ID TCDS-2023-0548.R1

Mental Workload Assessment Using Deep Learning Models from EEG Signals: A Systematic Review

Kunjira Kingphai and Yashar Moshfeghi
NeuraSearch Laboratory, University of Strathclyde Glasgow, UK
kunjira.kingphai@strath.ac.uk, yashar.mosfeghi@strath.ac.uk

Abstract— **Mental workload (MWL) assessment is crucial in information systems (IS), impacting task performance, user experience, and system effectiveness. Deep learning offers promising techniques for MWL classification using electroencephalography (EEG), which monitors cognitive states dynamically and unobtrusively. Our research explores deep learning's potential and challenges in EEG-based MWL classification, focusing on training inputs, cross-validation methods, and classification problem types. We identify five types of EEG-based MWL classification: within-subject, cross-subject, cross-session, cross-task, and combined cross-task and -subject. Success depends on managing dataset uniqueness, session and task variability, and artifact removal. Despite potential, real-world applications are limited. Enhancements are necessary in self-reporting methods, universal preprocessing standards, and MWL assessment accuracy. Specifically, inaccuracies are inflated when data is shuffled before splitting to train and test sets, disrupting EEG signals' temporal sequence. In contrast, methods like the time-series cross-validation or leave-session-out approach better preserve temporal integrity, offering more accurate model performance evaluations. Utilizing deep learning for EEG-based MWL assessment could significantly improve IS functionality and adaptability in real-time based on user cognitive states.**

Index Terms— **Cross-validation, Deep learning, EEG signals, Mental workload**

I. INTRODUCTION

In this paper, we propose a systematic review and meta-analysis of current research on deep learning techniques for classifying mental workload (MWL) levels using EEG data. This review focuses on identifying opportunities, challenges, and best practices within the field. MWL describes the cognitive resources required to engage in a particular task and encompasses mental effort, cognitive engagement, and the overall demands on an individual's cognitive system during task performance. MWL is crucial when designing and optimizing any system requiring human interaction, significantly affecting performance, safety, efficiency, and user satisfaction. It plays an important role in various aspects of human life, such as influencing attention disorders in children [1], driving fatigue [2], [3], and task performance [4].

An effective MWL prediction and management are crucial for optimizing human performance and preventing cognitive overload or underload. 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 [5]. Therefore, it is critical to find the right amount of work that helps people perform at their best without causing problems [6]. Specifically, MWL plays an indispensable role across all facets of information interaction, encompassing both retrieval and consumption. The workload level directly impacts the efficacy of Information Systems (IS), influencing the user's capacity to effectively locate, comprehend, and apply the acquired information [7]. By comprehending and accurately predicting MWL, we can facilitate the creation of adaptive IS systems [8]. Innovative systems designed to align with the user's cognitive state dynamically promise substantial improvements in productivity, precision, and user satisfaction [9]. 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 [10]. 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 [11], 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 [12].

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) [13], Subjective Assessment Technique (SWAT) [10], and the Workload Profile [14]. 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 [15]. Additionally, self-reporting can be complex, difficult to understand, and influenced by the participant's competence, talents, and effort, potentially increasing their MWL [16].

While physiological measures, such as electrooculography (EOG) [17], [18], electrocardiogram (ECG) [19], [20] heart rate, blood pressure, and skin conductance, are used to assess the body's physiological responses to stress [21], they have limitations. For instance, EOG and ECG are non-invasive and portable, but they are not directly related to brain activity [22], [23]. Moreover, changes in physiological measures can be caused by physical exertion, emotional arousal, and environmental stressors [24], 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) [25], functional magnetic resonance imaging (fMRI) [26], functional near-infrared spectroscopy (fNIRS) [27], and notably, electroencephalography (EEG) [17].

Each neurophysiological signal has its own set of advantages and limitations. For example, MEG and fMRI can measure 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 specialized equipment [28]. 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 blood flow and movement artifacts. 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. Among these methods, EEG is often the preferred method for measuring a subject's MWL level, particularly in the context of human-computer interaction, due to its noninvasive nature and high temporal resolution, which enable millisecond-scale measurements [29], [30]. Research also shows significant correlations between MWL and physiological factors derived from EEG data [31]. These findings enable continuous measurement and classification of MWL, facilitating the development of more responsive and adaptive IS. The ability to classify between MWL levels—low, medium, and high—is a key to understanding the effectiveness of IS [32].

Considerable progress has been made in this field as a result of previous notable review papers [33]–[41]. However, given the dynamic nature of this field, it is vital to stay updated with the latest findings, emphasizing the need for more current and inclusive research. In line with recent advances, sophisticated deep learning models have been designed to accurately capture variant characteristics within EEG signals, allowing for precise classification of an individual's MWL levels [42]–[46], [162]. Despite the promising potential of deep learning for classifying MWL levels from EEG signals, its application is not without several inherent limitations. Therefore, this study aims to uncover challenges and opportunities in MWL level classification from EEG signals using a deep learning model, drawing on existing literature, aiming to address the following research questions (RQs):

- **RQ1: "What input formulations have been utilized for training deep neural networks in MWL classification?";**
- **RQ2: "What cross-validation procedures are appropriate for EEG signals in the context of deep learning for MWL levels classification?";**
- **RQ3: "What types of MWL classification problems have been addressed using deep learning techniques?"**

To investigate our RQs, we have gathered and analyzed peer-reviewed publications focusing on deep learning for EEG-based MWL level classification to bridge this knowledge gap. Our review includes an examination of signal preprocessing, feature engineering, and model training methodologies employed in these studies. In addition, we discuss prospects, challenges, and directions for improvement in this area. The application of deep learning to EEG signals is usually hindered by small sample sizes, the absence of standardized protocols for data preprocessing, the lack of diversity in study populations, and difficulties with feature extraction and model training.

Through our systematic review, we have provided a comprehensive picture of the state-of-the-art MWL assessment and classification, as well as its applications in information systems. Our review has also identified a notable inconsistency in methodologies, particularly in the EEG data preprocessing step. Furthermore, essential aspects of model training, especially the implementation of cross-validation techniques for EEG-based MWL classification—a critical step in machine learning techniques—have not been sufficiently addressed in existing research. Additionally, there is a lack of categorization and explanation of MWL classification from EEG signals using deep learning model problems; these problems include within-subject, cross-subject, cross-session, cross-task, and combined cross-task and -subject issues. Their understanding is pivotal as they are intricately linked to the methodology, influencing the choice of cross-validation strategies used in model evaluation. This comprehensive approach will enable a deeper understanding of MWL classification using EEG signals, paving the way for more accurate and reliable research in this IS field.

II. Method

In the initial stages of our research procedure, we developed comprehensive search strategies for each database. Recognizing the distinct characteristics and capabilities of each platform, we adapted our strategy to maximize their advantages.

A. Search Strategies

The search strategy for each database is customized based on specific attributes. Particularly, some databases do not support Boolean operators, exact phrase searches, or wildcard usage and impose restrictions on string length, the number of search terms, and Boolean logic usage. Consequently, we modify the search terms for each database to align with their unique search procedures. The comprehensive search technique is detailed in Table I.

TABLE I
The Search Terms for Each Database

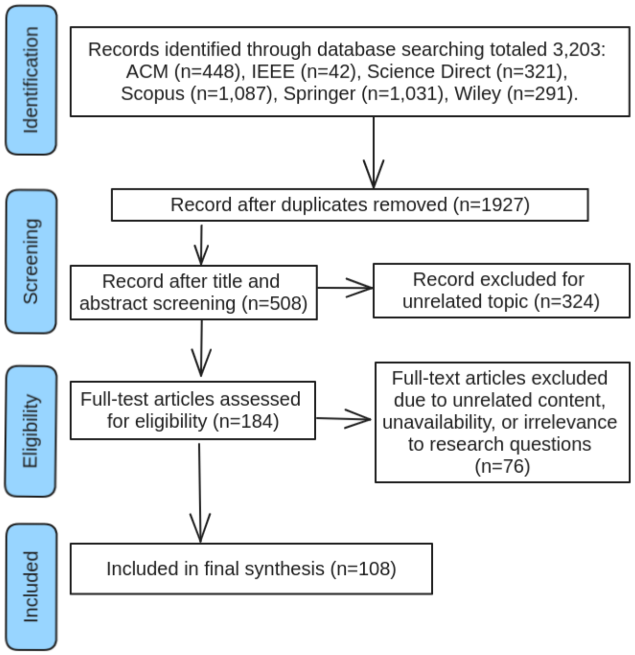
Databases	Search Terms
ACM Digital Library	AllField:(“deep learning” OR CNN OR RNN OR LSTM OR GRU OR “Reinforcement learning” OR Transformer*) AND AllField:((cognit* load*) OR (information load*)) AND AllField: (“EEG”) AND AllField: (Classif*)
IEEE Xplore	(“deep learning” OR CNN OR RNN OR LSTM OR GRU OR “Reinforcement learning” OR Transformer*) AND ((cognit* load) OR information load)) AND (“EEG”) AND Classif*
ScienceDirect	(“deep learning” OR “Reinforcement learning” OR Transformer) AND (cognitive load OR information load) AND (“EEG”) AND Classification
Scopus	(ALL ((“deep learning” OR cnn OR rnn OR lstm OR gru OR “Reinforcement learning” OR transformer*)) AND ALL (((cognit* AND load*) OR (information AND load*))) AND ALL ((“EEG”)) AND ALL (classif*))
Springer Link	(“deep learning” OR CNN OR RNN OR LSTM OR GRU OR “Reinforcement learning” OR Transformer*) AND (cognit* load OR information load) AND (“EEG”) AND Classif*
Wiley Online Library	“(“deep learning” OR CNN OR RNN OR LSTM OR GRU OR “Reinforcement learning” OR Transformer*) ” anywhere and “(cognit* load* OR information load*)” anywhere and “(“EEG”)” anywhere and “Classif*” anywhere

B. Inclusion and Exclusion Criteria

To ensure that the research focused on the effectiveness of deep learning models in classifying MWL levels using EEG signals, the eligibility criteria were based on the study's objective, methodology, and publication date. The primary aim was to investigate how deep learning models can be utilized for this purpose; only scholarly articles reporting original research were considered. Articles on the proposed devices, review articles, encyclopedia entries, book chapters, conference abstracts, editorials, short communications, software publications, and articles without full texts or abstracts were excluded to maintain consistency. These databases—ACM Digital Library [1], IEEE Xplore [2], ScienceDirect [3], Scopus [4], Springer Link [5], and Wiley Online Library [6]—were used to retrieve the papers.

C. Data Extraction and Quality Assessment

Fig. 1, PRISMA workflow summarises data extraction and quality assessment during the literature search process. The initial search yielded 3,220 articles. After excluding non-research articles, 1,927 remained. Following title and abstract screening, 508 articles were identified as EEG-related, and 184 were found relevant to the research questions. After full-text screening, 108 articles were found relevant to the research questions and were included in the study.


Identification
Records identified through database searching totaled 3,203: ACM (n=448), IEEE (n=42), Science Direct (n=321), Scopus (n=1,087), Springer (n=1,031), Wiley (n=291).
Screening
Record after duplicates removed (n=1927)
Record after title and abstract screening (n=508)
Record excluded for unrelated topic (n=324)
Eligibility
Full-test articles assessed for eligibility (n=184)
Full-text articles excluded due to unrelated content, unavailability, or irrelevance to research questions (n=76)
Included
Included in final synthesis (n=108)

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://www.sciencedirect.com/
[4] https://www.scopus.com/search/form.uri?display=basic#basic
[5] https://link.springer.com/
[6] https://onlinelibrary.wiley.com/

Fig. 1. Prisma flow diagram for the systematic review detailing database searches, the number of abstracts screened, and the full texts retrieved.

III. Literature Review

This study aims to advance our knowledge of EEG-based MWL classification in deep learning by investigating various neural network structures, training inputs, MWL issues, and appropriate cross-validation techniques. This section will start with the background of MWL, including its assessment and classification. We will also explore the interplay between MWL and IS to provide a comprehensive context. Following this foundational overview, we will proceed with a thorough systematic review of the relevant literature

A. Mental Workload (MWL)

The notion of MWL is understood by many, yet it can be challenging to articulate [47]. As a concept, MWL is essential in comprehending the cognitive demands placed on individuals during task performance. MWL is closely associated with stress and strain, reflecting two aspects of our interaction with challenging tasks [48]. 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 [49], [50]. 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 [51], [52]. Therefore, achieving an optimal balance between demands and cognitive resources is crucial when managing MWL effectively. This is because the MWL significantly impacts cognitive strain, which 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 to the design of educational programs [53], [54], from driving fatigue [2], [3] to performance across a broad spectrum of fields [4]. Research studies, such as Young et al. [15], have shown that excessive workloads often lead to decreased performance and increased errors. This is consistent with Kahneman's resource model [55], which suggests that our cognitive resources are limited. The graph in Fig. 2 illustrates the relationship between MWL and performance. 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. To understand the intricacies of MWL, it is crucial to employ precise measurement tools. Methods such as analyzing performance metrics and administering questionnaires have advantages and limitations. However, EEG signals have recently emerged as a preferred method for assessing MWL levels in human-computer interaction contexts. This is largely due to its non-invasive nature and high temporal resolution, allowing millisecond-level measurements [29]. Additionally, EEG signals have been found to correlate highly with a person's real-time MWL status [56], making it a valuable measuring tool for MWL research. However, these signals are often noisy and time-varying, which poses challenges for EEG-based MWL assessments.

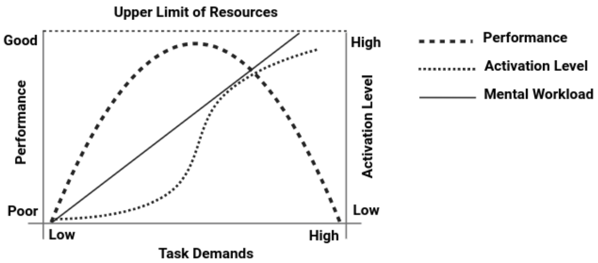


Fig. 2. The relationship between activation level, workload (task demands) and performance (adapted from de Waard 1996)

B. MWL Assessment

1) **Performance Metrics**

Measuring an individual's MWL is crucial, particularly in safety-critical scenarios such as driving. Typically, this involves assessing task performance, which is vital for evaluating the effectiveness and efficiency of an individual's abilities. Direct task performance measures are useful for determining an individual's MWL by assessing their performance on the primary task. For example, in a driving scenario, errors in steering or inconsistencies in following distance can indicate a higher MWL. 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 MWL increases on the primary task, performance on the secondary task declines [15]. 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 [57] is a secondary task measure of MWL 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 MWL is high, their ability to process peripheral information decreases. So, if the participant's MWL 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 MWL the participant is experiencing [15]. 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 MWL, even without phone use, as indicated by significantly poorer performance [50]. Another widely used method for measuring participants' MWL is cognitive tasks such as the n-back task [58]. The n-back task is a commonly used tool for MWL 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 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 MWL [59]. 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 [60].

2) Subjective Assessment

A self-report questionnaire is another method for measuring MWL, unlike an objective measure, which infers workload from task outcome. NASA-TLX [13], [61] is a widely used questionnaire that helps evaluate participants' workload 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 [10] 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 MWL assessment is the Workload Profile (WP) questionnaire [14]. This tool evaluates MWL 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 offer valuable qualitative feedback on a participant's experience, it is inherently subjective and can be influenced by factors such as the participant's mood, willingness to provide honest responses, and ability to self-assess. Furthermore, self-reporting may increase participants' MWL, especially in studies where participants are required to rate their MWL level after completing a task and immediately engaging in another task.

3) Physiological Measures and Neurophysiological Measures

Various physiological measurements are commonly used to assess MWL. For example, electrocardiac and cardiovascular activity can be measured from 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. [62] found that HR can be a suitable indicator of MWL under certain conditions. During the study, participants performed various 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 without 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 [63]. Although some studies have shown increased blood pressure with harder tasks, others have reported mixed results [21]. 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 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 [64]. 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 [64].

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 [65], 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 [66].

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. The signal used such as EOG [17], [18] and ECG [19], [20], as alternative measurement methods. Specifically, brain signal activity has been evaluated using various neuroimaging techniques such as MEG [25], fMRI [26], fNIRS [27], and notably, EEG [17].

Each neurophysiological signal has its own set of advantages and limitations. For instance, EOG and ECG are non-invasive and portable, but they are not directly related to brain activity [22], [23]. On the other hand, 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 specialized equipment [28]. fNIRS, relatively inexpensive and portable, can measure brain activity in different brain regions. However, it has low spatial resolution and is prone to blood flow and movement artefacts.

EEG, which is relatively inexpensive and portable, can measure brain activity with high temporal resolution, making it ideal for detecting subjects' MWL levels in real-time. Among these neurophysiological signals, EEG is frequently preferred in human-computer interaction contexts due to its non-invasive nature and high temporal resolution, allowing for millisecond-scale measurements [29]. Its popularity is further enhanced by its strong correlation with a person's real-time MWL status [56].

C. 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 III-B2, 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 MWLs 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 [67], 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 analyzed to reveal patterns in blood flow and respiration 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 categorize subjects' MWL. For instance, in one study [68], they utilized both task design (the 1-hour computerized letter recognition task) and questionnaires (the visual analogue scale of fatigue and the NASA-TLX) to categorize MWL. The task design induced an MWL 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 and 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 participants' subjective experience, which can vary individually. Researchers can create models that predict MWL levels based on physiological or neurophysiological data by combining subjective and objective measures. These models can be more accurate because they consider the individual variability in physiological or neurophysiological responses to MWL. This can be useful for tailoring assessments to the individual and for training classification models.

D. MWL and Information System

MWL, or the extent of cognitive resources required to complete a task, can range from low, requiring minimal cognitive effort, to high, exceeding an individual's cognitive capacity, potentially resulting in errors, a decline in performance, and stress [61], [69]. In the context of IS, which consists of hardware, programs, data, procedures, and people that operate together to produce information that supports the operation and management functions of an organization [70], addressing MWL is essential for enhancing user interfaces, optimising algorithms, and promoting usability [32]. A search interface with excessive information or intricate layouts can produce a high MWL [71], leading to user frustration and hindering search efficiency [72]. On the contrary, a well-designed search system within an IS that swiftly and accurately produces relevant results can diminish MWL, thus enhancing user satisfaction [73]. Understanding user MWL during the use of IS is crucial to improving overall system efficiency, ensuring universal accessibility, and driving beneficial design changes. Several challenges, such as inefficient information retrieval, complex user interfaces, irrelevant search results, slow response times, privacy concerns, and inaccurate or incomplete data, can intensify MWL [74]. These factors underline the necessity for continuous refinements in the design and functionality of IS [75].

Effective management of MWL within the scope of IS is crucial in optimizing user interfaces, enhancing algorithms, and elevating user satisfaction [76]. In the broader context of designing and operating adaptive systems, managing MWL becomes even more crucial, as it directly influences user experience and system performance. Emerging technologies, such as EEG and deep learning, offer great promise for the future. These tools can predict a user's MWL, enabling IS to be better adapted in real-time to the cognitive state and needs of the user.

E. Data Preprocessing

EEG features often display significant differences in scale, which can introduce bias in subsequent analyses. For example, the disproportionately large magnitude of theta waves compared to gamma waves may lead to an undue influence on model outcomes [77]. Additionally, the wide variability in EEG feature values can negatively affect the model classification performance [78]. Although data scaling techniques have been employed to transform features to a unified scale, there is no consensus on the most suitable scaling method. Researchers have utilized various techniques, such as z-score standardization [4], [53], [79]–[81], robust scaler [82], decibel (dB) conversion [77], and innovative algorithms, such as filter bank common spatial pattern (FBCSP) and optimal spatial-temporal pattern (OSTP) [83]. Nonetheless, the absence of a universally accepted approach and rigorous comparisons between these techniques may limit the generalizability of the findings. This led us to examine the general procedure of feature engineering, which is essential for creating effective models.

To answer **RQ1**: "**What input formulations have been utilised for training deep neural networks in MWL classification?**", this section focuses on the various strategies and deep learning models used for feature extraction, along with relevant references. In situations where a model serves both extraction and classification functions, feature extraction is prioritised.

1) **Data Scarcity in EEG-based MWL Levels Classification**

To train deep neural networks for MWL classification, thorough consideration must be given to data preparation, including data transformations and feature extraction, as these constitute the final input dataset. The limited availability of EEG-based MWL sample datasets presents substantial difficulty. Confidentiality and ambiguous factors inhibit the sharing of laboratory-collected EEG data, frequently resulting in a deficiency of models. This lack of data can result in subpar performance and over- or under-fitting problems.

While data augmentation techniques, such as shift, scale, rotation, and reflection, are commonly employed in machine learning to mitigate the problem of inadequate data, these methods are incompatible with continuous EEG signals. They tend to disrupt the signals' temporal characteristics [84]. Moreover, adding noise is also inadvisable due to the high randomness and temporal variability of the signals, potentially leading to local reformatting of EEG data.

Upon reviewing recent literature, several innovative approaches emerge, each attempting to tackle the issue of limited EEG samples differently. Sun et al. [85] developed a shallow version of a CNN called WLnet to detect EEG signal patterns. Compared with algorithms such as common spatial pattern feature extraction, temporally constrained sparse group spatial pattern feature extraction, and EEGnet, WLnet outperformed them in terms of detection accuracy under stress and non-stress conditions. Zhang et al. [86] utilized transfer learning as an alternative approach to address the problem of insufficient EEG data. Using a pre-trained Inception-v3 model, a CNN designed for image analysis and object detection allowed them to extract relevant features from the EEG data. They converted the 1D EEG signal into image-like data using a recurrence plot (RP) to train the model. While this study demonstrates the potential of transfer learning for MWL classification, it raises questions about the efficiency and scalability of converting 1D signals to image-like data and the possible loss of information during this process. Chavarriaga et al. [87] underscored the difficulties associated with labelling cognitive state data, arguing that existing computational models with limited label information for engagement assessment perform poorly due to overfitting. They explored a method that involved pre-training several deep learning models with unlabeled data and fine-tuning them with different proportions of labelled data (top 1%, 3%, 5%, 10%, 15%, and 20%). The goal was to discover new representations for engagement assessment. The results showed that the engagement assessment performance of the new data representations was comparable to the original EEG characteristics.

A thorough analysis of these studies reveals a progression in developing novel strategies to address the challenges posed by limited EEG samples in MWL classification. Approaches such as developing shallow CNNs, transfer learning, and pre-training with unlabeled data offer promising avenues for future research.

2) Deep Learning for EEG Feature Extraction

In the machine learning and deep learning domains, addressing the challenges associated with high-dimensional data, such as multi-dimensional EEG signals, is important. While recent studies have made significant improvements to mitigate computational demands by feature engineering and identifying robust sets of features that enhance model performance in subsequent analyses, this review critically analyses these approaches, identifying their benefits, limitations, and possibilities for future research. This study seeks to uncover technological advances, methodological biases, and challenges in current EEG feature extraction methodologies.

Firstly, manual feature extraction is the traditional method that people in the machine learning area have used. However, one of the biggest challenges of this method is that it is time-consuming and labour-intensive in practical applications. Nevertheless, this approach facilitates a critical assessment of which feature set and classifier best suit a specific dataset [88]. Hand-crafted feature engineering relies heavily on meticulous preprocessing work and advanced domain knowledge [89], making model performance dependent on the quality of feature selection techniques. The types of features extracted vary across studies, illustrating the adaptability of these methodologies. Mohamed et al. [77] concentrated on time- and frequency-domain features, while Diaz et al. [90] focused exclusively on frequency-domain and predefined features predicated on the potential of the theta frequency band for assessing MWL. Similary, Wu et al. [91] utilized the wavelet packet transform to decompose EEG signals into gamma, theta, alpha, and beta bands, employing the combined representation of the power spectrum curve area as the optimal features for assessing pilots' mental status. Likewise, Chen et al. [92] implemented Wavelet packet decomposition to identify the most promising frequency band for evaluating mental load and incorporated the Hilbert-Huang transform algorithm in their analysis. Consequently, these techniques enable comprehensive exploration and understanding, albeit at the expense of time and potentially losing some pertinent data.

a) CNN for EEG feature extraction

To avoid the mentioned problems, recent advancements are focusing on using end-to-end deep neural networks with self-adaptive feature learning capabilities. These advancements aim to address the challenge of automatic feature recognition more effectively. CNN, renowned for its success in image classification tasks, has emerged as a prevalent choice in EEG signal processing. CNN filters, serving as feature detectors, have shown effectiveness in directly capturing features from EEG signals. However, their application in EEG processing, as highlighted in Cao et al.'s study [93], needs careful analysis of the signal's distinctive properties compared to standard image data. In a comparative analysis, Almogbel et al. [80] used raw EEG signals, without pre-processing, as input for a CNN model. The capacity of this model to automatically extract relevant information from EEG data is a major advancement in the identification of different levels of cognitive workload.

b) 1D, 2D, and 3D CNNs

Zhang et al. [79]'s introduction of a one-dimensional CNN (1D-CNN) to capture frequency band information from EEG signals represents a pivotal development. Their architecture consists of various filter lengths, which capture information from different frequency bands. In this work, the authors posited that automatic feature extraction is superior to hand-crafted feature engineering because all the useful information is retained in the original data without distortion. The 1D-CNN model has also been employed in several studies [3], [53], [82], [94]–[96] to discover patterns in EEG signals and automatically assess the cognitive state of subjects. While 1D-CNN shows proficiency in capturing frequency band information, its comparability to multidimensional CNNs in terms of accuracy and efficiency remains under-explored. The adaptability of 2D and 3D CNNs opens up new avenues in EEG analysis and introduces complexity in data transformation and interpretation. The research by Qayyum et al. [97] utilizes a pre-trained 2D-CNN for analyzing EEG signals related to human mental states during repetitive multimedia learning tasks. To accommodate the 2D-CNN, the one-dimensional EEG signals were reshaped into two dimensions. This transformation was achieved through the short-time Fourier transform method, which enabled the 2D-CNN to extract significant information. The alpha brain wave, a key feature, was observed to display a uniform pattern across various cognitive tasks. Meanwhile, Kwak et al. [98], [99] reshaped the one-dimensional EEG signal into a 3D EEG image to allow the constructed 3D CNN to learn spectral and spatial information over the scalp. In this approach, multilevel features are retrieved in each layer of the proposed model, and each extracted feature is multiplied by a weighting factor to determine its usefulness in predicting the target variable.

c) Hybrid CNN architectures

Diverse aspects of signal analysis have been addressed by a number of CNN-based designs proposed in the literature to improve feature extraction. While some models, such as 1D-CNN, focus solely on the temporal aspect of the signal, others, such as 3D-CNN, take spatial and spectral aspects into account. Nevertheless, it is essential to recognise that existing models have certain limitations. In light of these limitations, researchers have embarked on the development of hybrid CNN architectures to capture EEG signals more effectively. For example, a ternary-task convolutional bidirectional neural turning machine (TT-CBNTM) for analysing the cognitive states of subjects was proposed in [100]. The TT-CBNTM is built around two basic model architectures: CNN and bidirectional neural turning machines (BNTM). The suggested model considers EEG variables from spatial, spectral, and temporal dimensions. The CNN section examines the spatial and spectral characterization of EEG, after which the recovered temporal information is supplied to the BNTM part. A neural network architecture that combines CNN and BiLSTM networks was proposed by Dewan et al. [101] to capture both spatial and temporal features from sequential data, such as time series or spatial data. The main idea behind combining CNN and BiLSTM in Conv-BiLSTM-NN is to leverage the strengths of both models. CNNs excel at extracting spatial patterns and features, while BiLSTM networks are proficient in modelling sequential dependencies and capturing temporal dynamics. By combining these two components, Conv-BiLSTM-NN aims to effectively capture both spatial and temporal information, enabling comprehensive analysis of the input data. In addition, Zhang et al. [102] also investigated the spatial and temporal dimensions of EEG signals and introduced a novel model, a two-stream neural network (TSNN), to autonomously integrate spectral and temporal EEG data. The results demonstrated that the TSNN extracted meaningful information from both EEG dimensions, thereby enhancing the evaluation of MWL. The hybrid CNN architectures represent an innovative fusion of methodologies and provide promise in integrating spatial, spectral, and temporal data, yet their real-world applicability and implementation complexity merit further investigation. The development and application of various CNN architectures, including 1D, 2D, and 3D CNNs, as well as hybrid models, have significantly advanced the field of EEG signal analysis. These models enable researchers to automatically extract features and assess the cognitive state of subjects more accurately and efficiently than traditional hand-crafted feature engineering methods.

In addition to developing and applying various CNN architectures that have significantly advanced EEG signal analysis, researchers have also explored alternative approaches to enhance feature learning further. A novel restricted Boltzmann machine (RBM) architecture has been employed in [103] and [84] for unsupervised feature learning of EEG signals, aiming to identify salient features for categorization. To pre-train the RBM model, the mean absolute difference (MAD) features were utilized. Building upon the need for improved spatial precision, Chakladar et al. [104] highlight the issue of limited spatial precision in EEG signals, leading to suboptimal classification results. To address this, they propose a deep variational autoencoder (VAE) combined with a spatial attention-based approach (CBAM) to improve EEG spatial resolution and derive noise-free robust features from latent space for better classification. CBAM extracts spatial-channel level attention features from localized VAE signals in topographical videos. Similarly, Saha et al. [105] emphasize that EEG signals are channel-based temporal-spatial signal sequences, and some brain regions are more deeply involved than others, resulting in EEG oscillations that require further research and improvement. To tackle these issues, they introduce a region-dependent and attention-driven bi-directional long short-term memory (RA-BiLSTM) to encode region-level features for classification.

The multi-branch long short-term memory with hierarchical temporal attention (MuLHiTA) [106] was designed for detecting MWL at an early stage. This model employs a unique strategy that enables parallel processing of interslice and interslice EEG samples by integrating two attention modules that are mutually advantageous. More precisely, the model was designed to analyze both the characteristics within individual segments of the EEG signal (interslice features) and the relationships or differences between different segments of the EEG signal (interslice features). This can provide a more comprehensive analysis of the EEG data.

Yin et al. [107] put forward a transfer learning-based method to extract the dynamic properties inherent in EEG signals. This approach uses a large EEG dataset, gathered in the context of emotional stimuli, to enhance the model's training stability with the help of a transferred MWL classifier. In another development, a unique feature creation network reported in [108] was able to identify the dynamic centre-based binary pattern (DCBT) and a multi-threshold ternary pattern (MTTP) within EEG signals. To further understand the interconnections among signal channels, frequency-domain features were employed, leading to the proposal of multi-channel networks and multi-threshold networks, as indicated in [109] and [110], respectively. This entailed the transformation of EEG time data into the frequency domain using spectrograms, which paved the way for the development of multi-channel and multi-threshold networks based on spatial distances. The networks were classified according to their structural properties. In addition, Ahmadi et al. [111] developed a novel characteristic known as the Gaussian copula mutual information (GCMI). This characteristic, computed from wavelet EEG coefficients, served to determine the relationships between various brain regions. Nevertheless, the fine-grained and multi-scale motif (FGMSM) method for feature extraction from raw EEG data was introduced by Shao et al. [112]. The method involves identifying patterns in the data at various time scales. By doing so, FGMSM aims to extract more pertinent information from the EEG signals, thereby enhancing the model's capacity to account for cross-subject variations.

The utilization of deep learning methods, particularly CNN and their variants, for EEG feature extraction, has shown promising results, as these techniques enable automatic and robust feature recognition. Hybrid CNN architectures and other advanced models, such as the restricted Boltzmann machine (RBM) and variational autoencoder (VAE), have further enhanced the process by accounting for spatial, spectral, and temporal dimensions of EEG signals. These advancements signify a paradigm shift from traditional manual feature extraction to automated approaches, enabling more efficient and comprehensive analysis of EEG data to evaluate subject's MWL levels.

In conclusion, feature extraction in EEG signal analysis is trending towards more sophisticated, automated techniques that take advantage of the most recent developments in deep learning. To be more precise, the new model aims to undertake a more thorough analysis by capturing spatial, spectral, and temporal dimensions [104]. In contrast, the previous model only captures one dimension, such as the frequency band, which is spectral dimension [79]. This change creates new opportunities for investigation and use in studying intricate brain functions. These studies collectively underscore the dynamic and expanding field of EEG research, showcasing the potential of advanced computational techniques to enhance understanding and application of brain activity data.

TABLE II
OVERVIEW OF INPUT FORMULATIONS IN DEEP NEURAL NETWORK FOR MWL CLASSIFICATION

Input Formulation	References
Raw EEG Signals	[80]
Transformed EEG Signals using CNNs	
□ 1D-CNN (Time-series Data)	[79], [94], [53], [95], [96], [3], [82]
□ 2D-CNN (Spectrograms)	[97]
Raw EEG Signals	[80]
□ 3D-CNN (3D EEG Images)	[98], [99]
□ CNN Hybrid Models (CNN-LSTM)	[100], [101], [102]
Feature-based Inputs	
□ Time/Frequency Domain Features	[77], [90]
□ Wavelet Transform Features	[91], [92]
□ Multi-Channel/Threshold Features	[109], [110]
□ Features via Transfer Learning	[86], [107]
Advanced Model Features	
□ Unsupervised Learning (RBM)	[103], [84]
□ Spatial Precision (VAE with CBAM)	[104]
□ Temporal-Spatial (RA-BiLSTM)	[105]
□ Temporal Attention (MuLHiTA)	[106]
□ Transfer Learning (Dynamic Properties)	[107]
□ Dynamic Pattern Features	[108]
□ Fine-Grained Motifs (FGMSM)	[112]

Table II provides a concise overview of various input formulations used in deep neural networks for MWL classification. It highlights the progression from basic raw EEG signals to sophisticated model features, illustrating the evolving application of deep learning in EEG data analysis.

F. Cross-Validation in MWL Classification

To answer **RQ2: "What cross-validation procedures are appropriate for EEG signals in the context of deep learning for MWL level classification?"**, this section will discuss the various cross-validation strategies used in the context of EEG data and deep learning. In the case of EEG data, the proper cross-validation technique can considerably improve a model's predictive ability, making it a crucial step in classifying MWL levels.

Cross-validation is an essential technique for evaluating deep learning models and assessing their performance [113]. Various cross-validation methods have been developed, including Hold-out, K-folds, Leave-one-out, Leave-p-out, and Repeated K-folds, each with its algorithm. The choice of cross-validation method depends on the experimental goal, which may be subject-dependent, task-dependent, or session-dependent.

1) **Hold-Out Cross-Validation**

Starting with the hold-out cross-validation, this straightforward technique splits the dataset into training and testing parts, typically in an 80:20 ratio. The model is trained and validated only once, with the limitation that performance may be inaccurate if the training and testing datasets have different characteristics. Fig. 3 demonstrates the data splitting process into training and testing sets employed in hold-out cross-validation. For a more robust evaluation, researchers often turn to K-fold cross-validation.

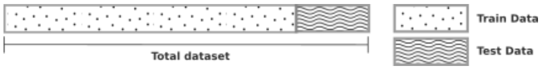


Fig. 3. Hold-out cross-validation technique

2) K-fold Cross-Validation

Building upon the hold-out cross-validation technique, the *K*-fold cross-validation technique randomly divides the dataset into *K* equal non-overlapping parts or folds. Common values for *K* are 5 or 10, with one fold reserved for testing and the remaining *K* □ 1. As illustrated in Table III, the K-fold cross-validation technique appears to be the most popular, and various studies have utilized it. However, partition strategies vary across studies. For instance, in the study by Yang et al. [4], 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. [114] 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 the combination of data from different tasks and subjects, ensuring generality across both subjects and tasks in their study [114]. Fig. 4 demonstrates the data splitting process into training and testing sets, as employed in K-fold cross-validation.

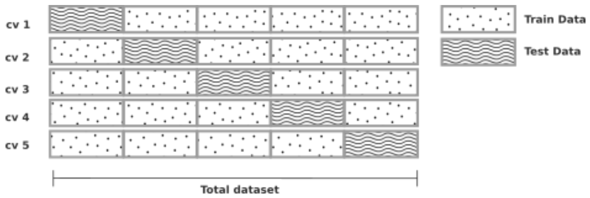


Fig. 4. *K*-fold cross-validation technique

3) Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV), a variation of the *K*-fold cross-validation, uses the number of samples in the dataset (*n*) as the number of folds (*K*). One sample or subject is selected as the test set, while the remaining *n* □ 1 samples form the training set. While LOOCV requires the building of n models instead of *K* models, thereby increasing computational overhead, its precision can prove invaluable in specific scenarios. Variants of LOOCV, such as Leave-subject-out, Leave-session-out, and leave-task-out cross-validation, have been developed to address specific experimental objectives.

a) Leave-subject-out cross-validation

The leave-subject-out cross-validation, a popular choice for cross-subject classification model evaluation, reserves data from one subject for testing while the rest are combined for training. This procedure is repeated until each subject is used at least once as a testing subject [42], [78], [115]–[118]. Fig. 5 demonstrates the data splitting process into training and testing sets, as employed in leave-subject-out cross-validation. To better suit different contexts, adaptations like leave-session-out cross-validation have been introduced.

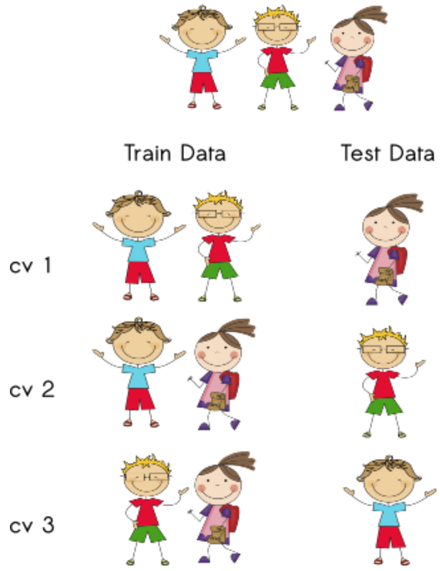


Fig. 5. Leave-subject-out cross-validation technique

b) Leave-session-out cross-validation:

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 [80], [85]. Another LOOCV variant, Leave-task-out cross-validation, focuses on tasks themselves for data separation.

c) Leave-task-out cross-validation:

This approach involves selecting training and test data from different tasks [112], [119]. 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.

4) Leave-p-Out Cross-Validation

Leave-p-out cross-validation (LPOCV), a method similar to LOOCV, reserves *p* samples/subjects for testing instead of just one. The remaining *n* □ *p* subjects are used for training [120]. 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. Other methods, such as Monte Carlo cross-validation, have been explored to address these limitations.

5) Monte Carlo Cross-Validation

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. In this approach, the model is trained for a specified number of iterations, denoted by *k*. During each iteration, the data is randomly divided into training and testing sets, which can lead to certain data points appearing multiple times in the test set or not appearing at all. This characteristic of Monte Carlo cross-validation introduces a degree of randomness that can help mitigate potential biases present in the data.

Saha et al. [105] adopted the Monte Carlo cross-validation technique with four folds, arguing that it offers higher optimization 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.

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. For instance, random shuffling may not adequately address the temporal nature of EEG signals before splitting the data into training and testing sets. If the goal is to predict a future event, such as a subject's MWL, disrupting the temporal characteristics could lead to unreliable classification model performance [121]. Fig. 6 shows the Monte Carlo cross-validation technique, which randomly splits the dataset into training and test sets multiple times.

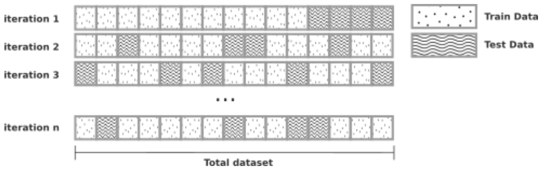


Fig. 6. Monte Carlo cross-validation technique

6) Time-Wise Cross-Validation

Considering the temporal nature of EEG signals, time-wise cross-validation has been suggested as a suitable strategy to accommodate these characteristics [123]. 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 [122]. To minimize 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. Fig. 7 illustrates the time-wise cross-validation technique, where the dataset is split based on the temporal order of the data points.

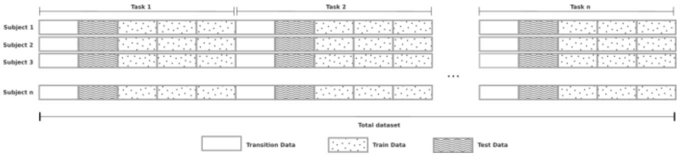


Fig. 7. Time-wise cross-validation technique

7) Time Series Cross-Validation

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 [124]. 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 generalizable models. Fig. 8 illustrates the time series cross-validation technique, where the dataset is split based on the temporal order of the data points.

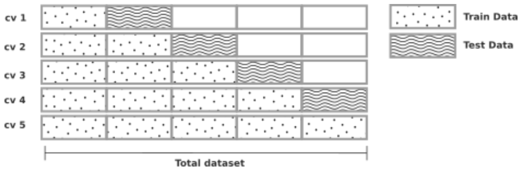


Fig. 8. Time series cross-validation technique

Cross-validation techniques aim to replicate real-world scenarios for training models. Despite the diversity of algorithms across studies, temporal characteristics of EEG signals are crucial for accurate MWL predictions. It is essential to recall that EEG signals are naturally time series as workload levels change with the introduction or modification of tasks over time. Table III provides a comprehensive overview of various cross-validation techniques employed in studies on the classification of MWL levels using EEG signals.

TABLE III
CROSS-VALIDATION TECHNIQUES USED IN THE ANALYSIS

Cross-validation technique	**References**
Hold-out	[2], [3], [53], [86], [97], [103], [54], [125], [126]
K-fold	
□ 5-fold	[79], [82], [94], [111], [127], [128
□ 10-fold	[77], [94], [96], [98], [99], [100], [102], [109]
□ Other *K*	[85], [86], [114]
LOOCV	
□ Leave-one(subject)-out	[42], [78], [115]–[118]
□ Leave-one(session)-out	[80], [85]
□ Leave-one(task)-out	[119], [112]
LPOCV	
□ Leave-p(session)-out	[120]
Monte Carlo	[105]
Time-wise	[122],[123]
Time-series	[124]

The literature review in this section focuses on the cross-validation technique used to evaluate the deep learning model for MWL classification. We have found a gap in many studies - they fail to specify how they trained their models or which cross-validation techniques they employed. Even when mentioned, the explanation is often vague, with just the technique's name provided. This lack of clarity makes reproducing these studies challenging, even when the machine learning architecture is known. Adding to the confusion is the inconsistency in using cross-validation methods for training MWL classification models with EEG data. Some studies have applied the k-fold cross-validation technique, which randomly splits EEG signals and feeds them into the model. However, this approach can disrupt the temporal integrity of EEG data, potentially leading to data leakage issues. Therefore, it is crucial to recognize that the cross-validation technique should align with the study's specific research objectives and classification problems.

G. Expanding Horizons in MWL Classification

Deep learning techniques have emerged as pivotal tools within the rapidly advancing domain of EEG-based MWL classification. In this section, we will delve deeper into the breadth of applications of these techniques, scrutinizing their efficacy and challenges and answering the **RQ3: "What types of MWL classification problems have been addressed using deep learning techniques?"**. This section offers a detailed exploration of how deep learning models have been utilized for such classification tasks. 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 the 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 Fig. 9.

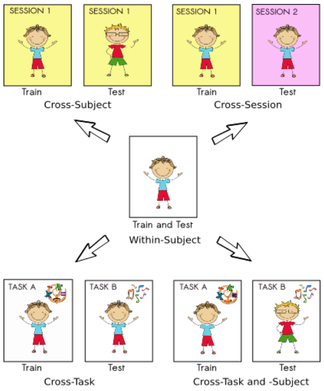


Fig. 9. EEG MWL Classification problems

1) **Within-Subject Problem**

This section focuses on the within-subject classification problem, the most popular study problem across papers related to EEG-based MWL classification. The "within-subject" approach focuses on charting the fluctuations in an individual's MWL 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 depreciation 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. This review evaluates these approaches and their contributions to the field in an effort to provide a thorough understanding of the topic. In this diverse landscape of models, traditional shallow models continue to hold their ground. 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 [126], [128], [129]. 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 inherent in EEG data.

The efficacy of CNN in extracting spatial features from EEG data has resulted in their increasing prominence in recent years [43], [46], [125], [126]. 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 [42], [119]. 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. [80] utilized 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. [130] 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. [97] employed a pre-trained 2D-CNN to categorize human mental states during recurrent multimedia learning tasks. By transforming one-dimensional EEG signals into a twodimensional 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. Stacked denoising autoencoders (SDAEs) have been introduced as a solution to reduce the dimensionality of EEG features while retaining the local information present in the data [4], [93]. 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. [131] 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 been utilized to tackle within-subject MWL classification problems. These include RNN [42], BiLSTM [45], [132], AConv-BiLSTM-NN [101], and BiLSTM-LSTM [43]–[45], [149]. 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) [43], bidirectional gated recurrent unit (BiGRU) [43], and a combination of BiGRU with GRU (BiGRU-GRU) [43]. Long Short-Term Memory (LSTM) networks [44] 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.

The within-subject classification problem in EEG-based MWL assessment has been tackled using an array of model architectures and algorithms. 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.

2) **Cross-Subject Problem**

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 normalization or standardization of features to counterbalance individual differences.

However, effectively transferring the EEG analysis model of existing subjects to other subjects' EEG signals has proven to be a complex task [116]. In response to this challenge, Hefron et al. [78] 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 an increase in classification accuracy and a decrease in variance across different participants, which underscores its potential effectiveness for addressing the cross-subject problem in EEG-based MWL classification. Meanwhile, Zheng et al. [115] 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. [117], 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 personalize a classifier for each subject, resulting in higher accuracy compared to other deep unsupervised domain adaptation (D-UDA) methods.

Jimenez et al. [133] also addressed the issue of disparate EEG signal distributions among different subjects by proposing a custom domain adaptation (CDA) method integrating adaptive batch normalization (AdaBN) and maximum mean discrepancy (MMD) into two separate deep neural networks. This method aimed to reduce both marginal and conditional distribution differences and personalize a classifier for each subject, achieving higher accuracy than other deep unsupervised domain adaptation (D-UDA) methods. Yin et al. [134] developed a switching deep belief network with 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. [2] employed a gradient boosting-based classifier, LightFD, 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 rendered it especially suitable for real-time EEG mental state prediction, underlining its potential utility in real-world applications. In a parallel effort, Shao et al. [112] 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. [116] utilized a domain-adversarial neural network (DANN), a model that has demonstrated superior performance in transfer learning, notably in document analysis and image recognition. However, it was not previously applied directly in 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 endeavored 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. In a related advancement, Ma et al. [135] further refined this domain by introducing a dynamic threshold distribution domain adaptation network (DTDDAN). This innovative model leverages domain discrimination and Jensen-Shannon loss to effectively address individual differences, thereby enhancing the learning of invariant EEG features for cross-subject driver fatigue recognition.

The cross-subject challenge in EEG-based MWL classification is primarily rooted in inter-individual variability. Specifically, the difference in data distribution from one subject to another presents a considerable obstacle to machine learning, particularly in training and applying models across these varying distributions. To address this, the researchers have developed a range of models tailored to navigate this variability effectively. Despite ongoing research, there's a notable trend in the current literature towards single-session studies, indicating a significant gap in developing cross-session models with broader generalizability and relevance. The next section will explore MWL classification through the "cross-session" approach to fill this research gap.

3) Cross-Session Problem

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 [136].

As discussed in the III-G2, 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 generalization. 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. For example, Yin et al. [81] 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 generalizability and applicability of EEG-based MWL classifiers across multiple sessions.

4) Cross-Task Problem

In Section III-G2 and III-G3, we examined the obstacles related to "cross-subject" and "cross-session" variations, respectively, in EEG-based MWL classification. An equally significant hurdle is the "cross-task" challenge. The complexity of this issue arises from the unique EEG patterns produced by different tasks. The goal is to devise models capable of assessing MWL across various tasks, utilizing training data from one task and applying it to test data from a different task performed by the same individual. While this approach promises wide applicability across tasks, its complexity is underlined by the distinct MWL types and levels different tasks evoke, resulting in varied EEG signatures.

In the realm of EEG-based emotion recognition, Li et al. [137] provide a structured literature review, categorizing transfer learning research into cross-subject, cross-session, and, crucially, cross-task domains as well. However, they explicitly identify a significant research gap: the absence of cross-task transfer learning studies directly relevant to EEG-based emotion recognition. This gap starkly contrasts with our research in the MWL domain, where cross-task applications are actively being pursued and developed. For instance, Lim et al. [138] investigated the application of consistent frequency features (alpha, beta, and theta) across diverse cognitive workload tasks using two distinct datasets. The first, a multitasking dataset, involved participants in simultaneous activities designed to induce three levels of MWL. The second, a Stroop test dataset, assessed MWL through a psychological test where participants name the color of a word that may spell a different color. They employed Transfer Component Analysis (TCA) for domain adaptation to reduce dataset distribution differences and enhance classification effectiveness. Despite its theoretical promise, their findings revealed an average classification accuracy of only 30.0%, a figure above chance levels but still insufficient for practical applications. This underscores the need for more advanced methods to improve cross-task performance. Alternatively, Shao et al. [112] introduced a novel concatenated structure combining deep recurrent and 3D convolutional neural networks (R3DCNNs) to learn EEG features across tasks. This method converted 1D EEG signals into 3D representations, enabling the R3DCNNs model to capture EEG features from spatial, spectral, and temporal perspectives. This approach showed potential in the binary classification of low and high MWL levels across tasks, marking progress in the cross-task challenge. However, the computational demands of this advanced model remain a concern. Other studies, such as those by Zeng et al. and Zhou et al., have also shown potential for tackling this issue [54], [114] by using CNN-based models. Furthermore, Kirchner et al. [139] developed two different classifiers to detect single-trial event-related potentials (ERP), with two types of transfer cases focused on applying models trained on one task to another task. This approach suggests a growing interest in addressing the cross-task problem in MWL classification.

Despite the complexity of the challenge, the literature review suggests that limited progress has been made in addressing it. Developing innovative neural network architectures that can discern EEG signals from various tasks is crucial, as current methods for detecting MWL are primarily laboratory-based and task-specific. However, cognitive overload can occur in diverse scenarios. Therefore, translating these models to practical applications could bridge the gap between laboratory settings and real-world environments, providing a more versatile approach to detecting workload levels across different tasks.

5) **Cross-Task and -Subject Problem**

A combination of "cross-task" and "cross-subject" approaches presents a significant challenge yet promises the highest level of robustness and generalizability. This model is expected to predict MWL across various tasks and individuals. This problem has only been tackled by a few researchers. A significant contribution to this topic was made by Zeng et al. [114], who developed two CNN-based EEG classifiers, EEG-Conv and EEG-Conv-R, to identify drivers' MWL. The EEG-Conv model is based on a traditional CNN architecture, while EEG-Conv-R combines the CNN approach with deep residual learning to enhance the model's 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.

In exploring the role of deep learning in EEG-based MWL classification, this section addresses the variety of classification problems tackled by these techniques. Deep learning's versatility is evident in managing the variability inherent in subject, session, and task data. The classification challenges are broadly categorized into "within" and "cross" groups, reflecting deep learning's capacity to adapt to complex EEG analysis requirements. For an overview of deep learning models applied across these MWL classification scenarios, see Table IV, highlighting the approaches and their efficacy in addressing specific challenges.

TABLE IV
DEEP LEARNING APPROACHES TO MWL CLASSIFICATION PROBLEMS

MWL Classification Challenge	**Deep Learning Approaches and References**
Within-subject	Traditional shallow models (LDA, SVM, KNN, Random Forest) [126], [128], [129], CNN-based models [43], [46], [125], [126], Hybrid CNN-LSTM models [42], [119]
Cross-subject	MPCRNN [78], ELM-based ensemble (ED-SDAE) [115], Deep neural network with GRU [117], Custom domain adaptation (CDA) [133], SDBN for OFS assessment [134], LightGBM framework (LightFD) [2], BiLSTM model [112], DANN and GDANN for fatigue detection [116]
Cross-session	Adaptive SDAE model [81]
Cross-task	TCA for domain adaptation [138], Deep R3DCNNs [112], CNN-based models for MWL across tasks [54], [114], ERP detection classifiers [139]
Cross-task and -subject	CNN-based classifiers (EEG-Conv and EEG-Conv-R) for MWL identification [114]

H. Synthesis of Finding in EEG-Based MWL Classification

Drawing from the existing literature, we envisage possible further combinations such as the "cross-subject" and "cross-session" methodologies, as well as the tripartite approach that incorporates "cross-subject", "cross-session", and "cross-task" elements. The dual method of "cross-subject" and "cross-session" is aimed at formulating a model capable of predicting MWL diversely across subjects and sessions. The most challenging yet potentially rewarding scenario lies within the all-encompassing "cross-subject", "cross-session", and "cross-task" approach. This ambitious strategy is designed to generate a model that can predict MWL across a spectrum of individuals, sessions, and tasks, creating a highly flexible tool with considerable real-world applicability. However, the existing literature on these complex problems is still limited. Thus far, the research community has not fully engaged with these two methodologies' inherent challenges, making them a promising avenue for future exploration and innovation in this ever-evolving 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 MWL decoders' performance. 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.

Addressing this research gap is crucial because it is the first step towards real-life application improvement. Once we can reliably capture the user's MWL while they perform the tasks, we can expedite the improvement of systems and applications. For instance, we would no longer need to wait for post-interaction feedback to evaluate its effectiveness in a search system. However, we can instead directly utilize the user's EEG signal to assess whether the system design matches the user's cognitive demands and expectations.

IV. Challenges and Opportunities for Future Research

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, frequently inspiring further investigation. This study goes beyond answering the three initial research questions noted in by uncovering challenges and gaps that offer opportunities for future studies. This section will look into the difficulties of using deep learning models to classify MWL levels from EEG signals and suggest areas for future research.

A. Dataset Diversity and Scarcity Challenge

EEG signal collection and developing 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. To overcome these challenges, increased availability of online datasets is necessary.

B. Self-Reporting Workload Challenges

In classifying MWL, we utilize neural networks that employ EEG signals as input, supplemented by labels from participant evaluations. These labels, indicative of self-reported workload levels, are gathered through post-experiment questionnaires [140]. This approach can be viewed as a secondary task [15]. To conduct post-task self-report feedback or performance evaluations, individuals must be trained to understand the instrument used for expressing their MWL [141]. These methods can increase subjects' burdens, making it harder for them to respond to new events. Future research would benefit from integrating measurements such as heart rate monitoring as labels for a more accurate reflection of the MWL of individuals. Due to the direct correlation between an increase in MWL and an increase in pulse rate, this is the case [142]–[144].

C. 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, thereby enhancing EEG signal preprocessing quality in laboratory and real-world contexts with dynamic soundscapes.

D. Enhancing Model Generalization and Minimizing Calibration Requirements

The practical utility of deep learning models for MWL estimation is based on their capacity to generalize effectively and require minimal calibration, enhancing their applicability in real-world settings. An ideal model should possess strong generalization 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. Prioritizing these attributes in model development can significantly improve the practicality and utility of deep learning models for EEG-based MWL classification.

E. Integration of Artifact Removal and Online Learning in Advanced Deep Learning Models

Future research could 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 that are 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.

F. Resource-Efficient Adaptive Modeling for Constrained Environments

Since a continuously adaptive model is needed, using cumbersome models can be inefficient in terms of energy efficiency and computational cost. Tiny machine learning (tinyML) [145] 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 algorithm and hardware co-design. The real-time control system is regarded as the modern vehicle's brain [146].

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

Exploring "cross-subject", "cross-session", and "cross-task" methodologies offer a pathway to refining EEG-based MWL classification. Integrating these approaches aims to develop versatile models capable of accurately predicting MWL across diverse subjects, time frames, and tasks. Despite its promise, current literature on these comprehensive strategies is sparse, marking an essential area for future exploration. Challenges in MWL decoding stem from physiological variabilities, session-to-session and task-to-task performance fluctuations, environmental influences, and personal attributes like gender and age. Addressing these complexities is vital for creating robust MWL predictors. Progress in this domain has significant practical implications. By effectively capturing MWL through EEG signals in real-time, systems and interfaces can be directly evaluated and improved based on cognitive demands, enhancing user interaction without relying on delayed feedback. In the future, to manage cross-subject, cross-session, and cross-task variability in MWL classification, we can leverage the co-teaching graph learning method, which is used in [147] to the approach used in EEG-based motor imagery recognition. This method enhances feature extraction and mitigates the impact of noisy data, making it a promising technique for developing versatile models that accurately predict mental workload across diverse conditions.

H. 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 [148]. 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 [107], future research should focus on developing models capable of capturing common properties found across subjects, sessions, and tasks.

I. Cross-Validation Issue in MWL Classification

In III-F, we address the use of CV techniques for evaluating deep learning models in MWL classification, identifying a critical gap in current practices. Many studies provide inadequate details on their model training and CV approaches, undermining reproducibility—especially problematic given the complex nature of machine learning models. Common methods such as k-fold cross-validation and random data division overlook EEG data's temporal and subject-specific characteristics in MWL classification, leading to potential overestimating of model performance. This oversight affects the integrity of research findings and the efficacy of MWL assessment tools. We argue for reevaluating CV strategies, advocating for methods that maintain EEG data's temporal sequence (e.g., time-series cross-validation [43]) and mirror real-world scenarios (e.g., subject-specific splits). Additionally, we recommend standardizing detailed documentation of data handling, model training, and validation processes to improve MWL research reproducibility and transparency. Establishing standardized CV guidelines could markedly enhance the field, ensuring research reliability and facilitating a unified literature for future work.

J. Other Recent Advances and Applications

Apart from investigating EEG-based MWL applications, we also delve into recent advancements in deep learning models that have significantly impacted various domains within EEG research. All of this work could be further investigated and expanded from the perspective of mental workload detection using EEG signals. Sun et al. [150] developed a novel Gating mechanism and Dilated Convolutional Neural Network (GDCNN) which efficiently decodes driving intentions from EEG signals, achieving superior accuracy compared to established benchmarks like EEGNet and DeepConvNet. On another front, Zhang et al. [151] systematically reviewed brain fingerprints as a novel biometric feature, extracting them through various neuroimaging technologies, including EEG, MRI, MEG, and fNIRS, and discussing their application in identity recognition. Furthermore, Chakladar et al. [104] introduced a Variational Autoencoder and Convolutional Block Attention Module (VAE-CBAM) based model for estimating cognitive workload from EEG, which significantly improved classification performance in mental arithmetic tasks. In the realm of emotion recognition, Liu et al. [152] compared the effectiveness of deep canonical correlation analysis (DCCA) and bimodal deep autoencoders (BDAE) across multiple datasets, demonstrating the robustness of these models under various noise conditions.

Additionally, Sun et al. [153] tackled the challenge of automatic sleep spindle detection using a convolutional neural network with a label refinement component, which helped optimize feature learning despite inaccuracies in label data. In a broad review, Gong [154] encapsulated a decade's progress in applying deep learning to EEG, highlighting applications in brain-computer interfaces and disease detection among others. The issue of seizure prediction was addressed by Zhang et al. [155], who introduced a transformer-based domain adversarial model to enhance generalization across different patient datasets. Moreover, Li et al. [156] proposed a trainable adjacency relation driven graph convolutional network (TARDGCN) to improve EEG-based emotion recognition by enhancing the correlation among multichannel EEG sets.

In an analysis of EEG data complexity, Lin et al. [157] presented a multistream 3D CNN with parameter sharing, which not only reduced the model's susceptibility to overfitting but also improved performance in tasks like lane-keeping and sleep monitoring. In clinical applications, Shahabi et al. [158] developed a hybrid model combining transfer learning, LSTM, and attention mechanisms to predict responses to antidepressants in patients with major depressive disorder, achieving exceptionally high classification accuracies. Li et al. [159] contributed a novel feedback capsule network for patient-specific seizure prediction, effectively capturing and integrating spatiotemporal properties from EEG signals. Addressing privacy concerns in motor imagery classification, Zhang et al. [160] introduced a lightweight source-free transfer learning approach, which maintained high classification performance while ensuring data privacy. Finally, Karnati et al. [161] and Wang et al. [147] respectively advanced deception detection and cross-subject motor imagery recognition using novel neural network models, demonstrating substantial improvements in accuracy and generalizability.

I. Discussion & Conclusion

This paper conducted a systematic review and meta-analysis of the current research on deep learning techniques for classifying MWL levels using EEG data and addressing associated challenges and limitations. We developed tailored search strategies based on criteria such as Boolean and string length of each database, and inclusion and exclusion criteria were summarised to ensure the relevance and accuracy of our data collection. After applying inclusion and exclusion criteria, we narrowed down our initial pool of 3,203 articles to 91 relevant articles for our study. By delving into the existing literature, we identified research gaps, set more precise goals, and investigated three key research questions on (i) the types of input formulations used, (ii) the cross-validation procedures deemed suitable for EEG signals in the context of deep learning, and (iii) the nature of specific MWL classification problems tackled. Furthermore, it outlined the encountered challenges and proposed directions for future research.

In conclusion, the measurement of MWL using EEG signals in the field of brain-computer interaction has seen growing interest and presents exciting challenges. Deep learning has shown promising results in forecasting mental effort, but its application in MWL classification varies across studies. These challenges include dataset uniqueness, self-reporting limitations, variability across sessions and tasks, artefact removal, energy efficiency, and resource constraints. To address these challenges, it is essential to establish a universal preprocessing pipeline, create models with integrated artefact removal, and consider within-subject, cross-subject, cross-session, cross-task, and cross-task and -subject variability. Incorporating temporal dynamics in cross-validation and refining self-reporting methods can improve the accuracy of model evaluation and workload assessment. Despite the challenges faced, deep learning algorithms have improved MWL classification. With further research and development, they may become more suitable for real-world applications, advancing the development of more effective brain-computer interfaces and related applications. Researchers are encouraged to build upon these findings and develop practical solutions for EEG-based MWL classification.

Acknowledgment

I extend my sincere gratitude to the Government of Thailand for funding my PhD, and to Chiang Mai Rajabhat University for their invaluable support and resources. Their combined assistance has been essential in my academic journey. Special thanks to the NeuraSearch Lab, University of Strathclyde, for providing the server resources for running my project.

References

[1] C. G. Lim, T. S. Lee, C. Guan, D. S. S. Fung, Y. Zhao, S. S. W. Teng, H. Zhang, and K. R. R. Krishnan, "A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder," PloS one, vol. 7, no. 10, p. e46692, 2012.

[2] H. Zeng, C. Yang, H. Zhang, Z. Wu, J. Zhang, G. Dai, F. Babiloni, and W. Kong, "A lightgbm-based eeg analysis method for driver mental states classification," Computational intelligence and neuroscience, vol. 2019, 2019.

[3] M. R. Islam, S. Barua, M. U. Ahmed, S. Begum, and G. 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. Springer, 2019, pp. 121–135.

[4] S. Yang, Z. Yin, Y. Wang, W. Zhang, Y. Wang, and J. Zhang, "Assessing cognitive mental workload via eeg signals and an ensemble deep learning classifier based on denoising autoencoders," Computers in biology and medicine, vol. 109, pp. 159–170, 2019.

[5] P. Thiffault and J. Bergeron, "Monotony of road environment and driver fatigue: a simulator study," Accident Analysis & Prevention, vol. 35, no. 3, pp. 381–391, 2003.

[6] P. A. Hancock, "A dynamic model of stress and sustained attention," Human factors, vol. 31, no. 5, pp. 519–537, 1989.

[7] L. M. Quiroga, M. E. Crosby, and M. K. Iding, "Reducing cognitive load," in 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the. IEEE, 2004, pp. 9–pp.

[8] J. Gwizdka, "Distribution of cognitive load in web search," Journal of the American Society for Information Science and Technology, vol. 61, no. 11, pp. 2167–2187, 2010.

[9] A. Darejeh, S. Mashayekh, and N. Marcus, "Cognitive-based methods to facilitate learning of software applications via e-learning systems," Cogent Education, vol. 9, no. 1, p. 2082085, 2022.

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

[11] K. R. Boff, L. Kaufman, and J. P. Thomas, Handbook of perception and human performance. Wiley New York, 1986, vol. 1.

[12] P. Wang, W. Fang, and B. Guo, "A measure of mental workload during multitasking: Using performance-based timed petri nets," International Journal of Industrial Ergonomics, vol. 75, p. 102877, 2020.

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

[14] P. S. Tsang and V. L. Velazquez, "Diagnosticity and multidimensional subjective workload ratings," Ergonomics, vol. 39, no. 3, pp. 358–381, 1996.

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

[16] L. Longo and M. C. Leva, Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers. Springer, 2017, vol. 726.

[17] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," Neuroscience & Biobehavioral Reviews, vol. 44, pp. 58–75, 2014.

[18] A. Giorgi, V. Ronca, A. Vozzi, N. Sciaraffa, A. Di Florio, L. Tamborra, I. Simonetti, P. Aricò, G. Di Flumeri, D. Rossi et al., "Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: A comparison with laboratory technologies," Sensors, vol. 21, no. 7, p. 2332, 2021.

[19] T. Heine, G. Lenis, P. Reichensperger, T. Beran, O. Doessel, and B. Deml, "Electrocardiographic features for the measurement of drivers' mental workload," Applied ergonomics, vol. 61, pp. 31–43, 2017.

[20] J. Zhang, Z. Yin, and R. 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, vol. 45, no. 2, pp. 200–214, 2014.

[21] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," Applied ergonomics, vol. 74, pp. 221–232, 2019.

[22] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey," Clinical neurophysiology, vol. 118, no. 3, pp. 480–494, 2007.

[23] C. Dussault, J.-C. Jouanin, M. Philippe, and C.-Y. Guezennec, "EEG and ECG changes during simulator operation reflect mental workload and vigilance," Aviation, space, and environmental medicine, vol. 76, no. 4, pp. 344–351, 2005.

[24] T. M. Spruill, "Chronic psychosocial stress and hypertension," Current hypertension reports, vol. 12, pp. 10–16, 2010.

[25] M. Tanaka, A. Ishii, and Y. Watanabe, "Neural effects of mental fatigue caused by continuous attention load: a magnetoencephalography study," Brain research, vol. 1561, pp. 60–66, 2014.

[26] . Lim, W.-c. Wu, J. Wang, J. A. Detre, D. F. Dinges, and H. Rao, "Imaging brain fatigue from sustained mental workload: an ASL perfusion study of the time-on-task effect," Neuroimage, vol. 49, no. 4, pp. 3426–3435, 2010.

[27] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS," Frontiers in human neuroscience, vol. 7, p. 935, 2014.

[28] Z. Liu, J. Shore, M. Wang, F. Yuan, A. Buss, and X. Zhao, "A systematic review on hybrid EEG/fNIRS in brain-computer interface," Biomedical Signal Processing and Control, vol. 68, p. 102595, 2021.

[29] J.-P. Lachaux, N. Axmacher, F. Mormann, E. Halgren, and N. E. Crone, "High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research," Progress in neurobiology, vol. 98, no. 3, pp. 279–301, 2012.

[30] A. Gevins and M. E. Smith, "Neurophysiological measures of cognitive workload during human-computer interaction," Theoretical issues in ergonomics science, vol. 4, no. 1-2, pp. 113–131, 2003.

[31] M. A. Hogervorst, A.-M. Brouwer, and J. B. Van Erp, "Combining and comparing EEG, peripheral physiology, and eye-related measures for the assessment of mental workload," Frontiers in neuroscience, vol. 8, p. 322, 2014.

[32] A. Kartali, M. M. Janković, I. Gligorijević, P. Mijović, B. Mijović, and M. C. 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. Springer, 2019, pp. 20–34.

[33] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10-year update," Journal of neural engineering, vol. 15, no. 3, p. 031005, 2018.

[34] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," Journal of neural engineering, vol. 16, no. 3, p. 031001, 2019.

[35] A. J. Bidgoly, H. J. Bidgoly, and Z. Arezoumand, "A survey on methods and challenges in EEG-based authentication," Computers & Security, vol. 93, p. 101788, 2020.

[36] G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for EEG data analytics: A survey," Concurrency and Computation: Practice and Experience, vol. 32, no. 18, p. e5199, 2020.

[37] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," Biomedical Signal Processing and Control, vol. 63, p. 102172, 2021.

[38] D. Merlin Praveena, D. Angelin Sarah, and S. Thomas George, "Deep learning techniques for EEG signal applications—a review," IETE Journal of Research, pp. 1–8, 2020.

[39] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," Journal of neural engineering, vol. 16, no. 5, p. 051001, 2019.

[40] X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung, and C.-T. Lin, "EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," IEEE/ACM transactions on computational biology and bioinformatics, 2021.

[41] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive workload recognition using EEG signals and machine learning: A review," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, pp. 799–818, 2021.

[42] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. 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). IEEE, 2018, pp. 2576–2580.

[43] K. Kingphai and Y. Moshfeghi, "On time series cross-validation for deep learning classification model of mental workload levels based on EEG signals," in Proc. Int. Conf. Machine Learning, Optimization, and Data Science, Cham, Switzerland, 2022, pp. 402-416.

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

[45] K. Kingphai and Y. Moshfeghi, "On EEG preprocessing role in deep learning effectiveness for mental workload classification," in International Symposium on Human Mental Workload: Models and Applications. Springer, 2021, pp. 81–98.

[46] D.-H. Lee, J.-H. Jeong, K. Kim, B.-W. Yu, and S.-W. Lee, "Continuous EEG decoding of pilots' mental states using multiple feature block-based convolutional neural network," IEEE Access, vol. 8, pp. 121 929–121 941, 2020.

[47] K. Smith and P. A. Hancock, "Situation awareness is adaptive, externally directed consciousness," Human factors, vol. 37, no. 1, pp. 137–148, 1995.

[48] F. Nachreiner, "International standards on mental work-load the ISO 10 075 series," Industrial Health, vol. 37, no. 2, pp. 125–133, 1999.

[49] K. F. Lui and A. C.-N. Wong, "Does media multitasking always hurt? A positive correlation between multitasking and multisensory integration," Psychonomic bulletin & review, vol. 19, pp. 647–653, 2012.

[50] J. Törnros and A. 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, vol. 9, no. 4, pp. 298–306, 2006.

[51] M. A. Vidulich and P. S. Tsang, "Mental workload and situation awareness," Handbook of human factors and ergonomics, pp. 243–273, 2012.

[52] C. J. Patten, A. Kircher, J. Östlund, L. Nilsson, and O. Svenson, "Driver experience and cognitive workload in different traffic environments," Accident Analysis & Prevention, vol. 38, no. 5, pp. 887–894, 2006.

[53] A. Qayyum, M. A. Khan, M. Mazher, and M. Suresh, "Classification of EEG learning and resting states using 1D-convolutional neural network for cognitive load assessment," in 2018 IEEE Student Conference on Research and Development (SCOReD). IEEE, 2018, pp. 1–5.

[54] Y. Zhou, T. Xu, S. Li, and R. Shi, "Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game," Universal Access in the Information Society, vol. 18, no. 3, pp. 551–563, 2019.

[55] D. Kahneman, Attention and effort. Citeseer, 1973, vol. 1063.

[56] M. Teplan et al., "Fundamentals of EEG measurement," Measurement science review, vol. 2, no. 2, pp. 1–11, 2002.

[57] 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.

[58] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." Journal of experimental psychology, vol. 55, no. 4, p. 352, 1958.

[59] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," Neuroimage, vol. 59, no. 1, pp. 36–47, 2012.

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

[61] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in Proceedings of the human factors and ergonomics society annual meeting, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[62] S. Mach, P. Storozynski, J. Halama, and J. F. Krems, "Assessing mental workload with wearable devices–reliability and applicability of heart rate and motion measurements," Applied ergonomics, vol. 105, p. 103855, 2022.

[63] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work," European journal of applied physiology, vol. 92, pp. 84–89, 2004.

[64] J. Veltman and A. Gaillard, "Physiological indices of workload in a simulated flight task," Biological psychology, vol. 42, no. 3, pp. 323–342, 1996.

[65] C. Wu, J. Cha, J. Sulek, T. Zhou, C. P. Sundaram, J. Wachs, and D. Yu, "Eye-tracking metrics predict perceived workload in robotic surgical skills training," Human factors, vol. 62, no. 8, pp. 1365–1386, 2020.

[66] B. Cain, "A review of the mental workload literature," DTIC Document, 2007.

[67] D. Jaiswal, A. Chowdhury, T. Banerjee, and D. 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). IEEE, 2019, pp. 2202–2206.

[68] S. Delliaux, A. Delaforge, J.-C. Deharo, and G. Chaumet, "Mental workload alters heart rate variability, lowering non-linear dynamics," Frontiers in physiology, vol. 10, p. 565, 2019.

[69] Y. Y. Yurko, M. W. Scerbo, A. S. Prabhu, C. E. Acker, and D. Stefanidis, "Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool," Simulation in healthcare, vol. 5, no. 5, pp. 267–271, 2010.

[70] N. Ahituv and S. Neumann, Principles of information systems for management. William C. Brown Publishers, 1986.

[71] Y. Chen, S. Yan, and C. C. Tran, "Comprehensive evaluation method for user interface design in nuclear power plant based on mental workload," Nuclear Engineering and Technology, vol. 51, no. 2, pp. 453–462, 2019.

[72] K. Bessiere, J. E. Newhagen, J. P. Robinson, and B. Shneiderman, "A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood," Computers in human behavior, vol. 22, no. 6, pp. 941–961, 2006.

[73] N. I. Arana-De las Casas, J. De la Riva-Rodríguez, A. A. Maldonado-Macías, and D. Sáenz-Zamarrón, "Cognitive analyses for interface design using dual n-back tasks for mental workload (MWL) evaluation," International Journal of Environmental Research and Public Health, vol. 20, no. 2, p. 1184, 2023.

[74] J. Gwizdka, "Assessing cognitive load on web search tasks," arXiv preprint arXiv:1001.1685, 2010.

[75] P. J.-H. Hu, P.-C. Ma, and P. Y. Chau, "Evaluation of user interface designs for information retrieval systems: a computer-based experiment," Decision support systems, vol. 27, no. 1-2, pp. 125–143, 1999.

[76] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during web browsing," Sensors, vol. 18, no. 2, p. 458, 2018.

[77] Z. Mohamed, M. El Halaby, T. Said, D. Shawky, and A. Badawi, "Characterizing focused attention and working memory using EEG," Sensors, vol. 18, no. 11, p. 3743, 2018.

[78] R. Hefron, B. Borghetti, C. Schubert Kabban, J. Christensen, and J. Estepp, "Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks," Sensors, vol. 18, no. 5, p. 1339, 2018.

[79] D. Zhang, D. Cao, and H. 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, 2019, pp. 1–5.

[80] M. A. Almogbel, A. H. Dang, and W. Kameyama, "Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning," in 2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019, pp. 1–6.

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

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

[83] M. Mahmoudi and M. Shamsi, "Multi-class EEG classification of motor imagery signal by finding optimal time segments and features using SNR-based mutual information," Australasian physical & engineering sciences in medicine, vol. 41, no. 4, pp. 957–972, 2018.

[84] S. Dutta and A. Nandy, "An extensive analysis on deep neural architecture for classification of subject-independent cognitive states," in Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 180–184.

[85] Z. Sun, B. Li, F. Duan, H. Jia, S. Wang, Y. Liu, A. Cichocki, C. F. Caiafa, and J. Sole-Casals, "Wlnet: towards an approach for robust workload estimation based on shallow neural networks," IEEE Access, vol. 9, pp. 3165–3173, 2020.

[86] Q. Zhang, Z. Yuan, H. Chen, and X. Li, "Identifying mental workload using EEG and deep learning," in 2019 Chinese Automation Congress (CAC). IEEE, 2019, pp. 1138–1142.

[87] R. Chavarriaga, M. Ušćumlić, H. Zhang, Z. Khaliliardali, R. Aydarkhanov, S. Saeedi, L. Gheorghe, and J. d. R. Millán, "Decoding neural correlates of cognitive states to enhance driving experience," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 4, pp. 288–297, 2018.

[88] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," Artificial intelligence in medicine, vol. 88, pp. 58–69, 2018.

[89] N. Mehdiyev, J. Lahann, A. Emrich, D. Enke, P. Fettke, and P. Loos, "Time series classification using deep learning for process planning: A case from the process industry," Procedia Computer Science, vol. 114, pp. 242 – 249, 2017, complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 –

November 1, 2017, Chicago, Illinois, USA. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917318707

[90] C. Diaz-Piedra, M. V. Sebastián, and L. L. Di Stasi, "EEG theta power activity reflects workload among army combat drivers: an experimental study," Brain sciences, vol. 10, no. 4, p. 199, 2020.

[91] E. Q. Wu, X. Peng, C. Z. Zhang, J. Lin, and R. S. Sheng, "Pilots' fatigue status recognition using deep contractive autoencoder network," IEEE Transactions on Instrumentation and Measurement, vol. 68, no. 10, pp. 3907–3919, 2019.

[92] D. Chen, D. Wu, H. Yu, and Z. Cui, "Assessment of pilot's mental load during traffic pattern with simulator EEG data," in 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT). IEEE, 2020, pp. 1077–1082.

[93] L. Cao, Z. Qian, H. Zareipour, Z. Huang, and F. 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, vol. 7, pp. 155 219–155 228, 2019.

[94] M. Bilalpur, M. Kankanhalli, S. Winkler, and R. Subramanian, "EEG-based evaluation of cognitive workload induced by acoustic parameters for data sonification," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 315–323.

[95] S. B. Shafiei, A. S. Elsayed, A. A. Hussein, U. Iqbal, and K. A. Guru, "Evaluating the mental workload during robot-assisted surgery utilizing network flexibility of human brain," IEEE Access, vol. 8, pp. 204 012–204 019, 2020.

[96] C. Hua, H. Wang, J. Chen, T. Zhang, Q. Wang, and W. Chang, "Novel functional brain network methods based on cnn with an application in proficiency evaluation," Neurocomputing, vol. 359, pp. 153–162, 2019.

[97] A. Qayyum, I. Faye, A. S. Malik, and M. 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). IEEE, 2018, pp. 600–605.

[98] Y. Kwak, K. Kong, W.-J. Song, B.-K. Min, and S.-E. Kim, "Multilevel feature fusion with 3d convolutional neural network for eeg-based workload estimation," IEEE Access, vol. 8, pp. 16 009–16 021, 2020.

[99] Y. Kwak, W.-J. Song, B.-K. Min, and S.-E. Kim, "3d cnn based multi-level feature fusion for workload estimation," in 2020 8th International Winter Conference on Brain-Computer Interface (BCI). IEEE, 2020, pp. 1–4.

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

[101] D. Dewan, L. Ghosh, B. Chakraborty, A. Chowdhury, A. Konar, and A. 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). IEEE, 2020, pp. 1–8.

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

[103] J. Zhang and S. Li, "A deep learning scheme for mental workload classification based on restricted boltzmann machines," Cognition, Technology & Work, vol. 19, no. 4, pp. 607–631, 2017.

[104] D. D. Chakladar, S. Datta, P. P. Roy, and A. Vinod, "Cognitive workload estimation using variational auto encoder & attention-based deep model," IEEE Transactions on Cognitive and Developmental Systems, 2022.

[105] A. Saha, V. Minz, S. Bonela, S. Sreeja, R. Chowdhury, and D. Samanta, "Classification of eeg signals for cognitive load estimation using deep learning architectures," in International Conference on Intelligent Human Computer Interaction. Springer, 2018, pp. 59–68.

[106] L. Xia, Y. Feng, Z. Guo, J. Ding, Y. Li, Y. Li, M. Ma, G. Gan, Y. Xu, J. Luo, Z. Shi, and Y. Guan, "Mulhita: A novel multiclass classification framework with multibranch lstm and hierarchical temporal attention for early detection of mental stress," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–14, 2022.

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

[108] T. Tuncer, S. Dogan, F. Ertam, and A. Subasi, "A dynamic center and multi threshold point based stable feature extraction network for driver fatigue detection utilizing eeg signals," Cognitive neurodynamics, vol. 15, pp. 223–237, 2021.

[109] J. Shang, W. Zhang, J. Xiong, and Q. Liu, "Cognitive load recognition using multi-channel complex network method," in International Symposium on Neural Networks. Springer, 2017, pp. 466–474.

[110] J. Shang and Q. Liu, "Cognitive load recognition using

multi-threshold united complex network," in International Conference on Neural Information Processing. Springer, 2017, pp. 490–498.

[111] A. Ahmadi, H. Bazregarzadeh, and K. Kazemi, "Automated detection of driver fatigue from electroencephalography through wavelet-based connectivity," Biocybernetics and Biomedical Engineering, vol. 41, no. 1, pp. 316–332, 2021.

[112] S. SHAO, T. WANG, C. SONG, Y. SU, Y. WANG, and C. YAO, "Fine-grained and multi-scale motif features for cross-subject mental workload assessment using bi-lstm," Journal of Mechanics in Medicine and Biology, p. 2140020, 2021.

[113] C. Schaffer, "Selecting a classification method by cross-validation," Machine Learning, vol. 13, no. 1, pp. 135–143, 1993.

[114] H. Zeng, C. Yang, G. Dai, F. Qin, J. Zhang, and W. Kong, "EEG classification of driver mental states by deep learning," Cognitive Neurodynamics, vol. 12, no. 6, pp. 597–606, 2018.

[115] Z. Zheng, Z. Yin, and J. Zhang, "An elm-based deep sdae ensemble for inter-subject cognitive workload estimation with physiological signals," in 2020 39th Chinese Control Conference (CCC). IEEE, 2020, pp. 6237–6242.

[116] H. Zeng, X. Li, G. Borghini, Y. Zhao, P. Aricò, G. Di Flumeri, N. Sciaraffa, W. Zakaria, W. Kong, and F. Babiloni, "An EEG-based transfer learning method for cross-subject fatigue mental state prediction," Sensors, vol. 21, no. 7, p. 2369, 2021.

[117] M. Jiménez-Guarneros and P. Gómez-Gil, "Cross-subject classification of cognitive loads using a recurrent-residual deep network," in 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017, pp. 1–7.

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

[119] W. Zhang and Q. 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). IEEE, 2018, pp. 578–582.

[120] Y. Ming, D. Pelusi, C.-N. Fang, M. Prasad, Y.-K. Wang, D. Wu, and C.-T. Lin, "EEG data analysis with stacked differentiable neural computers," Neural Computing and Applications, vol. 32, no. 12, pp. 7611–7621, 2020.

[121] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: An empirical study on performance estimation methods," Machine Learning, vol. 109, no. 11, pp. 1997–2028, 2020.

[122] X. Qu, P. Liu, Z. Li, and T. Hickey, "Multi-class time continuity voting for EEG classification," in International Conference on Brain Function Assessment in Learning. Springer, 2020, pp. 24–33.

[123] X. Qu, Y. Sun, R. Sekuler, and T. Hickey, "EEG markers of STEM learning," in 2018 IEEE Frontiers in Education Conference (FIE), 2018, pp. 1–9.

[124] K. Kingphai and Y. Moshfeghi, "On Time Series Cross-Validation for Mental Workload Classification from EEG Signals," Neuroergonomics conference, Sep. 2021, poster.

[125] Y. Liu and Q. Liu, "Convolutional neural networks with large-margin softmax loss function for cognitive load recognition," in 2017 36th Chinese control conference (CCC). IEEE, 2017, pp. 4045–4049.

[126] L. G. Hernández, O. M. Mozos, J. M. Ferrández, and J. M. Antelis, "EEG-based detection of braking intention under different car driving conditions," Frontiers in neuroinformatics, vol. 12, p. 29, 2018.

[127] Z. Cao, Z. Yin, and J. Zhang, "Recognition of cognitive load with a stacking network ensemble of denoising autoencoders and abstracted neurophysiological features," Cognitive Neurodynamics, vol. 15, no. 3, pp. 425–437, 2021.

[128] S. M. Salaken, I. Hettiarachchi, L. Crameri, S. Hanoun, T. Nguyen, and S. Nahavandi, "Evaluation of classification techniques for identifying cognitive load levels using EEG signals," in 2020 IEEE International Systems Conference (SysCon). IEEE, 2020, pp. 1–8.

[129] N. Sciaraffa, P. Aricò, G. Borghini, G. Di Flumeri, A. Di Florio, and F. 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. Springer, 2019, pp. 170–185.

[130] D.-H. Lee, J.-H. Jeong, K. Kim, B.-W. Yu, and S.-W. Lee, "Continuous EEG decoding of pilots' mental states using multiple feature block-based convolutional neural network," IEEE Access, vol. 8, pp. 121 929–121 941, 2020.

[131] J. Zhang, S. Li, and Z. Yin, "Pattern classification of instantaneous mental workload using ensemble of convolutional neural networks," IFAC-PapersOnLine, vol. 50, no. 1, pp. 14 896–14 901, 2017.

[132] S.-h. Zhong, A. Fares, and J. Jiang, "An attentional-LSTM for improved classification of brain activities evoked by images," in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1295–1303.

[133] M. Jiménez-Guarneros and P. Gómez-Gil, "Custom domain adaptation: A new method for cross-subject, EEG-based

cognitive load recognition," IEEE Signal Processing Letters, vol. 27, pp. 750–754, 2020.

[134] Z. Yin and J. Zhang, "Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights," Neurocomputing, vol. 260, pp. 349–366, 2017.

[135] C. Ma, M. Zhang, X. Sun, H. Wang, and Z. Gao, "Dynamic threshold distribution domain adaptation network: A cross-subject fatigue recognition method based on EEG signals," IEEE Transactions on Cognitive and Developmental Systems, 2023.

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

[137] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song, "Can emotion be transferred?—a review on transfer learning for EEG-based emotion recognition," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, pp. 833–846, 2021.

[138] W. L. Lim, O. Sourina, and L. Wang, "Cross dataset workload classification using encoded wavelet decomposition features," in 2018 International Conference on Cyberworlds (CW). IEEE, 2018, pp. 300–303.

[139] E. A. Kirchner and S. K. Kim, "Multi-tasking and choice of training data influencing parietal ERP expression and single-trial detection—relevance for neuroscience and clinical applications," Frontiers in neuroscience, vol. 12, p. 188, 2018.

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

[141] E. N. Wiebe, E. Roberts, and T. S. Behrend, "An examination of two mental workload measurement approaches to understanding multimedia learning," Computers in Human Behavior, vol. 26, no. 3, pp. 474–481, 2010.

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

[143] M. De Rivecourt, M. Kuperus, W. Post, and L. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," Ergonomics, vol. 51, no. 9, pp. 1295–1319, 2008.

[144] L. J. Mulder, "Measurement and analysis methods of heart rate and respiration for use in applied environments," Biological psychology, vol. 34, no. 2-3, pp. 205–236, 1992.

[145] P. Warden and D. Situnayake, TinyML. O'Reilly Media, Incorporated, 2019.

[146] Y. Jia, L. Guo, and X. Wang, "Real-time control systems," in Transportation Cyber-Physical Systems. Elsevier, 2018, pp. 81–113.

[147] B. Wang, H. Shen, G. Lu, Y. Liu et al., "Graph learning with co-teaching for EEG-based motor imagery recognition," IEEE Transactions on Cognitive and Developmental Systems, 2022.

[148] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," Information Sciences, vol. 191, pp. 192–213, 2012.

[149] K. Kingphai and Y. Moshfeghi, "On channel selection for EEG-based mental workload classification," in Proc. Int. Conf. Machine Learning, Optimization, and Data Science, Cham, Switzerland, 2023, pp. 403-417.

[150] J. Sun, Y. Liu, Z. Ye, and D. Hu, "A novel multi-scale dilated convolution neural network with gating mechanism for decoding driving intentions based on EEG," IEEE Transactions on Cognitive and Developmental Systems, 2023.

[151] S. Zhang, W. Yang, H. Mou, Z. Pei, F. Li, and X. Wu, "An overview of brain fingerprint identification based on various neuroimaging technologies," IEEE Transactions on Cognitive and Developmental Systems, 2023.

[152] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 715–729, 2021.

[153] X. Sun, Y. Qi, Y. Wang, and G. Pan, "Convolutional multiple instance learning for sleep spindle detection with label refinement," IEEE Transactions on Cognitive and Developmental Systems, vol. 15, no. 1, pp. 272–284, 2022.

[154] S. Gong, K. Xing, A. Cichocki, and J. Li, "Deep learning in EEG: Advance of the last ten-year critical period," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 348–365, 2021.

[155] Z. Zhang, A. Liu, Y. Gao, X. Cui, R. Qian, and X. Chen, "Distilling invariant representations with domain adversarial learning for cross-subject children seizure prediction," IEEE Transactions on Cognitive and Developmental Systems, 2023.

[156] W. Li, M. Wang, J. Zhu, and A. Song, "EEG-based emotion recognition using trainable adjacency relation driven graph convolutional network," IEEE Transactions on Cognitive and Developmental Systems, 2023.

[157] C.-T. Lin, J. Liu, C.-N. Fang, S.-Y. Hsiao, Y.-C. Chang, and

Y.-K. Wang, "Multistream 3-d convolution neural network with parameter sharing for human state estimation," IEEE Transactions on Cognitive and Developmental Systems, vol. 15, no. 1, pp. 261–271, 2022.

[158] M. S. Shahabi and A. Shalbaf, "Prediction of treatment outcome in major depressive disorder using ensemble of hybrid transfer learning and long short term memory based on EEG signal," IEEE Transactions on Cognitive and Developmental Systems, 2022.

[159] C. Li, Y. Zhao, R. Song, X. Liu, R. Qian, and X. Chen, "Patient-specific seizure prediction from electroencephalogram signal via multi-channel feedback capsule network," IEEE Transactions on Cognitive and Developmental Systems, 2022.

[160] W. Zhang and D. Wu, "Lightweight source-free transfer for privacy-preserving motor imagery classification," IEEE Transactions on Cognitive and Developmental Systems, 2022.

[161] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "Lienet: a deep convolution neural network framework for detecting deception," IEEE transactions on cognitive and developmental systems, vol. 14, no. 3, pp. 971–984, 2021.

[162] K. Kingphai and Y. Moshfeghi, "EEG-based mental workload level estimation using deep learning models," in Proc. Ergonomics & Human Factors; CIEHF, Birmingham, UK, 2022.