Journal of Neural Engineering

PURPOSE-LED PUBLISHING™

PAPER • OPEN ACCESS

The ‘Sandwich’ meta-framework for architecture agnostic deep privacy-preserving transfer learning for non-invasive brainwave decoding

To cite this article: Xiaoxi Wei *et al* 2025 *J. Neural Eng.* **22** 016014

View the article online for updates and enhancements.

You may also like

- Nanoparticle targeting strategies for traumatic brain injury
David E Flores-Prieto and Sarah E Stabenfeldt
- When neuromodulation met control theory
Roberto Guidotti, Alessio Basti, Giulia Pieramico et al.
- Enhancing mechanical properties of cellular core sandwich panels: a review of topological parameters and design improvements
Assil Charkaoui, Noha M Hassan and Zied Bahroun

This content was downloaded from IP address 165.120.83.26 on 11/08/2025 at 10:55

Journal of Neural Engineering

PAPER

The 'Sandwich' meta-framework for architecture agnostic deep privacy-preserving transfer learning for non-invasive brainwave decoding

Xiaoxi Wei[1], Jyotindra Narayan[2] and A Aldo Faisal[1,2,*]

[1] Brain & Behaviour Lab, Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom
[2] Chair in Digital Health, Universität Bayreuth, Kulmbach 95326, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: a.faisal@imperial.ac.uk

Keywords: deep learning, transfer learning, brain-computer-interfaces (BCI), electroencephalography (EEG), privacy-preserving, federated machine learning, heterogenous datasets

OPEN ACCESS

RECEIVED
22 April 2024

REVISED
18 September 2024

ACCEPTED FOR PUBLICATION
2 December 2024

PUBLISHED
23 January 2025

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Objective. Machine learning has enhanced the performance of decoding signals indicating human behaviour. Electroencephalography (EEG) brainwave decoding, as an exemplar indicating neural activity and human thoughts non-invasively, has been helpful in neural activity analysis and aiding paralysed patients via brain-computer interfaces. However, training machine learning algorithms on EEG encounters two primary challenges: variability across data sets and privacy concerns using data from individuals and data centres. Our objective is to address these challenges by integrating transfer learning for data variability and federated learning for data privacy into a unified approach. *Approach*. We introduce the 'Sandwich' as a novel deep privacy-preserving meta-framework combining transfer learning and federated learning. The 'Sandwich' framework comprises three components: federated networks (first layers) that handle data set differences at the input level, a shared network (middle layer) learning common rules and applying transfer learning techniques, and individual classifiers (final layers) for specific brain tasks of each data set. This structure enables the central network (central server) to benefit from multiple data sets, while local branches (local servers) maintain data and label privacy. *Main results*. We evaluated the 'Sandwich' meta-architecture in various configurations using the BEETL motor imagery challenge, a benchmark for heterogeneous EEG data sets. Compared with baseline models like *Shallow ConvNet* and *EEGInception*, our 'Sandwich' implementations showed superior performance. The best-performing model, the Inception SanDwich with deep set alignment (*Inception-SD-Deepset*), exceeded baseline methods by 9%. *Significance*. The 'Sandwich' framework demonstrates advancements in federated deep transfer learning for diverse tasks and data sets. It outperforms conventional deep learning methods, showcasing the potential for effective use of larger, heterogeneous data sets with enhanced privacy. In addition, through its diverse implementations with various backbone architectures and transfer learning approaches, the 'Sandwich' framework shows the potential as a model-agnostic meta-framework for decoding time series data like EEG, suggesting a direction towards large-scale brainwave decoding by combining deep transfer learning with privacy-preserving federated learning.

1. Introduction

Disabilities, paralysis, and conditions such as stroke exert a profound impact on individuals, significantly impairing their ability to engage with their environment. Brain-computer interfaces (BCIs) have emerged as a promising solution aimed at assisting individuals in regaining lost functionalities [1, 2]. BCIs represent interfaces designed to decode the intricate patterns of human brainwave activity and

© 2025 The Author(s). Published by IOP Publishing Ltd

translate them into computer commands. These commands are subsequently employed to oversee real-time systems, e.g. wheelchairs or video games. Advancements have been achieved in the field of invasive BCI methodologies [3]. However, invasive BCI typically necessitates surgical interventions, which are not universally accessible or desirable. Among the various non-invasive BCI techniques, Electroencephalography (EEG) stands out as a safe and relatively non-expansive choice. EEG entails the measurement of cortical field potentials emanating from the scalp's surface. Within this context, the application of deep learning has emerged as a potent approach, obviating the need for manual feature engineering while eliminating inherent biases [4, 5]. Nevertheless, EEG recordings often exhibit challenges of low signal-to-noise ratio and cross-subject or cross-dataset variabilities stemming from differences in scalp sensor placements, sensor impedance, individual brain anatomy, functional organisation and different tasks performed during data collection. Consequently, BCI systems typically mandate an extensive training period, sometimes for months, for new users to attain a personalised and stable decoding mechanism, thereby impeding the seamless integration of BCIs into real-world applications. To ameliorate these challenges, our study employs a combination of deep learning and transfer learning methodologies to enhance EEG-BCI decoding.

Deep learning has been widely used for EEG decoding [1, 2]. Convolutional neural networks (CNNs) have gained widespread adoption in the domain of EEG decoding across various fields, including applications like seizure detection, Parkinson's disease diagnosis, and motor imagery (MI) decoding, as evidenced by prior studies [6–8]. CNNs have risen to prominence and serve as benchmarks in fields like MI decoding in the existing literature [5]. CNNs offer a distinct advantage in their ability to extract relevant features from raw EEG data autonomously. This autonomous feature learning capability is especially valuable as it obviates the need for subjective decisions often made during handcrafted EEG processing, thereby mitigating the risk of biased feature extraction [9, 10].

Nonetheless, the utilisation of deep learning techniques in EEG analysis poses a notable challenge known as the negative transfer problem, as highlighted in previous research [11, 12]. This challenge manifests as a significant drop in the accuracy of CNNs when they are trained on EEG data sourced from multiple subjects or data sets, which is caused by the disparate distributions across data sets. The underlying source of this dissimilarity stems from inter-subject variability, encompassing differences in skull shapes, brain anatomies, and brain functionalities. Consequently, due to this negative transfer effect, it becomes impractical to directly apply EEG data from other individuals, data sets and tasks to compensate for the scarcity of relevant data. Moreover, CNNs are renowned for their requirement for large amounts of data. Consequently, most BCI solutions continue to necessitate the collection of extensive EEG data sets from users across numerous sessions to attain stable and desirable levels of accuracy. These prolonged training periods impede user motivation and stand as one of the principal challenges confronting the field of BCI.

Transfer learning has emerged as a promising solution to effectively merge data from various data sets, even those originating from domains with distinct data distributions. Transfer learning [13] encompasses a wide array of algorithms with the overarching goal of transferring knowledge and representations gleaned from source domains to a target domain. Some strategies have been developed in this context, primarily within fields such as computer vision and natural language processing. One straightforward approach is fine-tuning, frequently employed to transfer model representations from one specific task or data set to another [14]. In fine-tuning, neural networks are initially pre-trained on one domain and subsequently fine-tuned on the target domain, effectively adapting their learned features to the new data distribution. Another strategy involves training a shared network in early layers that simultaneously incorporates all available data sets but distinguishes and adapts the deeper layers to accommodate the unique characteristics of each data set [15]. Moreover, deep domain adaptation [16] represents a technique that maps source and target data distributions into a unified distribution using Neural Networks. With this approach, methods enable data from source domains to contribute to improving the model's performance on the target domain, bridging the gap between different data distributions.

Some studies have tried to apply these concepts to EEG decoding [17, 18]. For example [19], a deep transfer network was introduced to transfer knowledge representations from images to EEG images. Another study enhanced domain adaptation by incorporating an adversarial network into the architecture, leading to improved performance [20]. Additionally, in the realm of MI decoding, pre-training and fine-tuning strategies were leveraged in a study [21] to develop effective models. Moreover, a deep domain adaptation network was utilised to facilitate the transfer of data representations between different subjects in EEG decoding [22], with a focus on separating the deeper layers of the network to accommodate the individual subject differences. However, despite these advancements, two challenges remain in the literature. Firstly, while current domain adaptation networks are effective in

transferring data distribution from one data set to another (one-to-one transfer), successfully addressing the transfer of multiple data sets to a single target (multiple-to-one transfer) has not been fully resolved. Secondly, the field still faces the challenge of transfer learning involving multiple EEG decoding tasks. These challenges constrain the ability to leverage a diverse range of data for training deep-learning models with the highest possible accuracy.

A significant attempt to address these challenges was the international BEETL BCI competition, which took place at NeurIPS 2021 [23]. The competition centred around cross-dataset and cross-task EEG transfer learning and garnered substantial academic attention. Various successful solution algorithms using deep transfer learning were developed to tackle the BEETL challenge, offering insights for cross-dataset and cross-task EEG transfer learning. Some other studies in the realm of EEG transfer learning have explored various distribution alignment approaches to align data sets and domains from different sources. One of the techniques includes Maximum-Mean Discrepancy (MMD), which has been employed for inter-subject transfer learning in EEG decoding [12, 16, 24]. Maximum Density Divergence has also been utilised as a metric for measuring distributional dissimilarity in EEG data transfer scenarios [25, 26]. Moreover, set theory combined with deep learning was introduced into EEG transfer learning [23, 27, 28] to deal with variability across data sets.

However, another challenge arises for cross-dataset transfer learning: the imperative of privacy preservation. As machine learning algorithms increasingly demonstrate their capacity to harness heterogeneous EEG data sets sourced from various data centres and providers for large-scale applications, the safeguarding of EEG data privacy becomes a primary concern. It is crucial to acknowledge that brainwave data contains intricate and sensitive information, encompassing images, text, and even individual identities, all of which machine learning algorithms possess the potential to unveil. Consequently, the practice of sharing EEG data across data centres or international borders encounters substantial regulatory and legal restrictions, exemplified by stringent frameworks such as the EU's GDPR data policy and ongoing data disputes involving nations like China and the United States.

There are some strategies for privacy-preserving in the machine learning literature [29–33]. Federated Learning [29, 34] stands as an approach where models are trained on edge servers, eliminating the need for raw data exchange. While this method effectively safeguards privacy, it maintains the ability for collaborative model training, striking a balance between data security and model improvement. Data Encryption Methods [30] take the path of encrypting either the raw data or model parameters, yet this approach introduces an encoding–decoding procedure that comes at the cost of additional computational resources. Data Transformation Techniques [31], such as the introduction of cancelable noise to local gradients or parameters prior to their transmission to a central server, offer a means of safeguarding individual data privacy, which also introduces extra computational costs. These techniques obscure sensitive information while allowing for secure data sharing. Model Splitting Methods [32] allocate distinct model parameters to various data sets, ensuring the protection of each individual's model privacy. This approach contributes to enhanced data security within collaborative settings. Multi-party Computation [33] involves local model training by individual parties followed by secure aggregation of model updates conducted by a trusted third party. This method protects privacy while enabling the improvement of the collaborative model.

Nonetheless, the suitability of these methods for brainwave decoding has not been universally established. Furthermore, these approaches may introduce additional computational costs, which could impede training efficiency, particularly when dealing with large-scale EEG data sets. Balancing privacy with the practicalities of EEG data analysis remains an ongoing challenge in BCIs. Several studies within the EEG literature have delved into privacy preservation [35–38], leveraging the methods previously discussed. Additionally, there have been endeavours to enhance privacy preservation through improved protocols and user-level system design [39]. Recent investigations [12, 38] explored the deployment of distributed individual feature extractors, allowing for the accommodation of EEG differences of individual data sets while upholding data privacy without incurring the additional costs associated with encryption. However, it is important to note that the studies mentioned thus far could not fully address the challenge of cross-dataset transfer learning when handling different tasks while simultaneously maintaining the privacy of data, parameters and inference. This limitation can restrict the utilisation of large-scale EEG data in real life. Consequently, the development of privacy-preserving methods capable of accommodating diverse tasks within cross-dataset transfer learning remains an ongoing challenge in the EEG research landscape.

In summary, below we highlight three challenges in the literature that pose potential barriers to the advancement of large-scale EEG transfer learning. Firstly, addressing the variability in experiment setups in different data sets, including differences in EEG input shapes, impedance, and sensor positions, is still a challenge. A solution is needed to harmonize these

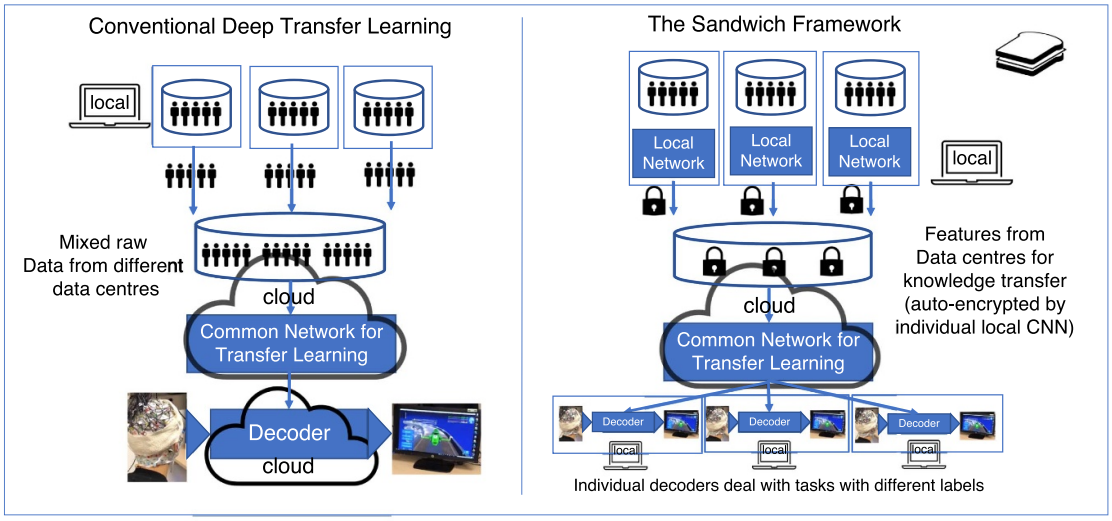


Figure 1. Demonstration of the privacy-preserving deep transfer learning. Conventional transfer learning involves pooling source data sets on a central server, whereas, in the Sandwich framework, data are encrypted as features in the local network before common knowledge training. Moreover, the privacy-preserving transfer learning framework enables predictions to be made on local branches, ensuring data privacy and security from the prediction level.

disparities for effective cross-dataset transfer learning. A second challenge is learning common knowledge across data sets as well as being able to differentiate the information of each individual data set. The key challenge is to find a solution that learns common knowledge across data sets and maintains their individual properties. Developing approaches that enable the incorporation of diverse tasks into a unified model is pivotal for comprehensive EEG transfer learning. The final challenge is to build a transfer learning model that integrates privacy-preserving properties. Ensuring the privacy of EEG data stored in various data centres is important for encouraging secure and seamless data sharing. Robust methods for privacy preservation should be integrated into the transfer learning process to facilitate collaborative research while safeguarding sensitive information. These challenges collectively underscore the complexity of large-scale EEG transfer learning and emphasise the need for innovative solutions that address these issues to unlock the full potential of EEG data in diverse applications.

In our research, we introduce a model-agnostic deep privacy-preserving meta-framework to tackle the aforementioned challenges, termed the 'Sandwich' due to its 3-layer structure, a demonstration of the concept is shown in figure 1. The Sandwich framework is specifically engineered to integrate privacy-preserving features into deep transfer learning for EEG decoding across multiple data sets and tasks. Our study presents various implementations of the meta-architecture to show its versatility, as well as demonstrating improvements in decoding accuracies on the BEETL BCI challenge when compared to baseline methods. Moreover, we introduce the enhanced privacy-preserving capabilities and the ability to utilise heterogeneous data sets, underscoring its potential to advance the field of BCIs towards larger-scale EEG transfer learning.

2. Methods

2.1. Dataset and experimental protocol

The proposed architectures are testing using the MI BCI task of the BEETL competition as an example of cross-dataset and cross-tasking real-world challenges. To provide a visual overview, figure 2 and table A1 depict the data sets employed in the BEETL MI task. The dataset comprises three distinct source data sets, each characterised by unique devices and data collection protocols. Specifically, the Cho2017 data set [40] encompasses 52 subjects engaged in left and right-hand MI tasks. Moving on, the BCICIV2a dataset [41] involves nine subjects participating in MI tasks involving left-hand, right-hand, feet, and tongue actions. Subjects 1, 3, 7, 8, and 9 were used in this study based on superior data quality, narrowly defined as embedding better discriminative features, as sources. As one of the pilot studies in cross-dataset EEG transfer learning, it is crucial to guarantee better quality sources for transfer learning in case of transferring noise from the source. Similarly, the PhysioMI dataset [14, 42], the third source dataset, contains a substantial 109 subjects. Based on the same standard of discriminative feature, subjects 1, 7, 17, 24, 28, 31, 33, 34, 35, 42, 49, 52, 54, 55, 56, 60, 62, 63, 68, 71, 72, 73, 85, 91, 93, 94, and 103 are used as source

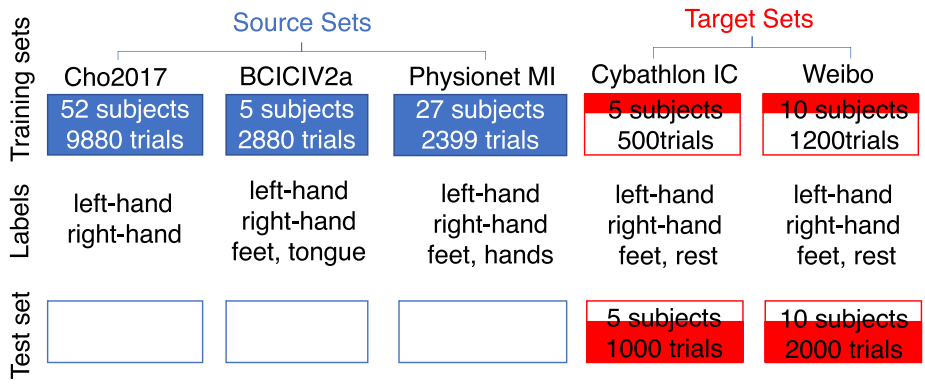


Figure 2. The source data sets (columns 1–3 in blue) and the target set (columns 4 and 5 in red) used in this study. All data from the source sets are used as the source domain and the first 100+ trials from each target subject (100 trials for CybathlonIC or 120 trials for Weibo) are used as the target domain. They constitute the training set. The test set comprises the remaining 200 trials of each target subject. The top row of rectangles identifies the size of the training sets. The middle row (labels) identifies the differing MI labels used in each data set. The bottom row identifies the test set.

sets. Data quality per subject with better discriminative features was determined from classification performance (independent of this study), as they were reported by previous studies or our replications of those studies [12, 42, 43]. In the target training set, the CybathlonIC dataset has 100 trials for each subject. The Weibo dataset has 120 trials per subject during training. 20 trials each were used as the validation set for model selection. Comprehensive details regarding the datasets can be accessed in the 2021 NeurIPS BEETL Competition Study [23].

The two baseline models and implementations of the Sandwich framework were evaluated using the BEETL BCI Challenge containing 5 subjects. This evaluation intends to align with the competition and facilitate comparison with previous and future studies. We also performed a scalability test with full CybathlonIC (test set A) and Weibo (test set B) datasets testing containing 15 subjects to show the stability and scalability of our proposed framework. In the test set, the subjects from two different data sets were included in the MI challenge, each contributing 200 trials with balanced labels. The CybathlonIC data set [12, 23] was collected through an online closed-loop format with a virtual car game, labelled as Rest (0), Left-hand (1), Right-hand (2), Feet (3). The Weibo2014 data set [44] contains trials as Left-hand (0), Right-hand (1), Feet (2), and Rest(3). It is important to note that, during training and prediction of models, it remains a 4-class classification problem. During scoring, as in the BEETL competition standard, we evaluate the final accuracies based on the weighted accuracy of three key classes by combining the last two labels as 'others'. In the BEETL MI challenge, to increase the variability of the test set, 'other' in set A represents labels 2(RH) and 3(Feet), whereas 2(Feet) and 3(Rest) in set B. This design of challenge increases the variety of the two data sets and requires the algorithm to be generalised well to different task scenarios. In this study, considering the heterogeneity of the target sets, we train set A and set B individually as two models during training and testing.

To ensure the alignment and effective utilisation of all data sets, three second windows were extracted for all trials, given that Cho2017 imposed a maximum trial length of three seconds. It's worth noting that this may not fully exploit the potential of the four second target test data, as observed in other solutions excluding Cho2017 within the BEETL competition. Another comparison experiment to the winning method in the competition will be conducted later with a four second window. From the four data sets, we selected 17 common EEG channels (Fz, FC1, FC2, C5, C3, C1, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2) to first evaluate the transferability of our models. Later, an additional experiment using different channel sizes as input will be conducted to show the flexibility of the model for variation of inputs. To maintain consistency, all trials were down-sampled to 200 Hz. Following a standard preprocessing as the Shallow ConvNet [5], a baseline network in MI to be introduced in the next section, a 5th-order bandpass filter was applied between 4 Hz and 32 Hz, where MI usually occurs. Normalisation across channels was applied, followed by temporal normalisation across time steps to a zero mean and standard deviation of one.

Given the disparate sample sizes in the three source sets (BCIC 2880 trials, Cho2017 9880 trials, PhysioMI 2399 trials), a potential imbalance in training individual feature extractors was anticipated. Since this study represents the initial attempts to test the Sandwich meta-framework systematically, efforts were made to minimise the impact of training imbalances during experimental design. While future

studies may explore imbalanced sets to leverage full data sets, for the current investigation, trial numbers across all four data sets were balanced to 2880 trials, aligning with the size of BCIC sources. The PhysioMI and target data sets were augmented with randomisation until reaching 2880 trials to address the varying sample sizes. 2880 trials of Cho2017 were randomly chosen. These above processing procedures try to use as many data sets as possible but reduce the information of channels, trials or samples, whereas the winning method in the competition tried to maximise the channel sizes and window length to get higher performance but did not use all source sets. A comparison between the winning method and the Sandwich models under the same setup will be shown in the result section.

2.2. The 'Sandwich' framework design

In this section, we will introduce our Sandwich meta-architecture, followed by the presentation of two exemplary implementation architectures, *Shallow-ConvNet-Sandwich* and *EEGInception-Sandwich*, on the BEETL BCI challenge. These examples will demonstrate how to seamlessly integrate existing EEG decoding models into the federated transfer learning meta-framework. Subsequently, we will delve into the demonstration of two transfer learning approaches employed within the Sandwich framework, the distributional alignment and deep set approach.

The general framework of the Sandwich is illustrated in figures 3 and 4. It begins with individual split branches functioning as feature extractors for each data set. These individual feature extractors serve two primary purposes. Firstly, they handle discrepancies in the original data setup, including variations in channel impedance, location and number of channels. Across subjects or datasets, even when the data sets share the same channel names, locations could be shifted, and impedance, as well as channel sizes, vary. Secondly, these individual feature extractors can be locally stored and trained within each data centre, effectively serving as auto-encryption layers alongside feature extraction. This process can be likened to an auto-encryption mechanism, where the individual feature extractors are exclusively utilised within their local server during training, ensuring that the 'key' to encryption remains solely in the possession of the local server. Only the features 'encrypted' are shared with the central server for common knowledge learning and transfer learning. In this way, the framework ensures raw-data level and parameter-level privacy to be secured.

Following the individual feature extractors are the common layers. The functionality of the common layers is to learn common knowledge across different data sets. These common layers receive features from each branch and perform a common feature extraction process, facilitating shared rules for feature extraction across different data sets. After common knowledge learning, additional deep transfer learning techniques can be applied within the Sandwich framework, such as distributional alignment [24] or the deep set approach [23]. These transfer learning layers could be integrated into or after the common knowledge transfer learning layers. A detailed description of network parameters of implementation and computations regarding these additional transfer learning layers will be presented later in this section.

Two strategies for the final classification are considered to address variations in labels. The first strategy involves using a unified classifier that encompasses all tasks from all data sets (figure 4). This approach potentially benefits the classifier by exposing it to a broader range of training examples, particularly if these examples are drawn from similar distributions after the alignment. The second strategy entails splitting the classifiers for different data sets, which may perform better when tasks exhibit very dissimilar distributions (figure 3). Additionally, the second strategy introduces an additional layer of privacy preservation, as these classifiers can be trained and make predictions on local servers, which protects model inference-level privacy.

To illustrate the implementation of the Sandwich meta-framework, we incorporate two example models into the architecture. These models serve as tangible demonstrations of how the Sandwich can be effectively utilised. For establishing a comparison baseline, we employ the Shallow ConvNet [5] and EEG InceptionNet [45], two benchmark models for MI BCI decoding, as baseline models.

We first introduce the implementation of the Sandwich framework using the EEG *InceptionNet*. We employ a feature extractor design similar to that utilised in the BEETL winning method, in which the detailed network design can be found [23, 28]. The main concept of the inception network layer involves channelling the data through multiple parallel layers with distinct kernel sizes. Subsequently, the filtered features from these parallel layers are concatenated before being forwarded to the next layer. Specifically, the individual inception feature extractor used in this study comprises a temporal inception block composed of three inceptions with kernel sizes of (1, 21) but varying dilation rates of (1, 4), (1, 2), and (1, 1). Then, a spatial convolutional layer with a kernel size of (channel size, 1), and a filter size of 48 is applied. Additional components afterwards include a batch normalisation layer, an average pooling layer, and a dropout layer with a dropout rate of 0.25. The common knowledge learning layer consists of another inception block with a kernel size of (1,5) and varying dilation rates of (1, 8), (1, 4), and (1, 2). The

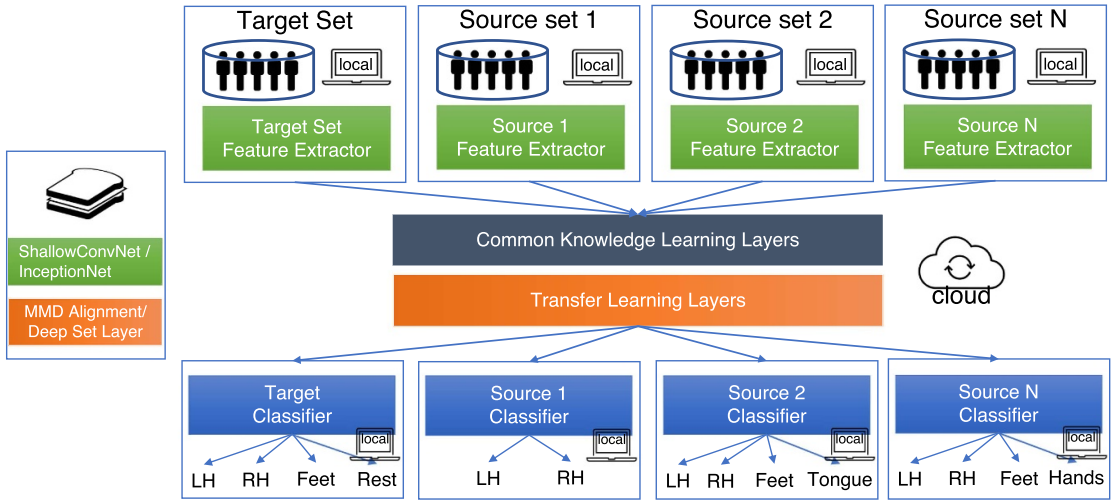


Figure 3. The network architecture of the Sandwich with multiple classifiers.

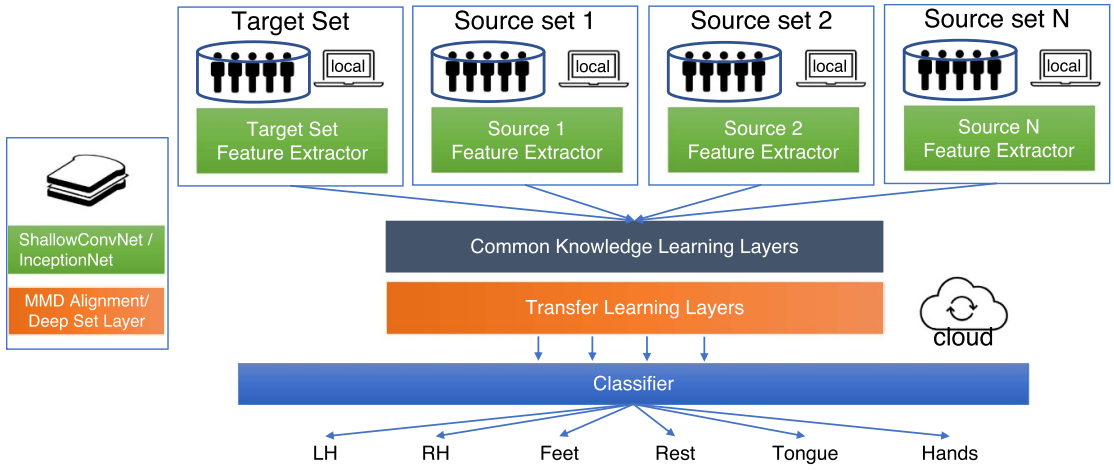


Figure 4. The network architecture of the Sandwich with a unified classifier.

transfer learning layers consist of two CNN blocks as learning parameters. The first CNN block has two convolutional layers of kernel size (1, 9) and (1,1) with a batch normalisation layer and a dropout of 0.25. The other CNN block has smaller kernel sizes of (1, 5) and (1,1) with a batch normalisation layer and a dropout of 0.25. In this study, two transfer learning approaches were tested in the transfer learning layers. If MMD alignment is applied, an MMD layer is conducted once after the second CNN block. If the deep set method is used, the deep set layers are applied both before and after the CNN blocks. We will introduce how MMD and Deep Set work in later paragraphs. Finally, a standard unified 6-class classification layer or multiple individual classifiers are used to perform classification.

Furthermore, we integrate the *Shallow ConvNet* as a second example of the feature extractor for the Sandwich implementation. This additional assessment demonstrates the versatility of Sandwich in accommodating various EEG decoding models. Similar to the Shallow ConvNet study [5], the Shallow ConvNet feature extractor comprises several key components: a temporal layer with a kernel size of (25,1) and 40 filters, a spatial layer with a kernel size of (1, channel size) and 40 filters, a batch normalization layer, a square non-linear layer, a pooling layer with a kernel size of (75,1) and a stride size of (15,1), a logarithmic non-linear layer for pooling, and finally, a dropout layer with a rate of 0.5. The above layers assume data format as (time samples, and channels). These architectural parameters remain consistent and can be found in the study of Shallow ConvNet [5]. The temporal layer is responsible for extracting time-relevant information, while the spatial layer performs spatial convolutions across EEG channels, collectively contributing to the feature extraction process. Before the common transfer layers, a feature reduction layer was applied to reduce the feature space to 50. The common knowledge learning layer consists of three

convolutional layers with a filter size of 50 and kernel size of (1,1). If transfer learning layers are applied, an MMD layer is conducted after the three convolutional layers. The deep set layers are added both before and after the convolutional layers if the deep set method is used. Finally, the network has a standard unified six class classification layer or multiple individual classifiers to perform prediction.

As for hyperparameters of the frameworks, a batch size of 20 and a weight decay factor of 0.0005 were used for both *shallowConv*-based Sandwich and *EEGInception*-based Sandwich models. A learning rate of 0.0005 was used aligning with their original studies. During the training of Sandwich models, four batches of size 20 from branches (80 in total with four branches) are delivered to the central layers simultaneously and features are distributed back to each separate local branch after common feature extraction.

Following the introduction of the network structures, we present the two transfer learning techniques in the subsequent paragraphs: domain adaptation based on MMD [24] for aligning distributions and deep set alignment [23] to enhance set invariance. These two transfer learning techniques are applied to the networks to boost transferability across different data sets.

In this study, we have devised a Sandwich-MMD architecture, incorporating Maximum Mean Discrepancy (MMD) alignment into the Sandwich framework. Long *et al* [46] have successfully employed MMD as a similarity constraint to reduce the distributional divergence between just two image domains in Deep Domain Adaptation Networks (DAN). In our study, we extend the DAN framework to our proposed network architectures by MMD computation among multiple EEG data sets. This adaptation enhances our network's ability to align distributions across diverse data sets, contributing to improved transfer learning capabilities. The detailed computation of MMD is elucidated in equation (1). X_t and X_{si} represent the target data and the i_{th} source data respectively. The selection of the kernel function Φ has been a prominent subject of investigation in the existing literature, and in line with prior research [46], we employ the Radial Basis Function kernel with the mean L2 distance of the data serving as the variance. MMD is then included as a component of the loss function. In the loss function, denoted by equation (2), L_c signifies the classification loss, N represents the count of source data sets, and λ serves as the weight parameter that governs the trade-off between the MMD loss and the classification loss. Within the Sandwich-MMD framework, we compute and accumulate MMD_i^2 between the output features of the target branch network and each of the source branch networks *i*. Our study concentrates on a common task across all data sets, specifically left-hand and right-hand MI. To this end, our model computes the MMD distance with respect to left-hand and right-hand samples between the source and target domains.

$$\mathrm{MMD}_i^2\left(X_{si}, X_t\right) = \Big|\Big|\frac{1}{X_{si}}\sum_{x_{si}\in X_{si}}\Phi\left(x_{si}\right) - \frac{1}{X_t}\sum_{x_t\in X_t}\Phi\left(x_t\right)\Big|\Big|. \quad (1)$$

To be specific about the implementation of the MMD layer mentioned in the architecture. It comprises a convolutional feature reduction layer, reducing dimensionality from the previous layer to 50, and another convolutional layer with 50 filters and a kernel size of (1,1). These layers act as learnable parameters, facilitating the projection of features into a comparable distribution. Subsequently, MMD distances, serving as distance losses, are computed based on equation (2) to align the distribution of the target data set with each source data set.

$$L = L_c + \lambda\sum_{i=1}^{N}\left(\mathrm{MMD}_i^2\left(X_{si}^L, X_t^L\right) + \mathrm{MMD}_i^2\left(X_{si}^R, X_t^R\right)\right). \quad (2)$$

The deep set alignment layer is implemented based on the Deep Set [27]. Deep Set is a technique that employs deep learning to process sets of variable-size inputs. It addresses the challenge of working with unordered and varying sets using permutation-invariant architectures to align these source sets. Deep Set has applications in various domains and enables tasks such as set classification, matching, generation, and similarity estimation. The deep set idea was proposed to be used in EEG decoding, and we employed the identical configuration for the deep set layer as outlined in [23, 28, 47]. The design of the deep set layers is grounded in theorem 1, in which each subject from different data sets is regarded as a 'set'. In deep learning, convolutional layers can be interpreted as the transformation function Φ, while the non-linear activation functions of these layers are denoted as nonlinearity ρ.

Theorem 1 (Set Theorem used in Deep Set [27]). *Function f(X) operating on a set X having elements from a countable Universe, is a valid set function, i.e. invariant to the permutation of instances in X, iff it can be decomposed in the form* $\rho(\sum_{x\in X}\Phi(x))$, *for suitable transformations* Φ *and* ρ.

In alignment with the findings from previous studies, the deep set layers serve as a mechanism for alignment, rendering each set invariant to trial orders and enhancing the robustness of predictions within each set. A comprehensive description of the deep

set layer implementation can be found in [23, 28, 47]. In these studies, EEG trials underwent processing through convolutional layers. Subsequently, the EEG features of each subject are organised as 'sets' and are then subjected to deep set layers to enhance the robustness of individual features. In our study, as illustrated in figures 3 and 4, we further extend this concept to multi-task scenarios. The common layers facilitate knowledge transfer learning across data sets, and the subsequent deep set layers treat each subject as a set, reinforcing the stability and robustness of these features.

In this study, we employed the same deep set alignment block as the above studies. The deep set block rearranges the three-dimensional training batch from (S*T, N, f) to subject-wise four-dimensional features (S, T, N, f). Here, S represents subjects, T represents subject trials, N represents filter size, and f represents feature size. The mean of each subject is computed, resulting in a feature space of (S, 1, N, f). These subject-specific features are then passed through linear layers of size 8 resulting in a feature vector of (S, 1, 8, f). This process aligns with the concept of the '$\sum$' operation in theorem 1. Subsequently, these subject-specific features are concatenated back to each trial of the corresponding subject, serving as 'global' information for that subject (S, T, N+8,f). The data is then passed through another linear layer to project the feature space back to the original size (S, T, N,f). This process corresponds to the Φ operation in theorem 1, and the activation function of the layers is regarded as the non-linear operation ρ in the theorem.

3. Results

3.1. Results on the BEETL MI Challenge

As depicted in figure 5, we analysed the topography showing the power map of channels for the five subjects within the test set in the BEETL MI challenge to elucidate distinctive characteristics of EEG data during various tasks. To facilitate this visualisation, we leveraged the MNE topography API, as outlined in [48]. Each topography map has been normalised by subtracting the average power across all channels, thereby ensuring a uniform range of values from -0.2 to 0.2 for meaningful comparisons. Across subjects, a consistent pattern emerges, wherein a reduction in power is observed on the contralateral side of the motor cortex, specifically at channels C1, CP1, C2, and CP2, during the execution of left-hand and right-hand MI tasks. When subjects engage in feet MI, we note an increase in power around the motor cortex. This phenomenon could be explained by considering the reference channel, Cz, situated at the centre. It corresponds to the feet MI task. Therefore, a decrease in power around the Cz channel could lead to an increase in power on both sides of the motor cortex, manifesting as the observed pattern during feet MI.

In the following experiments, we compared two implementations of the Sandwich architectures, the ShallowConv-Sandwich (*ShallowConv-SD*) models and the Inception-Sandwich (*Inception-SD*) models, to the baseline *Shallow ConvNet* and *EEGInception* on the above test set. Our approach encompasses two strategies addressing different labels across diverse tasks. The first involves a single classifier accommodating all potential labels within the data sets, while the second employs individual classifiers tailored to the specific labels of each dataset. We also tested two transfer learning approaches, the MMD and Deep Set, in the proposed framework. The decoding accuracy for various combinations of architectures and approaches is extensively presented in the appendix (see A4). Within this section, we draw attention to the models that exhibit superior performance compared to the baseline models, as illustrated in figure 6.

The *EEGInception* and *Shallow ConvNet* exhibit comparable average decoding accuracies of 47% and 48%, establishing a baseline for comparison. The Single-classifier Sandwich models with the MMD alignment (*SD-MMD-1Cls*) performed better than the baseline models, yielding a decoding accuracy of 52% for the Inception-based model (*Inception-SD-MMD-1Cls*), and 50% shallowConvNet-based model (*shallowConvNet-SD-MMD-1Cls*), as shown in figure 6. In terms of the performance of models with individual classifiers with deep-set layers (*SD-Deepset-MultiCls*), Specifically, the Inception-based model (*Inception-SD-Deepset-MultiCls*) exhibited a 9% improvement over the *EEGInception* baseline. A parallel trend was evident in the *Shallow ConvNet* example, with *ShallowConv-SD-Deepset-MultiCls* surpassing the *shallowConv* net baseline by 6%.

Additional comparative models, including federated networks without transfer learning layers, networks combining multiple classifiers with MMD, and combining a single classifier with deep set layers, are detailed in the supplementary figure A4. Compared to the baseline models, federated networks without MMD or deep set layers perform slightly better than the baseline, yielding decoding accuracy of 1% and 3% higher for the *EEGinception* and *Shallow ConvNet*. However, we can observe that in both cases, combining the deep set with a single classifier and combining MMD with multiple classifiers both result in lower performance. A clear example is that the *Inception-SD-Deepset-1Cls* is 10% lower than *Inception-SD-Deepset-MultiCls*, and the *Inception-SD-MMD-MultiCls* is 7% lower than the performance of *Inception-SD-MMD-1Cls*. We will provide insights into the rationale behind the superior performance of combining MMD with a single

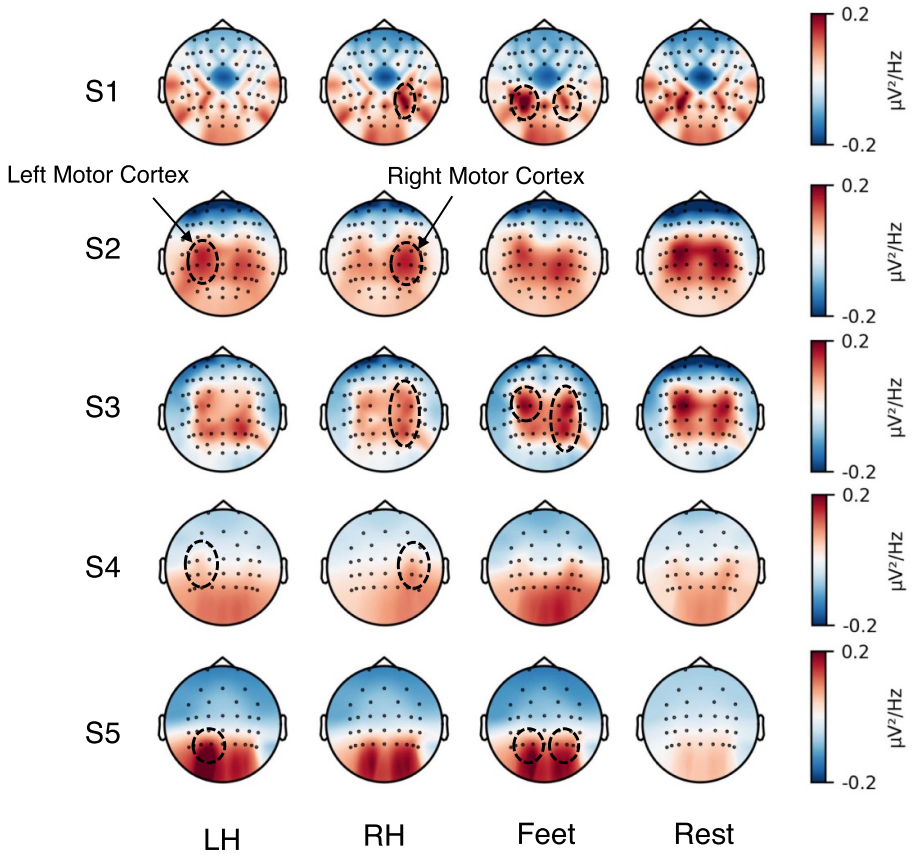


Figure 5. Raw Data topography map of individual subjects in the test set.

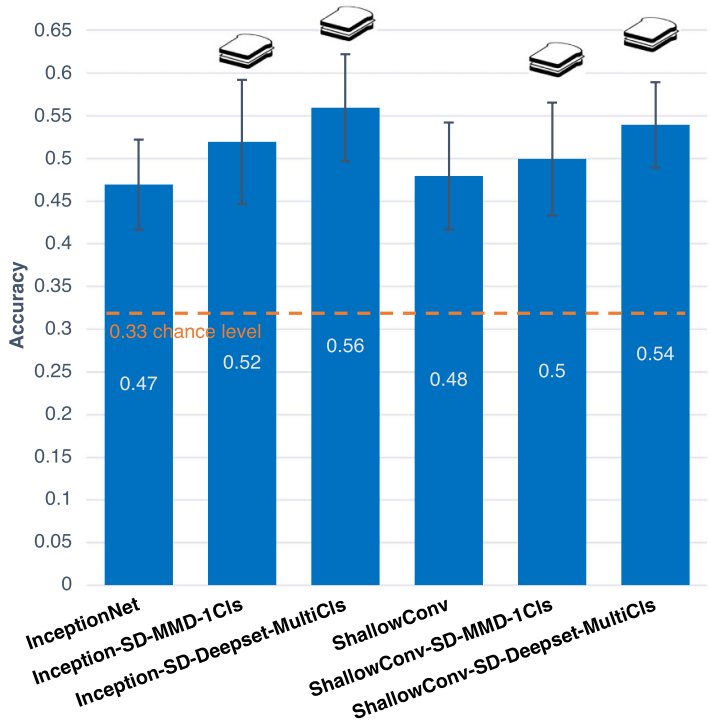


Figure 6. Averaged decoding accuracy of different models.

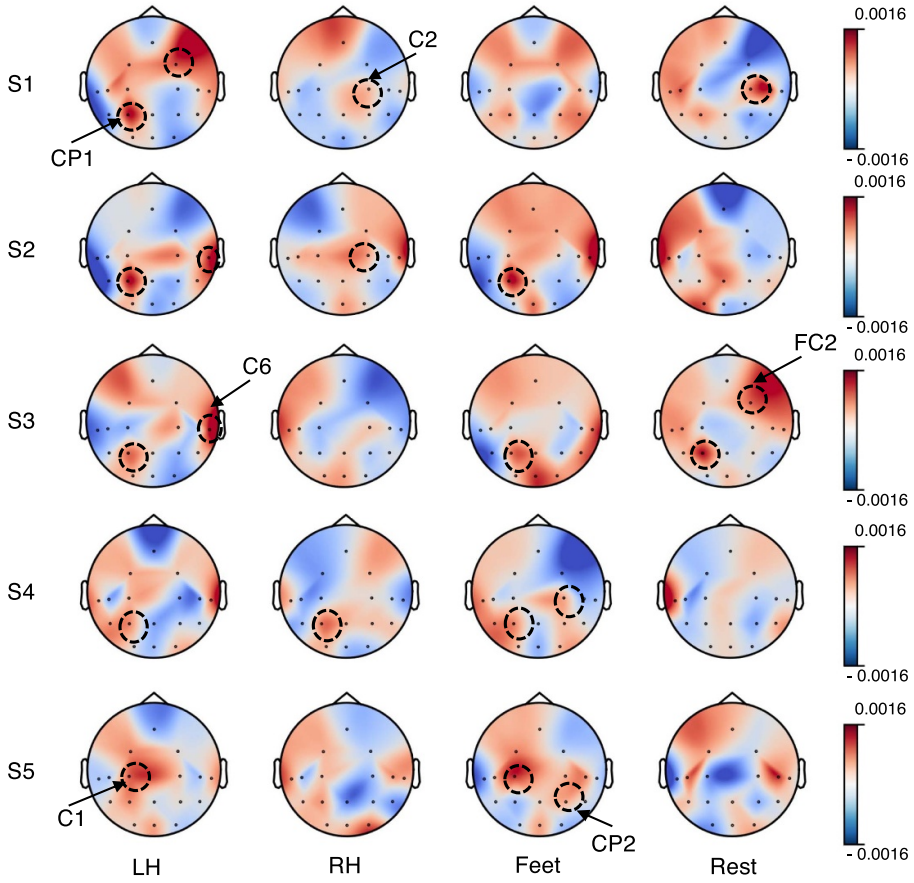


Figure 7. Example Network Focus Visualisation on Inception-SD-Deepset-MultiCls.

classification head and combining deep set with multiple classification heads in the Discussion section and explain why the transfer learning could not work by reserving the combination.

To provide interpretability for the model, we present a network-focused visualisation of tasks across five subjects in figure 7, utilising the *Inception-Deepset-MultiCls* architecture with the highest decoding accuracy as an illustrative example. This visualisation is generated through an input-output perturbation method, as outlined in [5]. The colour scheme employed reflects the relevance of data to the final prediction for each label, where red indicates a greater focus on specific regions. As depicted in the figure, a discernible trend emerges, mirroring the topographical map of raw data around the motor cortex (figure 5). Regions with relatively higher relevance, specifically C1, C2, CP1, and CP2, are circled around the motor cortex for the left hand (LH), right hand (RH), and feet across the five subjects. Additionally, the visualisation shows the algorithm sometimes tries to capture noisy channels during the learning process as features. Noteworthy instances include FP2 in the near frontal region for LH of subject 1, and the Rest state for subject 3, as well as C6 near the right ear for LH of subject 3. The discussion section will delve into the potential implications of motor-related features and noise-related features, offering a comprehensive exploration of the model's interpretability.

To elucidate the characteristics of the two proposed Sandwich architectures and provide empirical evidence of the impact of common transfer learning layers, we present the t-SNE [49] plots comparing EEG features before and after the computation through the common layers. Specifically, we focus on the *SD-MMD-1Cls* and *SD-Deepset-MultiCls* models which showed superior performance than baselines, with *ShallowConv-SD-MMD-1Cls* and *ShallowConv-SD-Deepset-MultiCls* serving as illustrative examples. As articulated in equation (2), the MMD-based model aims to align different distributions of common labels (LH and RH) across data sets. Therefore, we visualise the learned features for LH and RH across four data sets, where shapes represent labels, and colors represent data sets. These t-SNE plots offer insights into how the common transfer learning layers influence the distributional alignment of features for the specified labels, providing a tangible representation of the impact of MMD-based alignment in the context of the Sandwich framework.

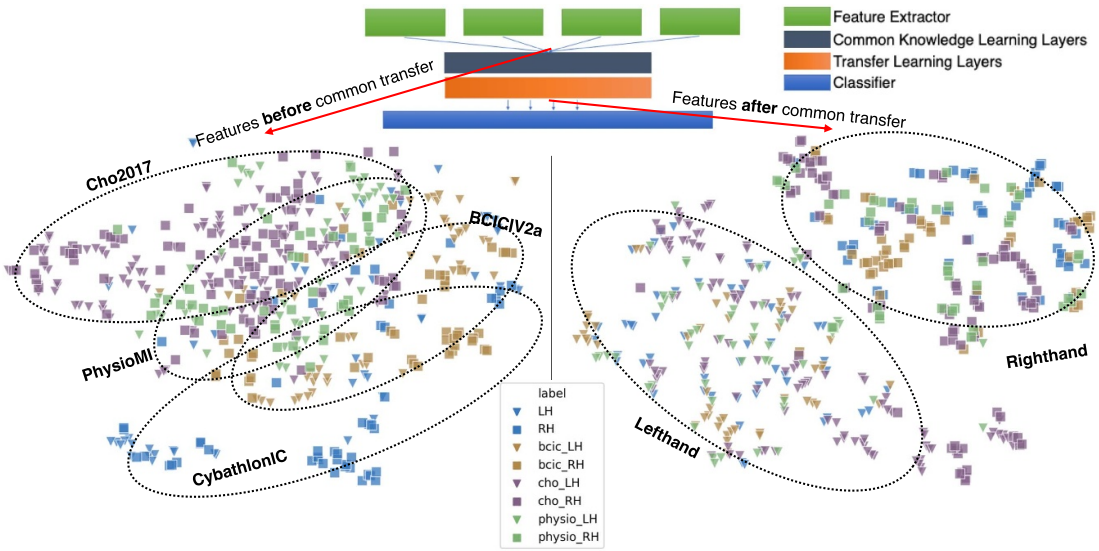


Figure 8. t-SNE feature plot across data sets before (left) and after (right) the common transfer learning layers on *ShallowConv-SD-MMD-1Cls*.

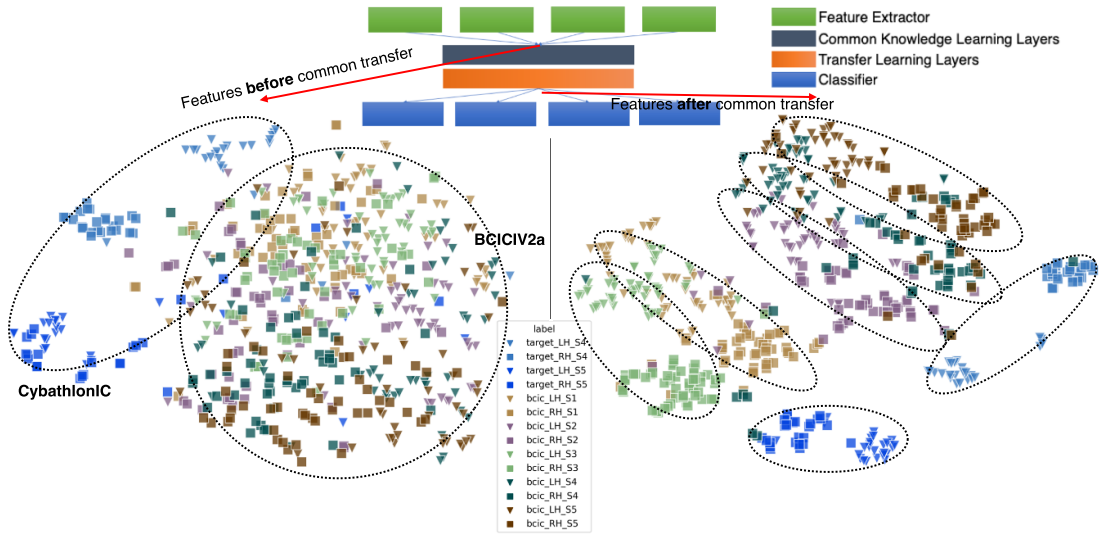


Figure 9. t-SNE feature plot across data sets before (left) and after (right) the common transfer learning layers on *ShallowConv-SD-Deepset-MultiCls*.

As depicted in the t-SNE plot for *ShallowConv-SD-MMD-1Cls* (figure 8), a distinct pattern emerges. Prior to the integration of common transfer learning layers, the four datasets exhibit separate clusters, indicating diverse distributions. However, after the application of common knowledge learning layers and MMD layers, two discernible effects come to light. One is that, post-alignment, the features from different data sets converge into relatively the same region, suggesting a harmonisation of distributions. This aligns with the goal of the MMD-based model to reduce the distributional gap among data sets. The other is the clustering of Left-Hand and Right-Hand labels. This outcome indicates that, despite the distributional alignment, the model has successfully retained the discriminative information between LH and RH labels, which is crucial for accurate classification.

As outlined in the methods and Theorem 1, the Deepset-based model aims to enhance robustness and invariance within a set, considering subjects as sets in our context. To illustrate this effect, we present t-SNE plots focusing on the LH and RH labels for subjects within the CybathlonIC and BCICIV2a data sets. The shapes in the plot represent labels, while colours denote individual subjects. To manage visual complexity, we showcase the effect on seven subjects. As evident in figure 9, the t-SNE plots showcase the

impact of the deepset-based model on the distribution of labels and subjects. Before the integration of common transfer layers, different labels and subjects did not exhibit clear clustering and classification patterns. However, post the application of common transfer layers, distinct clusters for different subjects become apparent. Moreover, different labels fall on different sides of these clusters, contributing to a clearer separation for classification purposes. The t-SNE plots for the inception-based models, presented in appendix figures A1 and A2, exhibit a similar trend as above, further validating the effectiveness of the proposed approach across different model architectures.

3.2. Comparison experiment results to the BEETL benchmark model

In this section, we compare our framework with the BEETL top-winning methods under the same setup and demonstrate comparable performance. The experiments in the previous section covered all three source data sets and all MI tasks to evaluate the 'Sandwich' framework's ability to utilise heterogeneous data sets. The common 17 channels and common three second window trial length across datasets had to be used to utilise all sources. On the contrary, the winning benchmark in the BEETL Competition [23] only used the PhysioNet data set as the source set to utilise longer EEG trials (four second window) and more channels (30 common channels across two datasets instead of 17 common channels across all datasets) to maintain richer dimensional information in the test set to benefit the prediction, as well as using same tasks across target and sources (LH, RH, feet, rest). This approach results in a better baseline while decreasing the diversity and complexity of setups and scale of datasets that could be utilised. Additional approaches like label smoothing loss function were applied to increase the performance [28]. Although the benchmark algorithm reduced the complexity of the setup which limits the use of broader heterogeneous data sets, here we applied the Sandwich framework to the same setup as the benchmark in the competition as a comparison to the benchmark transfer learning algorithm in the literature. We reproduced the Benchmark *EEGInception Deepset* model and got the same accuracy score as the benchmark in the competition (76%), which details of the experiment could be found in their original study [23, 28]. This baseline was given by an additional average voting procedure during prediction, which trains ten models with different data splits and performs average voting in the end to get the best possible accuracy.

Figure 10 shows the decoding accuracies of the benchmark model and the benchmark with the sandwich framework. In this comparison experiment, the main purpose is not to get the highest possible accuracy, here we show the comparison of the results of the models without the additional average voting approach. As shown in the figure, all transfer learning algorithms (benchmark and the sandwich networks) perform better than pure *EEGInception* network baseline without transfer learning (64%). The Sandwich network achieved comparable accuracy to the benchmark algorithm in the competition (67%) while maintaining privacy-preserving properties through the protection of raw data and parameters, securely stored in the local branches. An additional experiment was conducted by varying channel sizes (*DiffCh*), where the target set utilised 30 channels (consistent with the benchmark), and the source set employed all possible 64 channels, resulting in an accuracy of 68%. This highlights the Sandwich network's flexibility in accommodating different input sizes within data sets, potentially expanding the utilisation of heterogeneous data sets, with additional privacy-preserving properties.

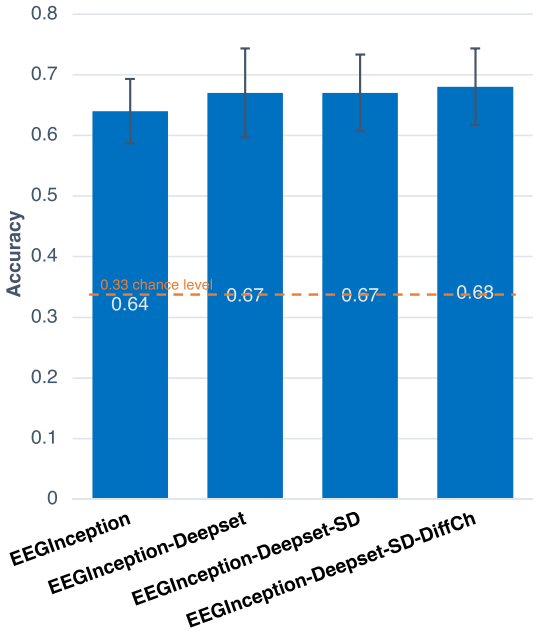


Figure 10. Accuracies of models applying the Sandwich framework to the Benchmark Transfer Learning Model.

3.3. Framework scalability test on full CybathlonIC and Weibo datasets

In this section, to further verify the scalability and stability of the proposed framework, we ran the framework on the full CybathlonIC (5 subjects) and Weibo (10 subjects) datasets, which tripled the size of the test set, using the same standard as in section 3.1. Additionally, we ran the pretraining-finetuning transfer learning strategy [19], as a comparison on the expanded datasets, to compare the sandwich framework's performance with the most commonly used

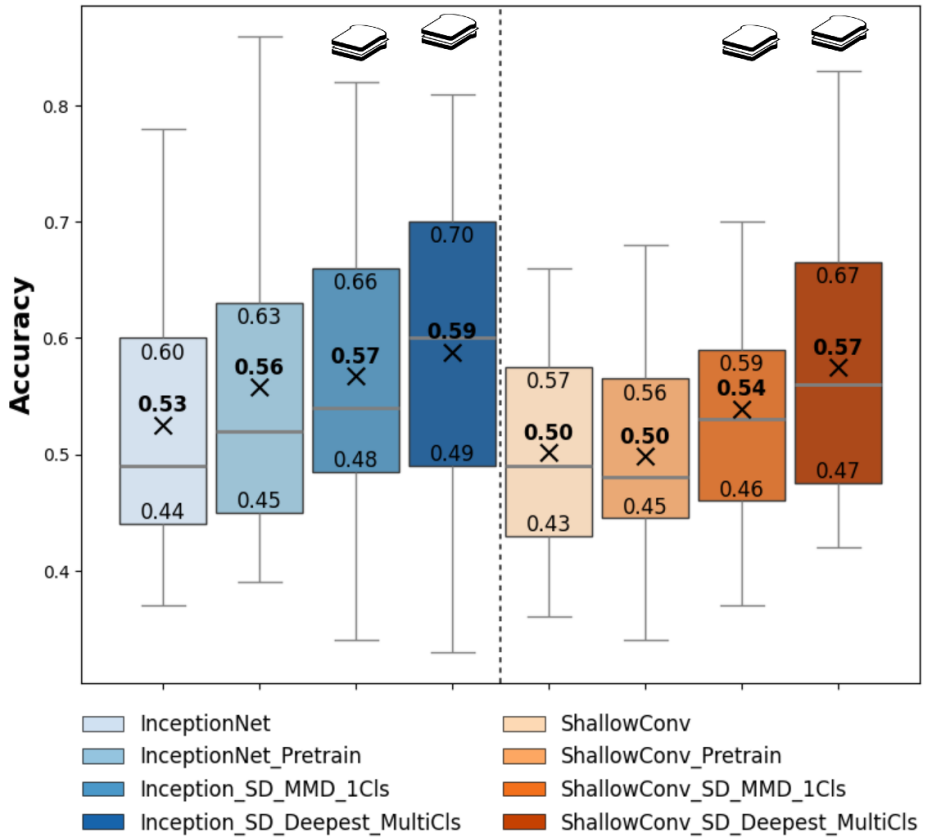


Figure 11. Average decoding accuracy of different models on 15 subjects ($n = 15$) in the full CybathlonIC and Weibo datasets (including the same 5 subjects in the BEETL test and using the same BEETL standard). The left group represents InceptionNet-based models, and the right group represents ShallowConvNet-based models. The cross mark and bolded number represent the average decoding accuracy.

deep transfer learning strategy in EEG literature. In the pretraining experiment, the model was first pretrained using the source datasets (with different classifiers handling the various labels) and then finetuned using the same network structure and training strategy as the baseline. The scalability and stability test was conducted using additional subjects rather than cross-validation due to the adaptive learning nature [50, 51] of new BCI users. In practice, BCI users gradually learn and adapt to the system in a mutual learning process during the session. Therefore, swapping the order of the session for cross-validation does not accurately reflect the model's real-world performance. Instead, using the initial collection of trials as the target domain and testing on more subjects provides a more realistic representation of how the model will be used in practice.

With a larger dataset containing more subject samples, we present a box and whiskers plot reflecting the overall decoding accuracy distribution of different subjects on our frameworks, as well as a standard paired t-test [52] to evaluate the statistical significance of the improvement between our models and the baselines. As shown in figure 11, the trend is consistent with the results in previous sections. Our sandwich frameworks exhibit higher mean accuracy, as well as higher median and quartile boundaries compared to the baselines. Specifically, our *ShallowConv-SD-MMD-1Cls* and *ShallowConv-SD-Deepset-MultiCls* performed significantly better than the *ShallowConvNet* baseline, with improvements of 4% and 7%, and p-values of 0.004 and 0.001 respectively ($p<0.05$ indicating significance). Similarly, compared to the *InceptionNet* baseline, our *Inception-SD-MMD-1Cls* and *Inception-SD-Deepset-MultiCls* performed 4% ($p = 0.013<0.05$) and 6% ($p = 0.001<0.05$) better than the *InceptionNet* baseline. To reflect the training property of the model, we attach the test set confusion matrices of the four sandwich models in the appendix figure A3. The matrices exhibit strong diagonal dominance, indicating that the majority of predictions for each class are correctly identified. The diagonal values are notably balanced. This suggests that the classifier performs consistently well across all classes, without a significant bias towards any particular class.

In addition, we observed that the traditional pretraining-finetuning transfer learning method did not constantly perform better performance than the baseline. Although The *Inception-Pretrain*

model performs slightly better than the *InceptionNet* baseline by 3% ($p = 0.05$), the *ShallowConv-Pretrain* model does not show significantly better accuracy than the baseline *ShallowConv* model with the same mean of 50% ($p = 0.85 > 0.05$). The results show that our proposed models not only significantly perform better than the baselines, but also show better stability than the standard traditional transfer learning method.

4. Discussion

In the previous section, we conducted a performance comparison of the proposed Sandwich framework, highlighting the outcomes of employing various network backbones or transfer learning approaches. As demonstrated earlier, both Sandwich implementations exhibited superior performance compared to baseline methods across the example architectures. This observation underscores the potential of the Sandwich framework to effectively leverage cross-dataset features from diverse tasks, thereby enhancing the overall performance of EEG decoding.

The Sandwich framework with multiple classifiers (*SD-Deepset-MultiCls*) and the Sandwich frame with a unified classifier (*SD-MMD-1Cls*) both outperformed the baseline with feasible transfer learning approaches. This has been detailed in both figure 6 using the BEETL test set on five subjects and figure 11 on 15 subjects using full CybathlonIC and Weibo datasets as the test. Here we highlight the commonalities and the differences between the two approaches. For both architectures, they used individual feature extractors which potentially handle differences in the raw data (e.g. impedance, sensor locations or channel sizes) by using different spatial and temporal kernels in the individual CNNs. Common knowledge learning layers are applied in both frameworks to learn the same set of network parameters (the same rules) to draw features from different data sets, which forces the network to learn the commonality across data sets.

In both frameworks, it is evident that the application of appropriate transfer learning approaches can enhance the performance of models, whereas the requirements by the unified classifier and multiple classifiers are shown to be different. Both unified classifiers and multiple classifiers present potential solutions for handling label variability in data sets. However, the unified classifier necessitates that the features received for decoding adhere to a similar distribution. This requirement is essential for learning common discriminative rules for identical labels across different sets using the same classifier. Achieving this similarity in distribution mandates an alignment approach that projects features of the same labels from different data sets onto a comparable distribution. This alignment objective aligns with the design purpose of methods such as MMD, as evidenced by the t-SNE plot in figure 8.

On the contrary, the Sandwich framework with multiple classifiers operates on a different premise. Unlike the unified classifier, it does not necessitate the labels of different data sets to conform to the same distribution. This is because individual data sets apply distinct decision boundaries for classification with individual classifiers. Therefore, after the common knowledge learning layer, simply aligning data sets and features with the same label may not help the decoding, in fact, it has the potential to diminish discriminative information by projecting features to another latent space. In the supplementary experiment, when MMD alignment was applied to models with multiple classifiers, we observed a decrease in performance compared to models without MMD alignment. For instance, as depicted in supplementary figure A4, the accuracy of *Inception-SD-MMD-MultiCls* (45%) was lower than that without using MMD, such as *Inception-SD-MultiCls* (48%), and notably, 11% lower than when using the Deepset approach, as seen in *Inception-SD-Deepset-MultiCls*. The observed benefit of applying the deep set approach to multiple classifiers may be attributed to the impact of deep set layers on enhancing feature invariance within each subject considered as a 'set'. This heightened feature invariance is reflected in the clustering of sets, as depicted in the t-SNE figure 9. Consequently, this increased feature invariance contributes to making the predictions of each classifier more robust. The deep layers seem to be crucial for extracting and representing invariant features, which enhances generalisation and performance in the context of multiple classifiers.

The observed characteristics of the two frameworks prompt a discussion on the criteria for selecting between them. Briefly, the MMD approach with a single classifier aims to align distributions across samples to enhance classifier performance, yet it may diminish the unique dataset-specific features due to the unified projection. Conversely, the deep set framework employing multiple classifiers preserves subject-specific variability and individual discriminative information, although it does not guarantee perfect alignment across data sets, as evidenced by t-SNE plots. The challenge lies in the variance of the original sample distributions; the more divergent they are, the tougher it is to map features into a similar distribution. Consider sleep stage decoding and MI as illustrative examples. Sleep stages are generally consistent across individuals, characterized by variations in brain energy levels, with most people adhering to this norm. In contrast, MI involves more intricate functionalities of the motor cortex, where individual

differences are more pronounced [23]. Hence, the decision between these approaches should hinge on the distributional disparities across data sets. For data sets with minor differences, an MMD-like method that projects them into a similar distribution could be beneficial. However, for more complex tasks that tap into intricate brain mechanisms, simple alignment might reduce each dataset's discriminative information. In such cases, a deep set framework with multiple classifiers could enhance invariance within sets while leveraging common knowledge layers for performance gains.

To gain insights into how a neural network perceives EEG data, a comparative analysis of the raw topography map in figure 5 and network relevance visualisation in figure 7 across different tasks and subjects can be conducted. By examining these plots, inferences can be drawn about how the network's prediction rules align with our understanding of the human brain's biology. Common regions of interest, such as C1, C2, CP1, and CP2, emerge in both figures, aligning with the expected activation of the motor cortex during MI tasks. However, there are still regions with high relevance that the network learnt that could be just noise of channels during sessions. For example, the high relevance of FP2 (near frontal region) for LH of subject 1 and Rest state for subject 3 could be just noise of eye movement or blinks that affect frontal channels. The high relevance of C6 near the right ear for LH of subject 3 could be ear or chewing movement during the experiment. Overfitting these noise features potentially decreases the generability of models on test sets. Based on these observations, further studies could explore methodologies for removing noisy features from deep learning models to enhance their generalisability.

In addition to the benefit of enhancing decoding accuracies by utilising diverse data sets, we also emphasize four significant **privacy-preserving attributes** of the Sandwich framework. To begin with, subject-specific and dataset-specific information, i.e. raw data and IDs, is locally stored with the respective data owners, ensuring the preservation of **data-level privacy**. Moreover, the feature extractors are stored locally to maintain **parameter-level privacy**. In addition, local feature extractors naturally encrypt the raw data during feature extraction so as to **eliminate additional encryption costs** in traditional privacy-preserving methods. Lastly, labels and classifiers could be stored and predicted locally in a multi-classifier manner, safeguarding **inference-level privacy**. All these privacy-preserving properties are seamlessly integrated into the deep transfer learning architecture design.

We also illustrated the advantages of the Sandwich framework in comparison to the benchmark transfer learning algorithm. While exhibiting a comparable capacity to leverage the source set for improving model performance, reaching (67%) as depicted in figure 10, the Sandwich network demonstrated two additional key advantages: enhanced privacy preservation and the capability to accommodate diverse input shapes. This adaptability facilitates the utilisation of heterogeneous data sets with varying configurations, as the '*DiffCh*' experiment in figure 10, which extends the Deep set sandwich to the different input channel sizes, with a comparable accuracy of 68%. The ability to adapt to diverse input shapes holds the potential to extend the applicability of algorithms in the literature, enabling their utilisation with a broader range of heterogeneous data sets.

It is important to allow dynamic input shapes in future studies to incorporate more datasets for transfer learning. As evidence, the BEETL winning benchmark establishes a higher Inception baseline, as shown in figure 6, compared to the Inception baseline in figure 10. This is achieved by using a single source dataset with identical tasks, a four second window, and 30 common channels, as opposed to a three second window and 17 common channels that are common across all possible source sets. While this setup increases baseline accuracy by expanding the data dimensions available for prediction, it does not fully reflect the primary scientific aim of developing solutions for utilising larger-scale datasets with more tasks and datasets. Therefore, to leverage both the use of more datasets and the higher and dynamic-dimensional data available for prediction, individual feature extractors, such as those in the sandwich framework, could be crucial in designing heterogeneous dataset transfer learning models in future studies.

Moreover, the Sandwich framework represents an architecture-agnostic meta-framework with the potential for application in domains beyond MI EEG decoding. As described in the methodology section and implied in the results, the feature extractors and task decoding layers can be readily substituted with different models, e.g. shallow CNNs, Inception networks, and potentially other task-specific decoders. These layers are flexible to be replaced with alternative time-series decoding layers, such as EMG decoding networks or domain-specific layers for ECoG, fMRI, etc. This flexibility highlights the versatility of the Sandwich meta-architecture, enabling its adaptation to various domains and tasks beyond EEG decoding.

It's important to highlight certain limitations and prerequisites associated with the use of the Sandwich framework. This approach, involving split individual feature extractors, is particularly beneficial when there exists a substantial difference in the distributions of various data sets, for example, in MI tasks. In scenarios where the distributional shift is relatively consistent, as in sleep stage decoding, opting for a common feature extractor may be a more advantageous design [23]. This is because a shared feature extractor allows the model to learn and adapt to small shifts in distribution autonomously. However, when

individual feature extractors are split, it can potentially reduce the amount of data available to each branch, leading to overfitting and diminishing generalisability, especially if the distributional shift is not pronounced. Moreover, it is crucial to consider the size gap among the target and source sets. In this study, there are several hundred examples in the target set and several thousand in each source set. It's worth noting that if the over-sampling of the target set is over-conducted to compensate for data shortage, the target feature extractor may overfit quickly compared to other feature extractors. This can result in unbalanced training, potentially undermining the effectiveness of transfer learning and alignment approaches. Therefore, careful consideration of data set characteristics and sizes is essential when applying the Sandwich framework to ensure optimal model performance. In addition, another limitation of the study is to test on a larger scale scalability of the target feature extraction branch. We tested with 5 and 15 subjects respectively, and the performance remained stable and outperformed the baseline. However, how the framework would perform on large-scale EEG datasets with potentially hundreds of subjects needs to be tested in future studies. Furthermore, the generalisation of our proposed models across broader brainwave datasets other than EEG poses a future direction. There also could be certain properties of other data types that EEG could not cover, fMRI, for example, is more similar to images instead of time series [53], in which the feasibility of our method needs to be tested and refined according to such specific domain when applied.

In conclusion, we introduced the Sandwich meta-architecture, exemplified through the *Inception-SD* and *ShallowConv-SD* models with transfer learning approaches. These implementations showcase a federated deep transfer learning technique designed for cross-data sets and cross-task applications. Our approach integrates privacy-preserving features with deep transfer learning methodologies, leveraging transfer learning approaches to enhance model performance. The results demonstrate that the proposed Sandwich framework, benefiting from both transfer learning and privacy-preserving characteristics, outperforms the baseline methods. This suggests that our proposed methods have the potential to effectively leverage larger, heterogeneous data sets encompassing diverse tasks for transfer learning. Notably, the Sandwich meta-framework exhibits improved privacy-preserving properties at the raw data level, parameter level and inference level. This indicates a promising avenue for advancing the application of deep transfer learning techniques in real-life scenarios with enhanced privacy considerations across borders. Furthermore, our findings suggest that the Sandwich framework holds promise as an architecture-agnostic meta-framework, capable of integrating decoding architectures and transfer learning approaches for time series data. This versatility opens up possibilities for broader applications and encourages exploration into how diverse decoding architectures could be incorporated within the Sandwich framework to improve the decoding of EEG and time series data.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.6084/m9.figshare.c.7584419 [54].

Acknowledgments

We acknowledge funding from UKRI Turing AI Fellowship (EP/V025449/1) and the High Tech Agenda Bayern to AAF.

Appendix.

Table A1. MI data sets in BEETL.

MI Data set	Subjects	Channels	Tasks
Cho2017	52	64	Left/Right hand
BNCI2014	9	22	Left/Right hand/Feet/Tongue
PhysionetMI	109	64	Left/Right hand/Feet/Both hands/Rest
Weibo2014	10	60	Left/Right hand/Feet/Rest
CybathlonIC	5	63	Left/Right hand/Feet/Rest

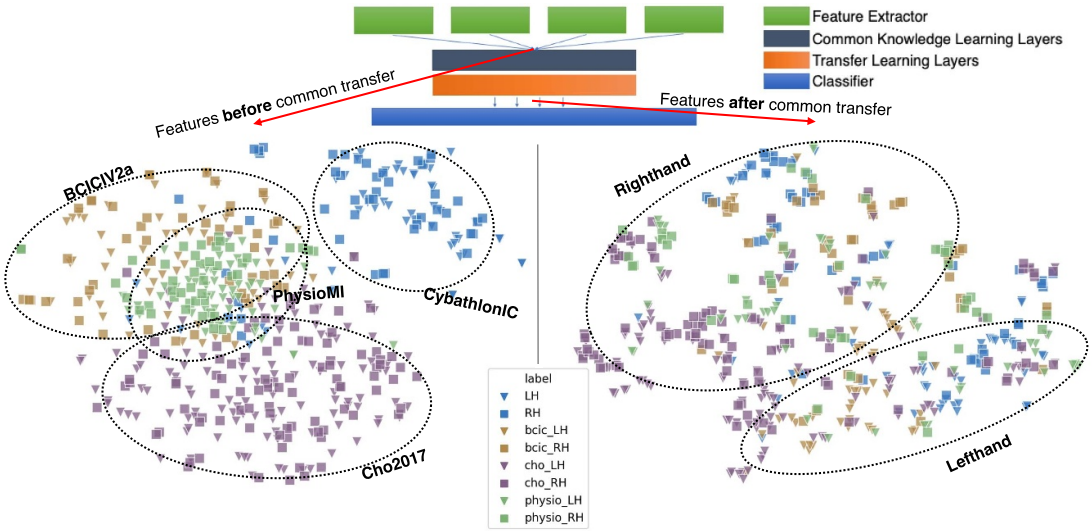


Figure A1. t-SNE feature plot across data sets before (left) and after (right) the common transfer learning layers on *Inception-SD-MMD-1Cls*.

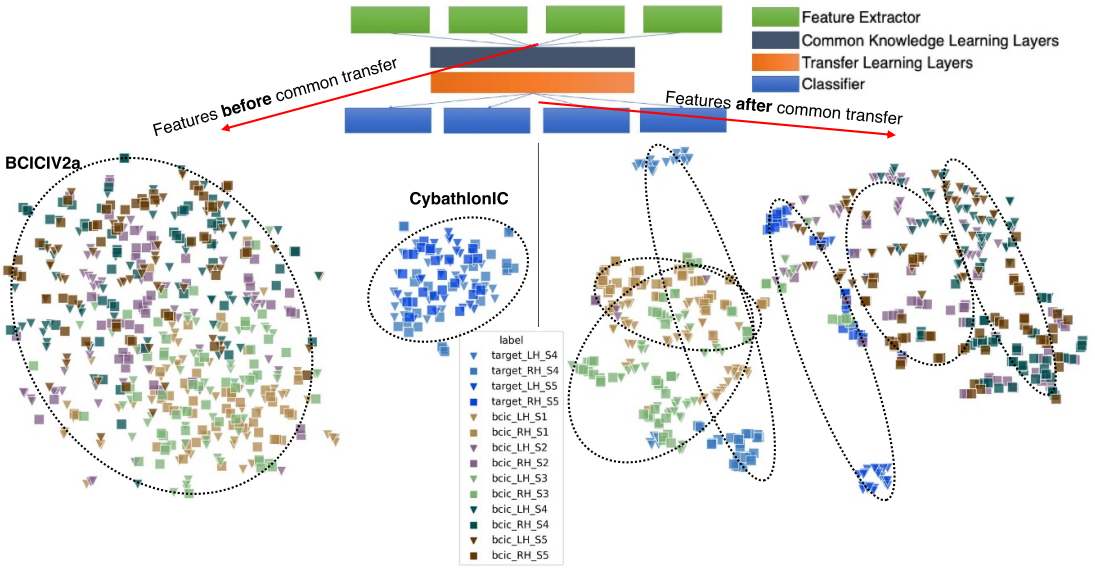


Figure A2. t-SNE feature plot across data sets before (left) and after (right) the common transfer learning layers on *Inception-SD-Deepset-MultiCls*.

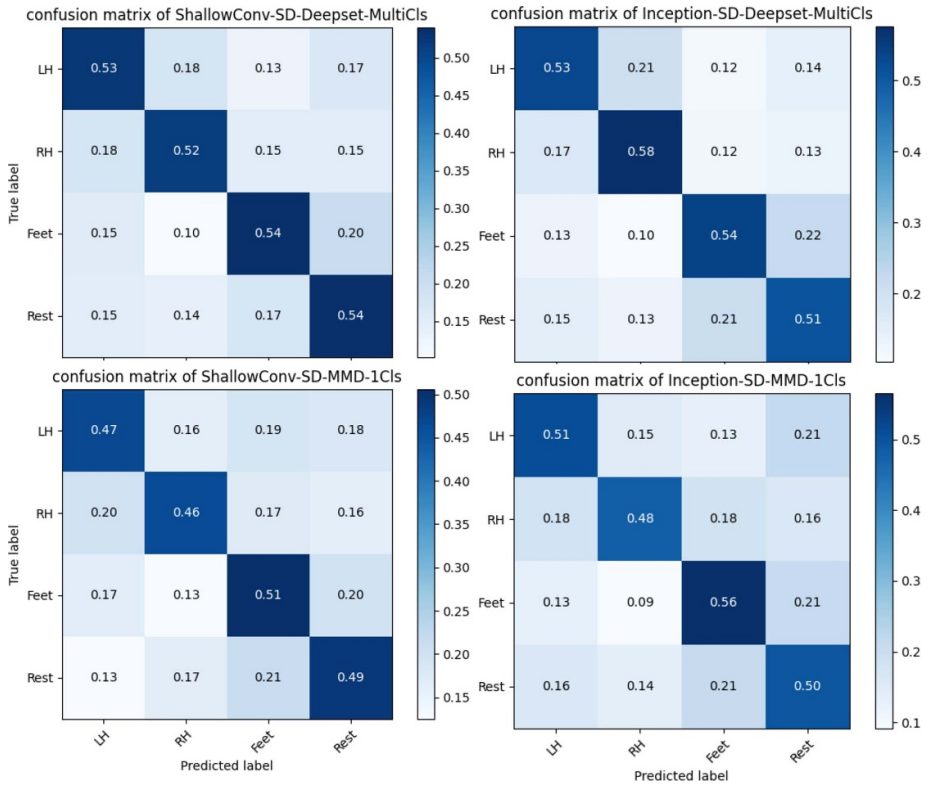


Figure A3. Confusion matrices of the sandwich models on the full CybathlonIC and Weibo test set. The confusion matrices are computed by comparing the 4-class prediction of sandwich models (before merging two classes into 'other' as in the BEETL standard) with true labels of the CybathlonIC and Weibo Dataset.

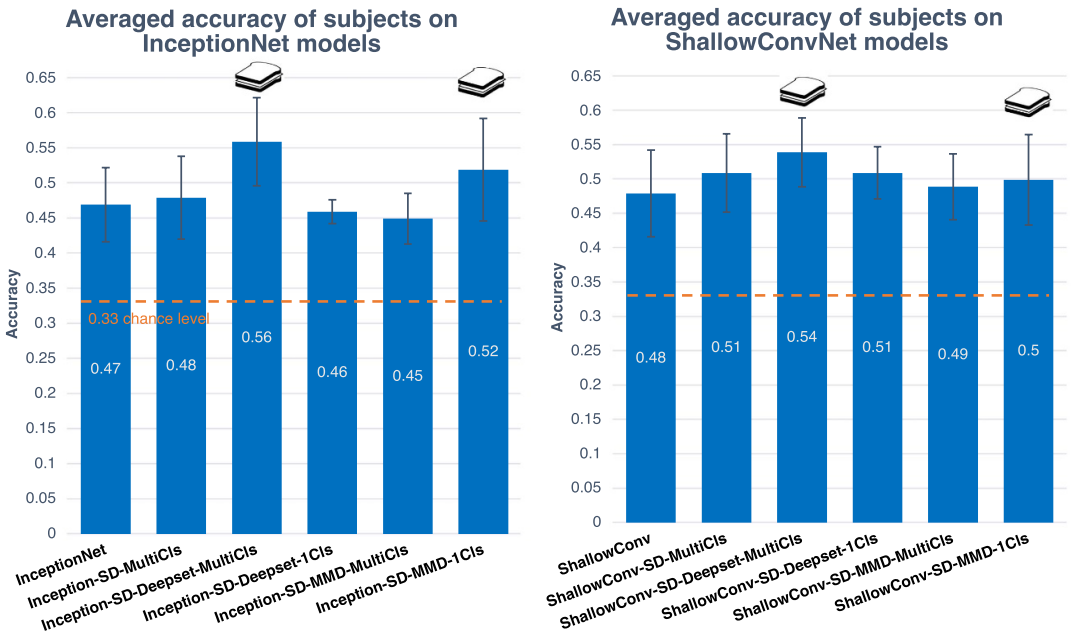


Figure A4. Averaged decoding accuracy of models with the combination of approaches on BEETL MI Test set.

ORCID iDs

Xiaoxi Wei https://orcid.org/0000-0002-9848-9179
Jyotindra Narayan https://orcid.org/0000-0002-2499-6039
A Aldo Faisal https://orcid.org/0000-0003-0813-7207

References

[1] Vaid S, Singh P and Kaur C 2015 EEG signal analysis for BCI interface: a review *5th Int. Conf. on Advanced Computing & Communication Technologies* (IEEE) pp 143–7
[2] Rashid M, Sulaiman N, Abdul Majeed P P, Musa A, Ab Nasir R M, Bari A F and Bari B S 2020 Current status, challenges and possible solutions of EEG-based brain-computer interface: a comprehensive review *Front. Neurorobot.* **14** 515104
[3] Ramadan R A and Vasilakos A V 2017 Brain computer interface: control signals review *Neurocomputing* **223** 26–44
[4] Walker I, Deisenroth M and Faisal A 2015 Deep convolutional neural networks for brain computer interface using motor imagery Imperial College *Technical Report* 68
[5] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K and Tangermann M 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[6] Acharya U R, Oh S L, Hagiwara Y, Tan J H and Adeli H 2018 Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals *Comput. Biol. Med.* **100** 270–8
[7] Oh S L *et al* 2018 A deep learning approach for Parkinson's disease diagnosis from EEG signals *Neural Comput. Appl.* **32** 1–7
[8] Dai M, Zheng D, Na R, Wang S and Zhang S 2019 EEG classification of motor imagery using a novel deep learning framework *Sensors* **19** 551
[9] Ortega P, Colas C and Faisal A A 2018 Compact convolutional neural networks for multi-class, personalised, closed-loop EEG-BCI *2018 7th IEEE Int. Conf. on Biomedical Robotics and Biomechatronics (Biorob)* (IEEE) pp 136–41
[10] Tayeb Z, Fedjaev J, Ghaboosi N, Richter C, Everding L and Qu X 2019 Validating deep neural networks for online decoding of motor imagery movements from EEG signals *Sensors* **19** 210
[11] Pan S J and Yang Q 2009 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
[12] Wei X, Ortega P and Faisal A A 2021 Inter-subject deep transfer learning for motor imagery EEG decoding *2021 10th Int. IEEE/EMBS Conf. on Neural Engineering (NER)* (IEEE) pp 21–24
[13] Weiss K, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *J. Big Data* **3** 1–40
[14] Goldberger A L *et al* 2000 PhysioBank, physiotoolkit and physionet: components of a new research resource for complex physiologic signals *Circulation* **101** e215–20
[15] Huang J T, Li J, Yu D, Deng L and Gong Y 2013 Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (IEEE) pp 7304–8
[16] Long M, Cao Y, Wang J and Jordan M I 2015 Learning transferable features with deep adaptation networks (arXiv:150202791)
[17] Jayaram V, Alamgir M, Altun Y, Scholkopf B and Grosse-Wentrup M 2016 Transfer learning in brain-computer interfaces *IEEE Comput. Intell. Mag.* **11** 20–31
[18] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M and Rakotomamonjy A 2018 A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update *J. Neural Eng.* **15** 031005
[19] Tan C, Sun F and Zhang W 2018 Deep transfer learning for EEG-based brain computer interface *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 916–20
[20] Tan C, Sun F, Fang B, Kong T and Zhang W 2019 Autoencoder-based transfer learning in brain–computer interface for rehabilitation robot *Int. J. Adv. Robot. Syst.* **16** 1729881419840860
[21] Sakhavi S and Guan C 2017 Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI *2017 8th Int. IEEE/EMBS Conf. on Neural Engineering (NER)* (IEEE) pp 588–91
[22] Hang W, Feng W, Du R, Liang S, Chen Y and Wang Q 2019 Cross-subject EEG signal recognition using deep domain adaptation network *IEEE Access.* **7** 128273–82
[23] Wei X *et al* 2022 BEETL competition: advancing transfer learning for subject independence & heterogenous EEG data sets *Proc. NeurIPS 2021 Competitions and Demonstrations Track. vol. 176 of Proc. Machine Learning Research* (PMLR) pp 205–19
[24] Smola A J, Gretton A and Borgwardt K 2006 Maximum mean discrepancy *13th Int. Conf., ICONIP* pp 3–6
[25] Li J, Chen E, Ding Z, Zhu L, Lu K and Shen H T 2020 Maximum density divergence for domain adaptation *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 3918–30
[26] Han J, Wei X and Faisal A A 2023 EEG decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks (arXiv:230613109)
[27] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R R and Smola A J 2017 Deep sets *Advances in Neural Information Processing Systems* p 30
[28] Bakas S *et al* 2022 Team cogitat at neurips 2021: benchmarks for EEG transfer learning competition (arXiv:220203267)
[29] Li L *et al* 2020 A review of applications in federated learning *Comput. Ind. Eng.* **149** 106854
[30] Hao M, Li H, Xu G, Liu S and Yang H 2019 Towards efficient and privacy-preserving federated deep learning *2019 IEEE Int. Conf. on Communications (ICC)* (IEEE) pp 1–6
[31] Lyu L, He X, Law Y W and Palaniswami M 2017 Privacy-preserving collaborative deep learning with application to human activity recognition *Proc. 2017 ACM on Conf. on Information and Knowledge Management* pp 1219–28
[32] Dong H, Wu C, Wei Z and Guo Y 2017 Dropping activation outputs with localized first-layer deep network for enhancing user privacy and data security *IEEE Trans. Inf. Forensics Secur.* **13** 662–70
[33] Du W and Atallah M J 2001 Secure multi-party computation problems and their applications: a review and open problems *Proc. 2001 Workshop on New Security Paradigms* pp 13–22
[34] Truong N, Sun K, Wang S, Guitton F and Guo Y 2021 Privacy preservation in federated learning: an insightful survey from the GDPR perspective *Comput. Secur.* **110** 102402
[35] Popescu A B *et al* 2021 Privacy preserving classification of EEG data using machine learning and homomorphic encryption *Appl. Sci.* **11** 7360
[36] Xia K, Duch W, Sun Y, Xu K, Fang W and Luo H 2022 Privacy-preserving brain–computer interfaces: a systematic review *IEEE Trans. Comput. Soc. Syst.* **10** 2312–24
[37] Ju C, Gao D, Mane R, Tan B, Liu Y and Guan C 2020 Federated transfer learning for EEG signal classification *Proc. IEEE* pp 3040–5

[38] Bethge D, Hallgarten P, Grosse-Puppendahl T, Kari M, Mikut R and Schmidt A 2022 Domain-invariant representation learning from EEG with private encoders *ICASSP 2022* (IEEE) pp 1236–40
[39] Kapitonova M, Kellmeyer P, Vogt S and Ball T 2022 A framework for preserving privacy and cybersecurity in brain-computer interfacing applications (arXiv:220909653)
[40] Cho H, Ahn M, Ahn S, Kwon M and Jun S C 2017 EEG datasets for motor imagery brain–computer interface *GigaScience* **6** gix034
[41] Tangermann M, Müller K R, Aertsen A, Birbaumer N, Braun C and Brunner C 2012 Review of the BCI competition IV *Front. Neurosci.* **6** 55
[42] Schalk G *et al* 2004 BCI2000: a general-purpose brain-computer interface (BCI) system *IEEE Trans. Biomed. Eng.* **51** 1034–43
[43] Hwaidi J F and Chen T M 2022 Classification of motor imagery EEG signals based on deep autoencoder and convolutional neural network approach *IEEE Access* **10** 48071–81
[44] Yi W, Qiu S, Wang K, Qi H, Zhang L and Zhou P 2014 Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery *PLoS One* **9** e114853
[45] Santamaria-Vazquez E, Martinez-Cagigal V, Vaquerizo-Villar F and Hornero R 2020 EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** 2773–82
[46] Long M, Cao Y, Wang J and Jordan M 2015 Learning transferable features with deep adaptation networks *Int. Conf. on Machine Learning* (PMLR) pp 97–105
[47] Bakas S, Ludwig S, Adamos D A, Laskaris N, Panagakis Y and Zafeiriou S 2023 Latent alignment with deep set EEG decoders (arXiv:231117968)
[48] Gramfort A, Luessi M, Larson E, Engemann D A, Strohmeier D and Brodbeck C 2013 MEG and EEG data analysis with MNE-python *Front. Neurosci.* **7** 267
[49] Van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 11
[50] Vidaurre C, Schlogl A, Cabeza R, Scherer R and Pfurtscheller G 2006 A fully on-line adaptive BCI *IEEE Trans. Biomed. Eng.* **53** 1214–9
[51] Perdikis S, Tonin L, Saeedi S, Schneider C and Millán J R 2018 The Cybathlon BCI race: successful longitudinal mutual learning with two tetraplegic users *PLoS Biol.* **16** e2003787
[52] Zar J H 1999 *Biostatistical Analysis* (Princeton-Hall)
[53] Glover G H 2011 Overview of functional magnetic resonance imaging *Neurosurg. Clin.* **22** 133–9
[54] Wei X and Faisal A 2024 2021 BEETL competition: advancing transfer learning for subject independence and heterogenous EEG sata sets figshare Collection (https://doi.org/10.6084/m9.figshare.c.7584419)