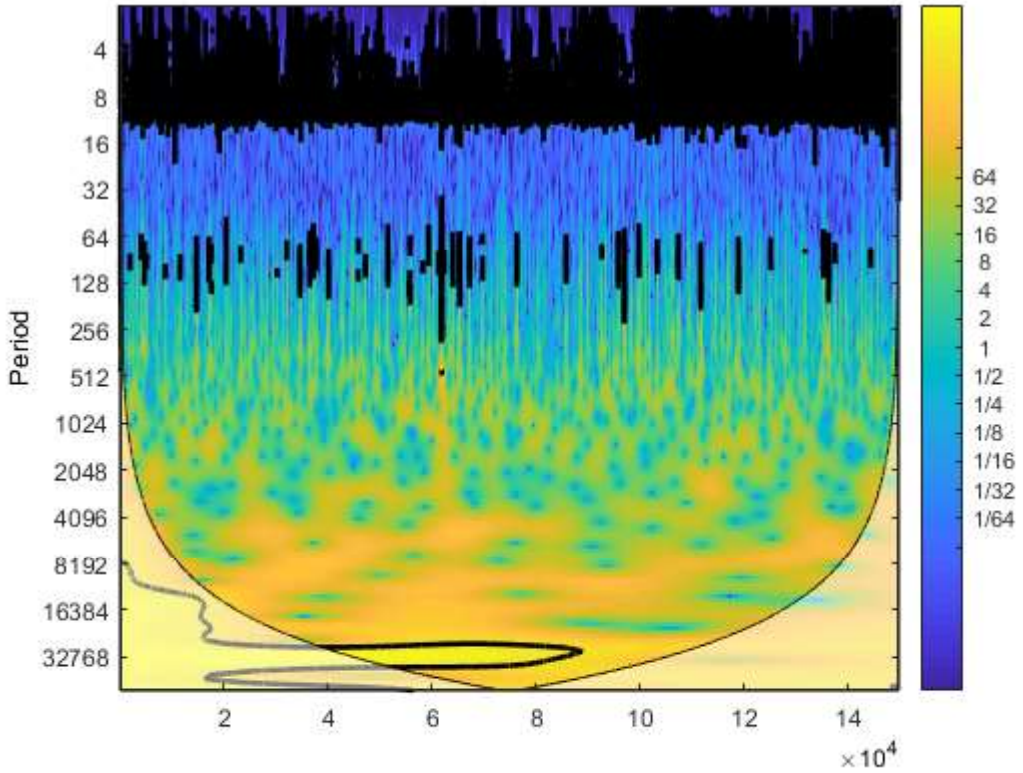

Figure 3.10: Wavelet Output using EEGLAB and MATLAB for Subject 2, Session 3, MATB Difficult Task, F4 Channel

In Figure 3.8, Figure 3.9, and Figure 3.10, it can be observed that the combination of EEG data extraction using EEGLAB and wavelet transform using the MATLAB continuous wavelet transform function (wt) resulted in much sharper and more distinguishable outputs for different task difficulties - easy, medium, and difficult. This indicates that this method yields distinct and clearly separable features for different levels of task difficulty, which is crucial for accurately classifying the mental workload levels. Hence, the study proceeded with the training using these outputs.

The dark regions seen in the images correspond to the higher frequency bands, which are beyond the scope of our analysis as we are primarily interested in the 1-31 Hz frequency range, commonly associated with EEG signals. These high-frequency components are considered noise for our analysis and are removed during the preprocessing stage. Specifically, this is done in the 'dataloader' section of the code (Section 3.2.1), where the images are cropped to retain only the relevant frequency bands. This ensures that the model is trained only on the pertinent features of the EEG signals, thereby improving its ability to accurately classify different levels of mental workload.

3.2 Training

In the present study, the primary objective of the training phase was to develop a convolutional neural network (CNN) capable of accurately classifying task difficulties based on wavelet transform outputs generated from EEG signal channels. The significance of this training lies in the potential to extract relevant features from the wavelet transform outputs, which can provide valuable insights into the underlying brain activity associated with different task difficulties. Such insights could be instrumental in understanding the cognitive demands of different tasks and could have implications for the optimization of task design and workload assessment.

The wavelet transform was chosen as a suitable method for time-frequency analysis of the EEG signals due to its ability to provide a high-resolution representation of the signal's frequency content over time. The wavelet transform outputs, in the form of images, were expected to capture variations in EEG signal characteristics that correspond to different levels of task difficulty. These images were then used as input data for training convolutional neural networks.

Training was conducted on a dataset that had been preprocessed to remove artifacts and noisy wavelet outputs. The dataset was divided into training and test sets, with the training set used to optimize the model parameters and the test set reserved for evaluating the model's performance. The training set consisted of wavelet transform images from 22 subjects, while the test set comprised images from four subjects.

The training process involved iteratively updating the CNN's parameters to minimize the discrepancy between the predicted and true task difficulty labels. Various CNN architectures, including ResNet18, ResNet50, and EfficientNet-B0, and optimization strategies were explored during training to identify the most suitable approach for this classification problem. The performance of the trained models was then evaluated on the test set to assess their ability to generalize to new, unseen data.

The successful classification of task difficulties based on wavelet transform outputs of EEG signals could provide a robust and non-invasive means of assessing cognitive workload. This could have applications in a wide range of fields, including human-computer interaction, ergonomics, and the design of adaptive systems that respond to changes in cognitive demand.

3.2.1 Preparing the Dataset for Training

During the removal of artifacts, it was found that Subjects 5, 9, and 14 had a high number of images with high artifact content in some of their sessions. Upon further examination, it was observed that the data from Subject 17 was indistinguishable from that of Subject 27, suggesting potential data duplication. Therefore, to ensure the integrity of the dataset, Subject 17 was omitted from further analysis. Consequently, these subjects (5, 9, 14, 17)

were completely removed from the dataset. As a result, the total number of images was reduced to 13,545, representing the combined contributions from the remaining 25 subjects.

The dataset was prepared for training by dividing it into distinct training and test sets based on subject numbers. For each of the training methodologies employed, 20 subjects were used for training while the remaining 5 subjects constituted the test set. Before any preprocessing, the dataset comprised 16,362 images spanning these subjects. In terms of class distribution, before the removal of artifacts, each difficulty class (easy, medium, difficult) represented 33.3% of the images. After artifact removal, this equal class distribution was retained in both the training and test sets. Specifically, in the dataset, each difficulty class was represented by 4,515 images, total of 13,545 images

The first method of training utilized was 5-fold cross-validation. In this technique, the dataset is divided into five equal parts. For each fold, four parts are used for training and the fifth part is set aside for testing. This cycle is repeated five times, ensuring each subset is used as a test set once.

The second method introduced was session-based training. For this approach, two sessions were selected for training purposes and the remaining session was designated for testing. This process was conducted for all possible combinations, resulting in three unique training-testing configurations.

Artifacts, defined as wavelet images with more than 15% black area in the region of interest (1-31Hz), were then removed from the dataset. An example of removed images can be seen in Figure 3.11. This figure displays the wavelet transform output for a MATB difficult task from the PO7 channel of Subject 9, Session 2.

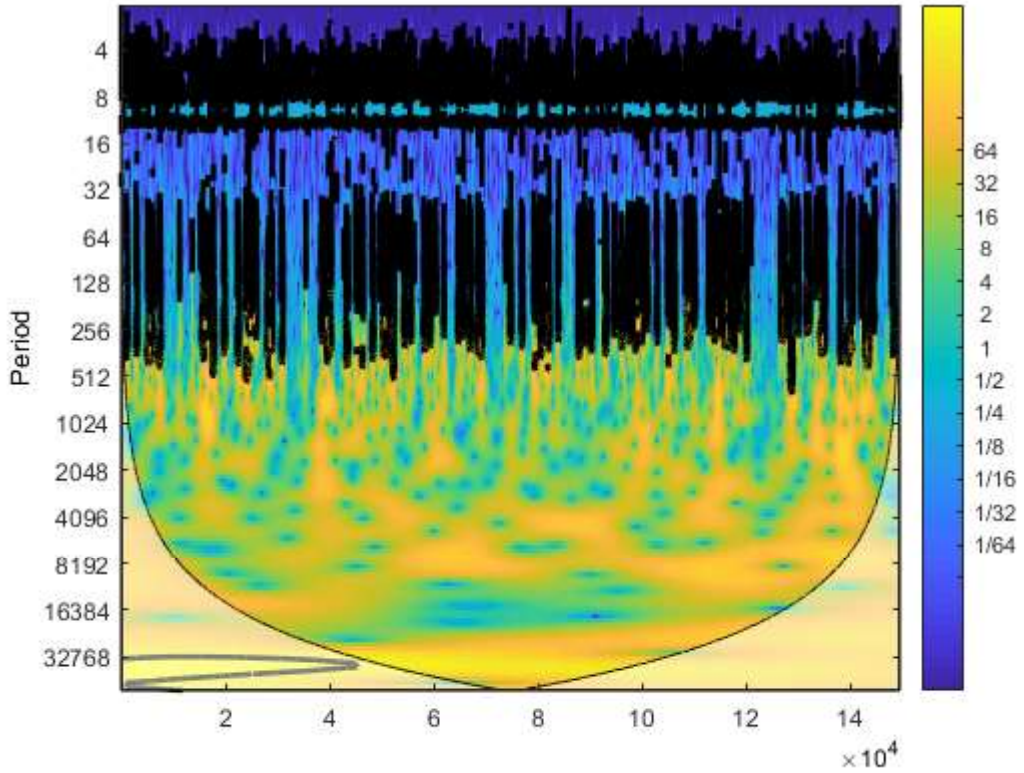

Figure 3.11: Wavelet Output using EEGLAB and MATLAB for Subject 9, Session 2, MATB Difficult Task, PO7 Channel

3.2.2 Dataloader and Dataset Structuring

The `CustomDataset` class was implemented as a subclass of the PyTorch Dataset class to facilitate the loading and preprocessing of the EEG data stored as image files. This custom dataset class is tailored to handle the specific structure and labeling of the wavelet transform images used in the study.

Upon instantiation, the `CustomDataset` class accepts three parameters: the root directory containing the images, the desired image size for resizing, and an optional image transformation. The class identifies all image files within the specified root directory and stores their file paths. The total number of files is recorded for subsequent reference.

The private function `_get_label` extracts the label for a given image based on its file name. It iterates through a predefined list of class names (MATBeasy, MATBmed, MATBdiff) and assigns a numerical label according to the class name present in the file name. The file naming convention is as follows: "sub-XX_ses-SY_MATBZZZ_CC.png",

where XX represents the subject number, SY denotes the session number, ZZZ indicates the task difficulty (easy, medium, or difficult), and CC corresponds to the EEG channel.

The `__getitem__` method retrieves a specific image and its corresponding label based on the provided index. It reads the image from disk, extracts a region of interest, and then converts it from the BGR to RGB color space, as this is the default preprocessing for all three models and is necessary to ensure the pretrained weights are relatable with our image color spaces. The region of interest corresponds to the approximate 1-31Hz region, which is between pixel 60 to 220 on the y-axis, as illustrated in Figure 3.12. This region is then resized as shown in Figure 3.13. The label is obtained using the `_get_label` method, and a one-hot encoded target vector is created. If an image transformation is provided, it is applied to the image; otherwise, the image is normalized by dividing it by 255.0 and converting it to a PyTorch tensor. The image is then permuted to match the PyTorch tensor shape (C, H, W) and converted to a NumPy array of type float32. The transformed image and one-hot encoded target vector are returned as a tuple.

In Figure 3.12, an example of a wavelet transform output is depicted. This specific output was generated from the EEG data of Subject 26, during Session 1, while performing the MATB difficult task on the AF3 channel. The frequency range represented in this wavelet transform spans from 1 to 31 Hz. As can be observed in the figure, the wavelet transform provides a visual representation of the EEG signal's time-frequency characteristics.

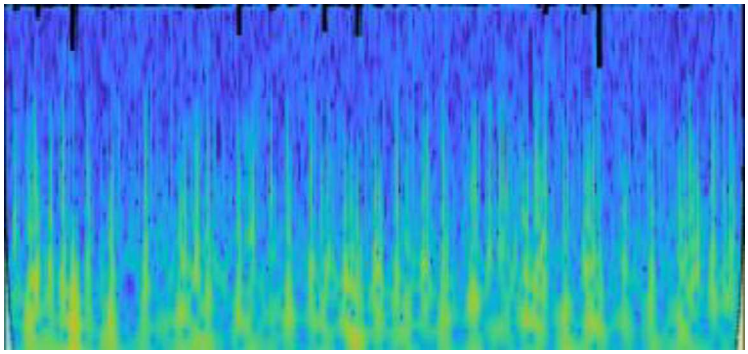

Figure 3.12: Example Wavelet Transform Output from Subject 26, Session 1, MATB Difficult Task AF3 Channel in approx. 1-31 Hz Frequency Range

Following the wavelet transformation, the images were resized to fit the input requirements of the neural network models. Figure 3.13 displays the resized format of the wavelet transform output shown in Figure 3.12. The image has been resized to 224x224 pixels, which is a standard input size for many deep learning models, including the ResNet and EfficientNet models used in this study. This resizing process ensures that the wavelet transform outputs are in a consistent format, suitable for training the neural network models.

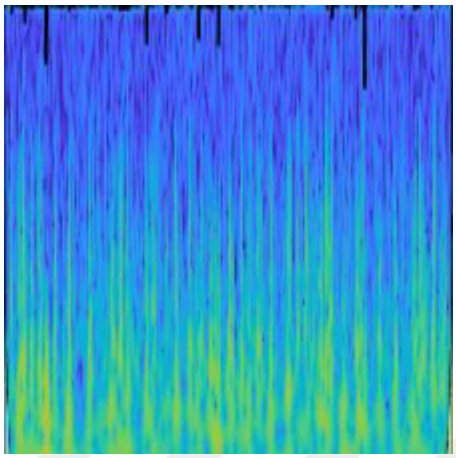

Figure 3.13: Resized Format of Figure 3.6 to 224x224 for Training Input

3.2.3 Model Architectures

In this study, three different convolutional neural network (CNN) architectures were utilized: ResNet18, ResNet50, and EfficientNetB0. Each architecture was trained from scratch (non-pretrained) and with pre-trained weights (transfer learning). The image input size for all models was set to 224x224 pixels.

For each model architecture, the fully connected layer at the end of the network was modified to suit the specific requirements of the task at hand. The original fully connected layer, which was designed for the ImageNet classification task with 1000 classes, was replaced with a new fully connected layer tailored for our three-class classification problem. Specifically, the output dimension of the fully connected layer was set to 3, corresponding to the three classes in our dataset: easy, medium, and difficult. This modification allows the network to produce class scores specifically for our classification task, enabling it to learn features that are relevant to the distinctions between the three difficulty levels in the MATB tasks.

- ResNet18

ResNet18 is a residual network consisting of 18 layers. It is a relatively lightweight model, making it suitable for applications where computational resources are limited. In this study, the ResNet18 model had a total of 11,178,051 parameters, all of which were trainable. The estimated total size of the model was 106.00 MB, with the model parameters accounting for 42.64 MB.

- ResNet50

ResNet50 is another residual network, but with a more complex architecture, consisting of 50 layers. It has a higher capacity for learning complex features compared to ResNet18. The ResNet50 model in this study had a total of 23,514,179 parameters, all of which were

trainable. The estimated total size of the model was 376.82 MB, with the model parameters accounting for 89.70 MB.

- EfficientNetB0

EfficientNetB0 is part of the EfficientNet family of models, which are designed for efficient training and inference. The EfficientNetB0 model used in this study had a total of 4,011,391 parameters, all of which were trainable. The estimated total size of the model was 228.67 MB, with the model parameters accounting for 15.30 MB.

3.2.4 Transfer Learning Purpose and Advantages

Transfer learning is a machine learning technique where a pre-trained model on a large dataset is used as a starting point for training on a smaller, domain-specific dataset. In this study, transfer learning was applied by initializing the models with weights pre-trained on the ImageNet1000 dataset.

The primary advantage of transfer learning is that it allows for faster convergence during training. The pre-trained weights provide a good starting point for the optimization process, and the model can fine-tune these weights to the specific task at hand. This is particularly beneficial when the available dataset is small, as it helps prevent overfitting.

In this study, transfer learning was explored for all three architectures: ResNet18, ResNet50, and EfficientNetB0. Interestingly, it was observed that the pretrained models achieved better results. This might be counterintuitive given that the wavelet transform images used in this study are quite different from the natural images in the ImageNet dataset. However, this outcome suggests that the initial layers of the pretrained models, which capture generic features, might be beneficial even when the target dataset (wavelet transform images of EEG) differs significantly from the images (natural images) the models were pretrained on. This implies that transfer learning can still be a valuable approach in scenarios where the target dataset does not share obvious similarities with the pre-training dataset.

3.2.5 Model Training Details

- Loss Functions

In our training process, we employed the Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) as the loss function. This loss function combines the sigmoid activation function and binary cross-entropy loss in one single class. It is well-suited for multi-label classification problems where each example can belong to more than one class.

- Training Hyperparameters

The training process utilized a range of learning rates from 0.01 to 0.0005. A decaying learning rate was employed using the StepLR scheduler from the PyTorch library, which reduces the learning rate by a factor of 0.9 every 3 epochs. This approach helps to gradually decrease the learning rate, allowing the model to converge more effectively to a local minimum. In terms of batch size, we used a value of 64 for training. Early stopping was considered but ultimately not implemented. We observed that the loss and accuracy

were not consistently stable across epochs, with the best accuracy sometimes achieved at later epochs (e.g., epoch 150) and at other times at earlier epochs (e.g., epoch 17). Thus, we decided to avoid using an early stopping algorithm in order to fully explore the potential of the models throughout the entire training process.

- Optimization Methods

Two optimization methods were employed in the training process: Adam and RMSprop. The Adam optimizer combines the advantages of both the AdaGrad and RMSprop optimization methods. It computes adaptive learning rates for each parameter by considering the first and second moments of the gradients. On the other hand, the RMSprop optimizer divides the learning rate for each parameter by an exponentially decaying average of squared gradients. Both optimizers were experimented with in our training process to determine their impact on the model's performance.

3.2.6 Evaluating Models during Training

During the evaluation process, the test function initially sets the model to evaluation mode by invoking model.eval(). This step is crucial, as it ensures that layers like dropout and batch normalization, which behave differently during training and evaluation, are set to the appropriate mode.

The loss function used in the evaluation is the Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss), which combines the sigmoid activation function and binary cross-entropy loss into a single class. This loss function is well-suited for binary classification problems like ours.

The function then iterates over the validation dataset, provided by the test_loader, and for each batch of data, it performs several steps. First, the input data and target labels are moved to the GPU, as all the computations are performed on the GPU for faster processing. The model then computes the output predictions for the input data. The loss between the predictions and target labels is computed and added to the total loss. The predictions are converted to class labels by selecting the index with the maximum probability. The accuracy is estimated by comparing the predicted class labels with the true labels, which are extracted from the image file names and have been numerically labeled in the dataloader. If the predicted and true labels match, the prediction is considered correct (1), otherwise, it is considered incorrect (0). The number of correctly predicted samples is updated accordingly.

After iterating over the entire validation dataset, the function calculates the average loss by dividing the total loss by the number of samples in the validation dataset. The accuracy is calculated as the percentage of correctly predicted samples out of the total samples in the validation dataset. Both the average loss and accuracy are then printed, offering insights into the model's performance on the validation dataset.

Finally, the function returns the average loss and accuracy, which can be used for further analysis, such as monitoring the performance over time or deciding when to stop training.

By employing this validation approach, we were able to observe the progress of the models throughout the training process. Monitoring the decrease in loss and increase in accuracy on the validation set provided valuable insights into the learning process and helped assess the effectiveness of the training hyperparameters and optimization strategies.

CHAPTER 4

4 RESULTS

4.1 Preprocessing Results from Python Libraries

It was observed that the images generated by the PyWavelets (PyWT), PyTorch Wavelet Toolbox (PTWT), and SqueezePy libraries were not very sharp, as illustrated in Figures 3.5, 3.6, and 3.7. Despite being experimented in training and testing, these libraries were not included in the final results due to their suboptimal performance, with accuracy rates ranging from 29% to 42%. Given that this dataset comprises three classes with an equal number of data points, the baseline accuracy is 33%. Therefore, the performances of PyWT, PTWT, and SqueezePy were deemed insufficient for inclusion in the final analysis. These libraries were initially tried in the first experiments but were later discarded as they were considered obsolete compared to the combination of the MATLAB function and EEGLAB.

4.2 Model Experiments

In our testing approach for model experiments, sessions 1 and 2 were designated for training, and session 3 was reserved for testing. We conducted seven distinct experiments, encompassing architectures such as EfficientNet-B0, ResNet18, and ResNet50. Each architecture was trained in two ways: using pretrained weights and starting from scratch. Additionally, we compared the performance of the Adam optimizer with that of the RMSprop optimizer.

Among our findings, the best accuracy of 58.53% was achieved by EfficientNet-B0 when trained with pretrained weights and using the Adam optimizer (Experiment 2). On the other hand, the lowest accuracy of 49.17% was observed with EfficientNet-B0 trained with pretrained weights but using the RMSprop optimizer (Experiment 1). This highlights the importance of the optimizer choice and suggests that pretrained models, especially EfficientNet-B0, can be particularly effective for classifying task difficulties based on EEG wavelet transform images.

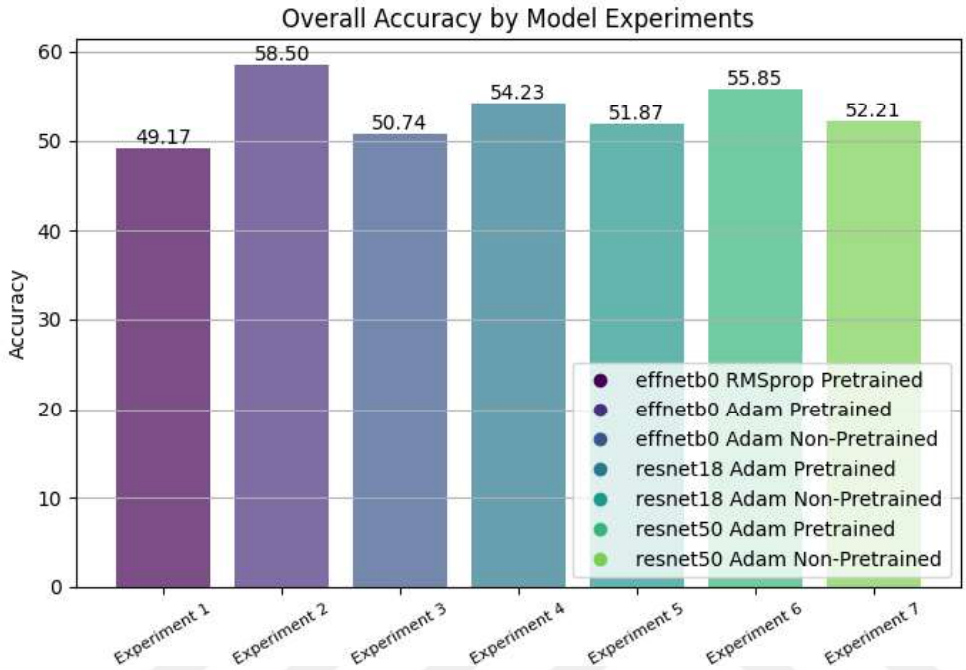

Figure 4.1: Overall Model Accuracy Comparison, Trained on Session 1-2, Tested on Session 3

Figure 4.1 displays the models based on their overall accuracy. For each model, the model properties (model type, whether it is pretrained, and the optimizer used) and the epoch at which the highest accuracy was achieved are provided. It was noted that the RMSprop optimizer performed significantly worse, hence it was only included in one model for comparison. In this summary, all models utilize the Adam optimizer, which has been observed to perform better in practice. The choice of optimizer can have a significant impact on the model's convergence and overall performance. In this case, Adam appears to be a suitable choice for the given models and dataset.

4.3 Analysis by Session

In our testing approach for model experiments, we further probed the performance of our most accurate model, EfficientNet-B0. Initially, we retrained the model using data from sessions 1 and 2 and subsequently evaluated its efficacy on session 3 data. This methodology aligns with the protocols set forth by the competition organized by Roy et al. (2022). In addition to the configuration mentioned earlier, we conducted two other rigorous tests: firstly, the model was trained using data from sessions 2 and 3, followed by an assessment of its performance on session 1 data. Secondly, we trained the model

with data from sessions 1 and 3 and then evaluated its accuracy on session 2 data. These supplemental tests were instrumental in further validating the model's adaptability and its proficiency in generalizing across varied session combinations.), which emphasized testing model resilience against cross-session variability.

Interestingly, when the model was trained on session 2 and 3 data and then tested on session 1 data, it registered a peak accuracy of 63.94%, as depicted in Figure 4.2. Across the three experiments, the obtained results varied, with accuracies of 52.96%, 58.50%, and 63.94%. This variation in performance further underscores the notion that session data can indeed differ, even for the same subjects.

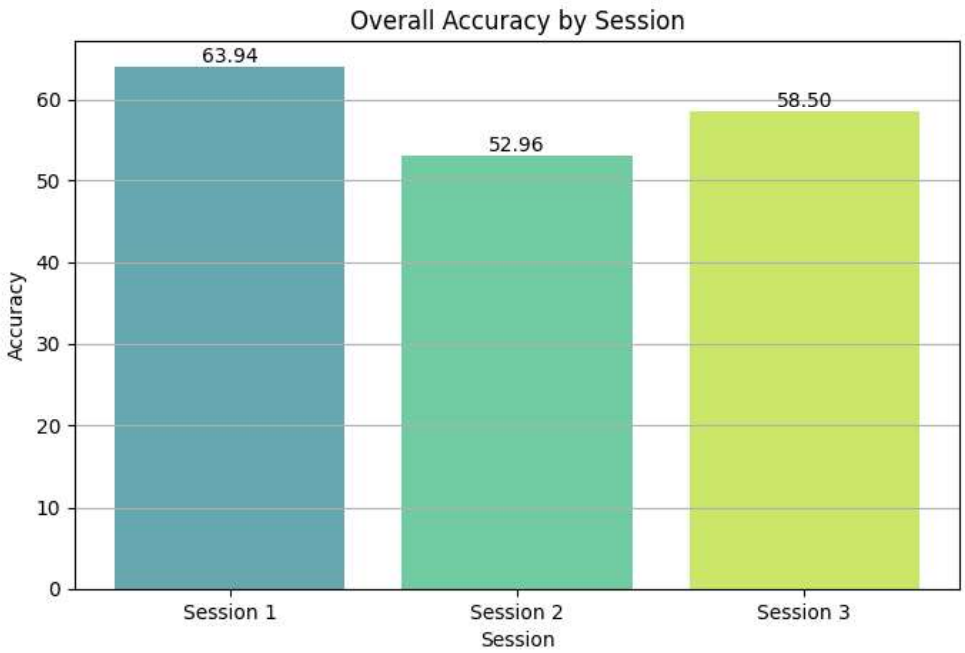

Figure 4.2: Overall Accuracy by Sessions for all 3 experiments where the x-axis denotes the session used as the test data.

In Figures 4.3, 4.4, and 4.5, bar plots delineate the relationship between accuracy and subject density for tests conducted on sessions 1, 2, and 3 respectively.

For Figure 4.3, the median accuracy was pegged at 61%, with an interquartile range of 26%, indicating the middle 50% spread of the data. Accuracies spanned from a minimum of 11% to a maximum of 98%.

Turning to Figure 4.4, the median accuracy was recorded at 50%. The data showcased a spread, as indicated by the interquartile range of 23%. Accuracies in this set fluctuated between a minimum of 26% and a peak of 86%.

In Figure 4.5, the central tendency of the data, as represented by the median, was 59%. The data exhibited an interquartile range of 17%, signifying variability in model performance. The results ranged from a low accuracy of 26% to an impressive high of 95%.

Collectively, these figures underscore the variability and challenges associated with EEG data. Intriguingly, while certain sessions showcased accuracy rates surpassing 90%, others exhibited rates as low as 10%, even though the experiments were conducted on identical subjects with sessions spaced only a week apart. This highlights the inherent variability of EEG data, even when derived from the same subjects under ostensibly comparable conditions.

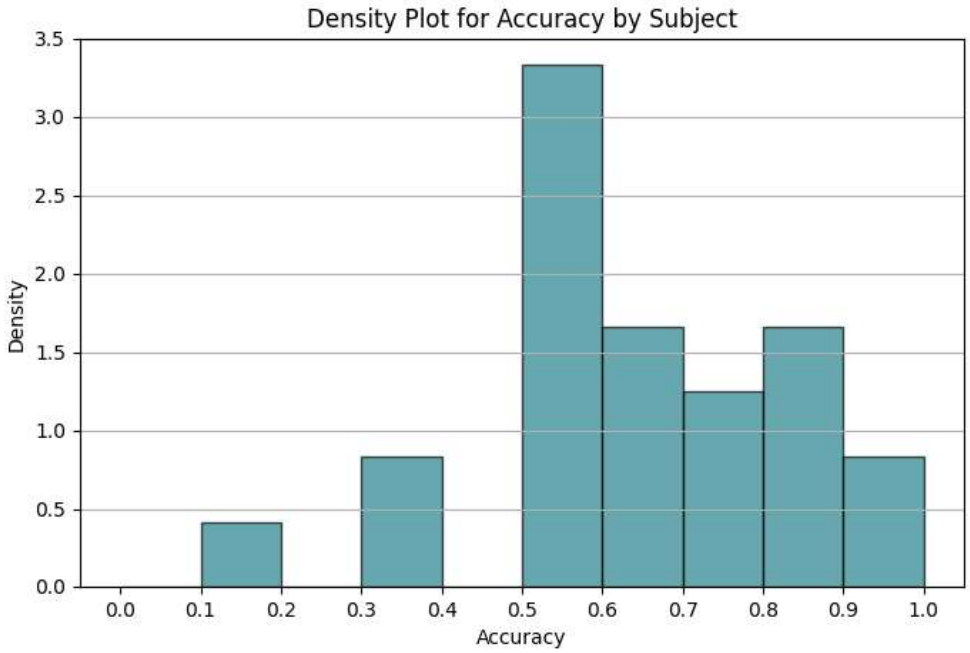

Figure 4.3: Accuracy vs Density for Session 1 Test Data, Session 2 and 3 used for training

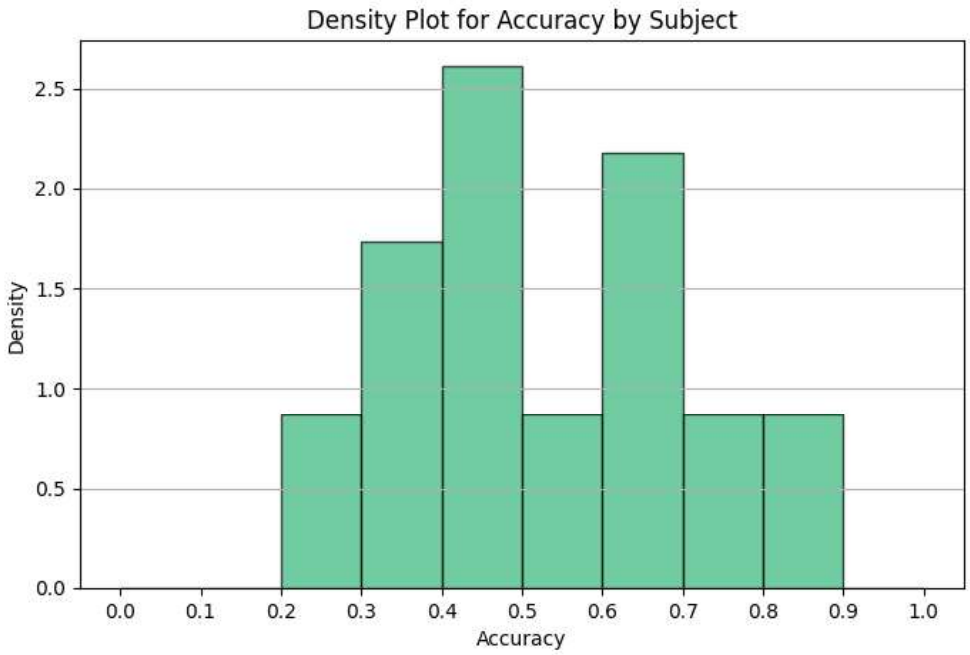

Figure 4.4: Accuracy vs Density for Session 2 Test Data, Session 1 and 3 used for training

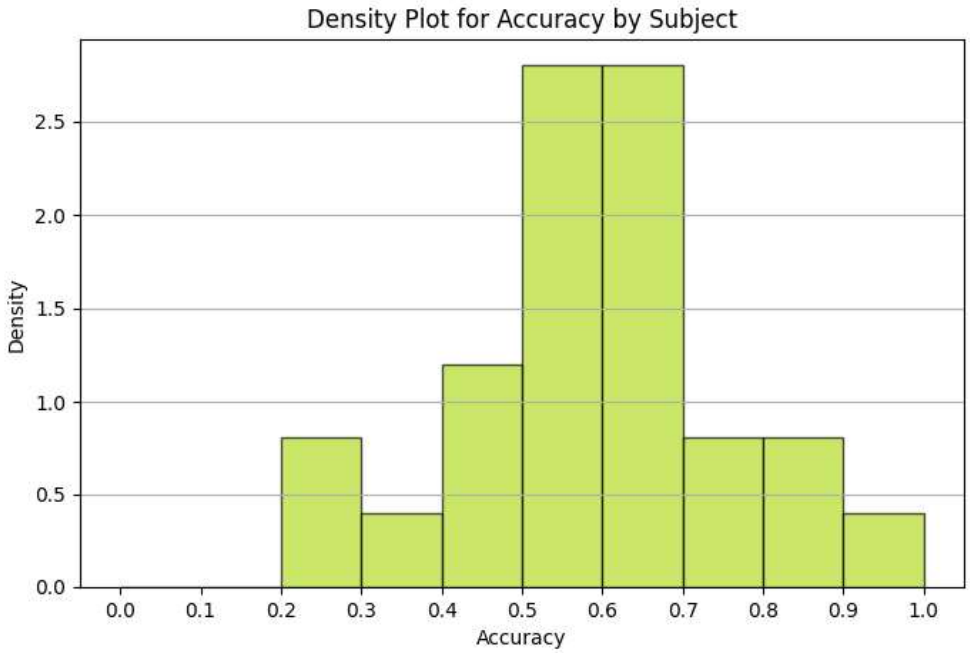

Figure 4.5: Accuracy vs Density for Session 3 Test Data, Session 1 and 2 used for training

4.4 Analysis by Subject with 5-Fold Cross Validation

Figure 4.6 illustrates the overall accuracy of EfficientNet-b0 model trained using the 5-fold cross-validation method with pretrained weights and Adam optimizer. In this scheme, each fold comprises data from 20 training subjects and 5 testing subjects, ensuring that every subject is incorporated into the training process in one of the five folds. Notably, this method leverages all three sessions directly from each subject, aiming to derive a generalized solution for unseen subjects based on the data from the training subjects. This methodology aligns with the approach employed by Hinss and colleagues (2023), from whom the 29-subject COG-BCI dataset originated.

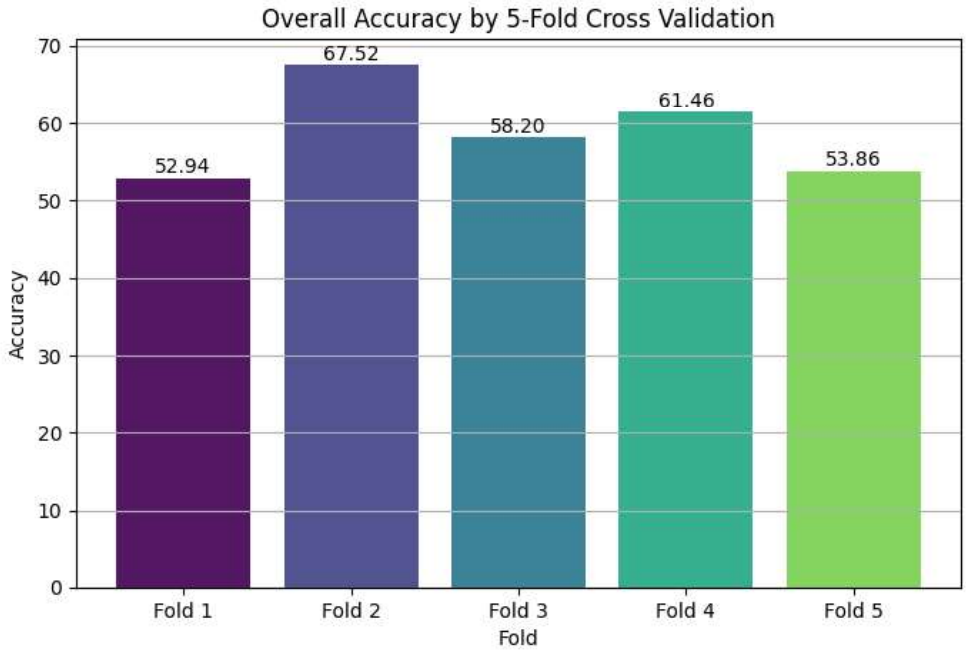

Figure 4.6: Overall Accuracy of 5-Fold Cross Validation Training, every fold includes 20 training and 5 test subjects

The results from this method present a range of accuracies: the highest accuracy achieved in one of the folds is 67.52%, while the lowest dips to 52.94%. Recall, precision, and F-1 ratio statistics and the confusion matrix for this model are provided in the Appendix.

Figure 4.7 provides a comprehensive boxplot representation that captures the distribution of overall accuracies when applying the 5-fold cross-validation. The median accuracy across the five folds stands at 58.20%. The interquartile range (IQR), which represents the middle 50% spread of the data, spans a range of 7.6%. The lowest accuracy observed is

52.94%, while the highest peaks at 67.52%. The spread of the data, as visualized by the range from the minimum to the maximum accuracy, highlights the variability in model performance across different folds, underscoring the importance of a robust evaluation strategy.

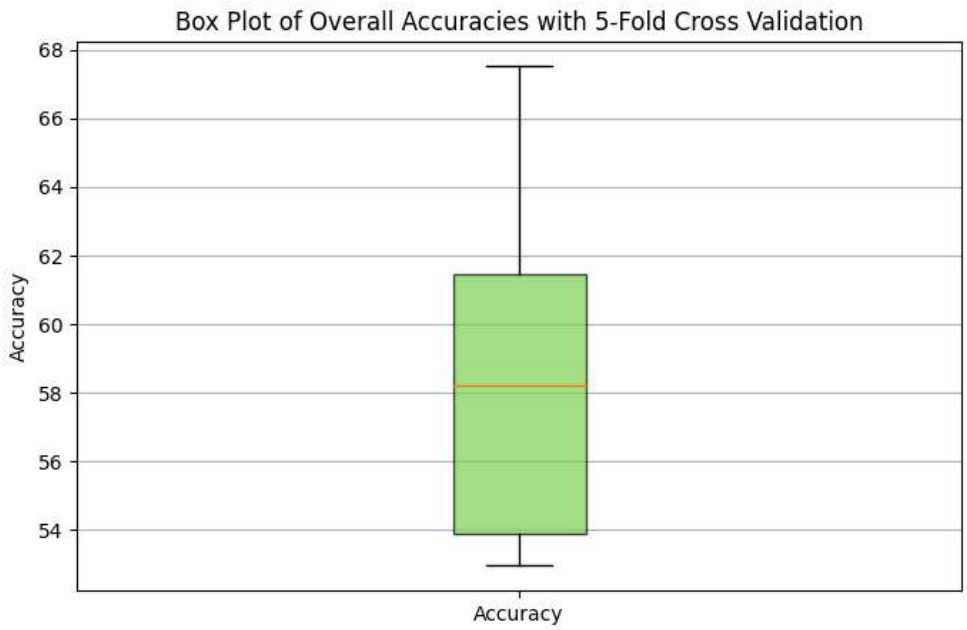

Figure 4.7: Box Plot of Overall Accuracies with 5-Fold Cross Validation Method

Figure 4.8 presents the distribution of model accuracies against subject density using the Fold-5 Cross Validation method. The median accuracy for this set of results is 0.56. The data showcases a spread with an interquartile range (IQR) of 0.14, suggesting variability in model performance across different subjects. The minimum accuracy value recorded is 0.36, while the model's performance peaks at an impressive accuracy of 0.84. The range between these extremes highlights the diverse performance outcomes based on subject-specific nuances, emphasizing the inherent variability in EEG data even under controlled experimental conditions.

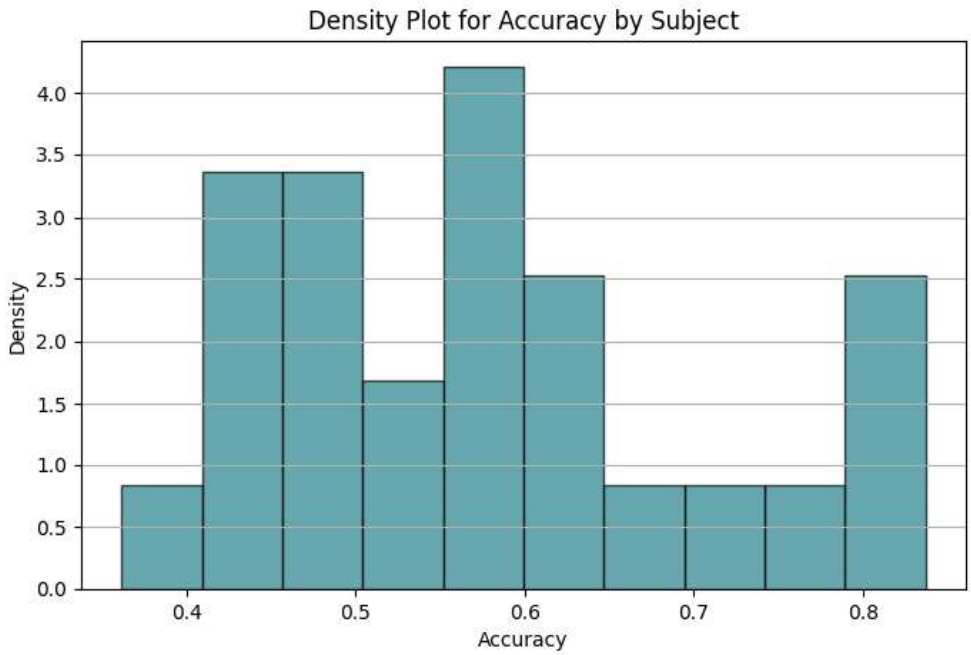

Figure 4.8: Model Accuracy vs Density by Subject from 5-Fold Cross Validation Method

Figure 4.9 provides a visual insight into the model's performance across three distinct sessions. By examining the boxplots, we observe that for all sessions, the accuracies predominantly hover above the 50% threshold, indicative of a model that often performs better than random guessing which is 33.3% for this experiment. Session 1's median accuracy approaches 60%, and its interquartile range suggests a relatively tight spread of results, although a few high-performing outliers are evident. Session 2 displays a median slightly above 50%, but its wider interquartile range implies greater variability in model performance for this session. In contrast, Session 3 stands out, not only boasting the highest median accuracy, close to 65%, but also displaying a compact spread of results, which underscores a consistent performance.

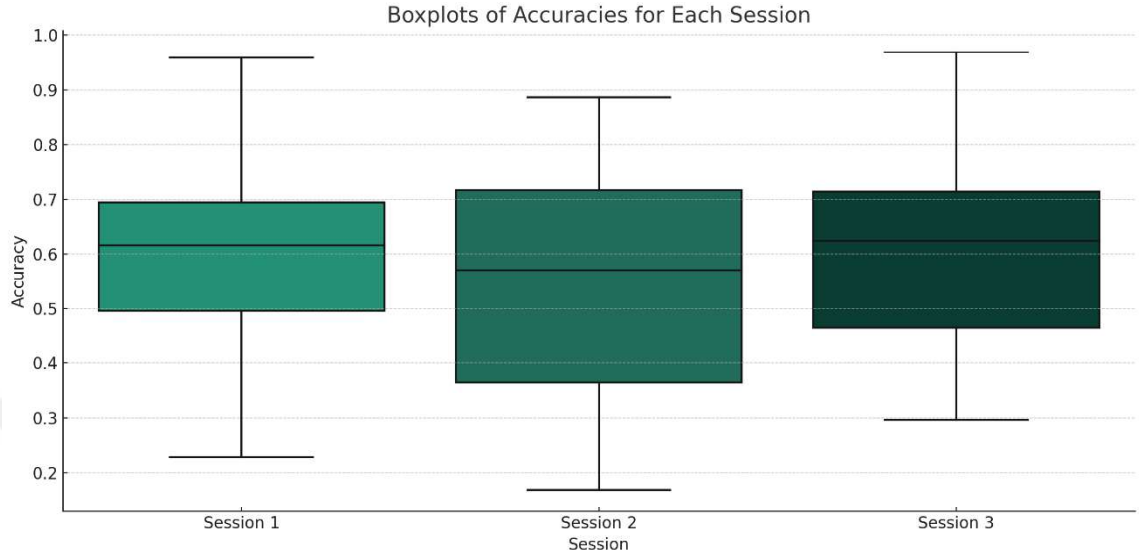

Figure 4.9: Boxplots of Accuracies for Each Session from 5-Fold Cross Validation Method

4.5 Analysis by Task Difficulty

In the results derived from the session 1 test data, as presented in Figure 4.10, utilizing the EfficientNet-B0 architecture with pretrained weights and the Adam optimizer, the model showcased a remarkable proficiency in classifying 'Easy' tasks, achieving an accuracy rate of 77%. This performance highlights the model's adeptness in recognizing features associated with tasks of lower cognitive demand. However, when confronted with 'Medium' tasks, its performance dipped significantly, registering an accuracy of just 40%. The accuracy for 'Difficult' tasks was intermediate, landing at 75%. It's noteworthy to mention that these findings are consistent across three distinct session training cases.

Turning to the session 2 test data, a consistent pattern is evident. The model's strength in discerning 'Easy' tasks persisted, reflected in an accuracy of 70%. Yet, its consistent challenge in accurately classifying 'Medium' tasks was again underscored, with an accuracy of merely 33%. The 'Difficult' tasks were classified with an accuracy of 56%, aligning with the observed trend.

Upon analyzing the session 3 test data, the model maintained its momentum in classifying 'Easy' tasks, achieving a striking accuracy of 83%. Conversely, its performance on 'Medium' tasks remained a bottleneck, with an accuracy of 34%. The 'Difficult' tasks were identified with a moderate accuracy of 59%. Precision, recall, and F-1 ratio statistics for these models together with confusion matrices are provided in the Appendix.

Cumulatively, across all sessions, the data illuminates the model's recurring challenge with the 'Medium' tasks classification. While it consistently excels in identifying 'Easy'

tasks, and moderately recognizes 'Difficult' tasks, the 'Medium' category remains a distinct hurdle. This highlights the model's potential difficulty in distinguishing the subtle complexities inherent to 'Medium' tasks when compared to the other levels of difficulty.

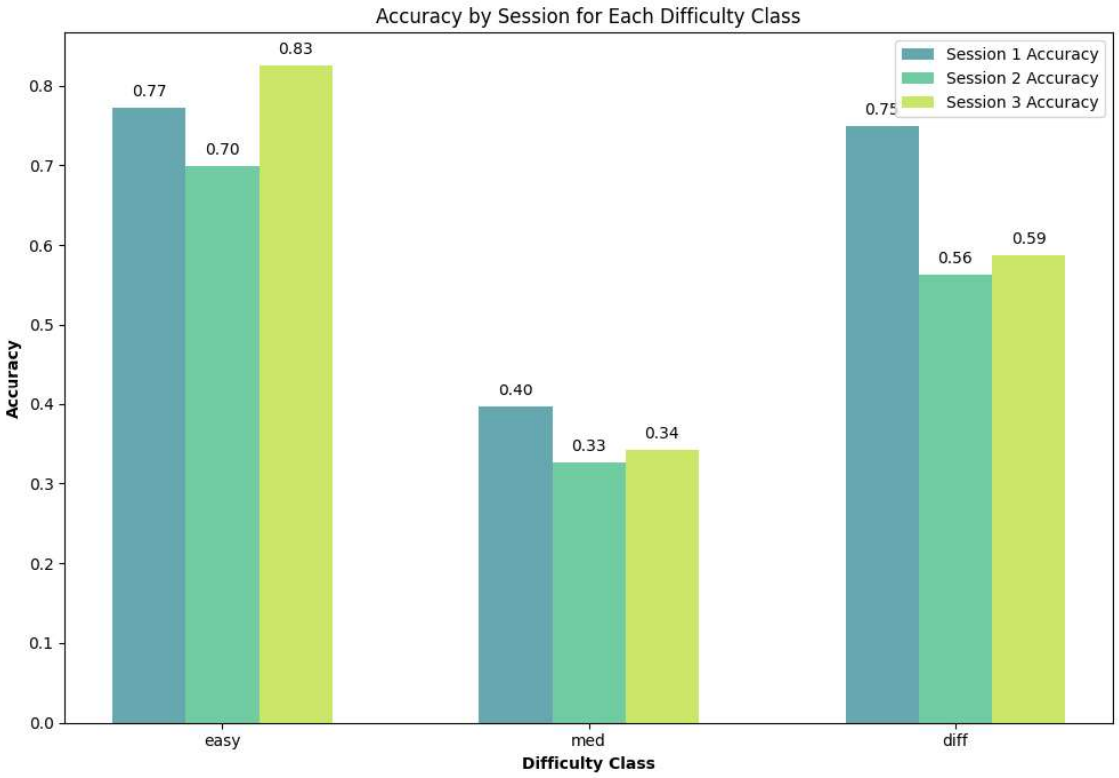

Figure 4.10: Model Accuracy Comparison by Task Difficulties for Each Session Training

In Figure 4.11, which presents results from the experiment utilizing the 5-fold cross-validation method, the model's performance characteristics echo those observed in previous session-based trainings.

Using the EfficientNet-B0 architecture equipped with pretrained weights and the Adam optimizer, the model demonstrated a noteworthy capability in classifying 'Easy' tasks, achieving an accuracy of 78%. This consistent high performance in recognizing 'Easy' tasks further accentuates the model's strength in identifying features associated with tasks that demand lower cognitive engagement.

However, just as in the session-based experiments, the 'Medium' tasks emerged as a challenge for the model. The accuracy for this category was notably lower at 43%. This continued disparity in the model's ability to accurately classify 'Medium' tasks raises

intriguing questions about the intricacies and potential overlaps in EEG signatures between 'Medium' and the other difficulty levels.

For 'Difficult' tasks, the model achieved an intermediate accuracy of 55%. This falls in line with the previously observed trend, where the accuracy for 'Difficult' tasks consistently positions itself between the accuracies for 'Easy' and 'Medium' tasks.

Overall, the consistency in the model's performance across both session-based and 5-fold cross-validation methods is striking. The repeated pattern — with 'Easy' tasks being most accurately classified, 'Difficult' tasks achieving moderate accuracy, and 'Medium' tasks being the most challenging — underscores the inherent complexities in distinguishing EEG patterns associated with varying cognitive demands. The model's consistent behavior cross different experimental setups also emphasizes the reproducibility and robustness of these findings.

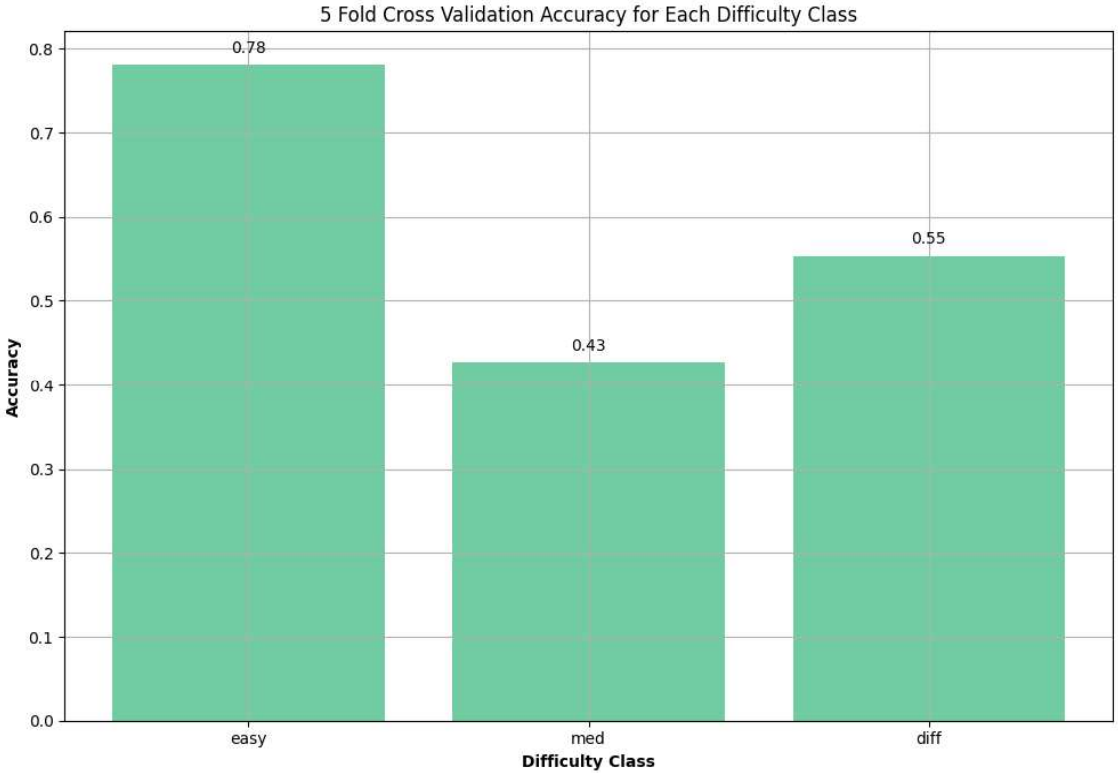

Figure 4.11: Model Accuracy Comparison by Task Difficulties for 5-Fold Cross Validation Method

CHAPTER 5

5 DISCUSSION

5.1 Comparison with Previous Studies

The present study utilized deep learning models to classify task difficulties based on wavelet transform images of EEG signals from 29 subjects and achieved the highest overall accuracy of 67.52% using a pretrained EfficientNet-B0 model with the Adam optimizer. This is significant given the complexity of the task; however, it is essential to compare these results with previous studies. For instance, Roy et al. (2022) organized a passive brain-computer interface (pBCI) competition focusing on cross-session workload estimation using EEG data from 15 subjects performing the Multi-Attribute Task Battery (MATB) tasks. This 15-subject dataset is the first version of the COG-BCI database, which was later completed with 29 subjects (Hinss et al., 2022). In the competition, Roy et al. (2022) used session 1 and 2 as training and validation datasets, and session 3 as the testing dataset. The best-performing algorithm in the Roy et al. (2022) study utilized Riemannian geometry principles and achieved an accuracy of just under 60%, well above the chance level of 33% (default accuracy for a 3-class classification problem with equally distributed class weights), highlighting the challenges in this research area. The use of deep learning methods in the Roy et al. (2022) study encountered generalization issues and significant overfitting problems. This is one of the reasons why the present study chose to use wavelet transform images.

In our extended exploration, distinct from the methodology adopted by Roy et al. (2022), we utilized data from all three sessions across 25 subjects. It's noteworthy to mention that four subjects were excluded from our analysis due to the presence of artifacts in their data, and one of the subject’s results were found to be identical to another subject, raising concerns about data integrity. This refined dataset provided a comprehensive understanding of the model's behavior across diverse configurations. Specifically, we trained our models on combinations of two sessions and tested them on the third. For our first experiment, sessions 1 and 2 served as the training dataset, with session 3 acting as the test set. In our subsequent experiment, sessions 2 and 3 were designated for training, and the model's performance was assessed on session 1 data. Lastly, in our third configuration, we trained using data from sessions 1 and 3 and tested on session 2.

These experiments yielded varying degrees of accuracy. We observed a peak accuracy of 63.94% when tested on session 1 data, 52.96% on session 2, and 58.50% on session 3 data. A consistent theme across these results was the distinct difference in accuracy based on task difficulty. Notably, 'Easy' tasks consistently emerged as the best-classified category across all sessions. In contrast, 'Medium' tasks presented a consistent challenge, registering the lowest accuracy. This recurring pattern underlines the complexities in interpreting EEG data to discern cognitive workload levels. It also suggests an inherent challenge in distinguishing 'Medium' tasks from their 'Easy' and 'Difficult' counterparts.

Although our method increases the complexity of the model, it can provide a high-resolution representation of the signal's frequency content over time, which is crucial for capturing variations in EEG signal characteristics that correspond to different levels of task difficulty. This approach aims to balance the need for a complex model that can capture the nuances in the EEG data while mitigating the risk of overfitting that was observed in the Roy et al. (2022) study.

Subsequently, Hinss et al. (2023) used the same database as our study, the COG-BCI database (Hinss et al., 2022). However, they took a different approach, using a Riemannian Minimum Distance to Mean (MDM) classifier and focusing mainly on the alpha and beta bandwidths to classify cognitive workload. They also used the 5-fold cross-validation method for dividing their data into training and testing sets. Hinss et al. reported an accuracy of 69.40% (±12.50%) and found different accuracy levels for each session. Specifically, they got the highest accuracy in session 2 (70.30 ± 12.73%), then session 1 (64.30 ± 12.52%), and finally session 3 (66.96 ± 12.87%). While their method has benefits, like reducing data size and possibly improving signal quality, it might miss out on useful information from other frequency bands.

In contrast, our study used a wider frequency range of 1-31 Hz for the initial EEG signals and worked with wavelet transform images. This way, we could gather data across many frequencies. This approach aimed to get more information from the EEG signals, even if it might introduce some noise or less relevant features into the model. The challenge is to find the right balance between getting as much information as possible and limiting noise.

In our study, we also used the 5-fold cross-validation method, similar to the approach taken by Hinss et al. (2023), but applied it to 25 subjects. Our accuracy results varied quite a bit: our highest accuracy was 67.52% and the lowest was 52.94%. The average accuracy across the five groups, as illustrated in Figure 4.7, stood at 58.20% (±7.6%). When examining the results from individual sessions, session 1 showed an average accuracy of 60.03% (±22.12%), session 2 had an average of 54.85% (±22.47%), and session 3 had an average accuracy of 61.11% (±21.35%). Additionally, as seen in Figure 4.8, when we assessed accuracy based on individual subjects, the mean accuracy was 0.56, ranging from 0.36 to 0.84. This variation underscores the distinctiveness of each subject's EEG data, even when gathered under the same conditions. While the best way might change based on the specific task and dataset, our results suggest that using wavelet transform images can lead to high accuracy. Still, more research is needed to see if this method works well for other tasks and datasets.

5.2 Model Performance

In our session-based experiment, paralleling the approach of Roy et al. (2022), we discerned consistent patterns in task classification. Specifically, 'Easy' tasks were most accurately classified, with a peak accuracy of 83%, while 'Difficult' tasks followed closely at 75%. However, 'Medium' tasks posed a consistent challenge across the experiments,

registering the lowest accuracy at 40%. This difficulty in distinguishing 'Medium' tasks was further echoed in our 5-fold cross-validation experiment, where the highest accuracies recorded were 78% for 'Easy' tasks, 55% for 'Difficult' tasks, and 43% for 'Medium' tasks. The recurring theme, both in the session-based and the 5-fold experiments, underscores the inherent challenge in discerning the nuanced EEG patterns of 'Medium' tasks as compared to their 'Easy' and 'Difficult' counterparts.

This variation in performance across subjects suggests that the model may be capturing subject-specific features that are not generalizable to other individuals. This is a common challenge in EEG-based classification tasks, as EEG signals can vary significantly between individuals due to differences in anatomy, physiology, and cognitive processes. Another issue particularly highlighted by the across session analysis is that learning or expertise development due to repeated exposure to a task (e.g. the task battery was repeated 3 times with 1-week separation between sessions) might have factored into the EEG responses, especially for the high difficulty conditions. Therefore, it is important to develop models that can generalize well across different individuals and across sessions. One potential approach to address this issue is to use transfer learning, where a model is pretrained on a large dataset from multiple subjects and then fine-tuned on a smaller, subject-specific dataset. This approach can help the model learn generalizable features from the larger dataset while also adapting to the specific characteristics of each individual. Another potential approach is to use subject-independent features, which are less likely to vary between individuals. Finally, the effects of changing task novelty due to learning on brain responses also need to be considered. However, further research is needed to determine the most effective approach to improve generalizability across subjects.

5.3 Potential Improvements

There are several potential ways to improve the performance of the models in the present study. First, experimenting with larger models with higher resolution may be beneficial. The present study used EfficientNet-B0, ResNet18, and ResNet50 models, which have relatively small input sizes. Larger models, such as EfficientNet-B7, may achieve higher classification accuracy, although they have more parameters and require more computational resources. Additionally, larger models may be more prone to overfitting, especially if the available dataset is small. Therefore, it is important to carefully consider the trade-offs between model size, computational resources, and classification accuracy when selecting a model architecture.

Another potential improvement is to experiment with different bandwidths and feature extraction methods. The present study used wavelet transform images, which capture information across a range of frequencies. However, focusing on specific bandwidths, such as the alpha and beta bands, as suggested by Hinss et al. (2023), may help reduce noise and improve the signal-to-noise ratio. Our approach also did not explicitly capitalized on the spatial information in the EEG signals, which is considered a strength

of Riemannian and spherical CNN approaches (Roy et al., 2022). Therefore, it may be beneficial to experiment with different bandwidths and feature extraction methods to determine the optimal approach for the specific task and dataset.

Experimenting with different model architectures, features, and training strategies may help optimize the model's performance. For example, the present study used convolutional neural networks (CNNs) with wavelet transform images as input features. Other model architectures, such as recurrent neural networks (RNNs) or attention-based models like Transformers, may offer advantages in capturing the temporal aspects of this task, with each bringing unique strengths to modeling sequential data. Additionally, other features, such as time-frequency representations or statistical features, may yield better classification accuracy. Finally, different training strategies, such as data augmentation or regularization, may help prevent overfitting and improve the model's generalization performance. Therefore, it is important to experiment with different approaches and determine the most effective strategy for the specific task and dataset.

5.4 Applications and Implications

The ability to accurately classify task difficulty based on EEG signals has a wide range of potential applications and implications. For example, it could be used to develop adaptive interfaces that adjust the difficulty of a task in real-time based on the user's cognitive workload. This could help prevent cognitive overload and improve performance in tasks that require sustained attention and cognitive effort. Additionally, it could be used to develop brain-computer interfaces (BCIs) that allow users to control devices or applications using their brain activity. This could be particularly useful for individuals with motor impairments or for applications that require hands-free control. Furthermore, it could be used to develop systems that monitor cognitive workload in real-time and provide feedback or interventions to optimize performance. For example, a system could monitor a user's cognitive workload while they are performing a task and provide breaks or adjust the task difficulty if the cognitive workload becomes too high. This could help prevent fatigue and improve overall performance. However, there are also potential challenges and ethical considerations that need to be addressed. For example, the variability in EEG signals across individuals may make it difficult to develop models that generalize well across different individuals. Additionally, the use of EEG-based systems may raise privacy and security concerns, as EEG signals can contain sensitive information about an individual's cognitive state and mental health. Therefore, it is important to develop robust and secure systems that protect user privacy and provide accurate and reliable results.

5.4.1 Operationalizing Cognitive Constructs: Implications and Comparisons

Understanding task difficulty and cognitive load involves distinguishing the inherent challenges presented by stimuli and the subsequent mental effort required to process them. The fact that accuracy levels as high as 70% can be achieved in classifying EEG signals suggests that EEG can be utilized to shed light on brain responses to varying task

complexities. While a higher task difficulty could correspond to increased cognitive load, the exact relationship remains intricate and multifaceted.

Task effort and task difficulty, though interrelated, have distinct connotations. Task difficulty pertains to the inherent challenge of a stimulus, while task effort denotes the mental or physical exertion expended in response. Metrics like response time often serve as indicators of task effort, with more challenging tasks potentially demanding greater effort and, consequently, longer response durations.

The scientific approach taken in this study mirrors practices in other domains, like Natural Language Processing (NLP). In NLP, word embeddings, which represent words in multi-dimensional spaces, have been utilized as proxies for semantic meanings. This doesn't necessitate a complete understanding of the intricacies of human linguistic cognition. Similarly, while EEG-based studies aim to decipher patterns indicative of cognitive states, they don't claim to fully unravel the complexities of human cognition. Instead, a more pragmatic approach is embraced, focusing on discernible outputs and patterns rather than attempting to decode every internal cognitive nuance. Such an approach, akin to treating certain components as a "black box", is valuable as it facilitates tangible outcomes and insights even when the entirety of the underlying processes isn't exhaustively understood.

5.5 Limitations

The present study has several limitations that need to be acknowledged. First, the performance of the models varied significantly across subjects and task difficulties, which suggests that they may not be generalizable across different individuals and tasks. This is a common challenge in EEG-based classification tasks, and further research is needed to develop models that can generalize well across different individuals and tasks. Second, the models' performance was lower for medium tasks, which suggests that they may have difficulty distinguishing between medium and other levels of difficulty. The confusion matrices presented in the Appendix further suggest that several medium difficulty instances are predicted by the corresponding models as difficult cases, especially for the models trained on sessions rather than subjects. This may be due to the similarity in EEG signals between medium and other levels of difficulty or due to the limitations of the models used in the study. Task learning effects may have also contributed to this blurring between medium and high difficulty levels. One may also problematize the distinction made by the experiment designers between medium and difficult episodes in terms of the number of parallel tasks that needs to be attended by the participant in the MATBII environment. A labeling of the dataset based on behavioral performance could mitigate some of these issues. Further research is needed to investigate these issues and develop models that can accurately classify all levels of task difficulty. The wavelet-based CNN approach followed in this thesis also did not explicitly took into account the spatial distribution of the EEG electrodes, which is a strength of the Riemannian and spherical CNN approaches. Finally, the present study used a relatively small dataset from a single

study, which may limit the generalizability of the results. It is important to validate the models on larger and more diverse datasets to ensure their robustness and generalizability.

5.6 Conclusion

In conclusion, the present study suggests that deep learning models can achieve high accuracy in classifying task difficulty based on wavelet transform images of EEG signals, but their performance may vary across subjects and task difficulties. This suggests that there is still room for improvement in developing models that are generalizable across different individuals and can accurately classify all levels of task difficulty. Potential ways to improve model performance include experimenting with larger models, focusing on specific bandwidths, experimenting with different model architectures, features, and training strategies, and validating the models on larger and more diverse datasets. Despite these limitations, the present study provides valuable insights into the development of EEG-based systems for classifying task difficulty and has important implications for the development of adaptive interfaces, BCIs, and systems for monitoring and optimizing cognitive performance.

REFERENCES

Al-Qazzaz, N. K., Bin Mohd Ali, S. H., Ahmad, S. A., Islam, M. S., & Escudero, J. (2015). Selection of Mother Wavelet Functions for Multi-Channel EEG Signal Analysis during a Working Memory Task. *Sensors, 15*(11), 29015–29035. https://doi.org/10.3390/s151129015

Anderson, E., Potter, K., Matzen, L., Shepherd, J., Preston, G., & Silva, C. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum*, 30(3), 791-800. https://doi.org/10.1111/j.1467-8659.2011.01928.x

Coyle, S. (2005). *Near-infrared spectroscopy for brain computer interfacing* (Doctoral dissertation, Faculty of Science and Engineering, Electronic Engineering, National University of Ireland Maynooth)

Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Engel, A., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, 2(10), 704-716. https://doi.org/10.1038/35094565

Grossman, A., & Morlet, J. (1985). Decomposition of functions into wavelets of constant shape, and related transforms. *Mathematics and Physics: Lectures on Recent Results*, *11*, 135-165.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(4), 647-658.doi: 10.1109/21.17382

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hinss, M. F., Jahanpour, E. S., Somon, B., Pluchon, L., Dehais, F., & Roy, R. N. (2022). COG-BCI database: A multi-session and multi-task EEG cognitive dataset for passive brain-computer interfaces. *Zenodo* https://doi.org/10.5281/zenodo, 6874128.

Hinss, M. F., Jahanpour, E. S., Somon, B., Pluchon, L., Dehais, F., & Roy, R. N. (2023). Open multi-session and multi-task EEG cognitive Dataset for passive brain-computer Interface Applications. *Scientific Data*, 10(1), 85.

Hinss, M. F., Somon, B., Dehais, F., & Roy, R. N. (2021, May). Open EEG datasets for passive brain-computer interface applications: Lacks and perspectives. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 686-689). IEEE. https://doi.org/10.1109/NER49283.2021.9441214

Hoonakker, P., Carayon, P., Gurses, A., Brown, R., McGuire, K., Khunlertkit, A., & Walker, J. (2011). Measuring workload of icu nurses with a questionnaire survey: the nasa task load index (tlx). *IIE Transactions on Healthcare Systems Engineering*, 1(2), 131-143. https://doi.org/10.1080/19488300.2011.609524

Issartel, J., Bardainne, T., Gaillot, P., & Marin, L. (2015). The relevance of the cross-wavelet transform in the analysis of human interaction–a tutorial. *Frontiers in Psychology, 5*, Article 1566.

Küskü, M. (2022). *Inter-brain synchronization patterns of cooperation in the prefrontal cortex during the Stag Hunt game via fNIRS hyperscanning* (Master's thesis, Department of Cognitive Science, Graduate School of Informatics, Middle East Technical University).

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings* (pp. 402-405). Springer.

Longo, L. (2015). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, 34(8), 758-786. https://doi.org/10.1080/0144929X.2015.1015166

Longo, L., Rusconi, F., Noce, L., and Barrett, S. (2012). "The importance of human mental workload in web design," in *WEBIST 2012-Proceedings of the 8th International Conference on Web Information Systems and Technologies*, Porto, Portugal, 18-21 April 2012, 403–409.

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, 13, 883321.

Luck, S. J. (2014). An Introduction to the Event-Related Potential Technique. *MIT Press*.

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, 8(5), 204-210. https://doi.org/10.1016/j.tics.2004.03.008

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674-693. https://doi.org/10.1109/34.192463

Miller, S. (2001). Workload Measures. Iowa City, IA: National Advanced Driving Simulator

Mitchell, D. K. (2000). Mental workload and ARL workload modeling tools. *US Army Research Laboratory: Aberdeen Proving Ground*, MD.

Murugappan, M., Ramachandran, N., & Sazali, Y. (2010). Classification of human emotion from eeg using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 03(04), 390-396. https://doi.org/10.4236/jbise.2010.34054

Myers, B. A. (1998). A brief history of human computer interaction technology. *interactions* 5(2), 44–54. doi: 10.1145/274430.274436

Rosanne, O., Albuquerque, I., Cassani, R., Gagnon, J. F., Tremblay, S., & Falk, T. H. (2021). Adaptive filtering for improved eeg-based mental workload assessment of ambulant users. *Frontiers in Neuroscience*, 15, 611962.

Roy, R. N., Hinss, M. F., Darmet, L., Ladouce, S., Jahanpour, E. S., Somon, B., ... & Lotte, F. (2022). Retrospective on the first passive brain-computer interface competition on cross-session workload estimation. *Frontiers in Neuroergonomics*, 3, 838342.

Senhadji L., G. Carrault, J. J. Bellanger and G. Passariello, "Comparing wavelet transforms for recognizing cardiac patterns," in *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 2, pp. 167-173, March-April 1995, doi: 10.1109/51.376755.

Splawn, J. M., & Miller, M. E. (2013, September). Prediction of perceived workload from task performance and heart rate measures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57*(1), 778-782. SAGE Publications. https://doi.org/10.1177/1541931213571170

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1), 61-78. https://doi.org/10.1175/1520-0477(1998)0792.0.co;2

Uhlhaas, P. J. and Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, 11(2), 100-113. https://doi.org/10.1038/nrn2774

Vural, M. (2018). *Online detection of pilot workload by using FNIR sensors* (Master's thesis, Department of Information Systems, Graduate School of Informatics, Middle East Technical University).

Wickens, C. D. (1984). Processing resources in attention. In *R. Parasuraman & D. R. Davies (Eds.), Varieties of attention* (pp. 63-102). Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science,* 3(2), 159-177.

Wickens, C. (2008). Multiple resources and mental workload. *Human Factors the Journal of the Human Factors and Ergonomics Society*, 50(3), 449-455. https://doi.org/10.1518/001872008x288394

Wilson, G. and Russell, C. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors the Journal of the Human Factors and Ergonomics Society*, 45(3), 381-389. https://doi.org/10.1518/hfes.45.3.381.27252

Zarjam, P., Epps, J., & Lovell, N. H. (2015). Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload. *IEEE Transactions on Autonomous Mental Development*, 7(4), 301–310. doi:10.1109/tamd.2015.2441960

Zhiwen, Z., Duan, F., Solé-Casals, J., Dinarès-Ferran, J., Cichocki, A., & Yang, Z. (2019). A novel deep learning approach with data augmentation to classify motor imagery signals. *IEEE Access*, 7, 15945-15954. https://doi.org/10.1109/access.2019.2895133

APPENDIX

This section provides the precision, recall and F-1 ratio statistics for the 5-fold cross validation and session-based train/test split models. All models were trained with Efficientnet-b0 with the Adam optimizer and pretrained weights.

	Easy			Medium			Difficult			Macro-Average		
	Precision	Recall	F1-ratio	Precision	Recall	F1-ratio	Precision	Recall	F1-ratio	Precision	Recall	F1-ratio
5-Fold	0.718	0.781	0.748	0.49	0.427	0.456	0.532	0.554	0.543	0.58	0.587	0.583
Session 1	0.832	0.772	0.801	0.578	0.397	0.471	0.541	0.749	0.628	0.65	0.639	0.633
Session 2	0.709	0.7	0.704	0.41	0.327	0.363	0.463	0.562	0.508	0.527	0.53	0.525
Session 3	0.676	0.825	0.743	0.467	0.342	0.395	0.562	0.587	0.574	0.568	0.585	0.571

Figure 7.1: Precision, Recall, F1-Ratio for 5-Fold Cross Validation, Session 1, Session 2, and Session 3 Tests

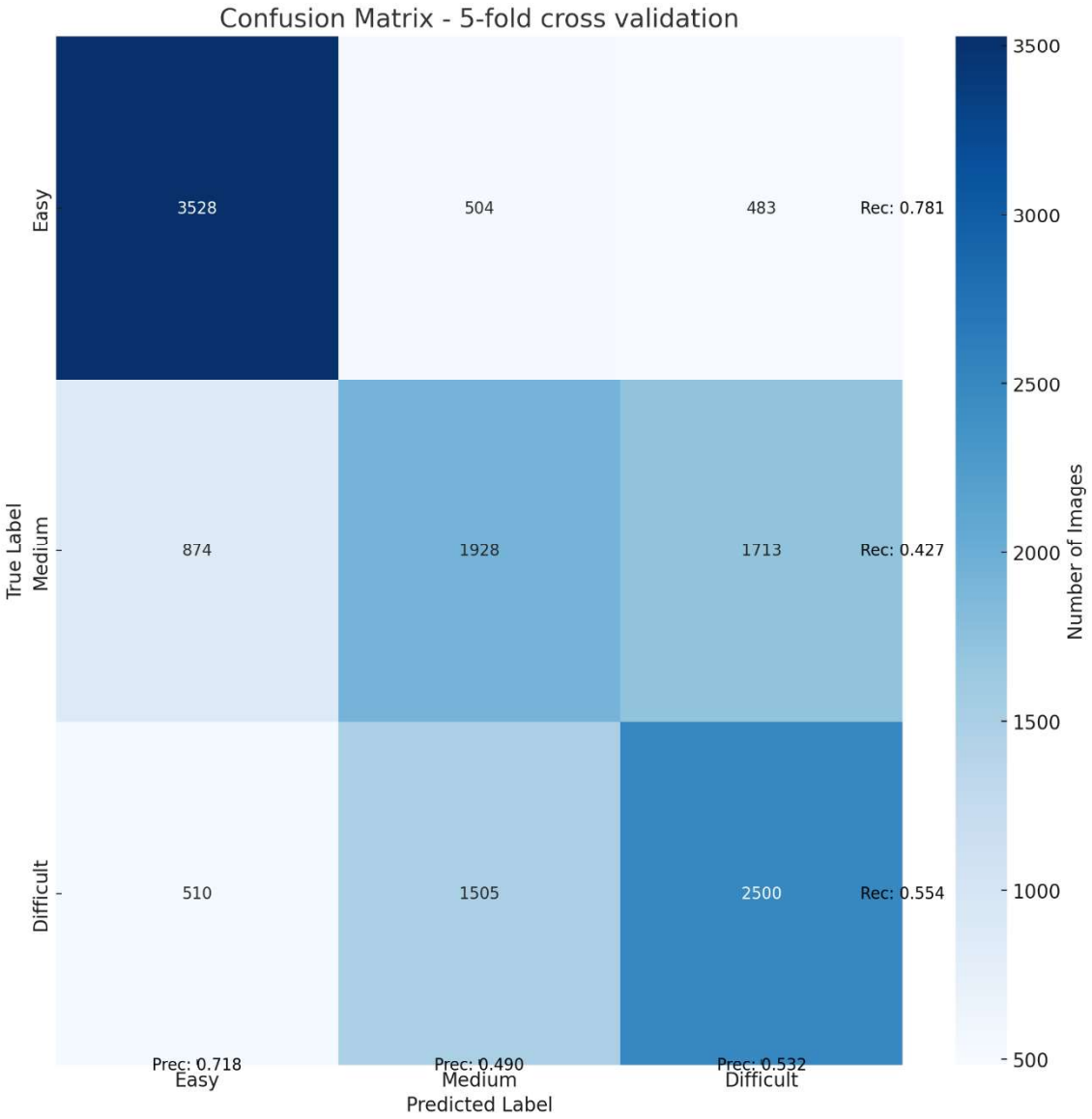

Figure 7.2: Confusion Matrix for 5-Fold Cross Validation

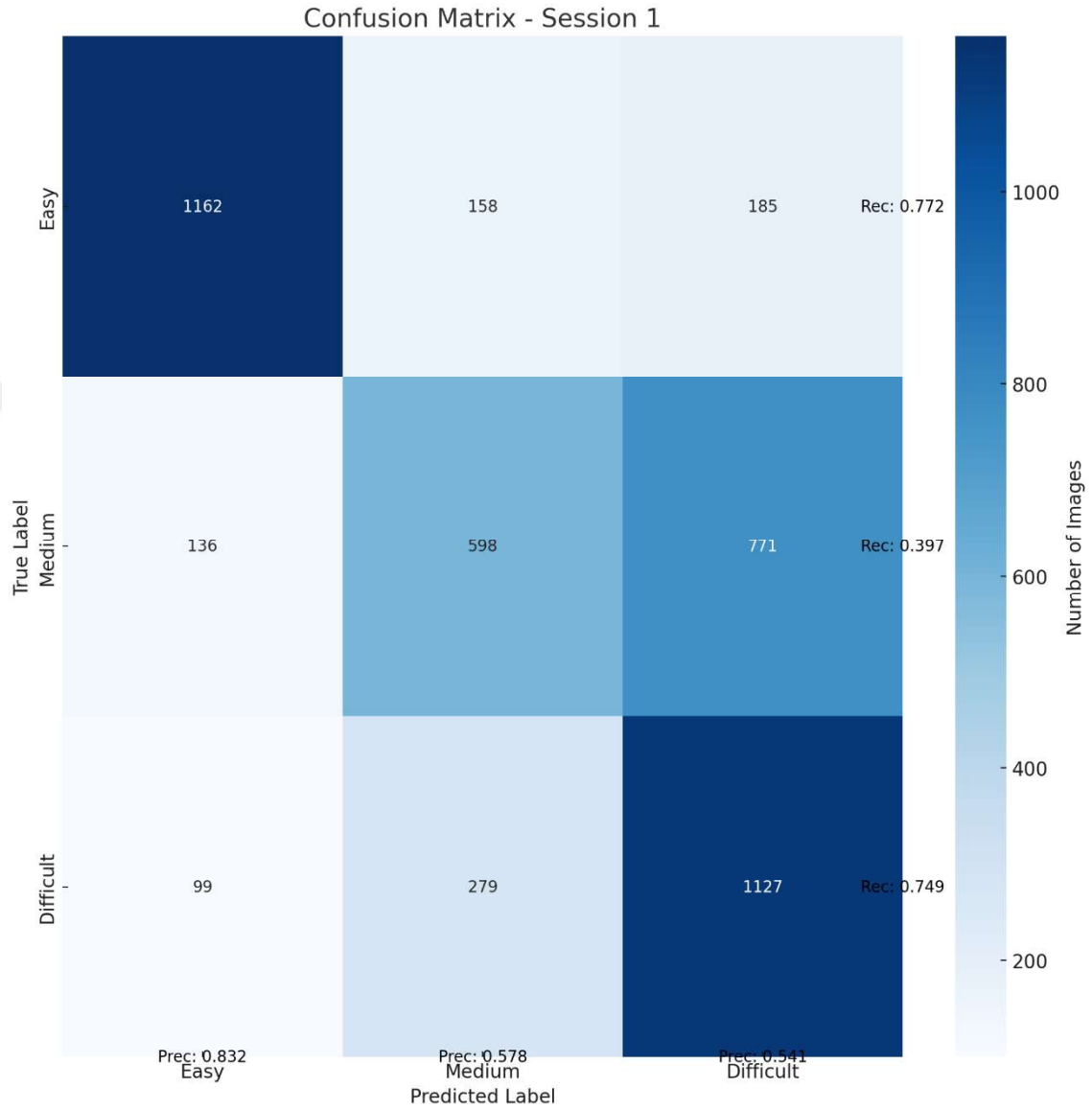

Figure 7.3: Confusion Matrix for Session 1 Test (Session 2, and 3 is used for training)

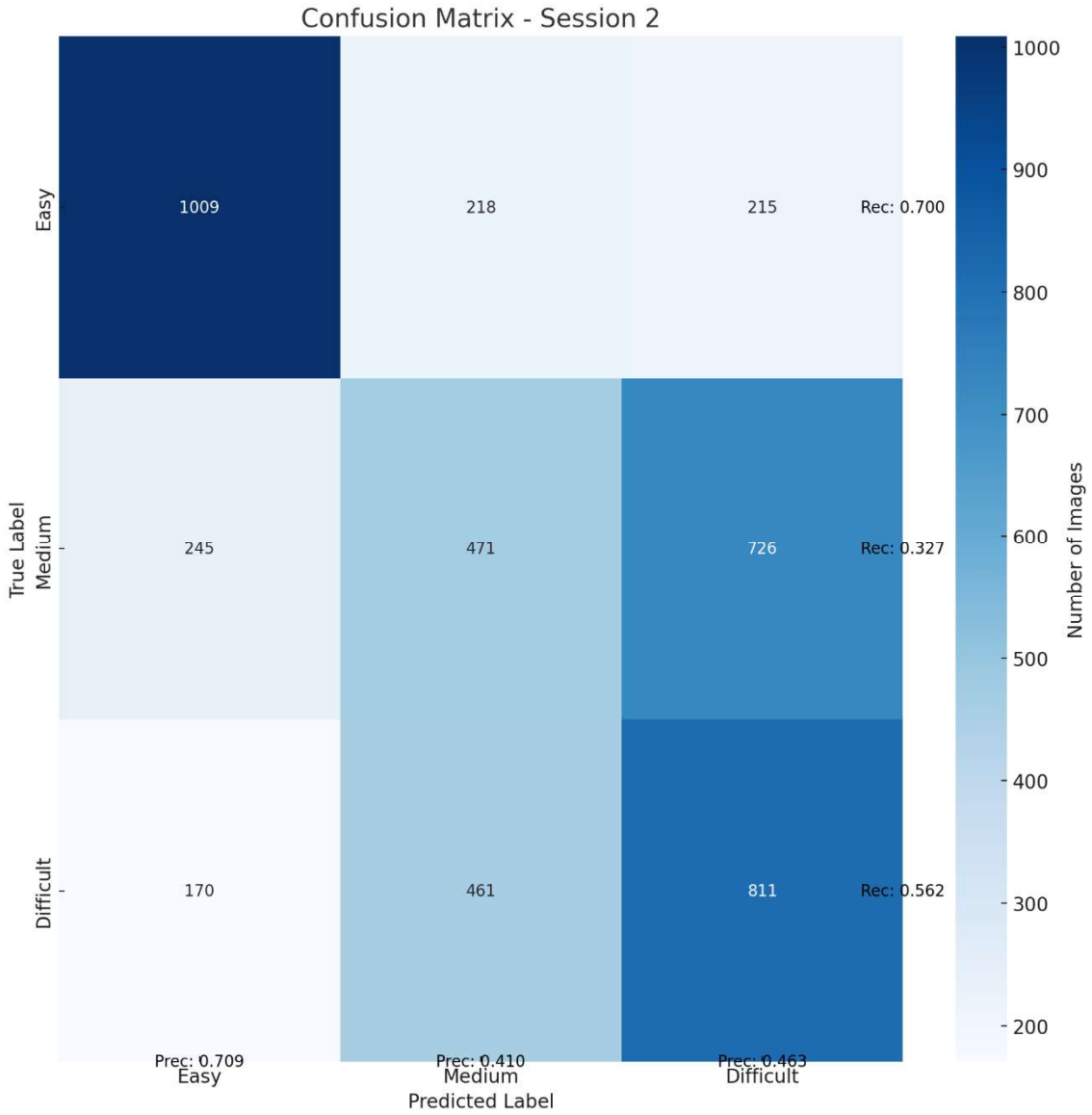

Figure 7.4: Confusion Matrix for Session 2 Test (Session 1, and 3 is used for training)

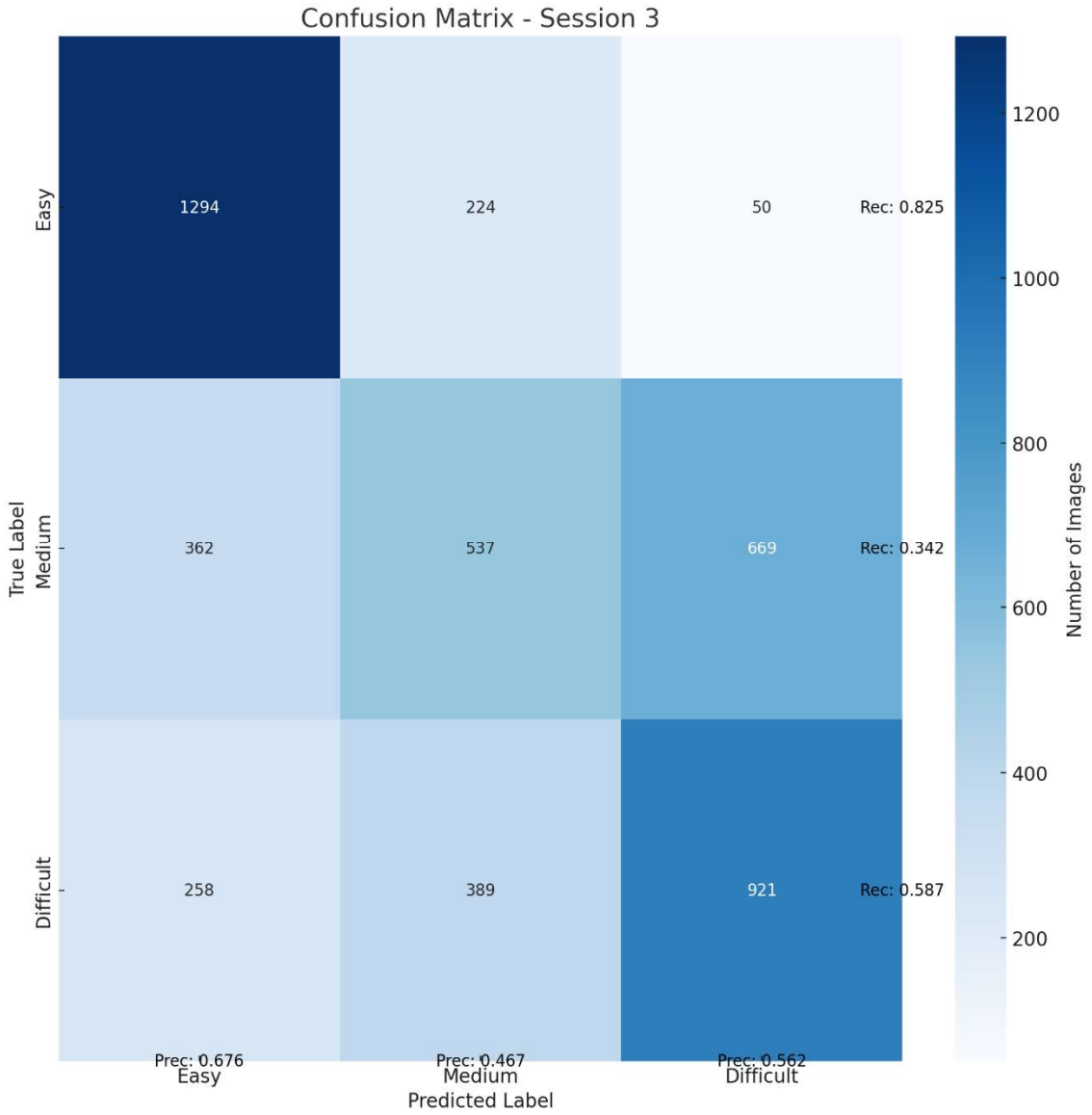

Figure 7.5: Confusion Matrix for Session 3 Test (Session 1, and 2 is used for training)