Beyond Within-Subject Performance: A Multi-Dataset Study of Fine-Tuning in the EEG Domain

Christina Sartzetaki*‡, Panagiotis Antoniadis*‡, Nick Antonopoulos‡, Ioannis Gkinis‡,
Agamemnon Krasoulis†§, Serafeim Perdikis[x], Vassilis Pitsikalis‡
‡Deeplab, Athens, Greece §Insilico Medicine AI Limited, Abu Dhabi, UAE [x] University of Essex, Colchester, UK

***Abstract*—There is a critical demand for BCI systems that can swiftly adapt to a new user and at the same time function with any user. We propose a fine-tuning approach for neural networks that serves a dual purpose; first, to minimize calibration times through requiring considerably less data - up to one-sixth - from the target subject than training from scratch, and second, to alleviate cases of user illiteracy by providing a substantial performance boost of over 11% in absolute accuracy from the features learned from other subjects. Ultimately, our adaptation method surpasses standard within-subject performance by a large margin in all subjects. We present ablation studies across three datasets, in which we demonstrate that fine-tuning outperforms other adaptation methods for BCI systems and that what matters most is the quantity of pre-training subjects, rather than their BCI-ability, achieving over 8% absolute increase in classification accuracy when scaling up the order of magnitude. Finally, we compare our approach to the state-of-the-art in EEG-based motor imagery and find it comparable, if not superior, to methods employing far more complex neural networks, obtaining 82.60% and 85.64% within-subject accuracy in the four-class BCIC IV-2a and binary MMI datasets respectively.**

***Index Terms*—BCI, EEG, Motor Imagery, Domain Adaptation, Fine-tuning, BCI-illiteracy**

I. Introduction

Controlling computer interfaces with brain signals is a complex and ambitious objective which holds great promise, and has the capacity to remarkably impact the world through applications in healthcare, communication, and entertainment. For individuals with disabilities, brain computer interfaces (BCIs) could provide an unprecedented level of independence, by allowing them to interact with the world around them using only their mind [1]. In other cases, BCIs can be employed to assist with recovery from neurological conditions, such as stroke or spinal cord injuries [2]. Even for able-bodied individuals, the ability to interact with computers and devices (e.g. in a virtual reality or drone control application) without any physical input could enhance the overall experience, making it more comfortable, seamless, and intuitive.

However, brain signals, specifically those recorded with non-invasive electroencephalogram (EEG) devices, are characterized by some unique traits when compared to other machine learning data domains, introducing multiple challenges that must be overcome for the successful development of a practical BCI. First and foremost, the signals recorded from different users constitute different domains [3]. Without adaptation measures, the adoption of the system by a new user results in skewed probability distributions and distinctly low performance [4]. Secondly, it would appear that not all users possess the ability to effectively control BCIs, particularly in the case of motor imagery (MI) BCIs [5], with low-performing users often being described as "BCI-illiterate". Lastly, large datasets are scarce in the EEG domain, as there is no readily accessible and abundant source of data and they need to be recorded with human effort. On top of these, the functionality of such a system in real-time commercial applications necessitates fast response times with low-cost, dry electrode devices which decrease the signal to noise ratio (SNR), as well as minimal calibration times when bootstrapping a new user.

In order to confront these challenges, various approaches have been proposed in the literature, including fine-tuning (FT) [6]–[8] for domain adaptation, hybrid BCIs [9], [10] to mitigate BCI-illiteracy, and advanced neural network architectures [11], [12] designed to learn more robust features.

In this work, we attempt to address the same issues in a unified manner, through evaluating the adaptation method of FT across MI datasets that differ in both their scale as well as the recording device's grade and resolution. Through extensive ablations and probings, we examine its robustness under constraints on the amount of data used in the different training stages and the performance level of the subjects used in training and evaluation. FT consistently surpasses within-subject performance by over 7% in terms of absolute accuracy. Most remarkably, we find that the maximal benefits of FT are reaped when scaling up the number of pre-training subjects and minimising calibration data, where it transcends within-subject by 8% and 34% respectively. Equally outstanding are cases of low-performing subjects who, after a boost of over 11% from FT, attain acceptable performance for the operation of a BCI ($> 70\%$ accuracy).

II. Related Work

In recent years, EEG-BCIs have become an area of focus in the machine learning and deep learning community. Neural networks have exhibited promise in reducing the necessity for expert-crafted features and manual subject-specific filters, thereby potentially supplanting traditional (classical) machine learning approaches. The latter most commonly consist of

*These authors contributed equally to this work.
†Work done by Agamemnon Krasoulis while at Deeplab.
Contact: {c.sartzetaki, p.antoniadis, nanton, g.gkinis, vpitsik}@deeplab.ai, a.krasoulis@insilicomedicine.com, serafeim.perdikis@essex.ac.uk

classifiers (e.g. random forest, linear discriminant analysis (LDA), and support vector machines) trained on features such as power spectral density [13], covariance in a Riemannian manifold [14], common spatial patterns [15], and filter bank common spatial patterns [16]. This effort has been succeeded by artificial neural networks (ANNs) which learn features and decision surfaces based solely on the time-domain input. Initial architectures such as DeepConvNets [17], C2CM [18], and EEGNet [19], were followed by more advanced ANNs that pushed the state-of-the-art to new highs such as EEG-TCNET [11] and EEG-ITNET [12]. It is worth noting that even the most complex of these architectures still tend to be more shallow when compared to those utilized in other domains. Although some deeper architectures like TIDNet [7] and Inception [20] have been proposed, larger networks often display signs of overfitting when trained on data from the EEG domain [17]. This can be attributed to the available data being insufficient to train these large architectures.

Numerous adaptation techniques have been developed to tackle the challenge of making a system operational across users, requiring a minimal calibration set (e.g. five minutes) for each new user instead of the full data collection process (approx. one hour). We proceed to further sub-categorize these based on the stage in which they are applied, into signal level, model level, and prediction level adaptation methods.

a) Signal level: When pre-processing the signal before input to a model, a common practice in the classical methods is to select user-specific filters [16], [21] depending on their efficacy, either manually or via feature selection. Another widely adopted signal level method is Euclidean alignment (EA) [22], a highly effective form of whitening, along with Riemannian alignment which performs the same operation in the Riemannian manifold; these can more accurately be viewed as domain generalization, rather than adaptation methods [7].

b) Model level: Inspired by other domains where deep learning has given breakthrough results, the concept of fine-tuning (FT) an ANN has been explored [6], [7], along with layer freezing [8]. In the same vein, the technique of knowledge distillation (KD) as introduced by Hinton et al. [23], has been adapted to the EEG domain with considerable success in the seizure detection [24] as well as MI [25] tasks, being further combined with FT in the latter. The model adaptation methods investigated for EEG also include self supervised learning (SSL), which leverages larger unlabeled datasets to pursue functionality across users [26], as well as alternative approaches attempting a subject distribution transfer [27].

c) Prediction level: After obtaining a trained model from a set of users, one easy technique to adapt the skewed probability distribution for a new user is to tune the decision surface, thus obtaining user-specific thresholds. This is referred to as threshold-moving in other domains, mostly applied for class imbalance [28], or as "unbiasing" for a BCI [29].

Correspondingly for the problem of BCI-illiteracy there is also a multitude of strategies that have been employed. One established approach focuses on adapting the user rather than the machine, and involves recording sessions with feedback from a real-time system that steers users into improving their imagined movement, as well as addresses cases of user inattentiveness, boredom, and restlessness [1], [9], [30]. Another strategy is multi-modality where another modality apart from EEG is used to make predictions more robust to user BCI-ability [31], [32]. Similarly, a hybrid BCI with paradigm [9] or target [10] switching is often utilized, where different paradigms such as MI, steady state visually evoked potential (SSVEP) and event-related potential (P300), and targets e.g. left, right, feet, tongue, are alternated and combined differently in different users to form a more robust BCI.

We now contrast our work to the presented literature, each time highlighting the unique approach taken and its primary benefits. Compared to the literature where a lot of effort is invested on more complex model architectures [11], [12], we utilize a simple EEGNet [19] model and achieve a competitive to the state-of-the-art accuracy of 82.60% in the BCIC IV-2a dataset [33] through user adaptation. After a multi-dataset comparison of adaptation methods from the signal, model, and prediction level categories, we propose utilizing FT, either as a stand-alone adaptation, or combined with EA. Our approach is in line with the FT in previous works [6], [7] to which we are competitive in the MMI dataset [34] with 85.64%, however, we distinctively probe this methodology with a unique set of experiments on the amount of data and the type of subjects used. We show for the first time, to the best of our knowledge, that FT can also be used to address the issue of BCI-illiteracy, where it surpasses a hybrid BCI [10] approach.

Ultimately we make the following key contributions:

- We compare adaptation methods from three categories across three datasets where the best-performing method, FT, proves competitive to the state-of-the-art and consistently surpasses the within-subject baseline by over 7%.
- We show that FT outperforms within-subject by an astonishingly larger margin of 34% with limited target data, indicating its fitness for use with a small calibration set. The quantity of subjects used for pre-training is more crucial than their performance level, achieving over 8% improvement when scaling up the order of magnitude.
- We demonstrate that FT can be used as an illiteracy mitigation strategy to benefit low-performing subjects substantially, over 11%.

III. Methodology

We choose the EEGNet [19] architecture as the core model for our experiments. We compare the proposed fine-tuning (FT) adaptation method with other BCI adaptation methods, namely the signal-level Euclidean alignment (EA) [22], model-level knowledge distillation (KD) [23], and prediction-level thresholding [29]. To examine the robustness of FT under constraints on the amount of data and the performance level of the subjects used, we design three evaluation experiments.

A. Deep Learning Architecture

We limit our experiments on a single EEG-specific deep learning architecture called EEGNet [19]. It consists of the

temporal, depthwise, and separable convolutional layers. The temporal convolution layer learns the temporal filters for each channel of the raw EEG signal, while the depthwise convolution layer learns the so called "spatial filters" across the EEG channels on the scalp. The separable convolution layer combines a depthwise and a pointwise convolution.

B. Baseline Adaptation Methods

1) Euclidean Alignment (EA): A domain generalization method proposed for EEG signals [22] that is based on Correlation Alignment [35]. In this method, each subject's data are mapped into a common feature space by performing whitening with their mean covariance matrix. For each trial X_i of a subject with n trials, we compute $\tilde{X}_i = \bar{R}^{-1/2} X_i$, where $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$. This results in the mean covariance of each subject being equal to the identity matrix.

2) Knowledge Distillation (KD): A technique used in deep learning to transfer the knowledge of a teacher model to a student model [23] that has also been investigated for EEG-BCIs [24], [25]. The student model is trained to minimize the loss function $\mathcal{L} = (1-\alpha)\mathcal{L}_{student} + \alpha\mathcal{L}_{distillation}$, in which the distillation term helps to mimic the behavior of the teacher model and generalize better on unseen data through measuring the similarity between the models' output probabilities.

3) Thresholding: A threshold adaptation method implemented on a post-processing level. This was inspired by [29] where a modified LDA classifier that computes a decision threshold ensuring equivalent error rates for both classes of a binary BCI task is proposed. In our case, the decision threshold applied to the trained network's output probabilities is optimized with respect to each subject's data. We also extend this method to multiple classes by performing a hierarchical one vs. rest thresholding for each class.

C. Proposed Fine-Tuning (FT) Adaptation

Fine-tuning (FT) is a standard transfer learning technique applied in deep learning [36]. In the case of domain adaptation, it involves taking a model trained on a source domain D_S called the pre-trained model, and re-training it on the target domain D_T. The model is initialized with the learned parameters from pre-training, which it then adjusts through training on the data from D_T. In this manner, FT realizes the core objective of transfer learning, i.e. to learn the target conditional probability distribution $P(Y_T|X_T)$ in D_T with the information gained from D_S, where $D_S \neq D_T$. Due to the nature of most application settings, the source domain D_S often contains a considerable amount of data, while the target domain D_T can only provide limited data, therefore FT can improve performance on D_T by leveraging the information obtained from D_S.

D. Evaluation strategy

The standard approach for evaluating an adaptation method is commonly limited to comparison with alternatives and the current state-of-the-art. In order to provide a more thorough analysis of the impact of FT, we present a series of experiments and evaluations that delve deeper into its effects and demonstrate its usefulness in different application scenarios.

1) Limited calibration data: The process of collecting subject-specific data for a BCI application is both expensive and time-consuming. Consequently, the minimum quantity of subject-specific data required for a BCI adaptation method is a crucial parameter. To investigate this aspect, we perform an experiment where the validation and test sets are kept constant while the size of the training data is successively reduced, initially to $\frac{1}{3}$, and then to $\frac{1}{6}$ of the original. Each time, we compare the performance of the adaptation method to a within-subject model trained with the same limited data.

2) Illiteracy mitigation: To evaluate the ability of a domain adaptation method to address the problem of BCI-illiteracy, we estimate the distribution of performance scores for all subjects before and after using the method, focusing on the low-performing subjects. For binary tasks, a threshold of 70% is commonly used for illiteracy, however there is no agreement on the accepted proficiency threshold for more than two classes [9]. Therefore, to err on the side of caution, we select the 70% threshold for both the binary and the 4-class tasks.

3) Pre-training subjects selection: Lastly, we devise an experiment aimed at examining the impact of the number and performance level of pre-training subjects on the performance of the fine-tuned model, enabled by the large number of subjects in the MMI dataset. Specifically, we select a hold-out test set of 17 subjects so that they represent the overall distribution of within-subject performance. We experiment with selecting subsets of the remaining 88 subjects for pre-training, based on their within-subject performance, namely 25 high-performing (HP) subjects of $> 90\%$ binary accuracy, 25 low-performing (LP) subjects of $< 70\%$ binary accuracy, and 25 "mixed" subjects that cover the entire range of performances. We also experiment with the number of pre-training subjects by selecting a subset of nine subjects out of each of the aforementioned subsets, as well as pre-training with the whole set of ("mixed") 88 subjects.

IV. Experimental Setup

A. Datasets

1) BCIC IV-2a (BCI Competition IV): The dataset 2a [33] from the BCI Competition IV [37] contains the EEG recordings of nine subjects. Each subject performed a total of 12 runs, each containing four MI tasks, namely left-hand, right-hand, both-feet, and tongue, in two separate sessions recorded on different days (six runs per session). Each run contains 48 imagined movement trials that last for 4s. The EEG recordings were collected using 22 Ag/AgCl electrodes in the international 10-20 system, sampled at 250Hz.

2) MMI (Physionet): The Motor Movement/Imagery (MMI) dataset [34] from the PhysioNet [38] repository consists of 109 subjects. Each subject performed three runs of imagined left or right hand movement, three runs of imagined both-hands or feet movement, and the non-imagined counterparts of both of these sets of runs, totaling to 12 recorded runs

TABLE I
ABLATION STUDY OF ADAPTATION METHODS ACROSS THREE DATASETS

	BCIC IV-2a 4-class	**MMI** 2-class	**DUEMI** 3-class
Baselines			
Within-subject k-fold	73.09 ± 11.23	77.83 ± 14.19	57.19 ± 17.37
Cross-subject LO/MSO[a]	57.29 ± 8.72	83.35 ± 12.54	54.07 ± 15.75
Chance level at $\alpha = 0.05$[b]	25.00 ± 6.06	50.00 ± 22.48	33.33 ± 15.84
Adaptation Methods			
EA [22]	62.31 ± 10.36	83.02 ± 11.24	60.96 ± 17.71
KD [23]	78.47 ± 8.34	81.11 ± 13.00	62.30 ± 15.60
Thresholding	70.50 ± 6.23	84.38 ± 13.95	61.85 ± 20.55
Fine-tuning (FT)	80.75 ± 7.03	$\mathbf{85.64 \pm 12.67}$	68.59 ± 15.34
FT + EA	$\mathbf{82.60 \pm 5.91}$	85.22 ± 11.15	$\mathbf{68.81 \pm 15.84}$
FT + KD	79.01 ± 8.05	80.73 ± 13.45	64.81 ± 17.12

[a]MMI is 5-fold LMSO while the others are LOSO.
[b]Calculated according to [40], 5% confidence intervals reported.

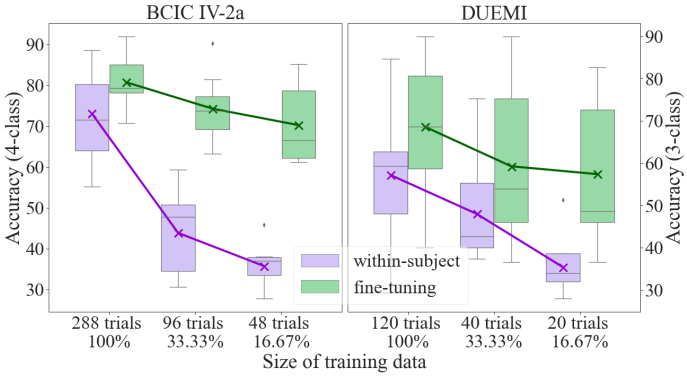

Fig. 1. The effect of the target subject's data size on training from scratch and fine-tuning, for two datasets. The performance gap between FT and the within-subject baseline increases for both datasets when the number of trials is limited.

per subject. In this work we focus on the three imagined left or right hand movement runs. Each run contains 15 imagined movement trials that last for 4s. The EEG recordings were collected using 64 electrodes in the international 10-10 system and were sampled at 160Hz. After processing the dataset, and similar to [7], we find that subjects S088, S090, S092 and S100 contain unusable data, leading to 105 subjects being used.

3) DUEMI (Ours): The Deeplab Unicorn EEG Motor Imagery (DUEMI) dataset is employed for the first time in this work and consists of EEG data from nine subjects acquired in-house. All participants signed informed consent and the study was conducted according to the mandates of the Declaration of Helsinki. Each subject performed 10 runs, each containing two MI tasks (left hand and right hand) and an additional rest task where the subject does not imagine any movement, but continues to avoid blinking and moving as during the imaginary movement. Each run contains 15 trials that last for 4s. The EEG recordings were collected with eight electrodes and sampled at 250 Hz, using the Unicorn Hybrid Black commercial EEG device by g.tec [39].

B. Preprocessing

1) Common Average Referencing (CAR): For all three datasets, we re-reference the data to the channel average with CAR [41], described by the simple operation $x'_j = x_j - \frac{1}{C}\sum_{j=1}^{C} x_j$ for each channel j where C is the total number of channels. By removing any common external factor from all channels, CAR can reduce noise by up to 30% [41].

2) Trial segmentation: In BCIC IV-2a and MMI we use the [0, 4] segment post cue onset while in DUEMI we use the [1.5, 5.5] segment post cue onset, performing no further windowing of the signal in either case.

C. Evaluation

For our metrics, given that in all datasets classes are fairly balanced, we use the binary accuracy for MMI, 3-class accuracy for DUEMI, and 4-class for BCIC IV-2a. We evaluate our models and adaptation methods using the following protocols.

1) Within-subject k-fold cross validation (CV): To calculate within-subject performance, we use the k-fold CV protocol. Within-subject performance refers to the performance of a model trained on a subject's data, and evaluated on a test set of the same subject's data. We also use the k-fold CV protocol when we adapt pre-trained models to the data of a target subject using one of the domain adaptation methods (FT, KD, thresholding).

2) LO/MSO: To calculate cross-subject performance, we use the Leave One Subject Out (LOSO) protocol, where we train a model on the $N-1$ out of the total N subjects, and test it on the one left-out subject. In the case of the MMI dataset we use the Leave Multiple Subjects Out (LMSO) protocol, where the N subjects are divided into k folds, and each time the model is trained on $\frac{(k-1)N}{k}$ subjects and tested on the remaining $\frac{N}{k}$ subjects. This is done for computational reasons due to the large number of subjects in the MMI dataset.

D. Hyperparameters

In all cases, the default hyperparameters of EEGNet-v4 [19] are utilized. We use the cross-entropy loss and AdamW optimizer, along with early stopping for training our models. In BCIC IV-2a and MMI datasets we use similar training hyperparameters as in the literature. In DUEMI we train our models using a learning rate of 0.001, a batch size of 32, and a weight decay of 0.001. For KD α is set to 0.7, and for KD+FT to 0.9. For the LMSO evaluation protocol we set $k = 5$, while for k-fold CV $k = 3$.

V. RESULTS & DISCUSSION

In this section we present our results on the three datasets, BCIC IV-2a, MMI, and our own DUEMI dataset.

A. Comparison with other adaptation methods

We first show an ablation study in Table I, comparing the proposed FT with the baseline adaptation methods described in Sec. III-B. In all three datasets, either the FT, or the FT+EA method is the best-performing one, with the two being very close in all cases ($< 1.9\%$ difference). Compared to the baseline within-subject performance, we achieve a maximum absolute increase of 9.51% in BCIC IV-2a, 7.81% in MMI, and 11.62% in DUEMI. This is an important outcome, as it shows

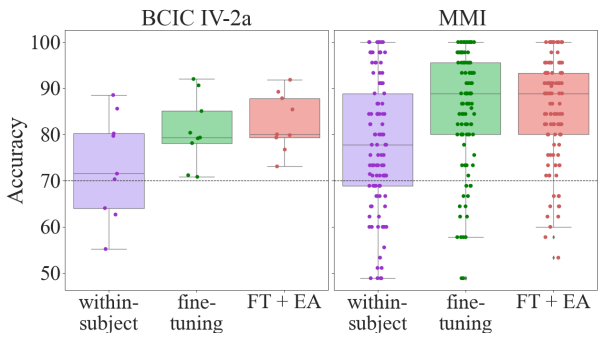

Fig. 2. The influence of FT on subjects with low performance. FT is especially effective for low performing subjects as it succeeds in raising the performance of numerous subjects above the 70% usability threshold. Indicatively in MMI, the absolute improvement with FT for the bottom 50% of subjects is 11.34%.

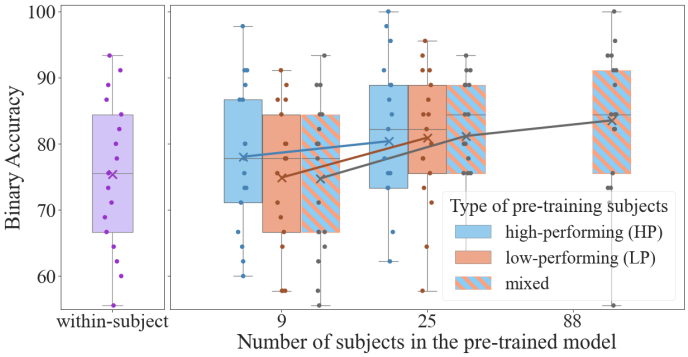

Fig. 3. The effect of the pre-training subjects' type and quantity on the performance of the fine-tuned model. Results are for the MMI dataset on a hold-out set of 17 subjects. The number of subjects explored ranges from the order of magnitude of 10^1, up to 10^2. The number of pre-training subjects is more important than their performance level.

that pre-training on multiple subjects before training a subject-specific model can greatly improve performance, indicating that neural networks can benefit from the diverse information provided by the different subjects.

An important observation is that in large datasets such as MMI, with a number of subjects in the order of magnitude of 10^2, cross-subject performance can out-of-the-box exceed within-subject performance, especially if the data available for each subject are limited as is the case in the MMI dataset. FT further improves the cross-subject baseline by 2.29% in MMI.

In BCIC IV-2a, both EA and thresholding provide an important boost over the cross-subject baseline, but fail to surpass within-subject performance without re-training on the target subject's data, just by observing the second order statistics and tuning the decision threshold respectively. In DUEMI they both manage to exceed within-subject performance, however in that case this can be attributed to the baseline being considerably low, as the dataset is recorded on roughly one third of the channels compared to BCIC IV-2a. Similarly in MMI, they both improve within-subject performance, with only thresholding surpassing the cross-subject baseline. Finally, in all datasets KD manages to improve within-subject performance but fails to reach the performance boost of FT, as well as the cross-subject baseline in MMI. However, the primary objective of KD is to compress the knowledge from a larger teacher model into a smaller student model, rather than perform domain adaptation explicitly. Its combination with FT can lead to better results than KD itself, especially in DUEMI, but still fails to reach the stand-alone FT performance, which is in line with the results by Sakhavi et al. [25].

B. Limited calibration data

In Fig. 1, we can see the results of the limited calibration data experiments described in Sec. III-D1. We are confining this experiment to the BCIC IV-2a and DUEMI datasets as MMI contains a very limited amount of per-subject data to begin with. We observe that FT is much more robust than within-subject when the amount of data per-subject is limited, which enables its use in a short calibration session before usage of a BCI system by a new user. Specifically, we can fine-tune a pre-trained model with just 48 four-second trials per subject in BCIC IV-2a (corresponding to 12 trials per class), which is equivalent to just five minutes of calibration, and retain an average target subject performance of 70.25%, whereas within-subject performance would drop to 35.69% rendering the system unusable. Similar results can be seen in our in-house DUEMI dataset, with the within-subject performance dropping to 35.41% and FT retaining a 57.41% when only 20 trials are used for fine-tuning the pre-trained model. Indeed, we note that the performance of FT with $\frac{1}{6}$ of the trials is equivalent to that of training from scratch with all the trials, with the clear advantage of the minimal calibration time.

C. Illiteracy mitigation

One of our most interesting findings can be observed in Fig. 2 concerning the evaluation of FT for illiteracy mitigation as explained in Sec. III-D2. We observe that the improvement on the low-performing subjects is larger than the average improvement, indicating that FT can be employed to specifically target this type of subjects. The bottom 50% of subjects in terms of performance in BCIC IV-2a and MMI attain average within-subject scores of 64.76% and 65.86% respectively, while after FT they achieve 78.02% and 77.20%, thus surpassing the illiteracy threshold of 70% to efficiently control a BCI. From another viewpoint, training from scratch results in 3/9 subjects below the 70% threshold in BCIC IV-2a and 31/105 subjects in MMI. These are reduced to 0/9 and 15/105 respectively after fine-tuning multi-subject models on the target subjects' data. This provides evidence that FT can be a highly effective method for the mitigation of the illiteracy problem, on top of all its other merits.

D. Pre-training subjects selection

In Fig. 3 we benchmark the fine-tuned model's performance for different subsets of pre-training subjects, as described in Sec. III-D3. We observe that the type of the pre-training subjects does not influence the overall performance, as much

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON BCIC IV-2A (4-CLASS)

	Split	Accuracy
Deep ConvNet [17]	Competition	70.90
Shallow ConvNet [17]	Competition	73.70
EEG-TCNET [11]	Competition	$\mathbf{77.35 \pm 11.57}$
FT+EA (Ours)	Competition	76.43 ± 8.55
EEGNet[a] [19]	k-fold	70.78 ± 9.09
C2CM [18]	k-fold	74.46 ± 14.43
EEG-ITNET [12]	k-fold	76.74 ± 11.48
FT+EA (Ours)	k-fold	$\mathbf{82.60 \pm 5.91}$

[a] The vanilla EEGNet result was reproduced by us.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON MMI (2-CLASS)

	Accuracy
A-cVAE [42]	63.80
EEGNet[a] [19]	76.15 ± 15.21
Dose et al. [6]	86.49 ± 10.09
TIDNet [7]	**88.19**
FT (Ours)	85.64 ± 12.67

[a] The vanilla EEGNet result was reproduced by us.

as their quantity. Specifically, when changing the order of magnitude from 10^1 pre-training subjects to 10^2, we perceive a 8.76% increase in the average performance of the FT method, whereas the variability between the average performance when using a different type of pre-training subject is $< 0.79\%$ for the 25 subjects. This is especially interesting when only pre-training on the LP subjects (defined in Sec. III-D3), who cannot achieve more than 70% in within-subject performance but manage to boost the average within-subject accuracy by 5.5%. In contrast, when pre-training with nine subjects, we observe that pre-training only with HP subjects resulted in a 3.14% absolute increase over the two other options (LP and "mixed"), potentially indicating that when few subjects are available for pre-training, their performance level becomes more important. In addition, the top HP subjects are benefited more when they are pre-trained on HP subjects, with this configuration surpassing the other two for the top five subjects by 3.11% and 3.56%, in the nine and 25 subject configurations respectively. These top five subjects achieve an overall 94.67% when 25 HP subjects are used, which also surpasses the 92% baseline which results from all 88 subjects being used for pre-training.

E. Comparison with state-of-the-art methods

In Table II we compare our best adaptation method, the combination of FT with EA, to the state-of-the-art in the BCIC IV-2a dataset. This comparison is done with the two most common within-subject evaluation protocols found in literature. First, with the split given for the purposes of the BCI Competition (which corresponds to one session for training and one for testing per subject) we achieve a performance of 76.43% 4-class accuracy, which is comparable to the state-of-the-art model EEG-TCNET [11]. Second, with k-fold cross-validation on both available sessions, our proposed method surpasses the best-performing model EEG-ITNET [12] by an absolute 5.86%. In both evaluation protocols, our method leads to considerably smaller standard deviation than the competition, attesting to a normalizing effect on subject variability achieved by the substantial improvement of LP subjects.

We now present a comparison to the target-switching hybrid BCI by Wang et al. [10] which reports 80.98% for binary classification on the BCIC IV-2a dataset. Their method specifically improves the low-performing subjects, similar to ours, but in doing so they sacrifice the multi-class capabilities of the BCI, limiting it to only two classes. Although we cannot compare exactly, we achieve 82.60% in the four-class task using a method less complex to implement in a real-life scenario.

In Table III our best performing FT approach is evaluated against the state-of-the-art in the MMI dataset. All the works under consideration here leverage some form of transfer learning for adaptation on the target subjects, presumably due to the limited amount of data available per subject in the MMI dataset. Our method attains an 85.64% binary accuracy that is slightly inferior if not competitive to the leading FT method in TIDNet [7], which is nonetheless enhanced by per-subject optimal combinations of EA and regularization. Compared to [6] where subjects are split into a single fold, we obtain results for the entirety of the subjects through fine-tuning the respective pre-trained models from the LMSO protocol.

The results of the competition in Tables II and III are those reported in their original publications, with the exception of the vanilla EEGNet which we reproduced. Compared to the vanilla EEGNet, we achieve a total boost of 11.82% in BCIC IV-2a and 9.49% in MMI by applying our proposed adaptation method and preprocessing pipeline.

VI. CONCLUSION

In conclusion, our proposed fine-tuning (FT) approach shows significant promise in addressing two challenges that are present in EEG data, namely adapting to new users with minimal calibration times and mitigating user illiteracy. Through extensive ablations and comparisons, we demonstrate that FT outperforms other adaptation methods and consistently surpasses within-subject performance by a large margin. Furthermore, the quantity of pre-training subjects appears to be the most important factor in achieving the maximal benefits from FT, rather than their literacy level. This observation emphasizes the crucial need for datasets with a larger number of subjects in the field of EEG-based BCIs.

In future work, we intend to benchmark FT and other adaptation methods using more sophisticated deep learning architectures. Additionally, we intend to investigate the possibility of fine-tuning a pre-trained model from a large public dataset such as MMI, on the subjects of the smaller datasets, in order to further leverage large-scale pre-training. Finally, we aim to explore a broader range of transfer learning techniques, including self-supervised learning (SSL) [26].

ACKNOWLEDGMENT

This project was funded by Deeplab (https://deeplab.ai/), as part of Deeplab's research activities.

REFERENCES

[1] S. Perdikis, L. Tonin, S. Saeedi, C. Schneider, and J. d. R. Millán, "The cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users," *PLoS Biol.*, vol. 16, no. 5, May 2018.

[2] A. Biasiucci, R. Leeb, I. Iturrate, S. Perdikis, A. Al-Khodairy, T. Corbet, A. Schnider, T. Schmidlin, H. Zhang, M. Bassolino, D. Viceic, P. Vuadens, A. G. Guggisberg, and J. D. R. Millán, "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke," *Nat. Commun.*, vol. 9, no. 1, p. 2421, Jun. 2018.

[3] H. Morioka, A. Kanemura, J. ichiro Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii, "Learning a common dictionary for subject-transfer decoding with resting calibration," *NeuroImage*, vol. 111, pp. 167–178, 2015.

[4] S. Saha and M. Baumert, "Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: A review," *Front. Comput. Neurosci.*, vol. 13, 2020.

[5] M. Ahn and S. C. Jun, "Performance variation in motor imagery brain–computer interface: A brief review," *J. Neurosci. Methods*, 2015.

[6] H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Syst. Appl.*, vol. 114, pp. 532–542, 2018.

[7] D. Kostas and F. Rudzicz, "Thinker invariance: enabling deep neural networks for BCI across more people," *J. Neural Eng.*, vol. 17, 2020.

[8] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Networks*, vol. 136, pp. 1–10, 2021.

[9] B. Z. Allison and C. Neuper, "Could anyone use a BCI?" *Brain-computer interfaces: Applying our minds to human-computer interaction*, pp. 35–54, 2010.

[10] T. Wang, S. Du, and E. Dong, "A novel method to reduce the motor imagery BCI illiteracy," *MBEC*, vol. 59, no. 11-12, Nov. 2021.

[11] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in *IEEE SMC*, 2020.

[12] A. Salami, J. Andreu-Perez, and H. Gillmeister, "EEG-ITNet: An explainable inception temporal convolutional network for motor imagery classification," *IEEE Access*, vol. 10, pp. 36 672–36 685, 2022.

[13] S. Bhattacharyya, A. Khasnobish, A. Konar, D. Tibarewala, and A. K. Nagar, "Performance analysis of left/right hand movement classification from EEG signal by intelligent algorithms," in *IEEE CCMB*, 2011.

[14] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[15] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topography*, vol. 2, no. 4, pp. 275–284, Jun. 1990.

[16] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter Bank Common Spatial Pattern algorithm on BCI competition IV datasets 2a and 2b," *Front. Neurosci.*, vol. 6, 2012.

[17] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, 2017.

[18] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, 2018.

[19] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, 2018.

[20] E. Santamaría-Vázquez, V. Martínez-Cagigal, F. Vaquerizo-Villar, and R. Hornero, "EEG-Inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabilitation Eng.*, vol. 28, no. 12, pp. 2773–2782, 2020.

[21] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Muller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE TBME*, vol. 53, 2006.

[22] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A euclidean space data alignment approach," *IEEE TBME*, vol. 67, 2020.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv*, 2015.

[24] D. Wu, J. Yang, and M. Sawan, "Bridging the gap between patient-specific and patient-independent seizure prediction via knowledge distillation," *J. Neural Eng.*, vol. 19, no. 3, 2022.

[25] S. Sakhavi and C. Guan, "Convolutional neural network-based transfer learning and knowledge distillation using multi-subject data in motor imagery BCI," *IEEE/EMBS NER*, pp. 588–591, 2017.

[26] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Front. Hum. Neurosci.*, vol. 15, 2021.

[27] B. Sun, Z. Wu, Y. Hu, and T. Li, "Golden subject is everyone: A subject transfer neural network for motor imagery-based brain computer interfaces," *Neural Networks*, vol. 151, pp. 111–120, 2022.

[28] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data. Eng.*, vol. 18, no. 1, pp. 63–77, 2005.

[29] S. Perdikis, M. Tavella, R. Leeb, R. Chavarriaga, and J. d. R. Millan, "A supervised recalibration protocol for unbiased bci," *5th Int. Brain-Computer Interface Conf.*, Jan. 2011.

[30] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain Topogr.*, vol. 23, no. 2, pp. 194–198, Jun. 2010.

[31] M.-H. Lee, S. Fazli, J. Mehnert, and S.-W. Lee, "Subject-dependent classification for robust idle state detection using multi-modal neuroimaging and data-fusion techniques in BCI," *Pattern Recognition*, vol. 48, no. 8, pp. 2725–2737, 2015.

[32] C. Zich, S. Debener, C. Kranczioch, M. Bleichner, M. Gutberlet, and M. de Vos, "Real-time eeg feedback during simultaneous eeg-fmri identifies the cortical signature of motor imagery," *NeuroImage*, vol. 114, Apr. 2015.

[33] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "BCI competition 2008–Graz data set A," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[34] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE TBME*, vol. 51, no. 6, pp. 1034–1043, 2004.

[35] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *Proc. Conf. AAAI Artif. Intell.*, vol. 30, no. 1, Mar. 2016.

[36] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[37] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz *et al.*, "Review of the BCI competition IV," *Front. Neurosci.*, p. 55, 2012.

[38] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[39] g.tec neurotechnology GmbH, "Unicorn Hybrid Black," https://www.unicorn-bi.com/brain-interface-technology/, 2023.

[40] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random? a closer look on bci results," *Int. J. Bioelectromagn.*, vol. 10, pp. 52–55, 01 2008.

[41] K. A. Ludwig, R. M. Miriani, N. B. Langhals, M. D. Joseph, D. J. Anderson, and D. R. Kipke, "Using a common average reference to improve cortical neuron recordings from microelectrode arrays," *J. Neurophysiol.*, vol. 101, no. 3, pp. 1679–1689, 2009.

[42] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," *IEEE/EMBS NER*, pp. 207–210, 2019.