Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS

Boyang Lyu,[a,*,†] **Thao Pham**,[b,†] **Giles Blaney**,[b] **Zachary Haga,**[c] **Angelo Sassaroli,**[b] **Sergio Fantini**,[b] **and Shuchin Aeron**[a]

[a]Tufts University, Department of Electrical and Computer Engineering, Medford, Massachusetts, United States

[b]Tufts University, Department of Biomedical Engineering, Medford, Massachusetts, United States

[c]Tufts University, Department of Computer Science, Medford, Massachusetts, United States

Abstract

Significance: We demonstrated the potential of using domain adaptation on functional near-infrared spectroscopy (fNIRS) data to classify different levels of *n*-back tasks that involve working memory.

Aim: Domain shift in fNIRS data is a challenge in the workload level alignment across different experiment sessions and subjects. To address this problem, two domain adaptation approaches—Gromov–Wasserstein (G-W) and fused Gromov–Wasserstein (FG-W) were used.

Approach: Specifically, we used labeled data from one session or one subject to classify trials in another session (within the same subject) or another subject. We applied G-W for session-by-session alignment and FG-W for subject-by-subject alignment to fNIRS data acquired during different *n*-back task levels. We compared these approaches with three supervised methods: multiclass support vector machine (SVM), convolutional neural network (CNN), and recurrent neural network (RNN).

Results: In a sample of six subjects, G-W resulted in an alignment accuracy of 68% ± 4% (weighted mean ± standard error) for session-by-session alignment, FG-W resulted in an alignment accuracy of 55% ± 2% for subject-by-subject alignment. In each of these cases, 25% accuracy represents chance. Alignment accuracy results from both G-W and FG-W are significantly greater than those from SVM, CNN, and RNN. We also showed that removal of motion artifacts from the fNIRS data plays an important role in improving alignment performance.

Conclusions: Domain adaptation has potential for session-by-session and subject-by-subject alignment of mental workload by using fNIRS data.

[DOI: 10.1117/1.JBO.26.2.022908]

Keywords: fNIRS; *n*-back task; machine learning; Gromov–Wasserstein; fused Gromov–Wasserstein; transient artifact reduction algorithm.

Paper 200209SSRR received Jul. 1, 2020; accepted for publication Nov. 25, 2020; published online Jan. 7, 2021.

1 Introduction

Functional near-infrared spectroscopy (fNIRS) is a noninvasive optical technique for monitoring regional tissue oxygenation based on diffusion and absorption of near-infrared light photons in human tissue. Continuous-wave (cw) fNIRS provides measurements of concentration changes in oxy-, deoxy-, and total-hemoglobin species ($\Delta[\mathrm{HbO_2}]$, $\Delta[\mathrm{Hb}]$, and $\Delta[\mathrm{HbT}]$, respectively) in

*Address all correspondence to Boyang Lyu, Boyang.Lyu@tufts.edu

†These authors contributed equally.

tissue with temporal sampling rate of on the order of 10 Hz.[1] Over the past three decades, fNIRS has been used in several brain imaging applications, including noninvasive imaging of cognitive tasks and brain functional activation,[1–4] and brain computer interface (BCI).[5]

Memory-based workload classification using fNIRS measurements has been demonstrated to be an ideal approach for a realistic adaptive BCI to measure human workload level.[6] In this paper, we study the problem of classification of fNIRS corresponding to different conditions of an n-back task (i.e., subjects are required to continuously remember the last $n \in \{1,2,3,\dots\}$ of rapidly changing letters or numbers). We performed fNIRS measurements on prefrontal cortex (PFC), which has been found to be a relevant area for memory-related tasks by positron emission tomography and functional magnetic resonance imaging.[7,8] Most n-back classification studies in the literature are based on supervised methods on fNIRS signals in within-session and within-subject basis (i.e., within single trial of data acquisition on a single subject).[9–11] While those studies showed promising results, subject- and session-dependent systems are not realistic for an interface system that can adapt to different users with a wide range of physiological conditions. With the aim of use in BCI, workload classifications based on fNIRS data across experiment sessions (session-by-session alignment) and across subjects (subject-by-subject alignment) are necessary.

There are several challenges that hamper accurate workload classification using fNIRS data. We outline them below and propose methods to mitigate them.

The first challenge, which is the main focus of this paper, is to deal with session-by-session and subject-by-subject variations in classification of n-back tasks. These problems are related to what is referred to as domain adaptation in machine learning.[12–14] More specifically, data from different sessions or different subjects are referred to as belonging to different domains, and the changes in data distributions across different domains (the session or subject that the data belongs to) are considered as a domain shift.[15] Due to this phenomenon, the knowledge we learned from one domain cannot be applied directly to another one. To address this problem, recent advances in the theory and methods of optimal transport (OT)[16] and metric measure space alignment[17–19] could be used to align data with a known labeled n-back condition from one session or one subject to the unlabeled data from a different session within the same subject or from a different subject. Though OT has been applied for domain adaptation with potential performance,[20,21] it has some limitations when two sets of data used for alignment do not share the same metric space in which case a meaningful notion of distance between two spaces does not exist. For example, for session-by-session alignment, data from some of the fNIRS channels are removed from one of the two sessions due to a poor signal-to-noise ratio (SNR). This will cause two sessions' data to be embedded in different dimensions in the two domains. A naïve solution is to remove the corresponding channels from the other session to guarantee that the two sessions have the same dimension. However, this has the disadvantage of causing loss of information. In this paper, we proposed that using Gromov–Wasserstein (G-W)[18,22] and fused Gromov–Wasserstein (FG-W) barycenter[23] would alleviate this problem and provide algorithms to align across domains for fNIRS n-back task classification.

The second challenge is motion artifacts in fNIRS signals. Motion artifacts in fNIRS are commonly due to the coupling changes of any source or detector from the scalp during the experiment. This causes sudden increases or decreases in measured light intensity and can affect the measured fNIRS signals. From a machine learning perspective, motion artifact detection and correction help remove any misleading correlation from the subject behavior (twitching, head movement, etc.) to what the classification model learns from fNIRS data. For example, a classification model may recognize when a subject presses a button as a requirement during the experiment by detecting spikes in the measured signals due to the subject's head movement, instead of detecting real hemodynamic responses from the brain signals. A number of approaches, inspired by statistical signal processing methods such as adaptive filtering, independent component analysis (ICA), and time-frequency analysis, have been proposed to remove or correct for motion artifacts in fNIRS signals.[24–30] Most of these techniques either depend on the use of auxiliary reference signals (e.g., accelerometry, etc.) or extraoptical channels or require certain assumptions on the characteristics of motion artifacts and cleaned fNIRS signals. In this paper, we used an off-the-shelf method based on sparse optimization for automatic detection and removal of spikes and steps anomalies, namely transient artifact reduction algorithm (TARA).[31]

We will apply the method TARA in the hope of improving the classification accuracy of n-back tasks.

The main contributions of this paper to the classification of different n-back task conditions include: (1) applying G-W to align fNIRS data during each n-back task condition across different experimental sessions for every single subject (session-by-session alignment); (2) applying FG-W barycenter to align fNIRS data during each n-back condition between different subjects (subject-by-subject alignment); and (3) demonstrating that alignment accuracy could be improved by applying motion artifact removal with TARA as a preprocessing step on fNIRS data.

2 Experiment

2.1 *Subjects and Experiment Design*

Six healthy human subjects (one female, five males; age range: 23 to 54 years) participated in this study. The Tufts University Institutional Review Board approved the experimental protocol, and the subjects provided written informed consent prior to the experiment.

During the n-back task, subjects were instructed to watch a series of rapidly flashing random one-digit numbers (from 0 to 9) shown on a computer screen placed ~50 cm in front of the subject. Subjects must continuously remember the last n numbers ($n = 0$, 1, 2, and 3) and were asked to press the space bar if the currently displayed number (target) matched the preceding n'th number. In the 0-back task, the subject pressed the space bar whenever numeral “0” appeared. With increasing n, the task difficulty is expected to increase, as the subjects must remember an increasing number of preceding digits and continuously shift the remembered sequence. The experiment was designed such that the targets appeared with 25% to 35% chance (i.e., 65% to 75% nontargets) in each task (chosen randomly). We measured the task performance by counting the number of missed targets (when the subject did not press the space bar for a target) and the number of wrong reactions (when the subject incorrectly identified a nontarget stimulus as a target).

Each subject performed a total of four separate experiment sessions in two days: two sessions per day, one in the morning shift (9 to 12 a.m.) and one in the afternoon shift (1 to 4 p.m.). The order of the n-back tasks was randomized among sessions, but the randomization order was kept the same among subjects (i.e., only four random sequences were used and each subject was shown each of the four after all of their sessions). A session started with 155 s of initial baseline with a countdown timer displayed on the screen. At the beginning of a task, an instruction was shown to inform the subject that the upcoming task was 0-, 1-, 2-, or 3-back. A task consisted of 100 displayed digits each lasting 2 s, during which the stimulus was displayed for 1.5 s and followed by a resting time of 0.5 s where a black screen was shown. Therefore, each task was a total of 200 s in length. Subsequently, the subject entered 30 s of baseline (rest) after finishing the task while the performance accuracy of the preceding task was displayed on the screen. This process was repeated for the four values of n. At the end of the fourth task, the subject rested for a 155-s baseline after which the experiment was completed. Figure 1(a) shows the experiment protocol. The entire experiment had a recording time of 20 min (four 200-s tasks, two 155-s baselines, and three 30-s rests in the middle).

2.2 *Data Acquisition*

During the entire experiment session, optical data were collected continuously with a cw fNIRS device (NIRScout, NIRx Medical Technology, Berlin, Germany). Eight light-emitting diode source pairs (at two wavelengths of 760 and 850 nm) and seven detector fiber bundles connected to photodiode detectors were arranged on a conformable fabric headset. The fNIRS headset can be quickly fixed to the forehead to enable high quality measurements of the PFC within the range of several minutes. A total of 20 channels at 3-cm source–detector distances were collected. A schematic diagram of the arrangement is shown in Figs. 1(b) and 1(c). Light intensities were collected at a sampling rate of 7.81 Hz. Linear detrending was applied to the collected changes in light intensity with respect to baseline to remove slow temporal drifts. Then, the detrended

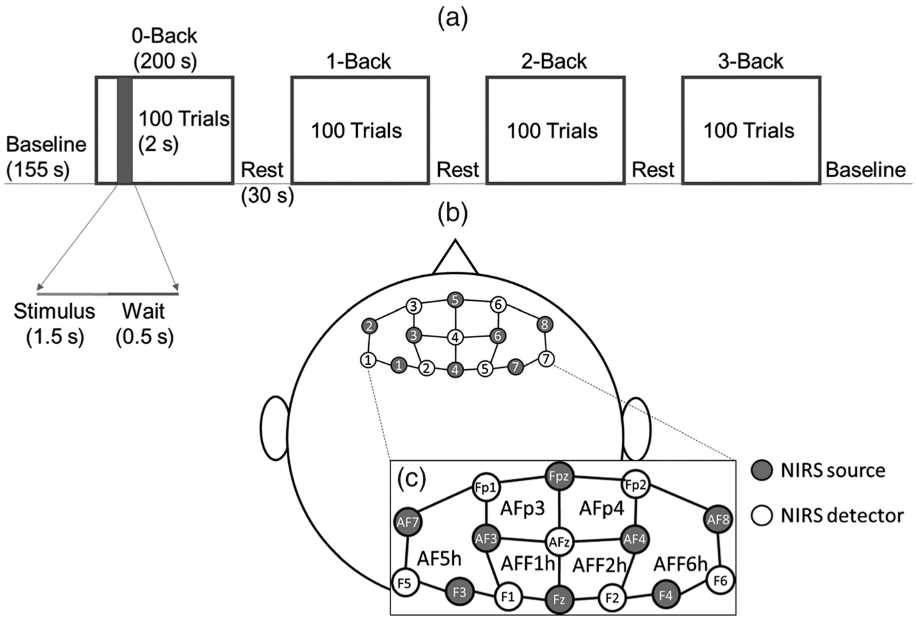


Fig. 1 (a) Experimental design for *n*-back tasks. (b) fNIRS headset with eight sources and seven detectors to give a total of 20 channels at source–detector distance of 3 cm. (c) An enlarged view of the schematic in (b) showing positions of the 10-10 system (Fp1, Fpz, Fp2, AF7, AF3, AFz, AF4, AF8, F5, F3, F1, Fz, F2, F4, and F6) and the 10-5 system (AFp3, AFp4, AF5h, AFF1h, AFF2h, and AFF6h) covered by the sources and detectors.

normalized intensities were converted into $\Delta[\mathrm{HbO}_2]$ and $\Delta[\mathrm{Hb}]$ using the modified Beer–Lambert law.[32] We assumed the wavelength-dependent differential pathlength factors (DPFs), which account for the increase in photon pathlength due to multiple scattering, equal to 9.1 and 8.0 for 760 and 850 nm, respectively.[33]

During the experiment, continuous arterial blood pressure (ABP) was collected with a beat-to-beat finger plethysmography system (NIBP100D, BIOPAC Systems, Inc., Goleta, California). ABP measurements were converted into mean arterial blood pressure (MAP, in units of mmHg) and heart rate (HR, in units of beats per minute, bpm).

2.3 *fNIRS Data Preprocessing by TARA*

Measured fNIRS data were checked manually to remove those noisy channels contaminated by high frequency noise (>1 Hz). Examples of removed and retained channels from two subjects are shown in Fig. 2, and the numbers of remaining channels are reported in Table 4 in the Appendix. The whole session will be removed if more than 60% of channels are identified as noisy. To further remove motion artifacts from the retained channels, we used the TARA algorithm[31] in which measured time series data are treated as a linear combination of a low-pass signal, motion artifacts, and white noise. The algorithm focuses on two types of motion artifacts: transient pulses (spike-like signals) and step discontinuities, and assumes both of them appear infrequently. A sparse optimization problem is then formulated to jointly estimate two types of motion artifacts. We refer the reader to Ref. 31 for more details. We used the code provided by the authors[34] and chose parameters for our fNIRS data as shown in Table 5 in the Appendix. Once the motion artifacts are detected, they can be removed from the original signal to obtain the cleaned data.

3 Domain Adaptation for fNIRS

After the removal of the channels with poor SNR and motion artifacts, a small time duration w is chosen as the window size to divide the remaining n-back data ($\Delta[\mathrm{HbO}_2]$ and $\Delta[\mathrm{Hb}]$) into

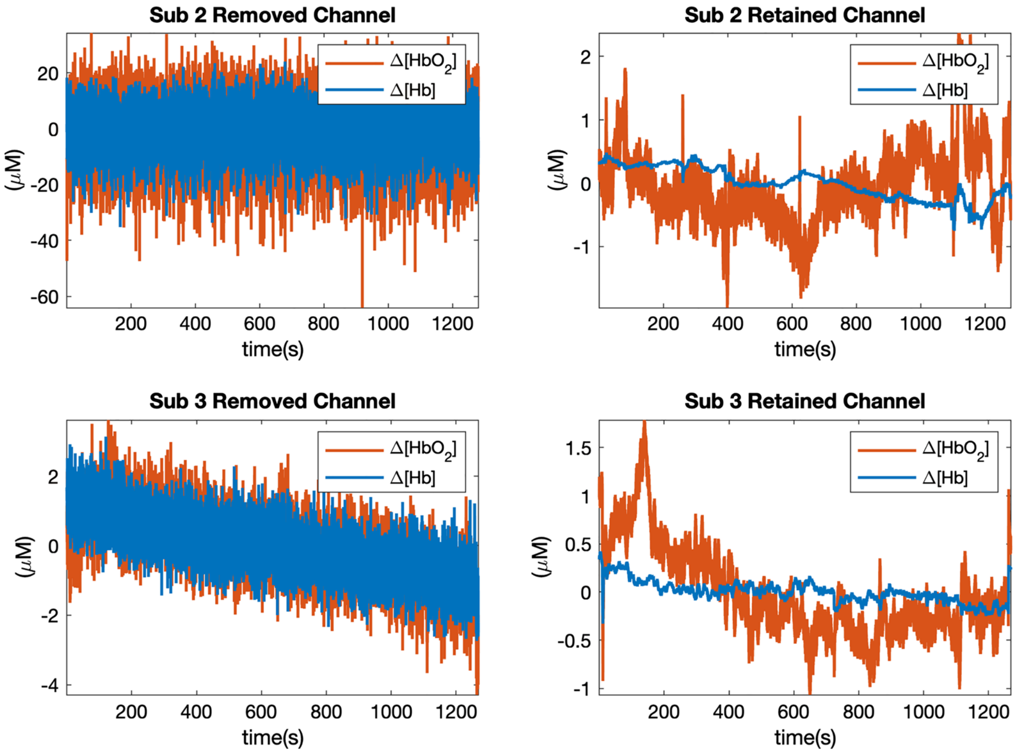


Fig. 2 Examples of removed and retained channels from two subjects (2 and 3). The first column shows the removed channels, the second column shows the retained channels. Time courses are shown for concentration changes in oxy-(Δ[HbO$_2$], shown in orange) and deoxy-hemoglobin (Δ[Hb], shown in blue).

N nonoverlapping small segments. Here, we use $w = 60$ samples (∼8 s). To concretely describe the proposed method, next we will set some notations that are used throughout the paper.

Notation: We will use lower-case boldface letters $\mathbf{x}$ to denote vectors and upper case boldface letters $\mathbf{X}$ to denote matrices. Unless otherwise stated, unbolded lower case letters denote scalars. $\{(\mathbf{X}^s_{m,i}, y^s_{m,i})\}^N_{i=1}$ stands for the collection of segmented data set of subject s in its m_{th} session, where N is the number of segments, integer $s \in [1,6]$, and integer $m \in [1,4]$. The i'th segment is denoted as $\mathbf{X}^s_{m,i} \in \mathbb{R}^{d\times w}$, where d is the number of channels and w is the window length. $y^s_{m,i} \in [0,3]$ is the corresponding n-back task label for subject s in session m and segment i, $\mathbf{y}^s_m = \mathbf{vec}(y^s_{m,i})$ is an N-dimensional vector of the label. The remaining notation will be introduced as needed.

3.1 *Session-by-Session Alignment*

3.1.1 *Optimal transport theory and Gromov–Wasserstein matching*

Consider two discrete sets of points $\{\mathbf{x}_i\}_{i\in 1\dots n}$, $\mathbf{x}_i \in \mathbb{R}^d$ in a metric space $\mathcal{X}$ with a metric $d_{\mathcal{X}}$, and $\{\mathbf{y}_j\}_{j\in 1\dots m}$, $\mathbf{y}_j \in \mathbb{R}^d$ in another metric space $\mathcal{Y}$ with the metric $d_{\mathcal{Y}}$. The main idea behind aligning two sets of points is by viewing them as two empirical distributions

$$\mathbf{a} = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{b} = \sum_{j=1}^{m} b_j \delta_{\mathbf{y}_j}, \tag{1}$$

where $\delta_{\mathbf{x}_i}$ and $\delta_{\mathbf{y}_j}$ are Dirac functions at the position of $\mathbf{x}_i$ and $\mathbf{y}_j$ and a_i and b_j are the corresponding probabilities. Without further information, a_i and b_j will be set as $\frac{1}{n}$ and $\frac{1}{m}$, respectively. The OT problem is proposed to find a plan $\mathbf{T} \in \mathbb{R}^{n\times m}$ that is the solution to

$$\arg \min_{\mathbf{T}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{T} \rangle, \tag{2}$$

where $\langle \mathbf{C}, \mathbf{T} \rangle = \sum_{i,j} \mathbf{C}_{i,j} \mathbf{T}_{i,j}$, $U(\mathbf{a},\mathbf{b}) = \{\mathbf{T} \in \mathbb{R}_+^{n\times m} : \sum_{j=1}^m \mathbf{T}_{i,j} = \mathbf{a}, \sum_{i=1}^n \mathbf{T}_{i,j} = \mathbf{b}\}$, $\mathbf{C} \in \mathbb{R}^{n\times m}$ with the i, j'th element $\mathbf{C}_{i,j}$ being the cost of associating (moving) the point $\mathbf{x}_i$ to the point $\mathbf{y}_j$. This is also known as the Kantorovich's relaxation[35] for the original Monge problem.[36] To reduce the computational cost of solving the linear program Eq. (2), an entropic regularization term is usually added to Eq. (2), leading to

$$\min_{\mathbf{T}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{T} \rangle - \lambda H(\mathbf{T}), \tag{3}$$

where $H(\mathbf{T}) = -\sum_{i,j} \mathbf{T}_{i,j}(\log \mathbf{T}_{i,j} - 1)$. This entropic OT problem[37] can be solved efficiently using the Sinkhorn Algorithm[38] or its variations such as the Greenkhorn algorithm,[39] both of which can achieve a near-linear time complexity.[40] This approach has been used in domain adaptation[20,21] for transfer of data in different domains.

Though widely used for domain adaptation, classic OT lacks the ability of mapping two different metric spaces. When the points have different dimensions, i.e., $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{y}_j \in \mathbb{R}^{d_2}$, where $d_1 \neq d_2$, a distance between $\mathbf{x}_i$ and $\mathbf{y}_j$ may not be meaningfully defined. Thus, instead of seeking a distance matrix between elements in different domains, the G-W method compares the dissimilarity between the pairwise distances in each domain. It poses a weaker assumption that if $\mathbf{x}_i$ should be aligned to $\mathbf{y}_j$ and $\mathbf{x}_{i'}$ should be aligned to $\mathbf{y}_{j'}$, then for two distance matrices $\mathbf{C}^{\mathcal{X}} \in \mathbb{R}^{n\times n}$ and $\mathbf{C}^{\mathcal{Y}} \in \mathbb{R}^{m\times m}$ in space $\mathcal{X}$ and $\mathcal{Y}$, $\mathbf{C}^{\mathcal{X}}_{i,i'}$ and $\mathbf{C}^{\mathcal{Y}}_{j,j'}$ should be similar.[17] Formally, the G-W distance is defined as

$$\mathrm{GW}[(\mathbf{a}, \mathbf{C}^{\mathcal{X}}), (\mathbf{b}, \mathbf{C}^{\mathcal{Y}})] = \min_{\mathbf{T}\in U(\mathbf{a},\mathbf{b})} \sum_{i,i',j,j'} L(\mathbf{C}^{\mathcal{X}}_{i,i'}, \mathbf{C}^{\mathcal{Y}}_{j,j'}) \mathbf{T}_{i,j} \mathbf{T}_{i',j'}, \tag{4}$$

where L is a cost function, which typically can be chosen as a quadratic function or Kullback–Leibler divergence. For our method, a squared loss function is applied. Equation (4) is a non-convex problem related to a quadratic assignment problem.[18] A regularized version of the GW problem is proposed in Ref. 17, written as

$$\mathrm{GW}_\lambda[(\mathbf{a}, \mathbf{C}^{\mathcal{X}}), (\mathbf{b}, \mathbf{C}^{\mathcal{Y}})] = \min_{\mathbf{T}\in U(\mathbf{a},\mathbf{b})} \sum_{i,i',j,j'} L(\mathbf{C}^{\mathcal{X}}_{i,i'}, \mathbf{C}^{\mathcal{Y}}_{j,j'}) \mathbf{T}_{i,j} \mathbf{T}_{i',j'} - \lambda H(\mathbf{T}). \tag{5}$$

The problem in Eq. (5) can be solved by projected gradient descent algorithm wherein each iteration solution is found by running Sinkhorn Algorithm.[22]

3.1.2 *Metric for G-W alignment for fNIRS data*

For electroencephalogram (EEG) and fNIRS processing, mean and covariance of the time segments have been considered as useful features.[41,42] Here, we use these features to compute the inner metric matrix of each session. Specifically, for data $\{\mathbf{X}^s_{m,i}\}_{i=1}^N$ from the m'th session of subject s, we compute its covariance matrices $\{\mathbf{P}^s_{m,i}\}_{i=1}^N$ and mean vectors $\{\mathbf{h}^s_{m,i}\}_{i=1}^N$, where $\mathbf{P}^s_{m,i} \in \mathbb{R}^{d\times d}$, $\mathbf{h}^s_{m,i} \in \mathbb{R}^d$. The distance matrix $\mathbf{C}^s_m \in \mathbb{R}^{N\times N}$ is then defined with the i, i''th element $(\boldsymbol{C}^s_m)_{ii'}$ set as

$$(\mathbf{C}^s_m)_{ii'} = (\rho_{\text{hellinger}}(\mathbf{P}^s_{m,i}, \mathbf{P}^s_{m,i'}) + \|\mathbf{h}^s_{m,i} - \mathbf{h}^s_{m,i'}\|_2)/d, \tag{6}$$

where $\rho_{\text{hellinger}}(\cdot)$ is the matrix version of the Hellinger distance,[43] written as

$$\rho_{\text{hellinger}}(\mathbf{A}, \mathbf{B}) = \{tr(\mathbf{A} + \mathbf{B}) - 2tr[\mathbf{A}^{1/2}(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})^{1/2}\mathbf{A}^{1/2}]\}^{1/2}, \tag{7}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite matrices. Since the number of channels d selected for different sessions' data are not necessarily the same, we normalize by the number of channels in each session.

Algorithm 1 Alignment between session *m* and session *n*.

```
Input: Source data and label $\{(\mathbf{X}_{m,i}, y_{m,i})\}_{i=1}^{N}$, target data $\{\mathbf{X}_{n,i}\}_{i=1}^{N}$
Output: Target label $\{y_{n,i}\}_{i=1}^{N}$
Calculate inner distance matrices $\mathbf{C}_m$ and $\mathbf{C}_n$ using Eq. (6) for $\{\mathbf{X}_{m,i}\}_{i=1}^{N}$ and $\{\mathbf{X}_{n,i}\}_{i=1}^{N}$.
Solve Eq. (5) to get the transport plan $\mathbf{T}$ between session $m$ and session $n$.
Choose the largest value of each column of $\mathbf{T}$ as 1 and set others to be 0 to get the coupling matrix $\mathbf{T}_{cp}$
Get target label $\{y_{n,i}\}_{i=1}^{N}$ by calculating $\mathbf{T}_{cp}^{\mathsf{T}}\mathbf{vec}(y_{m,i})$
```

3.1.3 *Domain adaptation for session-by-session alignment*

We assume the label is given for one session's data and aim to infer the label for all other sessions belonging to the same subject. Using the metric defined in Eq. (6), we show the pseudocode for the session-by-session alignment in Algorithm 1. Since we only consider data within the same subject, the upper index for the subject will be dropped in the algorithm.

3.2 *Subject-by-Subject Alignment*

When targeting subject-by-subject alignment, we assume data and the corresponding labels for all sessions of one subject are given and denote this subject as the source subject. Then, we will use this information to predict the labels of fNIRS data for all four sessions of other subjects (target subjects). Transferring labels between different subjects is a bigger challenge since there is a larger shift in domain. Directly using the same G-W alignment as discussed above will lead to a large variance in alignment accuracy. More importantly, we will lose the advantage of knowing all the features and structural information from multiple sessions of the source subject. To address this problem, we consider a recently proposed method named FG-W.[23] By computing an FG-W barycenter, which is the Fréchet mean of the FG-W distance, we summarize all the given information into a new representation of the source subject and then follow the same routine as session-by-session alignment to achieve the label alignment.

3.2.1 *Fused Gromov–Wasserstein barycenter*

FG-W, unlike the G-W, combines both feature and structural information and shows its advantage in graph classification.[23,44] Consider two sets of tuples $\{(\mathbf{x}_i, \mathbf{f}_i)\}_{i\in 1\cdots n}$ in space $(\mathcal{X}, \Sigma)$ and $\{(\mathbf{y}_j, \mathbf{g}_j)\}_{j\in 1\cdots m}$, in space $(\mathcal{Y}, \Sigma)$, here, $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{y}_j \in \mathbb{R}^{d_2}$ are the data points, $\mathbf{f}_i$ and $\mathbf{g}_j$ are their corresponding features, which are both in space Σ and share the same dimension. With a slight abuse of notation, we will use the same symbol as Eq. (1) to denote their empirical distribution

$$\mathbf{a} = \sum_{i=1}^{n} a_i \delta_{(\mathbf{x}_i, \mathbf{f}_i)}, \quad \mathbf{b} = \sum_{j=1}^{m} b_j \delta_{(\mathbf{y}_j, \mathbf{g}_j)}. \tag{8}$$

The FG-W distance between such two distributions with both data and the corresponding feature information included is then defined as

$$\mathrm{FGW}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{T}\in U(\mathbf{a},\mathbf{b})} \sum_{i,i',j,j'} [(1-\alpha)\rho(\mathbf{f}_i, \mathbf{g}_j)^q + \alpha|\mathbf{C}^{\mathcal{X}}_{i,i'} - \mathbf{C}^{\mathcal{Y}}_{j,j'}|^q]\mathbf{T}_{i,j}\mathbf{T}_{i',j'}, \tag{9}$$

where $\alpha \in [0,1]$ is a trade-off parameter, $q \geq 1$, $\rho(\mathbf{f}_i, \mathbf{g}_j)$ stands for the cost of matching feature $\mathbf{f}_i$ to feature $\mathbf{g}_j$ which in our case corresponds to the labels, i.e., scalar value n in the n-back task.

For multiple distribution setting like those related to multiple sessions, a natural extension of FG-W distance is its barycenter, which inherits the advantages of FG-W that leverages both structural and feature information. The FG-W barycenter can be obtained by minimizing the weighted sum of a set of FG-W distances. Let $\{\mathbf{C}^k\}_{k=1}^K$ be a set of distance matrices, where $\mathbf{C}^k \in \mathbb{R}^{N\times N}$, $\{\mathbf{f^k}\}_{k=1}^K$, $\mathbf{f}^k \in \mathbb{R}^{\mathbb{N}}$ is the corresponding feature vector. Here, K will correspond to the number of sessions for each subject in our case. We assume the base histograms $\{\mathbf{a}^k\}_{k=1}^K$ and the histogram $\mathbf{a}$ associated with the barycenter are known and fixed as uniform distributions. By calculating the Fréchet mean of the FG-W distance, we aim to find a feature vector $\mathbf{f}$ and a distance matrix $\mathbf{C}$ that represents the structure information, such that

$$\min_{\mathbf{C}\in\mathbb{R}^{N\times N},\mathbf{f}\in\mathbb{R}^N,(\mathbf{T}^k)_k\in U(\mathbf{a},\mathbf{a}^k)} \sum_k \sum_{i,i',j,j'} \zeta_k[(1-\alpha)\rho(\mathbf{f}_i,\mathbf{f}_j^k)^q + \alpha|\mathbf{C}_{i,i'} - \mathbf{C}_{j,j'}^k|^q]\mathbf{T}_{i,j}^k\mathbf{T}_{i',j'}^k, \tag{10}$$

where $\sum_k \zeta_k = 1$ are the weights for sessions and are chosen evenly for each session. $q \geq 1$ for the loss of features and a squared loss between features is used for our method. This problem can be solved by block coordinate descent algorithm described in Ref. 23. Note that after solving Eq. (10), only the distance matrix $\mathbf{C}$ and feature vector $\mathbf{f}$ will be used to form the new representation of the provided K sessions.

3.2.2 *Metric for FG-W barycenter alignment*

Unlike the metric defined in Eq. (6) for session-by-session alignment, we removed the L2 norm of the mean difference from the distance when considering the metric for subject-by-subject alignment. This is because the differences of the mean values are usually the same within the same subject but vary across different subjects. It is worth mentioning that after removing the L2 norm of the mean difference, the covariance matrices themselves can be viewed as points in a Riemannian space.[45] Formally, for the m'th session of subject s, the distance matrix $\mathbf{C}_m^s$ is defined using its covariance matrices $\{\mathbf{P}_{m,i}^s\}_{i=1}^N$, $\mathbf{P}_{m,i}^s \in \mathbb{R}^{d\times d}$, with the i, i''th element $(\mathbf{C}_m^s)_{ii'}$ computed via

$$(\mathbf{C}_m^s)_{ii'} = [\rho_{\text{hellinger}}(\mathbf{P}_{m,i}^s, \mathbf{P}_{m,i'}^s)]/d. \tag{11}$$

3.2.3 *Domain adaptation for subject-by-subject alignment*

The algorithm for subject-by-subject alignment is shown in Algorithm 2, where we only take two subjects (each with four sessions) as an example, but the algorithm can be easily adapted to all other subjects with different number of sessions.

3.3 *Comparison with Supervised Machine Learning Methods*

To further demonstrate the potential of domain adaptation methods, we compared our method with a convolutional neural network (CNN), a recurrent neural network (RNN), and a multi-

Algorithm 2 Alignment between subject *s* and subject *t*.

```
Input: Source data and label $\{(\mathbf{X}^s_{\{1...4\},i}, y^s_{\{1...4\},i})\}_{i=1}^N$, target data $\{\mathbf{X}^t_{\{1...4\},i}\}_{i=1}^N$
Output: Target label $\{y^t_{\{1...4\},i}\}_{i=1}^N$
1: For source and target data, calculate two lists of distance matrices $[\mathbf{C}_1^s, \mathbf{C}_2^s, \mathbf{C}_3^s, \mathbf{C}_4^s]$ and $[\mathbf{C}_1^t, \mathbf{C}_2^t, \mathbf{C}_3^t, \mathbf{C}_4^t]$, respectively, using Eq. (11).
2: Solve Eq. (10) using $[(\mathbf{C}_1^s, \mathbf{y}_1^s), (\mathbf{C}_2^s, \mathbf{y}_2^s), (\mathbf{C}_3^s, \mathbf{y}_3^s), (\mathbf{C}_4^s, \mathbf{y}_4^s)]$ to get the inner distance matrix and corresponding label vector of the barycenter for subject s, denoted as $\{\mathbf{C}^s_{\text{bary}}, \mathbf{y}^s_{\text{bary}}\}$.
3: Repeat steps 2 to 4 in Algorithm 1 using $\{\mathbf{C}^s_{\text{bary}}, \mathbf{y}^s_{\text{bary}}\}$ and $\mathbf{C}_1^t$, $\mathbf{C}_2^t$, $\mathbf{C}_3^t$, $\mathbf{C}_4^t$, respectively, as input to get the labels $\{y^t_{\{1...4\},i}\}_{i=1}^N$ for target data.
```

class support vector machine (SVM)-based classifiers applied without any domain adaptation techniques.

For the CNN model, we adapted the architecture structure of EEG-NET.[46] Due to paucity of the amount of data in our case compared to the original paper,[46] we simplified the structure to three convolutional layers followed by two dense layers. Details of the CNN structure can be found in Appendix, Table 3. To compare with the session-by-session alignment using the G-W method, $\Delta[HbO_2]$ and $\Delta[Hb]$ data were first separated and then stacked along a new dimension as input to the CNN. Since the removal of some noisy channels will lead to different input data points being in different dimensions, thereby causing a mismatch between input data and the fixed model structure, we replaced the discarded channels with the average of data from the remaining channels (separately for $\Delta[HbO_2]$ and $\Delta[Hb]$). We used data from one session as input to train the model with Adam optimizer[47] using cross entropy loss and tested the remaining sessions. The model was trained until severe overfitting occurred (300 epochs in our case) to guarantee the best test accuracy can be achieved within the training process. Test accuracy was recorded during the whole training process (i.e., after each training epoch) and the best result was selected among them. The training and testing processes were conducted five times and the average test accuracy was reported. To compare with subject-by-subject alignment using the FG-W method, data from all four sessions of one subject were combined and used as input and the classification model was trained to predict the task labels for the other five subjects in the same manner as discussed above.

For the RNN model, we used a basic three-layer long short term memory[48] model with the hidden size set as 20. The training and testing data were prepared in the same way as the CNN method except that $\Delta[HbO_2]$ and $\Delta[Hb]$ data were not separated, but were input together. For both session and subject prediction, the training and evaluation procedure followed the same routine as CNN.

Before applying SVM, a dimension reduction technique was applied to the segmented multichannel fNIRS data. Here, we used uniform manifold approximation and projection (UMAP)[49] to compress each piece of segmented data[50] into a 50-dimensional vector, with the distance matrix calculated using Eq. (6) for session-by-session alignment and Eq. (11) for subject-by-subject alignment. During the training procedure for session-by-session alignment, only one session's data was used for testing and all the remaining data were used for training. Hyperparameters were selected by leave-one-session-out cross-validation. This was similar for subject-by-subject alignment, where one subject's data (including all the sessions) was used for testing while all other data were used for training. Hyperparameters were again selected by leave-one-subject-out cross-validation.

The Student's t-test was used to investigate differences between alignment accuracy from G-W/FG-W methods with 25% chance levels (25% stands for the chance to assign any session as 0, 1, 2, or 3-back), between G-W/FG-W methods using raw and cleaned data from TARA, and between G-W/FG-W and comparison methods stated above. All values are reported as mean ± standard error weighted by the standard deviations of the alignment accuracy values from six subjects unless otherwise noted.

4 Results

4.1 *Subject Performance*

Figure 3 shows summary of subject performance analysis with the average percentages of wrong and missed responses, respectively, across four sessions and six subjects for each n-back task condition. The difficulty level, in terms of the amounts of wrong and missed responses, increases significantly for the 3-back task as compared to other n-back tasks ($p < 0.05$, paired-sample t-test). Next, the numbers of wrong and missed responses for the 2-back task in the four experiment sessions are significantly higher than those for the 1- and the 0-back tasks ($p < 0.05$, paired-sample t-test). Finally, there was no significant difference in the difficulty level between the 0- and the 1-back task in terms of wrong and missed responses ($p > 0.05$, paired-sample t-test).

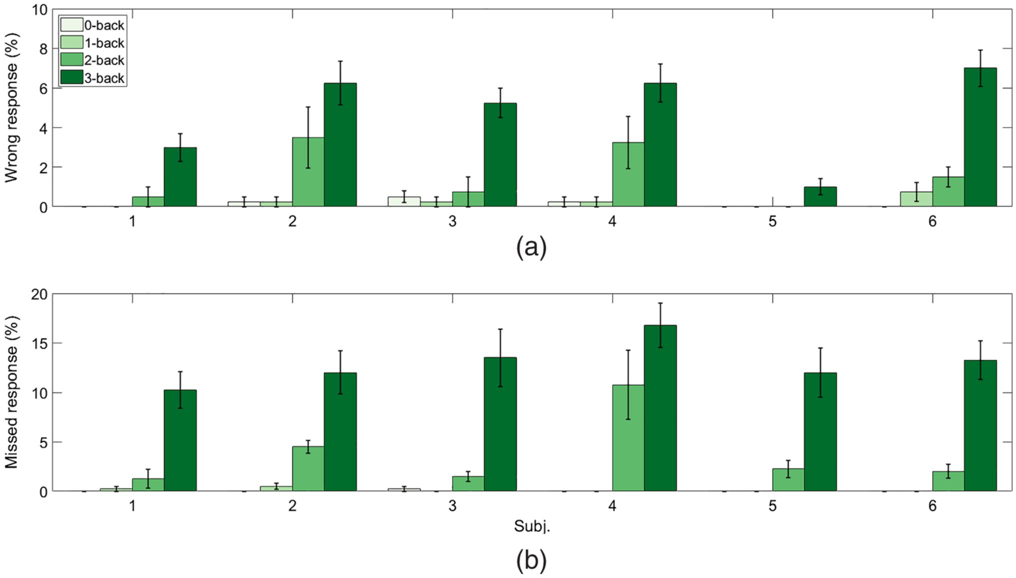


Fig. 3 Summary of subject performance for the *n*-back task: average percentages of (a) wrong responses and (b) missed responses for *n*-back task conditions across subjects. Bars represent the means, and error bars represent standard errors across four experimental sessions.

4.2 *Peripheral Physiological Measurements*

Figure 4 shows the examples of average time courses of changes in MAP and HR from three subjects (1, 2, and 4) across different measurement sessions of the 2-back task. We observe a greater variability in task-evoked changes in HR and MAP across subjects than across sessions. In particular, subjects 1 and 2 show negligible changes in MAP and HR during the task with respect to the initial baseline, with individual measurements from different sessions following the same trend. On the other hand, all the measurements from different sessions from subject 4 show

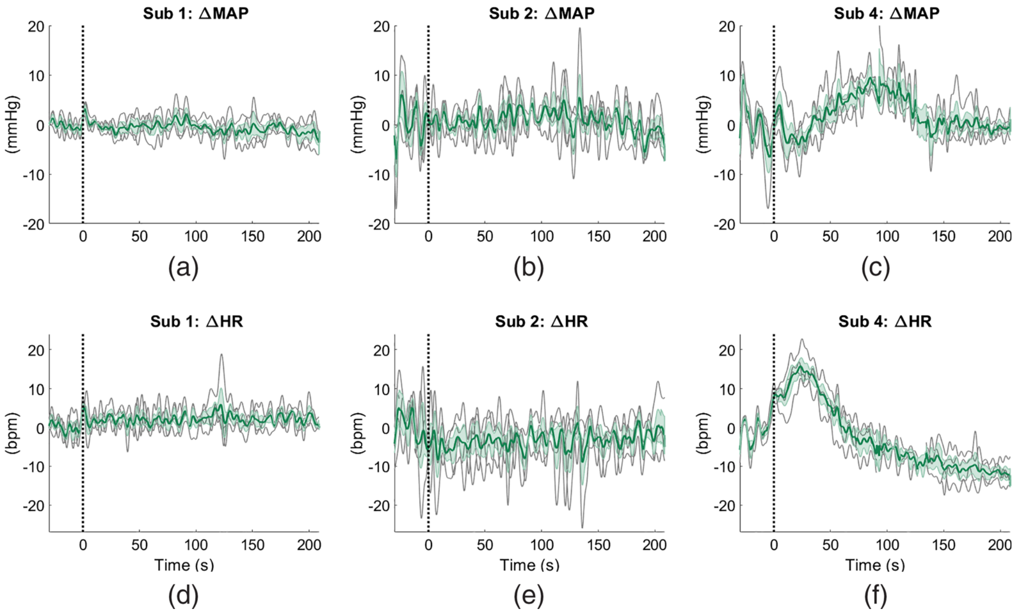


Fig. 4 Average changes in HR and MAP across all sections of 2-back task for three subjects (1, 2, and 4). The time traces are shown starting from 30 s before the task. Black dotted lines indicate time $t = 0$ s. Solid green lines are the averages across sessions in ΔMAP and ΔHR; standard errors of these averages are shown by the cyan shaded regions; solid gray lines depict the individual measurements.

totally different responses as compared to subjects 1 and 2. For subject 4, MAP increases during the middle of the task and then returns to the baseline in the last minute. The HR measurements from this subject feature an initial increase and immediate decrease at the onset of the task.

4.3 *Effects of Motion Artifact Removal Using TARA*

Figure 5 shows the effects of TARA in removing motion artifacts in the fNIRS signals (Δ[Hb]). As shown in the figure, the original signal is contaminated by the motion artifacts with spikes and steps. After applying the TARA algorithm, most of the motion artifacts have been removed. As compared to applying a low-pass filter to the original signal, TARA does not bring any further distortion to the cleaned signal and is more effective at removing step artifacts. The effect of this motion artifacts' removal as a preprocessing step before applying alignment algorithms is also shown in Table 1 and Fig. 8. An improvement for session-by-session alignment accuracy (by an average of $3 \pm 3\%$ across six subjects; $p < 0.005$, paired-sample t-test) and subject-by-subject alignment accuracy (by an average of $5 \pm 2\%$ across six subjects; $p < 0.0005$, paired-sample t-test) can be seen after applying TARA on fNIRS signals.

4.4 *Session-by-Session Alignment*

A low-dimensional UMAP visualization of the alignment for two sessions' data is shown in Fig. 6 for subject 4. In Fig. 6, the low-dimensional projection was generated individually from

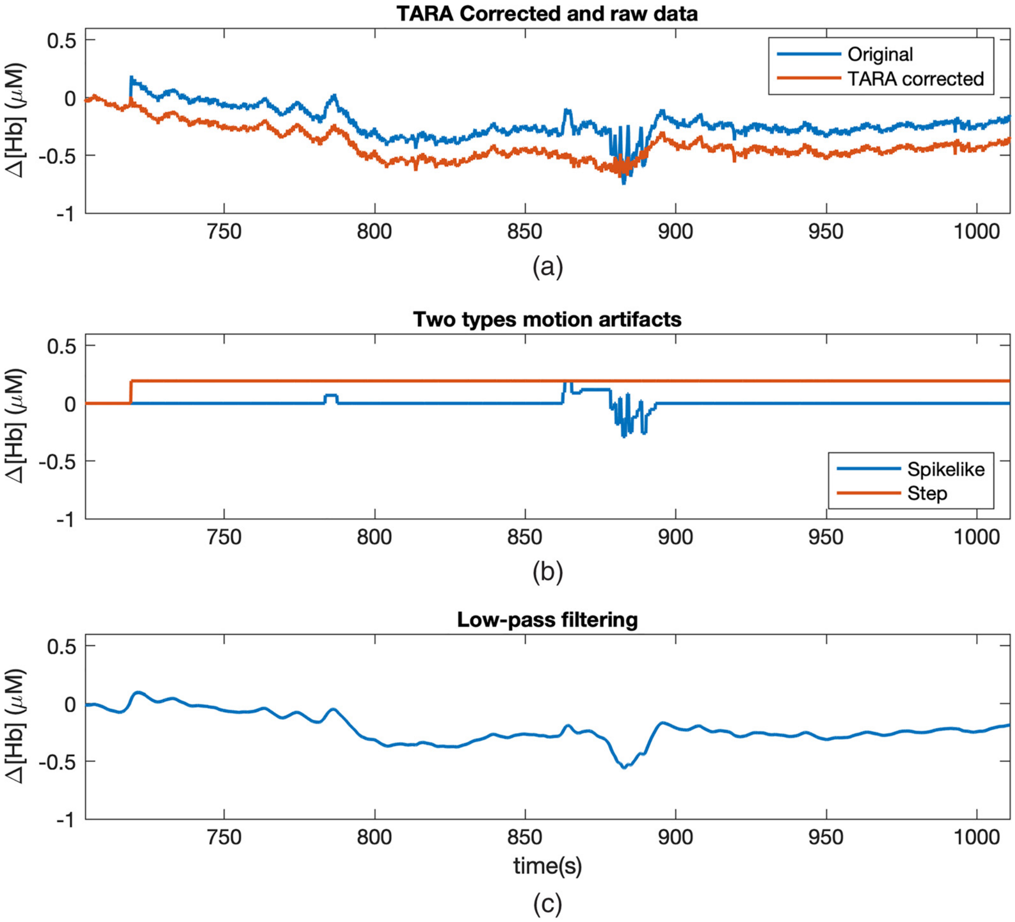


Fig. 5 Effect of using TARA on Δ[Hb] signal. (a) The original and cleaned signals. (b) The detected motion artifacts, including spike- and step-like features. The cleaned signal is obtained by subtracting these motion artifact features from the original signal. (c) The low-pass filtered signal of the original data. Distortion in the signal arising from the step discontinuity can still be observed from the low-pass filtered signal.

Table 1 Average session-by-session and subject-by-subject alignment accuracy (%) using G-W and FG-W, respectively, as compared with SVM, CNN, and RNN. G-W and FG-W barycenter alignment methods applied to both original data (org) and data cleaned by TARA algorithm, using data including and excluding data from subject 3. For other methods, only cleaned data were used as input. Averages and standard errors across all subjects are reported (Avg.).

		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Avg.
Sess-by-sess	SVM	25	31	34	38	23	20	29 ± 3**
	CNN	60 ± 7	37 ± 11	40 ± 12	58 ± 12	70 ± 7	52 ± 6	56 ± 4**
	RNN	68 ± 10	40 ± 11	43 ± 13	61 ± 14	70 ± 8	53 ± 11	58 ± 5*
	G-W org	54 ± 17	58 ± 12	48 ± 23	76 ± 12	75 ± 7	49 ± 26	68 ± 4*
	G-W	54 ± 8	63 ± 16	63 ± 21	76 ± 12	77 ± 6	62 ± 15	68 ± 4
Sub-by-sub	SVM	25	39	30	36	47	43	37 ± 3**
	CNN	45 ± 8	45 ± 10	50 ± 10	36 ± 10	38 ± 10	47 ± 11	44 ± 5**
	RNN	42 ± 6	49 ± 9	50 ± 10	41 ± 11	37 ± 9	49 ± 10	44 ± 2**
	FG-W org	51 ± 12	57 ± 10	48 ± 6	60 ± 18	57 ± 10	60 ± 16	53 ± 2**
	FG-W	55 ± 12	71 ± 16	50 ± 4	68 ± 18	67 ± 9	67 ± 16	55 ± 2
	FG-W w/o Sub 3	60 ± 3	77 ± 9	N/A	75 ± 9	69 ± 7	74 ± 9	64 ± 3

*$p < 0.005$ compared to G-W or FG-W.
**$p < 0.0005$ compared to G-W or FG-W.

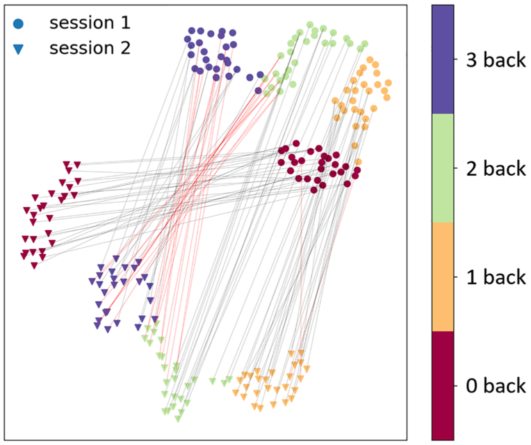


Fig. 6 Visualization of the alignment from session 1 to session 2 for subject 4. Circles indicate data from session 1 and triangles indicate the data from session 2. Four different colors represent 0- to 3-back experiments. Black lines indicate correct alignment and red lines indicate misalignment.

the distance matrix of each session's data. Therefore, the positions of the two groups of sessions' data are assigned randomly and their relative distances are not their true distances.

Figure 7 shows the confusion matrices of session-by-session alignment for four n-back tasks (0, 1, 2, and 3) of six subjects. Numbers reported in the confusion matrix are the average alignment accuracies of all the possible combinations of two out of all four sessions for each subject. Values in the main diagonal of each confusion matrix represent correct alignment between predicted and true labels, while the other values represent the misalignment results. Correct alignment results are significantly greater than chance level of 25% ($p < 0.0001$, one-sample t-test).

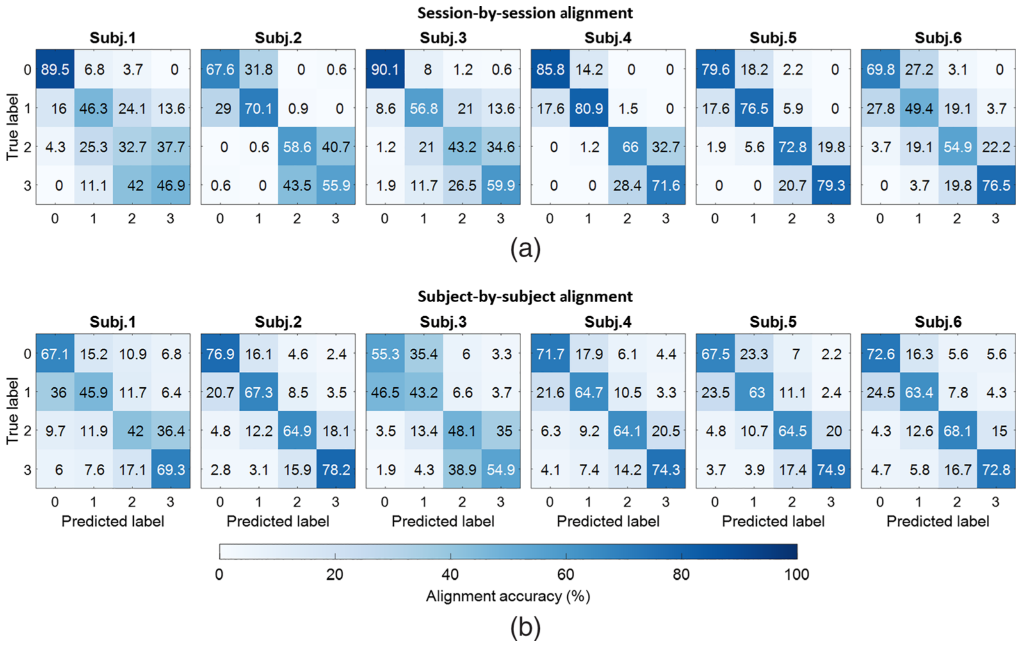


Fig. 7 Confusion matrices of session-by-session and subject-by-subject alignments in six subjects. (a) Session-by-session alignment accuracy within each subject. Each number reported in each confusion matrix is the average accuracy from the alignment of every two separate sessions among four sessions. (b) Subject-by-subject alignment accuracy. Each number reported is the average accuracy from the alignment between one source subject to the other five target subjects.

The averages and standard deviations of session-by-session alignment accuracy for six subjects are summarized in Table 1 and Fig. 8. For each subject, the value reported is calculated based on the alignment or prediction accuracy for all possible combination of session pairs. As compared to SVM, CNN, and RNN, alignment accuracy of G-W is greater by an average of 43% ± 5%, 7% ± 4%, and 5% ± 5%, respectively ($p < 0.005$, paired-sample t-test).

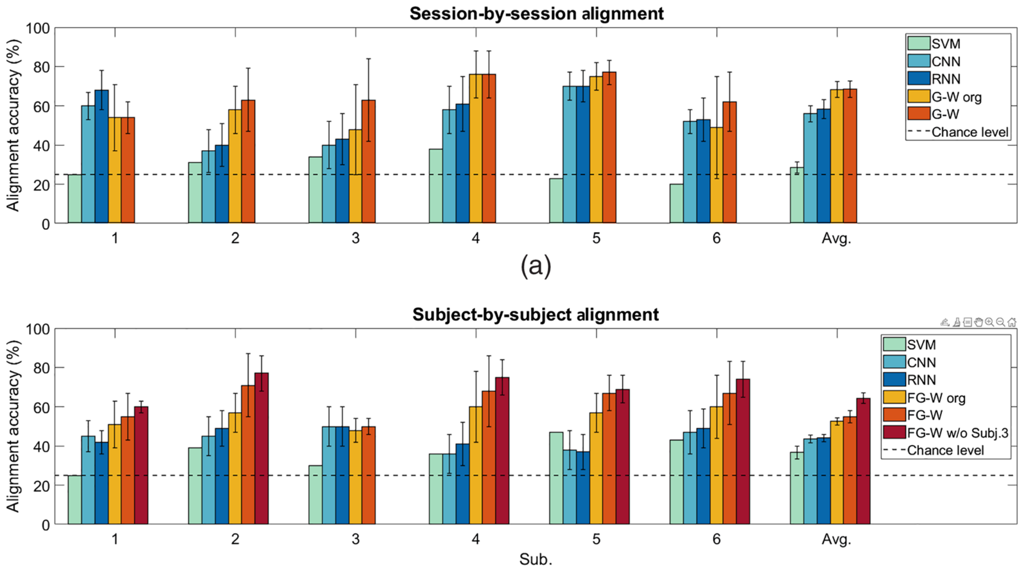
4.5 *Subject-by-Subject Alignment*

Figure 7 shows the confusion matrices of subject-by-subject alignment for four n-back tasks of six subjects. Each number in the reported confusion matrix is the average of alignment accuracies of different tasks from the source subject to five other subjects as the targets. Correct alignment results are significantly greater than chance level of 25% ($p < 0.0001$, one-sample t-test). Average subject-by-subject alignment accuracy is shown in Table 1 and Fig. 8. Each reported average accuracy value is the average of the alignment accuracy when considering one subject as the source and five other subjects as the targets. For each subject pair, accuracies are calculated between source subject and all sessions within the target subject and averaged to obtain the accuracy between source and target subject. As compared to SVM, CNN, and RNN, alignment accuracy of FG-W is greater by an average of 22% ± 2%, 15% ± 5%, and 15% ± 5%, respectively ($p < 0.0005$, paired-sample t-test).

Since data from subject 3 have poor SNR and are severely affected by motion artifacts in half of the fNIRS data (see appendix Table 4), we could treat this subject as an outlier. Alignment accuracy without using data from subject 3 is reported in Table 1 and Fig. 8.

4.6 *Combining n-Back Tasks in Session-by-Session and Subject-by-Subject Alignment*

The analysis of subject performance (Sec. 4.1) showed significant differences in the number of missed targets and wrong reactions depending on the n-back task conditions. Particularly,



(a)

(b)

Fig. 8 Average alignment accuracy (%) from six subjects. (a) Session-by-session alignment, values are shown for SVM, CNN, and RNN using cleaned data from TARA, G-W using original data, and cleaned data from TARA. (b) For subject-by-subject alignment, values are shown for SVM, CNN, and RNN using cleaned data from TARA, FG-W using original and cleaned data from TARA, and FG-W using cleaned data from TARA when data from subject 3 is excluded. Bars represent the means, and error bars represent standard errors.

Table 2 Average session-by-session and subject-by-subject alignment accuracy (%) from G-W and FG-W methods, respectively, when combining 0-back together with 1-back tasks, and 2-back together with 3-back tasks. Cleaned fNIRS data were used. Averages and standard errors across all subjects are reported (Avg.).

	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Avg.
Sess-by-sess using G-W	79 ± 8	99 ± 2	82 ± 17	99 ± 1	96 ± 2	87 ± 10	98 ± 3
Sub-by-sub using FG-W	81 ± 11	89 ± 9	89 ± 3	86 ± 13	88 ± 7	86 ± 13	88 ± 2

the subject performance suggests that 0- and 1-back tasks could be combined together in the alignment since they show similar brain activation behaviors. In this section, we showed that by combining data from 0-back together with 1-back and 2-back together with 3-back tasks, the alignment accuracy increased abruptly for both session-by-session and subject-by-subject alignment, as shown in Table 2. As compared to the results reported in Table 1, the session-by-session alignment accuracy increased by an average of 22% ± 2%, and the subject-by-subject alignment accuracy increased by an average of 33% ± 3%.

5 Discussion

In this study of six subjects, we showed that fNIRS signals measured from 20 channels on the PFC can be used to robustly discriminate subjects' mental workload between different n-back task levels across sessions within one subject and across different subjects. One limitation of our study is a small number of subjects. However, with the current number of subjects (six subjects), this paper still achieved the goals of demonstrating: (1) an alignment accuracy greater than that of chance (25%) for the majority of session-session and subject-subject combinations; and (2) greater accuracies than that obtained from multiclass SVM-, CNN-, and RNN-based models.

We thereby showed the potential of fNIRS as a modality for BCI and user state monitoring that can adapt to different users with various physiological states. Future works will address the extension of this study with a larger sample of subjects to further investigate the variability between sessions and subjects.

In regards to data preprocessing, we show that motion artifact removal in fNIRS signals is an important step for the following mental workload alignment. Specifically, we report that using TARA to remove motion artifacts from fNIRS signals increased alignment accuracy by an average of $3\% \pm 3\%$ for session-by-session alignment and by $5\% \pm 2\%$ for subject-by-subject alignment ($p < 0.005$). Future work could include addressing different types of artifacts that could arise in fNIRS time series, which were not considered by TARA, such as oscillatory transients. In addition, possible future improvements in TARA may be to investigate an automatic way for selection of regularization and nonconvexity parameters in TARA algorithm across subjects.

We introduced two approaches, G-W and FG-W barycenter, for session-by-session and subject-by-subject alignment of mental workload during n-back task. We proved that our methods could be generalized across different sessions and subjects' data. In particular, for session-by-session alignment, we used labeled fNIRS data with known n-back task conditions from one session to align with other unlabeled sessions from the same subject using the G-W method. We showed that most of the unlabeled sessions' data could be mapped correctly to their true labels, with the alignment accuracy ranging from 54% to 77% (with 25% representing chance alignment). Meanwhile, with multi-class SVM and simplified CNN and RNN models, the n-back task classification accuracy was lower (by an average of $43\% \pm 5\%$, $7\% \pm 4\%$, and $5\% \pm 5\%$, respectively). Note that CNN and RNN required the same amount of data as the proposed methods for training, while SVM required more data (from more than one session) for training. Similarly, for subject-by-subject alignment, we used labeled data from one subject as the source data for alignment. Labels and structural information of the source data were combined to generate a new representation (i.e. the FG-W barycenter). Following the same routine as the session-by-session alignment, we were able to use the barycenter from the source subject to predict the labels for data from different sessions for other subjects, with the alignment accuracy from 50% to 71% (also with 25% representing chance alignment). From the corresponding SVM, RNN, and CNN methods, n-back task classification performance achieved lower accuracy than the FG-W method (by an average of 15% to 22%). Again, CNN and RNN were trained from data from one subject (source data), while SVM was trained from data from five subjects for classification. Moreover, our methods of G-W and FG-W do not require the two subsets of data used for alignment to have the same dimension. Thus, they do not require data interpolation due to removing noisy fNIRS signals as for CNN and RNN methods. However, we note that even though G-W and FG-W methods are free from the dimension requirement for data, they could not achieve satisfying results when a large amount of data is missing (e.g., in the case of subject 3 when around half of the channels were discarded in the preprocessing step).

We found relatively higher alignment results for session-by-session alignment (average of $68\% \pm 4\%$) than subject-by-subject alignment (average of $55\% \pm 2\%$). One source of variation in fNIRS data across experiment sessions and across different subjects could come from the variability in systemic physiology, as seen in the variability in the task-evoked changes in MAP and HR (see Sec. 4.2). We observed that the variability of these two physiological measurements are larger across subjects than across sessions. This may explain a greater accuracy results for session-by-session alignment than subject-by-subject alignment. Another source of variability across sessions and subjects may also come from the variation in fNIRS optode placement on the subject's head. We anticipate the optode placement variation to be greater across subjects than across sessions due to different head geometry from different human subjects. From our results, we found that the new representation of the barycenter of the source subject still aligned well to data from other subjects even though subject-by-subject alignment was a more challenging problem. This is indicative that representations of different subjects may still share similar underlying structures even from different domains. Future work will explore generating barycenter from source data from multiple subjects' information for subject-by-subject alignment to account for the across-subject variations in the barycenter.

Based on our alignment results shown in confusion matrices in Fig. 7, the misalignment in session-by-session and subject-by-subject alignments are relatively high between 0-back and 1-back, and between 2-back and 3-back tasks. In particular, the misalignment is the highest between 2- and 3-back tasks (when the 2-back task is the true label and the 3-back is the predicted label and vice versa), ranging from 19.8% to 43.5%. The second highest misalignment is between 0- and 1-back tasks, ranging from 6.8% to 31.8%. Similarly, for subject-by-subject alignment, the highest misalignment came from 0- and 1-back tasks, ranging from 15.2% to 46.5%. The second highest misalignment is between 2- and 3-back tasks, ranging from 14.2% to 38.9%. This gave us an idea of combining 0- with 1-back tasks, and 2- with 3-back tasks in the alignment. Substantial increases in alignment performance (by an average of $22 \pm 2\%$ for session-by-session and $33 \pm 3\%$ for subject-by-subject alignment) suggests that future works could study workload classification between rest to low workload level (0- and 1-back tasks) versus high workload level (2- and 3-back tasks).

Finally, single-distance cw fNIRS measurements of intensity from source-detector pairs at 3-cm distance were used in this study. These measurements have been known to be more sensitive to hemodynamic changes in superficial tissues (i.e., scalp and skull) than in the brain.[51] Previous study[52] has shown that tasked-evoked superficial artifacts may arise during brain activation task due to systemic changes in peripheral physiology rather than the cerebral hemodynamics. This also confirms the claim that variations in our alignment results across sessions and subjects could be partially due to variability in systemic physiological origins. For the purpose of our aim, it is desirable to increase the sensitivity of our measurements to brain tissue to probe hemodynamic changes associated with brain activation. One approach, namely the dual-slope method, involves a simple implementation of a certain arrangement of sources and detectors to localize sensitivity of NIRS measurements to a deeper region,[53] thus suppressing confounding signals from superficial tissue. This approach could also help remove instrumental drifts and motion artifacts from measured signals as dual-slopes are unaffected by changes in optical coupling. Future extensions of this work may involve implementing the dual-slope configuration in such experiments as those described here. Another approach to correct for extracerebral contamination is to acquire measurements in a multidistance arrangement to incorporate short (<1 cm, sensitive to extracerebral tissue only) and long (>2.5 cm, sensitive to both extracerebral and brain tissues) source–detector separations[54] and apply a processing method such as adaptive filtering[55] to remove global interference from systemic physiology from fNIRS measurements.

6 Conclusions

To illustrate that fNIRS signals can be effectively used to identify subjects' mental workload between different n-back task levels across different sessions and subjects, we proposed two domain adaptation methods, G-W and FG-W, for session-by-session and subject-by-subject alignments, respectively. The proposed methods can achieve alignment accuracies greater than the chance level of 25%. At the same time, the proposed methods do not require the same subset of fNIRS channels or further data interpolation for classification across all subjects and sessions as opposed to some other supervised methods such as CNN and RNN. This will alleviate the pressure from having to exclude fNIRS channels that were noisy in one session but not in others, or from having to interpolate the signals to replace those noisy channels. Besides adapting the domain adaptation method, we explored the effect of using the TARA signal processing algorithm for removing motion artifacts and found an improvement in the alignment accuracy results. In the future, we plan to explore the effect of our method on a larger sample of subjects and make it applicable for multiple source subjects.

7 Appendix

Tables 3, 4, and 5 specify the structure of CNN, the number of retained channels for all subjects and the parameters used in TARA for motion artifact removal, respectively. "$*$" stands for multiplication operator and ReLU means the rectified linear activation function.

Table 3 CNN architecture, where d = number of channels (20 in our case), w = number of time points (60 in our case), T_1, T_2 = length of time points after applying the filter and C = number of classes (four in our case).

Layer	Operation	Output size
Input	—	$(2, d, w)$
Conv2D	20 $*$ filter (1, 10) + BatchNorm + ReLU + Dropout (0.2)	$(20, d, T_1)$
Conv2D	20 $*$ filter (1, 5) + BatchNorm + ReLU + Dropout (0.2)	$(20, d, T_2)$
DepthwiseConv2D	20 $*$ kernel $(d, 1)$ + BatchNorm + ReLU + Dropout (0.2)	$(20, 1, T_2)$
—	Flatten	$(20 * T_2)$
Dense $*$ 2	—	C

Table 4 Number of retained channels for six subjects. The total number of channels is 20. “0” indicates when the particular session is removed.

Subject	Number of retained channels
Sub 1	[20, 20, 20, 0]
Sub 2	[15, 17, 16, 16]
Sub 3	[11, 0, 14, 8]
Sub 4	[20, 20, 20, 20]
Sub 5	[20, 20, 20, 20]
Sub 6	[0, 20, 20, 20]

Table 5 Values of TARA parameters (f_c, cut-off frequency for the low-pass filter; d, order of the filter; θ and β, regularization parameters for TARA; σ, noise standard deviation). Parameter values were chosen differently for $\Delta[HbO_2]$ and $\Delta[Hb]$ due to different noise level. For each subject, values of σ and the choice of β vary among sessions, as reported in square brackets “[]”. “-” indicates when TARA is not applied or when the session is removed.

		Parameters				
Subject	Signal type	f_c (Hz)	d	θ	β	σ (μM)
Sub 1	$\Delta[HbO_2]$	0.15	1	0.01	[1.9, 1.9, 1.9, —]	[0.15, 0.15, 0.1, —]
	$\Delta[Hb]$	0.15	1	0.01	[1.9, 1.9, 1.9, 1.9]	[0.05, 0.05, 0.05, 0.025]
Sub 2	$\Delta[HbO_2]$	0.15	1	0.01	[1.7, 1.6, 1.3, 1.2]	[0.25, 0.23, 0.3, 0.15]
	$\Delta[Hb]$	0.15	1	0.01	[1.7, 1.7, 1.3, 1.3]	[0.06, 0.03, 0.04, 0.03]
Sub 3	$\Delta[HbO_2]$	0.15	1	0.01	[1.9, —, 1.5, 1.4]	[0.15, —, 0.13, 0.15]
	$\Delta[Hb]$	0.15	1	0.01	[1.9, —, 1.5, 1.6]	[0.04, —, 0.05, 0.1]
Sub 4	$\Delta[HbO_2]$	—	—	—	—	—
	$\Delta[Hb]$	—	—	—	—	—
Sub 5	$\Delta[HbO_2]$	0.15	1	0.01	[1.8, 1.3, 1.9, 1.9]	[0.1, 0.1, 0.15, 0.14]
	$\Delta[Hb]$	0.15	1	0.01	[1.8, 1.3, 1.9, 1.9]	[0.02, 0.015, 0.025, 0.02]
Sub 6	$\Delta[HbO_2]$	0.15	1	0.01	[—, 1.9, 1.9, 1.7]	[—, 0.16, 0.1, 0.16]
	$\Delta[Hb]$	0.15	1	0.01	[—, 1.9, 1.6, 1.9]	[—, 0.03, 0.019, 0.015]

Disclosures

The authors have no relevant financial interests in this manuscript and no potential conflicts of interest to disclose.

Acknowledgments

This material was based upon work supported by the Air Force Office of Scientific Research under Grant No. FA9550-18-1-0465. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the United States Air Force. Shuchin Aeron would like to acknowledge support by NSF CAREER Grant No. CCF:1553075.

Code, Data, and Materials Availability

No materials were used for the analysis. The code and data used to generate the results and figures are available in the Github repository: https://github.com/boyanglyu/nback_align.

References

1. M. A. Franceschini et al., "On-line optical imaging of the human brain with 160-ms temporal resolution," *Opt. Express* **6**, 49–57 (2000).
2. M. Wolf et al., "Different time evolution of oxyhemoglobin and deoxyhemoglobin concentration changes in the visual and motor cortices during functional stimulation: a near-infrared spectroscopy study," *NeuroImage* **16**(3), 704–712 (2002).
3. K. Bejm et al., "Influence of contrast-reversing frequency on the amplitude and spatial distribution of visual cortex hemodynamic responses," *Biomed. Opt. Express* **10**(12), 6296 (2019).
4. X. Cui, D. M. Bryant, and A. L. Reiss, "NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation," *NeuroImage* **59**(3), 2430–2437 (2012).
5. A. Bosworth, M. Russell, and R. J. K. Jacob, "Update of fNIRS as an input to brain–computer interfaces: a review of research from the Tufts Human–Computer Interaction Laboratory," *Photonics* **6**(3), 90 (2019).
6. K.-S. Hong, U. Ghafoor, and M. J. Khan, "Brain–machine interfaces using functional near-infrared spectroscopy: a review," *Artif. Life Rob.* **25**, 204–218 (2020).
7. A. M. Owen et al., "N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies," *Hum. Brain Mapp.* **25**(1), 46–59 (2005).
8. E. E. Smith and J. Jonides, "Working memory: a view from neuroimaging," *Cognit. Psychol.* **33**(1), 5–42 (1997).
9. C. Herff et al., "Mental workload during n-back task–quantified in the prefrontal cortex using fNIRS," *Front. Hum. Neurosci.* **7**, 935 (2014).
10. H. Aghajani, M. Garbey, and A. Omurtag, "Measuring mental workload with EEG + fNIRS," *Front. Hum. Neurosci.* **11**, 359 (2017).
11. J. Shin et al., "Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset," *Sci. Data* **5**(1), 580003 (2018).
12. S. Ben-David et al., "A theory of learning from different domains," *Mach. Learn.* **79**(1–2), 151–175 (2010).
13. W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1 (2019).
14. V. K. Kurmi and V. P. Namboodiri, "Looking back at labels: a class based domain adaptation technique," in *Int. Joint Conf. Neural Networks (IJCNN)*, IEEE, pp. 1–8 (2019).
15. H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Plann. Inference* **90**(2), 227–244 (2000).
16. G. Peyré and M. Cuturi, "Computational optimal transport: with applications to data science," *Found. Trends® Mach. Learn.* **11**(5–6), 355–607 (2019).

17. J. Solomon et al., “Entropic metric alignment for correspondence problems,” *ACM Trans. Graphics* **35**(4), 1–13 (2016).
18. F. Mémoli, “Gromov–Wasserstein distances and the metric approach to object matching,” *Found. Comput. Math.* **11**(4), 417–487 (2011).
19. D. Das and C. G. Lee, “Sample-to-sample correspondence for unsupervised domain adaptation,” *Eng. Appl. Artif. Intell.* **73**, 80–91 (2018).
20. N. Courty et al., “Optimal transport for domain adaptation,” *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(9), 1853–1865 (2016).
21. O. Yair et al., “Optimal transport on the manifold of SPD matrices for domain adaptation,” arXiv:1906.00616 (2019).
22. G. Peyré, M. Cuturi, and J. Solomon, “Gromov–Wasserstein averaging of kernel and distance matrices,” in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, Vol. 48, pp. 2664–2672 (2016).
23. T. Vayer et al., “Optimal transport for structured data with application on graphs,” in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, California, Vol. 97, pp. 6275–6284 (2019).
24. M. R. Siddiquee et al., “Movement artefact removal from NIRS signal using multi-channel IMU data,” *BioMed. Eng. OnLine* **17**(1), 120 (2018).
25. Q. Zhang, G. E. Strangman, and G. Ganis, “Adaptive filtering to reduce global interference in non-invasive NIRS measures of brain activation: how well and when does it work?” *NeuroImage* **45**(3), 788–794 (2009).
26. F. Scholkmann et al., “How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation,” *Physiol. Meas.* **31**(5), 649–662 (2010).
27. X. Cui, S. Bray, and A. L. Reiss, “Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics,” *NeuroImage* **49**(4), 3039–3046 (2010).
28. M. Izzetoglu et al., “Motion artifact cancellation in NIR spectroscopy using discrete Kalman filtering,” *BioMed. Eng. OnLine* **9**(1), 16 (2010).
29. A. V. Medvedev et al., “Event-related fast optical signal in a rapid object recognition task: improving detection by the independent component analysis,” *Brain Res.* **1236**, 145–158 (2008).
30. H. Sato et al., “Wavelet analysis for detecting body-movement artifacts in optical topography signals,” *NeuroImage* **33**(2), 580–587 (2006).
31. I. W. Selesnick et al., “Transient artifact reduction algorithm (TARA) based on sparse optimization,” *IEEE Trans. Signal Process.* **62**(24), 6596–6611 (2014).
32. A. Sassaroli and S. Fantini, “Comment on the modified Beer–Lambert law for scattering media,” *Phys. Med. Biol.* **49**(14), N255–N257 (2004).
33. I. J. Bigio and S. Fantini, *Quantitative Biomedical Optics: Theory, Methods, and Applications*, 1st ed., Cambridge University Press (2016).
34. http://eeweb.poly.edu/iselesni/TARA/index.html.
35. L. V. Kantorovich, “On the translocation of masses,” *J. Math. Sci.* **133**(4), 1381–1382 (2006).
36. G. Monge, *Mémoire sur la théorie des déblais et des remblais*, Histoire de l’Académie Royale des Sciences de Paris (1781).
37. M. Cuturi, “Sinkhorn distances: lightspeed computation of optimal transport,” in *Adv. Neural Inf. Process. Syst.*, pp. 2292–2300 (2013).
38. R. Sinkhorn, “Diagonal equivalence to matrices with prescribed row and column sums. II,” *Proc. Am. Math. Soc.* **45**(2), 195–198 (1974).
39. B. K. Abid and R. M. Gower, “Greedy stochastic algorithms for entropy-regularized optimal transport problems,” in *Proc. Twenty-First Int. Conf. Artificial Intelligence and Statistics*, Vol. **84**, pp. 1505–1512 (2018).
40. J. Altschuler, J. Niles-Weed, and P. Rigollet, “Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration,” in *Adv. Neural Inf. Process. Syst.*, pp. 1964–1974 (2017).
41. A. Barachant et al., “Classification of covariance matrices using a Riemannian-based kernel for BCI applications,” *Neurocomputing* **112**, 172–178 (2013).

42. D. Heger et al., "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Hum. Assoc. Conf. Affect. Comput. and Intell. Interact.*, pp. 832–837 (2013).
43. R. Bhatia, S. Gaubert, and T. Jain, "Matrix versions of the Hellinger distance," *Lett. Math. Phys.* **109**(8), 1777–1804 (2019).
44. T. Vayer et al., "Fused Gromov–Wasserstein distance for structured objects," *Algorithms* **13**(9), 212 (2020).
45. A. Barachant et al., "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.* **59**(4), 920–928 (2012).
46. V. J. Lawhern et al., "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.* **15**, 056013 (2018).
47. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *3rd Int. Conf. Learn. Represent.* (2015).
48. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
49. L. McInnes, J. Healy, and J. Melville, "UMAP: uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426 (2018).
50. M. Ali et al., "TimeCluster: dimension reduction applied to temporal data for visual analytics," *Vis. Comput.* **35**(6–8), 1013–1026 (2019).
51. I. Tachtsidis and F. Scholkmann, "False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward," *Neurophotonics* **3**(3), 031405 (2016).
52. E. Kirilina et al., "The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy," *Neuroimage* **61**(1), 70–81 (2012).
53. A. Sassaroli, G. Blaney, and S. Fantini, "Dual-slope method for enhanced depth sensitivity in diffuse optical spectroscopy," *J. Opt. Soc. Am. A* **36**(10), 1743–1761 (2019).
54. S. Fantini, B. Frederick, and A. Sassaroli, "Perspective: prospects of non-invasive sensing of the human brain with diffuse optical imaging," *APL Photonics* **3**, 110901 (2018).
55. Q. Zhang, E. N. Brown, and G. E. Strangman, "Adaptive filtering to reduce global interference in evoked brain activity detection: a human subject case study," *J. Biomed. Opt.* **12**(6), 064009 (2007).

Boyang Lyu is a PhD student at Tufts University under Professor Shuchin Aeron. Her research involves unsupervised domain adaptation and signal processing. She has applied some of these techniques to word alignment and mental workload identification.

Thao Pham is a PhD student in the Diffuse Optical Imaging of Tissue (DOIT) Lab at Tufts University, under Professor Sergio Fantini. Her research interests involve using near-infrared spectroscopy (NIRS) and coherent hemodynamics spectrocopy (CHS) model for noninvasive monitoring of cerebral blood flow (CBF) and cerebral hemodynamics in healthy human subjects and in clinical settings.

Giles Blaney is a PhD student in the Diffuse Optical Imaging of Tissue (DOIT) Lab at Tufts University under Professor Sergio Fantini. He received a Bachelor of Science in mechanical engineering and physics from Northeastern University in 2017, with minors in electrical engineering and mathematics. Currently, he is researching methods for depth discrimination and imaging for use in NIRS. This includes studying the sensitivity of various NIRS optode arrangements in heterogeneous media, and development of the dual-slope method.

Angelo Sassaroli received his PhD in physics from the University of Electro-Communications, Tokyo, Japan, in 2002. From 2002 to 2007, he was a research associate at Tufts University, Medford, Massachusetts, USA. Since 2007 he has been a research assistant professor at Tufts University. He is the coauthor of more than 80 peer reviewed publications. His research interests focus on diffuse optical imaging.

Sergio Fantini is a professor of biomedical engineering and principal investigator of the "Diffuse Optical Imaging of Tissue Laboratory" (DOIT Lab) at Tufts University. The DOIT

Lab aims to develop noninvasive applications of NIRS for medical diagnostics, monitoring of tissue oxygenation, quantitative assessment of tissue perfusion, and functional imaging. He coauthored with Dr. Irving Bigio the textbook *Quantitative Biomedical Optics*, published by Cambridge University Press. He is a fellow of OSA, SPIE, and AIMBE.

Shuchin Aeron is an associate professor in the Department of ECE at Tufts University. Prior to Tufts, he was a postdoctoral research scientist at Schlumberger Doll Research, Cambridge, Massachusetts, from 2009 to 2011. He was awarded the School of Engineering and Electrical and Computer Engineering Best Thesis Award. He is a recipient of the NSF CAREER award (2016). His main research interest lies at the intersection of information theory, statistical signal processing, optimization, and machine learning.