Deep Learning Classification Model of Mental Workload Levels using EEG Signals

PhD Thesis

Kunjira Kingphai

NeuraSearch Laboratory

Computer and Information Science

University of Strathclyde, Glasgow

May 8, 2024

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

Understanding and improving humance performance, especially in situations that require safety, productivity, and well-being, relies on categorising mental workload (MWL). Traditional methods for measuring MWL, such as in driving and piloting, have given us some understanding, but these methods must accurately distinguish between low and high workload levels. Excessive work can tyre participants, while insufficient work can make them bored and inefficient.

Traditional MWL assessment tools, such as questionnaires, sometimes make it harder for people to manage their MWL, especially when they struggle to express or understand their thoughts and feelings. The recent work shift to neurophysiological signals, specifically electroencephalogram (EEG), provides a promising way to measure brain activity related to MWL non-invasively. Advanced techniques such as deep learning have made it easier to study EEG signals in more detail.

Our goal was to develop a clear and consistent approach for using EEG signals to classify MWL effectively. Our approach focused on each process stage, from preparing the data to evaluating the model and addressing common mistakes and misunderstandings in current techniques.

The first study addresses the challenges of using EEG data contaminated by artefacts for assessing MWL. EEG signal artefacts, such as eye movement or muscle activity, can skew MWL assessment. Recently, there has been significant progress in using deep learning models to interpret EEG signals, but the challenge remains. The preprocessing pipeline for EEG artefact removal is broad and inconsistently adopted; some pipelines are time-consuming and contain human intervention steps, so they are unsuitable for automation systems. Therefore, this study focused on automatic EEG artefact removal for deep learning analysis. Furthermore, we examined the impact of various preprocessing techniques on the effectiveness of deep learning models in classifying MWL levels. We used state-of-the-art models such as Stacked LSTM, BLSTM, and BLSTM-LSTM, and found that certain techniques—specifically, the ADJUST algorithm—significantly enhanced model performance. However, the sophisticated models could extract relevant information from raw data, indicating a reduced need for preprocessing.

The second study shifted the focus to channel selection to refine the automation of MWL classification and reduce unnecessary computational expenses by using unnecessary electrodes, aligning more closely to real-world applications. We prioritised the best electrode setup focusing on brain activity related to MWL. We removed unnecessary data using Riemannian geometry, an effective method for EEG channel selection. We aimed to balance information sufficiency with computational efficiency and to reduce the number of electrodes. The study also evaluated covariance estimators for Riemannian geometry for their effectiveness in channel selection and impact on deep learning models for MWL classification, as the traditional Empirical Covariance (EC) has limitations for the EEG signal.

Finally, the third study tackled a critical but frequently overlooked aspect of MWL-level classification using machine learning or deep learning techniques: the temporal nature of EEG signals. We underscored that the traditional cross-validation technique violates the sequential nature of time series data, leading to data leakage, model overfitting, and inaccurate MWL assessment. Specifically, to predict the subject's MWL level, we could not randomly split data and use future data to train the model and predict the previous MWL level. To address this problem, this study focused on the model training phase, specifically on the importance of time series cross-validation methods. We adopted the expanding window and rolling window strategies, finding that using the expanding window strategy outperformed those using the rolling window strategy.

This research carefully developed a comprehensive and consistent method for classifying MWL using EEG signals. We aimed to correct misunderstandings and set a standard in brain-computer interface (BCI) systems. This will help guide future research and development efforts.

Contents

List of Figures

List of Tables

Acknowledgements

My PhD work could not have been completed without the assistance of others. Individuals who provide direction and counsel and are consistently present for me constitute an exceptional support network. To begin with, I express my sincere gratitude to my supervisor, Dr. Yashar Moshfeghi, for his continuous valuable guidance, assistance, and unwavering support, whether it is a suggestion for an experiment or research. All the experience I get from conducting my PhD work helps me be a better researcher who can work independently and more professionally. Furthermore, my supervisor also helped me overcome obstacles and stay engaged and productive in my research. Nonetheless, his enriching knowledge and research ideas that my supervisor shared with me could help me progress my PhD work and shed light on future research. I am also grateful to my supervisors from my bachelor's and master's degrees for their continuous encouragement and belief in my potential. Thanks to all of my colleagues in Thailand for their inspiring and shared experiences. I am appreciative of the Department of Computer and Information Sciences at the University of Strathclyde, particularly the NeuraSearch Laboratory, and all the students I met there. Their warm welcome and support in Scotland were remarkable. Special thanks to my PhD friends from Thailand, particularly Natee Tanman, for their companionship and emotional support during challenging times, including the pandemic. Sarawoot Kongyong deserves thanks for enriching discussions which have positively impacted my work. My PhD studies were made possible by financial support from the Thai government, which enabled my educational and international experiences. Lastly, I want to give my heartfelt thanks to my parents for their unwavering love, support, and belief in me. Their values and encouragement have been extremely important to me during this journey.

Chapter 1

Introduction

Mental workload (MWL) is a critical factor in many aspects of human life, including attention disorders in children [125], study design [168, 256], driving fatigue [88, 241], and task performance [230]. MWL arises from a variety of factors, such as multitasking [134]. When individuals engage in multiple tasks that require visual and auditory input simultaneously, their MWL increases [6]. Some multitasking tasks, such as walking and talking, are more manageable for people in normal health than others, such as using a phone while driving, which can significantly increase MWL and even cause road traffic incidents [162, 207]. Other factors that can strain attention and increase MWL include the complexity of a task, time pressure, and environmental distractions [14]. To cope with these demands, individuals use skills such as memory, planning, and experience [215].

The relationship between MWL and performance is complex and nonlinear. When faced with more demanding tasks, people tend to increase their effort and use more effective strategies to meet the challenge [82]. However, this compensatory behaviour has limits; too much workload can lead to distractions, decreased processing capacity, and divided attention. On the other hand, too little workload can lead to inattention, decreased alertness, and even drowsiness [205]. Therefore, the optimal MWL is neither too high nor too low, but rather at a level where performance is best [69].

To measure MWL level and determine whether it is too low, at a good level, or too high, we can use specific measurement tools, such as performance-based measures, subjective measures, physiological measures, and neurophysiological measures [173].

Performance-based measures assess performance on a task or set of tasks, such as the time it takes to complete a task or the number of errors made. A decrease in performance can indicate a high MWL [25], but performance-based measures can be affected by other factors, such as motivation and fatigue. Therefore, it is important to use them in conjunction with other measures, such as subjective measures and neurophysiological measures [218].

Subjective measures assess the participant's own perception of their MWL using a questionnaire. The most commonly used questionnaires include the Task Load Index (NASA-TLX) [71], Subjective Assessment Technique (SWAT) [173], and the Workload Profile [208]. These multidimensional questionnaires measure the overall workload during task performance. They require participants to evaluate and articulate their workload. However, subjective measures have some limitations. The boundary between too low and too high MWL is often blurred for some people, making it difficult to determine if the workload is excessive or inadequate [237]. Additionally, self-reporting can be complex, difficult to understand, and influenced by the participant's competence, talents, and effort, potentially increasing their MWL [131].

While physiological measures, such as electrooculography (EOG) [26, 66], electrocardiogram (ECG) [74, 245] heart rate, blood pressure, and skin conductance, are used to assess the body's physiological responses to stress [38], they have limitations. For instance, EOG and ECG are non-invasive and portable, but they are not directly related to brain activity [58, 61]. Moreover, changes in physiological measures can be caused by physical exertion, emotional arousal, and environmental stressors [194], making it challenging to distinguish between MWL and other sources of physiological arousal.

Despite these challenges, MWL assessment remains a valuable tool for researchers aiming to elucidate its characteristics. Thus, many have turned to neurophysiological measures to assess the activity of the brain and nervous system. Specifically, brain signal activity has been evaluated using various neuroimaging techniques such as magnetoencephalography (MEG) [200], functional magnetic resonance imaging (fMRI) [126], functional near-infrared spectroscopy (fNIRS) [75], and notably, electroencephalography (EEG) [26].

Each neurophysiological signal has its own set of advantages and limitations. For example, MEG and fMRI are capable of measuring brain activity and have high temporal and spatial resolution, respectively. Yet, they are not suitable for all environments, and they are not only cumbersome and expensive but also require specialised equipment [130]. fNIRS, which is relatively inexpensive and portable, can measure brain activity in different brain regions. However, it has low spatial resolution and is prone to artefacts from blood flow and movement. EEG, which is also portable, can measure brain activity with a high temporal resolution, making it ideal for detecting subjects' MWL levels in real-time. Additionally, when considering response time, EEG is generally superior to fNIRS. Among these neurophysiological signals, EEG is frequently preferred in human-computer interaction contexts with regard to its non-invasive nature and high temporal resolution, allowing for millisecond-scale measurements [117]. Its popularity is further enhanced by its strong correlation with a person's real-time MWL status [204].

EEG data is unique in that it is time series data [105], meaning that it consists of a sequence of data points recorded at successive, equally spaced intervals in time. This property captures the dynamic shifts in the brain's electrical activity over time [105]. Time series data, particularly EEG data, has the potential to yield valuable insights into underlying brain functions because it can capture temporal patterns and trends. This characteristic is particularly important in tasks where continuous monitoring of brain activity is needed, such as detecting seizures [181], monitoring sleep stage [90], or assessing cognitive workload [46]. In order to accurately analyse and interpret EEG data, it is crucial to have a good grasp of its time series nature and be able to utilise appropriate analysis methods accordingly.

Traditional machine learning techniques have been used to predict MWL level from EEG signals, such as linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbours (KNN), and random forest remain effective baselines in the literature [76, 180, 187]. These models offer the advantage of being generally easier to train and interpret than deep learning models. However, they might face challenges when dealing with the complex, non-linear relationships of EEG data. Artificial neural networks (ANNs) are the main component of deep learning, a specialised subfield of machine learning that takes cues from the structure and function of the human brain [232]. This method is particularly effective for monitoring, forecasting, and managing MWL in real-time because it can extract complex correlations from huge amounts of information. Deep learning models can be used to identify people's MWL, making systems more flexible and effective. So, deep learning models have recently emerged as a promising alternative. This enables continuous measurement and classification of MWL, paving the way for more responsive and adaptive brain-computer interfaces (BCI) in the future. Crucially, distinguishing between different MWL levels—low, medium, and high—is essential for measuring the effectiveness of BCI [108].

Traditional machine learning techniques, such as linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbours (KNN), and random forest [76, 180, 187], have been used to predict MWL levels from EEG signals. These models offer the advantage of being generally easy to train and interpret, but they do struggle to handle the complex, non-linear patterns in EEG data. Deep learning, a specialised subset of machine learning inspired by the human brain, utilises artificial neural networks (ANNs) [232]. The model excels in real-time monitoring, forecasting, and managing MWL due to its ability to extract complex correlations from large amounts of data. As a result, deep learning models have attained significant progress in MWL level prediction, advancing the adaptability and performance of systems [85, 98, 230]. These models are becoming a viable alternative, offering continuous MWL measurement and classification and setting the stage for the evolution of more intuitive and flexible BCI [108]. Recent advancements in deep learning have led to the development of sophisticated models capable of discerning subtle fluctuations in EEG signals, with the primary aim of accurately categorising MWL levels. Models such as the Recurrent Neural Network (RNN) [116], Long Short Term Memory [109], and Bidirectional Long Short-Term Memory-Long Short Term Memory (BLSTM-LSTM) [36] are leading the way in this domain.

However, employing EEG signals in deep learning is not straightforward. A major challenge is the susceptibility of EEG signals to noise and other disturbances [223], which can originate from both physiological and external sources [87]. Given the precision and sensitivity of EEG, maintaining data integrity becomes paramount. Artefact removal is, therefore, an indispensable preprocessing step, meticulously filtering out anomalies to preserve the data's authenticity and ensure a superior signal-to-noise ratio, which bolsters the performance of subsequent analytical models [209]. Despite the importance of artefact removal, numerous studies in deep learning that employ EEG signals often neglect to follow a standardised protocol for data cleaning. This lack of uniformity in data preprocessing raises two primary concerns. First, it makes it difficult to accurately assess the true effectiveness of deep learning models, as the results may be confounded by differences in the data preprocessing methods used. Second, it prevents researchers from comparing findings across different studies, even when they have used data from the same experiments. Therefore, the ongoing pursuit of a standardised approach for removing artefacts from EEG data in deep learning applications remains a significant and unresolved challenge.

Building models for MWL classification is challenging [234]. In deep learning, too much data can cause overfitting, where models perform well on the training data but poorly on new data. Conversely, too little data can prevent models from learning the underlying patterns effectively [55]. This balance between data abundance and scarcity is especially relevant when considering the sources of the data. In the context of EEG signals acquisition, data can be captured using portable devices or EEG caps with multiple electrodes (channels) [213]. Introducing too many variables or features from similar sources can introduce redundancy in deep learning, leading to issues such as multicollinearity [37]. Therefore, for optimal results, it is crucial to select only the EEG channels that are specifically relevant to MWL, which can improve both data quality and the efficiency of model training.

The challenge of redundancy data in deep learning models can often lead to overfitting [55], especially during the evaluation phase. To tackle this problem, cross-validation techniques are typically adopted. However, with different datasets and study objectives, each strategy needs to be carefully considered. In the case of EEG signals, which are time series data, traditional CV approaches such as shuffling and random splitting can lead to an unreliable model due to overfitting [35]. This challenge is even more pronounced in forecasting tasks, where the model must not be able to see the future data during training. Therefore, a tailored cross-validation strategy for time series EEG data in MWL level classification is urgently needed.

In the following chapters, we explore these facets in depth, aiming to advance our understanding of MWL through EEG data with sophisticated computational techniques. This research aims to identify MWL levels experienced by users during various cognitive tasks using EEG signals. It highlights the potential of EEG as an objective, real-time tool for MWL assessment. The study focuses on several aspects, including EEG preprocessing, deep learning model application, and thorough model evaluation using time series cross-validation.

1.1 Research Motivations and Aims

The increasing prevalence of high-stress professions and complex tasks requires accurate and objective MWL assessment. Consequently, there is a compelling need for non-invasive, objective methods to measure MWL. As mentioned earlier, existing methods to assess MWL, such as performance-based metrics, subjective evaluations, physiological indicators, and neurophysiological metrics [173], each have their strengths and weaknesses. Among these tools, EEG stands out as one of the most promising tools for MWL assessment due to its non-invasive nature and ability to measure brain activity directly. Inspired by the successes of deep learning across various fields [72,103,172,233], our objective is to leverage deep learning to analyse EEG, which is inherently dynamic and temporally nuanced, posing challenges for the precise classification of MWL levels. As such, our primary aim is to accurately classify individuals' MWL levels using deep learning models that utilise EEG signals. The development of such a model has the potential to greatly enhance how MWL is monitored in various professional fields, leading to improved productivity and well-being. In this section, we aim to create a deep learning model that can effectively distinguish between different levels of MWL using EEG data.

Utilising EEG signals effectively requires the removal of unwanted noise or artefacts. However, lacking a universally accepted pipeline for this process can be problematic, especially for those without specialised expertise. This is due to the inconsistency that arises as each researcher might use a distinct method in their pipeline [1, 23, 49, 63]. Determining the optimal method for data processing can be challenging and may lead to a decline in confidence in disseminating outcomes. This is because of the uncertainty surrounding the appropriateness of noise removal techniques or the possibility of excessive data cleansing. Moreover, methods that rely heavily on human involvement may be time-consuming, create biases that impact EEG data reliability [210] and are not optimal for routine usage [100]. Nonetheless, in the world of deep learning, it can be challenging to compare model performances due to the lack of a standardised pipeline. Although deep learning models are capable of processing raw datasets efficiently, the presence of noise can often be misleading. For example, during a high MWL task, if a subject moves simultaneously, the model may mistakenly interpret the noise as relevant data, leading to misclassification.

To address these issues, this study aims to create an automated pipeline that can efficiently and seamlessly remove artefacts in EEG data without human intervention. With this approach, we aim to provide a user-friendly tool that can standardise and simplify EEG data processing, thereby making it accessible to researchers across various fields. Additionally, this approach enables us to evaluate the impact of each preprocessing step on model performance and compare the performance of different models.

As mentioned earlier, traditional EEG recording for achieving high accuracy in EEG signal classification typically involves using an electrode-rich cap channel recording [213]. However, this method has its limitations when it comes to real-world applications. The bulkiness and potential inconvenience of the technique can make it sub-optimal for everyday use [5], and there is also a risk of capturing redundant or irrelevant data [5]. Moreover, research shows that indicators of MWL are usually localised, especially within the prefrontal cortex during sensory, motor, or cognitive activities [8]. Given these challenges, it becomes evident that a more strategic approach to EEG channel selection is necessary.

Therefore, our primary objective is to investigate and analyse Riemannian channel selection methodologies that offer the inherent ability to handle the space of covariance matrices. Riemannian geometry serves as the inspiration for this approach. Additionally, we aim to identify the optimal covariance estimator within the Riemannian framework that yields channels best suited for predicting MWL levels. We also aspire to determine the optimal EEG channel configuration that maintains high accuracy comparable to all available channels.

When it comes to evaluating deep learning models, it is crucial to consider the cross-validation approach used. However, standard cross-validation techniques can be problematic when applied to EEG data, as they tend to disrupt the temporal structure of the data and lead to biased performance metrics. The underlying issue is that EEG signals are time series data and exhibit temporal dependencies, which are not accounted for by traditional cross-validation techniques that assume data independence and identical distribution (i.i.d.) [80]. This can result in models that are unreliable and fail to reflect real-world conditions [35], particularly in forecasting tasks where future information should not be accessible during training. To address this challenge, our study also focuses on incorporating time series cross-validation into the model training process for EEG data. By maintaining the temporal integrity of the data, we aim to develop more accurate and reliable models for MWL prediction. This advancement could significantly improve the effectiveness of EEG-based models and pave the way for more robust deep learning models in the future.

1.2 Thesis Statement

The overarching goal of this research is to enhance the accuracy of MWL detection using EEG technology. This study aims to develop a comprehensive deep learning-based approach that addresses key challenges in EEG signal processing, model architecture, and validation methods to advance the field of EEG-driven MWL detection.

1.3 Research Objectives

MWL detection using EEG signals involves complex signal processing challenges and requires advanced classification techniques. This study focuses on leveraging deep learning advancements to improve the accuracy and reliability of MWL classification. The specific research objectives are:

1. to review and synthesize the existing literature on deep learning applications in EEG signal analysis to identify optimal input configurations and address classification challenges.
2. to enhance EEG signal processing by developing and implementing artifact removal techniques that improve the accuracy of MWL detection.
3. to determine the optimal number of EEG channels necessary for accurate classification and evaluate the effectiveness of different covariance matrix estimators in channel selection.
4. to investigate and compare the efficacy of various deep learning models, including Stacked LSTM, BLSTM, BLSTM-LSTM, Stacked GRU, BGRU, BGRU-GRU, and CNN, in predicting session-specific MWL levels.
5. to design and apply rigorous cross-validation methods tailored for EEG data to ensure the reliability and validity of the classification models developed.

Each of the experimental **Chapters 4 - 6** contains research questions specific to a particular investigation, and together, they contribute toward answering some of the overarching objectives. Finally, **Chapter 7** summarises our efforts to meet the objectives of this thesis, as well as the limitations of our work and directions for future research.

1.4 Publications

The research that resulted from this PhD has been published at or submitted to peer-reviewed venues. Each paper has a direct link to a particular chapter where the content of that paper is thoroughly discussed.

1. Kingphai, K. and Moshfeghi, Y., Mental Workload Assessment Using Deep Learning Models from EEG Signals: A Systematic Review, IEEE Transactions on Cognitive and Developmental Systems (TCDS), Submitted. The content of this paper is discussed in **Chapter 2**.

2. Kingphai, K. and Moshfeghi, Y., 2023, September. On channel selection for EEG-based mental workload classification. In International Conference on Machine Learning, Optimization, and Data Science (pp. 403-417). Cham: Springer Nature Switzerland. The content of this paper is discussed in **Chapter 5**.

3. Kingphai, K. and Moshfeghi, Y., 2022, September. On time series cross-validation for deep learning classification model of mental workload levels based on EEG signals. In International Conference on Machine Learning, Optimization, and Data Science (pp. 402-416). Cham: Springer Nature Switzerland. The content of this paper is discussed in **Chapter 6**.

4. Kingphai, K. and Moshfeghi, Y., 2022. EEG-based mental workload level estimation using deep learning models. In: Ergonomics & Human Factors 2022, Birmingham, UK: The Chartered Institute of Ergonomics & Human Factors (CIEHF), pp. 297-299. The content of this paper is discussed in **Chapter 4**.

5. Kingphai, K. and Moshfeghi, Y., 2021, September. Mental workload prediction level from EEG signals using deep learning models. In The 3rd Neuroergonomics Conference 2021. The content of this paper is discussed in **Chapter 5**.

6. Kingphai, K. and Moshfeghi, Y., 2021. On time series cross-validation for mental workload classification from EEG signals. In Neuroergonomics Conference. The content of this paper is discussed in **Chapter 6**.

7. Kingphai, K. and Moshfeghi, Y., 2021. On EEG preprocessing role in deep learning effectiveness for mental workload classification. In Human Mental Workload: Models and Applications: 5th International Symposium, H-WORKLOAD 2021,

Virtual Event, November 24–26, 2021, Proceedings 5 (pp. 81-98). Springer International Publishing. The content of this paper is discussed in **Chapter 4**.

1.5 Thesis Outline

This thesis is organised into subsequent parts, each corresponding to specific chapters.

Chapter 1 - Introduction. It provides the thesis outline and explains the motivation behind the thesis objectives. It also presents a thesis statement and overviews the research objectives and contributions.

Chapter 2 - Literature Review. This chapter provides a thorough background to the central themes of the thesis, namely, the classification of MWL using EEG signals and deep learning models. The chapter unfolds through the following sections:

Section 2.1 to **Section 2.1.7** begin with the general background of neurophysiological, which is the EEG signal in this study, and the definition of NeuraSearch is also described in **Section 2.1.10**. The initial step of EEG analysis—the preprocessing stage, setting the stage for future data examination—is provided in **Section 2.1.8** and **Section 2.1.9** focuses on channel selection methods. The background of deep learning models used in this study is provided in **Section 2.2**

The subsequent **Section 2.3** describes how we can assess MWL level using traditional methods and physiological and neurophysiological measurements. Lastly, **Section 2.4** presents a comprehensive literature review on how the signal can show subject MWL and the feasibility of using signal to predict MWL levels and **Section 2.4.2** explores how to evaluate models possessing temporal characteristics by employing cross-validation in machine learning.

Chapter 3 - Methodology. In this chapter, we begin with the details of the datasets used in this thesis, highlighting their characteristics and the tasks performed by the participants. The next section outlines the fundamental procedures applied across all sections of our thesis, focusing on the pivotal machine learning stages of data preprocessing, feature engineering, and model training. The methodology is structured as follows:

- Data Preprocessing: We provide a detailed discussion of our data preprocessing strategy, which is essential for our research.
- Feature Engineering: We exhaustively detail our approach to feature extraction, selection, and standardisation, as well as the meaning and formula of each feature.
- Deep Learning Model Evaluation: We present an overview of our evaluation techniques, emphasising the use of various cross-validation methods tailored to the unique attributes of each dataset and aligned with our experimental aims.
- Statistical Analysis: Our approach to descriptive and inferential statistics is discussed here, serving as the backbone for our hypotheses and the results of our experiments.

Chapter 4 - EEG Preprocessing and Its Effect on Deep Learning Models in MWL Prediction. In this chapter, we delve into the impact of various preprocessing techniques on the performance of deep learning models in predicting MWL using EEG signals. As EEG signals are susceptible to noise, we explore techniques such as high-pass filters, the ADJUST algorithm, and re-referencing. Our primary research question is to understand the effects of these techniques on the effectiveness of deep learning models in predicting MWL levels using EEG signals. To evaluate these techniques, we employ three state-of-the-art deep learning models - Stacked LSTM, BLSTM, and BLSTM-LSTM. These models are all variants of RNNs that capture temporal dependencies in sequential data. The "Bidirectional" variants process the data in both forward and backward directions, enabling the models to capture past and future information. The "Stacked" variants involve stacking multiple layers of the same model to create a "deeper" network. By effectively preprocessing the data, we can refine it for further analysis. (**Chapter 5 and 6**.)

Chapter 5 - EEG Channel Selection Enhancement with Covariance Estimators in Riemannian Geometry. In this chapter, we aim to investigate the measurement of MWL using EEG and optimise channel selection strategies to improve the computational efficiency and model performance of deep learning models for MWL classification. We focus on evaluating the effects of different covariance estimators on the Riemannian distance-based channel selection approach and their impact on various deep learning models.

Chapter 6 - Time Series Cross-Validation This chapter delves into evaluating models based on time series data obtained specifically from EEG signals. To accomplish this, we employ two time series cross-validation methods - the expanding and rolling windows. Within each strategy, we explore varying window sizes, which are crucial in influencing the sample size used for training the model. Our primary aim is to determine the most effective strategy and establish the optimal window size for cross-validation. This will provide us with valuable insights into the minimum amount of data required to predict a subject's MWL levels. This understanding is crucial, as our ultimate goal is to forecast a participant's MWL based on their records in the MWL task.

Chapter 7 - Conclusions and Further Work. The present concluding chapter provides an overview of the significant contributions of our thesis to the domain of EEG-based MWL classification. We highlight the key findings and acknowledge the limitations of our study. Additionally, we discuss the potential real-world applications of our research and propose future research directions.

Chapter 2

Background

In this chapter, we first explore neurophysiology, which has been used in the context of NeuraSearch. The background of EEG includes the definition and related aspects, and the advantages and limitations of EEG signals are explained in this chapter. Moving forward, we explore the initial step of EEG analysis—the preprocessing stage, which sets the stage for future data examination and focuses on channel selection methods. Additionally, we provided background on the deep learning models used in this study. We also describe the assessment of MWL levels using traditional methods and physiological and neurophysiological measurements. Furthermore, we present a literature review on how EEG signals can indicate subject MWL levels and the feasibility of using signals to predict MWL levels. Lastly, we explore ways to evaluate models possessing temporal characteristics by employing cross-validation in machine learning.

2.1 Neurophysiological

2.1.1 EEG

EEG technology has been pivotal in advancing our understanding of the brain's electrical activities since its development by Hans Berger in 1924, who termed it "Elektrenkephalogramm." He published his first paper detailing the recording of electrical activity in the human brain in 1929 [17]. Subsequently, EEG has been widely used in clinical settings for diagnosing neurological disorders such as epilepsy. The scope of EEG applications has broadened considerably over the decades. It has been used to understand brain activity and various sleep stages [195]. Between the 1950s and 1990s, technological advancements with more sophisticated recording equipment refined EEG's capabilities, allowing it to capture high-resolution temporal dynamics of brain activity, making it invaluable in both clinical and research settings [44].

In the modern invocation and application during the 2000s, digital signal processing and more advanced computational techniques have led to significant improvements in EEG analysis [184]. After that, the integration of EEG with other neuroimaging techniques, such as fMRI, has provided a more comprehensive understanding of the brain's structural and functional aspects [120]. Furthermore, the development of High-Density electrodes setups, which provide finer spatial resolution of brain activity, marked another advancement [156]. However, it is not practical in some applications; consequently, several studies have proposed lightweight, wearable EEG devices. These innovations have expanded the use of EEG from clinical settings to users' daily environments, facilitating continuous monitoring of brain health and functioning [34]. Advances in machine learning and artificial intelligence, especially deep learning, have also propelled EEG into new applications, including BCIs, enhancing EEG's analytical power [45, 84]. This progression illustrates EEG's versatility and adaptability to new scientific and technological demands.

EEG is a sophisticated method that records the brain's electrical activity using electrodes placed on the scalp. By analysing the resulting waveforms, we can gain insights into the functioning of the cerebral cortex, which plays a vital role in our thoughts, emotions, and behaviours [161]. EEG quantifies the electrical activity generated by the movement of electrical charges within the central nervous system, which is sustained by ionic gradients across neuronal membranes. When strategically placed scalp electrodes detect these subtle electrical signals, they indicate brain activity [202]. Once captured, these weak electrical signals will be amplified to a level where they can be analysed. The amplified electrical signals are subsequently converted into a digital format and stored in computer memory for further analysis. Acquiring this data from the scalp allows for examining the various brain waveform characteristics, such as frequency, voltage, morphology, and spatial distribution [204]. Figure 2.1 displays an EEG cap, which consists of electrodes placed on the scalp and connected to a computer that records signals.

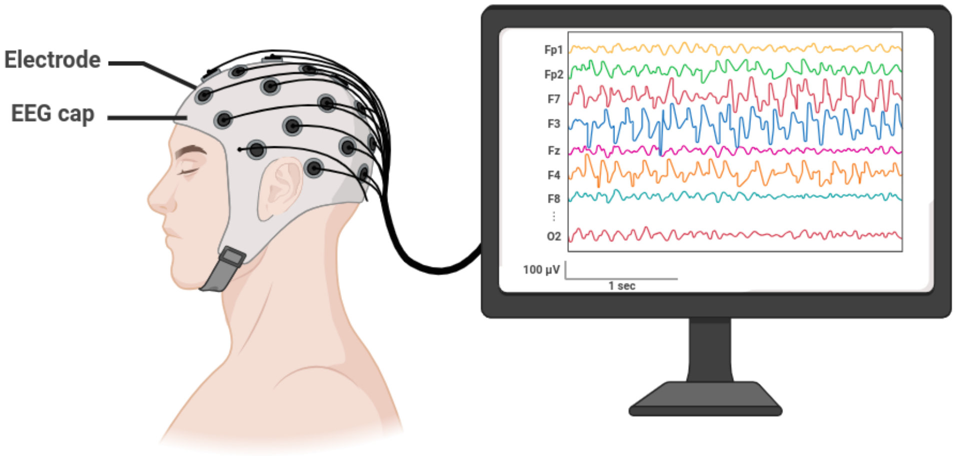


Figure 2.1: Illustrated diagram of EEG cap and brain signal

2.1.2 EEG Electrode Placement

The placement of these electrodes is critical for acquiring accurate and reliable data. Several standard electrode placement systems have been developed by researchers, including the 10/20 system, the 10/10 system, and the 10/5 system. The commonly used system is the 10/20 system; it is based on specific points of reference on the skull that are used to ensure consistent positioning of electrodes. The name of this system is derived from the distances between adjacent electrodes, which are either 10% or 20% of the total front-to-back or right-to-left distance of the skull [101, 114]. Initially, the 10/20 system involved placing 21 EEG electrodes [91]. The placement of the EEG electrodes according to the 10/20 system is shown in Figure 2.2

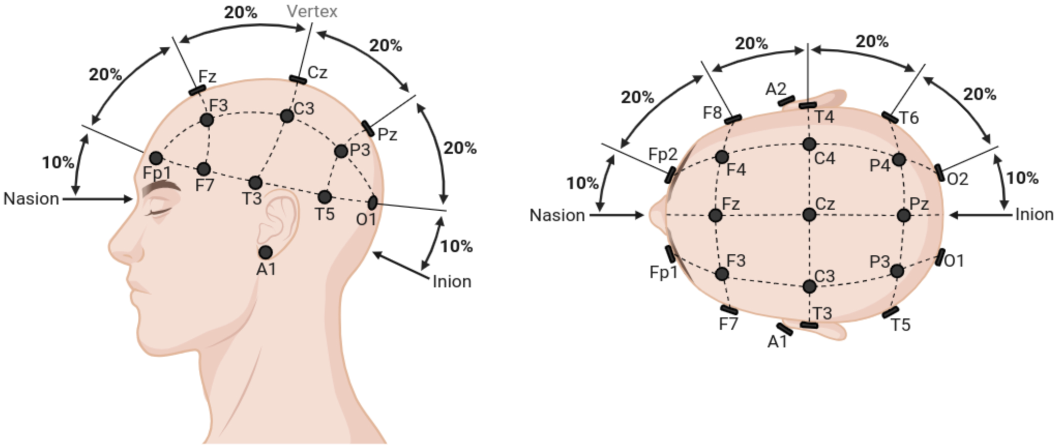


Figure 2.2: EEG electrode positioning (10/20 system)

In order to increase the channel density, a more fine-grained system known as the 10/10 system was proposed [39]; it is an extension of the 10/20 system. This system provides full coverage of the scalp with a higher density of 81 electrodes, achieved by adding 60 electrodes to the unmodified 21 electrodes. The additional electrodes are placed using a 10% division, which fills in intermediate sites halfway between those of the existing 10/20 system. This method ensures closely and evenly spaced electrodes, resulting in a more comprehensive and precise measurement of brain activity [39].

With advancements in EEG research, there has been a move towards even higher channel densities. Some studies have employed up to 256 channels to capture a more detailed picture of brain activity [188, 197]. Therefore, further refinement came with the introduction of the 10/5 system, which is designed for high-resolution EEG studies. This system allows for up to 345 electrode placements and uses proportional distances of 5% of the total length between skull reference points for electrode positioning. As a result, it is also called the 5% system or the 10-5 system [156].

2.1.3 EEG Electrode Types

EEG electrodes can be categorised into several types based on their material, design, and intended use. The common types of electrodes used in various applications are as follows.

1. Traditional Wet Ag/AgCl Electrodes (Wet) [193]. These electrodes require the application of a conductive gel that serves to bridge air gaps caused by hair or irregular scalp surfaces to reduce impedance and ensure a stable, high-quality connection between the electrodes and the scalp. The wet electrode typically provides lower impedance and better signal quality than dry electrodes. However, the preparation process can be more time-consuming and might cause discomfort or irritation for some individuals. These electrodes are the most commonly used in clinical and research EEG due to their stable and low-noise characteristics [77, 107].

2. Active Dry Single Gold Pin-Based Electrodes (BP Gold) [53] are designed to eliminate the need for conductive gel, simplifying the set-up process and enhancing individual comfort. This electrode consisted of a gold-coated single pin shaped like a mushroom. The gold pins gently penetrate through the hair to directly contact the scalp. These electrodes also include a built-in amplifier within the electrodes (active term). This amplifier helps boost the EEG signal at the source, reducing signal degradation caused by distance and external noise. Moreover, as it has a built-in amplifier, these electrodes are suitable for dynamic or mobile EEG applications such as ambulatory EEG monitoring or studies involving movement [3].

3. Passive Dry Solid-Gel Based Electrodes (BP Solid) offers a compromise between traditional wet electrodes and completely dry designs. These electrodes utilise a solid gel to establish direct contact with the scalp without requiring extensive skin preparation. Unlike the BP gold type, these electrodes do not contain built-in amplifiers, thus making them passive. They are easy to use and require minimal cleanup time. However, they rely on external amplification systems to boost

the EEG signal. BP Solid electrodes are suitable for standard EEG tests, sleep studies, and other scenarios where mobility is not a primary concern but where ease of use and patient comfort are valued. [53]

4. Hybrid Dry Multiple Spikes-Based Electrodes (Quasar) are an innovative type of electrode that combines the best features of dry and wet electrode technologies. They are made up of multiple tiny conductive spikes or micro-needles that gently penetrate the scalp and make direct contact without the need for conductive gel. As such, they are highly preferred in dynamic recording environments such as neurofeedback sessions, cognitive research, and mobile EEG monitoring [53]. However, it is important to note that due to the spikes, careful handling and maintenance are required [138].

Each electrode type has its own specific advantages and considerations, making them suitable for different applications and user preferences in EEG monitoring.

2.1.4 Electrode Labelling

EEG electrodes are placed around the head to detect electrical signals from the brain. These electrodes are carefully labelled to correspond to different brain regions. The brain has four main lobes, shown in Figure 2.3, but for the EEG, there is also the expansive pre-frontal area, which plays an important role in cognitive functions.

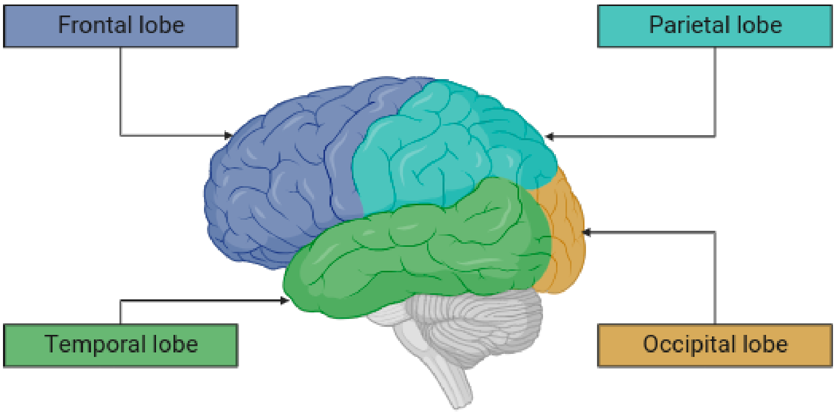


Figure 2.3: Brain lobes

The electrode names, such as pre-frontal (Fp), frontal (F), temporal (T), parietal (P), and occipital (O), help us understand which part of the brain they are recording from. For instance, a typical 21-electrode EEG system is based on the 10/20 system, consisting of electrodes such as Fp1, Fp2, F3, F4, F7, F8, Fz, T3, T4, T5, T6, C3, C4, Cz, P3, P4, Pz, O1, O2, A1, and A2 (M1, M2). The electrodes with "Z" in their names, such as FpZ, Fz, Cz, and Oz, are placed on the midline sagittal plane of the skull. The electrodes in the central area are represented by "C" and odd-numbered electrodes (1, 3, 5, 7) refer to electrodes placed on the left side of the head. In contrast, even-numbered electrodes (2, 4, 6, 8) refer to those on the right side. The "A" electrode, sometimes referred to as "M" for the mastoid process, refers to the bone found just behind the outer ear [91].

In the high-resolution 10/10 EEG system, the labelling of electrodes employs a two-letter combination system that represents intermediate contours between traditional placements of the 10/20 system. For example, electrodes between frontal-central are FC, frontal-temporal are FT, central-parietal are CP, and parietal-occipital are PO. Those between frontopolar-frontal are AF, and temporal-parietal is TP. Moreover, T3/T4 become T7/T8, while T5/T6 become P7/P8 [39]. Similarly, in the 10/5 EEG system, the naming follows this convention. Locations between the C and CP contours are labelled CCP, while the region between O and PO is designated POO [156].

The signal quality from each electrode can vary due to various factors. These can include the contact quality between the electrode and the scalp, the amount of hair in that area, and the impedance of the electrode. Some electrodes may need to be repositioned or excluded to ensure a good signal. Certain positions may be uncomfortable for long durations, negatively affecting signal quality.

2.1.5 EEG Recording

The process of EEG recording, whether using traditional caps or mobile headsets, can be broken down into several essential components. Firstly, electrodes with conductive media detect electrical signals from the scalp. These signals are then amplified using amplifiers to make them clearer and more accurate. Afterwards, the analogue signals or waveform of EEG signals are converted into a digital sequence of numerical values using an analogue-to-digital (A/D) converter [204]. Analogue signals are continuous waveforms, while digital signals are discrete and represented numerically with a limited set of possible values. To create digital signals, the ongoing waveform is periodically sampled, and each sample is discretised to correspond with a numeric value [216]. The sampling rate, typically measured in hertz (Hz) [133], is the conversion rate or the number of samples taken per second. For instance, a sampling rate of 512 Hz implies that 512 samples of data are taken every second, making the signals easier to store and process. Finally, the digital EEG signals are stored and displayed using a recording device.

2.1.6 EEG Waveform

EEG waveforms can be divided into different frequency bands, each associated with different states of brain activity. Common bands include Delta, Theta, Alpha, Beta, and Gamma, each with a specific range of frequencies. The Delta band, with a frequency of less than 4 Hz, is associated with deep sleep. The Theta band, with a frequency of 4-8 Hz, is associated with drowsiness, daydreaming, and memory consolidation. The Alpha band, with a frequency of 8-12 Hz, is associated with relaxed wakefulness and focused attention. The Beta band, with a frequency of 12-30 Hz, is associated with active wakefulness and thinking. Finally, the Gamma band, with a frequency of over 30 Hz, is associated with high-level cognitive processing and consciousness. Figure 2.4 illustrates these frequency bands and their typical applications in EEG analysis.

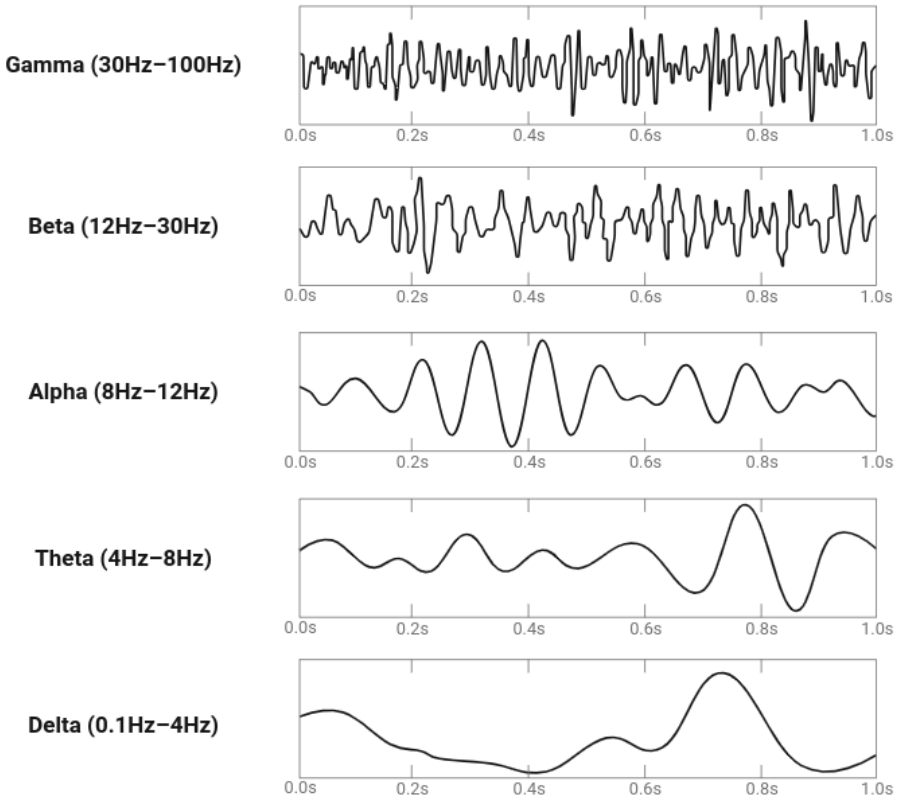


Figure 2.4: EEG frequency bands

2.1.7 Advantages and Limitations of EEG

Advantages

EEG is an exceptional tool with a high temporal resolution, enabling it to precisely track changes in brain activity to the millisecond [115]. This makes it an invaluable asset for real-time comprehension of sleep studies [164, 252], epilepsy monitoring [65], and cognitive processes and MWL in various applications [227]. One of the significant advantages of EEG, as elaborated in **Section 2.1.5**, is its non-invasive nature. It does not require penetration into the body, and the simple application of a cap or mobile device equipped with electrodes to the participant's scalp suffices for data collection. This ease of use, combined with the portability of EEG systems, allows for their application in various settings, from clinical [77] to field environments [83]. Furthermore, unlike techniques that involve ionising radiation, which can be harmful to participants' bodies, such as Computed Tomography (CT) or X-rays [22], which must be used judiciously due to potential risks, EEG is remarkably safe and can be utilised across a broad demographic spectrum. It is suitable for everyone from infants [157] to the elderly [217], accommodating individuals with health conditions [40, 231] and those in good health [50]. Overall, EEG's non-invasive nature, adaptability, and safety profile make it a versatile and powerful tool for various research applications.

Limitations

Despite its numerous advantages, EEG still faces certain challenges. One of the main obstacles is its lower spatial resolution compared to other modalities such as fMRI [250]. This is due to the electrical signals having to diffuse through the skull and other tissues before reaching the electrodes on the scalp. Additionally, EEG has to deal with noisy and time-varying signals, making the signal-to-noise ratio a critical factor in determining the quality of the recording [99]. The electrical activity generated by the neurons in the brain is what EEG aims to measure, as explained in **Section 2.1**, but unwanted electrical activity (noise) can also interfere with the signal. This susceptibility of EEG signals to noise or unwanted disturbances is a significant issue. There are several sources of disturbances that can affect EEG recordings, including physiological and non-physiological activities. Physiological activities, including eye movements like blinks and ocular adjustments, generate electrical potentials captured through EOG. Similarly, muscle activities—examples being chewing, clenching, frowning, or eyebrow movements—are monitored using Electromyography (EMG). Additionally, the heart's electrical impulses are measured via an ECG or EKG [223]. Besides physiological sources, EEG signals are also susceptible to disruption by various external factors. For instance, instrumental interference can arise in EEG equipment due to electrode displacement, inadequate grounding, or cable movement. Additionally, electrical noise from external sources and electromagnetic interference from devices emitting radio waves, visible light, or microwaves can introduce artefacts. Accurate EEG readings, therefore, require a controlled environment to mitigate the impact of these electrical interferences. Moreover, the subject's body movements represent another significant source of potential artefacts [87]. Given the precision and sensitivity of EEG, maintaining data integrity becomes paramount. Thus, artefact removal is an indispensable preprocessing step. By meticulously filtering out these anomalies, the data's authenticity is preserved and ensures a superior signal-to-noise ratio, bolstering the performance of subsequent analytical models [209].

2.1.8 EEG Artefact Removal

As mentioned in **Chapter 1**, we can use EEG signals to predict a person's MWL levels accurately. One of the primary difficulties in accurately analysing EEG data is managing artefacts. Preprocessing is very important in EEG data analysis. Knowing how to deal with artefacts before using the clean signal for further analysis is crucial. This is especially important for people who are not experts in neuroscience. Recently, deep learning has successfully been used in EEG analysis due to its capacity to capture good feature representation from data [177]. While some researchers have used noise reduction techniques as part of their EEG preprocessing stage, the effectiveness of each of these techniques on deep learning models for MWL classification has not yet been investigated. This has resulted in a lack of a uniform framework to be followed, and in turn, makes the comparisons of such models impossible. For example, Kurnar et al. [116] have applied a band-pass filter in the raw EEG to remove unwanted signals and employed a deep recurrent neural network (RNN) to classify four levels of the cognitive workload. As a result, they have gained an average accuracy of 92.5% in their classification. The band-pass filter has also been adopted into Maneesh Bilalpur et al. [24] study. However, the range of frequencies has been set at a different value; in this study, it has been set between 0.1 and 45 Hz. Moreover, the authors have also further rejected noisy epochs by visual inspection. Finally, noisy ICA components corresponding to eye blinks and movements have been manually removed. The artefact-removed data has been used to classify two levels of MWL induced by acoustic parameters by using a deep convolutional neural network (CNN) classifier. The F1-score of 0.64 has been obtained. The other type of filter which has been applied is a notch filter. Zhang et al. (2019) [246] have applied a 50 Hz notch filter in their study and a Butterworth band-pass filter between 0.5 Hz and 50 Hz. Then, the EEG signals from all channels have been re-referenced to the average of two ear electrodes. However, the authors have left ocular artefacts in their data. A three-class MWL induced by n-back tasks with easy, medium, and hard difficulty levels has been categorised by the proposed two-stream neural networks (TSNN). The proposed model has achieved an average accuracy of 91.9%. Similar to the previous work, the Butterworth filter has been employed in [230] as well; however, it has been used with a low-pass cutoff frequency of 40 Hz. To correct the artefacts from eye movements, the authors have employed ICA in their preprocessing procedure. Then, the cleaned data is put into an ensemble deep learning model (EL-SDAE) for the binary MW classification problem. The model has achieved 92% accuracy in MWL recognition. As observed from the literature, deep learning researchers working with the EEG have attempted to remove noise from their data using the existing preprocessing techniques proposed by the neurobiologist. However, there is no single procedure of preprocessing that everybody is following. In particular, the band-pass filter technique, which seems to be the most commonly used tool, has been defined in a diverse range.

Another aspect is that even when people perform classification using the same dataset, they apply different preprocessing techniques in their experiments. For example, Lim et al. [128] have performed MWL classification using their own STEW dataset. In the artefact removal stage, the authors used a high-pass filter at 1 Hz, a notch filter, and artefact subspace reconstruction (ASR). Then, they re-referenced data to the average for removing artefacts from muscle movement and cleaning the noise. Authors have obtained 69% MWL classification accuracy from a Support Vector Regression (SVR) model. The STEW dataset has also been adopted into Chakladar et al. [36] study for MWL level classification as well. However, in this study, the authors have removed the artefacts from EEG signals using only a band-pass filter technique. The filter has allowed signals between 4 and 32 Hz to pass. Then, the preprocessed data has been fed into various models for estimating human workload levels. They have performed an analysis in two Tasks: 1) “No task” and 2)“SIMKAP-based multitasking activity”. The proposed Bidirectional Long Short-Term Memory- Long Short-Term

Memory (BLSTM-LSTM) model has outperformed other models in their study. The model has reached 86.33% and 82.57% classification accuracy for studies 1) and 2), respectively. Moreover, from the literature, it can also be observed that the effect of the preprocessing procedure has not been considered in the analysis. The presence of diverse preprocessing techniques among comparable datasets and the underappreciated impact of these procedures on the outcomes highlight the necessity for a methodical approach and additional exploration in this domain.

While artefact reduction is a critical step in ensuring the reliability and quality of the EEG signal, knowing which EEG channel to use to obtain data for analysis is also critical. A large number of EEG electrodes are used to record the signal, and it is possible that a signal from the same brain area may be picked up by numerous electrodes, making the data redundant and the same information overlapped between channels. Utilising too many channels can introduce analysis complexity, slow data transmission, and increase experimental costs. Furthermore, it can lead to inefficiencies and practical challenges in real-world applications [206, 229]. Conversely, relying on too few channels can be problematic. For instance, when applying Independent Component Analysis (ICA) on a limited number of EEG channels, the resulting components might merely represent mixed sources rather than individual ones [174]. In essence, the transformation with ICA in such cases might change one set of mixed sources into another, offering no real insight or benefit. Thus, an excessive or insufficient number of channels is considered unsuitable for EEG analysis. [255]. To avoid the problem of using either too many or too few channels. In this regard, channel selection is just as important as preprocessing approaches. We will go deeper into the significance and methodology of EEG channel selection in the following section.

2.1.9 Channel Selection

As described in **Section 2.1.2**, the number of electrodes or EEG channels can vary from 21 to 345, with each electrode corresponding to a specific brain region. The electrode is also referred to as a “channel”. The selection of the number of channel configurations is customised based on the specific objectives of the experiment. Every choice entails trade-offs. While some EEG systems offer a predetermined set of channels, in practical scenarios, researchers often must decide which channels to use, balancing relevance and minimising data redundancy. Effective channel selection methods for EEG data analysis are essential to enhance classification accuracy [97]. To perform channel selection, wrapper or filtering techniques can be used [7]. Wrapper techniques optimise a channel subset using classification accuracy as the primary measure [10]. For example, Mzurikwao et al. [152] used a wrapper strategy with a convolutional neural network (CNN) to select channels for decoding multiple motor imagery intention classes from four amputees. They achieved a classification accuracy of 99.7% with a CNN model trained on 64-channel EEG data, and channel selection based on weights extracted from the trained model resulted in 8-channel models with 91.5±% accuracy. Despite offering potentially high performance, these techniques can be computationally demanding, requiring the model to be retrained for each subset evaluated and carrying the risk of overfitting due to their inherent exhaustive search nature [10]. In contrast, filtering techniques evaluate subsets of channels a search algorithm generates using independent evaluation criteria, such as distance, dependency, or information measures, offer speed, independence from the classifier, and scalability [190]. These methods aim to maintain the accuracy achieved with all channels by training the model with an optimal channel set [9]. For instance, the mutual information maximisation technique proposed in [119] ranks EEG channels based on their correlation with class labels, which lowers classification error. Similarly, the normalised mutual information technique proposed in [221] selects an optimal subset of EEG channels for emotion recognition, achieving high accuracy with a sliding window approach and short-time Fourier transform. The sparse common spatial pattern algorithm proposed in [7] optimises channel selection under classification accuracy constraints and outperforms several other methods, achieving up to 10% improvements over three channels.

2.1.10 NeuraSearch

There is a study called NeuraSearch [149] which has recently been on the edge of integrating neurophysiological signals into information systems. The NeuraSearch aims to combine computer and information science knowledge with neuroscience and cognitive science knowledge. Specifically, the research in NeuraSearch aims to leverage neurophysiological signals to enhance the effectiveness and user-centricity of information systems and to enhance the understanding of the user's intentions and cognitive state across different scenarios by analysing the signals. Different brain signals, such as fMRI and EEG, have been used.

Various studies under the NeuraSearch umbrella have shown promising diversity and depth [142–144,158–160,166]. For instance, the research by Lamprou et al. (2022) [118] focuses on using fMRI to evaluate and understand natural language processing (NLP). This work offers valuable insights into how the human brain processes text semantically and the potential of using this understanding to improve NLP models. Similarly, Michalkova et al. (2023) [144] employed EEG to assess users' cognitive states in information systems, clarifying the impact of different levels of knowledge on search behaviour.

Moreover, NeuraSearch has been expanding into the domain of MWL, which is the key factor in various areas and tasks, impacting performance and outcome. Kingphai and Moshgefhi (2021) [109] utilised EEG to classify levels of mental workload and investigated the importance of preprocessing for EEG data when used for MWL classification. The results show that preprocessing substantially improves the classification accuracy of machine learning models over that of the non-processed EEG data. Moreover, they also identify the lack of a commonly adopted preprocessing pipeline within the community and propose their preprocessing pipeline, which more recent works have subsequently adopted. Their research also delved into the efficacy of time series cross-validation in enhancing the performance of machine learning models that analyse EEG data for MWL classification [111]. This work highlights the importance of maintaining the temporal nature of the EEG signal when evaluating the model. In summary, NeuraSearch represents a cutting-edge interaction between neuroscience and information technology, leading to significant advancements and understanding of how neurophysiological signals can enhance and improve communication between humans and computers. One pivotal neurophysiology signal that has been extensively utilised and explored within the scope of NeuraSearch is the EEG.

2.2 Deep Learning (DL)

Artificial neural networks (ANNs) are the main component of deep learning, a specialised subfield of machine learning that takes cues from the structure and function of the human brain. This method is particularly effective for monitoring, forecasting, and managing MWL in real-time because it can extract complex correlations from huge amounts of information. Deep learning models can be used to identify people's MWL, making systems more flexible and effective. The feedforward neural network (FNN), the recurrent neural network (RNN), and the convolutional neural network (CNN) are the three main neural network architectures for deep learning that are introduced in this section. Each of these architectures has specific strengths and drawbacks and is designed for various types of data and tasks. A thorough grasp of their distinctions is essential for choosing the best approach for a given problem and creating effective deep learning models.

Recent studies have investigated the possibility of predicting MWL via brain activity captured using EEG. The main advantage of such a technique is that it is unobtrusive, allowing the MWL to be captured in real-time [21]. In line with recent advances, sophisticated deep learning models have been designed to accurately capture variance characteristics within EEG signals, allowing for precise classification of an individual's MWL levels [36, 109, 111, 116, 123]. Despite the promising potential of deep learning in classifying MWL levels from EEG signals, its application is not without several inherent limitations. Small sample sizes usually hinder the application of deep learning to EEG signals, the absence of standardised protocols for data preprocessing, the lack of diversity in study populations, and difficulties with feature extraction and model training. Removing these constraints will enhance the accuracy, consistency, and applicability of deep learning methods in EEG classification.

2.2.1 Feedforward Neural Networks (FNN)

In 1958, Frank Rosenblatt proposed the perception concept, which was initially a single-layer network used for binary classification tasks, and it is one of the earliest and most significant contributions to the development of Feedforward Neural Networks (FNNs) [175]. The FNN is the foundation of deep learning, and it is characterised by its sequential layer structure,. The connections between the nodes are acyclic, comprising an input layer, a hidden layer(s), and an output layer that connects forward to neurons in the subsequent layer. This allows the information to flow in one direction only, and that is why it is called feedforward. Moreover, there are no back-loops or connections between neurons within the same layer.

The general workflow in an FNN involves the following steps:

- **Input Layer:** This layer receives the raw input data. Each neuron in this layer represents one feature of the input data.
- **Hidden Layers:** These layers perform most computational heavy lifting. Neurons apply a weighted sum on their inputs, followed by a nonlinear activation function to introduce non-linear properties into the network. Common activation functions include ReLU (Rectified Linear Unit), Sigmoid, and Tanh.
- **Output Layer:** The final layer outputs the prediction of the network. The function of this layer varies depending on the nature of the task (e.g., classification, regression).

The operations within the network can be mathematically described by the following equations:

$$\mathbf{h} = \sigma(\mathbf{W}_{ih}\mathbf{x} + \mathbf{b}_h) \tag{2.1}$$

$$\mathbf{y} = \mathbf{W}_{ho}\mathbf{h} + \mathbf{b}_y \tag{2.2}$$

where $\mathbf{h}$ represents the hidden layer activations. $\mathbf{x}$ represents the input vector. $\mathbf{y}$ represents the output vector. $\mathbf{W}_{ih}$ and $\mathbf{W}_{ho}$ are the weight matrices for the input-to-hidden and hidden-to-output connections, respectively. $\mathbf{b}_h$ and $\mathbf{b}_y$ are the hidden layer's and output layers' bias vectors, respectively. σ is the activation function used in the hidden layer, such as the sigmoid, ReLU, or tanh function.

In equation 2.1, the hidden layer activations ($\mathbf{h}$) are computed based on the input vector ($\mathbf{x}$), using the weight matrix $\mathbf{W}_{ih}$ and the bias vector $\mathbf{b}_h$, passed through the activation function σ. In equation 2.2, the output vector ($\mathbf{y}$) is computed by transforming the hidden layer activations through the weight matrix $\mathbf{W}_{ho}$ and adding the output bias vector $\mathbf{b}_y$.

2.2.2 Recurrent Neural Network (RNN)

A recurrent neural network (RNN) [59,81] is a type of neural network architecture that can process sequential data using internal memory. Unlike FNN, which processes data unidirectionally, RNN can take past information into account and use it to inform the current output. The architecture of RNN is characterised by feedback connections that allow the network to process sequences of inputs with internal memory. An input vector is fed into the RNN at each time step, producing an output value. The RNN considers previous inputs by maintaining a "hidden state" that encodes relevant information from previous time steps. This allows the RNN to capture long-term dependencies in the input data and produce more accurate predictions. With every time step, the hidden state of the RNN experiences an update, incorporating data from both the current input and the previous hidden state. This feedback mechanism enables the RNN to retain pertinent historical information, subsequently impacting future predictions.

This neural network is widely recognised as one of the most popular models for analysing sequential data. In an RNN, consider ($\mathbf{X}_t$ is a sequence of input vectors ($\mathbf{X}_t, t = 1, 2, 3, \ldots, t_n$). The input $\mathbf{x}_t$ are fed one at a time into the RNN then an output value $\mathbf{h}_t$ is given. The future information at the next time t continuously flows into the model. The equations for a basic RNN are as follows:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h) \tag{2.3}$$

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y \tag{2.4}$$

where $\mathbf{h}_t$ represents the hidden state at time step t. $\mathbf{x}_t$ represents the input at time step t. $\mathbf{y}_t$ represents the output at time step t. $\mathbf{W}_{hh}$ and $\mathbf{W}_{xh}$ are the weight matrices for the hidden-to-hidden and input-to-hidden connections, respectively. $\mathbf{b}_h$ is the bias vector for the hidden layer. $\mathbf{W}_{hy}$ is the weight matrix for the hidden-to-output connections. $\mathbf{b}_y$ is the bias vector for the output layer. tanh is the hyperbolic tangent function used to activate the hidden state. In equation (2.3), the hidden state $\mathbf{h}_t$ is updated based on the previous hidden state $\mathbf{h}_{t-1}$, the current input $\mathbf{x}_t$, and the corresponding weights and biases. In equation (2.4), the output $\mathbf{y}_t$ is computed by multiplying the hidden state $\mathbf{h}_t$ with the weights $\mathbf{W}_{hy}$ and adding the bias term $\mathbf{b}_y$.

In theory, the model with the most data should be able to correctly identify correlations between events; in practice, training the RNN model can be challenging. This is because RNN can suffer from the vanishing and/or exploding gradient problem [16], in which information is rapidly lost over time, making it difficult for the model to recall distant information [186]. The vanishing gradient problem occurs when the gradient, which is used to update the model parameters during training, becomes so small that the model's performance does not improve. Conversely, the exploding gradient problem occurs when the gradient becomes too large, causing the model's performance to become unstable. Various modifications to the basic RNN architecture have been proposed to address these issues, including using long short-term memory (LSTM) and gated recurrent units (GRU).

2.2.3 Long Short-Term Memory (LSTM)

Long short-term memory (LSTM), which is an extension of the basic RNN architecture, was proposed by Hochreiter and Schmidhuber [81] to address the vanishing gradient problem. The LSTM unit comprises a memory cell (c_t) and three different gates that control the flow of information through the memory cell. By incorporating memory cells and different types of gates, LSTM can learn long-term dependencies and recognise patterns in sequential data. The first gate is a forget gate (f_t), which decides what information in the memory cell should be discarded. The input gate (i_t) controls the incoming inputs that might not be relevant or may be errors that could interfere with the current memory content in the memory cell. The output gate (o_t) protects other units from currently irrelevant content in the memory cell and controls the error flow from the current state to the next state.

The memory cell and all these different gates are updated over time. This allows the LSTM to selectively remember or forget information from previous time steps and incorporate new information as it becomes available. The following equations commonly describe the LSTM:

$$\mathbf{f}_t = \sigma(\mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{W}_{fx}\mathbf{x}_t + \mathbf{b}_f), \tag{2.5}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{W}_{ix}\mathbf{x}_t + \mathbf{b}_i), \tag{2.6}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{W}_{ox}\mathbf{x}_t + \mathbf{b}_o), \tag{2.7}$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{ch}\mathbf{h}_{t-1} + \mathbf{W}_{cx}\mathbf{x}_t + \mathbf{b}_c), \tag{2.8}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{2.9}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{2.10}$$

Where σ is the sigmoid function, $\odot$ represents element-wise multiplication. The term $\mathbf{x}_t$ is the input at time t. $\mathbf{h}$ is a vector of hidden value. $\mathbf{W}_{fh}, \mathbf{W}_{fx}, \mathbf{W}_{ih}$, $\mathbf{W}_{ix}, \mathbf{W}_{ho}, \mathbf{W}_{ox}, \mathbf{W}_{ch}$, and $\mathbf{W}_{cx}$ are weight matrices for different values used to calculate in variant gates. For example, $\mathbf{W}_{fh}$ refers to a weight metric for hidden value in forget gate. The terms $\mathbf{b}$ are bias vectors. $\mathbf{c}_t$ is a memory cell and $\tilde{\mathbf{c}_t}$ is a candidate updated in the memory cell. Despite the advances the LSTM represents in mitigating the vanishing and exploding gradient problems, studies such as the one conducted by DiPietro et al. [56] suggest that these challenges have not been fully overcome. Indeed, their experiments demonstrated that LSTMs can still encounter difficulties when processing extremely long sequences, indicating that the search for improved architectures must continue.

2.2.4 Gated Recurrent Unite (GRU)

Addressing the longstanding challenge of managing long-term dependencies effectively in the realm of deep learning, the gated recurrent unit (GRU) has emerged as an innovative solution. The GRU, a streamlined version of the traditional RNN, shares the goals of the LSTM model yet differentiates itself through its minimalist design. This design eliminates memory cells $\mathbf{c}_t$ and merges the forget and input gates into a singular update gate, resulting in a structure with fewer components but equal potency. The intent behind this simplified architecture is to address the two primary obstacles that have continually hindered the effectiveness of the LSTM model: the vanishing gradient problem and the complex task of learning long-term dependencies [41].

The GRU unit is composed of $\mathbf{u}_t$ and $\mathbf{r}_t$, which are the update gate and reset gate, respectively. The GRU is commonly described as follows:

$$\mathbf{u}_t = \sigma(\mathbf{W}_{uh}\mathbf{h}_{t-1} + \mathbf{W}_{ux}\mathbf{x}_t + \mathbf{b}_u), \tag{2.11}$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_{rh}\mathbf{h}_{t-1} + \mathbf{W}_{rx}\mathbf{x}_t + \mathbf{b}_r), \tag{2.12}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{hh}(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{W}_{hx}\mathbf{x}_t + \mathbf{b}_h), \tag{2.13}$$

$$\mathbf{h}_t = \mathbf{u}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{u}_t) \odot \tilde{\mathbf{h}}_t. \tag{2.14}$$

Where $\mathbf{u}_t$ and $\mathbf{r}_t$ are the reset gate and update gate, respectively. The definition of $\sigma, \mathbf{x}_t$, $\mathbf{h}$ and $\mathbf{W}$ term can be found in the previous section of the LSTM model.

The equations show how the GRU computes its hidden state $\mathbf{h}_t$ at each time step t, based on the input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. The reset gate controls how much past information to forget, while the update gate determines how much of the new information to incorporate. The candidate activation function $\tilde{\mathbf{h}}_t$ computes the new proposed memory content. Finally, the hidden state $\mathbf{h}_t$ is computed based on the reset gate, the current hidden state, and the candidate activation function.

2.2.5 Bidirectional Model

The bidirectional processing technique is a powerful tool for modelling sequential data with long-term dependencies and can understand the context from both directions. This technique can be applied to various types of RNN architectures. Bidirectional RNN (BiRNN), for example, was proposed by Schuster and Paliwal [186] to address the limited model power problem when dealing with long input sequences. The idea of BiRNN is to train two RNN models, one in the forward direction and one in the backward direction, and then merge the hidden states from each network to generate predictions. This approach enables the network to consider both past and future contexts when processing the input sequence, thereby enabling it to capture long-term dependencies more effectively.

Similarly, bidirectional long short-term memory (BiLSTM) and bidirectional gated recurrent unit (BiGRU) combine bidirectional processing with advanced gating mechanisms to capture long-term dependencies and generate accurate predictions. These models can effectively represent the intricate relationships between various events by processing the sequence in both forward and backward directions and using memory cells and gating mechanisms to retain or discard information selectively [43, 67]. Despite their strengths, bidirectional models have limitations; they need full sequences for predictions, making them unsuitable for real-time use. They also require more computational resources than unidirectional models due to dual-direction data processing [104].

2.2.6 Convolutional Neural Network (CNN or CovNet)

Utilising the strategy of iterative application of convolutional filters to input data, CNN facilitates the learning of progressively intricate features. These features prove instrumental for various tasks, notably image classification and object detection [15]. Despite their initial conception for image processing tasks, the flexible design of CNN has allowed for their application in time series analysis as well. This is accomplished by incorporating time as an additional dimension, thereby expanding their functional scope [121]. This model typically consists of three layers: convolution, pooling, and fully connected. In the convolution layer, features are extracted via filters made up of small kernels. The dimensionality of the features is then reduced in the pooling layers, and the fully connected layers combine data from the final feature maps to provide the final classification. In the field of time series analysis, pooling layers are utilised to downsample features and reduce the dimensionality of the data. This strategy closely resembles the application of pooling in the context of image processing [62, 122].

One challenge in using CNN for time series analysis is that the network must be able to handle variable-length sequences [106]. One approach is to use a sliding window approach, where the time series data is divided into fixed-length segments and fed into the CNN [220]. Another approach is to use dilated convolutions, which can operate on inputs of variable length and allow the network to learn features at different scales [238].

In general, CNN differs from FNN and RNN in their optimisation for spatial feature learning and their use of convolutional layers; however, they all share the same objective of identifying complex patterns and interdependencies among various categories of data. By considering time as an additional dimension, CNN can be adapted for time series analysis and serve as a potent instrument for analysing high-dimensional data [42].

2.3 Mental Workload (MWL) Measurment

The terms “mental workload” and “cognitive load” are often used interchangeably in psychological and neuroscientific research [64]. However, they have different definitions and contexts, so distinguishing between them is important. Cognitive load theory focuses on how information and learning tasks can be optimised to make the best use of working memory, and it is divided into intrinsic, extraneous, and germane cognitive load. Intrinsic load refers to the complexity inherent to the material or task, regardless of the learner’s other activities. Extraneous load relates to how information or tasks are presented to the learner and can impede or enhance learning. Germane load represents the cognitive effort to create lasting knowledge structures [185]. On the other hand, MWL is a broad term that encompasses the demand placed on a person’s cognitive system, including memory, attention, and executive functions, by a particular task or set of tasks. It often considers a task’s total demand, including cognitive and emotional aspects [211]. While both concepts are related to performance, cognitive load theory often aims to optimise learning efficiency, whereas MWL assessments aim to maintain performance levels and safety in operational settings.

MWL is widely understood, yet it can be challenging to articulate [191]. Nonetheless, it is a crucial concept in understanding the cognitive demands placed on individuals during task performance. MWL is also closely associated with stress and strain, reflecting two aspects of our interaction with challenging tasks [153]. Stress refers to the external challenges that drain our mental resources, such as the complexity of the task, time pressure, environmental conditions, and the need to juggle multiple tasks [134, 207]. Strain, on the other hand, represents how we process, manage, and adapt to the stressors of the task, which is demonstrated through the use of cognitive skills such as memory and planning, as well as our accumulated experience [163, 215]. Therefore, achieving an optimal balance between demands and cognitive resources is crucial when it comes to managing mental workloads effectively. This is because mental workload has a significant impact on cognitive strain, which in turn can greatly influence an individual's productivity and overall performance.

MWL is evident in various areas of life, impacting everything from children's attention spans [125] to the design of educational programmes [168, 256], from driving fatigue [88, 241] to performance across a broad spectrum of fields [230]. Mastery of MWL is pivotal for maximising human capability and warding off cognitive overload or insufficient stimulation. Its effective prediction and management are crucial for optimising human performance and preventing cognitive overload or under-stimulation.

While it is commonly assumed that there is a simple linear relationship between workload and performance, research suggests that the relationship is curvilinear [82]. People may try harder and use different approaches when faced with challenges, which can lead to improved performance despite an increased workload. However, excessive workload can lead to decreased performance due to being distracted, having limited mental resources, and juggling too many tasks. On the other hand, a low workload can result in not paying attention, being less alert, and even falling asleep, which can also negatively impact performance [205]. Therefore, it is critical to find the right amount of work that helps people perform at their best without causing problems [69].

Research studies, such as Young et al. [237], have shown that excessive workloads often lead to decreased performance and increased errors. This is consistent with Kahneman's resource model [102], which suggests that our cognitive resources are limited. The graph in Figure 2.5 illustrates the relationship between MWL and performance.

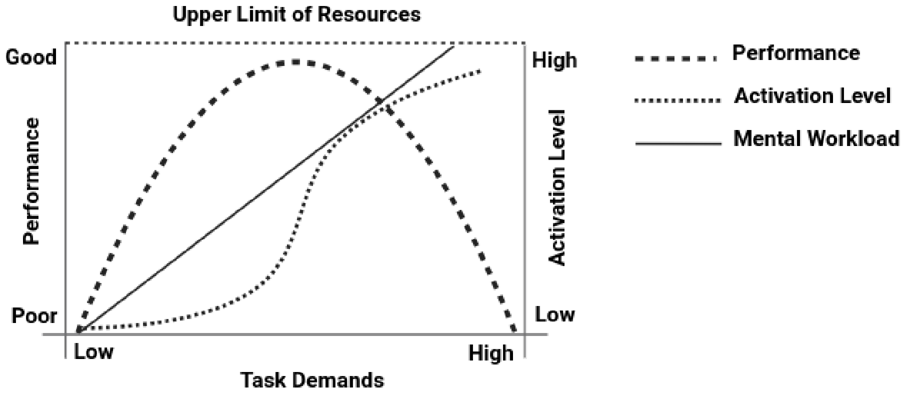


Figure 2.5: The relationship between activation level, workload (task demands) and performance (adapted from de Waard 1996 [48]).

The x-axis represents MWL, while the y-axis represents performance. Performance improves as MWL increases, but only up to a certain point, after which it begins to decline, forming an inverted U-shaped curve. The optimal MWL varies depending on the complexity of the task at hand: simpler tasks may require a lower MWL, while more complex tasks may demand a higher one.

In this study, MWL is defined as the measurable amount of cognitive demand that an individual's cognitive and emotional capacities experience while performing multiple tasks simultaneously. The assessment of MWL is important to anticipate and prevent mental overload or performance degradation in complex working environments and real-life situations. Chapter 3 will describe in detail a deep learning model that was developed to provide a reliable tool for assessing MWL across various operational settings and tasks.

2.3.1 Behaviour Measurement

The ability to measure an individual's MWL is crucial, especially in safety-critical scenarios like driving. Typically, this is done by assessing task performance, which is essential for evaluating the effectiveness and efficiency of an individual's abilities. Direct task performance measures can help determine an individual's MWL by evaluating how well they perform the primary task.

For example, in a driving scenario, errors in steering or inconsistencies in following distance can indicate a higher mental workload. Additionally, monitoring attention and workload from a primary task can be done by assessing performance on a secondary task, such as responding to peripheral visual signals, while performing the primary task. As mental workload increases on the primary task, performance on the secondary task declines [237]. One effective tool for assessing MWL in driving is the peripheral detection task (PDT), which measures response times and missed signals to visual cues. The PDT [212] is a secondary task measure of mental workload and visual distraction. With the PDT, drivers must respond to random targets presented in their peripheral view. It specifically assesses an individual's ability to detect and respond to stimuli presented in peripheral vision while engaged in a primary task, such as driving. During the primary task, if an individual's mental workload is high, their ability to process peripheral information decreases. So, if the participant's mental workload is high during the primary task (i.e., driving), their response times to the LED light will increase, and they may miss more signals. This change in PDT performance is used to infer the level of mental workload the participant is experiencing [237]. In a study examining the impact of mobile phone conversations (hands-free and handheld) on driving performance in various traffic environments, it was found that the complex urban environment presented the most demanding mental workload, even without phone use, as indicated by significantly poorer performance [207].

Another widely used method for measuring participants' mental workload is cognitive tasks such as the n-back task [113]. The n-back task is a commonly used tool for mental workload assessment and involves presenting participants with a sequence of stimuli such as letters, numbers, spatial positions, or sounds. Participants must identify whether the current stimulus matches the one presented "n" steps earlier in the sequence. The "n" factor can vary. Increasing numbers indicate a more demanding task, with common iterations including 1-back, 2-back, and 3-back. Performance in the n-back task is assessed based on two factors: the accuracy of the responses, the percentage of correct recognition of both targets and non-targets and the reaction times for correct responses. An interesting pattern emerges as the task's difficulty escalates: accuracy typically decreases and response times lengthen, indicating an increased mental workload [8]. Performance metrics are essential for measuring the effectiveness of a system or task. Common performance metrics include response time, completion time, efficiency, engagement, accuracy, and error rate [132].

2.3.2 Self-Report Measurement

A self-report questionnaire is another method for measuring MWL, unlike an objective measure, which infers workload from task outcome. NASA-TLX [70, 71] is a widely used questionnaire that helps to evaluate the workload of participants after performing a task. The questionnaire measures six different subscales of workload, including mental demands, physical demands, temporal demands, performance, effort, and frustration. Each subscale is rated on a 100-point scale with 5-point increments. The raw score obtained from the first part is then subjected to a weighting process via a pairwise comparison of subscales, where participants choose the subscale they perceive to be more relevant to their workload. The frequency of subscale selection serves as a weight for that subscale, which is multiplied by the participant's rating on each respective subscale to compute a weighted score for that subscale. The weighted scores are subsequently aggregated and divided by 15 (the number of paired comparisons) to derive an overall TLX score that reflects the participant's workload. SWAT [173] is a simpler alternative to NASA-TLX. It assesses participants on three subscales: time load, mental effort load, and psychological stress load. Participants choose from three levels - low, medium, and high - for each subscale. Another tool available for subjective mental workload assessment is the Workload Profile (WP) questionnaire [208]. This tool evaluates mental workload by asking individuals to assess the demand placed on them across eight distinct dimensions. These dimensions include perceptual/central processing, response selection and execution, spatial processing, verbal processing, visual processing, auditory processing, physical efforts related to manual tasks, and speech production. By gathering ratings on these dimensions, the WP questionnaire offers a comprehensive profile of the workload, highlighting how it is distributed across various cognitive and physical resources, providing a more nuanced understanding of workload beyond the overall intensity of demand.

While self-reporting can provide useful qualitative feedback on a participant's experience, it can be subjective and influenced by factors such as the participant's mood, willingness to provide honest responses, and ability to self-assess. Moreover, it may further increase participant's MWL, particularly in studies that require participants to rate their MWL level after completing a task and then immediately perform another task.

2.3.3 Physiological Measurement

Various physiological measurements are commonly used to assess MWL. For example, electrocardiac and cardiovascular activity can be measured by heart rate (HR), heart rate variability (HRV), and blood pressure (BP). However, the effectiveness of these measures can vary depending on the nature of the task being performed.

In a recent study, Mach et al. [135] found that HR can be a suitable indicator of MWL under certain conditions. During the study, participants performed a range of tasks with varying levels of mental effort while their HR was monitored. The researchers observed that HR increased with MWL when participants were sitting but not walking. This could be explained by the fact that physical exertion from walking can raise HR even in the absence of mental exertion. Thus, while HR is a reliable indicator of MWL when participants are stationary, its validity diminishes when they are mobile.

HRV is another important measure of the heart's rhythm, and recent research has shown that it changes during periods of stress. Specifically, the part of HRV linked to relaxation tends to decrease, while the ratio indicating stress increases. Interestingly, while blood pressure also increases during stressful tasks, it does not fully return to baseline even after a break, particularly the diastolic pressure (the lower number in a blood pressure reading). These findings suggest that HRV may be a more sensitive and accurate indicator of mental stress than blood pressure, which can be influenced by physical factors such as muscle activity. This highlights the importance of considering HRV as a potential biomarker for stress in both clinical and research settings [79]. Although some studies have shown increased blood pressure with harder tasks, others have reported mixed results [38]. Blood pressure has limitations in measuring MWL because it does not consistently rise with the complexity of tasks. Therefore, other measures such as HR and HRV may be more suitable for assessing MWL than blood pressure.

Another measure which has been adopted is respiratory measures such as respiration rate, which indicates the number of breaths per unit of time. The respiratory pattern is expected to change with an increase in MWL, resulting in slower and deeper breathing [214]. In a recent study, raw photoplethysmogram (PPG) data was collected to reconstruct respiratory signals while participants performed tasks. Using the respiratory pattern, the study effectively classified the MWL level [214].

Eye-tracking measures are also well-established for assessing MWL. These measures are based on eye activities such as blink rate, blink closure rate, gaze angle, pupil size, diameter, and pupillary responses. In a recent study [226], pupil diameter and gaze entropy were used to distinguish differences in workload between task difficulty levels. The study found that both metrics increased as task difficulty levels increased. However, it should be noted that this method has a key drawback, in that it is unresponsive after overload occurs and is highly sensitive to changes in environmental illumination [31].

While physiological signals can be used to assess a subject's stress [38], they have some limitations. Therefore, researchers have used more neurophysiological measures to access the subject's MWL.

2.3.4 Neurophysiological Measurement

Brain signal activity has been evaluated using various neuroimaging techniques such as MEG [200], functional magnetic fMRI [126], fNIRS [75], and EEG [26]. While MEG can show brain activity with a high temporal resolution, fMRI has the strength that it can measure brain activity mapping with high spatial resolution. However, they are not suitable for all environments. Both signals require a lot of professional lab settings and specialised equipment and are cumbersome, which is not practical in real-life environments [130]. While fNIRS is portable, can show hemodynamic responses associated with neural activity, and measure brain activity in different brain regions, it has low spatial resolution. The portable EEG device can measure brain activity with a high temporal resolution, making it ideal for detecting subjects' MWL levels in real-time. EEG is frequently preferred among these neurophysiological signals in human-computer interaction contexts due to its non-invasive nature and high temporal resolution, allowing for millisecond-scale measurements [117]. Its popularity is further enhanced by its strong correlation with a person's real-time MWL status [204].

2.4 MWL Classification

Classifying MWL levels using physiological or neurophysiological measures requires precise labels for each response category. This can be done through two primary methods. Firstly, the self-report measures, as described in **Section 2.3.2**, involve participants providing their subjective assessments of their MWL levels using a questionnaire. This approach provides valuable insight into participants' own perceptions of their MWL levels. Then, participants' physiological data can be classified into discrete levels of MWL — low, medium, or high — and any changes or patterns in the data can be observed. Mapping this objective measure against self-reported data helps us better understand the correlation between personal experience and physiological and neurophysiological markers of workload.

Task design offers an alternative yet equally systematic approach. In this method, researchers meticulously craft tasks expected to elicit varying levels of MWL. These tasks are typically employed during calibration to establish the baseline or reference point for low, medium, and high MWL levels. For example, a straightforward task is used to establish a baseline (low workload), a more intricate task for a medium workload, and the most challenging task for a high workload. The ensuing physiological and neurophysiological responses induced by these tasks help us to construct a profile of what low, medium, and high mental workloads look like for each individual.

In practical applications of these concepts, the n-back task, a well-established cognitive challenge, is often adjusted to induce varying levels of MWL. In this study [89], researchers modified a standard n-back task to create different levels of cognitive demand. The 1-back version represented a low cognitive load, while the 3-back represented a high cognitive load. During the experiment, while participants performed the tasks, their photoplethysmogram (PPG) signals were recorded and analysed to reveal patterns in blood flow and respiration in relation to the imposed cognitive demands.

Building further on this empirical foundation, recent studies have demonstrated an inclination towards using multifaceted criteria to gain a more nuanced understanding of MWL. In recent work, researchers have employed more than one criterion to categorise subjects' mental workload. For instance, in one study [51], they utilised both task design (the 1-hour computerised letter recognition task) and questionnaires (the visual analogue scale of fatigue and the NASA-TLX) to categorise mental workload. The task design induced a mental workload of a certain intensity, while the subjective questionnaires allowed participants to self-report their perceived stress level or workload. The monitored physiological signal was the ECG from which heart rate variability (HRV) was derived, as well as blood pressure waveforms captured using the finger volume clamp method. Combining these methods provides a more comprehensive assessment, as the task design ensures that MWL is being imposed. Simultaneously, the questionnaires measure the subjective experience of the participants, which can vary individually.

By combining subjective and objective measures, researchers can create models that predict mental workload levels based on physiological or neurophysiological data. These models can be more accurate because they consider the individual variability in physiological or neurophysiological responses to mental workload. This can be useful for tailoring assessments to the individual and for training classification models.

The MWL level labels for our study were obtained using two methods: self-reported data from questionnaires for the first dataset and task design data for the second dataset. A more detailed description of the datasets used in this thesis is provided in

Section 3.1.

2.4.1 Limitation of Current Works

The field of deep learning, known for its swift growth and potential, has shown particular promise when applied to EEG studies, especially in the classification of MWL. Traditional shallow models continue to hold their ground in this diverse landscape of models.

Comparative Analysis of Traditional and Deep Learning Models

Despite the growing popularity of deep learning models, techniques such as linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbours (KNN), and random forest remain effective baselines in the literature [76, 180, 187]. These models offer the advantage of being generally easier to train and interpret than deep learning models. However, they might face challenges when dealing with EEG data's complex, non-linear relationships.

Complexity in Hybrid and Ensemble Models

The efficacy of CNN in extracting spatial features from EEG data has resulted in their increasing prominence in recent years [76, 111, 123, 129]. However, the limitations of CNN to capture temporal dynamics have caused researchers to investigate hybrid architectures that combine CNN with recurrent networks such as the LSTM network [116, 248]. These hybrid models have demonstrated promising results in addressing the temporal aspect of EEG data. Nevertheless, the complexity of these models may pose challenges during the training process and result in reduced interpretability.

Almogbel et al. (2019) [4] utilised raw EEG signals without preprocessing as input to their developed CNN model. The model was engineered to automatically extract key information and discern three gradations of a vehicle driver's cognitive workload and driving environment. The classification model proved adept at identifying the low MWL level but faced challenges when attempting to consistently discriminate between medium and high workload levels. This reveals room for improvement in the model's ability to differentiate between higher levels of MWL. Similarly, with an emphasis on MWL classification, Lee et al. (2020) [124] implemented a CNN-based model. The research team constructed a multiple-feature block-based CNN (MFB-CNN) that harnessed temporal-spatial EEG filters to illustrate the current mental states of pilots, thereby enabling accurate classification. With a similar classification-centric approach, Qayyum et al. (2018) [167] employed a pre-trained 2D-CNN to categorise human mental states during recurrent multimedia learning tasks. By transforming one-dimensional EEG signals into a two-dimensional format using the short-time Fourier transform technique, the researchers enabled the use of the 2D-CNN for classification. This methodology consistently tracked the behaviour of alpha brain waves across different cognitive tasks, thus successfully classifying each distinct mental state.

Issues with Specific Deep Learning Implementations

Stacked denoising autoencoders (SDAEs) have been introduced to reduce the dimensionality of EEG features while retaining the local information present in the data [32,230]. SDAEs provide an alternative method to address the within-subject classification issue, addressing certain limitations associated with CNN-based models. However, they bring their own challenges, including the computational costs and their sensitivity to hyperparameter tuning, which demand further exploration.

Ensemble models, which combine the strengths of multiple classifiers to boost performance, have also emerged. The ensemble CNN (ECNN) model proposed by Zhang et al. (2017) [244] is a testament to such an approach. Although ensemble models can potentially enhance performance, they could also introduce increased complexity and longer training times, which might be problematic in certain applications. Other deep learning models have also tackled within-subject MWL classification problems. These include RNN [116], BiLSTM [109,254], AConv-BiLSTM-NN [52], and BiLSTM-LSTM [36, 109, 111]. The primary focus of these models is on capturing the spatial and temporal features present in EEG data. However, these models may require large amounts of data for effective training and are also susceptible to overfitting. Further models used in this context include the Gated Recurrent Unit (GRU) [111], bidirectional gated recurrent unit (BiGRU) [111], and a combination of BiGRU with GRU (BiGRU-GRU) [111]. Long Short-Term Memory (LSTM) networks [36] have also been applied. These models, too, focus on capturing the dynamics in EEG data but may bring their own challenges, such as the need for comprehensive data and the risk of overfitting.

Researchers must weigh the trade-offs among accuracy, computational complexity, and interpretability when selecting an appropriate model for their problem. Future efforts should be devoted to developing more efficient and robust models that effectively capture spatial and temporal EEG data features. Additionally, exploring innovative approaches to address the distinct challenges of EEG-based MWL classification will remain a significant area of research.

Cross-Subject and Cross-Session Challenges

Effective transfer of EEG analysis models from one subject to another has proven to be a complex task [239]. In response to this challenge, Hefron et al. [73] developed a novel approach that entailed training a model on a specific subject and then applying this model to other subjects for classification. This model, termed a multi-path convolutional recurrent neural network (MPCRNN), was tested in a non-stimulus-locked multi-task environment to predict a subject's cognitive workload levels. Notably, the MPCRNN demonstrated increased classification accuracy and decreased variance across different participants, underscoring its potential effectiveness for addressing the cross-subject problem in EEG-based MWL classification. Meanwhile, Zheng et al. (2020) [253] proposed an extreme learning machine (ELM)-based ensemble, the ED-SDAE, to classify cross-subject cognitive workload levels, aiming to reduce subject-independent variation and discover time-variant EEG signal properties. Alternative methods were proposed by Jimenez et al. (2017) [95], who introduced a unique deep neural network architecture that merges the strengths of residual networks and GRU. This model effectively captured patterns across various regions and frequencies and interpreted changes over time. In another study, they proposed a custom domain adaptation (CDA) method designed to reduce both marginal and conditional distribution differences and personalise a classifier for each subject, resulting in higher accuracy compared to other deep unsupervised domain adaptation (D-UDA) methods.

Jimenez et al. (2020) [96] also addressed the issue of disparate EEG signal distributions among different subjects by proposing a custom domain adaptation (CDA) method integrating adaptive batch normalisation (AdaBN) and maximum mean discrepancy (MMD) into two separate deep neural networks. This method aimed to reduce both marginal and conditional distribution differences and personalise a classifier for each subject, achieving higher accuracy compared to other deep unsupervised domain adaptation (D-UDA) methods. Yin et al. (2017) [235] developed a switching deep belief network with an adaptive weights (SDBN) model for assessing the subject's operator functional states (OFS). The model architecture consisted of two sets of deep belief networks (DBNs): static and dynamic. The static DBNs aimed to eliminate higher-level representations of EEG features, while the dynamic DBNs were designed to capture novel EEG feature characteristics from unseen testing subjects. Zeng et al. (2019) [241] employed a gradient boosting-based classifier, LightFD, which was developed using the LightGBM framework. This model was particularly effective in identifying variations in drivers' mental states. The LightFD model, as proposed by the researchers, showcased robust transfer learning capabilities coupled with minimal time consumption. These characteristics render it especially suitable for real-time EEG mental state prediction, underscoring its potential utility in real-world applications. In a parallel effort, Shao et al. (2021) [189] employed a BiLSTM model for their investigation, demonstrating the application of recurrent neural networks in handling the complexities and temporal dynamics of EEG data for cross-subject MWL analysis.

Finally, Zeng et al. (2021) [239] utilised a domain-adversarial neural network (DANN), a model that has demonstrated superior performance in transfer learning, notably in the areas of document analysis and image recognition. However, it was not previously applied directly to EEG-based cross-subject fatigue detection. They proposed a novel model, a generative domain-adversarial neural network (GDANN), which integrated DANN with generative adversarial networks (GAN) for EEG-based cross-subject fatigue mental state prediction. The GDANN model aimed to address the problem of different EEG distributions across subjects. It attempted to balance disparities in the sample sizes between the source and target domains, selected the most appropriate Top N source domain subjects for experimentation, and endeavoured to extract as many invariant features of the target domain as possible. The model allowed transfer learning to be conducted across various domains and data tasks. Experimental results revealed that the performance of GDANN surpassed that of DANN, SVM, and EasyTL.

Given that the assessment of MWL is vital for individuals in both daily life and work situations, it is crucial to construct models capable of effectively managing cross-subject variations. Most studies in the current literature have primarily focused on single-session experiments, underscoring the need for additional research on cross-subject models for improved generalisability and applicability in diverse contexts.

Yin et al. (2017) [234] introduced an adaptive stacked denoising autoencoder (SDAE) model. This model was designed to train a static pattern classifier with EEG signals recorded on separate days for both training and testing. The aim was to adaptively update the weights of the shallow hidden neurons during the testing phase, thereby enabling more accurate classification across sessions.

Despite these initial efforts, current literature suggests that the estimation of cross-session cognitive workload levels using deep learning models has not been thoroughly explored. This area calls for further research to enhance the generalisability and applicability of EEG-based MWL classifiers across multiple sessions. Lim et al. (2018) [127] explored the feasibility of using the same features for different cognitive workload tasks by employing two independent datasets. Despite the promising premise, the results indicated that the average accuracy, although higher than chance levels, was too low for practical use. This underscores the need for more sophisticated methods to achieve better cross-task performance. To tackle this challenge, Shao et al. (2021) [189] proposed a concatenated structure of deep recurrent and 3D convolutional neural networks (R3DCNNs) to learn EEG features across different tasks. By converting the 1D EEG signal into a 3D representation, the R3DCNNs model could simultaneously capture EEG features from spatial, spectral, and temporal dimensions.

After evaluating the current literature, it is evident that the field employs a variety of deep-learning models. Each model is complicated and suited to specific problems. For the analysis in this study, the focus is placed on the RNN family of models. This decision is motivated by RNNs' inherent strengths in processing sequential data. This is especially relevant for the task at hand: the categorisation of MWL utilising deep learning techniques applied to EEG data. Moreover, because the nature of the problem and the features of the data are both time-dependent, RNNs are an appropriate choice for our investigation.

The next section will explore the cross-validation techniques that have been utilized in various research projects for MWL classification.

2.4.2 Cross-Validation in MWL Classification

As shown in the previous section, we aim to use machine learning techniques, specifically deep learning, to capture the variance characteristics of EEG signals and make a classification. Cross-validation is an essential technique for evaluating deep learning models and assessing their performance [183]. Different cross-validation techniques have been developed, each with its own algorithm. The experimental purpose, which may be subject-, task-, or session-dependent, determines which cross-validation approach is used.

A traditional CV technique splits an entire dataset into K equal-sized subsets or folds. The model is trained on $K-1$ folds, which are called the training set. One fold, which is kept apart and not seen by the model, is used as the test set. Every fold takes its turn as a test set [196]. The model training process is repeated K times, with a different fold preserved for model evaluation each time. The fundamental idea underlying a CV technique is that a collection of random variables is being drawn from a given probability distribution; these variables are statistically independent of each other, satisfying the independent and identically distributed (i.i.d.) property in probability theory and statistics [80]. However, the conventional method presents challenges when used with EEG signals, which change over time and represent time series data. Therefore, applying the traditional CV approach (i.e. shuffling and randomly splitting the data into K-folds) can violate the i.i.d. assumption [19]. A violation can make a model unreliable due to overfitting [35]; this issue becomes more serious in forecasting tasks in which future information should not be available to the model during training. In response to the unique characteristics of EEG data and the objectives of the research, different types of CV strategies were developed.

For example, in the study by Yang et al. [230], the data from all subjects were combined, shuffled, and then randomly divided into subsets to establish a subject-generic paradigm. On the other hand, the study conducted by Zeng et al. [240] adopted a task-generic paradigm, where data from different tasks were mixed before performing K-fold cross-validation on each subject. This approach allowed for combining data from different tasks and subjects, ensuring generality across both subjects and tasks in their study [240]. Figure 2.6 demonstrates the data splitting process into training and testing sets, as employed in K-fold cross-validation.

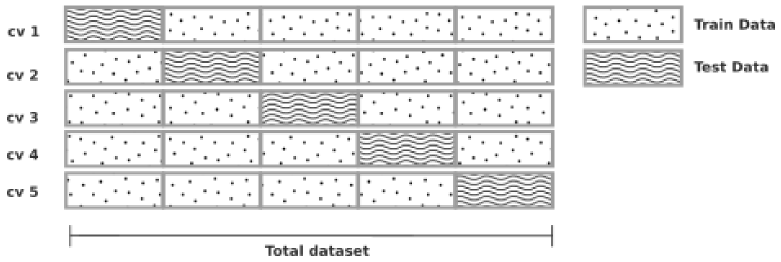


Figure 2.6: K-fold cross-validation technique

When looking more closely at validation methods for EEG, leave-subject-out cross-validation becomes a popular option, particularly for cross-subject classification model evaluation. Here, data from one subject is reserved for testing, while data from all other subject(s) are combined for training. This procedure is repeated until each subject is used at least once as a testing subject [73,94,95,116,239,253]. Figure 2.7 demonstrates the data splitting process into training and testing sets, as employed in leave-subject-out cross-validation. Different variations, including leave-session-out cross-validation, have also been developed to adapt this method to various contexts.

Diving deeper into these variations, leave-session-out cross-validation is applicable when training and testing. EEG signals are recorded in separate sessions or days; this strategy reserves one session for testing and uses the remaining sessions for training [4, 198].

In a parallel vein, another LOOCV offshoot—the leave-task-out cross-validation—shifts the focus, pivoting on tasks for data segregation. The leave-task-out cross-validation approach involves selecting training and test data from different tasks [189, 248]. The dataset is divided into two subsets: a training set and a test set, with separation carried out randomly or based on specific rules. As an alternative to these LOOCV variants, leave-p-out cross-validation has been proposed. The leave-p-out cross-validation, often abbreviated as LPOCV, is similar to LOOCV but with a twist. It reserves p samples/subjects for testing instead of just one. The remaining $n - p$ subjects are used for training [146]. Unlike K-fold and LOOCV, which have independent test sets in each iteration, some parts of the testing set might overlap in LPOCV, potentially causing the model to remember the training set pattern. This issue prompted the investigation of alternatives such as Monte Carlo cross-validation.

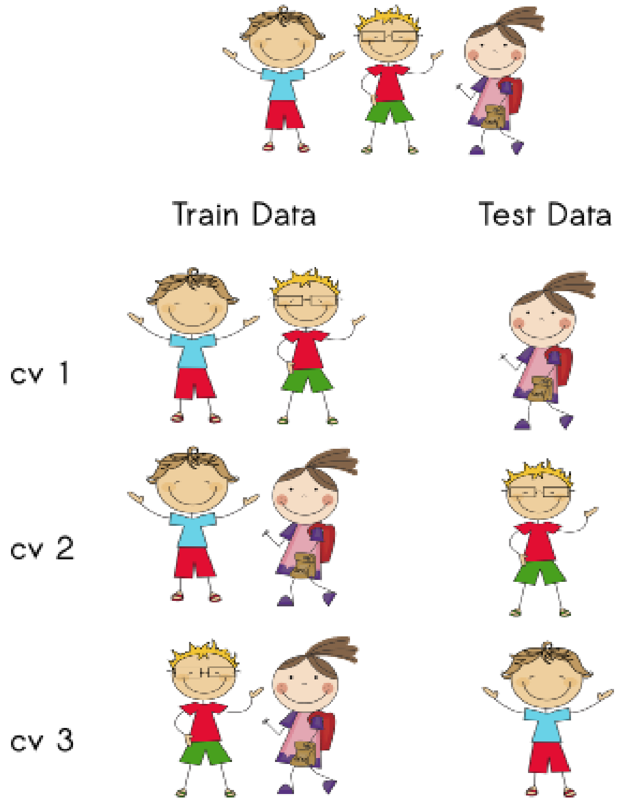


Figure 2.7: Leave-subject-out cross-validation technique

Monte Carlo cross-validation, also referred to as repeated K-fold cross-validation or repeated random sub-sampling cross-validation, is a variation of the K-fold method that aims to address some of its limitations. Saha et al. [178] adopted the Monte Carlo cross-validation technique with four folds, arguing that it offers higher optimisation than traditional K-fold and hold-out cross-validation methods. This study randomly divided each fold into training and testing datasets with a ratio of 60:40. The predictive accuracy obtained through this method was averaged across the splits to derive the final results.

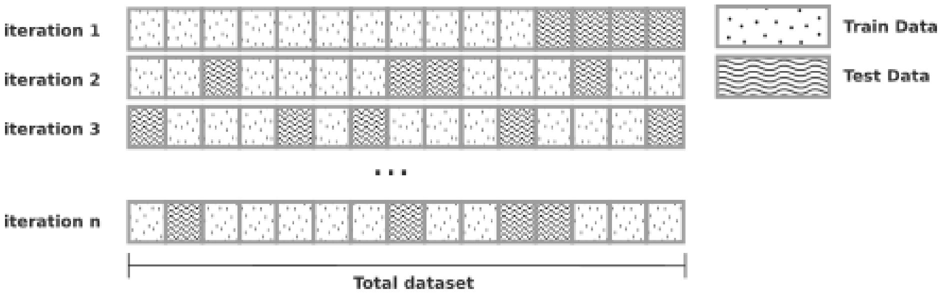


Figure 2.8: Monte Carlo cross-validation technique (adaptive from [178])

Although Monte Carlo cross-validation provides a more robust approach than traditional cross-validation techniques, it is essential to consider the data's specific characteristics. To be more precise, random shuffling does not adequately address the temporal nature of EEG signals before splitting the data into training and testing sets. Specifically, when the goal is predicting future events, such as a subject's MWL, disregarding this temporal characteristic could lead to unreliable classification model performance [35]. Figure 2.8 shows the Monte Carlo cross-validation technique, which randomly splits the dataset into training and test sets multiple times.

In addition to this, it is clear that thoroughly studying different cross-validation techniques is very important, especially when considering the type of data we have. Different approaches can be used to preserve the temporal structure while ensuring that the model evaluation is reliable when working with time series datasets such as EEG signals [110,170,171]. Thus, researchers must consider these options in alignment with their study's specific characteristics and goals.

Considering the temporal nature of EEG signals, time-wise cross-validation has been suggested as a suitable strategy to accommodate these characteristics [171]. This approach partitions the samples from each task and session into n evenly distributed, contiguous segments. The model is trained on $n - p$ segments from all tasks and validated on the remaining segments [170]. To minimise the impact of task transitions, some data from each task's initial and final segments may be excluded from the analysis. This approach provides a more tailored solution to the unique challenges of EEG signals. Figure 2.9 illustrates the time-wise cross-validation technique, where the dataset is split based on the temporal order of the data points.

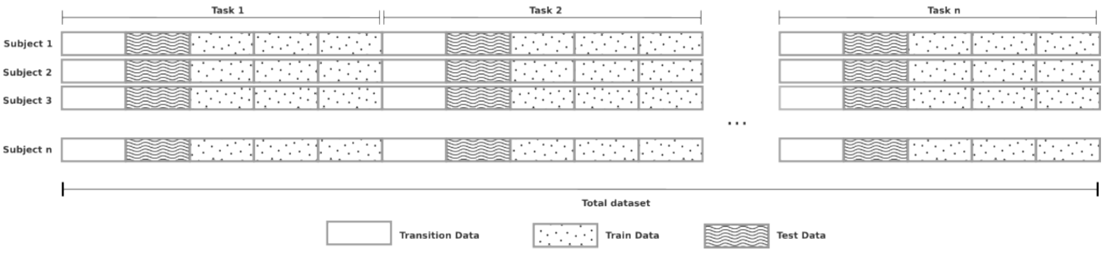


Figure 2.9: Time-wise cross-validation technique (adaptive from [170])

Time series cross-validation is another method that considers the temporal characteristics of time series data, such as EEG signals. This approach preserves the temporal structure of the data by reserving a final part of the series as the testing dataset. Importantly, the corresponding training set only includes observations that occurred before those in the test set [110]. By preserving the sequential arrangement of the data, time series cross-validation effectively precludes the leakage of information from future observations into the present prediction period, ultimately resulting in a more dependable and precise evaluation of the model's performance. In this way, time series cross-validation addresses the unique challenges of time series data and contributes to developing robust and generalisable models. Figure 2.10 illustrates the time series cross-validation technique, where the dataset is split based on the temporal order of the data points.

Due to the inherent temporal nature of EEG data, in our deep learning model evaluation step, It is crucial that we modify how we evaluate deep learning models in light of this. This highlights the significance of the time series cross-validation method for our MWL levels classification, which is discussed in **Chapter 6**. We will discuss the cross-validation technique in **Section 2.4.2**. At this point, we want to emphasise how crucial cross-validation is in the field of machine learning. It is crucial to evaluate the effectiveness and dependability of our deep learning model when predicting the intensities of MWL.

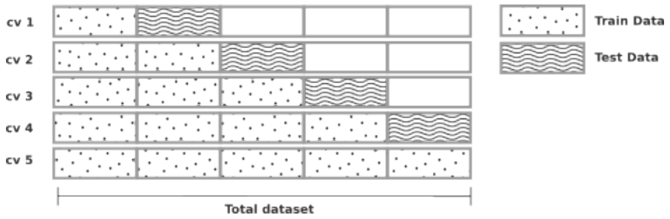


Figure 2.10: Time series cross-validation technique

2.5 Chapter Summary

This chapter provides a comprehensive review of the literature on EEG-based MWL classification. It sheds light on the intricacies and challenges of the field and identifies underexplored areas that this thesis will delve into in subsequent chapters. In addition to framing the classification problem, the chapter establishes a robust technical foundation by delving into essential components such as artefact removal, feature extraction, sequential model-based classifiers, CNN-based classification, and channel selection techniques. By examining these elements together, the chapter presents a holistic approach to the MWL assessment problem. Through careful analysis and synthesis, the chapter sets the stage for a nuanced understanding of EEG-based MWL classification, laying the groundwork for innovative contributions and insights.

Chapter 3

Methodology

The study utilised two distinct datasets to conduct experiments: the Simultaneous Task EEG Workload (STEW) dataset and the BCI Hackathon Grand Challenge dataset. The first dataset was the STEW dataset, which was used to answer research questions regarding EEG preprocessing and time series cross-validation. The methodology involved four key steps: data preprocessing, feature extraction, feature selection, and classification. The initial focus was on the intricate EEG preprocessing step, which involved only automated techniques. This step was crucial in capturing the unique characteristics of the EEG signal. Features were then calculated using a sliding window approach, extracting features from each window. To evaluate the models, a time series cross-validation technique was employed, incorporating both rolling and expanding window strategies. The classification was then performed on two tasks. The first task involved binary classification to categorise the EEG signal between resting and working states. In the second task, three MWL levels (low, moderate, and high) were classified from subjective ratings using objective EEG spectral data. Overall, the four preprocessing scenarios and time series cross-validation techniques were verified in deep learning models. The diagram provides a comprehensive overview of the process for EEG-based MWL classification using deep learning for the STEW dataset, as shown in Figure 3.1

The BCI hackathon dataset was utilised to perform the channel selection experiment, which contains more data from 62 EEG channels. This allowed for a more comprehensive experiment than the STEW dataset, which only has data from 14 EEG channels. The overall process was similar to the previous diagram, which included data preprocessing, feature extraction, feature selection, and classification. However, in this experiment, channel selection was performed after noise removal, and the cross-validation used was stratified sampling. The diagram provides a comprehensive overview of the EEG-based MWL classification process using deep learning for the BCI hackathon grand challenge, shown in Figure 3.2. The detailed methods will be described in the subsequent section, facilitating a clear understanding of the approach used in the study.

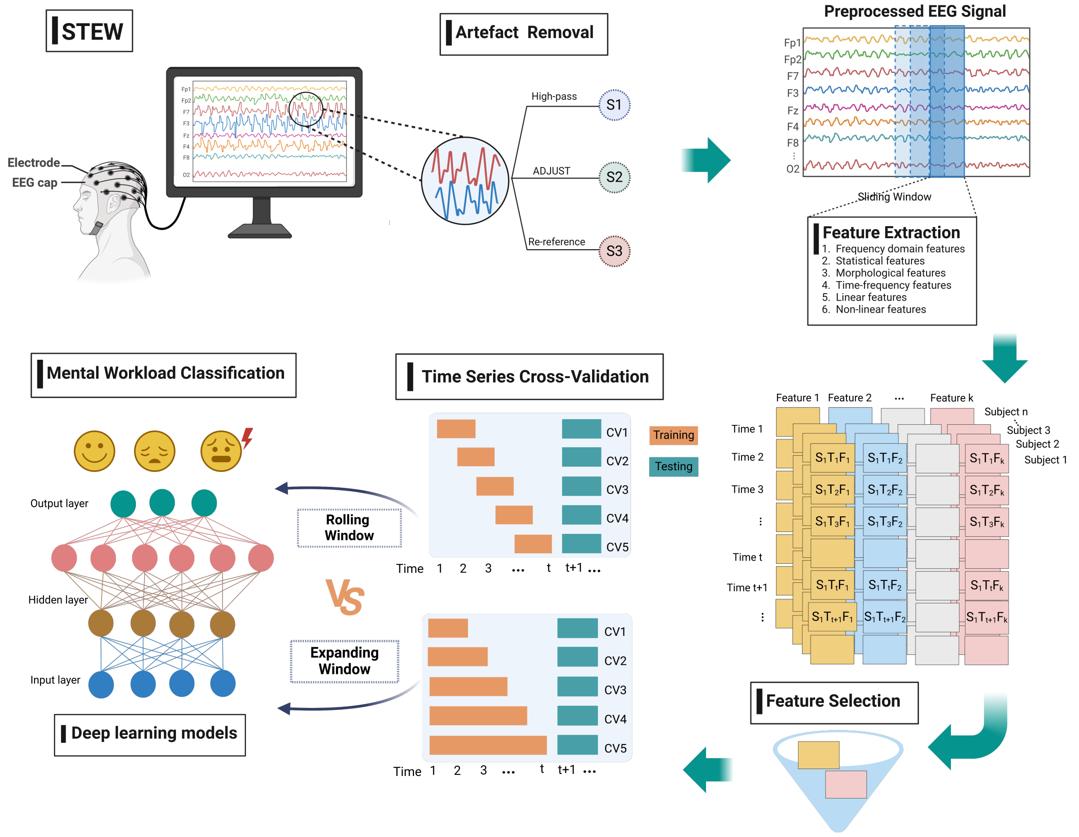


Figure 3.1: Methodological Overview for EEG Data Processing in the STEW Dataset

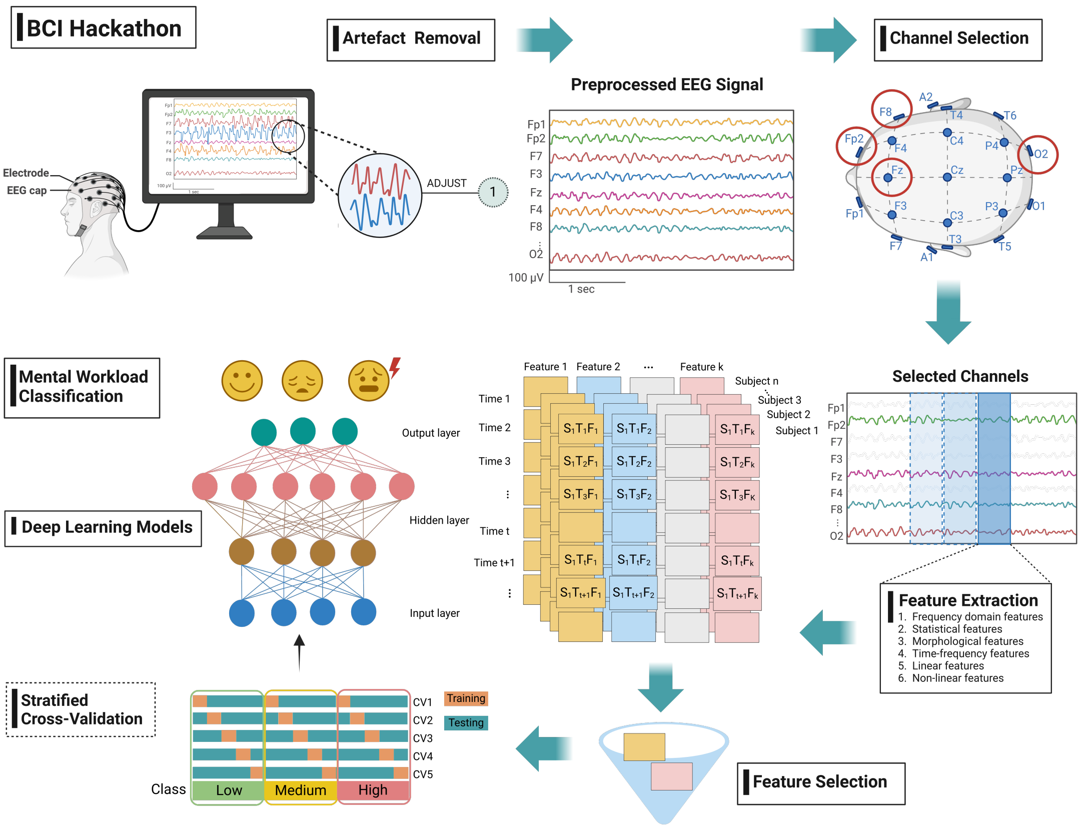


Figure 3.2: Methodological Overview for EEG Data Processing in the BCI Hackathon Dataset

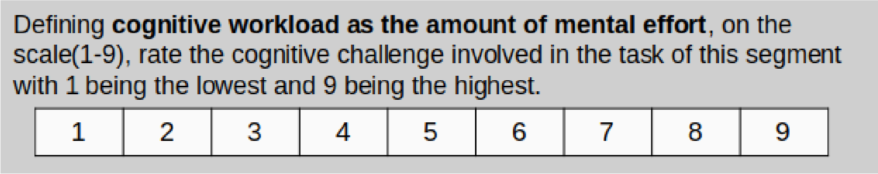
3.1 Data Source Description

In order to fully comprehend the analyses presented in this thesis, it is crucial to examine the data sources in detail and understand their underlying foundation. This section describes the open-access datasets used in this thesis.

3.1.1 Simultaneous Task EEG Workload (STEW)

The STEW dataset records multitasking MWL activity generated through a single-session simultaneous capacity (SIMKAP) experiment. Introduced by Lim et al. [128] and available as open access, this dataset is purpose-built to facilitate EEG studies and offers a framework for MWL level classification. It contains EEG signals from 48 male subjects who are university graduate students recruited via open email. They must not have neurological, psychiatric, or brain-related diseases and have never participated in an EEG experiment before. Data were recorded using the EMOTIV EPOC EEG headset, a wireless device equipped with 14 electrodes (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4). These electrodes were strategically positioned according to the 10–20 international system to ensure accurate signal capture. The sampling frequency was set at 128 Hz with a 16-bit A/D resolution. The data was transmitted to a paired PC desktop via wireless Bluetooth, and the raw data was recorded using Emotiv's 'TestBench' software.

The EEG signals were recorded in two states: resting and working. In the resting state, subjects sat in a chair for 3 minutes without performing any task. Their EEG was recorded during this time and used as the resting state data. In the working state, subjects performed the SIMKAP multitasking activity. The details of the SIMKAP task will be described in **Section 3.1.1**. Only the final 3 minutes of the EEG recording were used as the working state data. To reduce the effects of any between-task activity, the first and last 15 seconds of data were excluded from each recording, resulting in recordings of 2.5 minutes (150 seconds). The sample size of signals in each state is 19200 samples. Notably, the dataset does not provide any markers to indicate specific activities or events presented to the subjects. After each experiment state, participants were prompted to rate their MWL on a 9-point scale. This scale is analogous to the 1-21 scale used in the NASA-TLX [71], which is shown in Figure 3.3.



Defining **cognitive workload as the amount of mental effort**, on the scale(1-9), rate the cognitive challenge involved in the task of this segment with 1 being the lowest and 9 being the highest.

1	2	3	4	5	6	7	8	9

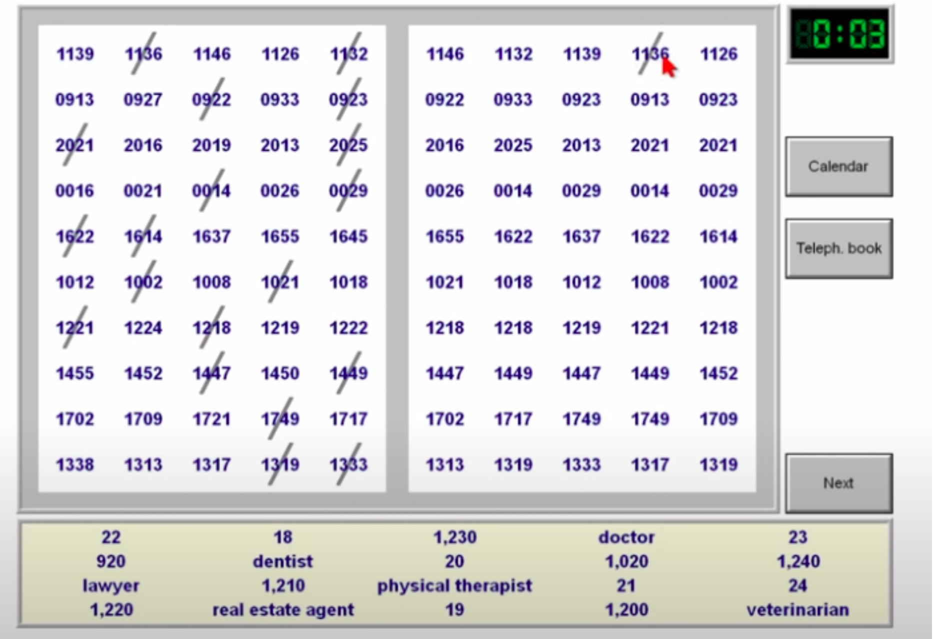
Figure 3.3: Questionnaire on a 1-9 scale for rating MWL, which subjects were required to fill out after completing each experiment segment (modified from Lim et al., 2018 [127])

For analysis purposes, the 9-point rating scale was categorised into three MWL levels: low (1-3), moderate (4-6), and high (7-9). Therefore, there is a single MWL rating for both the resting and working states.

Task: SIMKAP

The SIMKAP multitasking activity is primarily designed for job and career placement. It measures subjects' ability to concentrate and perform multiple tasks simultaneously, such as completing a routine task while answering questions. SIMKAP consists of three parts: routine tasks, problem-solving, and the SIMKAP test itself, which measures stress tolerance.

In the first part, participants must identify and mark a certain number, letters, and figures by comparing two separate panes shown in the upper part of Figure 3.4. In the problem-solving section, participants answer questions posed orally by selecting the answer from a multiple-choice selection displayed on the screen (shown in the bottom part of Figure 3.4). In the third section, participants combine all previous tasks with additional data look-up tasks, which require them to look up information in a simulated telephone directory or diary (see Figures 3.5 and 3.6 respectively).

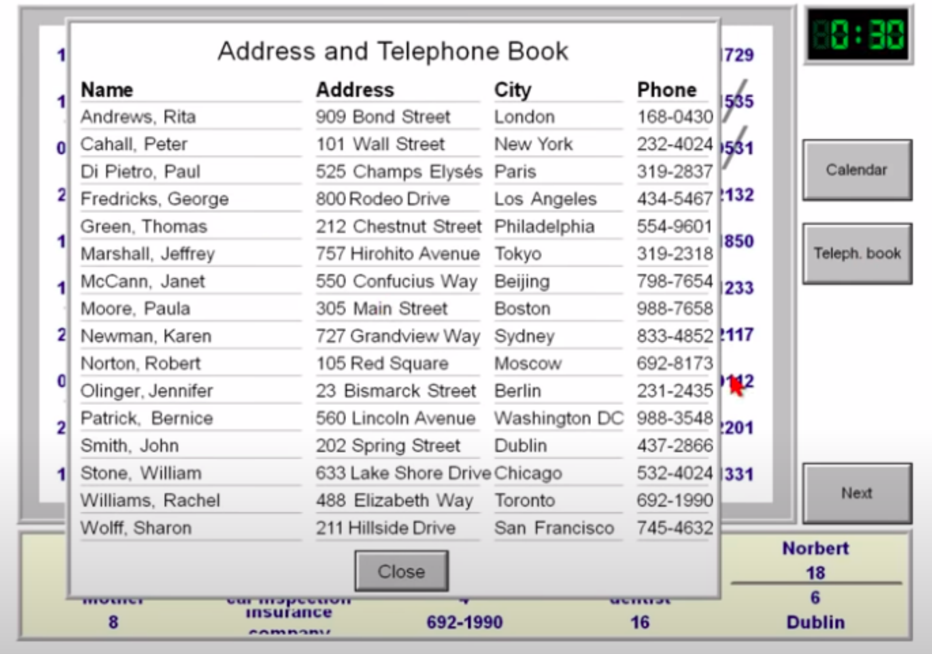


1139	~~1136~~	1146	1126	~~1132~~
0913	0927	~~0922~~	0933	~~0923~~
~~2021~~	2016	2019	2013	~~2025~~
0016	0021	~~0014~~	0026	~~0029~~
~~1622~~	~~1614~~	1637	1655	1645
1012	~~1002~~	1008	~~1021~~	1018
~~1221~~	1224	~~1218~~	1219	1222
1455	1452	~~1447~~	1450	~~1449~~
1702	1709	1721	~~1749~~	1717
1338	1313	1317	~~1319~~	~~1333~~

1146	1132	1139	~~1136~~	1126
0922	0933	0923	0913	0923
2016	2025	2013	2021	2021
0026	0014	0029	0014	0029
1655	1622	1637	1622	1614
1021	1018	1012	1008	1002
1218	1218	1219	1221	1218
1447	1449	1447	1449	1452
1702	1717	1749	1749	1709
1313	1319	1333	1317	1319

22	18	1,230	doctor	23
920	dentist	20	1,020	1,240
lawyer	1,210	physical therapist	21	24
1,220	real estate agent	19	1,200	veterinarian

Figure 3.4: Screenshot from the SIMKAP task. The top panel displays participants comparing numbers between two panes. The bottom panel presents a multiple-choice problem-solving task.

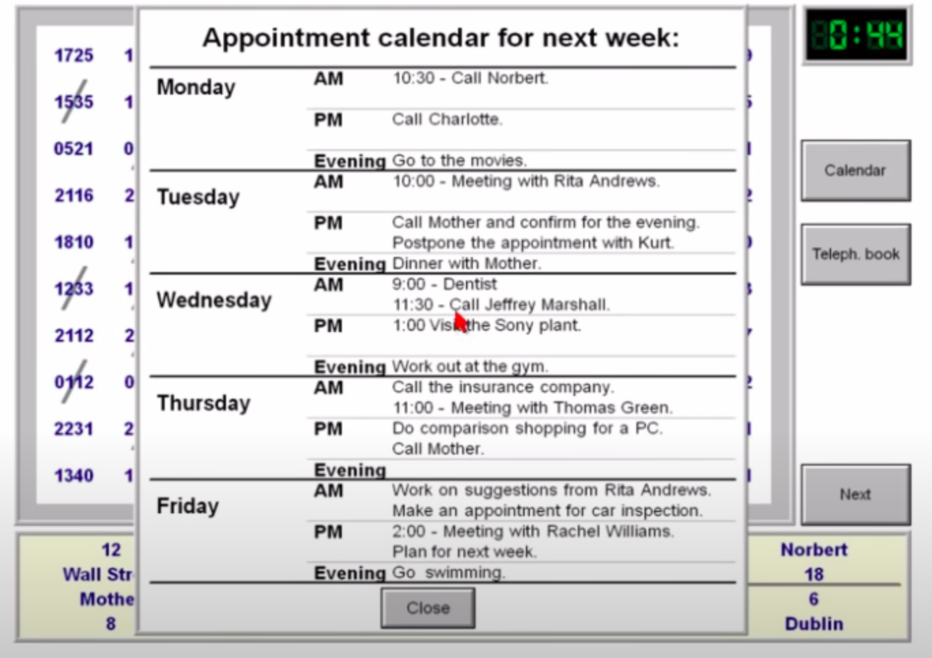


Address and Telephone Book

Name	Address	City	Phone
Andrews, Rita	909 Bond Street	London	168-0430
Cahall, Peter	101 Wall Street	New York	232-4024
Di Pietro, Paul	525 Champs Elysés	Paris	319-2837
Fredricks, George	800 Rodeo Drive	Los Angeles	434-5467
Green, Thomas	212 Chestnut Street	Philadelphia	554-9601
Marshall, Jeffrey	757 Hirohito Avenue	Tokyo	319-2318
McCann, Janet	550 Confucius Way	Beijing	798-7654
Moore, Paula	305 Main Street	Boston	988-7658
Newman, Karen	727 Grandview Way	Sydney	833-4852
Norton, Robert	105 Red Square	Moscow	692-8173
Olinger, Jennifer	23 Bismarck Street	Berlin	231-2435
Patrick, Bernice	560 Lincoln Avenue	Washington DC	988-3548
Smith, John	202 Spring Street	Dublin	437-2866
Stone, William	633 Lake Shore Drive	Chicago	532-4024
Williams, Rachel	488 Elizabeth Way	Toronto	692-1990
Wolff, Sharon	211 Hillside Drive	San Francisco	745-4632

Close

Figure 3.5: A Screenshot of the SIMKAP task: participants engage in a simulated task of searching through a telephone directory.



Appointment calendar for next week:

Monday	AM	10:30 - Call Norbert.
	PM	Call Charlotte.
	Evening	Go to the movies.
Tuesday	AM	10:00 - Meeting with Rita Andrews.
	PM	Call Mother and confirm for the evening. Postpone the appointment with Kurt.
	Evening	Dinner with Mother.
Wednesday	AM	9:00 - Dentist 11:30 - Call Jeffrey Marshall.
	PM	1:00 Visit the Sony plant.
	Evening	Work out at the gym.
Thursday	AM	Call the insurance company. 11:00 - Meeting with Thomas Green.
	PM	Do comparison shopping for a PC. Call Mother.
	Evening	
Friday	AM	Work on suggestions from Rita Andrews. Make an appointment for car inspection.
	PM	2:00 - Meeting with Rachel Williams. Plan for next week.
	Evening	Go swimming.

Close

Figure 3.6: Screenshot of the SIMKAP task where the participant searches for information in a simulated diary.

Data Formatting: STEW

The STEW dataset was captured using Emotiv EPOC [60], a device equipped with a filtering technique. It features a digital notch filter that effectively eliminates electrical interference at 50 and 60 hertz, frequencies that are commonly associated with electrical noise. Additionally, the device has a digital 5th-order Sinc filter which selectively allows brainwave frequencies between 0.16 and 43 Hz to pass through, further refining the signal. As a result, the STEW dataset has been preprocessed by a notch filtering technique that effectively removes 50/60 Hz.

Each participant's EEG signal from the resting and working states is saved in separate files, one for each state. The MWL ratings for these states are stored in a separate file, with each entry consisting of the participant's number followed by their ratings for the resting and working states. For example, the entry "1, 2, 8" indicates that participant 1 received a rating of 2 during the resting phase and 8 during the multitasking test.

3.1.2 BCI Hackathon Grand Challenge

This dataset was curated as part of the Passive BCI hackathon grand challenge at the Neuroergonomics Conference 2021, specifically for the cross-session MWL estimation problem [78][1]. It comprises EEG signals from 15 participants (6 female; 9 average 25 years old), captured using a 64-active Ag-AgCl Electrode system (ActiCap, Brain Products Gmbh) at a 500 Hz sampling rate according to the international 10-20 system and an ActiCHamp amplifier (Brain Products, Gmbh). In this dataset, the signal from one electrode could not be used. One electrode was dedicated to recording cardiac activity, resulting in 62 electrodes placed according to the international 10-20 system. Therefore, the final set of electrodes includes Fp1, Fz, F3, F7, FT9, FC5, FC1, C3, T7, CP5, CP1, Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, CP6, CP2, FCz, C4, T8, FT8, FC6, FC2, F4, F8, Fp2, AF7, AF3, AFz, F1, F5, FT7, FC3, C1, C5, TP7, CP3, P1, P5, PO7, PO3, POz, PO4, PO8, P6, P2, CPz, CP4, TP8, C6, C2, FC4, FT10, F6, AF8, AF4, F2. The effort was made to maintain impedances below 10 kΩ. Event markers

[1] `https://www.neuroergonomicsconference.um.ifi.lmu.de/pbci/`

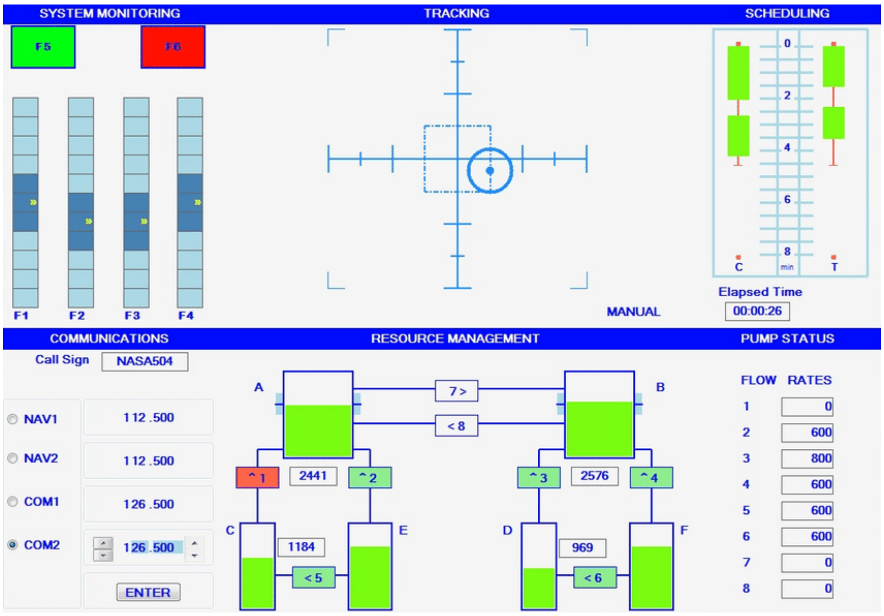
during tasks were recorded and synchronised via the LabStreamingLayer. In addition, the precise electrode locations were determined using a STRUCTURE 3D camera and the specially developed chanlocs plug-in for accurate EEG electrode placement. Details are available on github.com/sccn/get_chanlocs/wiki.

Data collection spanned three separate experimental sessions, each a week apart. Participants engaged in both a resting and a working state, echoing the STEW dataset's structure. In the resting state, participants sat with open eyes for a minute, and their EEG was recorded. Subsequently, in the working state, they tackled NASA's Multi-Attribute Task Battery II (MATB-II)[2]. This 15-minute task is segmented into three 5-minute blocks, each representing a distinct difficulty or MWL level (easy, medium, or difficult). The complexity of tasks was determined through a pseudo-random procedure. EEG data from sessions were labelled in terms of these MWL levels.

Task: MATB-II

MATB-II is a software tool for evaluating cognitive workload and performance [182]. It is used in research to assess the workload capacities of human subjects. MATB-II presents subjects with a set of tasks to perform simultaneously, which include system monitoring (SYTSMON), tracking (TRACK), resource management (RESMAN) and communications (COMM). The tasks are designed to be similar to those that might be performed in real-world operational systems, making MATB-II a valuable tool for studying human performance in complex environments. The SYTSMON task is crucial in ensuring the effective functioning of any information system and in responding to system failures. Similarly, the TRACK task simulates keeping track of multiple data streams, system states, or project progress in the field of information systems. The RESMAN task mimics resource allocation and load balancing tasks in information system management, while the COMM task is vital in collaborative information systems environments, project management, and inter-departmental communications. A screenshot of the interface of MATB-II can be viewed in Figure 3.7.

[2] `https://software.nasa.gov/software/LAR-17835-1`



Elapsed Time 00:00:26

Call Sign NASA504

○ NAV1 112.500
○ NAV2 112.500
○ COM1 126.500
● COM2 126.500

2441
2576
1184
969

FLOW	RATES
1	0
2	600
3	800
4	600
5	600
6	600
7	0
8	0

Figure 3.7: The screenshot of MATB-II task

The number of sub-tasks and their difficulties varied based on the workload condition. In the easy condition, participants engaged in two tasks - the SYTSMON and the TRACK task. The SYSMON task is displayed in the upper left of Figure 3.7. During the SYTSMON task, the subject monitored the four moving pointer dials for deviation from the midpoint and responded to the absence of the green light and the presence of the red light. At the same time, in the TRACK task presented in the upper middle window, participants were required to sustain a target at the centre of an inner box using a joystick.

In the medium condition, a third task called RESMAN was introduced, which increased the complexity of the tasks. The RESMAN window is located in the lower part of Figure 3.7, which displays six large rectangular tanks that indicate fuel levels and fluctuate in real-time via green indicators. Participants had to maintain tanks A and B at 2500 units each by activating and deactivating a set of pumps - eight pumps shown in the lower right of Figure 3.7. A red area on the failed pump indicated any pump failures.

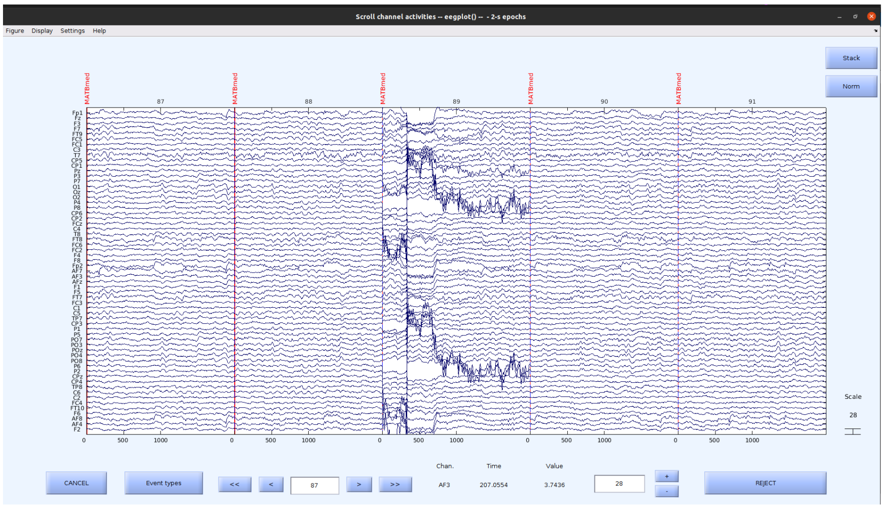
Lastly, in the difficult condition, the COMM task was added to the three previous tasks. Here, participants listened to pre-recorded auditory messages to operate the frequencies of different radios, which are displayed in the lower left of Figure 3.7. However, not all of the messages were relevant to the operator. The subject had to determine which messages were relevant and respond by selecting the appropriate radio and frequency on the communications task window.

The order of each condition was randomised, which means participants could start with something other than the easy task and end with the difficult one. Notably, the participant did not rate their workload after completing the tasks.

Data Formatting: BCI

The dataset underwent an initial preprocessing phase conducted by the hackathon organiser. Initially, data from the resting state and various tasks were isolated from the comprehensive recording. The ECG signal was meticulously removed, and the residual data were segmented into 2-second non-overlapping epochs. The criterion for data epoching was executed using workload level labelling, essentially dividing the entire dataset into sequential 2-second epochs without incorporating any 'pre-stimulus' or 'post-stimulus' data within the epochs. Consequently, they also perform artefact removal by applying the following techniques in the dataset: high-pass filtering at 1 Hz and low-pass filtering at 40 Hz using an FIR filter, electrode, and noisy independent component (IC) from muscle, heart, and eye rejection. Moreover, they also employed the average re-referencing (CAR) technique, helping to reduce common noise present across all channels. Finally, the signal was down-sampled to 250 Hz. Down-sampling reduces the data size and computational requirements by reducing the sampling rate while preserving sufficient information for analysis.

After rigorous preprocessing by the hackathon organiser, the refined BCI dataset structured the three distinct MWL levels determined by the assigned task. Each subject's EEG signal is saved in a singular directory, with each session having its own distinct directory. However, the data is fragmented due to the organisation of the different MWL levels into distinct folders, which disrupts the dataset's temporal structure. For example, Figure 3.8 shows that the MWL conditions are marked every 2 seconds, but it depicts only the medium level.



Chan. AF3

Time 207.0554

Value 3.7436

Figure 3.8: The example of EEG signal with MWL condition is marked every 2 seconds

3.1.3 Comparison of MWL between STEW and BCI Hackathon Grand Challenge Datasets

While experiments in both datasets aim to assess subjects' MWL, they exhibit significant differences in several aspects:

1. **Task Nature and Complexity**: The STEW dataset features MWL generation through a single-session SIMKAP experiment designed for job and career placement, incorporating multitasking activities that require subjects to concentrate and perform multiple tasks simultaneously. In contrast, the BCI Hackathon dataset involves the MATB-II, comprising tasks of varying complexity that simulate real-world operational systems. These tasks increase in complexity across three conditions (easy, medium, and hard), each adding more sub-tasks.

2. **Experimental Conditions**: In STEW, all activities occur within a single session, with EEG recordings taken during resting and working states. For the BCI dataset, data collection spans three sessions, with tasks segmented into different difficulty levels, influencing how workload is assessed across sessions.

3. **MWL Measurement**: STEW measures MWL directly through participant self-assessment using a 9-point post-task scale, reflecting subjective workload expe-

riences. Conversely, BCI infers MWL from task complexity. The tasks are predefined as easy, medium, or hard without direct MWL ratings from participants. This approach relies solely on the designed difficulty of the tasks rather than participant feedback.

4. **Data Labelling and MWL Levels**: In the STEW dataset, participants evaluate and assign MWL levels solely upon completing the working stage, resulting in each subject having a singular MWL rating for that stage. Meanwhile, the BCI dataset labels MWL based on task conditions, with three distinct levels (corresponding to the task difficulties) for each subject, as detailed in **Section 3.1.1**.

These distinctions highlight the varied methodologies and approaches to measuring and analysing MWL in different experimental contexts. The STEW dataset provides a subjective assessment of MWL, whereas the BCI Hackathon dataset employs a task-based evaluation method, leading to fundamental differences in data interpretation and application.

3.2 Key Components of Deep Learning Approach

In each chapter of this thesis, we employ a consistent experimental procedure for deep learning models, consisting of three key stages: data preprocessing, feature engineering, and model evaluation. We use deep learning models to assess the effectiveness of our studies in each chapter. For example, in **Chapter 4**, we compare different preprocessing techniques. In **Chapter 5**, we explore various EEG channel selection configurations and **Chapter 6** examines diverse time series cross-validation strategies. We will provide details of the features, models, and evaluation matrices used in all chapters in this section for clarity.

3.2.1 Data Preprocessing

The initial step in EEG data analysis is to address signal noise or artefacts from sources such as eye movements, muscle activity, or external electrical interference. These artefacts can contaminate EEG data, leading to inaccurate MWL assessments. This subsection describes the preprocessing methods employed in this study to remove artefacts tailored to two distinct datasets.

For the STEW dataset, a more detailed discussion of the preprocessing process is presented in **Chapter 4**. In this chapter, we introduce a key innovation: an automated pipeline for noise mitigation that improves analysis reliability while reducing expert intervention requirements. The preprocessing of BCI data is explained in **Section 3.1.2**.

3.2.2 Feature Extraction

After thoroughly cleansing and optimising the EEG data, the study progresses to the vital feature extraction stage. This process identifies key elements within the EEG signals that are closely linked with changes in MWL. Features can encompass time-domain elements, such as mean and variance, and frequency-domain elements, such as power spectral density in specific EEG bands. The selection process is meticulously designed to hone in on features that are most indicative of MWL classification, laying the groundwork for precise model training and in-depth data analysis. A comprehensive description of the feature extraction process, including the rationale for each chosen feature, will be provided in the respective experiment chapters.

In machine learning, high data dimensionality can lead to intensive computational demands. To address this, many researchers employ feature extraction to capture only pertinent EEG signal characteristics [151]. Feature selection then refines this by curating an optimised set for enhanced model performance. Various features are employed for MWL classification in EEG studies. In this research, we have delineated features into six categories: frequency, statistical, morphological, time-frequency, linear, and nonlinear. Details on each group are elaborated subsequently.

Frequency Domain

We calculated the signal power for each channel at four well-known power spectral density bands by using a fast Fourier transformation (FFT) [155]; delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz) bands were used. Each PSD band represents a different state of the human brain [87]. The delta shows deep sleep and not dreaming; thata happens when people are drowsy and drift down into sleep and dream. Alpha shows a very relaxed and deepens into a meditation. Beta appears when people are busily engaged in activities and conversation, and gamma reveals a hyper brain active and great for learning [87]. We also computed signal power features in every non-overlapping 2-Hz interval from 4–40 Hz because the non-overlapping 2-Hz could provide finer power spectrum information [219].

PSD alpha and PSD theta were extracted by a Fast Fourier transformation (FFT) [155].

Statistical Domain

The distribution of the signal can be determined from the time-domain features, i.e. the mean, standard deviation, skewness, and kurtosis.

Mean is calculated as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{3.1}$$

For each EEG channel, we calculate the mean of the signal data over a specific time window. This average electrical activity can provide insights into the brain's overall state during that period.

Standard deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - X)^2} \tag{3.2}$$

The dataset's standard deviation quantifies the values' dispersion around the mean. A low standard deviation indicates that the values are close to the mean, while a high standard deviation indicates that the values are spread out over a wider range.

Skewness [68], which is the degree of asymmetry in the distribution, was evaluated by (3.3)

$$Skewness = \frac{1}{N}\sum_{i=1}^{N}\frac{(X_i - X)^3}{\sigma^3} \tag{3.3}$$

If all the samples in a channel are uniformly distributed around the mean, they have a

Gaussian distribution.

Kurtosis [68] showing the degree of peakedness in the distribution can be represented using (3.4). Channel value with highly-tailed or high kurtosis refers to the presence of noise in the data.

$$Kurtosis = \frac{1}{N}\sum_{i=1}^{N}\frac{(X_i - X)^4}{\sigma^4} \tag{3.4}$$

Again, in this formula, N represents the total number of observations, X_i represents each individual observation, X represents the mean, and σ represents the standard deviation.

Morphological Domain

Curve length In 1D (one-dimensional) space, the curve length can be calculated by integrating over the curve. Given a function $f(x)$, the curve length L from $x = a$ to $x = b$ can be expressed as:

$$L = \int_a^b \sqrt{1 + [f'(x)]^2} dx \tag{3.5}$$

The number of peaks To count the number of peaks in a discrete signal, one can compute a second-order difference and count the number of sign changes from positive to negative, which indicate peaks. Let's denote the signal as $x[n]$ and the number of peaks as N_{peaks}. This concept isn't typically represented with a specific formula, but you can explain it using the pseudo formula:

$$N_{\text{peaks}} = \sum_{n=2}^{N-1} [(x[n] - x[n-1]) > 0 \,\&\, (x[n] - x[n+1]) < 0] \tag{3.6}$$

Average non-linear energy In signal processing, the non-linear energy operator E for a discrete-time signal $x[n]$ is typically defined as:

$$E[n] = x[n]^2 - x[n+1]x[n-1] \tag{3.7}$$

Then, the average non-linear energy E_{avg} can be computed as:

$$E_{\text{avg}} = \frac{1}{N} \sum_{n=1}^{N} E[n] \tag{3.8}$$

where N is the length of the signal.

Time-frequency Domain

Wavelet transform The Wavelet Transform is a powerful mathematical tool used for analysing localised power variations within a signal or a dataset. It decomposes a signal into different frequency sub-bands and then analyses each sub-band with a resolution matched to its scale. This allows for multi-resolution analysis, which is not possible with other traditional methods like the Fourier Transform.

The Continuous Wavelet Transform (CWT) of a function $f(t)$ with respect to a real-valued wavelet $\psi(t)$ is defined as:

$$CWT_x(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\psi^* \left(\frac{t-b}{a}\right) dt \tag{3.9}$$

where a is the scale factor. b is the translation factor. The star denotes the complex conjugate. Integration is over the entire line. The scale factor affects the width of the wavelet. The translation factor b affects the location of the wavelet. The Discrete Wavelet Transform (DWT) is a sampled version of the CWT and is computed for discrete values of the scale and translation parameters. DWT can be implemented efficiently using filter banks. In practice, DWT is used more often than CWT due to its computational efficiency.

Linear Domain

Autoregressive coefficient (AR) with p = 6 [249] was calculated in the linear domain to describe time-varying processes. An Autoregressive (AR) model represents a type of random process. It is autoregressive in that the value at a given time point is a function of previous values. The general form of an AR(p) model (p-order autoregressive model) is:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t \tag{3.10}$$

where: X_t is the time series data at time t, c is a constant, ϕ_1, ϕ_2, ...,ϕ_p are the parameters of the model, $X_{t-1}, X_{t-2}, ..., X_{t-p}$ are the values of the time series at previous p points, and ϵ_t is white noise.

To estimate the coefficients ϕ_1, ϕ_2, ...,ϕ_p, we can use the Yule-Walker equations, which are based on autocorrelations of the time series data. The ACF at lag k, denoted as $\gamma(k)$, can be represented as:

$$\gamma(k) = \frac{1}{N-k} \sum_{t=k+1}^{N} (X_t - \mu)(X_{t-k} - \mu) \tag{3.11}$$

The system of Yule-Walker equations at lag p can be represented as:

$$\begin{aligned} \gamma(0)\phi_1 + \gamma(1)\phi_2 + \ldots + \gamma(p-1)\phi_p &= \gamma(p) \\ \gamma(1)\phi_1 + \gamma(0)\phi_2 + \ldots + \gamma(p-2)\phi_p &= \gamma(p-1) \\ &\vdots \\ \gamma(p-1)\phi_1 + \gamma(p-2)\phi_2 + \ldots + \gamma(0)\phi_p &= \gamma(1) \end{aligned} \tag{3.12}$$

Which can be represented in matrix form as $[R]\Phi = \gamma$, where Φ is a vector of parameters, $[R]$ is a matrix of autocorrelations and γ is a vector of autocorrelations at lags 1 through p. The solution to the system of equations gives the estimates of the AR coefficients. In linear algebra, this can typically be solved using a method like Gaussian elimination or Cramer's rule. In practice, software packages can handle these computations.

Non-Linear Domain

The approximate entropy (ApEn) and Hurst exponent (H), treated as non-linear features, are used to quantify the unpredictability of fluctuations over time series and to measure the self-similarity of the time series, respectively.

Approximate entropy (ApEn) [165] is used to quantify the regularity and the unpredictability of fluctuations over time series. It can be represented in (3.13).

$$ApEn(m, r, N) = \phi^m(r) - \phi^{(m+1)}(r) \tag{3.13}$$

Where r is a parameter usually referred to as the filtering level, and m represents the length of compared run of data. For EEG signals data, the value of m is usually set at 2, and the value of r is set between 0.1 and 0.25 times the standard deviation of the original time series [249]. In this experiment, we set m at 2, and the r value was arbitrarily chosen at 0.2.

Hurst exponent (H) [86] is used for measuring the self-similarity of the time series. When H is equal to 0.5, it indicates no correlation in the time series; H lies between 0 to 0.5 means there are long-term anti-correlations, and H lies between 0.5 to 1 means time series have long-term correlations. H can be evaluated by (3.14).

$$H = \frac{log(R/S)}{log(T)} \tag{3.14}$$

where R denotes the range and standard deviation of the first n samples of time series data. T represents the time series data.

3.2.3 Feature Standardisation

Inherent intra- and inter-subject variability are undeniable characteristics of the EEG signal due to time-variant factors and psychological and neurophysiological parameters [219]; they can cause a data distribution shift problem [179]. Consequently, this would cause the extracted features to have poor generalisability. In this paper, a personalised feature standardisation method was applied to alleviate this problem [30, 219]. The extracted features were converted into the same scale across subjects by F_{scaled}. Assume the raw feature value is F_{raw}. L_w and U_w are the upper and lower whisker (limits), respectively, they are the measure's value distribution for generating box plot [225]. L_w = max (minimum feature value, lower quartile − 1.5 ∗ interquartile range) and U_w = min (maximum value, supper quartile + 1.5 ∗ interquartile range. The scaled feature value F_{scaled} is acquired from

$$F_{scaled} = \frac{F_{raw} - L_w}{U_w - L_w}. \tag{3.15}$$

3.3 Model Architectures

With the extracted features in hand, the study employed a range of deep learning models, including stacked LSTM, BLSTM, BLSTM-LSTM, stacked GRU, BGRU, and BGRU-GRU. These models can learn complex patterns and correlations within the data, making them powerful tools for predicting mental MWL levels. Details of these architectures are presented in Table 3.1.

Table 3.1: Deep learning model architectures

Model	Layers
Stacked LSTM	L128-L64-L40-D32-D1(D3)
BLSTM	BL128-D32-D1(D3)
Stacked GRU	G128-G64-G40-D32-D1(D3)
BGRU	BG128-D32-D1(D3)
BLSTM-LSTM	BL256-L128-L64-D32-D1(D3)
BGRU-GRU	BG256-G128-G64-D32-D1(D3)
CNN	1D-CNN(filters = 64, kernel = 3)-MaxPooling-Flatten-D32-D1(D3)

In Table 3.1, L, G, BL, BG and D refer to LSTM, GRU, BLSTM, BGRU and dense layers, respectively. For example, L128-L64-L40-L32 indicates an LSTM layer with 128 units, followed by a second LSTM layer with 64 units, a third LSTM layer with 40 units, and a dense layer with 32 units. While D1 in the last layer indicates the dense layer with 1 unit used in Task 1, D3 indicates the dense layer with 3 units used in Task 2. In this study, a dropout rate of 0.2 was applied to prevent overfitting, and the Adam optimiser was employed with an initial learning rate of 1e-04 to train all the models. An early stopping mechanism was also implemented, where the training would halt if there were no improvements in the model's performance for 30 consecutive epochs.

3.3.1 Deep Learning Model Training

Due to the distinct natures of the two datasets employed in this study (detailed in **Section 3.1**), we utilise two different cross-validation techniques.

For the STEW dataset, which has a temporal structure, we employ time series cross-validation. Unlike conventional cross-validation, time series cross-validation respects the sequential ordering of data points, strengthening the model's ability to predict MWL in real-world, time-sensitive scenarios accurately. A comprehensive elucidation of this distinct validation technique will be furnished in **Chapter 6**.

For the BCI hackathon dataset, we apply stratified cross-validation, which provides a reliable assessment of the model's performance on independent subsets of the data.

3.4 Metrics

The evaluation metrics used in this study are shown in Table 3.2. Where true positives (TP_i) for each class L_i represent the number of cases correctly predicted as belonging to that class, while false negatives (FN_i) represent the number of cases that belong to class L_i but were incorrectly predicted as belonging to another class. Conversely, true negatives (TN_i) represent the number of cases correctly identified as not belonging to class L_i, while false positives (FP_i) represent the number of cases that were incorrectly predicted as belonging to class L_i when they actually belong to a different class. Each individual class L_i, with i ranging from 1 to the total number of classes, is evaluated separately.

FRR and FAR are usually applied for measuring the performance of a biometric system [203]; this measurement is also known as Type I and Type II errors, respectively. That means FRR is the issue of the valid occasion that should be accepted are rejected and FAR occur when an unauthorised case which should actually be rejected are accepted.

Table 3.2: Evaluation matrics

Measure	Formula	Evaluation focus
Sensitivity (Recall)	$\left[\sum_{i=1}^{n} \frac{TP_i}{(TP_i + FN_i)}\right]/n$	Average per-class effectiveness of the classifier in identifying positive labels.
Specificity	$\left[\sum_{i=1}^{n} \frac{TN_i}{(TN_i + FP_i)}\right]/n$	Average per-class effectiveness of the classifier in identifying negative labels.
Precision	$\left[\sum_{i=1}^{n} \frac{TP_i}{(TP_i + FP_i)}\right]/n$	Average per-class agreement between the positive class labels and those predicted by the classifier.
Accuracy	$\left[\sum_{i=1}^{n} \frac{TP_i + TN_i}{(TP_i + TN_i + FP_i + FN_i)}\right]/n$	Average per-class overall effectiveness of the classifier.
Fscore	$\left[\sum_{i=1}^{n} \frac{2 * (Precision_i * Recall_i)}{(Precision_i + Recall_i)}\right]/n$	The average per-class balance between the model's precision and recall.
FAR	$\left[\sum_{i=1}^{n} \frac{FP_i}{(FP_i + TN_i)}\right]/n$	Average per-class proportion of incorrect acceptance of the invalid inputs by the system.
FRR	$\left[\sum_{i=1}^{n} \frac{FN_i}{(FN_i + TP_i)}\right]/n$	Average per-class proportion of incorrect rejection of the valid inputs by the system.

3.5 Statistic Analysis

The final step of each experiment is statistical analysis, which we perform to identify significant differences between groups and to investigate the hypotheses proposed in our study. The Kruskal-Wallis test, a non-parametric alternative to the ANOVA test for multi-group comparisons, is advantageous because it does not require statistical assumptions about normal population distribution, equal variances, or independent data.

If we observe significant differences between groups, we further perform pairwise post-hoc comparisons using the Wilcoxon Rank-Sum test (also known as the Mann-Whitney U test).

3.6 Chapter Summary

In this chapter, we meticulously describe the study's methodology. We begin with a detailed exploration of the foundational dataset, providing vital context regarding its scope, participant tasks, and task relevance to information system management. Next, we provide an exhaustive description of the experimental procedure, including a comprehensive feature list to facilitate a robust understanding of the analysed characteristics. This feature list will be pivotal in subsequent discussions, specifically in **Chapters 4, 5**, and **6**.

Chapter 4

EEG Preprocessing and Its Effect on Deep Learning Models in MWL Prediction

This chapter explores the effect of various EEG preprocessing techniques on the performance of deep learning models for predicting MWL levels. Given that EEG signals are prone to noise, the study examines preprocessing methods such as high-pass filters, the ADJUST algorithm, and re-referencing. The core research question investigates how these techniques influence the effectiveness of deep learning models in MWL prediction.

4.1 Introduction

Assessing MWL through EEG data is crucial, but it is challenging. One of the biggest hurdles is the presence of signal noise or ‘artefacts’ from various sources, such as eye movements, muscle activity, or external electrical interference. These artefacts not only contaminate the EEG data but also lead to inaccurate MWL assessments.

Fortunately, recent advancements in deep learning models, such as the Convolutional Neural Network (CNN) [123], RNN [116], and BLSTM-LSTM [36], have greatly improved our ability to extract information from EEG signals. These models are designed to capture the variance characteristics inherent in EEG signals, with the aim of classifying MWL status with greater accuracy.

Despite these advancements, the field still faces challenges. While neuroscientists have provided several EEG preprocessing guidelines [1, 23, 49], they remain broad and lack universal adoption, leaving researchers to decide which technique to employ for effective noise removal. Moreover, certain existing pipelines integrate visual inspection and manual labelling for noise reduction [29]. While these methods can be very useful to reduce the noise in signals, they suffer from three issues. First, these methods are time-consuming, particularly with large datasets. Second, it can introduce bias in the analysis [210]. Finally, they limit the use of such pipelines in automated processes.

Moreover, the influence of preprocessing steps on EEG analysis within deep learning remains largely unexplored, making it challenging to compare outcomes across different studies. To address a gap in the current literature by replicating the study conducted by Chakladar et al. [36], which evaluated the use of a publicly acknowledged MWL scenario and is considered state-of-the-art within the domain. However, the approach to artefact removal differs from the original paper, focusing on an automatic EEG artefact removal framework suitable for deep learning analyses. More information on the utilised scenario and the distinctive framework will be provided in **Section 4.3**.

To achieve the objective of investigating the effects of different preprocessing techniques on the effectiveness of deep learning models using EEG signals to predict MWL levels, the main research question for this chapter is posed as "What are the effects of different preprocessing techniques on the effectiveness of deep learning models using EEG signals to predict MWL levels?". The focus is on those preprocessing techniques that can be executed automatically, namely a high-pass filter, the ADJUST algorithm, and re-referencing, as they can be incorporated into deep learning models without any human intervention.

Three state-of-the-art deep learning models, specifically Stacked LSTM, BLSTM, and BLSTM-LSTM [36], have been selected to investigate the impact of these preprocessing techniques.

4.2 Research Questions

Our research questions for this chapter are as follows:

- **RQ1.1:** How do different preprocessing techniques, such as high-pass filtering, the ADJUST algorithm, and re-referencing, influence the performance of deep learning models in MWL prediction, and what preprocessing techniques are most effective for accurately interpreting EEG signals related to MWL?
- **RQ1.2:** How do preprocessing decisions affect the performance of specific deep learning models, such as Stacked LSTM, BLSTM, and BLSTM-LSTM?

4.3 Artefact Removal Techniques

Artefact removal is a critical step in ensuring the accuracy and reliability of EEG data interpretation. Noise, including muscle activity, eye movements, and heartbeats, can interfere with EEG signals. Various techniques are used to eliminate these artefacts, and we will describe them in detail in this section.

In digital signal processing, filtering is a common technique used to eliminate frequencies that are not of interest. The filtering process typically involves using low-, high-, and band-pass filters. Low-pass filters are used to remove high-frequency noise, while high-pass filters are used to remove low-frequency noise. Band-pass filters, on the other hand, are used to isolate the frequencies of interest. High-pass filtering is a key technique in signal processing that allows frequencies higher than a certain cutoff frequency to pass through while reducing the amplitude of frequencies lower than the cutoff frequency. This technique is particularly useful in applications such as audio processing, image processing, and telecommunications, where you may need to filter out low-frequency noise or other unwanted components.

ADJUST [147] is a state-of-the-art tool that uses Independent Component Analysis (ICA) to automatically detect and remove non-brain signals from EEG data. These signals, known as artefacts, can be caused by various sources, such as eye movements, heartbeats, and muscle activity. Artefacts can have a significant impact on the quality and interpretation of EEG data. ADJUST identifies which components likely represent artefacts by analysing the statistical properties of the independent components. Once the artefacts are identified, they can be removed, and the remaining components are recombined to produce a clean EEG signal. Figure 4.1 displays the topographies of all Independent Components (ICs). Those identified as artefacts by ADJUST are highlighted with a red box. Specifically, ICs 2, 3, 4, 6, and 10 are classified as artifacts by ADJUST.

Figure 4.1 shows an example of an Independent Component (IC) map computed by ADJUST.

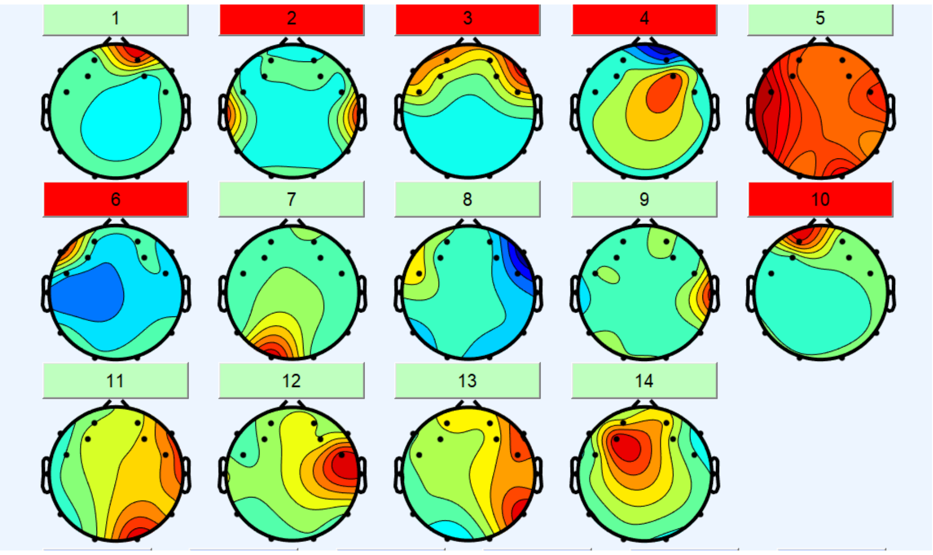


Figure 4.1: Independent Component (IC) map computed by ADJUST

In EEG analysis, re-referencing is a technique for altering the reference electrode used for the recordings. The choice of reference electrode can significantly affect the EEG signals observed. Hence, researchers may digitally re-reference the collected data to a different electrode or an average of multiple electrodes. This helps minimise the impact of reference electrode activity on the recorded signals and provides a more precise representation of brain activity. Depending on the specific needs of a study or experiment, various re-referencing methods can be employed.

4.3.1 Datasets: STEW

The dataset used for this study is the STEW dataset, which can be referred to in **Section 3.1.1**.

4.3.2 Preprocessing Scenario: STEW

EEG preprocessing consists of several techniques. Some can automatically eliminate noise from data, while others must be performed manually. This study investigates the effect of preprocessing techniques that can only be performed automatically, i.e., without any human intervention. The advantage of an automatic processing analysis is that it avoids the problem of bias from manually marking artifacts by visual inspection [210]. Therefore, the effect of three main preprocessing techniques is investigated: a high-pass filter, the ADJUST algorithm, and re-referencing. The reason for using only high-pass filtering follows the guidelines found in the literature [128].[1]

The STEW dataset, which forms the basis for this analysis, presents several limitations that restrict the use of certain preprocessing steps commonly used in EEG data analysis.

First, each stage of the experiment in the dataset is associated with a single MWL label, based on participant feedback. This provides a general overview of participants' MWL during each phase but does not capture the nuances of their mental state. Notably, the dataset lacks stimulus data, which prevents an understanding of the precise tasks or challenges participants faced.

Second, while participants in the SIMKAP experiment performed various tasks in a seemingly random order, the dataset does not provide the exact timing or sequencing of these tasks. This gap significantly hinders the ability to perform nuanced preprocessing tasks, such as data epoching, which relies on granular events or stimulus timings.

Third, as mentioned in **Section 3.1.1**, each participant's data is labeled with only one MWL level denoting their working state. This unique feature necessitates merging data from different participants to ensure a comprehensive representation of all three

[1] All preprocessing techniques were performed using EEGLAB v12, running under the cross-platform MATLAB environment.

MWL levels for model evaluation. This necessity also influences the decision to retain all channels, forgoing channel exclusions to ensure uniformity in data dimensions across all participants.

As a result of these considerations, the study defines four experimental scenarios as follows:

Table 4.1: Experimental scenarios

Scenario	Preprocessing process
1	None (Raw data)
2	High-pass filtering
3	High-pass filtering and ADJUST
4	High-pass filtering, ADJUST and Re-reference

- **Scenario 1 - Raw Data:** No preprocessing has been conducted on the data.
- **Scenario 2 - High-pass filter:** In this scenario, the EEG signal was filtered using a 1 Hz high-pass filter to remove slow linear trends. Signals with frequencies greater than a certain value were kept. A default of zero-phase FIR filter was used.
- **Scenario 3 - ADJUST:** In this scenario, Independent Component Analysis (ICA) was applied to the EEG signals previously filtered using a 1 Hz high-pass filter (from Scenario 2) employing the Runica function [136]. The artifact components identified through ICA analysis were automatically inspected by ADJUST [147], and the identified artifact components were removed without manual correction. Although general guidelines proposed by Mognon et al. [147] for running ADJUST include steps that require manual intervention, such as visual inspection, these steps were omitted to maintain an automated process.
- **Scenario 4 - Re-referencing:** In this scenario, the EEG signal was re-referenced by averaging electrical activity measured across all scalp channels. Re-referencing can typically be performed by averaging with all channels or specific reference channels, which are usually attached at locations like the earlobe or around the eye as the EOG channel. For this step, averaging was performed with all channels, as the adopted dataset did not contain any designated reference channel.

The four experimental scenarios are summarised in Table 4.1.

Figure 4.2 and 4.3 show sample data before and after preprocessing techniques applied in Scenario 4. It was observed that the artifacts were effectively removed by a high-pass filter, the ADJUST algorithm, and re-referencing techniques.

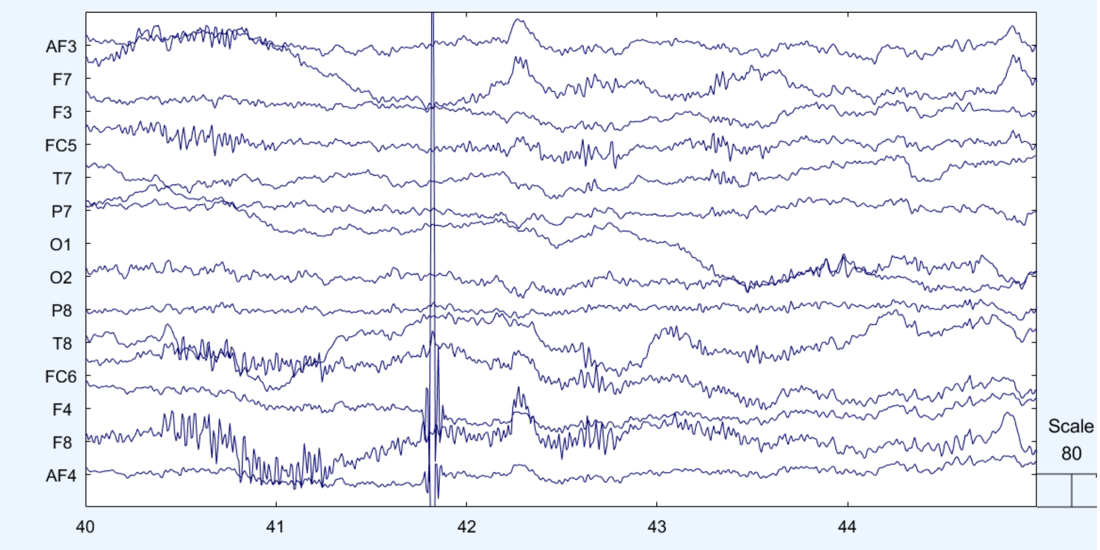


Figure 4.2: Sample continuous time EEG channel data before preprocessing

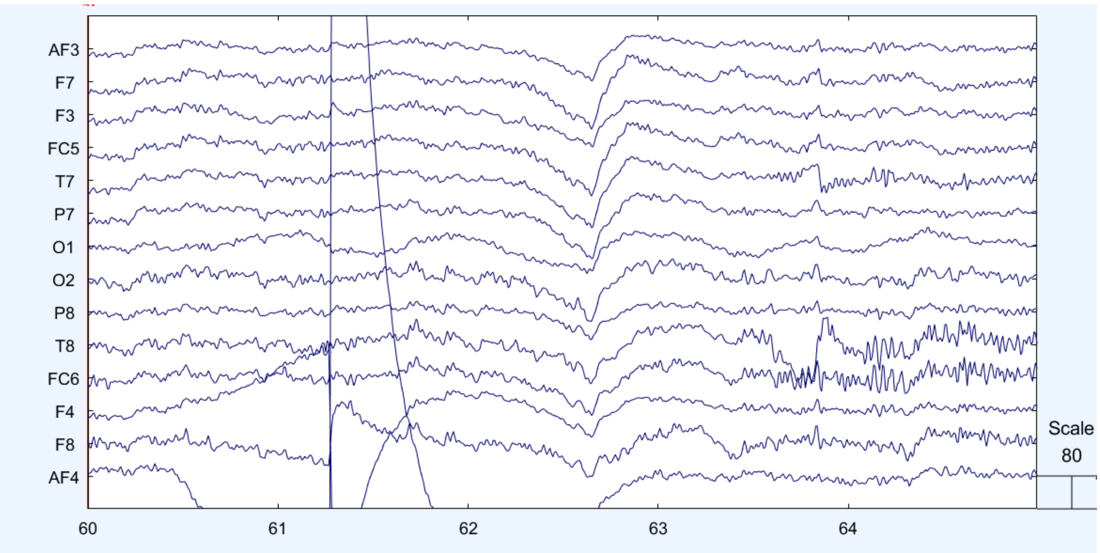


Figure 4.3: Sample continuous time EEG channel data after preprocessing

4.3.3 EEG Feature Extraction and Selection

The purpose of feature extraction is to capture characteristics of EEG signals. In the literature, various features have been utilised in EEG classification. Power spectral density (PSD) [137] is the most widely used feature. These features can be divided into sub-bands of delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz), each corresponding to different states of the human brain. The delta band indicates deep sleep without dreaming, theta occurs as individuals become drowsy and drift into sleep and dreams, alpha is associated with deep relaxation and meditation, beta with active engagement in activities and conversation, and gamma with heightened brain activity beneficial for learning [87]. The time domain, linear domain, and non-linear domain are other features that have been employed.

Features such as PSD alpha, PSD theta, skewness, kurtosis, approximate entropy, and Hurst exponent were extracted from 14 channels, as they were reported to be the best feature set for this collection, tasks, and models [36]. Skewness and kurtosis are time-domain features, while approximate entropy and Hurst exponent are non-linear domain features. All features were extracted from each EEG channel, using the entire data length for calculations. Feature extraction was performed using a sliding window approach to capture temporal dynamics. Specifically, a window of 512 samples with a 128-sample overlap was applied, ensuring a thorough analysis of signal variations. This technique is visualised in Figure 4.4, demonstrating how the sliding windows overlap to analyse the EEG data continuously.

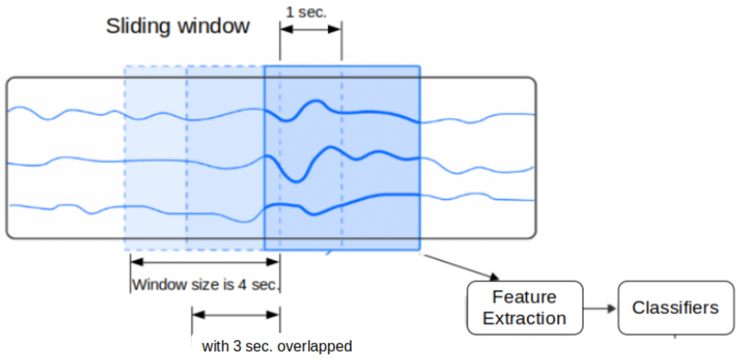


Figure 4.4: Overlapping sliding window

The details of each feature are elaborated in **Section 3.2.2**. Finally, the total number of extracted features of a 14-channel data is 84 (14 × 6). All features were standardised before further analysis. The standardised method is going to be explained in the following section.

4.3.4 Feature Standardisation

Inherent intra- and inter-individual variability in EEG signals can lead to poor generalisability of extracted features and potential data artefacts, complicating the task of building a cross-subject model for MWL recognition. This issue was mitigated using a personalised feature standardisation method, converting features into a unified scale across subjects, as detailed in 3.2.3.

4.3.5 Deep Learning Models

To investigate the effect of preprocessing techniques, the scenarios were verified using three state-of-the-art deep learning models named Stacked LSTM, BLSTM, and BLSTM-LSTM adopted from [36]. These models have been widely used in EEG signal processing for MWL classification [92, 154, 213] and are particularly useful for learning sequential data with long-term dependencies [251]. The deep learning model architectures are explained in **Section 3.3**.

4.3.6 Evaluation Metrics

The efficacy of preprocessing techniques on deep learning model performance is assessed using six metrics: sensitivity, specificity, precision, accuracy, False Acceptance Rate (FAR), and False Rejection Rate (FRR). Detailed descriptions of these evaluation metrics are provided in **Section 3.4**.

4.3.7 Experimental Procedure

In the study, classification was performed in two tasks; the first task involved binary classification, categorising EEG signals between resting state or no task and working state or during the subject performing the SIMKAP task. In the second task, three

MWL levels from subjective ratings, which are composed of low, moderate, and high, were classified using objective EEG spectral data. It was a multiclass classification. For Task 1, data from all 48 subjects were used, but for Task 2, data from subjects S05, S24, and S42 were excluded as rating data was unavailable for these subjects.

As shown in **Section 4.3.2**, in each preprocessing scenario, EEG features were extracted, and then a set of 84 optimised features based on the original paper [36] was utilised. In the model evaluation step, a cross-validation technique was applied. Initially, 80% of the dataset was used for model training, and 20% of the data was set aside as an unseen test dataset. Then, the training dataset was further split into five folds of approximately equal size, constituting a 5-fold cross-validation. Within each loop of classification model training, one fold was treated as a validation set, and the model was trained on the remaining four folds. Inside the loop, the selected features were standardised by using the F_{scaled} (3.15). Once the model was trained using 5-fold cross-validation, model performance was evaluated by comparing predicted levels with the true labels of the unseen dataset.

4.4 Results and Discussion

This section discusses different experimental results along with the performance analysis of the proposed framework.

4.4.1 Scenario 1

For Scenario 1, where only raw data without any preprocessing were used, it is observed from Table 4.2 and 4.3 that all adopted deep learning models are capable of capturing relevant information and classifying with good model performance scores. Specifically, the least sophisticated model, such as the Stacked LSTM, provided a good starting point compared to chance.

Table 4.2: The effect of the four preprocessing scenarios on MWL classification in Task 1: resting state vs working state. The numbers in the parentheses indicate the percentage change in model performance for each scenario compared to Scenario 1. An asterisk (*) indicates a statistically significant difference (p-value < 0.05) in model performance under three comparison conditions (Scenario 1 versus 2, Scenario 1 versus 3, and Scenario 1 versus 4).

Model	Scenario	Sensitivity	Specificity	Precision	Accuracy	FAR	FRR
Stacked LSTM	1	79.63	76.07	76.90	77.85	29.23	20.37
	2	81.78* (2.70)	77.11* (1.37)	78.13* (1.60)	79.44* (2.04)	22.89* (-21.69)	18.22* (-10.5)
	3	85.78* (7.72)	85.26* (12.08)	85.34* (10.98)	85.52* (9.85)	14.74* (-49.57)	14.22* (-30.19)
	4	**87.26*** **(9.58)**	**87.78*** **(15.39)**	**87.71*** **(14.06)**	**87.52*** **(12.42)**	**12.22*** **(-58.19)**	**12.74*** **(-37.46)**
BLSTM	1	78.81	76.96	77.38	77.89	23.04	21.19
	2	82.89* (5.18)	76.97 (0.01)	78.25* (1.12)	79.93* (2.62)	23.03 (-0.04)	17.11* (-19.25)
	3	87.19* (10.63)	85.78* (11.46)	85.98* (11.11)	86.48* (11.03)	14.22* (-38.28)	12.81* (-39.55)
	4	**88.30*** **(12.04)**	**86.74*** **(12.71)**	**86.94*** **(12.35)**	**87.52*** **(12.36)**	**13.26*** **(-42.45)**	**11.70*** **(-44.79)**
BLSTM-LSTM	1	83.85	79.56	80.40	81.70	20.04	16.15
	2	82.89* (-1.14)	79.78 (0.28)	80.39 (-0.01)	81.33 (-0.45)	20.22 (0.90)	17.11* (5.94)
	3	86.89* (3.63)	88.44* (11.16)	88.26* (9.78)	87.67* (7.31)	11.56* (-42.32)	13.11* (-18.82)
	4	87.93* (4.87)	90.96* (14.33)	90.68* (12.79)	89.44* (9.47)	9.04* (-54.89)	12.07* (-25.26)

Task 1: Resting vs Testing State

Data in Table 4.2 shows that the highest accuracy scores are from the BLSTM-LSTM model, followed by the BLSTM and Stacked LSTM models, with scores of 81.70%, 77.89%, and 77.85%, respectively. It could be implied that the more sophisticated model provided higher model accuracy. However, examining the sensitivity scores reveals a different pattern; the sensitivity of the BLSTM model was slightly lower than that of the most unsophisticated model architecture, like the Stacked LSTM. This indicates that the BLSTM model more frequently misclassified subjects who were in a resting state as being in a working state than other models. Conversely, the highest specificity score of 76.96% in Task 1 was obtained by the BLSTM model, indicating its proficiency in correctly identifying non-resting states.

Table 4.3: The effect of four preprocessing scenarios on MWL classification in Task 2: low vs moderate vs high MWL. The numbers in the parentheses indicate the percentage change in model performance for each scenario compared to Scenario 1. An asterisk (*) indicates a statistically significant difference (p-value < 0.05) in model performance under three comparison conditions (Scenario 1 versus 2, Scenario 1 versus 3, and Scenario 1 versus 4).

Model	Scenario	Sensitivity	Specificity	Precision	Accuracy	FAR	FRR
Stacked LSTM	1	66.52	84.59	68.34	78.57	15.41	33.48
	2	70.96* (6.67)	82.11* (-2.93)	66.48* (-2.72)	78.40 (-0.22)	17.89* (16.09)	29.04* (-13.26)
	3	81.81* (22.99)	90.44* (6.92)	81.06* (18.61)	87.57* (11.45)	9.56* (-37.96)	18.19* (-45.67)
	4	**83.56*** **(25.62)**	**91.83*** **(8.56)**	**83.65*** **(22.40)**	**89.07*** **(13.36)**	**8.17*** **(-46.98)**	**16.44*** **(-50.90)**
BLSTM	1	22.52	95.94	73.52	71.41	4.06	77.48
	2	17.11* (-24.02)	96.63 (0.72)	71.74* (-2.42)	70.12* (-1.81)	3.37* (-17.00)	82.89* (6.98)
	3	24.11* (7.06)	97.33* (1.45)	81.89* (11.38)	72.93* (2.13)	2.67* (-34.24)	75.89* (-2.05)
	4	**29.04*** **(28.95)**	**97.44*** **(1.56)**	**85.03*** **(15.66)**	**74.64*** **(4.52)**	**2.56*** **(-36.95)**	**70.96*** **(-8.42)**
BLSTM-LSTM	1	68.22	85.74	70.52	79.90	14.26	31.78
	2	71.74* (5.16)	87.17* (1.67)	73.65* (4.44)	82.02* (2.65)	12.83* (-10.03)	28.26* (-11.08)
	3	85.56* (25.42)	92.09* (7.41)	84.40* (19.68)	89.91* (12.53)	7.91* (-44.53)	14.44* (-54.56)
	4	86.59* (26.93)	93.43* (8.97)	86.82* (23.11)	91.15* (14.08)	6.57* (-53.93)	13.41* (-57.80)

Task 2: Low vs Moderate vs High MWL Level

When examining model accuracy in Table 4.3, it is observed that the best model remains the BLSTM-LSTM, followed by the Stacked LSTM, with the BLSTM being the least effective. Model accuracies for the BLSTM-LSTM, LSTM, and BLSTM are 79.90%, 78.57%, and 71.41%, respectively. In this task, a concerning trend is noted in the lower sensitivity score of the more sophisticated BLSTM model compared to the Stacked LSTM. Specifically, while the Stacked LSTM achieved a sensitivity of 66.52%, the BLSTM only reached 22.52%. This indicates that the BLSTM often failed to classify the true level of the subject's MWL accurately. For instance, when a subject was at a low level of MWL, the model tended to misclassify the low MWL level as medium or high. Moreover, a very high specificity score of 95.94% was observed for the BLSTM, suggesting its effectiveness at correctly identifying when subjects were not at a low level of MWL. For example, when subjects were not at a low MWL level, the BLSTM model was more accurate in marking medium or high levels compared to other models.

4.4.2 Scenario 2

In Scenario 2, the artefact components were removed from the dataset for the first time using a high-pass filter technique. The same pattern of results is observed from Table 4.2 and 4.3 as in the previous scenario, with a few exceptions in Task 1.

Task 1: Resting vs Testing State

As seen in Table 4.2, the highest model accuracy is from the BLSTM-LSTM, followed by the BLSTM and Stacked LSTM. The accuracy of the BLSTM and Stacked LSTM increased by 2.62% and 2.04%, respectively, compared to Scenario 1. However, the accuracy of the BLSTM-LSTM model decreased by 0.45% compared to Scenario 1. It was observed that there was no decrease in the sensitivity score of the BLSTM compared to the Stacked LSTM in this scenario.

Task 2: Low vs Moderate vs High MWL Level

As indicated in Table 4.3, the best model remains the BLSTM-LSTM; however, the second-best model becomes the Stacked LSTM and the least effective model to classify three levels of MWL is the BLSTM. The accuracy of the BLSTM-LSTM increased by 2.65%. However, for the Stacked LSTM and BLSTM, the accuracy decreased by 0.22% and 1.81%, respectively, compared to Scenario 1. In this task of Scenario 2, a sharp decrease in the sensitivity score of the BLSTM model compared to the Stacked LSTM was still observed. While the Stacked LSTM achieved 70.96% sensitivity, the BLSTM reached only 17.11% sensitivity. Nevertheless, the highest specificity score of 96.63% in this task was observed for the BLSTM.

From these data, it is observed that there was some improvement in model performance, but there were also some decreases. It appears that a high-pass filter technique does not contribute significantly to model performance. Moreover, it might distort the EEG signal somehow and cause a loss of some information from the EEG data. Furthermore, the problem of the BLSTM model still appears in this scenario.

The unexpected sensitivity underperformance of the BLSTM model in this scenario might be attributed to the high-pass filtering technique used. While removing unwanted low-frequency noise, high-pass filters also eliminate essential information crucial for accurate classification. This loss could disproportionately affect the BLSTM's sensitivity due to its bidirectional approach, making it particularly vulnerable to signal quality variations caused by such filtering.

4.4.3 Scenario 3

In Scenario 3, the artefact components were further removed from the dataset by using the ADJUST algorithm. A similar pattern of results is observed from Table 4.2 and 4.3 as seen in Scenario 2, with a few exceptions. The exception is noted in Task 1.

Task 1: Resting vs Testing State

From the data in Table 4.2, it is noted that the highest accuracy scores are from the BLSTM-LSTM, followed by the BLSTM and Stacked LSTM; their accuracy increases by 7.31%, 11.03%, and 9.85%, respectively, compared with those in Scenario 1. It was observed that the BLSTM obtained the highest sensitivity score in this task, followed by the BLSTM-LSTM and Stacked LSTM.

Task 2: Low vs Moderate vs High MWL Level

Considering the accuracy score in Table 4.3, the best model remains the BLSTM-LSTM, followed by the Stacked LSTM and the BLSTM. Model accuracy of the BLSTM-LSTM, Stacked LSTM, and BLSTM rose by 12.53%, 11.45%, and 2.13%, respectively, compared with those in Scenario 1. In this scenario, a significant decrease in sensitivity score in the BLSTM, compared with the Stacked LSTM, was still observed. While the Stacked LSTM enhanced to 81.81% sensitivity, the BLSTM acquired 24.11% sensitivity. Furthermore, it was observed that the BLSTM model obtained the highest specificity score of 97.33% in this task.

Generally, substantial progress in model performance was seen in this scenario after applying the ADJUST algorithm for further artefact removal from EEG signals. The ADJUST has significantly contributed to model performance by filtering irrelevant information from the data. However, the problem of low sensitivity in the BLSTM model was still observed in this scenario.

4.4.4 Scenario 4

In Scenario 4, additional preprocessing techniques were incorporated into the pipeline, specifically re-referencing.

Task 1: Resting vs Testing State

Considering the accuracy scores in Table 4.2, the best model remains the BLSTM-LSTM, which shows a 9.47% improvement compared to Scenario 1. In this scenario, both the BLSTM and Stacked LSTM achieved the same accuracy, albeit with different improvement rates. The accuracy score of the BLSTM increased by 12.36%, while the Stacked LSTM model improved by 12.42%, compared with their performances in Scenario 1. The sensitivity pattern in this scenario was similar to Scenario 3; however, the numbers were slightly increased.

Task 2: Low vs Moderate vs High MWL Level

In Task 2, the performance behaviour observed was the same as that found in Scenario 3, as shown in Table 4.3. Model accuracy of the BLSTM-LSTM, Stacked LSTM, and BLSTM increased by 14.08%, 13.36%, and 4.52%, respectively, compared to their performance in Scenario 1. A significant decrease in sensitivity score of the BLSTM model compared to the Stacked LSTM was still evident. While the Stacked LSTM model achieved a sensitivity of 83.56%, the BLSTM reached 29.04% sensitivity. The BLSTM model continued to obtain the highest specificity score of 97.44%, although the numbers did not differ significantly from Scenario 3.

To address the first research question (RQ1.1), various preprocessing techniques were investigated, revealing that integrating more steps across different tasks resulted in a noticeable improvement in model performance. This behaviour was consistent across all evaluated metrics, showing a linear decrease in error (FAR and FRR), with some exceptions. The exception occurred in Scenario 2 when a high-pass filter was used. In this scenario, an increase in error numbers was observed in Task 1 of the BLSTM-LSTM model and Task 2 of the Stacked LSTM and BLSTM. As shown in Table 4.2, FAR from Task 1 of the BLSTM-LSTM model starts at 20.04%; however, after high-pass filtering, the number climbs to 20.22%. FRR, which starts at 16.15%, changes to 20.22% after filtering. Considering Task 2 of the Stacked LSTM, as results in Table 4.3 indicate, an increase in error numbers was found in FAR, which starts at 15.41% and climbs to 17.89%, and FRR from the BLSTM rises by 6.98% after filtering. Smaller FAR and FRR values indicate better model performance. Conversely, the model captures the data structure more effectively when it exhibits larger sensitivity, precision, and accuracy values. Therefore, filtering EEG signals to remove artefacts might be a common preprocessing step, but it could introduce temporal distortions in the signals. Another important observation is that after EEG signals were preprocessed using the ADJUST algorithm, performances notably increased across the state-of-the-art deep learning models. The highest classification performance across the deep learning models was achieved when using all preprocessing techniques. Hence, there are opportunities for deep learning models to achieve higher performance by enhancing artefact removal techniques in the preprocessing stage.

RQ1.2: Evaluating the impact of preprocessing on specific deep learning models reveals that the less sophisticated model, i.e., Stacked LSTM, already provided a good starting point compared to chance when using raw data. This indicates that even simple deep learning model architectures can capture relevant information.

However, the BLSTM exhibited very low sensitivity, especially in Task 2: low vs medium vs high MWL. The bidirectional approach, which processes data forwards and backwards, did not seem to contribute significantly. The training strategy of the bidirectional model, which involves concatenating two independent neural networks—one processing inputs in chronological order and the other in reverse—might be problematic [186]. Generally, the model requires input from both past and future contexts. In practical scenarios where future values of time series are not available at the time of prediction, this could explain why the model struggles to categorize MWL levels effectively. Therefore, this study raises an important question: “How can a bidirectional neural network be effectively applied to time series analysis?”.

Furthermore, the BLSTM-LSTM had the highest starting point and also achieved the best performance. Notably, results from Table 4.2 show that the accuracy for the BLSTM-LSTM in Task 1 increased from 81.70% to 89.44%. The difference in performance from raw data in Scenario 1 compared with Scenario 4 became less significant. In Task 2, the accuracy of the BLSTM-LSTM climbed from 79.90% to 91.15%, as shown in Table 4.3.

In this analysis, an analysis of variance on ranks was performed using the Kruskal-Wallis H test [139] to test whether there were statistically significant differences between the four scenarios of EEG preprocessing on model performance. It was found that there was a statistically significant difference among the four scenarios across three models ($p < 0.05$). Consequently, a pair-wise comparison was conducted using the Mann-Whitney U test [140] to identify differences between model performance trained by raw data and preprocessed data. The Mann-Whitney U test indicated statistically significant differences in the performance of models at Scenarios 2, 3, and 4 compared to Scenario 1, which serves as the baseline for every model, with few exceptions. The asterisk in Table 4.2 and 4.3 indicates the statistically significant results where the p-value < 0.05. This demonstrates that the preprocessing techniques applied have a significant effect in improving the effectiveness of the deep learning models on EEG signals. Overall, even though the model is sophisticated, a suitable preprocessing pipeline still provides an advantage. Thus, there are opportunities for deep learning models to achieve higher performance by enhancing artifact removal techniques in the preprocessing stage.

4.5 Conclusion

In this chapter, the effect of preprocessing techniques defined by neuroscientists on the effectiveness of deep learning models was explored. The focus was placed on automated techniques including a high-pass filter, the ADJUST algorithm, and re-referencing. The effect of these preprocessing techniques was then verified across three state-of-the-art deep learning models: Stacked LSTM, BLSTM, and BLSTM-LSTM, using a publicly available MWL Scenario, STEW [128]. Findings indicate that the ADJUST algorithm had the most significant impact on performance across the investigated deep learning models compared to other techniques. Additionally, results demonstrated that EEG signals preprocessed using a combination of a high-pass filter, the ADJUST algorithm, and re-referencing provided the highest classification performance across the models. However, it was also observed that raw signals were sufficient for classification, as evidenced by each model's performance, particularly the BLSTM-LSTM. This model exhibited a strong starting point in both tasks. As models become more sophisticated, their potential to extract relevant information from raw data increases, reducing the need for preprocessing. Therefore, future work should focus on developing a deep learning model sophisticated enough to automatically incorporate preprocessing within its architecture.

The investigation into the effect of three artifact removal techniques, namely a high-pass filter, the ADJUST algorithm, and re-referencing, yielded positive findings, encouraging further research to explore the impact of other artifact removal techniques. Additionally, the effects of these techniques on a broader range of deep learning models, such as GRU, will be evaluated. Finally, future studies will explore how to integrate EEG preprocessing techniques into the deep learning model architecture.

4.6 Chapter Summary

- This chapter contextualizes the user study within the first research goal, addressing prevalent issues surrounding EEG preprocessing procedures in machine learning literature. Notably, the complexity and inconsistency of preprocessing techniques affect the efficiency and accuracy of deep learning models interpreting EEG data.

- Precise research questions are formulated to explore the consequential relationship between various preprocessing techniques and the accuracy of MWL state predictions through deep learning models, providing a strategic pathway for subsequent experimental scenarios and methodologies.

- Experimental scenarios are curated to scrutinise preprocessing techniques across two distinct datasets, forming a comparative study to discern the impacts of various preprocessing methods, specifically filtering, ADJUST algorithm application, and different re-referencing strategies on the output of three deep learning architectures. These scenarios serve as practical groundwork to probe the formulated research questions.

- Methodologies are elucidated, detailing the preprocessing steps and the mechanics of the deep learning models utilised, namely, Stacked LSTM, BLSTM, and a hybrid BLSTM-LSTM. This section offers comprehensive insights into the practical and theoretical underpinnings of the employed techniques, setting a foundation for subsequent analyses and findings. Furthermore, this structured flow not only provides crucial insights into the vital role of preprocessing in EEG signal interpretation through deep learning models but also paves a path forward for future research, particularly in honing methodologies for enhanced predictive accuracy and model robustness in EEG analyses.

- Data analysis and its resultant findings are presented, combining statistical rigour with insightful interpretations. Preprocessing methodologies are analysed in tandem with performance metrics of deep learning models across four structured scenarios, revealing substantial insights into their interplay and impact on prediction accuracy.

- The concluding sections sift through major findings, providing a critical review and an insightful discussion on how different preprocessing techniques influence model performance, elucidating the nuanced impacts of preprocessing on model efficiency and subsequent implications on EEG signal interpretation and MWL state prediction.

- Limitations in the dataset were also identified, highlighting areas beneficial for the community in data collection, preparation, and release. A robust benchmark dataset can expedite community advancements, facilitating progress without inheriting issues. During the preprocessing phase, several restrictions were observed in both datasets, which subsequently influenced the processes involved in experimental design. The STEW dataset, as mentioned in **Section 3.1.1**, presented several issues. These challenges stemmed from the dataset's limitation of providing solely EEG signals in a CSV file without any accompanying stimulus or event information. The lack of presence hindered the process of epoching the data, a crucial step in preparing EEG data. Epoching allows researchers to focus on specific periods when stimuli were presented and facilitate the study of event-related potentials, provided that event markers are available. Moreover, the absence of the mastoid channel in the dataset impeded the capacity to execute the reference procedure using average referencing—an approach that carries inherent risks since it may lead to data distortion if a channel contains outlier data or high noise.

- Additionally, the modelling technique was limited because each subject was assigned only one label for the MWL level. The execution of model training for individual participants was impossible and could only be achieved by concatenating data. From a favourable standpoint, this methodology effectively safeguarded the model from being limited to a certain subject, resulting in a more universally applicable model.

Chapter 5

EEG Channel Selection Enhancement with Covariance Estimators in Riemannian Geometry

This chapter explores optimising channel selection to increase the computational efficiency and performance of deep learning models for MWL classification. It examines the consequences of various covariance estimators on the Riemannian distance-based channel selection strategy and their implications on various deep learning models.

5.1 Introduction

Striving to balance data richness with computational efficiency, this study embarks on the crucial stage of channel selection. The goal is to pinpoint the least number of EEG channels that still retain the efficacy needed for precise MWL prediction. By optimising this selection, the study ensures a streamlined yet robust dataset, thereby enhancing the feasibility of the entire analytical process without sacrificing model accuracy.

Recently, the Riemannian geometry approach has become a popular method for channel selection in EEG analysis. This technique utilises the covariance matrices of

EEG signals as features, which are then manipulated and classified directly. By examining the covariance properties of these features, researchers can determine the most meaningful channels for further analysis. For example, in the study by Barachant et al. [11], Riemannian geometry was used to select fewer electrodes for brain signal analysis. The method assessed how well different electrodes could distinguish between classes by measuring the Riemannian distance between their spatial covariance matrices.

This method was applied to a two-class motor imagery paradigm, utilising the sample spatial covariance matrix. Similarly, Qu et al. [169] employed Riemannian geometry to minimise information redundancy, extracting key features from the most relevant time-frequency bands of the selected channels to enhance decoding for BCI. The EC estimator was utilised to analyse EEG signals in this binary classification problem, focusing on the left- and right-hand motor imagery tasks. This technique successfully reduced the number of electrodes from 61 to 18-32 using the LW estimator by sequentially pruning channels to maximise the Riemannian distance between the class-conditional covariance matrices [176].

Prior research primarily advanced binary classification. This study extends Riemannian channel selection to more complex multiclass classification across easy, medium, and difficult MWL levels, requiring multiple class comparisons. Additionally, past studies utilised various covariance estimators without thoroughly exploring their advantages and drawbacks, potentially affecting technique effectiveness. Consequently, the primary aim is to assess different covariance estimators' impacts on channel selection and multiclass classification performance. By identifying the optimal number of channels for model accuracy, this study aims to enhance the efficiency and effectiveness of channel selection in EEG analysis.

5.2 Research Questions

RQ2: How can EEG channel selection be optimised using covariance estimators in Riemannian geometry for MWL level classification?

- **RQ2.1:** How do different covariance estimators influence classification performance?

- **RQ2.2:** How does the reduction of EEG channels affect the classification model's performance?

RQ3: How can the selection and optimisation of EEG channels, including the targeting of specific brain regions, enhance the classification of MWL levels?

- **RQ3.1:** Which specific regions of the brain are indicative of the MWL level, and how can they be targeted effectively?
- **RQ3.2:** How can channel selection methods be employed to focus on these specific regions?

5.3 Experimental Set-Up

In this study, channel selection is undertaken using the BCI Hackathon dataset. As illustrated by the results presented in that chapter, model performance is enhanced when employing scenario 2, in which the data is further denoised using ADJUST. Therefore, in this experiment, the dataset processed according to that scenario is adopted for EEG channel selection.

Four distinct covariance estimators are incorporated, the descriptions of which are provided in **Section 5.5**. The targeted numbers of channels for selection have been set to 4, 8, 16, and 32, which represent a low-density EEG channel configuration [192]. For comparative purposes, a model utilising all EEG channels, i.e., 62, is also trained. The deep learning models employed in this experiment are elucidated in **Section 3.3**.

Channel selection was not performed on the STEW dataset because the method requires three MWL levels per subject to measure the distance between channels, but each subject in the STEW dataset only has one workload level.

5.3.1 Dataset: BCI hackathon

The effect of different covariance estimators on Riemannian channel selection was investigated using a publicly available EEG MWL dataset from the 2021 Neuroergonomics

Conference Passive BCI hackathon [1] [78]. The dataset is explained in 3.1.2

5.3.2 Data Preprocessing

Before processing, artefacts in the EEG signals needed to be removed. Hackathon organisers initially preprocessed the dataset by splitting the data from the task and resting states of the complete EEG recording. Subsequently, the heart activity electrode was removed, and the data was segmented into two-second non-overlapping epochs. The dataset was then further subjected to high-pass filtering at 1 Hz and low-pass filtering at 40 Hz using FIR filters, resulting in the rejection of electrodes and noisy independent components from muscle, heart, and eye activity. Average re-referencing downsampled the signal to 250 Hz. Based on the findings presented in **Chapter 4**, ADJUST is highlighted as a potent noise removal technique. To guarantee optimal data quality, it was employed [147] to eliminate any residual artefact components.

5.3.3 EEG Feature Extraction and Selection

In machine learning, high-dimensional data can pose significant challenges, including time-consuming and computationally expensive calculations. Traditional feature extraction techniques were employed to address this issue, capturing only relevant signal characteristics from EEG data. This strategy enables a compact, fast model tailored to the specific use case through customised features and interpretability. In this study, a set of features broadly classified into six groups was calculated. Details of each feature are described in **Section 3.2.2**. Rather than performing feature selection on individual features, a different approach was adopted where entire channels of data were selected. By choosing this method, all the features within those selected channels are utilised.

5.3.4 Feature Standardisation

The features were standardised as detailed in **Section 3.2.3**.

[1] `https://www.neuroergonomicsconference.um.ifi.lmu.de/pbci/`

5.3.5 Deep Learning Models

This research employs the Gated Recurrent Unit (GRU) family, including Stacked GRU, Bidirectional GRU (BGRU), and Bidirectional Stacked GRU (BGRU-GRU), to analyse sequential and time series data effectively. These models have demonstrated efficacy in MWL classification [45, 111, 213]. The architectures of the models are described in **Section 3.3**

5.3.6 Evaluation Metrics

In the experiments, the performance of the three-class classification is evaluated using various metrics, including Accuracy, Sensitivity (Recall), Precision, and F1-score.

5.3.7 Experimental Procedure

Cross-validation is a crucial technique for evaluating a machine learning model's performance, and the choice of cross-validation method depends on the objectives of the analysis [112]. The hackathon organiser organised the EEG data into distinct folders, including easy, medium, and difficult levels. This rigorous technique resulted in the data's temporal sequence being reorganised. To mimic a real-world experiment, where each difficulty level would be performed randomly, the features were shuffled before splitting them into training and validation sets. Stratified sampling was used to assign 80% of the data for model training and 20% for validation, thereby ensuring unbiased performance evaluation by equally representing labels in each class.

5.4 Riemannian Geometry in EEG Data Processing

The seminal work by Barachant et al. (2010) [12] introduced Riemannian geometry to the realm of EEG data processing, marking a transformative approach in this field. This method offers a robust and interpretable framework for channel selection, enhancing signal interpretation and discrimination. Within this framework, EEG data are represented as covariance matrices, residing in the space of symmetric positive-definite (SPD) matrices—a curved Riemannian manifold. Channel selection involves assessing the discriminative power of channels based on their influence on Riemannian distances between these covariance matrices. These distances quantify the dissimilarity between matrices derived from different EEG conditions. By mapping these matrices to a tangent space—a flat space where mathematical operations are more tractable—classical methods can be employed to evaluate channel relevance. This geometric perspective not only streamlines the channel selection process but also enhances the robustness and interpretability of EEG-based analyses, particularly in tasks such as BCI classification, where optimal channel combinations are crucial.

5.5 Covariance Estimation Techniques

The comprehension of interrelationships between multiple variables is of paramount significance in numerous scientific studies and applications. The covariance matrix serves as a standard tool to capture these interrelationships. This section delves into various approaches to estimating covariance, particularly the Riemannian approach. This method utilises covariance matrices to identify the links between MWL levels. The focus is on four distinct estimators, all based on the fundamental concept of Riemannian distance.

5.5.1 Empirical Covariance (EC)

Empirical Covariance, often termed as the sample covariance matrix or EC, is established using:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{5.1}$$

Here, n signifies the number of observations, x_i is the i-th observation, and $\bar{x}$ is the sample mean. While widely used, this method has drawbacks like susceptibility to outliers, noise amplification, and multicollinearity tendency, possibly affecting the accuracy of subsequent analyses. Alternative covariance estimators such as SC, LW, and OAS have been developed to resolve these issues.

5.5.2 Shrunk Covariance (SC)

The Shrunk Covariance matrix estimator addresses the limitations of the sample covariance matrix by combining it with a structured target matrix (T). The idea is to "shrink" the sample covariance matrix towards the target matrix to obtain a more stable and robust estimate. The SC matrix is computed as follows:

$$Shrunk\_Cov = \alpha T + (1 - \alpha)S \tag{5.2}$$

In the equation above, S is the sample covariance matrix, T is the target matrix, and α is a shrinkage parameter between 0 and 1. The target matrix is typically an identity matrix or a diagonal matrix with the average of the variances on the diagonal. The choice of α can be made using cross-validation or by minimising some criterion, such as the mean squared error.

5.5.3 Ledoit-Wolf (LW)

Further refining the concept of shrinkage, the Ledoit-Wolf estimator aims at deriving an optimised covariance matrix by minimising the difference (mean squared error) between the actual covariance matrix and the SC matrix. The LW estimator is described by:

$$Ledoit - Wolf = \beta I + (1 - \beta)S \tag{5.3}$$

In this formula, I is an identity matrix scaled by the average of the diagonal elements of the sample covariance matrix and β is the shrinkage factor. Similarly, the OAS estimator aims to find a shrinkage factor that minimises the mean squared error in an oracle setting, where the true covariance matrix is known.

5.5.4 Oracle Approximating Shrinkage (OAS)

The Oracle Approximating Shrinkage estimator is a modern approach to covariance estimation that approximates the ideal "oracle" estimator using available data. Its unique adaptability makes it the superior choice for high-dimensional data, providing superior performance compared to other shrinkage estimators.

$$OAS = \gamma I + (1 - \gamma)S \tag{5.4}$$

Here, I is an identity matrix, and γ is the shrinkage factor computed based on the trace and Frobenius norm of the sample covariance matrix.

5.5.5 Channel Number

In this study, 4-, 8-, 16-, and 32-channel configurations were evaluated to optimize the number of channels for each estimator. The goal was to identify the best covariance estimator and optimal channel number for high model performance. Additionally, six neural network models—LSTM, BLSTM, BLSTM_LSTM, GRU, BGRU, and BGRU_GRU—were compared to further evaluate the effectiveness of these covariance estimators.

The Riemannian distance within this framework, which quantifies the shortest distance between two points following a curved trajectory, is defined by equation 5.6. The Riemannian mean is expressed by equation 5.7.

$$\delta_R(C_1, C_2) = \log \|C_1^{-1} C_2\|_F = \left[\sum_{i=1}^{N} \log^2 \lambda_i \right]^{\frac{1}{2}} \tag{5.5}$$

where C_1, C_2 are two different covariance matrices respectively, λ_i denotes the i^{th} eigenvalue of $C_1^{-1}C_2$, $\| \cdot \|_F$ denotes the Frobenius norm, and $\log(\cdot)$ is the log-matrix operator.

$$\bar{C} = \arg\min_{C} \sum_{i=1}^{N} \delta_R^2(C, C_i) \tag{5.6}$$

$$\text{Crit} = \delta_R(\bar{C}_i, \bar{C}_j) = \|\log(\bar{C}_i^{-1}\bar{C}_j)\|_F \tag{5.7}$$

The pseudo-code employed for channel selection is shown in Algorithm 5.1.

Algorithm 5.1 Pseudo code for Riemannian distance-based channel selection.

```
Input: The preprocessed N-channel EEG signals $X_i$, the total number of channels $N$, the number of selected channels $N_{Ch}$, Number of MWL levels $N_{levels}$
Output: $N_{Ch}$ selected channel subset
procedure CHANNEL SELECTION
    Compute the covariance matrix $C_i$ of $X_i$;
    Compute the Riemannian means of each level $\bar{C}_1$, $\bar{C}_2$,...,$\bar{C}_{N_{levels}}$;
    for n = 1: $N_{Ch}$ do
        for k = 1: N do
            Remove $k^{th}$ channel by reducing the $k^{th}$ row and column from matrix $\bar{C}_1$, $\bar{C}_2$,...$\bar{C}_{N_{levels}}$ to $\bar{C}'_1$, $\bar{C}'_2$,...,$\bar{C}'_{N_{levels}}$;
            Compute the sum of pairwise Riemannian distances between all classes' Riemannian means $D_{Ksum}$;
        end for
        Select the channel corresponding to a minimum $D_{Ksum}$ value;
    end for
    return $N_{Ch}$ selected channels;
end procedure
```

To compute the sum of pairwise Riemannian distances for multiple classes, distances between Riemannian mean covariance matrices of each pair in the three MWL levels were calculated: $\bar{C}'_1$ and $\bar{C}'_2$, $\bar{C}'_1$ and $\bar{C}'_3$, and $\bar{C}'_2$ and $\bar{C}'_3$. The summed distances serve as an overall measure of how dissimilar the classes are. Channels that, when removed, result in the smallest summed distance are presumably the least informative for distinguishing between classes and, therefore, are selected for removal first.

5.6 Results and Discussion

Table 5.1: Pairwise comparisons of mean accuracies across various estimators

Factor	**Comparison**	**Mean**	**p-value**
Estimators	EC vs SC	(93.14, **93.74**)	0.0121*
	EC vs LW	(93.14, **93.66**)	0.0239*
	EC vs OAS	(93.14, **93.70**)	0.0101*
	SC vs LW	(**93.74**, 93.66)	0.7885
	SC vs OAS	(**93.74**, 93.70)	0.9790
	LW vs OAS	(**93.66**, 93.70)	0.7757

In the analysis, the Kruskal-Wallis H test was utilised to analyse variance on ranks to probe the impact of various factors on the accuracy of models designed for MWL level classification [139]. The aim was to determine:

1. Whether there were statistically significant differences in model performance across four covariance estimators used for Riemannian channel selection.

2. Whether significant variations existed across four channel configurations for Riemannian channel selection regarding model performance.

3. If there were any discernible performance differences among six distinct models.

The findings indicated statistically significant differences across all three points:

1. A statistically significant difference was identified among the covariance estimators with p-values < 0.05.

2. Significant variations in their impact on model performance were also observed among the four channel configurations.

3. The six distinct models showed discernible performance differences as well.

To delve deeper into these differences, pairwise comparisons were conducted using the Mann-Whitney U test (also known as the Wilcoxon Rank-Sum Test) [140], aiming to pinpoint specific discrepancies in model performance associated with each factor.

The results from Table 5.1 showed significant performance differences when comparing SC, LW, and OAS to EC, the conventional covariance estimator. Despite initial assumptions that EC might underperform, it trailed the other three. Specifically, SC, LW, and OAS achieved accuracies of 93.74%, 93.66%, and 93.70%, respectively, outperforming EC. As presented in Table 5.1, results marked with an asterisk denote statistical significance with p-values < 0.05. However, SC, LW, and OAS showed no significant performance differences among themselves. Notably, the SC estimator stood out as the top performer, suggesting a preference for SC, LW, or OAS over traditional methods.

These insights address **RQ2.1**: How do different covariance estimators influence classification performance? In conclusion, the findings underscore the importance of selecting the appropriate covariance estimator in the success of deep learning models for EEG signals. While a robust underlying model is crucial, the optimal covariance estimator for Riemannian channel selection can enhance performance. This sheds light on the potential for improved results by fine-tuning channel selection methodologies in EEG-focused deep learning models.

Table 5.2: Pairwise comparisons of mean accuracies across different numbers of channels

Factor	Comparison	Mean	p-value
Number of channels	4 vs 8	(92.67, 94.00)	0.0000*
	4 vs 16	(92.67, 93.91)	0.0003*
	4 vs 32	(92.67, 93.66)	0.0054*
	4 vs all	(92.67, 92.83)	0.9300
	8 vs 16	(94.00, 93.91)	0.1984
	8 vs 32	(94.00, 93.66)	0.0520
	8 vs all	(94.00, 92.83)	0.0045*
	16 vs 32	(93.91, 93.66)	0.4391
	16 vs all	(93.91, 92.83)	0.0215*
	32 vs all	(93.66, 92.83)	0.0716

To address the research question **RQ2.2:** "How does the reduction of EEG channels affect the classification model's performance?", various channel configurations were compared pairwise. The results in Table 5.2 offer several insightful observations. Firstly, there was a clear enhancement in model accuracy as the number of channels increased from 4 to 8, 16, and 32 ($p < 0.05$). Interestingly, when contrasting the 4-channel configuration with a 62-channel setup, model accuracy improved, but the difference was not statistically significant. After the 8-channel mark, consistent performance gains were not observed. Model accuracy tended to decrease as channels increased to 16 and 32, although these declines were not statistically significant ($p > 0.05$). The decrease became significant when the 8-channel configuration was compared with the 62-channel configuration. A similar pattern was noted with the 16-channel setup.

These insights are crucial as they confidently emphasise a fundamental point: simply increasing the number of channels does not always guarantee better performance. While scaling up from 4 to 8, 16, and 32 channels can offer measurable benefits, adding more channels indiscriminately can lead to diminishing, if not adverse, returns. Thus, carefully selecting and optimising channel usage is paramount, striking a perfect balance between performance goals and computational constraints.

Tables 5.3 through 5.6 show a comprehensive picture of how specific models behave under various channel configurations. The models LSTM, BLSTM, BLSTM_LSTM, GRU, BGRU, and BGRU_GRU provide results when tested across different covariance estimators and channel configurations.

Table 5.3: Performance metrics of different models under Empirical Covariance (EC) estimator with varied channel configurations

Scenario	Model	Accuracy	Sensitivity	Precision	F1-Score
EC(4)	LSTM	91.84	87.76	89.92	87.45
	BLSTM	92.97	89.46	90.57	89.31
	BLSTM_LSTM	92.24	88.36	90.04	88.12
	GRU	92.31	88.46	90.21	88.19
	BGRU	93.01	89.52	90.67	89.35
	BGRU_GRU	92.41	88.62	90.24	88.39
EC(8)	LSTM	93.25	89.88	91.79	89.53
	BLSTM	93.96	90.93	92.24	90.72
	BLSTM_LSTM	93.56	90.33	92.01	90.08
	GRU	92.34	88.51	90.79	88.11
	BGRU	93.91	90.86	92.20	90.64
	BGRU_GRU	93.00	89.50	91.45	89.14
EC(16)	LSTM	92.71	89.07	91.84	88.57
	BLSTM	94.40	91.61	93.10	91.38
	BLSTM_LSTM	92.33	88.49	91.40	87.96
	GRU	92.33	88.49	91.40	87.89
	BGRU	94.99	92.48	93.73	92.29
	BGRU_GRU	93.23	89.84	92.24	89.32
EC(32)	LSTM	92.87	89.31	92.07	88.75
	BLSTM	95.15	92.73	93.96	92.49
	BLSTM_LSTM	91.85	87.78	91.30	87.16
	GRU	93.40	90.10	92.52	89.47
	BGRU	94.87	92.31	93.83	91.99
	BGRU_GRU	92.47	88.70	91.61	87.99

In the context of empirical covariance (EC) estimation, Table 5.3 demonstrates that the performance of various models can differ significantly across different channel configurations. Specifically, as the number of channels increases from 4 to 32, some models show improvement, while others either stagnate or experience a slight reduction in performance.

Among the models examined, BLSTM consistently performed well as channels increased, with a peak accuracy of 95.15% achieved with the 32-channel configuration. Likewise, BGRU achieved its highest accuracy of 94.99% under the 16-channel configuration. Notably, BGRU had the highest accuracy of 93.01% for 4-channels, while BLSTM showed the highest accuracy of 93.96% for 8-channels configuration.

Moreover, BGRU and BLSTM demonstrated the most consistent top performance across the various channel configurations, often leading or closely following in accuracy. In contrast, LSTM, BLSTM_LSTM, and GRU exhibited performance fluctuations as channel configurations changed, with their results not always being consistent across all channel configurations.

Overall, the models displayed a good balance between precision and sensitivity, which is crucial for achieving a high F1-Score. This balance indicates that the models are effective at correctly classifying positive instances (high precision) and capturing most of the positive instances (high sensitivity/recall).

It is essential to note that the choice of model and channel configuration can significantly impact performance metrics. Depending on the application's requirements, such as prioritising accuracy or sensitivity, one might opt for a specific model-channel configuration over others. From a broader perspective, BLSTM and BGRU models appear to be the most promising candidates in terms of their performance metrics, especially when considering higher channel configurations.

Table 5.4: Performance metrics of different models under Shrunk Covariance (SC) estimator with varied channel configurations

Scenario	Model	Accuracy	Sensitivity	Precision	F1-Score
SC(4)	LSTM	92.56	88.84	91.00	88.60
	BLSTM	93.83	90.75	91.51	90.68
	BLSTM_LSTM	92.73	89.10	90.84	88.88
	GRU	93.22	89.84	91.38	89.70
	BGRU	93.99	90.98	91.84	90.84
	BGRU_GRU	92.85	89.28	90.76	89.00
SC(8)	LSTM	93.42	90.13	92.46	89.76
	BLSTM	94.84	92.26	93.39	92.12
	BLSTM_LSTM	94.26	91.39	92.80	91.16
	GRU	94.55	91.83	93.07	91.65
	BGRU	95.23	92.85	93.93	92.73
	BGRU_GRU	94.28	91.41	92.85	91.28
SC(16)	LSTM	93.49	90.24	92.54	89.68
	BLSTM	94.71	92.06	93.66	91.69
	BLSTM_LSTM	93.15	89.73	92.38	89.14
	GRU	93.18	89.77	92.28	89.29
	BGRU	94.96	92.44	93.84	92.18
	BGRU_GRU	93.41	90.11	92.48	89.40
SC(32)	LSTM	92.85	89.27	92.04	88.73
	BLSTM	94.53	91.80	93.32	91.54
	BLSTM_LSTM	92.65	88.98	91.91	88.38
	GRU	92.77	89.15	91.90	88.59
	BGRU	95.47	93.20	94.32	92.95
	BGRU_GRU	92.71	88.96	92.05	88.46

The performance of different machine learning models under the Shrunk Covariance (SC) Estimator with varying channel configurations is presented in Table 5.4. As the number of channel configurations increases from 4 to 32 channels, interesting patterns emerge in the performance of the models. While some models show an improvement in their metrics, others exhibit a plateau or minor decline in performance.

Among the models, BLSTM consistently demonstrates strong metrics, with its highest accuracy of 94.84% achieved in the 8-channels configuration. Conversely, BGRU attains its peak accuracy of 95.47% in the 32-channels setup. Notably, BGRU showcases consistent high performance across all channel configurations, often leading the pack or being a close contender. For instance, during the 4-channels setting, BGRU secured the top position with an accuracy of 93.99%. Similarly, in the 8-, 16-, and 32-channels configurations, BGRU continued to exhibit outstanding performance, achieving accuracies of 95.23%, 94.96%, and 95.47%, respectively.

While BLSTM displayed promising metrics, models like LSTM, BLSTM_LSTM, and GRU showed variable performance depending on the channel configuration. However, most models maintained a commendable balance between precision and sensitivity, indicative of high F1-Scores, suggesting that the models accurately classified positive instances and captured most of them.

Considering their robust metrics across various channel configurations, BGRU and BLSTM emerge as the clear front-runners.

Table 5.5: Performance metrics of different models under Ledoit-Wolf (LW) estimator with varied channel configurations

Scenario	Model	Accuracy	Sensitivity	Precision	F1-Score
LW(4)	LSTM	91.15	86.72	89.45	86.29
	BLSTM	93.27	89.90	90.73	89.74
	BLSTM_LSTM	91.90	87.85	89.91	87.52
	GRU	92.40	88.60	90.33	88.22
	BGRU	93.07	89.60	90.79	89.38
	BGRU_GRU	92.48	88.72	90.29	88.53
LW(8)	LSTM	93.21	89.81	92.15	89.37
	BLSTM	94.59	91.88	93.21	91.64
	BLSTM_LSTM	93.94	90.90	92.65	90.57
	GRU	93.97	90.95	92.43	90.76
	BGRU	95.12	92.68	93.58	92.56
	BGRU_GRU	93.76	90.64	92.48	90.34
LW(16)	LSTM	94.23	91.34	93.09	91.00
	BLSTM	95.47	93.21	94.34	93.04
	BLSTM_LSTM	93.49	90.23	92.74	89.67
	GRU	93.28	89.92	92.48	89.35
	BGRU	95.19	92.78	93.94	92.56
	BGRU_GRU	93.77	90.66	92.85	90.25
LW(32)	LSTM	93.28	89.92	92.63	89.34
	BLSTM	95.51	93.26	94.57	93.02
	BLSTM_LSTM	92.55	88.83	91.92	87.93
	GRU	93.60	90.41	92.74	89.92
	BGRU	95.58	93.37	94.56	93.20
	BGRU_GRU	92.96	89.44	92.13	89.01

The performance metrics of various models under the Ledoit-Wolf (LW) Estimator with different channel configurations are presented in Table 5.5. The table shows that the models exhibit a diverse range of performance metrics as the channel configurations transition from 4 to 32 channels. However, most models seem to show an upward trend in their performance as the number of channels increases, although there are exceptions where the performance remains stagnant or slightly decreases.

Two models, namely the BLSTM and BGRU, showcase impressive results across various channel configurations. The BLSTM model consistently exhibits remarkable results with increased channels, achieving a peak accuracy of 95.51% at 32-channels. Similarly, the BGRU model stands out, particularly at 8-channels and 32-channels, with top accuracies of 95.12% and 95.58%, respectively.

Interestingly, the top-performing model in each channel configuration varies for this covariance estimator. For instance, in the 4-channels scenario, the BLSTM model emerges as the top performer with an accuracy of 93.27%. Meanwhile, for the 8-channels configuration, the BGRU model leads the roster with an accuracy of 95.12%. In the 16-channels setup, the BLSTM model takes the lead again, achieving an accuracy of 95.47%. Lastly, under the 32-channels framework, the BGRU model marginally outperforms its counterparts, registering an accuracy of 95.58%.

Remarkably, both the BGRU and BLSTM models demonstrate commendable consistency across various channel configurations, frequently topping the list or being strong contenders. Moreover, they also emerge as standout models in terms of their performance metrics across the different channel configurations.

On the other hand, models like LSTM, BLSTM_LSTM, and GRU display some variability in their performance metrics depending on the channel setup.

Table 5.6: Performance metrics of different models under Oracle Approximating Shrinkage (OAS) estimator with varied channel configurations

Scenario	Model	Accuracy	Sensitivity	Precision	F1-Score
OAS(4)	LSTM	92.27	88.41	90.18	88.24
	BLSTM	**93.08**	89.62	90.69	89.40
	BLSTM_LSTM	92.11	88.16	90.32	87.88
	GRU	92.38	88.56	90.30	88.28
	BGRU	93.13	89.70	90.95	89.55
	BGRU_GRU	92.84	88.27	90.66	89.06
OAS(8)	LSTM	93.85	90.77	92.45	90.52
	BLSTM	**94.92**	92.39	93.49	92.29
	BLSTM_LSTM	92.99	89.49	91.76	89.10
	GRU	94.15	91.22	92.78	91.03
	BGRU	95.26	92.90	93.70	92.82
	BGRU_GRU	93.72	90.59	92.35	90.39
OAS(16)	LSTM	93.72	90.59	92.82	90.16
	BLSTM	**94.75**	92.12	93.74	91.80
	BLSTM_LSTM	93.46	90.19	92.49	89.74
	GRU	93.79	90.68	92.90	90.22
	BGRU	95.51	93.26	94.31	93.05
	BGRU_GRU	94.21	91.31	93.14	90.96
OAS(32)	LSTM	93.40	90.10	92.76	89.47
	BLSTM	**95.08**	92.61	94.10	92.36
	BLSTM_LSTM	91.69	87.54	91.09	86.77
	GRU	93.57	90.35	92.65	89.93
	BGRU	95.77	93.65	94.61	93.52
	BGRU_GRU	93.17	89.76	92.46	89.19

The Oracle Approximating Shrinkage (OAS) method has yielded some interesting results, as presented in Table 5.6. Observations show that the performance of most models improves as the channel configurations evolve from 4 to 32 channels, with some slight variances. One of the models exhibiting a trend of consistent performance is the BLSTM model, especially as the channel configurations increase, reaching its highest accuracy of 95.08% at 32 channels. On the other hand, the BGRU model stands out across all configurations, achieving an impressive accuracy of 95.77% at the 32-channel configuration.

Similarly, a trend analogous to that observed with the SC estimator is noted, where the BGRU model provides the highest accuracy across channel configurations, with

93.13%, 95.26%, 95.51%, and 95.77% accuracy for the 4-, 8-, 16-, and 32-channel configurations, respectively.

Table 5.7: Performance metrics of different models under with 62 channels configuration

Scenario	Model	Accuracy	Sensitivity	Precision	F1-Score
62	LSTM	92.37	88.55	91.59	87.80
	BLSTM	94.77	92.15	93.91	91.82
	BLSTM_LSTM	90.84	86.26	90.21	85.30
	GRU	92.48	88.72	91.92	87.98
	BGRU	**95.80**	93.70	94.74	93.50
	BGRU_GRU	90.70	86.05	90.53	85.17

In order to evaluate the performance of different covariance and channel configurations, comparisons will be made between the top-performing model from each category and the one that utilises data from all EEG channels. The results of this comparison are summarised in Table 5.7.

It is evident that models such as BGRU consistently deliver outstanding performance across several configurations, highlighting their adaptability and robustness. Similarly, the reliability of the BLSTM model is reinforced by its consistent results across diverse configurations. However, the interplay between the model's architecture and the number of channels employed is crucial. Certain models perform optimally with specific channel configurations, revealing a synergistic relationship. The analyses underline the importance of optimisation, emphasising that a more judicious approach to selecting the right channel configuration tailored for the chosen model is indispensable. This approach offers a twofold advantage: first, it prevents the model from getting bogged down by redundant information; second, it can potentially reduce computational overheads. In essence, while the channel number is a pivotal factor, the combination of the right model with an optimal channel configuration ultimately unlocks peak performance. This synthesis of findings provides actionable insights for current EEG classification tasks and lays down a guiding framework for future endeavours in this domain.

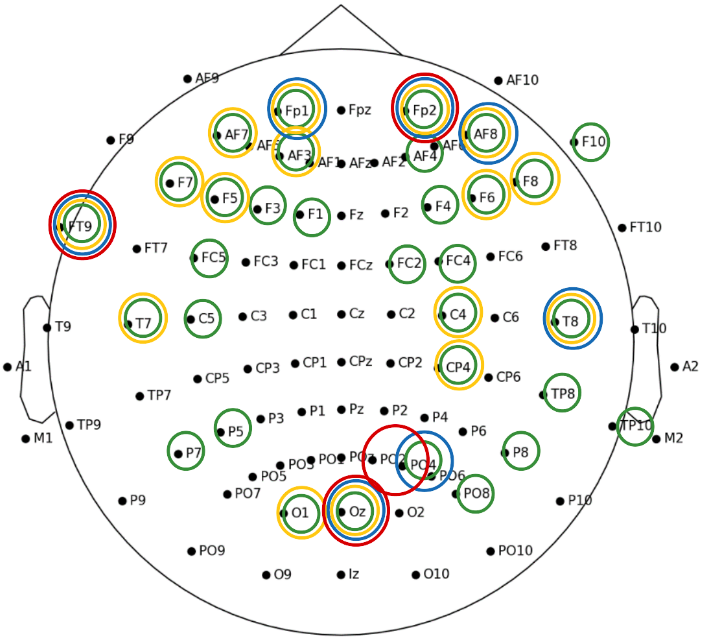


Figure 5.1: Selected EEG channel using Riemannian Geometry with Shrunk covariance estimator

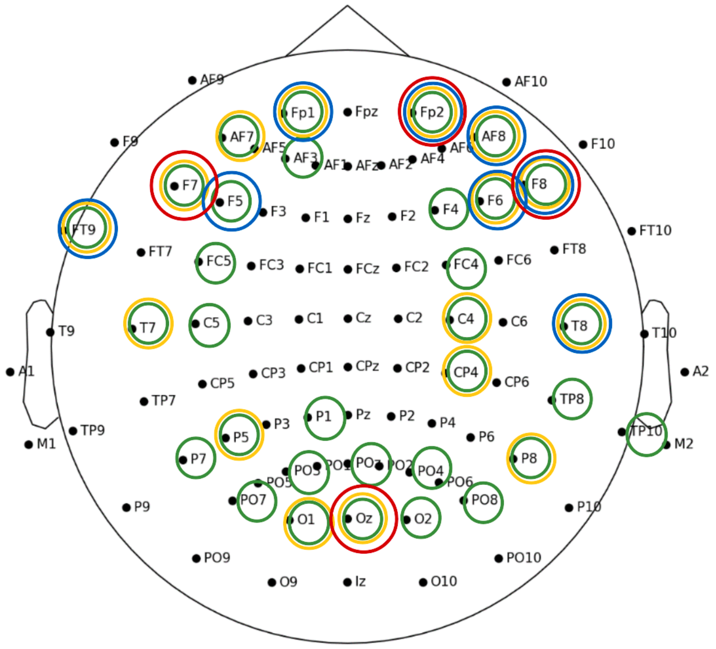


Figure 5.2: Selected EEG channel using Riemannian Geometry with Ledoit-Wolf covariance estimator

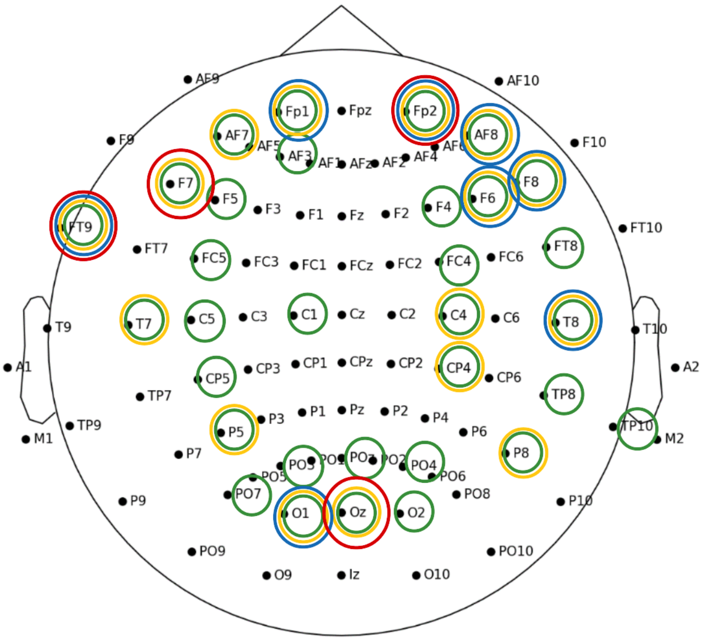


Figure 5.3: Selected EEG channel using Riemannian Geometry with OAS covariance estimator

The selected channels from the Riemannian technique under the SC, LW, and OAS under different numbers of channel configurations are depicted in Figure 5.1, 5.2 and 5.3, respectively. The results illustrated in the figures show the configuration with four channels, represented by red circles, spans across occipital, frontal, and prefrontal brain regions, which are significant for analysing EEG signals about visual perception, attention, and cognitive functions—key aspects of MWL [145]. Meanwhile, the 8-channel configuration, denoted by blue circles, encompasses frontal, central, and temporal regions, making it suitable for a broader range of EEG signals related to MWL. Expanding further to a 16-channel configuration, symbolised by yellow circles, entails capturing EEG signals from the frontal, central, parietal, and occipital areas. Lastly, the 32-channel configuration, indicated by green circles, spans across frontal, central, parietal, occipital, and temporal areas. Each channel combination is autonomously selected by the Riemannian algorithm, ensuring optimal coverage and data acquisition from targeted cerebral regions.

The results discern key similarities in various configurations, revealing that channels

Fp1 and AF8 were uniformly selected in 8, 16, and 32 configurations across the LW, OAS, and SC covariance estimators. Notably, Fp2 was a consistent choice across all configurations (4, 8, 16, and 32) for the LW, OAS, and SC estimators, while F7 was chosen in the 4, 16, and 32 configurations under the LW and OAS covariance estimators. The FT9 channel was versatile, being chosen in all configurations with OAS and SC estimators and in 8, 16, and 32 configurations with the LW estimator. Additionally, the Oz channel demonstrated significant applicability, being a common selection in 4, 16, and 32-channel configurations for LW and OAS estimators and was universally selected in all configurations when utilising the SC estimator. These repeatedly selected channels highlight the importance of specific brain regions, regardless of the configuration size used. Channels AF8, Fp1, Fp2, F7, FT9, and F8, associated with the frontal and prefrontal regions, are important in influencing MWL [145]. In contrast, the Oz channel, which is linked to the occipital region, highlights the importance of visual perception in tasks related to MWL. The findings support existing studies on the role of certain brain regions in mental effort [28], validating the results in exploring brain function and cognition.

5.7 Conclusion

MWL is a cognitive construct that measures the mental effort needed to perform tasks. Assessing MWL is essential for optimising human performance and decision-making, and for designing efficient human-computer interactions. EEG has become popular for estimating MWL due to its high temporal resolution and non-invasiveness. However, current EEG devices are complex and involve many channels, making them unsuitable for practical use. Selecting the optimal number of channels is important, e.g., in BCI applications. This study evaluated different covariance estimators for Riemannian geometry-based channel selection and assessed their effectiveness with deep learning models to classify MWL levels. Four covariance estimators were examined: EC, SC, LW, and OAS. The OAS estimator consistently delivered the best performance across all models, as did the covariance estimation technique. The study demonstrated that using as few as four channels can achieve an accuracy of 0.940 (±0.036), improving practicality for real-world applications. It was also found that the BGRU model, combined with OAS covariance estimators and a 32-channel configuration, outperforms other estimators for MWL classification tasks. The approach supports the development of user-friendly, efficient, and accurate BCI for various purposes, such as cognitive assessment and neurorehabilitation, by reducing the number of channels while retaining high classification accuracy. This has significant implications for enhancing EEG-based BCI in real-world settings.

In the future, it is possible that advanced hybrid techniques may be explored to refine channel selection processes further. These techniques could combine Riemannian geometry with other dimensionality reduction methods to improve the accuracy of classification. Additionally, there is potential to explore more nuanced channel configurations beyond the fixed sets of 4, 8, 16, and 32 channels selected from 62. This could provide deeper insights into the optimal balance between system complexity and classification accuracy, and an adaptive selection technique that adjusts the number of channels based on specific tasks could be implemented. For tasks that require more precise signal capture, this technique could adjust the number of electrodes accordingly. Such studies could help tailor EEG-based BCI systems for use in real-time monitoring and other varied settings.

5.8 Chapter Summary

- This chapter clarifies the second research goal, which addresses concerns about redundant information caused by using too many EEG channels in machine learning studies. The goal is to gather important information from relevant sources and accurately represent the changing patterns of neural activity. Primary concerns include having too much repetitive information and using an excessive number of EEG channels to measure MWL, which is impractical for real-life scenarios.

- The significance of the channel selection step is highlighted, introducing a useful methodology known as the Riemannian geometry technique. The importance and effectiveness of this technique in choosing EEG channels are described, along with the crucial role of the covariance estimator in Riemannian channel selection and how different characteristics of covariance estimators can impact this process.

- Different experimental settings are carefully selected, each with distinct ways of estimating covariance and the number of channels used. These specific settings provide a practical basis for investigating the research questions.

- The chapter also details the methodology used, including the specific setup for selecting channels and providing insights into how the deep learning models operate. The models include Stacked LSTM, BLSTM, a combination of BLSTM and LSTM, Stacked GRU, BGRU, and a combination of BGRU and GRU.

- The research investigates how various factors, such as the method used to estimate covariance and the number of channels, affect the performance of deep learning models. The study examines how these factors relate to interpreting EEG signals and predicting MWL states. Statistical testing is utilized to validate the findings and ensure the accuracy of the results. Insights are provided on how these factors interact and affect classification accuracy, highlighting the complex relationship between them.

- The final parts of the study show that using more EEG channels may not always improve the model's performance. This emphasises that only certain brain regions are relevant to the MWL task, as supported by neuroscientific theories.

Chapter 6

Time Series Cross-Validation

This chapter focuses on evaluating deep learning models trained for MWL level classification using EEG signals. Given the temporal nature of EEG data, the limitations of traditional cross-validation methods used in existing work are discussed. To address this, time series cross-validation methods, specifically the expanding and rolling windows strategies, are implemented to validate the models more effectively.

6.1 Introduction

The concept of MWL plays an important role in human life, influencing areas ranging from study design [168, 256] to driving fatigue [88, 241], pilot performance [123], and performance on various tasks [230]. To measure individuals' MWL, EEG signals have been thoroughly investigated due to their strong correlation with real-time MWL status [204]. Recently, machine learning techniques have garnered significant attention and have been developed to capture the variance characteristics of EEG signals to classify MWL levels [109]. Such classification models are often trained and evaluated using the cross-validation (CV) technique [183].

In a traditional CV technique, an entire dataset is split into K equally sized subsets (also known as folds). The model is trained on $K-1$ folds (called the training set), while the remaining fold is kept apart, unseen by the model, to be used as the test set [196]. The model training process is repeated K times, with a different fold preserved for model evaluation each time. The fundamental idea underlying a CV technique is that it assumes a collection of random variables is being drawn from a given probability distribution; these variables are statistically independent of each other, satisfying the independent and identically distributed (i.i.d.) property in probability theory and statistics [80]. However, EEG signals, which are generated over time, represent time series data. Therefore, applying the traditional CV approach—shuffling and randomly splitting the data into K-folds—can violate the i.i.d. assumption [19] and lead to an unreliable model [35] due to overfitting. This is especially true for a forecasting task where future information should not be available to the model during training.

Since the standard cross-validation (CV) approach is not directly applicable to time series data, researchers have proposed various modifications of CV techniques for this type of data [20, 35]. For example, a blocked form of CV with an expanding window strategy is proposed by Bergmeir and Benítez [19]. This CV strategy is implemented by mimicking a real-life scenario, where the test dataset is sequentially moved into the training dataset, and the forecast origin is changed accordingly. Another strategy that can be used in the blocked form of CV is the rolling window strategy. Bergmeir and Benítez [19] stated that this strategy could be beneficial when the characteristics of the previous observation dynamically change over time and tend to interrupt model generation. However, discarding some parts of the time series might cause the models to lose important information, potentially affecting their performance.

Machine learning models, especially deep learning models, have drawn the attention of researchers in neuroscience, who have used them to classify MWL levels based on EEG signals [109]. CV is a statistical technique for evaluating and comparing machine learning models. It involves training the models on a subset of the available input data and evaluating them on the complementary subset. The application of CV is also task-dependent. For example, in scenarios not aimed at predicting future outcomes, the traditional CV technique, which randomly shuffles data by splitting it into K-folds, could effectively evaluate the model. However, in cases focusing on predicting a future value, such as the subject's MWL level, the temporal aspect of the data must be considered since the time spent on a task typically affects the subjects' performance, which deteriorates over the period of task engagement [126].

6.1.1 Cross-Validation for Deep Learning Model

Various studies aiming to predict a subject's MWL have not considered the temporal aspect of the EEG signal [76]. Additionally, some studies have not provided sufficient details on how they perform CV [33, 247]. For instance, Ahmadi et al. [2] aimed to detect driver fatigue using an expert automatic method based on brain region connectivity. They utilized an EEG dataset that recorded fatigue and alert states. In the model evaluation step, the dataset was randomly divided into five subsets (folds). One of these folds was kept as the validation fold, and the others formed the training set used in the feature selection and hyper-parameter tuning stage. Five-time five-fold cross-validation was applied for each subject. However, the method of randomly defining the dataset is implausible in practical settings. Human fatigue develops over time [228], e.g., a driver's fatigue level at the beginning of a drive may be low, but it increases as time progresses. Therefore, training a fatigue detection model using future fatigue levels to detect previous fatigue levels might not result in an accurate model. Five-fold cross-validation was also adopted by Zhang et al. [242], who utilized a one-dimensional convolutional neural network (1D-CNN) to automatically capture information from different frequency bands and read the subject's mental states. They used an EEG dataset containing 31 recording sessions from five subjects and randomly divided independent recording sessions into five groups for CV. As in previous work, the subjects' MWL continuously increases with time [199]. Therefore, arbitrarily assigning data from various sessions to training and test datasets may be considered inappropriate. Zeng et al. [241] also aimed to identify the mental states of subjects. They proposed a light-weight classifier, LightFD, which is based on a gradient boosting framework. In the model evaluation step, they randomly extracted 80% of the EEG signals of each subject to create a training set, and the remaining 20% of the signals were used as a test set. The EEG time series used for MWL classification were also randomly divided into training and test sets to evaluate the proposed deep learning models in [243] and [247].

As evident from the literature, researchers studying EEG signals using machine learning methods have often overlooked the characteristics of time series and the i.i.d. assumption of the elements of time series while performing cross-validation. They applied traditional strategies by shuffling their EEG signals and then randomly dividing them into training and test sets, failing to account for the temporal information of time series data. To the best of current knowledge, there are only two studies that consider the temporal characteristics of EEG data in the cross-validation step, cited in [170] and [171]. In [170], the authors noted that evaluating the model using test samples that are chronologically near one of the training samples could introduce an overfitting problem due to EEG signal changes. Consequently, the authors of [171] adopted time-wise cross-validation strategies in their study. In this strategy, the samples of each task in each session are divided into n parts evenly and continuously. In each fold, the model is trained using $n - p$ parts of all the tasks and tested on the left-out parts of all the tasks as well. Some parts of the data at the beginning and the end of each task might be cut off to lessen the effect of task transition.

As the temporal aspect of time series has not been considered a significant factor in choosing a CV method, it is important to raise awareness of this potential pitfall in the community. This work aims to demonstrate how time series TSCV can be applied to deep learning models using EEG signals and to investigate the effectiveness of applying TSCV to various deep learning models.

This chapter investigates the effect of time series blocked CV strategies and the size of their training and testing data on deep learning models for MWL classification using EEG signals. Background information on time series and time series CV is described in this section.

Time Series

Time series data consist of a series of data points measured in time order; these data points can be measured every millisecond, every minute, hourly, daily, or annually. The data can be represented by either continuous or discrete datasets. In time series applications, the available past and present data are used to forecast the future values of X. A function $\mathbf{F}$ is calculated to do this. The estimated value $\hat{X}_{t+\tau}$ of X at time $t+\tau$ can be obtained from a function $\mathbf{F}$, where the function $\mathbf{F}$ is computed from a given value of X up to time t (plus additional time-independent variables in multivariate time series analysis):

$$\hat{X}_{t+\tau} = \mathbf{F}(X_t, X_{t-1}, ...) \tag{6.1}$$

where τ is the lag for prediction. Then, the function of the continuous time series will be mapped onto binary values of N classes for a classification approach:

$$\mathbf{F}_c(X_t, X_{t-1}, ...) \rightarrow \hat{c}_i \in C \tag{6.2}$$

$\mathbf{F}_c(X_t, X_{t-1}, ...) \rightarrow \hat{c}_i \in C$ where C is the set of class labels [57].

EEG signals, which are used to display human brain activity, are measured over specific time periods and are considered time series. Each signal from an electrode can be viewed as a univariate time series, with MWL levels serving as class labels. For MWL level classification, EEG signals from various electrodes can be mapped into specific classes corresponding to low, moderate, or high MWL levels using machine learning models. CV is commonly employed to evaluate such models and test their performance. However, a crucial distinction in forecasting is that the future is completely unavailable and must be estimated solely based on past occurrences. Thus, time series data cannot be randomly shuffled as done in traditional CV methods. For instance, in an autoregressive (AR(p)) model, the parameter p, known as the order, indicates the maximum separation that can exist between events that are related to each other in the AR process.

The model can be written as $\hat{X}_t = (c + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + \varepsilon)$, where $\beta_1, \beta_2, \ldots, \beta_p$ are the parameters of the model, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ are the past time series values, ε represents white noise, and c is a constant. It is clear that the traditional CV method could cause an overfitting problem [19] because it allows for potential leakage of future information into the training dataset. A method for performing CV on a time series dataset is explained in Section 6.3.7.

6.2 Research Questions

In this study, we aim to answer two research questions:

- **RQ3.1:** Which time series blocked CV strategies should be applied for the EEG classification task?
- **RQ3.2:** How does the size of the block in a TSCV influence the effectiveness of the models?

6.3 Experimental Set-Up

6.3.1 Datasets: STEW

The dataset used for this study is the STEW dataset, which can be referred to in **Section 3.1.1**.

6.3.2 Data Preprocessing

EEG signals can easily become contaminated by unwanted artefacts; therefore, removing artefacts from EEG signals is usually a prerequisite for signal analysis [209]. In this study, a noise removal technique outlined in **Chapter 4** was utilised. The technique involves high-pass filtering, independent component analysis based on ADJUST (ICA-ADJUST) [147], and re-referencing, and has been proven effective for the STEW dataset [109], showing significant improvement in model performance. The preprocessed data obtained from this technique was used for the analysis.

6.3.3 EEG Feature Extraction and Selection

In this section, the feature extraction and selection strategies outlined in **Chapter 4** are followed. All features underwent standardisation analysis, which is explained in the subsequent section.

6.3.4 Feature Standardisation

EEG signals are inherently variable within and between subjects due to time-variant factors, psychological parameters, and neurophysiological parameters [219]. This variability can lead to data distribution shifts [179], which can degrade the generalisability of extracted features. To mitigate this problem, a personalised feature standardisation method was applied [30,219]. This method converts the extracted features so that they all have the same scale across subjects (Equation 3.15), as detailed in **Section 3.2.3**.

6.3.5 Deep Learning Model

The TSCV strategies were evaluated on two distinct classification tasks: Task 1 and Task 2. A variety of deep learning models were employed, including Stacked LSTMs, BLSTMs, Stacked GRUs, BGRUs, BGRU-GRUs, BLSTM-LSTMs, and CNNs. The details of these model architectures can be found in Table 3.1.

6.3.6 Evaluation Metrics

The impact of preprocessing techniques on deep learning model performance was evaluated using six metrics: sensitivity, specificity, precision, accuracy, FAR, and FRR. For in-depth explanations of these metrics, refer to **Section 3.4**.

6.3.7 Classification and Model Training

From the literature, it is observed that the temporal character of the EEG signal is often not considered a critical criterion for choosing the cross-validation technique. For this dataset, the goal is to predict a subject's MWL in the next time step, treating the EEG signals measured over a period of time as time series data. Consequently, the deep learning models were trained using a 5-fold TSCV strategy [110]. Initially, 80% of the beginning part (i.e., time = 1, 2, 3, ..., j) of EEG signals from each subject was used to conduct model training, and 20% of the data (i.e., time = j+1, j+2, j+3, ..., T) was kept aside as an unseen test dataset. How the dataset was split for TSCV is illustrated in Figure 6.1 and 6.2. Subsequently, the training dataset was further divided into five folds of approximately equal size, constituting a 5-fold cross-validation. Once the model was trained using 5-fold, the classification model performance was evaluated by comparing the predicted levels with the true labels of the unseen dataset. More details about how each strategy works will be described in the next section.

Time Series Cross-Validation (TSCV)

Traditionally, using the entire time series dataset to evaluate machine learning or deep learning models leads to a theoretical statistical violation [19]. Therefore, in the evaluation procedure for time series models, a final part of the time series should be reserved for testing, ensuring that the corresponding training set consists only of observations occurring before those in the test set. This approach prevents information leakage from using future observations [19]. Tashman [201] has suggested four primary strategies for time series forecasting: fixed-origin evaluation, rolling-origin-recalibration evaluation (or expanding window), rolling-origin-update evaluation, and rolling-window evaluation.

In this study, the data were split by ensuring that the test set consists of data recorded after the data in the training set; additionally, expanding window and rolling window strategies were employed in model evaluation. In the expanding window strategy, the test dataset is sequentially moved into the training set, and the forecast origin is changed accordingly. The classification model is recalibrated until all available samples are used. Thus, the test data of the previous fold is used as a validation set to tune the parameters of the deep learning model. Figure 6.1 illustrates how data are split into training, validation, and test sets using the expanding window strategy.

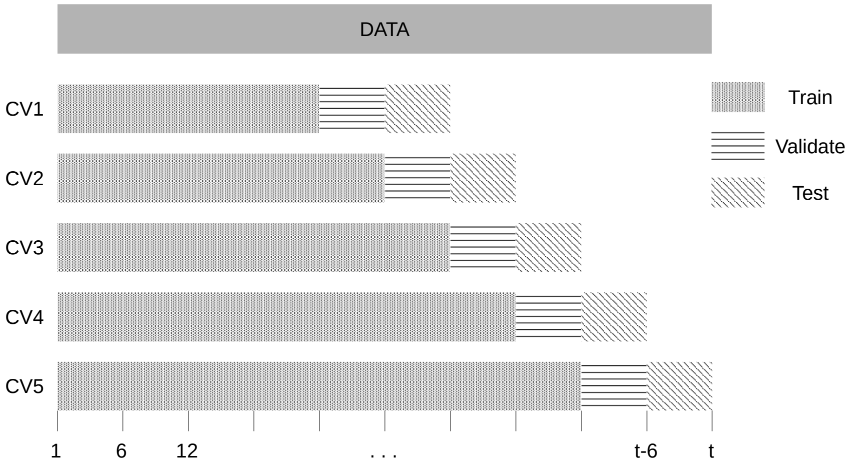


Figure 6.1: Expanding window strategy

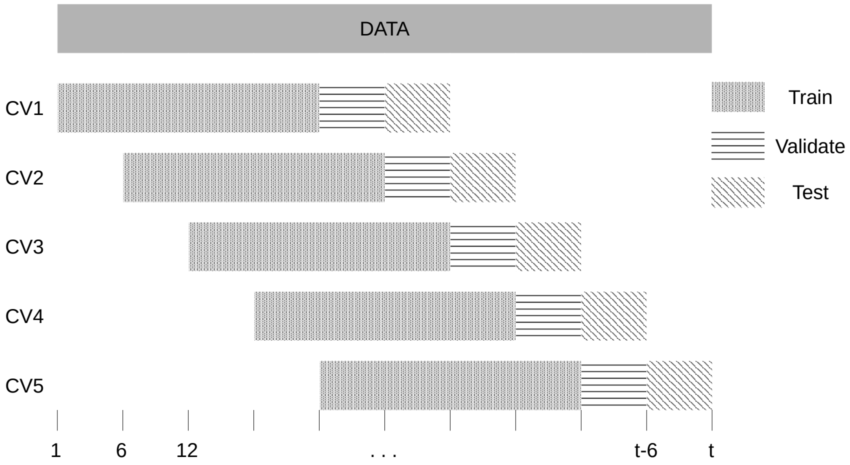


Figure 6.2: Rolling window strategy

Since the subject MWL level might change over time, various scenarios are set in the analysis. Firstly, the rolling window strategy is initiated with the training size starting at 20% of the dataset, increasing to 40%, and then to 90%. As the training dataset varies, this can affect the sizes of the validation and testing windows.

In the rolling window strategy, the amount of data used for training is kept constant. The window is implemented by shifting the training and test data forward by a constant window size of m seconds at each fold. As a result, new data enters the series, while old data from the beginning of the series is discarded. This method is suitable for time series data, as the model is rebuilt in every window, ensuring that the temporal fluctuation of the data does not disrupt model creation [201]. Figure 6.2 illustrates how data are split into training, validation, and test sets using the rolling window strategy. In this study, a 5-fold time series cross-validation with both expanding and rolling window approaches was applied. Two parameters must be optimised for time series cross-validation, namely the sizes of the training and testing windows.

The size of the training window for each time series TSCV strategy is investigated by varying it from 20% to 90% of the data with incremental steps of 10%. Meanwhile, the size of the validation and testing window, set at $(10\% \times n)/k$, is fixed at the end of the series, as this window size has been shown to provide good model performance for this dataset [110]. To generalise the model, data split from all subjects are fed into the models simultaneously.

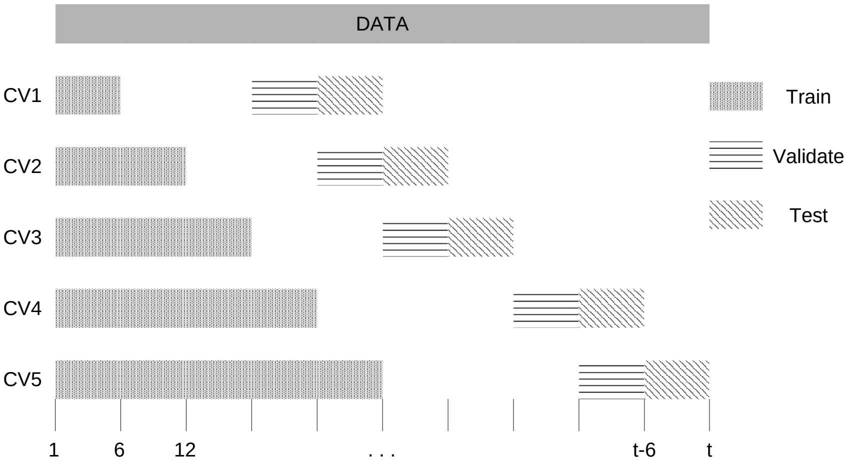


Figure 6.3: Expanding window strategy with a validation and test window size of 6 fixed at the end of series

6.4 Result and Discussion

6.4.1 Type of Time Series: Expanding vs Rolling and Training Size

To evaluate the classification models on EEG data and perform a comprehensive study of their effectiveness with several state-of-the-art deep learning models, a modification

Table 6.1: The average accuracy of deep learning models evaluated using different TSCV strategies and trained using different amounts of data for Task 1 and Task 2

		Task 1		Task 2	
TSCV strategy	Training Size	Accuracy	S.D	Accuracy	S.D
Expanding	20%	84.07	2.65	71.92	1.90
	40%	89.07	2.38	76.96	1.60
	60%	91.01	2.50	78.86	1.72
	80%	92.87	2.14	80.72	1.99
	90%	**93.53**	**1.80**	**81.38**	**2.19**
Rolling	20%	78.39	4.67	67.04	4.73
	40%	85.61	4.36	74.27	4.73
	60%	87.58	3.92	76.23	4.59
	80%	90.51	2.27	79.17	3.14
	90%	**91.78**	**1.55**	**80.44**	**1.90**

of the CV technique known as blocked CV with expanding window and rolling window strategies was adopted. The effect of different block sizes on each strategy was also investigated.

Task 1: Resting vs Testing State

The results shown in Table 6.1 indicate that models evaluated using the expanding window strategy achieved higher accuracy than those evaluated using the rolling window strategy for different training block sizes. For the expanding window strategy, model accuracy steadily improves as the training block size increases. The highest accuracy score of 93.53% was obtained when 90% of the data were used for model training, while the lowest accuracy score of 84.07% was recorded when 20% of the data were used. Similarly, for the rolling window strategy, a pattern of increasing accuracy scores with larger block sizes is observed; however, the accuracy scores in this scenario are slightly lower than those achieved with the expanding window strategy. The highest accuracy score recorded is 91.78%, with the lowest being 78.39%. It was also observed that both the expanding window and rolling window strategies demonstrate a similar pattern in the standard deviation (S.D.) of model accuracy, where the S.D. decreases as the block size of the training data increases. Additionally, models trained using the expanding window strategy consistently exhibit a lower S.D. than those trained using the rolling window strategy.

Task 2: Low vs Moderate vs High MWL Level

In this task, a pattern similar to that observed in Task 1 is noted, with a few exceptions related to the expanding window strategy. In this strategy, the highest model accuracy was still achieved when models were trained using 90% of the EEG dataset. However, the pattern of the S.D. score differs. It is noted that for models trained using 60% of the data, the S.D. score slightly increases compared to Task 1. For the rolling window strategy, the trend in the S.D. is consistent with that observed in Task 1.

In the analysis, the Kruskal-Wallis H test, a non-parametric test for the analysis of variance on ranks, was used to probe the impact of training size on the accuracy of models designed for MWL level classification [140]. Kruskal-Wallis tests were performed to compare the model accuracy of five groups for the training percentages of 20%, 40%, 60%, 80%, and 90%, for each TSCV strategy (expanding and rolling windows) in Task 1 and Task 2. In all scenarios, the Kruskal-Wallis test indicated statistically significant differences between the groups. To delve deeper into these differences, pairwise comparisons were conducted using the Mann-Whitney U test (also known as the Wilcoxon rank-sum test) to identify specific discrepancies in model performance associated with each percentage level.

Through the pairwise comparison analysis, it was found that there were no significant performance differences among the 40% to 60%, 60% to 80%, and 80% to 90% training data ranges in both Task 1_Expanding and Task 2_Expanding. Similarly, in Task 1_Rolling and Task 2_Rolling, non-significant pairwise differences were observed between 40% and 60%, 40% and 80%, 60% and 80%, and 80% and 90% of the training data. Therefore, it can be concluded that although some statistically significant differences in model performances occur as the percentage of training data changes, certain ranges (e.g., 40%-60%, 60%-80%) often do not show significant differences in performance across tasks. This finding has important implications for determining the optimal amount of training data to use in future projects.

Consequently, in response to RQ3.1 regarding the optimal TSCV strategy for EEG classification, it was discovered that the expanding window strategy, with a model trained on 90% of the data, performed the best. For RQ3.2, which investigates how block size in TSCV affects model performance, it was found that adding more training data generally improved model performance, although there were instances when the improvements were not significant. Essentially, while model performance consistently increased with the inclusion of more training data, the expanding window strategy demonstrated a reduced standard deviation compared to the rolling window approach. The fluctuations in model performance as the training dataset changed for the rolling window strategy indicate that the model was influenced by the data in some parts of the time series. Specifically, models trained using the rolling window strategy were more sensitive to variance in the data than those trained with the expanding window strategy. Therefore, to mimic a real-life scenario and obtain models with high accuracy and robustness to the EEG signal, which changes over time, a blocked form of CV with the expanding window strategy should be used in the model evaluation step for the classification of subjects' MWL levels.

6.4.2 Model Comparison

In the previous subsection, it was observed that the model trained using 90% of the data with the expanding window strategy exhibited the best performance for both Task 1 and Task 2. Therefore, the effectiveness of using a blocked form of CV with the expanding window strategy and performing model training using 90% of the data was investigated for several state-of-the-art deep learning models.

Task 1: Resting vs Testing State

As evidenced by the data in Table 6.2, the top three models in terms of model accuracy are the BGRU-GRU, BLSTM-LSTM, and Stacked GRU models, which achieved accuracies of 95.90%, 95.12%, and 94.79%, respectively. The lowest model accuracy of 91.53% was recorded by the CNN model. This may suggest that GRU-based models perform better than LSTM-based models. Additionally, the BGRU-GRU model also recorded the highest specificity score of 97.11% for Task 1, indicating that it was more effective

Table 6.2: The deep learning model performace, TSCV with the expanding window strategy, and training using of 90% of the data on MWL classification for Task 1 and Task 2

Task	Model	Sensitivity	Specificity	Precision	Accuracy	FAR	FRR
1 : resting vs testing state	Stacked LSTM	90.28	97.08	96.87	93.68	2.92	9.72
	BLSTM	90.60	93.01	91.31	91.81	6.99	9.40
	Stacked GRU	93.33	96.25	96.14	94.79	3.75	6.67
	BGRU	90.67	93.08	92.02	91.88	6.92	9.33
	BLSTM-LSTM	96.30	94.04	95.12	95.12	5.96	3.70
	BGRU-GRU	94.70	97.11	96.19	**95.90**	2.89	5.30
	CNN	89.03	94.03	93.71	91.53	5.97	10.97
2 : low vs moderate vs high MWL level	Stacked LSTM	71.17	91.17	81.31	82.34	8.83	28.83
	BLSTM	70.23	90.23	79.44	80.46	9.77	29.77
	Stacked GRU	71.72	91.72	82.42	83.45	8.28	28.28
	BGRU	70.27	90.27	79.51	80.53	9.73	29.73
	BLSTM-LSTM	66.30	84.04	67.50	78.12	15.96	33.70
	BGRU-GRU	72.28	92.28	83.53	**84.56**	7.72	27.72
	CNN	70.09	90.09	79.16	80.19	9.91	29.91

at identifying non-resting states compared to other models. However, when examining the sensitivity scores, a different pattern emerged; the sensitivity of the BGRU-GRU model was slightly lower than that of the BLSTM-LSTM model, indicating that the BGRU-GRU model was less effective at classifying the true level of the subject's MWL compared to the BLSTM-LSTM model. The lowest sensitivity score was also recorded by the CNN model. Despite this, it was observed that GRU-based models slightly outperformed LSTM-based models in terms of accuracy, with the Stacked GRU model performing better than the Stacked LSTM model, the BGRU model outperforming the BLSTM model, and the BGRU-GRU model surpassing the BLSTM-LSTM model in accuracy.

Task 2: Low vs Moderate vs High MWL Level

According to the model accuracies shown in Table 6.2, the BGRU-GRU model remains the best-performing model, followed by the Stacked GRU model, and then the Stacked LSTM model. The accuracies for the BGRU-GRU, GRU, and Stacked LSTM models are 84.56%, 83.45%, and 82.34%, respectively, indicating that GRU-based models continue to outperform LSTM-based models for this task.

Furthermore, the BGRU-GRU model achieves the highest sensitivity and specificity scores of 72.28% and 92.28%, respectively. This suggests that the BGRU-GRU model is effective at correctly classifying the true level of the subjects' MWL. It also performs well in identifying incorrect MWL levels; for instance, when subjects were not at a low MWL level, the BGRU-GRU model more accurately indicated a medium or higher level compared to other models.

Conversely, the least effective model proved to be the BLSTM-LSTM model, which yielded the lowest accuracy, specificity, and sensitivity scores of 78.12%, 84.04%, and 66.30%, respectively. This suggests that the BLSTM-LSTM model was less capable of accurately classifying subjects' MWL levels for this dataset. Overall, GRU-based models demonstrate superior performance compared to LSTM-based models in this task.

In summary, the results indicate that the BGRU-GRU model provided the best performance for both tasks, and GRU-based models outperformed LSTM-based models. The primary distinction between GRU and LSTM models lies in their complexity. GRUs are less complex than LSTMs as they incorporate only two gates—reset and update—compared to the three gates—input, output, and forget—found in LSTMs. Consequently, the less sophisticated GRU-based models have demonstrated the capability to capture relevant information effectively, making them preferable for this dataset.

6.5 Conclusion

MWL is a crucial concept in understanding human performance. High or low MWL can negatively impact performance, leading to stress, mood disorders, and illness. EEG is used to determine MWL levels. Deep learning models are employed to classify MWL accurately using EEG signals, with the CV technique being a common method for training and testing. However, the CV technique used for such models does not take into account the time series nature of EEG signals. Therefore, this chapter explores a modification of the CV technique that can be applied to EEG data, i.e., a blocked form of CV with expanding window and rolling window strategies. Additionally, the effect of the sizes of the training, validation, and testing window for each TSCV strategy was investigated. The TSCV strategies and the window sizes were then verified using the following deep learning models: the stacked LSTM, BLSTM, stacked GRU, BGRU, BGRU-GRU, BLSTM-LSTM, and CNN models. A publicly available MWL dataset called STEW [128] was used to carry out the experiment.

The findings show that the deep learning models evaluated using TSCV with the expanding window strategy provide better classification performance than those evaluated using TSCV with a rolling window strategy. Moreover, for both classification tasks, every deep learning model achieved the highest accuracy when trained using 90% of the data, with a window size of 3 for validation and testing.

Moreover, the results also show that the BGRU-GRU model achieved the highest accuracies of 95.90% and 84.56% for Task 1 and Task 2, respectively. Hence, in future work, it would be advisable to consider applying the BGRU-GRU model and TSCV with the expanding window strategy on other MWL datasets, which are more complex. For example, some datasets were collected from different sessions/days, which usually causes a non-stationary or co-variate shift problem. Finally, this proposed CV method can also be applied to other types of time series data, such as stock prices, annual retail sales, or monthly subscribers, since all of these data are sequential data that were measured at successive times and have a natural temporal ordering, just like EEG signals.

6.6 Chapter Summary

- In this chapter, the ultimate objective of the research is explained. This goal focuses on using the appropriate cross-validation technique to assess MWL classification using EEG signals. The aim underscores the importance of effectively training models for time series datasets, an endeavour where traditional cross-validation techniques — which shuffle all data together before partitioning into folds — prove unsuitable. Such a method violates statistical assumptions and disrupts the inherent temporal nature of EEG signals. Therefore, it is essential to use time series cross-validation, which accurately replicates real-world situations by not using future data for past predictions.

- Time series cross-validation is emphasised, and two helpful strategies, expanding and rolling windows, are introduced. The background of the time series is also described.

- Different experimental settings are carefully established, each utilising various strategies for time series cross-validation and window sizes. These specific settings provide a practical foundation for exploring the research questions.

- A comprehensive overview of the utilised methodology is provided, elucidating the specific process for time series cross-validation and imparting an understanding of the functionality of the deep learning models. The models explored include Stacked LSTM, BLSTM, a hybrid of BLSTM and LSTM, Stacked GRU, BGRU, and a combination of BGRU and GRU.

- This chapter examines how various factors, including the method used for time series cross-validation and the window size, impact the effectiveness of deep learning models. The study investigates the connection between these factors and the interpretation of EEG signals and classification of MWL levels. Statistical testing is employed to confirm the findings and ensure the precision of the results.

- The study presents empirical evidence suggesting that utilising an extended time series dataset can enhance the model's accuracy in classifying MWL. The rationality behind this result is based on the assumption that obtaining a larger historical dataset enables the model to improve its learning effectiveness.

- However, it is crucial to recognize a limitation in this study, specifically the time constraint of the data, which only covers a duration of 2.5 minutes. The temporal limitation inevitably hinders the ability to predict the cognitive effort of individuals in longer-duration circumstances. Nevertheless, it is a noteworthy starting point, especially in circumstances requiring rapid decision-making.

Chapter 7

Conclusions and Further Work

This chapter begins with a discussion and a recap of key achievements in using deep learning for MWL classification and EEG-based information retrieval. It then addresses the constraints of these breakthroughs. The future work section explores the potential application of these contributions to a practical online system and possible research directions in MWL classification.

7.1 Discussion

The experiments in this thesis systematically investigated various aspects of EEG-based MWL classification using deep learning models, emphasising the critical roles of preprocessing techniques, channel selection strategies, and TSCV approaches. This comprehensive examination factors influencing model performance, paving the way for future exploration in this area.

In addressing the challenges posed by EEG signals due to artifacts such as eye movements, muscle activities, and external interferences, researchers have proposed numerous preprocessing techniques, some manual [29] and others automated [147]. Previous studies [1, 23, 49] have identified a significant challenge in accurately assessing EEG signal information due to the presence of artifacts. These artifacts can distort EEG signals, leading to unreliable MWL classifications. The problem with manual preprocessing is that it is time-consuming and also susceptible to bias [210].

Therefore, in this thesis, particularly in **Chapter 4**, a focused examination of automated EEG preprocessing techniques was conducted. The examination of EEG preprocessing techniques, including high-pass filtering, the ADJUST algorithm, and re-referencing, was central to this study. These methods are crucial for integrating EEG preprocessing into automated deep learning frameworks, significantly impacting the classification of MWL [29, 210]. Building upon the foundations laid by Chakladar et al. (2020) [36], which demonstrated the effectiveness of deep learning models such as CNNs, RNNs, and BLSTM-LSTMs in extracting meaningful information from EEG signals, this research further explores the preprocessing phase [116,123]. Unlike manual methods, these automated techniques offer a more objective and scalable solution.

The findings from this experiment underscore the effectiveness of these techniques in enhancing the performance of the aforementioned deep learning models, corroborating the notion that sophisticated preprocessing can significantly mitigate noise and improve model accuracy. This is aligned with recent research advocating for the integration of refined preprocessing steps to bolster the accuracy of EEG-based classifications [116,123]. However, this study uniquely contributes by demonstrating the comparative effectiveness of automated techniques over manual methods, addressing a gap noted in previous guidelines which lacked specificity and universal application [1, 23, 49]. Future studies may continue to build on these findings, exploring further the optimal combinations of these techniques and their applicability in diverse real-world scenarios.

The enhanced clarity achieved through automated preprocessing provides a robust foundation for subsequent stages of EEG analysis, such as channel selection. Effective preprocessing is indispensable as it directly influences the accuracy of channel optimization techniques.

In **Chapter 5**, on channel selection, the study tackled the optimization of EEG channel selection using Riemannian geometry, a method increasingly recognized for its effectiveness in identifying the most informative EEG channels by leveraging the properties of covariance matrices. This approach builds on the work of Barachant et al. (2011) [11] and Qu et al. (2022) [169], who demonstrated the utility of Riemannian geometry in reducing the dimensionality of EEG data while maintaining necessary information for accurate classification in binary tasks. This thesis extends previous applications of Riemannian geometry from binary to multiclass classification tasks, addressing a gap in existing studies which have primarily focused on simpler task structures. By exploring this approach in the context of distinguishing between easy, medium, and difficult levels of MWL, this experiment not only tests the method's robustness but also its adaptability to more complex analytical challenges. The use of different covariance estimators like EC and LW, as shown in previous works [176], was critically evaluated to determine their efficacy in enhancing model performance through optimal channel selection. This study provides crucial insights into how these estimators can influence the effectiveness of channel selection and, consequently, the overall accuracy of the classification models for BCI hackathon dataset. However, the effectiveness of different covariance estimators can vary significantly depending on the characteristics of the EEG data and the specific MWL tasks being analyzed. Moreover, despite the advantages of Riemannian geometry, its computational complexity and the need for specific expertise in its application might limit its widespread adoption.

Given the promising results, future research could further investigate the applicability of Riemannian geometry in other areas of EEG analysis, such as affective computing or neurofeedback systems. Additionally, studies might explore the integration of these channel selection techniques with other forms of neural data processing, potentially creating more robust and versatile models for real-world applications.

Chapter 6, focuses on refining the CV technique for deep learning models analyzing EEG data, a crucial component in assessing MWL. Traditional CV methods, as noted in studies by Schaffer (1993) [183] and Stone (1974) [196], often mishandle EEG data by ignoring its time series nature, leading to potential overfitting and misestimation of model generalizability. This thesis draws from the modifications suggested by Bergmeir and Benítez (2012) [19], who advocated for time series-specific CV strategies to maintain the temporal integrity of the data. Unlike traditional CV, which randomly partitions data and can inadvertently mix training and testing data from different temporal contexts, time series CV ensures that the model is only ever trained on past data, mirroring real-world learning and prediction scenarios. This study not only addresses the gap in applying appropriate CV techniques to EEG data for MWL classification but also demonstrates the practical implications of selecting between expanding and rolling window strategies. By meticulously comparing these strategies, the experiment highlights how each strategy impacts the model's ability to generalize across time, with the expanding window strategy typically providing more robust and realistic assessments of model performance. This approach is crucial for tasks where understanding temporal dynamics—such as fatigue progression in pilots or drivers—is vital for the model's application. Therefore, it significantly contributes to the methodology of evaluating deep learning models for EEG data by providing a clearer framework for handling time series data in machine learning, ensuring that future studies can achieve more accurate and reliable results when predicting MWL. This enhancement is particularly relevant for fields where real-time monitoring and prediction of cognitive load can inform safety and performance, such as in aviation and automotive contexts.

7.2 Contributions & Conclusions

This thesis makes distinct contributions to the domain of EEG-based mental workload classification by addressing key challenges and presenting innovative solutions across three primary areas.

7.2.1 Impact of EEG Preprocessing Techniques

The first research goal was centered on the inconsistencies in EEG data preparation procedures and their impact on the effectiveness and precision of deep learning models in reading EEG data. The goal was to investigate and demonstrate the critical relationship between various preprocessing strategies and the accuracy of MWL state predictions using deep learning models. The researchers emphasized the importance of the preprocessing stage in machine learning domains. The study highlights that raw EEG data is noisy, hindering model training and classification accuracy. While some proponents argue that the CNN model can detect features in raw data, larger datasets are required for model training to capture delicate and nuanced information. Moreover, sophisticated models are needed to comprehend these complex features, which may not be suitable for situations requiring rapid decision-making, such as MWL detection in dynamic systems.

Therefore, this study contributes to the broader understanding of EEG preprocessing by focusing on classic preprocessing methods that remove noise and employ procedures that can be executed automatically without human intervention, such as visual inspection. Experimental scenarios are created to test preprocessing techniques across different datasets. This helps in understanding the impact of various preprocessing methods such as filtering, ADJUST algorithm application, and re-referencing strategies. The focus is particularly relevant in the context of integrating these techniques with deep learning models (Stacked LSTM, BLSTM, and BLSTM-LSTM).

The key finding is that the ADJUST algorithm significantly impacts the performance of the investigated deep learning models compared to other preprocessing techniques. Moreover, when combining all preprocessing techniques for optimal performance, the results indicate that using a combination yielded the highest classification performance across the models. This finding highlights the benefit of using multiple preprocessing techniques to improve deep learning model performance. Therefore, the study proved that models trained on preprocessed EEG signals significantly improve classification accuracy. Moreover, it was also found that raw EEG signals, without preprocessing, were still sufficient for MWL-level classification, particularly in the BLSTM-LSTM model. This finding reveals that more sophisticated models have the potential to extract relevant information from the raw signals.

7.2.2 Optimising EEG Channel Selection for Mental Workload Classification

The detection of MWL can be hindered by excessively large datasets, particularly in situations that require swift decision-making. This often leads to a slowdown in the process. To address this issue, the study aimed to selectively include only the most pertinent data related to MWL in the analysis. To ensure the feasibility of future MWL systems, Riemannian geometry was utilised to perform channel selection, ensuring practicality in future MWL systems.

In this study, the practical limitation of MWL measurement using too many EEG channels is highlighted, a key concern in real-world applications. Moreover, the introduction and implementation of the Riemannian geometry technique for EEG channel selection are discussed, as well as an effective method for identifying relevant channels for EEG MWL classification. A novel insight into the impact of covariance estimators in the process of EEG channel selection using Riemannian geometry is also provided. This aspect contributes to a better understanding of how different characteristics of covariance estimators affect EEG data analysis.

The study also presents the use of various deep learning models (such as Stacked LSTM, BLSTM, GRU, and their combinations) in EEG MWL-level classification. It contributes empirically to understanding how the covariance estimator and the number of EEG channels influence the performance of deep learning models. This knowledge is important for optimising EEG channel selection, which can affect the accuracy of MWL-level prediction. Finally, to validate the results, reliable research findings were utilised in this study.

The research reveals that excess EEG channels can lead to impracticality in real-life scenarios due to the redundant information they provide, which has the potential to cause overfitting problems. Interestingly, it was observed that when data from more electrodes are added to the model, the model performance drops in some scenarios. Therefore, a pivotal finding is that using more EEG channels does not necessarily enhance the model's performance. The findings also indicate that data obtained from the frontal and prefrontal lobes strongly correlates with an individual's MWL. This observation aligns with previous studies indicating that an increase in EEG channels can introduce noise and redundancy, detracting from model accuracy due to overfitting [10]. Thereby validating the existing research in neuroscience. Consequently, it is suggested to target these specific brain regions and channels relevant to MWL level classification [145].

7.2.3 Evaluating Time Series Cross-Validation Strategies in Deep Learning Models for EEG-Based MWL Classification

Cross-validation is an indispensable stage in training deep learning models, as it effectively mitigates the issue of model overfitting. The conventional approach involves partitioning data into k subsets, randomising their order, and assigning them to training and testing sets. However, within the framework of time series data analysis, especially in the domain of MWL classification, utilising future data to predict the present MWL state is not feasible.

Therefore, our primary contribution is to explore a modified cross-validation technique suitable for the EEG signal, which is time series data. It was found that there are TSCV strategies, including expanding window and rolling window approaches, in the field of time series analysis. The expanding the window strategy involves gradually increasing the window size as the analysis progresses. In the first step, the window starts at the minimum size and incrementally includes more data, expanding until it encompasses the entire training dataset. In the rolling window strategy, the window size for cross-validation is fixed. It involves moving windows along the signal. A wider range of deep learning models (Stacked LSTM, BLSTM, Stacked GRU, BGRU, BGRU-GRU, BLSTM-LSTM, CNN) were also utilised for MWL prediction using the TSCV approach. This contributes significantly to the fields of deep learning and time series analysis.

It was also found that deep learning models evaluated using TSCV with an expanding window strategy significantly outperformed those using a rolling window strategy. Specifically, models trained with 90% of the data, as demonstrated by the BGRU-GRU model, achieved the highest accuracies of 95.90% in Task 1 and 84.56% in Task 2. This suggests that a larger historical dataset enhances model performance, providing a guideline for choosing TSCV strategies in similar contexts. For the STEW dataset used in this study, the results underline the benefits of leveraging extensive historical data for training. In light of this, the introduction of time series cross-validation has been a vital development for practitioners in the BCI domain who seek to utilise machine learning or deep learning models to classify subject MWL levels. The successful implementation of this technique ensures that the temporal integrity of the data is maintained, which is crucial for achieving accurate and reliable predictions in time-sensitive environments.

In conclusion, MWL measurement utilising EEG signals in brain-computer interaction gained popularity and presented incredible challenges. Deep learning demonstrated promising results in mental effort forecasting; however, its application in MWL categorization varied across studies. This thesis addressed the crucial task of accurately classifying MWL levels. The proposed approach offered a comprehensive method for utilising EEG for effective MWL classification, focusing on each process stage, from preprocessing to model evaluation.

1. In the preprocessing stage, exploration of automatic EEG artifact removal techniques and their impact on deep learning models was conducted. The findings suggested that the ADJUST algorithm had the most significant impact on model performance compared with others, and the more sophisticated models could capture the relevant information from raw data, potentially reducing the need for extensive preprocessing.

2. Channel selection was another focus; the aim was to reduce redundant information and avoid using the cumbersome EEG cap to pave the way for automation of MWL level classification in practical applications. Using Riemannian geometry, the process was successfully performed by focusing only on electrodes that capture MWL-related brain activity, balancing computational efficiency and information sufficiency. However, there is no one-size-fits-all optimal number of channels for all datasets, but for the BCI Hackathon dataset, an optimal range might start at around 8 to 16 channels for simpler setups and can be extended to 32 channels for more detailed analyses.

3. Existing works that employed a machine learning approach to perform MWL level classification critically neglected the temporal character of EEG signals in the model evaluation step. These studies typically employed the traditional CV technique, which could lead to data leakage and model overfitting issues. To help resolve these issues, the significant importance of TSCV was emphasised, and

two TSCV strategies were adopted in this work: expanding and rolling. Through analysis of STEW data, it was found that the expanding window technique outperformed the rolling window strategy.

In summary, this thesis contributes a refined, efficient, and automated approach to classifying MWL using EEG signals, paving the way for improved performance and safety in critical environments. Further research and development could enhance their suitability for real-world applications, advancing the development of more effective BCI and related applications.

7.2.4 Limitations

Although the study acknowledges compelling insights, it also highlights numerous limitations. One notable drawback is the use of secondary datasets, which means that the study lacks control over the structure and accessibility of the data and may lack crucial information as a result. Improvements are needed regarding the presence of a stimulus marker in both datasets, and including a single label for each individual in the first dataset presents a limiting factor. The temporal aspect of the second dataset is distorted due to its pre-existing epoching based on MWL levels. Furthermore, it is important to note that the datasets used in this study have a limited duration, specifically spanning only 2.5 and 15 minutes. This temporal constraint is a potential barrier regarding the amount of data available for analysis.

The first dataset, referred to as STEW (Simultaneous Task EEG Workload), is mentioned (**Section 3.1.1**). The retroactive MWL labeling was conducted after each experimental phase, disregarding any potential variations in MWL that may have occurred throughout the experiment. Due to the brief duration of each experimental phase, which encompassed resting and working periods lasting only three minutes, it remains unclear whether any noteworthy alterations in workload occurred within these limited time intervals.

In contrast, the hackathon dataset (see **Section 3.1.2**) exhibited a higher frequency of labeling, as measurements of MWL were recorded at intervals of two seconds. However, the labeling process was conducted pseudo-randomly, without a direct association with particular activities or stimuli. The current framework fails to consider the subjective variability in task complexity, which can significantly vary among people. An activity deemed effortless for one person may present challenges for another.

Moreover, the data organisation in the hackathon dataset further complicates matters. Each workload level has been classified as easy, medium, or high and then stored separately. This not only divorces the data from the original time context but also oversimplifies the inherently dynamic nature of MWL, which changes fluidly over time.

In summary, while beneficial, these datasets present distinct challenges due to their secondary nature and specific organisation. The dynamic, subjective, and temporal aspects of MWL are critical considerations often oversimplified or overlooked in the current data, limiting their potential for accurate interpretation and application.

Much progress has been made in interpreting EEG signals for assessing people's MWL levels. However, the complexity of these signals presents an intriguing challenge to those unfamiliar with the discipline, often inspiring further investigation. This section will discuss the challenges of using deep learning models to classify MWL levels based on EEG signals and possible future research directions.

EEG signal collection and the development of deep learning models for MWL classification face numerous obstacles. Diverse datasets employed by distinct research groups and a shortage of publicly accessible datasets hinder experiment replication and comparison of results. Additionally, the distinctive nature of each dataset and insufficient data obstruct the determination of relationships between input and output data. One specific challenge is underfitting, which can arise due to the distinct characteristics of each dataset. Insufficient data makes identifying connections between input and output data difficult, ultimately leading to underfitting in the models. Increasing the availability of online datasets is necessary to overcome these challenges.

7.3 Future Work

In this section, potential future research directions will be described that, if pursued, could significantly enhance the classification of MWL levels from EEG signals using deep learning models.

EEG Preprocessing and Noise Removal Challenges

A comprehensive EEG preprocessing pipeline is essential and empowering for machine learning practitioners without a neuroscience background. Artefact removal toolboxes are becoming increasingly sophisticated, with the capacity to autonomously cleanse EEG data of ocular, muscle, and cardiac signals based on identifiable patterns. The future development of pattern recognition algorithms for various environmental noises, such as traffic, trains, and aeroplanes, is essential and thrilling. This innovation will enable even more effective noise removal, enhancing EEG signal preprocessing quality in laboratory and real-world contexts with dynamic soundscapes.

Enhancing Model Generalisation and Minimising Calibration Requirements

The practical utility of deep learning models for MWL estimation is based on their capacity to generalise effectively and require minimal calibration, enhancing their applicability in real-world settings. An ideal model should possess strong generalisation properties, enabling its use across different subjects performing the same task. Additionally, the model should exhibit adaptability to mental and environmental fluctuations during a session, ensuring its relevance and accuracy in various contexts. Prioritising these attributes in model development can significantly improve the practicality and utility of deep learning models for EEG-based MWL classification.

Self-Reporting MWL Challenges

In classifying MWL, neural networks use EEG signals as input, supplemented by labels from participant evaluations. These labels, indicative of self-reported workload levels, are gathered through post-experiment questionnaires [128]. This approach can be viewed as a secondary task [237]. To conduct post-task self-report feedback or performance evaluations, individuals must be trained to understand the instrument used for expressing their MWL [224]. These methods can increase subjects' burdens, making it harder for them to respond to new events. To gain deeper insights into MWL in future studies, it could be beneficial to incorporate heart rate monitoring as a metric. This is because an increase in MWL often correlates with a corresponding rise in heart rate, as supported by research [27, 47, 150].

Integration of Artefact Removal and Online Learning in Advanced Deep Learning Models

Future research could also investigate the development of deep learning models that incorporate an integrated artifact removal layer. This approach could facilitate the direct input of raw data during the model training phase, thereby streamlining the overall process. Furthermore, creating models capable of continuous adaptation through online learning is essential for maintaining their relevance and accuracy in real-world applications. This combination of cutting-edge techniques can significantly improve the performance and utility of deep learning models for EEG-based MWL estimation.

Resource-Efficient Adaptive Modelling for Constrained Environments

Since a continuously adaptive model is needed, using cumbersome models can be inefficient regarding energy efficiency and computational cost. Tiny machine learning (tinyML) [222] is a cutting-edge field that applies machine learning to performance- and power-constrained devices. For example, devices that detect a pilot's MWL must be small and housed within a flight helmet. Operating neural networks on devices with limited resources requires algorithms and hardware co-design. The real-time control system is regarded as the modern vehicle's brain [93].

Temporal Dynamics in Cross-Validation for MWL EEG Analysis

Researchers investigating EEG signals in the context of MWL levels can enhance their studies by considering the inherent time series characteristics. This includes incorporating the assumption of independently and identically distributed (i.i.d.) time series elements into their cross-validation procedures, which can improve the robustness and reliability of their findings [18]. Traditional cross-validation approaches involve randomly splitting EEG signals into training and test sets, disregarding the temporal dynamics of MWL levels. To address this limitation and improve model accuracy, it is crucial to emphasize the importance of considering the temporal component when selecting cross-validation methods for EEG analysis. Since physiological signals are influenced by previous time steps and their statistical properties vary across individuals and types of mental tasks [236], future research should focus on developing models capable of capturing common properties found across subjects, sessions, and tasks.

Manual Feature Extraction

Manual feature extraction from EEG signals is time-consuming and labor-intensive in practical applications. Nevertheless, this approach facilitates a critical assessment of which feature set and classifier best suit a specific dataset [13]. Hand-crafted feature engineering relies heavily on meticulous preprocessing work and advanced domain knowledge [141], making model performance dependent on the quality of feature selection techniques. The type of features extracted varies across studies, illustrating the adaptability of these methodologies. Mohamed et al. [148] concentrated on time- and frequency-domain features. In contrast, Diaz et al. [54] focused exclusively on frequency-domain and predefined features predicated on the potential of the theta frequency band for assessing MWL. Consequently, these techniques enable comprehensive exploration and understanding, albeit at the expense of time and the potential loss of some pertinent data.

Managing Cross-Subject, Cross-Session, and Cross-Task Variability in MWL Classification

Deep learning, known for its swift growth and potential, has shown particular promise when applied to EEG studies, especially in classifying MWL. This potential, however, is coupled with substantial challenges posed by the inherent variability between subjects, sessions, and tasks. To effectively manage these multi-dimensional variables, we have grouped them under two broad classifications: “within” and “cross”, as illustrated in Figure 7.1. Details of each multi-dimensional variable will be explained in this section.

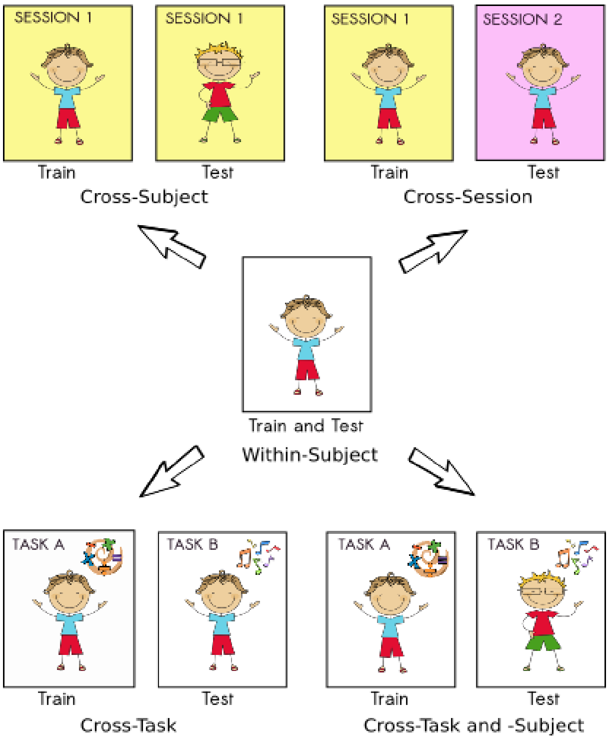


Figure 7.1: EEG MWL classification problems

1. **Within-Subject Variability.** From the literature, it is evident that the within-subject classification problem is the most popular study problem across papers related to EEG-based MWL classification. The "within-subject" approach focuses on charting an individual's MWL fluctuations as they engage in a singular task during one recording session. According to the literature, this methodology reduces the confounding effects of inter-individual variability by concentrating solely on intra-individual changes. This approach allows for an isolated exploration of an individual's MWL, which can be particularly useful in understanding individual physiological responses. Various model architectures and algorithms have been utilized to resolve this issue.

2. **Cross-Subject Variability.** The "cross-subject" approach, also known as the between-subjects or inter-subjects approach, is more complex. It strives to construct a predictive model using data from several subjects to forecast the MWL of unseen subjects. This strategy requires the model to be trained on data from

a cohort of subjects and then tested on its ability to classify the MWL of different individuals not included in the training phase. According to several studies, while this approach is fraught with challenges due to the inherent variability in EEG signals between individuals, it offers broader applicability. It necessitates a meticulous selection of machine learning algorithms and potentially requires the normalisation or standardisation of features to counterbalance individual differences.

Given that the assessment of MWL is vital for individuals in both daily life and work situations, it is crucial to construct models capable of effectively managing cross-subject variations. Most studies in the current literature have primarily focused on single-session experiments, underscoring the need for additional research on cross-subject models for improved generalisability and applicability in diverse contexts.

3. **Cross-Session Variability.** The "cross-session" approach involves tracking an individual's MWL across multiple sessions. This strategy seeks to develop a model capable of predicting the MWL from one session and then applying this model to data from different sessions. The model undergoes training during one session (the training set) and is then tested for its ability to classify MWL in a different session (the test set). While this approach allows for a more longitudinal assessment of an individual's MWL, it is challenged by the potential intra-individual variability in EEG signals between sessions, which might not be related to changes in MWL but other confounding factors such as fatigue or stress [176].

   Numerous studies have proposed innovative approaches for addressing cross-subject problems. However, the issue of cross-session variability remains relatively unexplored and presents unique challenges in EEG signal classification. The dataset may display substantial variation even when collected from the same participant during distinct sessions. As a result, models trained exclusively on EEG signals from one session may struggle with generalisation. Additionally, static pattern classifiers may not be suitable for classifying dynamic data, such as EEG signals recorded on different days. Several methodologies have recently been proposed to tackle the cross-session problem in response to these challenges.

4. **Cross-Task Variability.** In previous sections, the focus was on challenges associated with cross-subject and cross-session variations in EEG-based MWL classification. However, another crucial challenge in this field is the cross-task problem. The "cross-task" approach involves a single subject engaging in multiple tasks. The aim is to develop a model capable of predicting MWL across different tasks, training on data from one task and testing data collected during another task performed by the same subject. Despite its appeal for its potential to be generalisable across various tasks, the literature indicates that this method is complex due to the potential for different tasks to elicit varying types and levels of MWL, thus producing unique EEG signatures. This model is expected to predict MWL across various tasks and individuals.

5. **Cross-Task and -Subject Variability.** A combination of "cross-task" and "cross-subject" approaches presents a significant challenge yet promises the highest level of robustness and generalisability. This model is expected to predict MWL across various tasks and individuals. This problem has only been tackled by a few researchers.

   To the best of my knowledge, one significant contribution to this problem was made by Zeng et al. [240]. They developed two CNN-based EEG classifiers, EEG-Conv and EEG-Conv-R, to identify drivers' MWL. The EEG-Conv model employs a traditional CNN architecture, while EEG-Conv-R combines the CNN approach with deep residual learning to enhance performance. This combination addressed cross-task and cross-subject challenges, marking an innovative approach to EEG-based MWL classification. The potential for the development of more robust and versatile models was demonstrated through this research, signifying a significant step forward in handling cross-task and cross-subject variations. Nevertheless, the scarcity of studies investigating these combined problems indicates that further research is needed to establish more effective methods for managing such variations in real-world applications.

Guided by the existing literature, further combinations are envisioned, such as the “cross-subject” and “cross-session” methodologies, as well as the tripartite approach that combines the ‘cross-subject”, “cross-session”, and “cross-task” elements. The dual method of “cross-subject” and “cross-session” aims to develop a model that can predict MWL across subjects and sessions. “Cross-subject”, “cross-session”, and “cross-task” approaches present the most difficult but potentially most rewarding scenario. This ambitious strategy aims to develop a model capable of predicting MWL across a spectrum of individuals, sessions, and tasks, resulting in a highly adaptable tool with extensive practical applications. However, the literature on these complex interconnections remains sparse. The research community has yet to fully address the inherent challenges presented by these methodologies, rendering them a promising avenue for future exploration and innovation in this dynamic field.

Decoding MWL levels from EEG signals is difficult. This task presents many difficulties, primarily due to the intricate and numerous factors involved, all contributing to the overall difficulty of accurate MWL decoding. These challenges include cross-subject physiological variability arising from differences in individuals’ brain activities and physical responses. Additionally, cross-session variability refers to fluctuations in a single subject’s performance across different sessions, while cross-task variability highlights the differences that emerge when subjects perform various tasks. Moreover, the vast diversity of real-world environmental variables, such as ambient noise, lighting conditions, and external stressors, can also impact the performance of MWL decoders. To create more robust and accurate models, it is crucial to consider individual factors like gender, expertise, age, experience, and emotions during model training. These factors can significantly influence a person’s MWL, and by accounting for them, the models can better capture the nuances of MWL across different contexts and individuals.

Future research endeavours may seek to overcome these restrictions by designing experiments and collecting datasets independently. Moreover, it is valuable to pursue investigations into the integration of cross-task and MWL in practical contexts, such as education. Furthermore, it is important to develop task designs that build upon the existing research findings in order to provide support and guidance.

The observations obtained in this study establish a fundamental structure for BCI in tasks involving multitasking, which could potentially enhance the development of closed-loop systems. Despite accurately implementing the proposed model and following each stage of the approach, the application of our findings to real datasets in authentic MWL scenarios is still an area that requires further exploration. Future endeavours are focused on achieving practical implementation, particularly in the context of recognising the workload of individuals in driving scenarios. In the event that a person demonstrates elevated MWL or manifests indications of drowsiness, the system has the capability to notify the driver to cease or temporarily suspend their travel.

The existing framework of my thesis is grounded on an open-loop approach, which currently does not incorporate real-time feedback based on the user's brain signals. This setup serves as the foundation for the investigation and development pursued throughout this research. In the future, the objective is to integrate the circuits, thereby advancing our model into a closed-loop BCI system. This system will integrate a feedback mechanism, enabling real-time responses to the user's brain activity and promoting an interactive and dynamic exchange between the user and the system.

7.4 Chapter Summary

This chapter examines the contributions of this thesis, identifying and analysing the inherent constraints associated with these contributions. Additionally, it outlines research directions that could be investigated in the future.

- The chapter begins with a recap of the thesis's contributions, particularly in using EEG signals for MWL-level classification.

- This thesis highlights the importance of the EEG preprocessing step and reveals how different techniques impact the effectiveness of deep learning models in MWL level classification. The key contribution, explained in **Section 7.2**, concerns EEG channel selection for MWL classification. The results reveal that signals from specific EEG channels in certain brain regions can yield accurate MWL-level classifications. Furthermore, this thesis delves into the time series cross-validation strategy used in deep learning models for EEG-based MWL classification, showing that the expanding window strategy is superior to the rolling window strategy.

- The limitations are acknowledged, including challenges related to dataset diversity and scarcity, as well as the challenges of self-reporting MWL.
- Future research directions are outlined, highlighting the need for new and independent data collection for MWL-level classification and exploring “cross-task” and “cross-task and subject” applications. A closed-loop system in BCI is seen as a crucial advancement, enhancing real-time responsiveness and adaptability. This approach builds on the findings of the thesis, highlighting the dynamic nature of MWL and its implications for user-centered design in BCI technologies.

Bibliography

[1] Makoto's preprocessing pipeline.

[2] Amirmasoud Ahmadi, Hanieh Bazregarzadeh, and Kamran Kazemi. Automated detection of driver fatigue from electroencephalography through wavelet-based connectivity. *Biocybernetics and Biomedical Engineering*, 41(1):316–332, 2021.

[3] Ali Al-Saegh, Shefa A Dawwd, and Jassim M Abdul-Jabbar. Deep learning for motor imagery eeg-based classification: A review. *Biomedical Signal Processing and Control*, 63:102172, 2021.

[4] Mohammad A Almogbel, Anh H Dang, and Wataru Kameyama. Cognitive workload detection from raw eeg-signals of vehicle driver using deep learning. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pages 1–6. IEEE, 2019.

[5] Turky Alotaiby, Fathi E Abd El-Samie, Saleh A Alshebeili, and Ishtiaq Ahmad. A review of channel selection algorithms for eeg signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015:1–21, 2015.

[6] Xingwei An, Johannes Höhne, Dong Ming, and Benjamin Blankertz. Exploring combinations of auditory and visual stimuli for gaze-independent brain-computer interfaces. *PloS one*, 9(10):e111070, 2014.

[7] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing the channel selection and classification accuracy in eeg-based bci. *IEEE Transactions on Biomedical Engineering*, 58(6):1865–1873, 2011.

[8] Hasan Ayaz, Patricia A Shewokis, Scott Bunce, Kurtulus Izzetoglu, Ben Willems, and Banu Onaral. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47, 2012.

[9] Muhammad Zeeshan Baig, Nauman Aslam, and Hubert PH Shum. Filtering techniques for channel selection in motor imagery eeg applications: a survey. *Artificial intelligence review*, 53:1207–1232, 2020.

[10] Muhammad Zeeshan Baig, Nauman Aslam, Hubert PH Shum, and Li Zhang. Differential evolution algorithm as a tool for optimal feature subset selection in motor imagery eeg. *Expert Systems with Applications*, 90:184–195, 2017.

[11] Alexandre Barachant and Stephane Bonnet. Channel selection procedure using riemannian distance for bci applications. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*, pages 348–351. IEEE, 2011.

[12] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to bci classification. In *International conference on latent variable analysis and signal separation*, pages 629–636. Springer, 2010.

[13] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Lung sounds classification using convolutional neural networks. *Artificial intelligence in medicine*, 88:58–69, 2018.

[14] Ioannis Bargiotas, Alice Nicolaï, Pierre-Paul Vidal, Christophe Labourdette, Nicolas Vayatis, and Stéphane Buffat. The complementary role of activity context in the mental workload evaluation of helicopter pilots: a multi-tasking learning approach. In *International Symposium on Human Mental Workload: Models and Applications*, pages 222–238. Springer, 2018.

[15] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

[16] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[17] Hans Berger. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

[18] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[19] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192 – 213, 2012. Data Mining for Software Trustworthiness.

[20] Christoph Bergmeir, Mauro Costantini, and José M Benítez. On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, 76:132–143, 2014.

[21] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78(5), 2007.

[22] Pratibha R Bhise, Sonali B Kulkarni, and Talal A Aldhaheri. Brain computer interface based eeg for emotion recognition system: A systematic review. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 327–334. IEEE, 2020.

[23] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics*, 9:16, 2015.

[24] Maneesh Bilalpur, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. Eeg-based evaluation of cognitive workload induced by acoustic pa-

rameters for data sonification. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 315–323, 2018.

[25] Kenneth R Boff, Lloyd Kaufman, and James P Thomas. *Handbook of perception and human performance*, volume 1. Wiley New York, 1986.

[26] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[27] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377, 1996.

[28] Anne-Marie Brouwer, Maarten A Hogervorst, Jan BF Van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. Estimating workload using eeg spectral power and erps in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[29] Marco Buiatti and Andrea Mognon. *ADJUST: An Automatic EEG artifact Detector based on the Joint Use of Spatial and Temporal features, A Tutorial*, 2014 (accessed August 3, 2020). `https://www.nitrc.org/docman/view.php/739/2101/ADJUST%20Tutorial`.

[30] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V Elst. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):1–30, 2012.

[31] Brad Cain. A review of the mental workload literature. *DTIC Document*, 2007.

[32] Lixiao Cao, Zheng Qian, Hamidreza Zareipour, Zhenkai Huang, and Fanghong Zhang. Fault diagnosis of wind turbine gearbox based on deep bi-directional long short-term memory under time-varying non-stationary operating conditions. *IEEE Access*, 7:155219–155228, 2019.

[33] Zixuan Cao, Zhong Yin, and Jianhua Zhang. Recognition of cognitive load with a stacking network ensemble of denoising autoencoders and abstracted neurophysiological features. *Cognitive Neurodynamics*, 15(3):425–437, 2021.

[34] Alexander J Casson, David C Yates, Shelagh JM Smith, John S Duncan, and Esther Rodriguez-Villegas. Wearable electroencephalography. *IEEE engineering in medicine and biology magazine*, 29(3):44–56, 2010.

[35] Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020.

[36] Debashis Das Chakladar, Shubhashis Dey, Partha Pratim Roy, and Debi Prosad Dogra. Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm. *Biomedical Signal Processing and Control*, 60:101989, 2020.

[37] Jireh Yi-Le Chan, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, 10(8):1283, 2022.

[38] Rebecca L Charles and Jim Nixon. Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, 74:221–232, 2019.

[39] Gian Emilio Chatrian, Ettore Lettich, and Paula L Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked eeg activities. *American Journal of EEG technology*, 25(2):83–92, 1985.

[40] Zixu Chen, Guoliang Lu, Zhaohong Xie, and Wei Shang. A unified framework and method for eeg-based early epileptic seizure detection and epilepsy diagnosis. *IEEE Access*, 8:20080–20092, 2020.

[41] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase repre-

sentations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[42] Leon O Chua and Tamas Roska. The cnn paradigm. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 40(3):147–156, 1993.

[43] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[44] Thomas F Collura. History and evolution of electroencephalographic instruments and techniques. *Journal of clinical neurophysiology*, 10(4):476–504, 1993.

[45] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[46] Alex Dan, Miriam Reiner, et al. Real time eeg based measurements of cognitive load indicates mental states during learning. *Journal of Educational Data Mining*, 9(2):31–44, 2017.

[47] Michel De Rivecourt, MN Kuperus, WJ Post, and LJM Mulder. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, 51(9):1295–1319, 2008.

[48] Dick de Waard. The measurement of drivers' mental workload, 1996. s.n.

[49] Ranjan Debnath, George A Buzzell, Santiago Morales, Maureen E Bowers, Stephanie C Leach, and Nathan A Fox. The maryland analysis of developmental eeg (made) pipeline. *Psychophysiology*, 57(6):e13580, 2020.

[50] Frédéric Dehais, Alban Duprès, Sarah Blum, Nicolas Drougard, Sébastien Scannella, Raphaëlle N Roy, and Fabien Lotte. Monitoring pilot's mental workload using erps and spectral power with a six-dry-electrode eeg system in real flight conditions. *Sensors*, 19(6):1324, 2019.

[51] Stéphane Delliaux, Alexis Delaforge, Jean-Claude Deharo, and Guillaume Chaumet. Mental workload alters heart rate variability, lowering non-linear dynamics. *Frontiers in physiology*, 10:565, 2019.

[52] Dipayan Dewan, Lidia Ghosh, Biswadeep Chakraborty, Abir Chowdhury, Amit Konar, and Atulya K Nagar. Cognitive analysis of mental states of people according to ethical decisions using deep learning approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[53] Gianluca Di Flumeri, Pietro Aricò, Gianluca Borghini, Nicolina Sciaraffa, Antonello Di Florio, and Fabio Babiloni. The dry revolution: Evaluation of three different eeg dry electrode types in terms of signal spectral features, mental states classification and usability. *Sensors*, 19(6):1365, 2019.

[54] Carolina Diaz-Piedra, María Victoria Sebastián, and Leandro L Di Stasi. Eeg theta power activity reflects workload among army combat drivers: an experimental study. *Brain sciences*, 10(4):199, 2020.

[55] Tom Dietterich. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327, 1995.

[56] Robert DiPietro and Gregory D Hager. Deep learning: Rnns and lstm. In *Handbook of medical image computing and computer assisted intervention*, pages 503–519. Elsevier, 2020.

[57] Georg Dorffner. Neural networks for time series processing. In *Neural network world*. Citeseer, 1996.

[58] Caroline Dussault, Jean-Claude Jouanin, Matthieu Philippe, and Charles-Yannick Guezennec. Eeg and ecg changes during simulator operation reflect mental workload and vigilance. *Aviation, space, and environmental medicine*, 76(4):344–351, 2005.

[59] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[60] Emotiv. Emotiv epoc x – 14 channel mobile brainwear. `https://www.emotiv.com/product/emotiv-epoc-x-14-channel-mobile-brainwear/`, 2023. Accessed: [2020-05-05].

[61] Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. Emg and eog artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3):480–494, 2007.

[62] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[63] Laurel J. Gabard-Durnam, Adriana S. Mendez Leal, Carol L. Wilkinson, and April R. Levin. The harvard automated processing pipeline for electroencephalography (happe): Standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, 12:97, 2018.

[64] Edith Galy, Magali Cariou, and Claudine Mélan. What is the relationship between mental workload factors and cognitive load types? *International journal of psychophysiology*, 83(3):269–275, 2012.

[65] Yunyuan Gao, Bo Gao, Qiang Chen, Jia Liu, and Yingchun Zhang. Deep convolutional neural network-based epileptic electroencephalogram (eeg) signal classification. *Frontiers in neurology*, 11:375, 2020.

[66] Andrea Giorgi, Vincenzo Ronca, Alessia Vozzi, Nicolina Sciaraffa, Antonello Di Florio, Luca Tamborra, Ilaria Simonetti, Pietro Aricò, Gianluca Di Flumeri, Dario Rossi, et al. Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: A comparison with laboratory technologies. *Sensors*, 21(7):2332, 2021.

[67] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[68] Richard A Groeneveld and Glen Meeden. Measuring skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 33(4):391–399, 1984.

[69] Peter A Hancock. A dynamic model of stress and sustained attention. *Human factors*, 31(5):519–537, 1989.

[70] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[71] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[72] James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

[73] Ryan Hefron, Brett Borghetti, Christine Schubert Kabban, James Christensen, and Justin Estepp. Cross-participant eeg-based assessment of cognitive workload using multi-path convolutional recurrent neural networks. *Sensors*, 18(5):1339, 2018.

[74] Tobias Heine, Gustavo Lenis, Patrick Reichensperger, Tobias Beran, Olaf Doessel, and Barbara Deml. Electrocardiographic features for the measurement of drivers' mental workload. *Applied ergonomics*, 61:31–43, 2017.

[75] Christian Herff, Dominic Heger, Ole Fortmann, Johannes Hennrich, Felix Putze, and Tanja Schultz. Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, 7:935, 2014.

[76] Luis G Hernández, Oscar Martinez Mozos, José M Ferrández, and Javier M Antelis. Eeg-based detection of braking intention under different car driving conditions. *Frontiers in neuroinformatics*, 12:29, 2018.

[77] Hermann Hinrichs, Michael Scholz, Anne Katrin Baum, Julia WY Kam, Robert T Knight, and Hans-Jochen Heinze. Comparison between a wireless dry electrode eeg system with a conventional wired wet electrode eeg system for clinical applications. *Scientific reports*, 10(1):5218, 2020.

[78] Marcel F. Hinss, Ludovic Darmet, Bertille Somon, Emilie Jahanpour, Fabien Lotte, Simon Ladouce, and Raphaëlle N. Roy. An EEG dataset for cross-session mental workload estimation: Passive BCI competition of the Neuroergonomics Conference 2021, July 2021. The project was validated by the local ethical committee of the University of Toulouse (CER number 2021-342).

[79] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92:84–89, 2004.

[80] Bruce Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics*, pages 1977–1991, 1971.

[81] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[82] G Robert J Hockey. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological psychology*, 45(1-3):73–93, 1997.

[83] Daniel Hölle, Joost Meekes, and Martin G Bleichner. Mobile ear-eeg to study auditory attention in everyday life: Auditory attention in everyday life. *Behavior Research Methods*, 53(5):2025–2036, 2021.

[84] Mohammad-Parsa Hosseini, Amin Hosseini, and Kiarash Ahi. A review on machine learning for eeg signal processing in bioengineering. *IEEE reviews in biomedical engineering*, 14:204–218, 2020.

[85] Xinmei Hu, Shasha Yuan, Fangzhou Xu, Yan Leng, Kejiang Yuan, and Qi Yuan. Scalp eeg classification using deep bi-lstm network for seizure detection. *Computers in Biology and Medicine*, 124:103919, 2020.

[86] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799, 1951.

[87] Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. Methods for artifact detection and removal from scalp eeg: A review. *Neurophysiologie Clinique/Clinical Neurophysiology*, 46(4-5):287–305, 2016.

[88] Mir Riyanul Islam, Shaibal Barua, Mobyen Uddin Ahmed, Shahina Begum, and Gianluca Di Flumeri. Deep learning for automatic eeg feature extraction: an application in drivers' mental workload classification. In *International Symposium on Human Mental Workload: Models and Applications*, pages 121–135. Springer, 2019.

[89] Dibyanshu Jaiswal, Arijit Chowdhury, Tanushree Banerjee, and Debatri Chatterjee. Effect of mental workload on breathing pattern and heart rate for a working memory task: A pilot study. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2202–2206. IEEE, 2019.

[90] S Janjarasjitt, MS Scher, and KA Loparo. Nonlinear dynamical analysis of the neonatal eeg time series: the relationship between sleep state and complexity. *Clinical neurophysiology*, 119(8):1812–1823, 2008.

[91] Herbert H Jasper. Ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol*, 10:371–375, 1958.

[92] Ji-Hoon Jeong, Baek-Woon Yu, Dae-Hyeok Lee, and Seong-Whan Lee. Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional lstm network using electroencephalography signals. *Brain sciences*, 9(12):348, 2019.

[93] Yunyi Jia, Longxiang Guo, and Xin Wang. Real-time control systems. In *Transportation Cyber-Physical Systems*, pages 81–113. Elsevier, 2018.

[94] Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, and Haojun Xu. Deep convolutional neural networks for mental load classification based on eeg data. *Pattern Recognition*, 76:582–595, 2018.

[95] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil. Cross-subject classification of cognitive loads using a recurrent-residual deep network. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2017.

[96] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil. Custom domain adaptation: A new method for cross-subject, eeg-based cognitive load recognition. *IEEE Signal Processing Letters*, 27:750–754, 2020.

[97] Jing Jin, Yangyang Miao, Ian Daly, Cili Zuo, Dewen Hu, and Andrzej Cichocki. Correlation-based channel selection and regularized feature optimization for mi-based bci. *Neural Networks*, 118:262–270, 2019.

[98] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

[99] Don H Johnson. Signal-to-noise ratio. *Scholarpedia*, 1(12):2088, 2006.

[100] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

[101] Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage*, 34(4):1600–1611, 2007.

[102] Daniel Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.

[103] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

[104] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE, 2018.

[105] N Kannathal, U Rajendra Acharya, Choo Min Lim, and PK Sadasivan. Characterization of eeg—a comparative study. *Computer methods and Programs in Biomedicine*, 80(1):17–23, 2005.

[106] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Insights into lstm fully convolutional networks for time series classification. *IEEE Access*, 7:67718–67725, 2019.

[107] Philippa J Karoly, Vikram R Rao, Nicholas M Gregg, Gregory A Worrell, Christophe Bernard, Mark J Cook, and Maxime O Baud. Cycles in epilepsy. *Nature Reviews Neurology*, 17(5):267–284, 2021.

[108] Aneta Kartali, Milica M Janković, Ivan Gligorijević, Pavle Mijović, Bogdan Mijović, and Maria Chiara Leva. Real-time mental workload estimation using eeg. In *Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, November 14–15, 2019, Proceedings 3*, pages 20–34. Springer, 2019.

[109] Kunjira Kingphai and Yashar Moshfeghi. On eeg preprocessing role in deep learning effectiveness for mental workload classification. In *International Symposium on Human Mental Workload: Models and Applications*, pages 81–98. Springer, 2021.

[110] Kunjira Kingphai and Yashar Moshfeghi. On Time Series Cross-Validation for Mental Workload Classification from EEG Signals. Neuroergonomics conference, September 2021. Poster.

[111] Kunjira Kingphai and Yashar Moshfeghi. On time series cross-validation for deep learning classification model of mental workload levels based on eeg signals. In *Machine Learning, Optimization, and Data Science: 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19–22, 2022, Revised Selected Papers, Part II*, pages 402–416. Springer, 2023.

[112] Kunjira Kingphai and Yashar Moshfeghi. On time series cross-validation for deep learning classification model of mental workload levels based on eeg signals. In Giuseppe Nicosia, Varun Ojha, Emanuele La Malfa, Gabriele La Malfa, Panos Pardalos, Giuseppe Di Fatta, Giovanni Giuffrida, and Renato Umeton, editors, *Machine Learning, Optimization, and Data Science*, pages 402–416, Cham, 2023. Springer Nature Switzerland.

[113] Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[114] George H Klem. The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.*, 52:3–6, 1999.

[115] Thomas Koenig, Leslie Prichep, Dietrich Lehmann, Pedro Valdes Sosa, Elisabeth Braeker, Horst Kleinlogel, Robert Isenhart, and E Roy John. Millisecond by millisecond, year by year: normative eeg microstates and developmental stages. *Neuroimage*, 16(1):41–48, 2002.

[116] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and Kamisetty R Rao. Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2576–2580. IEEE, 2018.

[117] Jean-Philippe Lachaux, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E Crone. High-frequency neural activity and human cognition: past, present and possible future of intracranial eeg research. *Progress in neurobiology*, 98(3):279–301, 2012.

[118] Zenon Lamprou, Frank Pollick, and Yashar Moshfeghi. Role of punctuation in semantic mapping between brain and transformer models. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 458–472. Springer, 2022.

[119] Tian Lan, Deniz Erdogmus, Andre Adami, Misha Pavel, and Santosh Mathan. Salient eeg channel selection in brain computer interfaces by mutual information maximization. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7064–7067. IEEE, 2006.

[120] Helmut Laufs, Andreas Kleinschmidt, Astrid Beyerle, Evelyn Eger, Afraim Salek-Haddadi, Christine Preibisch, and Karsten Krakow. Eeg-correlated fmri of human alpha activity. *Neuroimage*, 19(4):1463–1476, 2003.

[121] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[122] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[123] Dae-Hyeok Lee, Ji-Hoon Jeong, Kiduk Kim, Baek-Woon Yu, and Seong-Whan Lee. Continuous eeg decoding of pilots' mental states using multiple feature block-based convolutional neural network. *IEEE Access*, 8:121929–121941, 2020.

[124] Dae-Hyeok Lee, Ji-Hoon Jeong, Kiduk Kim, Baek-Woon Yu, and Seong-Whan Lee. Continuous eeg decoding of pilots' mental states using multiple feature block-based convolutional neural network. *IEEE Access*, 8:121929–121941, 2020.

[125] Choon Guan Lim, Tih Shih Lee, Cuntai Guan, Daniel Shuen Sheng Fung, Yudong Zhao, Stephanie Sze Wei Teng, Haihong Zhang, and K Ranga Rama Krishnan. A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder. *PloS one*, 7(10):e46692, 2012.

[126] Julian Lim, Wen-chau Wu, Jiongjiong Wang, John A Detre, David F Dinges, and Hengyi Rao. Imaging brain fatigue from sustained mental workload: an asl perfusion study of the time-on-task effect. *Neuroimage*, 49(4):3426–3435, 2010.

[127] Wei Lun Lim, Olga Sourina, and Lipo Wang. Cross dataset workload classification using encoded wavelet decomposition features. In *2018 International Conference on Cyberworlds (CW)*, pages 300–303. IEEE, 2018.

[128] WL Lim, O Sourina, and Lipo P Wang. Stew: Simultaneous task eeg workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2106–2114, 2018.

[129] Yuetian Liu and Qingshan Liu. Convolutional neural networks with large-margin softmax loss function for cognitive load recognition. In *2017 36th Chinese control conference (CCC)*, pages 4045–4049. IEEE, 2017.

[130] Ziming Liu, Jeremy Shore, Miao Wang, Fengpei Yuan, Aaron Buss, and Xiaopeng Zhao. A systematic review on hybrid eeg/fnirs in brain-computer interface. *Biomedical Signal Processing and Control*, 68:102595, 2021.

[131] Luca Longo and M Chiara Leva. *Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers*, volume 726. Springer, 2017.

[132] Luca Longo, Christopher D Wickens, Gabriella Hancock, and Peter A Hancock. Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, 13:883321, 2022.

[133] Steven J Luck. *An introduction to the event-related potential technique.* MIT press, 2014.

[134] Kelvin FH Lui and Alan C-N Wong. Does media multitasking always hurt? a positive correlation between multitasking and multisensory integration. *Psychonomic bulletin & review*, 19:647–653, 2012.

[135] Sebastian Mach, Pamela Storozynski, Josephine Halama, and Josef F Krems. Assessing mental workload with wearable devices–reliability and applicability of heart rate and motion measurements. *Applied ergonomics*, 105:103855, 2022.

[136] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in neural information processing systems*, pages 145–151, 1996.

[137] Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, 9(5):504–512, 2001.

[138] Robert Matthews, Neil J McDonald, Harini Anumula, Jamison Woodward, Peter J Turner, Martin A Steindorf, Kaichun Chang, and Joseph M Pendleton. Novel hybrid bioelectrodes for ambulatory zero-prep eeg measurements using multichannel wireless eeg system. In *Foundations of Augmented Cognition: Third International Conference, FAC 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 3*, pages 137–146. Springer, 2007.

[139] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

[140] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[141] Nijat Mehdiyev, Johannes Lahann, Andreas Emrich, David Enke, Peter Fettke, and Peter Loos. Time series classification using deep learning for process planning: A case from the process industry. *Procedia Computer Science*, 114:242 – 249, 2017. Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA.

[142] Dominika Michalkova, Mario Parra Rodriguez, and Yashar Moshfeghi. Drivers of information needs: a behavioural study–exploring searcher's feeling-of-knowing. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 171–181, 2022.

[143] Dominika Michalkova, Mario Parra Rodriguez, and Yashar Moshfeghi. Confidence as part of searcher's cognitive context. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 510–524. Springer, 2022.

[144] Dominika Michalkova, Mario Parra Rodriguez, and Yashar Moshfeghi. Understanding feeling-of-knowing in information search: An eeg study. *ACM Transactions on Information Systems*, 2023.

[145] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[146] Yurui Ming, Danilo Pelusi, Chieh-Ning Fang, Mukesh Prasad, Yu-Kai Wang, Dongrui Wu, and Chin-Teng Lin. Eeg data analysis with stacked differentiable neural computers. *Neural Computing and Applications*, 32(12):7611–7621, 2020.

[147] Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti. Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240, 2011.

[148] Zainab Mohamed, Mohamed El Halaby, Tamer Said, Doaa Shawky, and Ashraf Badawi. Characterizing focused attention and working memory using eeg. *Sensors*, 18(11):3743, 2018.

[149] Yashar Moshfeghi. Neurasearch: Neuroscience and information retrieval. In *CEUR Workshop Proceedings*, volume 2950, pages 193–194, 2021.

[150] Lambertus JM Mulder. Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology*, 34(2-3):205–236, 1992.

[151] Hiroyasu Murakami and BVK Vijaya Kumar. Efficient calculation of primary images from a set of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):511–515, 1982.

[152] Deogratias Mzurikwao, Oluwarotimi Williams Samuel, Mojisola Grace Asogbon, Xiangxin Li, Guanglin Li, Woon-Hong Yeo, Christos Efstratiou, and Chee Siang

Ang. A channel selection approach based on convolutional neural network for multi-channel eeg motor imagery decoding. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 195–202. IEEE, 2019.

[153] Friedhelm Nachreiner. International standards on mental work-load the iso 10 075 series. *Industrial Health*, 37(2):125–133, 1999.

[154] Perattur Nagabushanam, S Thomas George, and Subramanyam Radha. Eeg signal classification using lstm and improved neural network algorithms. *Soft Computing*, pages 1–23, 2019.

[155] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981.

[156] Robert Oostenveld and Peter Praamstra. The five percent electrode system for high-resolution eeg and erp measurements. *Clinical neurophysiology*, 112(4):713–719, 2001.

[157] Alison O'Shea, Rehan Ahmed, Gordon Lightbody, Elena Pavlidis, Rhodri Lloyd, Francesco Pisani, Willian Marnane, Sean Mathieson, Geraldine Boylan, and Andriy Temko. Deep learning for eeg seizure detection in preterm infants. *International Journal of Neural Systems*, 31(08):2150008, 2021.

[158] Sakrapee Paisalnan, Yashar Moshfeghi, and Frank Pollick. Neural correlates of realisation of satisfaction in a successful search process. *Proceedings of the Association for Information Science and Technology*, 58(1):282–291, 2021.

[159] Sakrapee Paisalnan, Frank Pollick, and Yashar Moshfeghi. Towards understanding neuroscience of realisation of information need in light of relevance and satisfaction judgement. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 41–56. Springer, 2021.

[160] Sakrapee Paisalnan, Frank Pollick, and Yashar Moshfeghi. Neural correlates of satisfaction of an information need. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 443–457. Springer, 2022.

[161] Alvaro Pascual-Leone, Catarina Freitas, Lindsay Oberman, Jared C Horvath, Mark Halko, Mark Eldaief, Shahid Bashir, Marine Vernet, Mouhshin Shafi, Brandon Westover, et al. Characterizing brain cortical plasticity and network dynamics across the age-span in health and disease with tms-eeg and tms-fmri. *Brain topography*, 24:302–315, 2011.

[162] Christopher JD Patten, Albert Kircher, Joakim Östlund, and Lena Nilsson. Using mobile telephones: cognitive workload and attention resource allocation. *Accident analysis & prevention*, 36(3):341–350, 2004.

[163] Christopher JD Patten, Albert Kircher, Joakim Östlund, Lena Nilsson, and Ola Svenson. Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention*, 38(5):887–894, 2006.

[164] Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, 2022.

[165] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[166] Zuzana Pinkosova, William J McGeown, and Yashar Moshfeghi. Revisiting neurological aspects of relevance: an eeg study. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 549–563. Springer, 2022.

[167] Abdul Qayyum, Ibrahima Faye, Aamir Saeed Malik, and Moona Mazher. Assessment of cognitive load using multimedia learning and resting states with deep learning perspective. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 600–605. IEEE, 2018.

[168] Abdul Qayyum, MKA Ahamed Khan, Moona Mazher, and M Suresh. Classification of eeg learning and resting states using 1d-convolutional neural network for cognitive load assesment. In *2018 IEEE Student Conference on Research and Development (SCOReD)*, pages 1–5. IEEE, 2018.

[169] Tingnan Qu, Jing Jin, Ren Xu, Xingyu Wang, and Andrzej Cichocki. Riemannian distance based channel selection and feature extraction combining discriminative time-frequency bands and riemannian tangent space for mi-bcis. *Journal of Neural Engineering*, 19(5):056025, 2022.

[170] Xiaodong Qu, Peiyan Liu, Zhaonan Li, and Timothy Hickey. Multi-class time continuity voting for eeg classification. In *International Conference on Brain Function Assessment in Learning*, pages 24–33. Springer, 2020.

[171] Xiaodong Qu, Yixin Sun, Robert Sekuler, and Timothy Hickey. Eeg markers of stem learning. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–9, 2018.

[172] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

[173] Gary B Reid and Thomas E Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*, volume 52, pages 185–218. Elsevier, 1988.

[174] Izabela Rejer and Pawel Górski. Benefits of ica in the case of a few channel eeg. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7434–7437. IEEE, 2015.

[175] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[176] Raphaëlle N Roy, Marcel F Hinss, Ludovic Darmet, Simon Ladouce, Emilie S Jahanpour, Bertille Somon, Xiaoqi Xu, Nicolas Drougard, Frédéric Dehais, and

Fabien Lotte. Retrospective on the first passive brain-computer interface competition on cross-session workload estimation. *Frontiers in Neuroergonomics*, 3, 2022.

[177] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[178] Anushri Saha, Vikash Minz, Sanjith Bonela, SR Sreeja, Ritwika Chowdhury, and Debasis Samanta. Classification of eeg signals for cognitive load estimation using deep learning architectures. In *International Conference on Intelligent Human Computer Interaction*, pages 59–68. Springer, 2018.

[179] Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, page 87, 2020.

[180] Syed Moshfeq Salaken, Imali Hettiarachchi, Luke Crameri, Samer Hanoun, Thanh Nguyen, and Saeid Nahavandi. Evaluation of classification techniques for identifying cognitive load levels using eeg signals. In *2020 IEEE International Systems Conference (SysCon)*, pages 1–8. IEEE, 2020.

[181] Kaveh Samiee, Peter Kovacs, and Moncef Gabbouj. Epileptic seizure classification of eeg time-series using rational discrete short-time fourier transform. *IEEE transactions on Biomedical Engineering*, 62(2):541–552, 2014.

[182] Yamira Santiago-Espada, Robert R Myer, Kara A Latorella, and James R Comstock Jr. The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide. Technical report, 2011.

[183] Cullen Schaffer. Selecting a classification method by cross-validation. *Machine Learning*, 13(1):135–143, 1993.

[184] Michael Scherg, Nicole Ille, Harald Bornfleth, and Patrick Berg. Advanced tools for digital eeg review:: virtual source montages, whole-head mapping, correlation, and phase analysis. *Journal of Clinical Neurophysiology*, 19(2):91–112, 2002.

[185] Wolfgang Schnotz and Christian Kürschner. A reconsideration of cognitive load theory. *Educational psychology review*, 19:469–508, 2007.

[186] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[187] Nicolina Sciaraffa, Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Antonio Di Florio, and Fabio Babiloni. On the use of machine learning for eeg-based workload assessment: algorithms comparison in a realistic task. In *International Symposium on Human Mental Workload: Models and Applications*, pages 170–185. Springer, 2019.

[188] Margitta Seeck, Laurent Koessler, Thomas Bast, Frans Leijten, Christoph Michel, Christoph Baumgartner, Bin He, and Sándor Beniczky. The standardized eeg electrode array of the ifcn. *Clinical neurophysiology*, 128(10):2070–2077, 2017.

[189] SHILIANG SHAO, TING WANG, CHUNHE SONG, YUN SU, YONGLIANG WANG, and CHEN YAO. Fine-grained and multi-scale motif features for cross-subject mental workload assessment using bi-lstm. *Journal of Mechanics in Medicine and Biology*, page 2140020, 2021.

[190] Jian Shen, Xiaowei Zhang, Xiao Huang, Manxi Wu, Jin Gao, Dawei Lu, Zhijie Ding, and Bin Hu. An optimal channel selection for eeg-based depression detection via kernel-target alignment. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2545–2556, 2020.

[191] Kip Smith and Peter A Hancock. Situation awareness is adaptive, externally directed consciousness. *Human factors*, 37(1):137–148, 1995.

[192] Andres Soler, Eduardo Giraldo, Lars Magne Lundheim, and Maria Marta Molinas Cabrera. Relevance-based channel selection for eeg source reconstruction: An

approach to identify low-density channel subsets. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022)-Volume 1*. SciTePress, 2022.

[193] Marios Sophocleous and John K Atkinson. A review of screen-printed silver/silver chloride (ag/agcl) reference electrodes potentially suitable for environmental potentiometric sensors. *Sensors and Actuators A: Physical*, 267:106–120, 2017.

[194] Tanya M Spruill. Chronic psychosocial stress and hypertension. *Current hypertension reports*, 12:10–16, 2010.

[195] James L Stone and John R Hughes. Early history of electroencephalography and establishment of the american clinical neurophysiology society. *Journal of Clinical Neurophysiology*, 30(1):28–44, 2013.

[196] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[197] E Suarez, MD Viegas, M Adjouadi, and A Barreto. Relating induced changes in eeg signals to orientation of visual stimuli using the esi-256 machine. *Biomedical sciences instrumentation*, 36:33–38, 2000.

[198] Zhe Sun, Binghua Li, Feng Duan, Hao Jia, Shan Wang, Yu Liu, Andrzej Cichocki, Cesar F Caiafa, and Jordi Sole-Casals. Wlnet: towards an approach for robust workload estimation based on shallow neural networks. *IEEE Access*, 9:3165–3173, 2020.

[199] James L Szalma, Joel S Warm, Gerald Matthews, William N Dember, Ernest M Weiler, Ashley Meier, and F Thomas Eggemeier. Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human factors*, 46(2):219–233, 2004.

[200] Masaaki Tanaka, Akira Ishii, and Yasuyoshi Watanabe. Neural effects of mental fatigue caused by continuous attention load: a magnetoencephalography study. *Brain research*, 1561:60–66, 2014.

[201] Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450, 2000.

[202] William O Tatum IV. *Handbook of EEG interpretation.* Springer Publishing Company, 2021.

[203] Laura P. Taylor. Chapter 20 - independent assessor audit guide. In Laura P. Taylor, editor, *FISMA Compliance Handbook*, pages 239–273. Syngress, Boston, 2013.

[204] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[205] Pierre Thiffault and Jacques Bergeron. Monotony of road environment and driver fatigue: a simulator study. *Accident Analysis & Prevention*, 35(3):381–391, 2003.

[206] Laiyuan Tong, Jinchuang Zhao, and Wenli Fu. Emotion recognition and channel selection based on eeg signal. In *2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pages 101–105. IEEE, 2018.

[207] Jan Törnros and Anne Bolling. Mobile phone use–effects of conversation on mental workload and driving speed in rural and urban environments. *Transportation Research Part F: Traffic Psychology and Behaviour*, 9(4):298–306, 2006.

[208] Pamela S Tsang and Velma L Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.

[209] Jose Antonio Urigüen and Begoña Garcia-Zapirain. EEG artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, apr 2015.

[210] Swati Vaid, Preeti Singh, and Chamandeep Kaur. Eeg signal analysis for bci interface: A review. In *2015 fifth international conference on advanced computing & communication technologies*, pages 143–147. IEEE, 2015.

[211] Bram B Van Acker, Davy D Parmentier, Peter Vlerick, and Jelle Saldien. Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, technology & work*, 20:351–365, 2018.

[212] W Van Winsum, L Herland, and M Martens. *The effects of speech versus tactile driver support messages on workload, driver behaviour and user acceptance.* TNO Human Factors Research Institute, 1999.

[213] Abhishek Varshney, Samit Kumar Ghosh, Sibasankar Padhy, Rajesh Kumar Tripathy, and U Rajendra Acharya. Automated classification of mental arithmetic tasks using recurrent neural network and entropy features obtained from multichannel eeg signals. *Electronics*, 10(9):1079, 2021.

[214] JA Veltman and AWK Gaillard. Physiological indices of workload in a simulated flight task. *Biological psychology*, 42(3):323–342, 1996.

[215] Michael A Vidulich and Pamela S Tsang. Mental workload and situation awareness. *Handbook of human factors and ergonomics*, pages 243–273, 2012.

[216] Robert H Walden. Analog-to-digital converter survey and analysis. *IEEE Journal on selected areas in communications*, 17(4):539–550, 1999.

[217] Xin Wan, Kezhong Zhang, S Ramkumar, J Deny, G Emayavaramban, M Siva Ramkumar, and Ahmed Faeq Hussein. A review on electroencephalogram based brain computer interface for elderly disabled. *IEEE Access*, 7:36380–36387, 2019.

[218] Peng Wang, Weining Fang, and Beiyuan Guo. A measure of mental workload during multitasking: Using performance-based timed petri nets. *International Journal of Industrial Ergonomics*, 75:102877, 2020.

[219] Shouyi Wang, Jacek Gwizdka, and W Art Chaovalitwongse. Using wireless eeg signals to assess memory workload in the n-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, 2015.

[220] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[221] Zhong-Min Wang, Shu-Yuan Hu, and Hui Song. Channel selection method for eeg emotion recognition using normalized mutual information. *IEEE Access*, 7:143303–143311, 2019.

[222] Pete Warden and Daniel Situnayake. *TinyML*. O'Reilly Media, Incorporated, 2019.

[223] Kevin Whittingstall, Gerhard Stroink, Larry Gates, JF Connolly, and Allen Finley. Effects of dipole position, orientation and noise on the accuracy of eeg source localization. *Biomedical engineering online*, 2(1):1–5, 2003.

[224] Eric N Wiebe, Edward Roberts, and Tara S Behrend. An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, 26(3):474–481, 2010.

[225] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[226] Chuhao Wu, Jackie Cha, Jay Sulek, Tian Zhou, Chandru P Sundaram, Juan Wachs, and Denny Yu. Eye-tracking metrics predict perceived workload in robotic surgical skills training. *Human factors*, 62(8):1365–1386, 2020.

[227] Edmond Q Wu, XY Peng, Caizhi Z Zhang, JX Lin, and Richard SF Sheng. Pilots' fatigue status recognition using deep contractive autoencoder network. *IEEE Transactions on Instrumentation and Measurement*, 68(10):3907–3919, 2019.

[228] CD Wylie, T Shultz, JC Miller, MM Mitler, RR Mackie, et al. Commercial motor vehicle driver fatigue and alertness study: Technical summary. 1996.

[229] Liying Yang, Si Chao, Qingyang Zhang, Pei Ni, and Dunhui Liu. A grouped dynamic eeg channel selection method for emotion recognition. In *2021 IEEE In-*

ternational Conference on Bioinformatics and Biomedicine (BIBM), pages 3689–3696. IEEE, 2021.

[230] Shuo Yang, Zhong Yin, Yagang Wang, Wei Zhang, Yongxiong Wang, and Jianhua Zhang. Assessing cognitive mental workload via eeg signals and an ensemble deep learning classifier based on denoising autoencoders. *Computers in biology and medicine*, 109:159–170, 2019.

[231] Sana Yasin, Syed Asad Hussain, Sinem Aslan, Imran Raza, Muhammad Muzammel, and Alice Othmani. Eeg based major depressive disorder and bipolar disorder detection using neural networks: A review. *Computer Methods and Programs in Biomedicine*, 202:106007, 2021.

[232] Bayya Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[233] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

[234] Zhong Yin and Jianhua Zhang. Cross-session classification of mental workload levels using eeg and an adaptive deep learning model. *Biomedical Signal Processing and Control*, 33:30–47, 2017.

[235] Zhong Yin and Jianhua Zhang. Cross-subject recognition of operator functional states via eeg and switching deep belief networks with adaptive weights. *Neurocomputing*, 260:349–366, 2017.

[236] Zhong Yin, Mengyuan Zhao, Wei Zhang, Yongxiong Wang, Yagang Wang, and Jianhua Zhang. Physiological-signal-based mental workload estimation via transfer dynamical autoencoders in a deep learning framework. *Neurocomputing*, 347:212–229, 2019.

[237] Mark S Young, Karel A Brookhuis, Christopher D Wickens, and Peter A Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.

[238] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[239] Hong Zeng, Xiufeng Li, Gianluca Borghini, Yue Zhao, Pietro Aricò, Gianluca Di Flumeri, Nicolina Sciaraffa, Wael Zakaria, Wanzeng Kong, and Fabio Babiloni. An eeg-based transfer learning method for cross-subject fatigue mental state prediction. *Sensors*, 21(7):2369, 2021.

[240] Hong Zeng, Chen Yang, Guojun Dai, Feiwei Qin, Jianhai Zhang, and Wanzeng Kong. Eeg classification of driver mental states by deep learning. *Cognitive neurodynamics*, 12(6):597–606, 2018.

[241] Hong Zeng, Chen Yang, Hua Zhang, Zhenhua Wu, Jiaming Zhang, Guojun Dai, Fabio Babiloni, and Wanzeng Kong. A lightgbm-based eeg analysis method for driver mental states classification. *Computational intelligence and neuroscience*, 2019, 2019.

[242] Dongdong Zhang, Dong Cao, and Haibo Chen. Deep learning decoding of mental state in non-invasive brain computer interface. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, pages 1–5, 2019.

[243] Jianhua Zhang and Sunan Li. A deep learning scheme for mental workload classification based on restricted boltzmann machines. *Cognition, Technology & Work*, 19(4):607–631, 2017.

[244] Jianhua Zhang, Sunan Li, and Zhong Yin. Pattern classification of instantaneous mental workload using ensemble of convolutional neural networks. *IFAC-PapersOnLine*, 50(1):14896–14901, 2017.

[245] Jianhua Zhang, Zhong Yin, and Rubin Wang. Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. *IEEE Transactions on Human-Machine Systems*, 45(2):200–214, 2014.

[246] Pengbo Zhang, Xue Wang, Junfeng Chen, Wei You, and Weihang Zhang. Spectral and temporal feature learning with two-stream neural networks for mental workload assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1149–1159, 2019.

[247] Qiankun Zhang, Ziqian Yuan, He Chen, and Xiaoli Li. Identifying mental workload using eeg and deep learning. In *2019 Chinese Automation Congress (CAC)*, pages 1138–1142. IEEE, 2019.

[248] Wenxiang Zhang and Qingshan Liu. Using the center loss function to improve deep learning performance for eeg signal classification. In *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, pages 578–582. IEEE, 2018.

[249] Yong Zhang, Bo Liu, Xiaomin Ji, and Dan Huang. Classification of eeg signals based on autoregressive model and wavelet packet decomposition. *Neural Processing Letters*, 45(2):365–378, 2017.

[250] Yu-Dong Zhang, Zhengchao Dong, Shui-Hua Wang, Xiang Yu, Xujing Yao, Qinghua Zhou, Hua Hu, Min Li, Carmen Jiménez-Mesa, Javier Ramirez, et al. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020.

[251] Rui Zhao, Ruqiang Yan, Jinjiang Wang, and Kezhi Mao. Learning to monitor machine health with convolutional bi-directional lstm networks. *Sensors*, 17(2):273, 2017.

[252] Wenrui Zhao, Eus JW Van Someren, Chenyu Li, Xinyuan Chen, Wenjun Gui, Yu Tian, Yunrui Liu, and Xu Lei. Eeg spectral analysis in insomnia disorder: A systematic review and meta-analysis. *Sleep Medicine Reviews*, 59:101457, 2021.

[253] Zhanpeng Zheng, Zhong Yin, and Jianhua Zhang. An elm-based deep sdae ensemble for inter-subject cognitive workload estimation with physiological signals. In *2020 39th Chinese Control Conference (CCC)*, pages 6237–6242. IEEE, 2020.

[254] Sheng-hua Zhong, Ahmed Fares, and Jianmin Jiang. An attentional-lstm for improved classification of brain activities evoked by images. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1295–1303, 2019.

[255] Bangyan Zhou, Xiaopei Wu, Jing Ruan, LV Zhao, and Lei Zhang. How many channels are suitable for independent component analysis in motor imagery brain-computer interface. *Biomedical Signal Processing and Control*, 50:103–120, 2019.

[256] Yun Zhou, Tao Xu, Shaoqi Li, and Ruifeng Shi. Beyond engagement: an eeg-based methodology for assessing user's confusion in an educational game. *Universal Access in the Information Society*, 18(3):551–563, 2019.