applied sciences

MDPI

Article

Optimal Transport and Graph Neural Networks for Cross-Session Mental Workload Classification

Güliz Demirezen [1,*], Anne-Marie Brouwer [2,3] and Tuğba Taşkaya Temizel [4]

[1] Department of Information Systems, Graduate School of Informatics, Middle East Technical University, Ankara 06800, Türkiye
[2] Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), 3769 DE Soesterberg, The Netherlands
[3] Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 GD Nijmegen, The Netherlands
[4] Department of Data Informatics, Graduate School of Informatics, Middle East Technical University, Ankara 06800, Türkiye
* Correspondence: guliz@demirezen.tech

Abstract

Electroencephalography (EEG) offers a noninvasive, high-temporal-resolution modality for estimating mental workload. However, session-to-session variability limits the generalizability of workload classifiers, and few systematic cross-session evaluations are reported in the literature. This study systematically evaluates domain adaptation methods for cross-session mental workload classification using the publicly available COG-BCI dataset within an evaluation framework that may guide future studies on EEG-based classification models. We make four contributions: (i) integration of Optimal Transport (OT) with Graph Neural Networks (GNNs) to model spatial relationships and align feature distributions under strict session-wise separation; (ii) a data-centric evaluation pipeline incorporating Self-Organizing Map (SOM) visualizations for data exploration and a heuristic loss function for model selection; (iii) a strict cross-session protocol examining the effects of graph construction, feature selection, and data splits; and (iv) comparison of OT with CORrelation ALignment (CORAL) and GNN with EEGNet. Incorporating OT improved test accuracies across all experimental configurations. SOM visualizations confirmed enhanced feature alignment after OT. Our results highlight the potential of OT for mitigating session-to-session variability and underscore the importance of a data-centric approach and rigorous cross-session evaluation when developing classifiers for complex cognitive state estimation. Future work should explore semi-supervised OT strategies and scalable implementations for real-time applications.

Keywords: mental workload classification; EEG; cross-session variability; domain adaptation; optimal transport; passive brain–computer interface

Check for updates

Academic Editor: Jing Jin

Received: 7 April 2026
Revised: 10 May 2026
Accepted: 19 May 2026
Published: 1 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Mental workload refers to the cognitive effort required to perform a task and has become increasingly relevant as modern work environments place greater demands on mental rather than physical performance [1]. Evaluating mental workload may help ensure that systems do not impose excessive cognitive strain on users, especially in safety-critical domains such as air traffic control, driving, and aviation, and for informing system designs that balance user experience and operational demands.

Mental workload is a multidimensional construct influenced by task complexity, individual capacity, and contextual factors [2,3]. Among diverse measurement techniques, electroencephalography (EEG) offers a noninvasive, high-temporal-resolution window into brain activity, making it in principle well suited for assessing mental workload [4–6]. Its portability and cost-effectiveness allow application in both laboratory and real-world settings. Workload-related neural responses are reflected in event-related potentials (ERPs), spectral power variations, and connectivity measures. ERPs such as the P300 component decrease in amplitude and increase in latency with rising workload [7,8]. Other ERP components, including P200 and N200, have also been associated with cognitive demand [9]. Spectral features, notably increased frontal theta and decreased parietal alpha power with increasing mental effort, have been associated with workload [5,10]. The frontal theta to parietal alpha ratio has also been associated with workload, increasing in response to greater cognitive demand [6]. Broader frequency modulations, including delta and beta activity, and connectivity-based metrics such as fronto-parietal coherence may further enrich workload characterization [11,12].

Translating these neural signatures into practical workload assessments requires computational models capable of robust classification. Traditional machine learning approaches, including K-nearest neighbors, logistic regression, and support vector machines, have been employed to classify workload levels from EEG data [13–16]. More recent developments in neural network architectures, such as artificial neural networks [17,18], CNNs [14,19], and LSTMs [20,21], have achieved improved performance by capturing nonlinear and temporal dynamics.

EEG signals, recorded across multiple brain regions, exhibit spatial dependencies that conventional machine learning models often overlook by treating EEG signals as independent time series. Graph Neural Networks (GNNs) address this limitation by representing EEG channels as graph nodes, enabling spatial modeling and enhancing cognitive state classification [22]. Graph-based and Transformer-based models have recently shown strong performance in tasks such as emotion recognition and motor imagery [23,24], yet their application to workload classification remains limited [25].

1.1. Cross-Session Challenges and Optimal Transport Domain Adaptation Approach for EEG Signals

A significant challenge in EEG-based mental workload classification is generalizability. EEG signals exhibit substantial variability across recording sessions, subjects, and experimental contexts due to factors such as electrode placement, skin impedance, environmental conditions, and fluctuations in participants' cognitive and affective states [26]. This nonstationarity leads to domain shifts, causing models to underperform on unseen data. Cross-session modeling aims to address this nonstationarity by ensuring that models trained on data from one session remain effective on others [27]. Previous studies have documented severe performance degradation when models trained on one session are applied to another [28,29]. Addressing this challenge is essential for developing practical brain–computer interface (BCI) systems that require minimal recalibration and offer robust performance in real-world applications.

Domain adaptation techniques aim to mitigate the variability by aligning distributions across sessions or subjects. Among these, Optimal Transport (OT) has emerged as a mathematically grounded method that finds cost-efficient transformations between probability distributions while preserving the underlying data geometry [30]. OT has demonstrated utility in various EEG tasks. For instance, in sleep staging, OT has been applied to achieve cross-dataset, cross-subject, and cross-session transfer [31,32]. In emotion recognition, OT-based adaptation has helped address inter-subject variability and improve model robustness [33]. OT has also been widely applied in motor imagery classification, where it

supports generalization across subjects or sessions [34–36]. However, its application to cross-session mental workload classification remains unexplored.

1.2. Study Objectives and Contributions

This study establishes a systematic evaluation framework for cross-session EEG-based mental workload classification, comprising a strict session-wise data processing and model evaluation protocol, a data-centric analysis pipeline incorporating SOM visualizations and a heuristic loss-based model selection procedure, and a structured comparison of domain adaptation and model architectures. While domain adaptation techniques have demonstrated success in motor imagery, emotion recognition, and sleep staging, their effectiveness for mental workload estimation remains largely unexplored. Mental workload classification presents challenges including continuous cognitive states, label ambiguity, and high inter-class overlap that differ from discrete motor imagery paradigms.

Despite progress in EEG-based mental workload classification, many studies rely on within-session evaluation, where training and testing data originate from the same recording session [37,38]. Such settings may overestimate real-world performance and do not adequately address session-to-session distribution shifts. This highlights the need for methods designed to address session-to-session distribution shift.

Furthermore, while Graph Neural Networks can model spatial electrode relationships and Optimal Transport can reduce domain mismatch, their combined use for cross-session mental workload classification has not been systematically investigated. This motivates the proposed OT–GNN framework.

Our investigation makes four contributions.

First, we integrate OT domain adaptation with GNNs for cross-session mental workload classification. GNNs model spatial relationships among EEG channels through graph convolution operations, capturing topological dependencies that conventional architectures treat independently. OT aligns feature distributions across sessions by computing cost-efficient transformations between probability distributions while preserving underlying data geometry. We apply backward OT adaptation using the Sinkhorn algorithm, transporting validation and test features into the training domain without requiring model retraining. This integration is evaluated under strict session-wise separation where models trained on one session are tested on completely unseen sessions. Our results show that OT with GNNs achieves statistically significant improvements over nonadapted baselines (mean gains 6.46% three-class, 5.85% binary, $p < 0.01$ via Wilcoxon tests). To our knowledge, this is the first study combining GNNs with OT for EEG classification.

Second, we introduce a data-centric methodology addressing model selection under high EEG variability. EEG nonstationarity creates inconsistent performance across subjects, with some models overfitting and others underfitting. We propose a heuristic loss function that jointly optimizes validation performance and train-validation stability, preventing selection of overfit models or models that are otherwise not representative. SOM visualizations trained exclusively on training data serve as fixed projections for validation and test sets, providing interpretable assessment of cross-session distribution shifts and class separability before and after domain adaptation. These tools enable consistent model selection under conditions where standard validation-based selection fails and reveal why adaptation methods succeed or fail at the feature level. All preprocessing steps, experimental configurations, and model architectures are described in detail following the reproducibility guidelines established in [37], with complete source code publicly available (https://github.com/gdemirezen/gnn-ot-crosssession-eeg accessed on 8 May 2026) and all experiments conducted on the public COG-BCI dataset [39].

Third, we implement a rigorous cross-session evaluation protocol reflecting real-world deployment scenarios. Models are trained on one session and tested on completely unseen sessions, contrasting with common within-session evaluation that masks generalization failures. Our protocol includes multiple data split strategies isolating session effects, systematic ablations of graph construction methods (single connected component vs. k-nearest neighbor graphs), feature selection approaches (frequency bands, cortical regions, mutual information-based node selection), and epoch parameters (length and overlap). We apply nonparametric statistical testing (Friedman, Wilcoxon) across 26 subjects with multiple experimental configurations to establish robust performance comparisons.

Fourth, we conduct systematic comparisons with comparable, standard methods. We evaluate OT against covariance alignment (CORAL) as a computationally efficient second-order baseline and compare GNNs against the established EEGNet CNN architecture. CORAL deteriorates performance (33.07% test accuracy vs. 50.54% for OT), and EEGNet suffers similar cross-session degradation ($\leq$40.92%), confirming that architectural choice alone does not overcome session variability. Replication of FOOOF periodic features yields no improvement in our setting, suggesting context-dependent effectiveness. SOM visualizations reveal persistent label overlap despite adaptation, particularly between medium and difficult workload conditions. These results demonstrate fundamental differences from motor imagery benchmarks where within-session accuracies reach 70–90% and show that unsupervised domain adaptation alone cannot fully address label ambiguity inherent to mental workload tasks.

Across 26 subjects, we comprehensively evaluate 30 total experimental configurations ((2 baseline (multiclass and binary classification), 3 data split variants, 3 epoching variants, 8 GNN variants, 1 heuristic loss weight variant, 5 OT variants, 1 CORAL) $\times$ 150 trials as well as 8 EEGNet variants) $\times$ 26 subjects, amounting to 89,908 individual model runs.

This work provides methodological infrastructure for evaluating domain adaptation in mental workload estimation through rigorous protocols, interpretable visualization tools, and systematic baselines with transparent reporting of both successes and limitations.

1.3. Unique Challenges of Mental Workload Classification

EEG-based mental workload classification poses challenges that are fundamentally different from those encountered in more established EEG-BCI paradigms such as motor imagery (MI) or event-locked cognitive tasks. These differences arise from the nature of the cognitive processes involved, the definition of ground-truth labels, and the evaluation protocols typically used.

Motor imagery paradigms are characterized by discrete, experimenter-defined mental states with clear task instructions and well-established neural correlates. In within-session evaluations, MI classification achieves accuracies in the range of approximately 70–90%, depending on task complexity and subject variability, with performance degrading substantially in cross-subject or cross-session settings due to EEG nonstationarity [40]. Importantly, many widely cited MI benchmarks report within-session performance, where training and test data are drawn from the same recording session.

Mental workload classification, by contrast, encompasses heterogeneous task designs. Some workload paradigms—such as n-back or mental arithmetic—involve experimenter-defined difficulty levels aligned with trial onsets. In these settings, within-session binary classification between low and high load can reach relatively high accuracies (often 80–90%) under controlled conditions [7,41]. However, even in these discrete paradigms, performance decreases markedly when extending to multiclass workload levels or when evaluating across sessions, reflecting increased overlap between neural representations of adjacent workload states.

More complex multitasking paradigms, such as the Multi-Attribute Task Battery (MATB), differ substantially from both MI and discrete workload tasks. In these scenarios, workload emerges as a continuous cognitive construct, influenced by task coordination, strategy, fatigue, and learning effects, rather than being tied to isolated event triggers. As a result, EEG signatures of workload are more distributed and less standardized, leading to lower separability between workload levels. Systematic reviews report that classification accuracy tends to drop significantly in complex task settings such as multitasking when quantitative task load is used as the label, compared to single-task paradigms and subjective workload ratings [42].

A further challenge lies in label ambiguity. While motor imagery labels directly correspond to an explicitly instructed mental action, mental workload labels are typically inferred from task difficulty manipulations. Mental workload is widely recognized as a subjective and multidimensional construct that depends not only on task demands but also on individual characteristics or states, such as cognitive capacity, skill, and motivation. As a result, identical task conditions may induce different levels of experienced workload across individuals and sessions, weakening the correspondence between nominal task difficulty and underlying cognitive state [43]. This source of variability is intrinsic to workload estimation and is largely absent in paradigms with clearly defined, externally cued labels such as motor imagery.

These challenges are amplified in cross-session classification, where EEG feature distributions shift due to nonstationarity, changes in baseline cognitive state, and recording variability. Prior work has shown that cross-session MW classification without adaptation often approaches chance-level performance, with substantial improvements achievable only through explicit domain alignment techniques [44]. Consequently, absolute accuracy values in cross-session MW classification are not directly comparable to within-session MI or workload benchmarks.

Taken together, these factors indicate that performance expectations derived from MI or event-locked cognitive tasks cannot be directly transferred to mental workload classification—particularly in multiclass, cross-session settings. Proper interpretation of results therefore requires careful consideration of task design, labeling assumptions, and evaluation protocol.

1.4. Background

This subsection provides background on the methodologies employed in this study, including GNNs and EEGNet for classification, as well as CORAL and OT for domain adaptation.

1.4.1. Graph Neural Networks Classification Model

GNNs, initially introduced in [45], are neural architectures specifically designed to operate on graph-structured data. A graph is typically denoted as $G(N, E)$, where N represents the set of nodes and E the set of edges. For graph classification tasks, the graph G is transformed into an m-dimensional vector of integers in N^m through a mapping function $\tau(G)$, where m corresponds to the number of classes. GNNs are effective in capturing and utilizing the relational and structural information encoded in graphs by preserving the topological relationships among nodes.

Training GNNs involves a process known as Message Passing, where node features are propagated through the graph based on the edge connectivity and, if available, edge weights or directions. Each node aggregates feature information from its neighbors in a step called Neighborhood Aggregation to form updated node representations (embeddings). By iteratively exchanging information through message passing, each node incorporates

contextual information from its local neighborhood along with its own attributes. This mechanism enhances the representational capacity of the network. Each graph convolution layer facilitates information sharing from nodes within a 1-hop neighborhood, meaning that the number of graph convolution layers determines the effective receptive field in terms of node distance. Variations among GNN models are mainly attributed to differences in their approaches to message propagation and feature aggregation.

To perform classification at the graph level, node embeddings are combined into a single graph-level representation using an order-invariant pooling operation, such as global mean or max pooling. This pooled graph embedding is then passed through a classifier to produce the final class predictions.

1.4.2. EEGNet Classification Model

EEGNet has become a standard baseline in EEG classification research due to its effectiveness across multiple BCI paradigms and efficient parameter usage.

EEGNet employs a two-block convolution strategy optimized for EEG signal characteristics and a final softmax classification. An average pooling layer is used at the end of each block for dimension reduction.

- Block 1 consists of a temporal convolution layer that captures temporal patterns at different frequency scales. Temporal convolutions are followed by spatial depthwise convolutions with a kernel size of (C, 1), where C is the number of channels. Spatial convolutions learn frequency-specific spatial patterns across electrodes. Dropout is used to regularize the model.
- Block 2 entails a Separable Convolution in the form of Depthwise Convolution followed by pointwise convolution. This reduces the number of parameters to fit and summarizes and combines features across maps.

1.4.3. Optimal Transport-Based Domain Adaptation

Optimal Transport (OT) theory, originating from Gaspard Monge's 1781 formulation, addresses the problem of finding cost-efficient transformations between probability distributions while preserving the underlying data geometry [46]. The Kantorovich formulation relaxes the original Monge problem by allowing probabilistic transport plans, yielding a convex optimization problem whose solution defines the Wasserstein distance, a metric that captures distributional differences even when supports do not overlap [46–48]. Entropic regularization introduces an entropy term into the objective, enabling efficient computation via the Sinkhorn algorithm while smoothing the transport plan [49–51]. Full mathematical details are provided in Appendix A.

1.4.4. Correlation Alignment (CORAL) Domain Adaptation

CORAL [52] is another unsupervised, feature-based domain adaptation method. CORAL provides a computationally efficient solution by aligning distributions through their second-order statistics. While OT operates on the full distributional geometry, CORAL specifically targets covariance structure, offering a closed-form solution without iterative optimization. The transformation aims to minimize the Frobenius norm between the transformed source data and the target data given by the following optimization problem:

$$\min_{A} \left\| A^T C_S A - C_T \right\|_F^2$$

where A is the transformation matrix, C_S is the correlation matrix of source data, and C_T is the correlation matrix of target data.

CORAL assumes that domain shift primarily affects feature correlations while leaving means approximately stable across domains and transforms source features to match target covariance structure. For numerical stability, covariance matrices are computed with regularization, where λ is a small regularization term to ensure positive definiteness.

$$C_S = \mathrm{Cov}(X_S) + \lambda I_p$$
$$C_T = \mathrm{Cov}(X_T) + \lambda I_p$$

2. Materials and Methods

2.1. Dataset Description

The COG-BCI dataset, which contains raw EEG data from 29 subjects (11 Female, 18 Male; mean age: 23.9 (std: 3.2) years), was used in this study [39]. The raw recordings include 64-electrode EEG data from the Psychomotor Vigilance Task (PVT), Flanker task, n-back task (n = 0, 1, 2), and the Multi-Attribute Task Battery (MATB) task with three difficulty levels. Data was collected using an EEG system with 64 active Ag-AgCl electrodes (ActiCap, Brain Products GmbH, Germany) and an ActiCHamp amplifier (Brain Products GmbH, Germany) and following the standard 10–20 system for the positioning of electrodes. Data was collected with 500 Hz sampling rate and no online filtering was applied. Fpz was used as the reference electrode. All electrode impedances were kept below 25 kΩ. TP9 was used to record ECG. Three sessions were recorded at one-week intervals, with each session lasting approximately two hours. Resting-state recordings include one minute of eyes-open and one minute of eyes-closed conditions at the start and end of each session. Tasks were performed in a pseudo-randomized order during each session. Further details are presented in [53].

This study focuses on data from the MATB-II task, as it offers a closer representation of real-world scenarios compared to tasks like the n-back. Developed by NASA, the MATB-II is designed to assess task-switching and mental workload in environments that simulate realistic piloting conditions [54]. In the COG-BCI dataset, the MATB-II task includes three five-minute runs of varying difficulty, corresponding to easy, medium, and difficult workload conditions. It incorporates combinations of four subtasks: tracking (TRACK), system monitoring (SYSMON), communication (COMM), and resource management (RESMAN). Higher difficulty levels involve either a greater number of simultaneous tasks or increased task complexity. Further details are available in [53]. The classification problem that is focused on using this data is three-class, within-participant, and cross-session.

2.2. Preprocessing

Exploratory data analysis on the raw dataset showed that subjects 17 and 27 have the same data. Moreover, subjects 25 and 28 have the same data for session 1. Therefore, data from subjects 25, 27, and 28 was excluded from this study. As a result, data from 26 subjects was kept after removing duplicate or inconsistent recordings.

Preprocessing of the EEG signals was performed using the open-source EEGLAB toolbox (v2024.2) [55] that runs on MATLAB (R2023a) [56]. After data import, the ECG channel was excluded from further analysis. Electrode locations were assigned using spherical coordinates based on the Boundary Element Model (BEM) Dipfit template available in the EEGLAB toolbox. The EEG signals were then re-referenced to the right mastoid (*TP10*) and in preparation for Independent Component Analysis (ICA) computation, high-pass filtered at 1 Hz [57]. Channels exhibiting a standard deviation greater than two standard deviations of the average standard deviation of all channels were subsequently rejected. Across subjects and sessions in the MATB tasks, an average of 3.37 $\pm$ 0.88 EEG channels were discarded.

Artifact removal was performed using ICA, a widely adopted approach for denoising EEG signals. EEGLAB's built-in ICA functionality was utilized [58], and components were extracted using the Picard algorithm [59]. The resulting components were automatically labeled via the ICLabel toolbox [60], which classifies them into categories such as muscle, eye, heart, brain, or other. Components identified as muscle, eye, or heart artifacts with a probability exceeding 95% were excluded. An average of 2.32 $\pm$ 2.17 independent components were rejected per subject, per session.

Since channel interpolation introduces dependency and results in a rank-deficient data matrix, performing ICA after interpolation may lead to inaccurate results [61]. Therefore, spherical interpolation was applied only after ICA and component rejection. Finally, the data was low-pass filtered at 40 Hz and downsampled to 250 Hz. Both high-pass and low-pass filters were implemented as finite impulse response (FIR) filters using EEGLAB's `eegfiltnew()` function with default settings. This preprocessing pipeline was applied consistently across all subjects and sessions for the MATB task.

2.3. Feature Extraction

The EEG data was transformed into a graph representation by extracting node and edge features, which were then integrated with an adjacency matrix. Before feature extraction, the *Cz* channel, which was not recorded for the first nine subjects, was removed from the data for all participants to ensure consistency.

2.3.1. Segmentation and Labeling

A varying range of epoch sizes are used for EEG classification models in the literature [62]. Effects of epoch length can change with respect to task or data. Specifically, mental state classification studies have employed epochs spanning from short windows of a few seconds to multisecond durations [63].

In this study, EEG data was segmented into 4 s 50% overlapping epochs as a baseline. Overlapping was employed to increase the information content, especially to account for the tapered edges of convolution signals during spectral feature calculation. Each epoch was labeled based on the MATB task difficulty levels as Easy, Medium, or Difficult for three-class classification, corresponding to low, medium, and high mental workload conditions, respectively. Epoching into 4 s overlapping segments resulted in 444 samples for each of the training, validation, and test sets, where classes were balanced and there were 148 samples for each class and each participant. Additionally, binary classification experiments were performed to differentiate between low and high mental workload conditions.

2.3.2. Node Features

EEG Power Spectral Density (PSD) is frequently used as a feature for mental workload estimation due to its established relevance [4,5,10,42]. In this study, we used PSD mean power values, which were computed for each epoch using Welch's method, implemented with EEGLAB's `spectopo()` function. A 1-second Welch window with 50% overlap was selected to improve the information content. The node features consisted of the mean power values for the following frequency bands: theta (4–8 Hz), alpha (8–13 Hz), low beta (13–20 Hz), and high beta (20–30 Hz). The node features were MinMax scaled for normalization using only train dataset statistics. Furthermore, in Section 2.6.8, we empirically compared raw PSD against FOOOF-decomposed periodic features and found no performance gain, which supports the use of standard Welch-based PSD as the feature representation.

2.3.3. Edge Features

Edge features determine the influence of neighboring nodes during message passing. To reflect spatial relationships, the inverse of the three-dimensional Euclidean distances

between electrode positions were employed as edge weights, assigning higher importance to closer nodes. These weights were then normalized by the maximum value, as shown in Equation (1), where n denotes the total number of nodes, d_{ij} and w_{ij} represent the Euclidean distance and the corresponding edge weight between nodes i and j, respectively.

$$w_{ij} = \frac{1}{d_{ij}.max(w_{ij})} \quad (1)$$

where

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

2.3.4. Graph Representation

The graph structure is defined by specifying the adjacency matrix in conjunction with the extracted node and edge features. Two methods were used to construct the graph:

1. Single Connected Component: A predefined percentage, p, of edges with the highest weights was selected to ensure that the graph remained as a single connected component.
2. k-Nearest Neighbor Graph (k-NNG) [64]: The graphs were constructed by connecting each node to its k closest nodes based on Euclidean distances between the electrodes. `knn_graph()` function from PyTorch Geometric was used, ensuring that each node had exactly k incoming edges.

Graph construction in EEG-based GNN models determines how spatial relationships between electrodes are represented during message passing. We adopted distance-based sparse graphs because neighboring electrodes capture related cortical activity through volume conduction, avoiding the noise and computational burden of fully connected graphs. Crucially, electrode spatial distance is fixed by the cap layout and therefore session-invariant, whereas functional connectivity metrics such as correlation or coherence are estimated from the EEG signal itself and vary across sessions, introducing session-dependent graph topology that would conflict with the cross-session adaptation objective.

2.3.5. Node Selection

In certain experiments, a subset of nodes was selected based on the highest channel-wise mutual information between node features and ground truth labels, aiming to improve classification performance by reducing model complexity. This selection was performed exclusively on the training set to avoid data leakage into the validation and test sets and assumes cross-session distribution of workload-relevant EEG channels does not shift.

2.3.6. Data Split

To investigate cross-session workload estimation, the data was split into training, validation, and test sets based on sessions. Two strategies were employed for creating the training and validation sets. In the first strategy, sessions were kept separate between the training and validation sets to evaluate cross-session training performance. In the second strategy, data from two sessions was combined, shuffled, and then split 50/50 to assess performance when cross-session variation was minimized between the training and validation sets. The training set was used for model training, while the validation set was employed for hyperparameter optimization and model selection. The test set, containing only unseen data, was reserved for evaluating the model's performance and assessing its generalizability. The test set always comprised a distinct third session to ensure that generalizability was evaluated without cross-session data leakage.

2.4. Classification Model

2.4.1. Model Architecture and Training

The architecture of the employed GNN is depicted in Figure 1. The input graph was passed through several graph convolution layers, each followed by a nonlinear activation function, to compute node-level embeddings. At the readout stage, a permutation-invariant global pooling operation was applied to obtain a graph-level embedding, which was then fed into a classifier composed of fully connected layers with activation functions. In all experiments, global max pooling was used for graph-level readout.

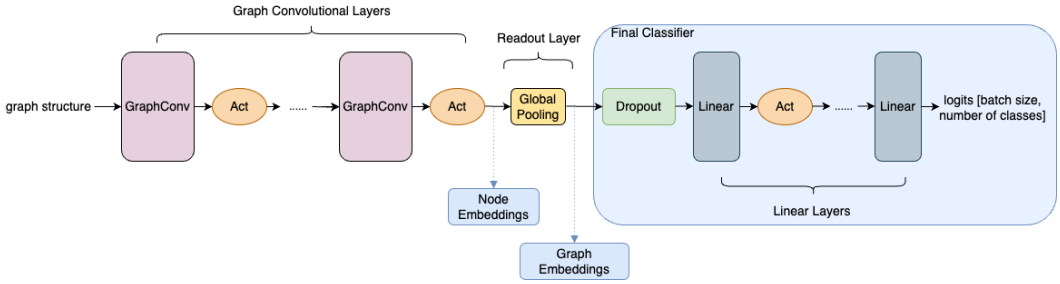

Figure 1. GNN architecture.

The graph convolution layers were based on the higher-order Weisfeiler–Lehman (WL) framework. This was preferred over simpler first-order aggregation schemes (e.g., Graph Convolutional Network (GCN)) because the WL framework's enhanced representational capacity has been shown to better capture complex structural patterns in graph-structured data [65]. Higher-order WL was implemented using the `GraphConv()` class provided by the PyTorch Geometric library [66]. All models were implemented using the open-source PyTorch Geometric framework, which is widely used for GNN development and offers a comprehensive set of tools for graph processing and learning.

Model training was conducted using the cross-entropy loss function, optimized with the Adam optimizer in PyTorch. To prevent overfitting, which was particularly important considering the limited dataset size, regularization techniques such as dropout at the classifier input, weight decay, and learning rate scheduling were employed.

2.4.2. Hyperparameters and Model Selection

Hyperparameter optimization was performed using the Optuna framework [67], which enables efficient exploration of the search space through Bayesian optimization. To ensure reproducibility across different runs, a fixed random seed was applied within the Optuna configuration. The set of hyperparameters tuned during the training of the GNN included activation functions, batch size, learning rate, dropout rate, weight decay, number of layers, and the number of neurons in each graph and fully connected layer. Hyperparameter ranges are given in Appendix B. The COG-BCI dataset is balanced, with each subject contributing an equal number of samples per mental workload level. Therefore, classification accuracy was adopted as the primary performance metric.

Model selection is challenging due to the small sample size and substantial EEG variability. Pronounced inter-subject differences can lead to inconsistent performance across subjects, with some models overfitting, achieving high training accuracy but poor generalization, and others failing to capture relevant patterns. Therefore, relying solely on peak validation accuracy can be misleading, ultimately resulting in unreliable validation performance. To address this, a weighted heuristic loss function (Equation (2)) was employed, combining validation loss with the discrepancy between training and validation losses. This approach penalized poor generalization and enabled automatic selection of the optimal Optuna trial and epoch.

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

The term $loss_{diff}$ reflects the training–validation loss gap, with a penalty ($1e6$) imposed when training loss exceeds validation loss, which corresponds to failure to learn from data and thus exclusion of such models. This formulation discourages both overfitting and nonrepresentative models, promoting those with better generalization across sessions. In cross-session EEG settings, validation loss may be lower than training loss due to noise, session mismatch, or limited sample sizes, which can lead to unstable or nonrepresentative model selection. When this occurs early in training, it reflects a condition where the model has not yet learned meaningful representations that generalize to either set. The apparently lower validation loss is a result of domain mismatch rather than generalization. In such cases, the model has not captured relevant signal structure from the training session. The heuristic loss is designed to exclude such pathological cases from model selection. The weighting parameter $w_{validation}$ was fixed to 0.75 in all GNN experiments to balance validation performance with generalization stability during model selection. To assess the effect of this choice, additional runs were conducted with $w_{validation}$ = 1.0, corresponding to conventional validation-only model selection. This comparison was not intended to optimize test performance but rather to examine the stability of the model selection process under high EEG variability. Qualitative inspection of training and validation loss trajectories for representative subjects was used to verify that the chosen weighting reduced extreme discrepancies and supported more consistent model selection across subjects.

Experiments were tracked and performance metrics were visualized utilizing the Weights and Biases framework [68]. Each trial, corresponding to a distinct set of hyperparameters, was trained for 500 epochs, with the best epoch being logged based on the heuristic loss.

2.4.3. Statistical Significance of Models

To assess the statistical significance of performance differences among models in this study, nonparametric tests were employed due to potential violations of normality assumptions in the data. The Friedman test, serving as a nonparametric alternative to repeated-measures ANOVA, was used to evaluate whether multiple models exhibit statistically significant differences in performance across subjects by ranking model results and testing the null hypothesis of equivalent performance. For pairwise comparisons between models, the Wilcoxon Signed-Ranks test was applied as a nonparametric alternative to the paired t-test, analyzing whether one model consistently outperforms another based on the ranks of performance differences rather than raw values. These nonparametric methods offer robust statistical analysis without assuming normal distribution of the dependent variables, ensuring reliable inference under the variability inherent in EEG-based mental workload classification [69].

2.4.4. EEGNet Classification Model

We compared our graph-based approach against EEGNet [70], a compact convolutional neural network specifically designed for EEG-BCI applications, to serve as a baseline. We implemented EEGNet using the Braindecode Python package [71].

2.5. Domain Adaptation

2.5.1. Optimal Transport-Based Domain Adaptation

Given the inherent variability in EEG data across sessions, OT offers a principled approach for domain adaptation in mental workload estimation by aligning feature distributions from different sessions. In this study, a backward OT strategy, inspired by [34], was implemented to transform features from the validation and test sets into the distributional domain of the training set. This enabled compatibility with the pre-trained GNN models without requiring any model retraining, as predictions were performed directly on the transported features unlike a forward strategy, which transforms the training set to validation or test distributions and requires retraining whenever new session data is collected. The backward OT approach was preferred for its ability to apply domain adaptation purely through feature-space transformations while maintaining robust performance under the nonstationarity of EEG signals. The OT mapping minimizes the Wasserstein distance between source and target distributions, thereby reducing session-to-session variability and enhancing generalization to unseen data. Computation of the OT plan was performed using the Sinkhorn algorithm implemented via the Python OT (POT) library [72]. The regularization parameter was initialized to 1×10^{-3} and doubled until convergence was achieved, selecting the minimum value that satisfied convergence criteria to balance computational efficiency and solution stability.

2.5.2. Correlation Alignment (CORAL)

We compared the OT approach against CORAL [52]. Following the same backward adaptation strategy as OT, we applied CORAL to transform validation and test features into the training domain rather than adapting the model to new sessions. This enables direct use of pre-trained classifiers without retraining. We implemented CORAL using the adapt Python package [73].

2.6. Experimental Design

Several classification experiments were conducted to investigate the effect of various factors on cross-session EEG-based mental workload classification performance. The experiment names are denoted by "{DatasetName}{Model Identifier}{i}{x}", where {DatasetName} is "COG_BCI", {Model Identifier} is "g" for GNN, {i} is the baseline experiment number, and {x} denotes variations from the baseline.

To ease navigation of the experimental design, Table 1 provides an overview of all experiment categories, the research question each addresses, and the corresponding subsection. Full details of each configuration are given in the subsections that follow.

2.6.1. Variations in Data Split

In the baseline experiment (COG_BCIg1), sessions 1 (S1), 2 (S2), and 3 (S3) were used as the training, validation, and test sets, respectively. To examine session effects, a second experiment (COG_BCIg3) swapped the validation and test sessions while keeping session 1 for training. Additional experiments followed the procedure described in Section 2.3.6. A 50/50 split was chosen to match the baseline configuration and keep the class distribution balanced across the training, validation, and test sets. The test session remained completely separate from the other sets, allowing an unbiased assessment of the model's ability to generalize to unseen sessions. The corresponding model identifiers are given in Table 2. The baseline experiment was also carried out in binary setting (COG_BCIg1_binary).

Table 1. Overview of experiments.

Category	Research Question	Experiments	Section
Baseline	What is the cross-session classification performance of the proposed GNN pipeline?	COG_BCIg1	Section 2.6.1
Data split variants	How does session assignment as train/val/test set affect performance?	COG_BCIg2–4 COG_BCIg1_binary	Section 2.6.1
Epoch and graph ablations	How sensitive is performance to epoch length, overlap, and graph topology?	COG_BCIg1_2sec, _4sec_no, _2sec_no COG_BCIg1_3nn, _5nn	Sections 2.6.2 and 2.6.3
Feature and node selection ablations	Does reducing feature dimensionality improve generalization?	COG_BCIg1_theta_alpha, _frontal, _parietal COG_BCI_50p_3nn, _50p_5nn, _25p_5nn	Section 2.6.4
Architecture comparison	Does a graph-based model outperform a CNN baseline under cross-session conditions?	COG_BCIe1–8	Section 2.6.5
Domain adaptation comparison	Does OT improve cross-session accuracy, and how does it compare to CORAL?	COG_BCIg5–8 COG_BCIg5_binary COG_BCIg5_CORAL	Sections 2.6.6 and 2.6.7
FOOOF replication	Do periodic PSD features improve workload classification as reported in prior work?	FOOOF variants on COG_BCIg1	Section 2.6.8

Table 2. Experiments with variations in data split.

Experiment Name	Train Set	Validation Set	Test Set
COG_BCIg1	S1 (100%)	S2 (100%)	S3 (100%)
COG_BCIg2	S1 (50%) + S2 (50%)	S1 (50%) + S2 (50%)	S3 (100%)
COG_BCIg3	S1 (100%)	S3 (100%)	S2 (100%)
COG_BCIg4	S1 (50%) + S3 (50%)	S1 (50%) + S3 (50%)	S2 (100%)

2.6.2. Epoch Length and Overlap

The baseline experiment (COG_BCIg1) segmented the EEG into 4-second epochs with 50% overlap. To examine how epoch length and overlap affect model performance, 2-second epochs with 50% overlap and no overlap, as well as 4-second nonoverlapping epochs, were additionally tested. The corresponding model identifiers are given in Table 3.

Table 3. Experiments with variations in epoch length and overlap.

Experiment Name	Epoch Size (Seconds)	Overlap
COG_BCIg1	4	50%
COG_BCIg1_2sec	2	50%
COG_BCIg1_4sec_no	4	None
COG_BCIg1_2sec_no	2	None

2.6.3. Graph Structures

Graph structures varied as described in Section 2.3.4. The baseline (COG_BCIg1) formed a single connected component by including the strongest 6% of edges when every EEG channel was treated as a node. In addition, k-NN graphs were created with $k = 3$ and $k = 5$. The specific graph structures evaluated are listed in Table 4.

Table 4. Experiments with variations in graph structures.

Experiment Name	Graph Structure
COG_BCIg1	Single Connected Component (threshold = 6%)
COG_BCIg1_3nn	knn graph ($k = 3$)
COG_BCIg1_5nn	knn graph ($k = 5$)

2.6.4. Feature and Node Selection

High feature dimensionality can lead to overfitting and poor generalization. To investigate this behavior, additional variants with reduced feature sets or node selections were evaluated. In the first three variants, only those cortical regions or frequency bands reported in the literature to correlate with mental workload were kept. Other variants selected a fixed proportion of nodes as explained in 2.3.5, thereby keeping only the most informative channels and reducing model complexity. The specific feature and node selections that are evaluated are listed in Table 5.

Table 5. Experiments with variations in feature and node selection.

Experiment Name	Feature/Node Selection
COG_BCIg1_theta_alpha	Only theta and alpha band features
COG_BCIg1_parietal	Only parietal channels
COG_BCIg1_frontal	Only frontal channels
COG_BCIg1_50p_3nn	50% of the nodes were selected in knn graph ($k = 3$)
COG_BCIg1_50p_5nn	50% of the nodes were selected in knn graph ($k = 5$)
COG_BCIg1_25p_5nn	25% of the nodes were selected in knn graph ($k = 5$)

2.6.5. EEGNet Experiments

We implemented EEGNet using the architecture from the original paper [70]. To ensure fair comparison with the optimized GNN architecture, we performed grid search over the following EEGNet hyperparameters:

- Temporal filters $F1$: [8, 16]
- Learning rate: [0.0005, 0.001]
- Weight decay: [0.0001, 0.001]

The experiments that were carried out using EEGNet are tabulated in Table 6. Experiment names are denoted by "{DatasetName}{Model Identifier}{i}", where {DatasetName} is "COG_BCI", {Model Identifier} is "e" for EEGNet, and {i} is the experiment number.

We used the following default values from the original paper for the remaining parameters:

- Depth multiplier ($D = 2$)
- Number of pointwise filters $F2 = F1 \times D$
- Dropout probability = 0.5 (for within-subject classification)
- Temporal kernel length = 125 (half the sampling frequency)

Unlike the GNN, which operates on graph-structured PSD features, EEGNet requires epoched time-series data with a shape of (channels × timepoints). For a fair comparison, we used the same preprocessed EEG epochs employed to compute PSD features for the GNN in the baseline COG_BCIg1 experiments. Specifically, 4-second epochs with 50% overlap were extracted, and the data was split in a session-wise manner, with Session 1 used for training, Session 2 for validation, and Session 3 for testing, consistent with the GNN experimental protocol.

Table 6. EEGNet experiments.

Experiment Name	Learning Rate	Weight Decay	F1
COG_BCIe1	0.0005	0.001	16
COG_BCIe2	0.0005	0.0001	16
COG_BCIe3	0.001	0.001	16
COG_BCIe4	0.001	0.0001	16
COG_BCIe5	0.0005	0.001	8
COG_BCIe6	0.0005	0.0001	8
COG_BCIe7	0.001	0.001	8
COG_BCIe8	0.001	0.0001	8

2.6.6. Optimal Transport Domain Adaptation

OT domain adaptation was applied to the experiments listed in Section 2.6.1, Table 2. Table 7 maps each of those experiments to its corresponding OT-adapted variant.

Table 7. Optimal transport domain adaptation experiments.

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

2.6.7. gCorrelation Alignment (CORAL) Domain Adaptation

We evaluated CORAL in the same baseline experimental configuration as OT (COG_BCIg5). For a fair comparison, this CORAL experiment (COG_BCIg5_CORAL) used Session 1 for training, Session 2 for validation, and Session 3 for testing with PSD power features extracted from 4-second epochs with 50% overlap. Because CORAL operates on centered second-order statistics, z-score normalization is required to satisfy the assumptions of the closed-form solution. The normalization was applied prior to adaptation and using only train dataset statistics. To ensure numerical stability of covariance whitening, CORAL was initialized with $\lambda = 10^{-5}$.

2.6.8. Comparison with the State-of-the-Art Literature on the COG-BCI Dataset

Publications utilizing the COG-BCI dataset were examined to assess their suitability for comparison. Many were not directly comparable due to methodological and experimental design differences. For instance, one study performed cross-task binary classification while including 20 s of data from the test session for calibration during training, which limits its relevance for session-independent evaluation [74]. Another study trained models using data from multiple sessions of the same subject combined with data from other subjects [75]. Although this approach does not explicitly violate session-independent separation, it differs from the present design, which enforces subject- and session-specific training–test splits to evaluate subject-dependent performance. Additionally, neither of these studies provided publicly available source code, making independent validation challenging.

One study of particular relevance applied the FOOOF (Fitting Oscillations and One-Over-F) algorithm [76] to decompose EEG power spectral density (PSD) into periodic and aperiodic components and evaluated their use in cross-session classification, where one day's data was used for cross-validation and second day data was used for test [77]. In that work, binary classification (rest vs. task) and three-class classification (rest vs. low MWL vs. high MWL) were performed on self-collected data. The results demonstrated that periodic features yielded higher accuracies than both raw PSD and aperiodic features. Specifically, periodic features achieved 84.2% ± 11.0% in the binary case and 58.9% ± 11.5% in the three-class case, whereas raw PSD features reached 69.9% ± 18.5% and 52.7% ± 14.0%, respectively. Additionally, cross-day performance on the COG-BCI dataset was reported using only binary classification between rest and MATB task conditions, with periodic features achieving substantially higher accuracy (93.8% ± 4.6%) than raw PSD features (73.4% ± 25.4%) when using an RBF kernel SVM. These results suggest that separating oscillatory activity from the 1/f-like background may improve workload classification. The study reported the use of the HAPPE preprocessing pipeline for the COG-BCI dataset with parameter settings adapted to their data, although the precise configuration was not described in detail. As a result, exact replication and controlled comparison are not straightforward. Previous evaluations of the HAPPE pipeline suggest that its performance is context-dependent, with no single preprocessing pipeline consistently outperforming others across studies [78,79]. Accordingly, while preprocessing choices may contribute to the observed results, further systematic investigation would be valuable to clarify the specific factors driving the reported improvements.

To explore the potential utility of the features within a controlled and reproducible setting, we extracted periodic and aperiodic components using the FOOOF approach and incorporated them into our graph-based classification pipeline. Importantly, we preserved our preprocessing pipeline and maintained the same train/validation/test session splits, which were consistently applied in all other experiments in this study, allowing us to isolate the contribution of the features rather than changes in data handling. Unlike [77], we applied both three-class (easy, medium, difficult) and binary (low vs. high workload) classification setups, reflecting the workload distinctions of the MATB task rather than the rest-versus-task design.

Another recent study proposes a deep learning architecture called EEGMeNet with four key components: local feature learner, global feature learner, feature preserver module, and supervised learning module [80]. The performance is assessed using four datasets, where one of them is MATB task data of COG-BCI dataset. The study first reports cross-subject (where all subjects except one are used for training) and within-subject (each subject's data is split in a 70-20-10 ratio without session-specific separation) conditions. Finally, they report a cross-session condition in a binary setting of low vs, high mental workload. They achieve a mean accuracy of 0.7177 ± 0.0378. Our binary classification with

OT reaches 0.7216 ± 0.1840 under strict session-wise separation, yielding comparable mean accuracy, though differences in preprocessing pipelines and experimental protocol may prevent direct quantitative conclusions.

2.7. Computational Resources

In this study, the classification of mental workload using GNN models was carried out and evaluated in Python (v3.8) within a notebook-based environment. Experiments were conducted on four workstations. COG_BCIg1 experiments were performed on a workstation equipped with Intel® Xeon® CPU, 32 GB RAM (Intel Corporation, Santa Clara, CA, USA), and an NVIDIA GeForce RTX 4070 Ti SUPER GPU (NVIDIA Corporation, Santa Clara, CA, USA). Each experiment was executed for all 26 subjects, with 150 Optuna trials per subject, and the run time of every trial was recorded. For the baseline COG_BCIg1 setup, a single trial took an average of 98.6 ± 34.0 s, calculated across all subjects and trials. The cumulative training time for COG_BCIg1 amounted to approximately 106.9 h.

3. Results

All performance metrics reported in this section represent results from the best model of each experimental setup. Values are presented as (mean ± standard) deviation across all 26 subjects. The aggregated statistics therefore reflect both model performance and inter-subject variability.

3.1. Exploratory Data Analysis on Features

In this study, SOMs [81,82] were employed to visualize the high-dimensional distributions of EEG feature vectors and to assess baseline feature alignment across sessions before any model training. The SOM serves as an unsupervised dimensionality reduction technique that preserves topological relationships from the original feature space, projecting multidimensional feature vectors onto a two-dimensional grid of nodes for ease of visualization. Importantly, the SOM was trained exclusively on data from Session 1 and then used as a fixed projection to map features from Sessions 2 and 3. This approach allows direct visual comparison of feature distributions across sessions, revealing potential domain shifts and highlighting challenges for cross-session generalization.

Although class labels are not used during SOM training, they are overlaid on the resulting plots to assess class separability visually. In the plots, the x- and y-axes represent the relative grid positions within the SOM lattice and do not correspond to physical EEG channels or specific feature dimensions. Each data point is shown as a marker colored according to its true class label as Easy, Medium, or Difficult, enabling inspection of whether workload classes form distinct clusters or exhibit overlap in the feature space. Figure 2 present SOM visualizations for Subject 1, with each column displaying data from Sessions 1, 2, and 3, respectively. In this figure, the top row shows the SOM mappings of the original features before applying OT, while the bottom row displays the mappings after OT. Well-separated clusters of similarly colored points indicate effective feature discrimination and good alignment across sessions, while overlapping regions suggest distributional differences that may limit classification performance. Thus, SOM plots provide valuable insights into the intrinsic structure of the EEG features and inform the potential benefits of domain adaptation strategies such as OT.

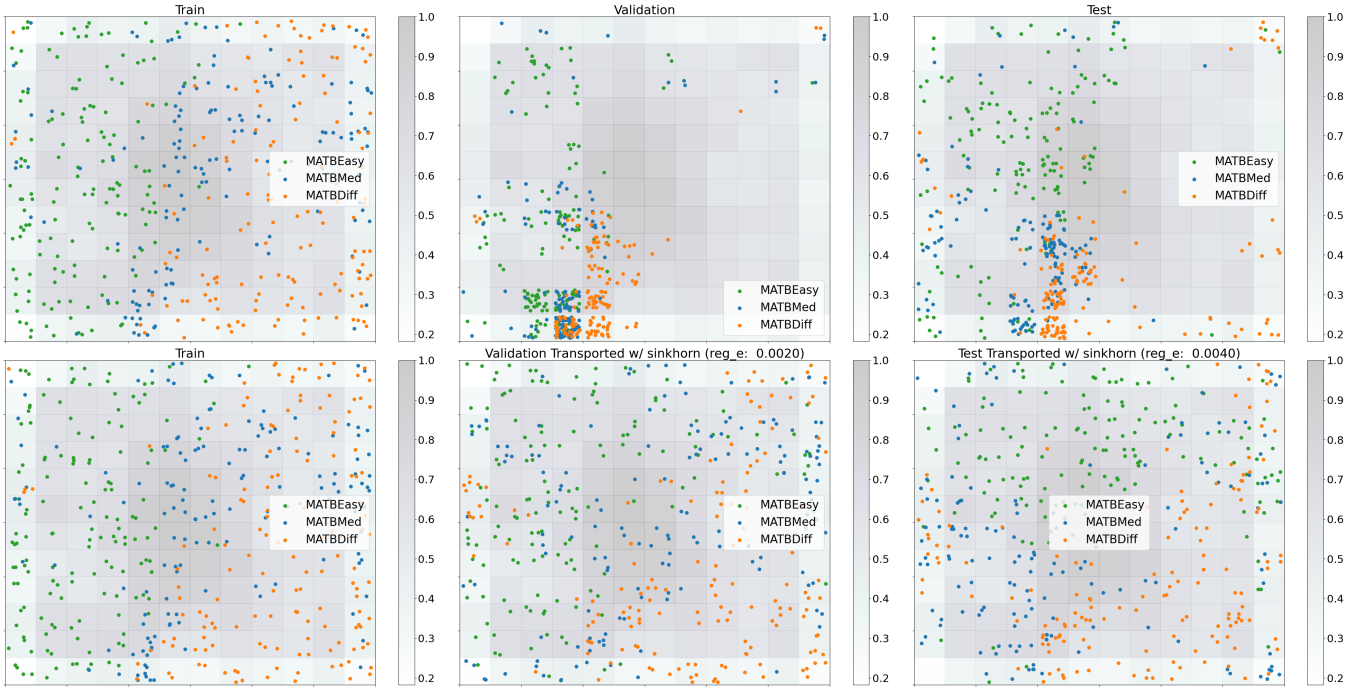

Figure 2. SOM visualizations of EEG features for subject 1. Each panel displays the projection of EEG feature vectors onto a two-dimensional SOM grid trained only on data from Session 1 (train data). The panels correspond to the train (left column), validation (middle column), and test (right column) sets. The top row shows original features and the bottom row shows features after OT. Colors indicate workload classes: MATBEasy (green), MATBMed (blue), and MATBDiff (orange).

3.2. Model Performance

Table 8 lists the mean train, validation, and test accuracies as well as precision and recall, averaged across subjects, for the best models of experiments involving different data-split strategies, while Table 9 reports the corresponding results for all other variants. Precision and recall are reported only in Table 8 for brevity. Since the dataset is balanced, classification accuracy is used as the performance metric in the remaining tables.

When averaged across subjects, all models achieved test accuracies above the adjusted chance level of 38% (for three-class classification with 148 samples per class, at the 95% confidence level [83,84]), with values ranging from 42% to 49%. Test accuracies consistently show a significant drop compared to train and validation accuracies, underscoring the impact of cross-session variability.

The Friedman test results showed no statistically significant differences in test accuracy across any of these experiments for a 95% confidence interval. This indicates that variations in data splits, epoch lengths and overlap, graph structures, or node selection strategies alone were insufficient to overcome the challenge of cross-session variability.

The contrast between cross-session performance (training accuracy 68%, validation accuracy 62% for COG_BCIg1) and the shuffled-split experiments (training accuracy > 98%, validation accuracy > 95%) is a direct quantitative indicator of cross-session distribution shift rather than a limitation of model capacity. These results confirm that the primary challenge in this task is domain mismatch across recording sessions, not insufficient model expressiveness. The high standard deviations in performance also reflect inter-subject variability.

Table 8. Results of experiments with variations in data split.

Experiment Name	Train (Mean ± Std)	Validation (Mean ± Std)	Test (Mean ± Std)	
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049	Accuracy
	0.6736 ± 0.2241	0.5940 ± 0.2091	0.4245 ± 0.1675	Precision
COG_BCIg2	0.9815 ± 0.0295	0.9535 ± 0.0374	0.4238 ± 0.1762	Accuracy
	0.9825 ± 0.0272	0.9542 ± 0.0364	0.4299 ± 0.2258	Precision
COG_BCIg3	0.5796 ± 0.1493	0.5500 ± 0.1329	0.4724 ± 0.1235	Accuracy
	0.5378 ± 0.2234	0.5013 ± 0.2039	0.4412 ± 0.2021	Precision
COG_BCIg4	0.9866 ± 0.0198	0.9551 ± 0.0342	0.4902 ± 0.1557	Accuracy
	0.9869 ± 0.0193	0.9555 ± 0.0339	0.4685 ± 0.1765	Precision

On the balanced COG-BCI dataset, macro-averaged recall equals accuracy by definition.

Table 9. Results of experiment variations compared to COG_BCIg1.

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

The performance of the best model from the binary classification experiment is presented in Table 10 together with the baseline model, COG_BCIg1, averaged across subjects. It achieves a mean test accuracy of 0.6631 for distinguishing Easy and Difficult conditions.

Table 10. Results of binary classification experiments compared to COG_BCIg1.

Experiment Name	Train Accuracy (Mean ± Std)	Validation Accuracy (Mean ± Std)	Test Accuracy (Mean ± Std)
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg1_binary	0.8744 ± 0.1567	0.8576 ± 0.1568	0.6631 ± 0.1647

3.3. Comparison with Heuristic Loss Weights

Table 11 presents the mean train, validation, and test accuracies for the baseline COG_BCIg1 experiment under two heuristic loss weightings: $w_{validation} = 0.75$ and

$w_{validation} = 1.0$, the latter corresponding to conventional validation-only model selection. The weight selection reasoning is explained in Appendix C.

The mean test accuracies are comparable between the two settings (44.08% vs. 44.51%). The standard deviation of test accuracies of multiple runs is higher under $w_{validation} = 1.0$ (0.1348 vs. 0.1049), which may indicate less consistent model selection across subjects.

Table 11. Results of heuristic loss weight experiments.

Experiment Name	Train Accuracy (Mean ± Std)	Validation Accuracy (Mean ± Std)	Test Accuracy (Mean ± Std)
COG_BCIg1 $w_{validation} = 0.75$	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg1 $w_{validation} = 1.0$	0.6811 ± 0.1898	0.6104 ± 0.1743	0.4451 ± 0.1348

3.4. Comparison with EEGNet

Table 12 lists the mean train, validation, and test accuracies, averaged across subjects, for the best models of EEGNet experiments compared to the best model of COG_BCIg1. All EEGNet models achieved lower mean test accuracies than our proposed pipeline experiment, COG_BCIg1, with the highest value being 40.92%. The session-to-session variability is also evident in EEGNet experiments with test accuracies showing a significant decline compared to train and validation accuracies. The consistent near-chance test accuracy across all eight EEGNet configurations, despite high training accuracy exceeding 93%, indicates that cross-session performance degradation is not specific to graph-based or spectral representations. As the EEGNet pipeline relied on validation accuracy for model selection, the selected models exhibit limited generalization, as suggested by the observed train–validation performance gap and consistently high training-set accuracies.

Table 12. Results of EEGNet classification experiments compared to COG_BCIg1.

Experiment Name	Train Accuracy (Mean ± Std)	Validation Accuracy (Mean ± Std)	Test Accuracy (Mean ± Std)
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	**0.4408 ± 0.1049**
COG_BCIe1	0.9654 ± 0.0430	0.5918 ± 0.1318	0.3912 ± 0.1036
COG_BCIe2	0.9589 ± 0.0451	0.5906 ± 0.1387	0.3914 ± 0.1034
COG_BCIe3	0.9877 ± 0.0145	0.6011 ± 0.1406	0.3691 ± 0.0869
COG_BCIe4	0.9863 ± 0.0136	0.5949 ± 0.1354	0.3836 ± 0.0652
COG_BCIe5	0.9346 ± 0.0726	0.6022 ± 0.1387	0.4018 ± 0.1163
COG_BCIe6	0.9284 ± 0.0735	0.6033 ± 0.1414	0.3767 ± 0.1021
COG_BCIe7	0.9728 ± 0.0424	0.5914 ± 0.1332	0.3876 ± 0.0959
COG_BCIe8	0.9667 ± 0.0489	0.5963 ± 0.1347	0.4092 ± 0.1258

Bold number in the "Test Accuracy" column indicate the highest test accuracy within the experiments.

3.5. Comparative Effectiveness of Domain Adaptation Methods

Having established the necessity of domain adaptation (Section 3.2), we now demonstrate OT performance for aligning feature distributions across sessions under variations in data split. We also compare the results to those with CORAL domain adaptation for the baseline experiment.

3.5.1. Optimal Transport Model Performance

Table 13 reports the subject-averaged train, validation, and test accuracies for the best models of session-based data-split experiments, both with and without OT.

Table 13. Results of experiments with and without optimal transport.

Experiment Name	Train Accuracy (Mean ± Std)	Validation Accuracy (Mean ± Std)	Test Accuracy (Mean ± Std)
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg1_binary	0.8744 ± 0.1567	0.8576 ± 0.1568	0.6631 ± 0.1647
COG_BCIg2	0.9815 ± 0.0295	0.9535 ± 0.0374	0.4238 ± 0.1762
COG_BCIg3	0.5796 ± 0.1493	0.5500 ± 0.1329	0.4724 ± 0.1235
COG_BCIg4	0.9866 ± 0.0198	0.9551 ± 0.0342	0.4902 ± 0.1557
COG_BCIg5 (COG_BCIg1+OT)	0.6483 ± 0.1428	0.6114 ± 0.1336	**0.5054 ± 0.1213**
COG_BCIg5_binary (COG_BCIg1_binary+OT)	0.8526 ± 0.1560	0.8316 ± 0.1505	**0.7216 ± 0.1840**
COG_BCIg6 (COG_BCIg2+OT)	0.9111 ± 0.0731	0.8604 ± 0.0755	**0.5236 ± 0.1116**
COG_BCIg7 (COG_BCIg3+OT)	0.5742 ± 0.1385	0.5274 ± 0.1149	**0.5289 ± 0.1171**
COG_BCIg8 (COG_BCIg4+OT)	0.9096 ± 0.0739	0.8539 ± 0.0695	**0.5612 ± 0.0953**

Bold numbers in the "Test Accuracy" column indicate the highest test accuracy within each pair of experiments—specifically, comparing the baseline experiment (without OT) and the corresponding experiment with OT applied.

OT consistently improves test accuracy across all four data split configurations. This consistency is notable given that the configurations differ in how training and validation sessions are assigned. Nevertheless, the absolute test accuracies after OT remain modest, ranging from 50.54% to 56.12% in the three-class setting. OT operates without label information during transport, meaning it aligns the overall feature distributions across sessions but cannot selectively correct misalignment at the class level.

The Friedman test indicated that classification accuracy differed between eight experiments (COG_BCIg1–COG_BCIg8) for train, validation, and test accuracies (all p-values < 0.001). To observe pairwise effects, Wilcoxon signed-rank tests concluded with the results given in Table 14. COG_BCIg1, COG_BCIg2, COG_BCIg3, and COG_BCIg4 experiments were compared with their respective counterparts where OT was applied, namely COG_BCIg5, COG_BCIg6, COG_BCIg7, and COG_BCIg8. The pairwise Wilcoxon tests (95% confidence interval) show that applying OT generally leads to significantly higher test accuracies compared to not using OT, with the exception of the COG_BCIg3–COG_BCIg7 pair, where the difference was not significant (p = 0.07). As presented in Table 13, no significant differences were observed in training and validation accuracies for the COG_BCIg1–COG_BCIg5 and COG_BCIg3–COG_BCIg7 pairs. In contrast, significant decreases in accuracy were found for the COG_BCIg2–COG_BCIg6 and COG_BCIg4–COG_BCIg8 pairs when OT was applied.

Table 14. Wilcoxon test results (p-values) for best models with and without OT.

Experiments	Train Accuracy	Validation Accuracy	Test Accuracy
COG_BCIg1 - COG_BCIg5	0.22	0.97	**<0.005**
COG_BCIg2 - COG_BCIg6	**<0.001**	**<0.001**	**<0.01**
COG_BCIg3 - COG_BCIg7	0.89	0.30	0.07
COG_BCIg4 - COG_BCIg8	**<0.001**	**<0.001**	**<0.05**

Bold numbers indicate statistical significance.

For a more comprehensive assessment, subject-level performance is presented in Appendix C for Subject 1 and experiments COG_BCIg1 and COG_BCIg5. The results highlight the challenge of cross-session generalization, with the OT approach helping performance.

3.5.2. Comparison with CORAL

Table 15 summarizes the mean train, validation, and test accuracies of the baseline model without adaptation (COG_BCIg1), the model with OT adaptation (COG_BCIg5), and the model with CORAL adaptation (COG_BCIg5_CORAL), averaged across subjects. Among the evaluated approaches, OT achieves the highest mean test accuracy (50.54%), indicating improved cross-session generalization compared to the baseline. In contrast, CORAL adaptation results in lower test performance than the nonadapted baseline, suggesting that covariance alignment alone is insufficient to mitigate session mismatch under the present experimental conditions. Across experiments, a clear train–test performance gap persists, reflecting substantial session-to-session variability in EEG data.

Table 15. Results of experiments without adaptation, OT adaptation, and CORAL adaptation.

Experiment Name	Train Accuracy (Mean ± Std)	Validation Accuracy (Mean ± Std)	Test Accuracy (Mean ± Std)
COG_BCIg1	0.6786 ± 0.1826	0.6174 ± 0.1659	0.4408 ± 0.1049
COG_BCIg5 (COG_BCIg1+OT)	0.6483 ± 0.1428	0.6114 ± 0.1336	**0.5054 ± 0.1213**
COG_BCIg5_CORAL (COG_BCIg1+CORAL)	0.5058 ± 0.1655	0.4305 ± 0.1106	0.3307 ± 0.0438

Bold number in the "Test Accuracy" column indicate the highest test accuracy within the experiments.

3.6. Results of the Replication Experiment

In the three-class setting (easy, medium, difficult), our best model of the baseline experiment (COG_BCIg1) achieved 44.08% ± 10.49% test accuracy with raw PSD features and when session 1, session 2, and session 3 data were used as train, validation, and test sets, respectively (Table 8). Incorporating periodic and aperiodic features yielded comparable results, with test accuracies of 41.97% ± 7.26% and 42.57% ± 10.78%, respectively. In the binary classification setup (easy vs. difficult MATB level), raw PSD features achieved 66.31% ± 16.47% test accuracy, while periodic features obtained 64.35% ± 11.56%. These outcomes differ substantially from the results reported by Ke et al. [77], where periodic features consistently outperformed raw PSD, particularly in binary rest-versus-task classification. In our setting, improvements from periodic features were not observed.

The discrepancy likely reflects differences in task formulation (rest vs. task versus workload-level classification), modeling frameworks (SVM versus graph-based models), and preprocessing strategies. Since [77] did not provide detailed parameterization of the HAPPE preprocessing pipeline, exact replication is not straightforward, and the ex-

tent to which preprocessing choices contributed to their findings remains unclear. Our findings suggest that while FOOOF-derived features are a promising idea, further systematic investigation is warranted to clarify their robustness under session-independent and workload-level classification setups.

4. Discussion

4.1. Exploratory Data Analysis on Features

The SOM plot for Subject 1 reveals that the distributions of the validation and test sets differ markedly from that of the training set, highlighting the session-specific variability in the data. Moreover, regions of overlap show the hardship in distinguishing between classes. Samples labeled as "Easy" predominantly cluster toward the left region of the SOM lattice for the train session, indicating relative consistency in this workload level's feature distribution. In contrast, "Medium" workload samples are more diffusely scattered throughout the map, with noticeable overlap into regions also occupied by "Difficult" samples. This suggests lower separability between Medium and Difficult classes for Subject 1. For the validation and test sessions, Medium and Difficult classes noticeably overlap. After the application of OT, the validation session shows better separability for the Difficult class; however, some overlap between Easy and Medium classes prevail. The Difficult class in the test session also shows better separability; however, Easy and Medium conditions swap considerably on the SOM grid. These visualizations underscore why model generalization remains challenging despite domain adaptation, particularly for Medium workload conditions.

4.2. Model Performance

COG_BCIg1 attains mean accuracies of 68% on the training set and 62% on the validation set, while COG_BCIg3 achieves 58% and 55%, respectively. In both experiments, the models fail to generalize, with test accuracies dropping to 44% for COG_BCIg1 and 47% for COG_BCIg3. When training and validation samples are shuffled and randomly re-split, the average training and validation accuracies exceed 98% and 95%, respectively, and the across-subject standard deviations shrink from 18% and 15% to 3% and 2%. This sharp contrast indicates that the weak training and validation performances of COG_BCIg1 and COG_BCIg3 stems mainly from cross-session variability between the original training and validation sessions, a conclusion supported by the SOM feature maps in Section 4.1.

Similarly, COG_BCIg2 and COG_BCIg4 exhibit reduced test accuracies compared to their training and validation counterparts, with no substantial improvement over the test performance of COG_BCIg1 and COG_BCIg3, further reinforcing the effect of cross-session variability. The discrepancy between unshuffled and shuffled results also highlights the importance of evaluating models on an unseen test session to obtain a realistic measure of generalizability.

The impact of cross-session variability remains evident in the other COG_BCIg1 variants, which again fail to generalize to the test set. Although test accuracies remained relatively low (42–49%), they consistently exceed the adjusted chance level of 38%, suggesting that the models captured some degree of relevant signal structure. These results illustrate the importance of systematic experimental design and rigorous evaluation, a key contribution of this study's data-centric approach. Our use of a heuristic loss function contributed to the comprehensive evaluation strategy by enabling consistent selection of models that avoided overfitting or yielded more representative results. Table 11 shows that while mean test accuracies are comparable between $w_{validation} = 0.75$ and $w_{validation} = 1.0$, the former yields lower variance (std: 0.1049 vs. 0.1348), indicating more stable model selection across subjects. As illustrated for Subject 5 (Figure A1), conventional validation-only selection can favor early-stopping points where validation loss falls well below training loss before meaningful

learning has occurred. The heuristic loss penalizes such cases by incorporating the validation loss weight and train–validation discrepancy.

The difficulty of cross-session mental workload classification is further supported by the performance of EEGNet, which operates on epoched time-domain EEG signals. Despite its widespread use and strong within-session performance reported in prior studies, EEGNet exhibited a substantial reduction in accuracy under the session-wise evaluation protocol employed in this work. This finding indicates that cross-session performance degradation is not specific to graph-based or spectral representations but instead reflects the broader nonstationarity of workload-related EEG patterns across recording sessions.

4.3. Optimal Transport Model Performance

OT was applied for domain adaptation using EEG data in sleep staging [31,32] and emotion recognition for transfer across subjects. In motor imagery, it has been used to successfully adapt for cross-subject or cross-session estimation [34–36]. In this study, we examined its application to cross-session mental workload classification.

In line with these studies, applying OT raised the mean test accuracy, demonstrating improved generalization in every case. The Friedman test reveals statistically significant differences in classification accuracies for the various models. The pairwise Wilcoxon tests indicate statistically significantly higher test accuracies when OT was applied compared to when it was not for all pairs except COG_BCIg3–COG_BCIg7. In this case, significance was marginal with $p = 0.07$. As shown in Table 14, for train and validation accuracies, no significant differences were observed for the COG_BCI g1–COG_BCIg5 and COG_BCIg3–COG_BCIg7 pairs. However, significant changes were detected for the COG_BCIg2–COG_BCIg6 and COG_BCIg4–COG_BCIg8 pairs where adding OT results in lower accuracies. This behavior aligns with expectations, as the training and validation sets for COG_BCIg2 and COG_BCIg4 were generated by shuffling two sessions, thereby minimizing variability between these sets and resulting in already well-aligned train and validation data. In these cases, applying OT without label information led to artificially increased distributional differences.

The visual findings from SOM plots align with the confusion-matrix results given in Table A2 for subject 1. For the train set, the nonadapted model (COG_BCIg1) achieved an accuracy of 67.79%, with high true positives for the Easy (117) and Difficult (127) classes but lower correct predictions for Medium (57), reflecting the observed overlap on the SOM. With OT, COG_BCIg5 further improved train accuracy to 73.20%, with increased true positives for both Easy (134) and Medium (92) classes and reduced misclassifications compared to COG_BCIg1, indicating better feature separability after OT. In the validation set, COG_BCIg1 showed a decreased accuracy of 58.11%, with substantial confusion between Medium and Difficult classes, while COG_BCIg5 increased validation accuracy to 68.47%, with clearer clusters for the Difficult class as observed in the SOM plots. For the test set, both models showed diminished accuracies (47.52% for COG_BCIg1 and 48.20% for COG_BCIg5). Although the Difficult class maintained higher true positives in COG_BCIg5 (113), Easy and Medium samples were more frequently misclassified, consistent with the overlapping and swapping patterns evident in the SOM plots. Overall, the SOM visualizations and confusion matrices collectively highlight the challenge of achieving consistent separability across sessions, particularly between Easy and Medium workload conditions, despite the improvements conferred by OT. OT improves class separability, yet a residual label shift remains because true class labels are not known during the transport step. Similar results were observed for the other participants.

Comparisons with covariance-based domain adaptation further clarify the role of OT in mitigating session mismatch. The performance drop under CORAL below

the nonadapted baseline warrants analysis. With 244-dimensional PSD features and 444 training samples per subject, the covariance matrix is substantially underdetermined, meaning the whitening transformation may amplify noise directions rather than suppress them, actively distorting the feature geometry. Furthermore, CORAL's core assumption that session shift is predominantly captured by second-order covariance structure may not hold for EEG workload data, where distributional shifts involve complex higher-order differences that covariance alignment alone cannot address. In contrast, OT operates on the full distributional geometry via the Wasserstein distance, making it better suited to the complex, high-dimensional nature of cross-session EEG shifts.

The unsupervised nature of OT limits complete alignment for certain classes, because no source or target labels are used during transport. Incorporating a small proportion of labeled data in a future semi-supervised OT scheme could improve alignment for the more confounded classes and further boost cross-session generalization. Overall, this analysis emphasizes the value of integrating SOM-based data exploration with iterative refinements in preprocessing and modeling to develop more interpretable and robust EEG-based workload classification models. The analysis shows how SOM plots complement confusion matrices by exposing class-level separability and session effects. Such exploratory visualization informs feature-engineering and model-selection decisions, ultimately enhancing interpretability and robustness.

4.4. Replication Experiment

The replication experiments show that FOOOF-derived features can be incorporated in graph-based models; however, performance improvements of the magnitude reported by [77] were not observed under the present task and modeling framework. The replication was carried out under the assumption that the reported benefits should generalize beyond the specific question and classifier employed in the original study. For this reason, we evaluated these features within our proposed graph-based modeling approach to explore our question of differentiating between different MATB-II levels rather than re-implementing the support vector machine (SVM) used by [77].

Similarly, the original preprocessing pipeline was not adopted; instead, we applied our preprocessing approach to maintain consistency across all conditions examined in this study. While these design decisions introduce some divergence from the original setup, they also reduce potential confounding factors and make the comparisons within our framework more transparent. These choices ensured comparability with the rest of our experiments and allowed assessment of whether the gains extend to more complex, representation-learning architectures.

Contrary to the original findings, periodic features did not outperform either raw PSD or aperiodic features. This divergence suggests that the utility of such features is not intrinsic but contingent on the modeling framework and task design. The discrepancy illustrates that findings that appear compelling within a particular pipeline may not translate directly to alternative approaches. By situating our results alongside the prior literature, it becomes evident that while FOOOF-derived features remain a promising methodological option, their contribution should be interpreted with respect to context.

5. Conclusions

This study addressed the challenge of cross-session generalization in EEG-based mental workload classification by integrating OT domain adaptation with GNNs. To our knowledge, this is the first work to combine OT and GNNs for this application, leveraging GNNs to model spatial EEG relationships while using OT to align feature distributions across sessions without retraining model parameters. Comprehensive experiments on the publicly available COG-BCI dataset demonstrated that models suffered notable performance drops on unseen

sessions, highlighting session-related domain shifts. SOM visualizations further revealed variability and overlap in class structures across sessions, underscoring the necessity of domain adaptation. The inclusion of EEGNet and CORAL baselines further demonstrates that both time-domain deep learning models and covariance-based alignment methods remain sensitive to session-induced distribution shifts in mental workload decoding.

Incorporating OT consistently improved test accuracies across all experimental configurations, yielding mean accuracy gains of up to 6.46% in three-class classification and 5.85% in binary classification, demonstrating OT's potential for improving cross-session generalization. Under strict session-wise evaluation, where test-session data does not inform training or adaptation, our method for three-class classification achieved a 12.5% improvement over the adjusted chance level of 38% and a statistically significant 6.46% gain over the nonadapted baseline ($p < 0.005$, Wilcoxon). This improvement is both statistically significant and practically meaningful. In operational BCI contexts where recalibration is costly, a 6.5% gain in cross-session accuracy without any model retraining represents a valuable reduction in session-induced error.

Importantly, this study contributed a data-centric and reproducible evaluation pipeline, including rigorous exploratory analyses, SOM-based visualizations, a heuristic loss function for robust model selection, and systematic assessments across varied experimental conditions. These methodological contributions offer valuable tools for researchers aiming to develop EEG-based workload classifiers that can operate reliably in real-world settings.

Nevertheless, practical challenges remain. EEG datasets are costly to acquire and label due to the time, expertise, and equipment required. Moreover, mental workload is difficult to quantify objectively. Task difficulty does not always reflect user effort due to individual factors such as fatigue or disengagement, which introduces variability and noise into both training and evaluation. These factors complicate model development, especially in real-world applications where such variability is harder to control, often leading to baselines with low performance.

Furthermore, mutual information-based node selection, while performed strictly on training set labels, implicitly assumes cross-session stationarity of workload-relevant EEG channels. If the spatial distribution of informative channels shifts across sessions, the selected feature subspace may not generalize optimally.

Future research should validate these findings on independent datasets and explore semi-supervised or adaptive OT strategies that incorporate limited labeled data to further improve feature alignment and generalization, as well as investigate scalable solutions suitable for real-time BCI deployment. Inter-subject variability remains a significant open challenge, as reflected in the standard deviation of test accuracies across subjects and the modest absolute accuracy achieved even after OT adaptation. This suggests that session-level alignment alone is insufficient for all individuals. Subject-specific adaptation strategies, which tailor the transport to individual neural profiles rather than applying a uniform alignment, represent a promising direction for future work. Additionally, architecture-level ablation is a limitation of this work, as we used one graph convolution architecture. Future work should compare the convolution operators itself. Overall, the proposed framework represents a rigorous step toward robust, session-invariant EEG-based mental workload classification, contributing to the advancement of practical brain–computer interface systems.

Author Contributions: Conceptualization, G.D., A.-M.B., and T.T.T.; methodology, G.D. and T.T.T.; software, G.D.; validation, G.D. and T.T.T.; formal analysis, G.D.; writing—original draft preparation, G.D.; writing—review and editing, G.D., A.-M.B., and T.T.T.; visualization, G.D.; supervision, A.-M.B. and T.T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: COG-BCI dataset is publicly available in [53].

Acknowledgments: The author(s) declare financial support was received for the research of this article. This study was supported by Yüksek Öğretim Kurulu (YÖK) 100/2000 scholarship.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CORAL	CORrelation ALignment
GNN	Graph Neural Network
ICA	Independent Component Analysis
OT	Optimal Transport
SOM	Self-Organizing Map

Appendix A. Mathematical Background on Optimal Transport

OT theory, originating from Gaspard Monge's 1781 formulation, addresses the problem of optimally reallocating mass from one distribution to another while minimizing a given cost function [46]. Monge's original problem seeks a transport map $T : \mathcal{X} \to \mathcal{Y}$ that pushes a source measure α onto a target measure β, minimizing the total cost:

$$\inf_{T_{\#}\alpha=\beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{A1}$$

However, this formulation is often ill-posed due to the nonlinearity and nonconvexity of the constraint set and may not always admit a solution [46,85]. To address this, Leonid Kantorovich [47] introduced a relaxed version, allowing for transport plans $\pi \in \Pi(\alpha, \beta)$, which are joint distributions with marginals α and β. The Kantorovich problem is then written as:

$$\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y), \tag{A2}$$

Unlike Monge's formulation, the Kantorovich approach is guaranteed to be feasible and yields a solution under relatively mild assumptions [47,85]. This relaxation leads to a convex optimization problem and admits a dual formulation:

$$\sup_{(f,g)} \left\{ \int_{\mathcal{X}} f(x) d\alpha(x) + \int_{\mathcal{Y}} g(y) d\beta(y) \mid f(x) + g(y) \leq c(x, y) \right\}, \tag{A3}$$

The Kantorovich approach makes it possible to define a metric on probability measures, called the Wasserstein distance. This metric can capture differences between distributions even when their supports do not overlap and can better capture geometric properties between distributions compared to other divergence metrics, such as Kullback–Leibler divergence [46,48]. For a cost function $c(x, y) = d(x, y)^p$, where $d(\cdot, \cdot)$ is a metric on $\mathcal{X}$, the p-Wasserstein distance between α and β is defined as:

$$W_p(\alpha, \beta) = \left(\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} d(x, y)^p d\pi(x, y) \right)^{1/p}. \tag{A4}$$

The Wasserstein metric possesses a special differentiable structure, known as the Riemannian structure of Wasserstein space, which makes it well suited for optimization tasks involving probability measures. This property enables the use of gradient flow-based

algorithms, establishing OT as a valuable tool for continuous optimization over probability distributions [86]. The Wasserstein distance enables meaningful comparisons between distributions, even when they lack overlapping supports or exhibit label shifts [87,88].

Kantorovich formulation is limited in its computational cost, especially for large datasets. Entropic regularization addresses this by adding an entropy term to the objective, leading to the regularized OT problem:

$$L_\epsilon(\alpha, \beta) = \inf_{\pi \in \Pi(\alpha,\beta)} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) + \epsilon H(\pi) \right], \tag{A5}$$

where $H(\pi) = -\int_{\mathcal{X}\times\mathcal{Y}} \pi(x,y) \log \pi(x,y)\, dx\, dy$ is the entropy of the transport plan π, and $\epsilon > 0$ controls regularization. Regularized variants such as entropy-constrained OT improve computational efficiency for high-dimensional data [49]. This smooths the transport plan and enables efficient algorithms like Sinkhorn's method, which leverages strict convexity of the objective term when regularization is introduced [49]. Entropic regularization is valuable in machine learning, allowing faster computation of distances between distributions and integration into models such as neural networks, while offering a balance between accuracy of the transport plan and computational efficiency with the ϵ regularization parameter [50].

The Sinkhorn algorithm efficiently solves the entropic OT problem by iteratively scaling the rows and columns of a kernel matrix $K = e^{-C/\epsilon}$, where C is the cost matrix. Starting with scaling vectors u and v initialized to $\mathbf{1}$, the updates proceed as:

$$u \leftarrow \frac{\alpha}{Kv}, \quad v \leftarrow \frac{\beta}{K^T u}, \tag{A6}$$

with divisions performed elementwise. The iterations continue until convergence, yielding an approximate transport plan more efficiently than traditional linear programming [50]

Sinkhorn's method is differentiable, parallelizable, and its convergence behavior is well understood, though it may struggle with very large distributions [51]. Its entropy term also adds robustness to noise by smoothing the transport plan, beneficial for applications with data prone to noise like EEG analysis in brain–computer interfaces.

Appendix B. Implementation Details

Table A1. Hyperparameter ranges.

Parameter Name	Range
activation function	['gelu', 'relu', 'tanh']
batch size	(16, 64)
learning rate	$(1 \times 10^{-4}, 1 \times 10^{-2})$
weight decay	$(1 \times 10^{-4}, 1 \times 10^{-1})$
number of graph layers	(1, 4)
number of nodes in each graph layer	(4, 32)
number of linear layers	(2, 5)
number of nodes in each linear layer	(8, 64)

Appendix C. Individualized and Extended Results

Appendix C.1. Heuristic Loss Weight Selection

Figure A1 illustrates the loss and accuracy plots for Subject 5. Under $w_{validation} = 1.0$, the selected epoch (55) occurs early in training where validation loss falls below training loss, indicating the model has not yet learned meaningful representations, a selection case that the heuristic loss is designed to penalize. Under $w_{validation} = 0.75$, a later epoch (181) is selected where training loss has converged and validation loss is only slightly higher, reflecting a more stable and generalizable model.

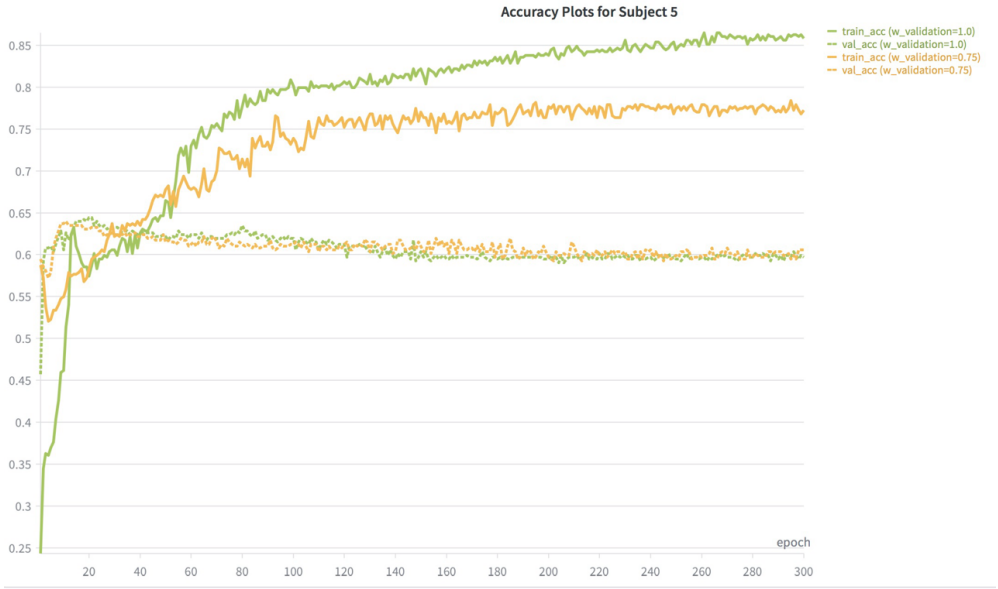

Figure A1. COG_BCIg1 performance over epochs for Subject 5.

Appendix C.2. Subject-Level Performance

Subject-level accuracy values as well as confusion matrices for train, validation, and test sets in experiments COG_BCIg1 and COG_BCIg5 are presented in Table A2 for Subject 1, Table A3 for Subject 8, and Table A4 for Subject 12. In these tables, accuracy and confusion matrices for the best models of the corresponding experiments are tabulated. SOM plots before and after OT are presented in Figure A2 for Subject 8 and in Figure A3 for Subject 12.

For Subject 1, experiments show accuracy drops on the validation and test sets compared to training set. While COG_BCIg5 achieved higher training and validation accuracies compared to COG_BCIg1, both experiments exhibited comparable performance on the test set. The confusion matrices reveal that certain classes are consistently misclassified across experiments, particularly in the test data, suggesting persistent class overlap or subject-specific variability in EEG features.

Table A2. Experiment results for Subject 1.

Experiment	Accuracy			Confusion Matrix		
	Train	Validation	Test	Train	Validation	Test
COG_BCIg1	0.6779	0.5811	0.4752	$\begin{bmatrix} 117 & 13 & 18 \\ 27 & 57 & 64 \\ 3 & 18 & 127 \end{bmatrix}$	$\begin{bmatrix} 136 & 7 & 5 \\ 62 & 35 & 51 \\ 37 & 24 & 87 \end{bmatrix}$	$\begin{bmatrix} 102 & 25 & 21 \\ 90 & 35 & 23 \\ 33 & 41 & 74 \end{bmatrix}$
COG_BCIg5	0.7320	0.6847	0.4820	$\begin{bmatrix} 134 & 14 & 0 \\ 25 & 92 & 31 \\ 2 & 47 & 99 \end{bmatrix}$	$\begin{bmatrix} 105 & 38 & 5 \\ 28 & 78 & 42 \\ 1 & 26 & 121 \end{bmatrix}$	$\begin{bmatrix} 48 & 85 & 15 \\ 56 & 53 & 39 \\ 12 & 23 & 113 \end{bmatrix}$

For Subjects 8 and 12, experiments show accuracy drops on the validation and test sets compared to training set. While COG_BCIg5 achieved more balanced training and validation accuracies compared to COG_BCIg1, the performance on the test set increased.

Table A3. Experiment results for Subject 8.

Experiment	Accuracy			Confusion Matrix		
	Train	Validation	Test	Train	Validation	Test
COG_BCIg1	0.8063	0.7568	0.4144	$\begin{bmatrix} 141 & 7 & 0 \\ 4 & 133 & 11 \\ 1 & 63 & 84 \end{bmatrix}$	$\begin{bmatrix} 145 & 3 & 0 \\ 1 & 90 & 57 \\ 0 & 47 & 101 \end{bmatrix}$	$\begin{bmatrix} 142 & 6 & 0 \\ 136 & 12 & 0 \\ 44 & 74 & 30 \end{bmatrix}$
COG_BCIg5	0.7635	0.7275	0.6419	$\begin{bmatrix} 135 & 13 & 0 \\ 6 & 102 & 40 \\ 1 & 45 & 102 \end{bmatrix}$	$\begin{bmatrix} 138 & 10 & 0 \\ 3 & 105 & 40 \\ 0 & 68 & 80 \end{bmatrix}$	$\begin{bmatrix} 94 & 45 & 9 \\ 47 & 92 & 9 \\ 0 & 49 & 99 \end{bmatrix}$

Table A4. Experiment results for Subject 12.

Experiment	Accuracy			Confusion Matrix		
	Train	Validation	Test	Train	Validation	Test
COG_BCIg1	0.8243	0.6937	0.4032	$\begin{bmatrix} 122 & 19 & 7 \\ 25 & 113 & 10 \\ 11 & 6 & 131 \end{bmatrix}$	$\begin{bmatrix} 129 & 15 & 4 \\ 25 & 68 & 55 \\ 3 & 34 & 111 \end{bmatrix}$	$\begin{bmatrix} 59 & 60 & 29 \\ 0 & 12 & 136 \\ 0 & 40 & 108 \end{bmatrix}$
COG_BCIg5	0.6937	0.6779	0.5338	$\begin{bmatrix} 116 & 25 & 7 \\ 39 & 66 & 43 \\ 6 & 16 & 126 \end{bmatrix}$	$\begin{bmatrix} 133 & 13 & 2 \\ 25 & 67 & 56 \\ 10 & 37 & 101 \end{bmatrix}$	$\begin{bmatrix} 120 & 16 & 12 \\ 15 & 48 & 85 \\ 32 & 47 & 69 \end{bmatrix}$

Overall, the confusion matrices reveal that certain classes are consistently misclassified across experiments, particularly in the test data, suggesting persistent class overlap or subject-specific variability in EEG features. Overlapping regions can also be observed in the SOM plots before and after OT, consistent with the corresponding confusion matrices.

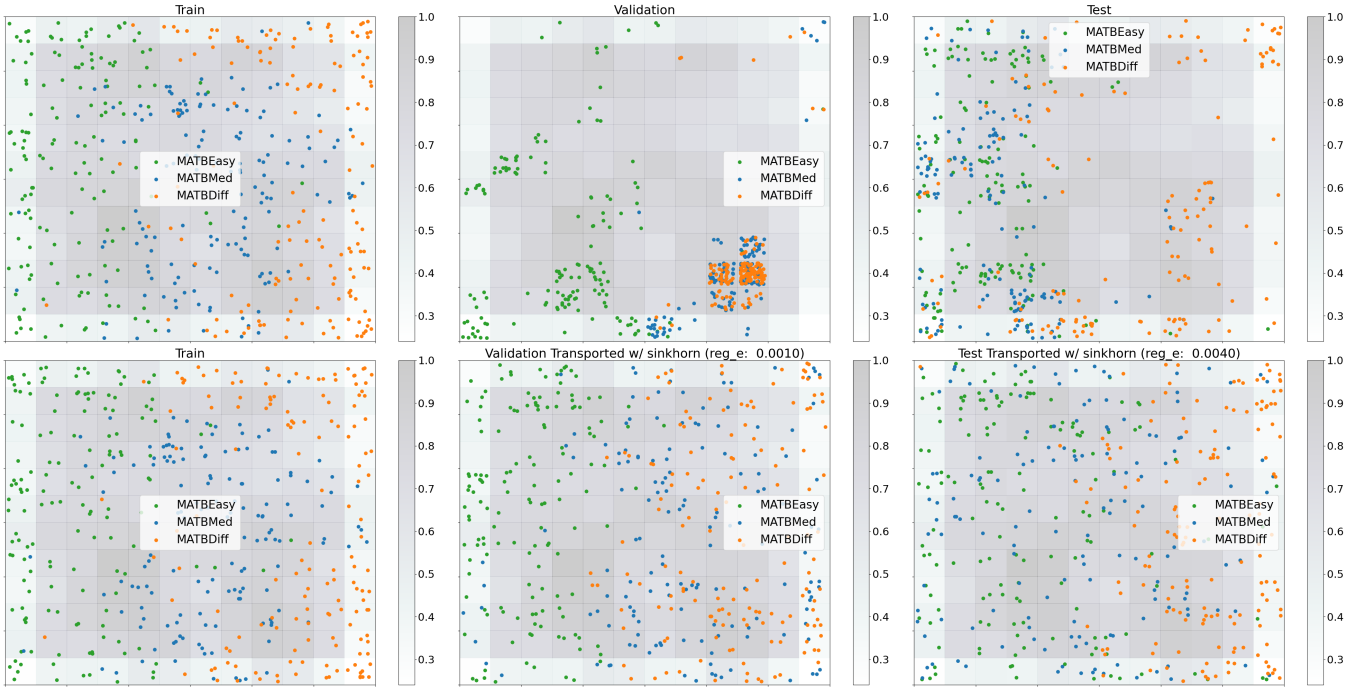

Figure A2. SOM visualizations of EEG features for Subject 8.

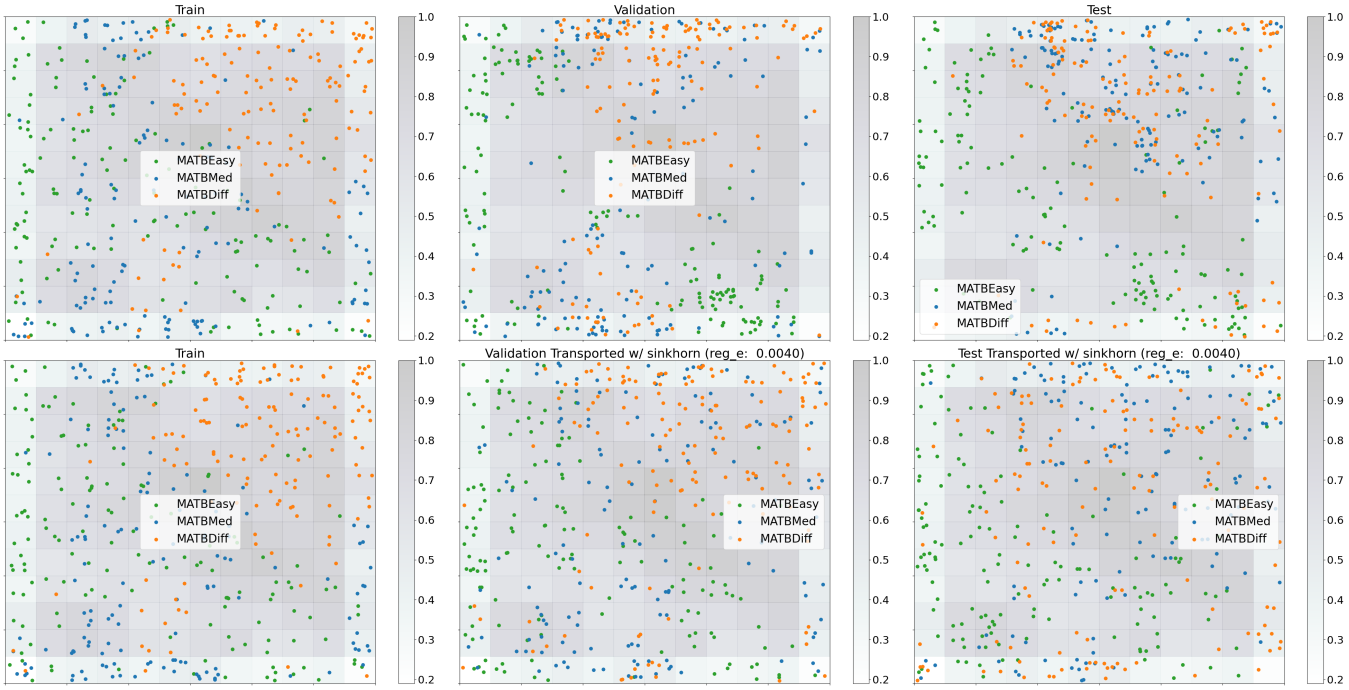

Figure A3. SOM visualizations of EEG features for Subject 12.

References

1. Young, M.S.; Brookhuis, K.A.; Wickens, C.D.; Hancock, P.A. State of science: Mental workload in ergonomics. *Ergonomics* **2015**, *58*, 1–17. [CrossRef]
2. Wickens, C.D. Multiple resources and mental workload. *Hum. Factors* **2008**, *50*, 449–455. [CrossRef]
3. Galy, E.; Cariou, M.; Mélan, C. What is the relationship between mental workload factors and cognitive load types? *Int. J. Psychophysiol.* **2012**, *83*, 269–275. [CrossRef]
4. Gevins, A.; Smith, M.E.; Leong, H.; McEvoy, L.; Whitfield, S.; Du, R.; Rush, G. Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods. *Hum. Factors* **1998**, *40*, 79–91. [CrossRef] [PubMed]
5. Klimesch, W. EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Res. Rev.* **1999**, *29*, 169–195. [CrossRef] [PubMed]
6. Holm, A.; Lukander, K.; Korpela, J.; Sallinen, M.; Müller, K.M. Estimating brain load from the EEG. *Sci. World J.* **2009**, *9*, 639–651. [CrossRef] [PubMed]
7. Brouwer, A.M.; Hogervorst, M.A.; Van Erp, J.B.; Heffelaar, T.; Zimmerman, P.H.; Oostenveld, R. Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* **2012**, *9*, 045008. [CrossRef]
8. Yu, S.; Tian, L.; Wang, G.; Nie, S. Which ERP components are effective in measuring cognitive load in multimedia learning? A meta-analysis based on relevant studies. *Front. Psychol.* **2024**, *15*, 1401005. [CrossRef] [PubMed]
9. Ghani, U.; Signal, N.; Niazi, I.K.; Taylor, D. ERP based measures of cognitive workload: A review. *Neurosci. Biobehav. Rev.* **2020**, *118*, 18–26. [CrossRef]
10. Smith, M.E.; Gevins, A.; Brown, H.; Karnik, A.; Du, R. Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Hum. Factors* **2001**, *43*, 366–380. [CrossRef]
11. Lim, W.; Sourina, O.; Wang, L.P. STEW: Simultaneous task EEG workload data set. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 2106–2114. [CrossRef]
12. Sauseng, P.; Klimesch, W.; Schabus, M.; Doppelmayr, M. Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory. *Int. J. Psychophysiol.* **2005**, *57*, 97–103. [CrossRef]
13. Plechawska-Wójcik, M.; Tokovarov, M.; Kaczorowska, M.; Zapała, D. A three-class classification of cognitive workload based on EEG spectral data. *Appl. Sci.* **2019**, *9*, 5340. [CrossRef]
14. Benerradi, J.; A. Maior, H.; Marinescu, A.; Clos, J.; L. Wilson, M. Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks. In *Proceedings of the Halfway to the Future Symposium 2019*; Association for Computing Machinery: New York, NY, USA, 2019; HttF '19; pp. 1–11. [CrossRef]
15. So, W.K.; Wong, S.W.; Mak, J.N.; Chan, R.H. An evaluation of mental workload with frontal EEG. *PloS ONE* **2017**, *12*, e0174949. [CrossRef]

16. Aghajani, H.; Garbey, M.; Omurtag, A. Measuring mental workload with EEG + fNIRS. *Front. Hum. Neurosci.* **2017**, *11*, 359. [CrossRef]
17. Baldwin, C.L.; Penaranda, B. Adaptive training using an artificial neural network and EEG metrics for within-and cross-task workload classification. *NeuroImage* **2012**, *59*, 48–56. [CrossRef] [PubMed]
18. Chandra, S.; Sharma, G.; Verma, K.; Mittal, A.; Jha, D. EEG based cognitive workload classification during NASA MATB-II multitasking. *Int. J. Cogn. Res. Sci. Eng. Educ.* **2015**, *3*, 35–42. [CrossRef]
19. Saadati, M.; Nelson, J.; Ayaz, H. Convolutional neural network for hybrid fNIRS-EEG mental workload classification. In *Proceedings of the Advances in Neuroergonomics and Cognitive Engineering*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 221–232. [CrossRef]
20. Chakladar, D.D.; Dey, S.; Roy, P.P.; Dogra, D.P. EEG -based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. *Biomed. Signal Process. Control.* **2020**, *60*, 101989. [CrossRef]
21. Asgher, U.; Khalil, K.; Khan, M.J.; Ahmad, R.; Butt, S.I.; Ayaz, Y.; Naseer, N.; Nazir, S. Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Front. Neurosci.* **2020**, *14*, 584. [CrossRef]
22. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef] [PubMed]
23. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* **2020**, *11*, 532–541. [CrossRef]
24. Zhong, P.; Wang, D.; Miao, C. EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1290–1301. [CrossRef]
25. Klepl, D.; Wu, M.; He, F. Graph Neural Network-Based EEG Classification: A Survey. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2024**, *32*, 493–503. [CrossRef]
26. Saha, S.; Baumert, M. Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review. *Front. Comput. Neurosci.* **2020**, *13*, 87. [CrossRef]
27. Wu, D.; Xu, Y.; Lu, B.L. Transfer learning for EEG -based brain–computer interfaces: A review of progress made since 2016. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 4–19. [CrossRef]
28. Ma, J.; Yang, B.; Qiu, W.; Li, Y.; Gao, S.; Xia, X. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Sci. Data* **2022**, *9*, 531. [CrossRef]
29. Huang, G.; Zhao, Z.; Zhang, S.; Hu, Z.; Fan, J.; Fu, M.; Chen, J.; Xiao, Y.; Wang, J.; Dan, G. Discrepancy between inter- and intra-subject variability in EEG-based motor imagery brain-computer interface: Evidence from multiple perspectives. *Front. Neurosci.* **2023**, *17*, 1122661. [CrossRef]
30. Montesuma, E.F.; Mboula, F.M.N.; Souloumiac, A. Recent Advances in Optimal Transport for Machine Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1161–1180. [CrossRef]
31. Chambon, S.; Galtier, M.N.; Gramfort, A. Domain adaptation with optimal transport improves EEG sleep stage classifiers. In Proceedings of the 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), Singapore, 12–14 June 2018; pp. 1–4. [CrossRef]
32. Gnassounou, T.; Flamary, R.; Gramfort, A. Convolution Monge Mapping Normalization for learning on sleep data. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 10457–10476.
33. Hamdan, B.; Aghdasi, H.S.; Salehpour, P. Semi-supervised EEG-based emotion recognition method by optimal transport distribution matching. *Eng. Res. Express* **2025**, *7*, 015272. [CrossRef]
34. Peterson, V.; Nieto, N.; Wyser, D.; Lambercy, O.; Gassert, R.; Milone, D.H.; Spies, R.D. Transfer Learning Based on Optimal Transport for Motor Imagery Brain-Computer Interfaces. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 807–817. [CrossRef] [PubMed]
35. Chen, P.; Wang, H.; Sun, X.; Li, H.; Grebogi, C.; Gao, Z. Transfer learning with optimal transportation and frequency mixup for EEG-based motor imagery recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 2866–2875. [CrossRef] [PubMed]
36. Li, J.; She, Q.; Meng, M.; Du, S.; Zhang, Y. Three-stage transfer learning for motor imagery EEG recognition. *Med. Biol. Eng. Comput.* **2024**, *62*, 1689–1701. [CrossRef] [PubMed]
37. Demirezen, G.; Taşkaya Temizel, T.; Brouwer, A.M. Reproducible machine learning research in mental workload classification using EEG. *Front. Neuroergon.* **2024**, *5*, 1346794. [CrossRef] [PubMed]
38. Kingphai, K.; Moshfeghi, Y. Mental Workload Assessment Using Deep Learning Models From EEG Signals: A Systematic Review. *IEEE Trans. Cogn. Dev. Syst.* **2025**, *17*, 40–60. [CrossRef]
39. Hinss, M.F.; Jahanpour, E.S.; Somon, B.; Pluchon, L.; Dehais, F.; Roy, R.N. *COG-BCI Database: A Multi-Session and Multi-Task EEG Cognitive Dataset for Passive Brain-Computer Interfaces*; Online Dataset; Zenodo, CERN: Geneva, Switzerland, 2022. [CrossRef]
40. Lotte, F.; Bougrain, L.; Cichocki, A.; Clerc, M.; Congedo, M.; Rakotomamonjy, A.; Yger, F. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *J. Neural Eng.* **2018**, *15*, 031005. [CrossRef]

41. Chin, Z.Y.; Zhang, X.; Wang, C.; Ang, K.K. EEG-based discrimination of different cognitive workload levels from mental arithmetic. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 1984–1987. [CrossRef]
42. Pušica, M.; Mijović, B.; Leva, M.C.; Gligorijević, I. Machine learning performance in EEG-based mental workload classification across task types: A systematic review. *Front. Neuroergon.* **2025**, *6*, 1621309. [CrossRef]
43. Cain, B. *A Review of the Mental Workload Literature*; Technical Report No. RTO-TR-HFM-121-Part-II; Defence Research and Development Canada (DRDC), Human Systems Integration Section: Toronto, ON, Canada, 2007.
44. Qu, H.; Zhang, M.; Pang, L. Mental Workload Classification Method Based on EEG Cross-Session Subspace Alignment. *Mathematics* **2022**, *10*, 875. [CrossRef]
45. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef]
46. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 338.
47. Kantorovich, L.V. On the Translocation of Masses. *J. Math. Sci.* **2006**, *133*, 1381–1382. [CrossRef]
48. Lv, J.; Yang, H.; Li, P. Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: New York, NY, USA, 2024; Volume 37, pp. 65445–65475.
49. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Proceedings of the Advances in Neural Information Processing Systems*; Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., Eds.; Curran Associates, Inc.: New York, NY, USA, 2013; Volume 26.
50. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
51. Genevay, A.; Cuturi, M.; Peyré, G.; Bach, F. Stochastic Optimization for Large-scale Optimal Transport. In *Proceedings of the Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2016; Volume 29.
52. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. *Proc. Aaai Conf. Artif. Intell.* **2016**, *30*, 1. [CrossRef]
53. Hinss, M.F.; Jahanpour, E.S.; Somon, B.; Pluchon, L.; Dehais, F.; Roy, R.N. Open multi-session and multi-task EEG cognitive Dataset for passive brain-computer Interface Applications. *Sci. Data* **2023**, *10*, 85. [CrossRef]
54. Santiago-Espada, Y.; Myer, R.R.; Latorella, K.A.; Comstock, J.R., Jr. *The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide (NASA/TM-2011-217164)*; Technical Report; National Aeronautics and Space Administration, Langley Research Center: Hampton, VA, USA, 2011.
55. Delorme, A.; Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef] [PubMed]
56. MathWorks Inc. *MATLAB*, Version: 9.14.0.2206163 (R2023a); The MathWorks Inc.: Natick, MA, USA, 2023.
57. Winkler, I.; Debener, S.; Müller, K.R.; Tangermann, M. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 4101–4105. [CrossRef]
58. Delorme, A.; Sejnowski, T.; Makeig, S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage* **2007**, *34*, 1443–1449. [CrossRef]
59. Ablin, P.; Cardoso, J.F.; Gramfort, A. Faster Independent Component Analysis by Preconditioning With Hessian Approximations. *IEEE Trans. Signal Process.* **2018**, *66*, 4040–4049. [CrossRef]
60. Pion-Tonachini, L.; Kreutz-Delgado, K.; Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* **2019**, *198*, 181–197. [CrossRef]
61. Delorme, A.; Makeig, S. This is no "ICA bug": Response to the article, "ICA's bug: how ghost ICs emerge from effective rank deficiency caused by EEG electrode interpolation and incorrect re-referencing". *Front. Neuroimaging* **2023**, *2*, 1331404. [CrossRef]
62. Ouyang, D.; Yuan, Y.; Li, G.; Guo, Z. The Effect of Time Window Length on EEG-Based Emotion Recognition. *Sensors* **2022**, *22*, 4939. [CrossRef]
63. Zanetti, R.; Arza, A.; Aminifar, A.; Atienza, D. Real-Time EEG-Based Cognitive Workload Monitoring on Wearable Devices. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 265–277. [CrossRef] [PubMed]
64. Elinas, P.; Bonilla, E.V.; Tiao, L. Variational Inference for Graph Convolutional Networks in the Absence of Graph Data and Adversarial Settings. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 18648–18660.
65. Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W.L.; Lenssen, J.E.; Rattan, G.; Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 4602–4609. [CrossRef]
66. Fey, M.; Lenssen, J.E. Fast graph representation learning with PyTorch Geometric. In Proceedings of the ICLR Workshop on Representation Learning on Graphs and Manifolds, New Orleans, LA, USA, 6–9 May 2019.

67. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, 15–18 August 2019; KDD '19; pp. 2623–2631. [CrossRef]
68. Biewald, L. Experiment Tracking with Weights and Biases. 2020. Available online: https://wandb.com (accessed on 30 January 2023).
69. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
70. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]
71. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef] [PubMed]
72. Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M.Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; et al. Pot: Python optimal transport. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
73. de Mathelin, A.; Deheeger, F.; Richard, G.; Mougeot, M.; Vayatis, N. ADAPT: Awesome Domain Adaptation Python Toolbox. *arXiv* **2021**, arXiv:2107.03049.
74. Wang, T.; Ke, Y.; Huang, Y.; He, F.; Zhong, W.; Liu, S.; Ming, D. Using Semi-supervised Domain Adaptation to Enhance EEG-Based Cross-Task Mental Workload Classification Performance. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 7032–7039. [CrossRef]
75. Wang, S.; Liu, Y.; Hu, D. An Early Warning Approach for Pilots' Cognitive Tipping Points Based Multi-Modal Signals. *IEEE Trans. Autom. Sci. Eng.* **2024**, *22*, 4670–4681. [CrossRef]
76. Donoghue, T.; Haller, M.; Peterson, E.J.; Varma, P.; Sebastian, P.; Gao, R.; Noto, T.; Lara, A.H.; Wallis, J.D.; Knight, R.T.; et al. Parameterizing neural power spectra into periodic and aperiodic components. *Nat. Neurosci.* **2020**, *23*, 1655–1665. [CrossRef]
77. Ke, Y.; Wang, T.; He, F.; Liu, S.; Ming, D. Enhancing EEG-based cross-day mental workload classification using periodic component of power spectrum. *J. Neural Eng.* **2023**, *20*, 066028. [CrossRef]
78. Auger, E.; Berry-Kravis, E.M.; Ethridge, L.E. Independent evaluation of the harvard automated processing pipeline for Electroencephalography 1.0 using multi-site EEG data from children with Fragile X Syndrome. *J. Neurosci. Methods* **2022**, *371*, 109501. [CrossRef]
79. Papatzikis, E. How to Process Infant Electroencephalography (Eeg) Data: A Critical Review of 10 Preprocessing Pipelines on Challenges, Strengths, and Future Directions. *Preprints* **2025**. [CrossRef]
80. Kongwudhikunakorn, S.; Ponwitayarat, W.; Kiatthaveephong, S.; Polpakdee, W.; Yagi, T.; Senanarong, V.; Ittichaiwong, P.; Wilaiprasitporn, T. EEGMeNet: End-to-End Multitask Neural Network for Brain-Based Mental Workload Classification. *IEEE Internet Things J.* **2025**, *12*, 42573–42589. [CrossRef]
81. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
82. Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*, 52–65. [CrossRef]
83. Roy, R.; Hinss, M.; Darmet, L.; Ladouce, S.; Jahanpour, E.; Somon, B.; Xu, X.; Drougard, N.; Dehais, F.; Lotte, F. Retrospective on the First Passive Brain-Computer Interface Competition on Cross-Session Workload Estimation. *Front. Neuroergon.* **2022**, *3*, 838342. [CrossRef] [PubMed]
84. Müller-Putz, G.; Scherer, R.; Brunner, C.; Leeb, R.; Pfurtscheller, G. Better than random: A closer look on BCI results. *Int. J. Bioelectromagn.* **2008**, *10*, 52–55.
85. Bertrand, J.; Puel, M. The optimal mass transport problem for relativistic costs. *Calc. Var. Partial. Differ. Equ.* **2013**, *46*, 353–374. [CrossRef]
86. Benamou, J.D.; Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **2000**, *84*, 375–393. [CrossRef]
87. Chen, Q.; Liu, Y.; Wang, Z.; Wassell, I.; Chetty, K. Re-Weighted Adversarial Adaptation Network for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7976–7985.
88. Shen, J.; Qu, Y.; Zhang, W.; Yu, Y. Wasserstein Distance Guided Representation Learning for Domain Adaptation. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 11784. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.