Taking a Deeper Look at the Brain: Predicting Visual Perceptual and Working Memory Load from High-Density fNIRS Data

Jiyang Wang, Trevor Grant, Senem Velipasalar*(Senior Member, IEEE)*, Baocheng Geng and Leanne Hirshfield

***Abstract*— Predicting workload using physiological sensors has taken on a diffuse set of methods in recent years. However, the majority of these methods train models on small datasets, with small numbers of channel locations on the brain, limiting a model's ability to transfer across participants, tasks, or experimental sessions. In this paper, we introduce a new method of modeling a large, cross-participant and cross-session set of high density functional near infrared spectroscopy (fNIRS) data by using an approach grounded in cognitive load theory and employing a Bi-Directional Gated Recurrent Unit (BiGRU) incorporating attention mechanism and self-supervised label augmentation (SLA). We show that our proposed CNN-BiGRU-SLA model can learn and classify different levels of working memory load (WML) and visual processing load (VPL) across participants. Importantly, we leverage a multi-label classification scheme, where our models are trained to predict simultaneously occurring levels of WML and VPL. We evaluate our model using leave-one-participant-out (LOOCV) as well as 10-fold cross validation. Using LOOCV, for binary classification (off/on), we reached an F1-score of 0.9179 for WML and 0.8907 for VPL across 22 participants (each participant did 2 sessions). For multi-level (off, low, high) classification, we reached an F1-score of 0.7972 for WML and 0.7968 for VPL. Using 10-fold cross validation, for multi-level classification, we reached an F1-score of 0.7742 for WML and 0.7741 for VPL.**

***Index Terms*— Cognitive Load, Working Memory Load, Workload, Classification, Deep Learning, fNIRS, self-supervision**

I. Introduction

IT is known in Human Computer Interaction (HCI) that optimum human performance can be achieved with systems that help users to maintain an ideal level of workload (WL). Too little WL can result in low productivity, boredom and complacency, while too much WL can result in human error, shedding of tasks, and frustration [1]–[4]. A plethora of recent research has used real-time behavioral [5] and physiological [6], [7] measures to make real-time predictions of WL with a common goal of building adaptive systems that can regulate users' workload. These adaptive systems would benefit not only from information about a user's overall workload, but by information gleaned from taking a more fine-grained view of workload by differentiating between the load on one's perceptual resources and the load on one's working memory resources. This way, an adaptive system could change the modality by which support is presented (visual channel), based on information about the auditory or visual perceptual load of the person. A person who has a dangerously high level of working memory load (WML) while staring at a radar monitor could be assisted through his or her auditory channel (by hearing information), while a person who has a high WML with very low current visual channel demands may benefit best by the assistance provided on his/her monitor, through that 'available' visual channel.

Despite the large body of literature devoted to the topics of WL, it has proven difficult to build robust and intelligent systems capable of predicting WL outside of tightly controlled laboratory experiments [8]. Differences in theoretical grounding of WL lead to differences not only in WL manipulations in experimental paradigms, but also competing evidence as to which measures, both behavioral and physiological, are most effective at measuring WL [7], [9]. Recently, increasingly effective strides have been made in the prediction and modulation of WL levels in tasks by using both behavioral [10] and, more recently, physiological and psychophysiological [1], [2], [6] measures. Even though these achievements have elucidated some of the underlying challenges with classifying and predicting WL, we still lack a clear picture of what an ideal approach to creating more accurate predictions might be [7].

Although many early successes were achieved using machine learning on brain data [11], several notable challenges have arisen, which significantly limit the impacts of these early successes [9]. In particular, earlier work trained models per individual, on very small datasets, leading to model overfitting and inflated accuracy rates. When these models were tested on

"The information, data, or work presented herein was funded in part by National Science Foundation (NSF) under Grant 1739748, Grant 1816732 and by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000940. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof."

Jiyang Wang, Senem Velipasalar are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, 13244 USA (e-mail: {jwang127, svelipas, bageng}@syr.edu). Trevor Grant and Leanne Hirshfield are with the Institute of Cognitive Science, University of Colorado, Boulder, CO, 80309 USA (e-mail: {Trevor.grant, Leanne.hirshfield}@colorado.edu). Baocheng Geng is with the Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL 35294 USA (e-mail: bgeng@uab.edu).

new participants, or even on the same participant during a different measurement session, the model performance degraded significantly [9], [12], [13]. The performance degradation may be in part due to the assumptions made by traditional machine learning models about the structure of the underlying data. Whereas traditional machine learning models assume that all training data samples are independent and identically distributed (i.i.d.), data obtained from brain measurements does not exhibit these characteristics.

To address the aforementioned challenges, the research outlined in this paper involves the creation and testing of advanced deep learning approaches that are well suited to pair with high-density Functional Near Infrared Spectroscopy (fNIRS) data, where the measurement channels are spatially and temporally intertwined. As shown in Fig. 1, high-density fNIRS devices can capture the temporal and spatial neural correlates of the target states of interest, making the resulting data well suited as input into deep learning models. Deep learning-based approaches have been successfully used in the video analysis domain to consider the inter-dependencies among the spatial and temporal relations within the data.

Our approach enables us to go beyond the classification of general WL, which Hart and Staveland [14] describe as "the perceived relationship between the amount of mental processing capability or resources and the amount required by the task". Our approach can delineate between different levels of working memory load as well as *concurrently occurring* visual perceptual load, using a multi-label classification schema. The vast majority of research to date has shown success at predicting overall levels of WL, which correlate with overall task difficulty. It is less common to further delineate WL into more specific sub-components of WL, such as working memory load and visual perceptual load. This added knowledge could be used to better inform adaptive systems on how to adapt in order to support users during human-computer interaction. For example, a technology user who is experiencing high working memory load can be assisted by an adaptive system, which may choose to assist via visual (e.g., computer display) or auditory (e.g., speakers) channels depending on the additional knowledge about the person's visual perceptual load. For example, a pilot whose helmet is embedded with neurophysiological sensors could be assisted in real-time while he/she flies the aircraft. If the pilot becomes overwhelmed by a complex task, we may predict that his/her working memory load is very high. To support and help the over-taxed pilot, we may provide real-time decision support information. In that case, we could choose the modality of the information that we present (through a visual cockpit display versus through an auditory channel into the pilots headphones) based on the pilot's current visual perceptual load.

Compared to the images and videos in the datasets commonly used by the computer vision community, fNIRS data has much more limited spatial resolution and a lower signal-to-noise-ratio (SNR). For example, the high density fNIRS used in these data collections is the Hitachi ETG4000, which can measure up to 52 channel locations. Convolutional layers will eliminate the spatial information and enhance the semantic information [15]. As a result, the classification model to be developed cannot be too deep, and it can run into problems with overfitting. In order to extract semantically rich features with a limited number of convolutional layers, we introduce self-supervised learning into the classification scheme with newly proposed transformations, which are well-suited for fNIRS data. Self-supervised learning has the ability to extract semantic features based on the spatial context of images. It was first studied in unsupervised learning [16], where annotated or labeled data is not used. Instead, the input data is projected to a latent space, where similar semantic features are close to each other. The basic idea behind self-supervised learning is defining a pre-task and generating corresponding artificial labels. Then, the model is trained with artificial labels in a supervised manner to solve the pre-task. One of the simplest, yet effective pre-tasks is predicting a transformation, such as rotation, colorization or patch context [17], self-supervised using a discriminative loss on transformations. This type of loss may help the model to learn a better representation. For semi-supervised learning [18], [19], on the other hand, the common set up is based on learning the original task and pretask with two classifiers. Lee et al. [20] introduced self-supervised label augmentation (SLA), which learns a joint distribution of supervised and self-supervised signals.

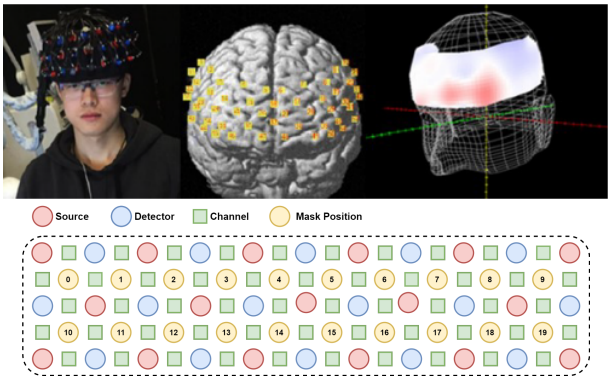

Fig. 1: The probe configuration for the fNIRS data. Red, blue and yellow circles represent sources, detectors and mask locations (Sec. IV-F), respectively. Green squares are the fNIRS channels.

In our proposed approach, we employ and modify a Bidirectional GRU (BiGRU) to apply it to fNIRS data, and introduce a learning framework, which modifies and adopts SLA providing a self-supervision signal to multi-branch and joint classifier schemes. The primary contributions of this paper include the following: (i) different from prior art, we are able to further delineate WL into its more specific sub-components, namely Working Memory Load (WML) and Visual Perceptual Load (VPL), by using a multi-level multi-label deep learning framework using the spatial and temporal information encapsulated in fNIRS data; (ii) we employ and compare two classifiers, namely multi-branch and joint classifiers; (iii) we introduce an attention mechanism on BiGRU network to process and weigh the sequential features; (iv) we incorporate and study the impact of self-supervised label augmentation (SLA) by introducing a new transformation for the pre-task, which uses different ordering of the controlled rest and task data, instead of rotation and permutation, and is more suitable for fNIRS data, and use this SLA with both multi-branch and joint

classifiers; (v) we perform not only binary but also multi-level (off, low, high) classification of WML and VPL; (vi) we perform data balancing in terms of different tasks (task-balanced data) as well as for different labels (label-balanced data) and compare the performances in detail; (vii) we demonstrate our approach on a dataset consisting of 44 sessions of data collection (22 participants completing two identical sessions on different days) using leave-one-participant-out and 10-fold cross validation; (viii) we compare our proposed approach with three types of commonly used classifiers and show that our method provides the highest performance.

II. Related Work

Predicting WL has been of interest since at least 1908, when the Yerkes-Dodson Law of arousal suggested that task performance is optimized when cognitive workload is neither too high nor too low [21]. Since that time, a dearth of research has aimed to define, operationalize, and measure the construct. More recently, large bodies of work have attempted to harness the power of machine learning techniques to make predictions of WL in single-trial and real-time settings. In this section, we i) provide an overview of non-invasive brain measurement techniques; ii) summarize the progress to date (and challenges encountered) using feature-based machine learning techniques on brain data to predict WL; and iii) highlight the more nascent research applying deep learning techniques to brain data.

A. Non-Invasive Brain Measurement and Utility of fNIRS

Functional magnetic resonance imaging (fMRI) is the most widely used neuroimaging tool in the neuroscience literature [22] due to its high level of spatial resolution (3Tesla fMRI scanners have resolution of 2-3mm voxels [22]) and relatively high level of temporal resolution (3Tesla fMRI scanners have temporal resolution of 2-4 sec. [22]). fMRI, however, is limited as a research modality within the HCI domain, since it is expensive, restricts participant's movements and interactions, and has a high sensitivity to motion artifacts, requiring participants to lie perfectly still while their brains are being measured.

Researchers in HCI have therefore turned to other devices, such as electroencephalography (EEG) and functional near infrared spectroscopy (fNIRS), for conducting experiments in more ecologically valid settings that may require participant movement. EEG, which has been actively used as a research tool for over 100 years, uses electrodes placed on the scalp to measure the electrical potential caused by neural activation across the brain's surface. EEG benefits from high temporal resolution, but suffers from a poor signal-to-noise ratio and has low spatial resolution [23]. fNIRS, on the other hand, operates by the use of near-infrared light, which can penetrate through scalp and skull to reach the cortical surface of the brain. Optical fibers are placed on the surface of the head for illumination, while detection fibers measure the light that is reflected back from the brain tissue. The change in light intensity can allow the device to detect concentration changes in oxy- and deoxy- hemoglobin [24] [25]. A review of the history of fNIRS is provided in [26]. fNIRS has the benefit of a higher spatial resolution than EEG, making it possible to localize specific functional brain regions of activation, as could be done with the constrictive fMRI device [23]. The ability to spatially locate specific functional brain regions of interest enables high-density fNIRS sensors to identify specific neural correlates of WL and other mental states of interest. However, fNIRS' temporal resolution (2-4 seconds) is relatively slow compared to EEG, especially in the context of adaptive systems. For this reason, multiple researchers have explored a hybrid approach whereby they merge EEG and fNIRS sensors together, with the goal of maximizing both temporal and spatial resolution, and ideally getting complementary data signals, for fast and precise classification of mental states in adaptive systems [27], [28]. This work contributes to this long-term goal, by investigating the suitability of high density fNIRS for the classification of states that utilize the high spatial resolution of high density fNIRS, and the ability to measure specific regions of functional regions of interest in the brain.

B. Workload Classification on Brain Data

Despite a series of incremental successes in researchers' ability to classify WL, the field has also converged on a number of challenges that may be hampering the impact of the successes achieved thus far. Some of these challenges include needing clear definitions of the mental states being measured, connecting those states with neurophysiology, eliminating confounding factors, and providing insight into machine learning models [9]. At the turn of the last century, researchers successfully trained neural networks to predict cognitive workload across participants from EEG data on a small set of participants [11]. However, these researchers employed hand selected and heavily pre-processed features generated from a small dataset. They also recorded a reduced accuracy when generalizing across participants. Since then, improvements to datasets, methods, and research designs have aimed to increase model robustness and prediction generalization across participants, tasks, and contexts. Several studies have opted to perform feature extraction by hand for classifying workload [2], [29], which often requires domain specific knowledge of the sensor and context. Models such as those mentioned above require that the brain data be represented by a feature-set, defined a priori. However, identifying the 'correct' features a priori for generating accurate predictions may not always be realistic.

Another issue is that many models are built per-participant, resulting in very small datasets that may not accurately represent the extremely high feature space of the brain data, and leading to the model overfitting to an individual. Since collecting training data with brain measurement devices is costly and time-consuming, researchers have noted the need to build models across participants [30], and it is becoming more commonplace to build and evaluate models using 'leave-one-participant-out cross-validation' [25], [31], [32]. Notably, several studies have demonstrated transfer learning between participants [33], [34]. Yet, despite these advancements, transfer learning between participants usually results in low classification accuracy. For example, in the case of [33], the best performing model between participants only achieved a mean accuracy of 63%.

Another challenge for the development of real-time adaptive systems is that relatively few studies have attempted to predict the perceptual load modality (audio, visual, etc.) associated with a participant's workload, despite this type of information being very valuable for intelligent adaptive interfaces to act upon. Some have successfully classified audio and visual workload conditions across participants [28], but this method again used hand-crafted features on a small dataset. Another area that shows promise is experimenting with augmenting EEG with fNIRS to improve classification generalization [2], [28], [34]. Increased accuracies in these cases may indicate that this combination may provide complementary information to models trained on both measurements. The majority of the work outlined above does classification on windows of time that are greater than 30 sec. per instance. In all cases, the prediction accuracy of cognitive workload decreases as brain data time windows are shortened [2], [35]. If real-time adaptive systems are to be developed, models that are able to make predictions in shorter time windows must be explored.

C. Deep Learning on Neurophysiological Data

In this section, relevant research from the video classification domain is described, since fNIRS data is spatio-temporal in nature just like video data. A common network structure uses Convolutional Neural Networks (CNNs) to process spatial information, and then hands over to recurrent units to learn temporal information. Researchers showed that Long Short Term Memory (LSTM) networks could robustly classify human activities, such as running and walking, from videos [36]. With 25 actors in the videos, they showed that LSTMs were robust to noise, and could generalize across actors. However, instead of using CNNs to extract features from the videos, they opted for hand-crafted features using an optical flow algorithm. Gated Recurrent Unit (GRU) [37] is another type of recurrent unit that can capture dependencies among different time steps. Similar to the LSTM unit, the GRU has gating units that control the information flow inside the unit without having a separate memory cell. It has a simpler information flow with less parameters compared to the LSTM unit. We employ GRU units to build our proposed sequential module, which will be described in more detail below.

It has been shown that deep convolutional LSTM networks can perform end-to-end feature extraction of EEG signals with minimal domain-specific human knowledge required [38]. Appriou et al. [33] showed that convolutional networks outperformed other machine learning methods for workload classification from EEG signals. Still, few works have focused on using CNNs on fNIRS data. In particular, CNNs have been applied to fNIRS data for gender classification [39], locating Regions of Interest (ROI) [40], and classification of affect [41]. Researchers have used deep learning to classify fNIRS signals acquired during different Brain-Computer interfacing cognitive tasks [42], [43]. Most recently [44], CNNs were used to classify workload levels from a seven channel fNIRS input.

III. Proposed Method

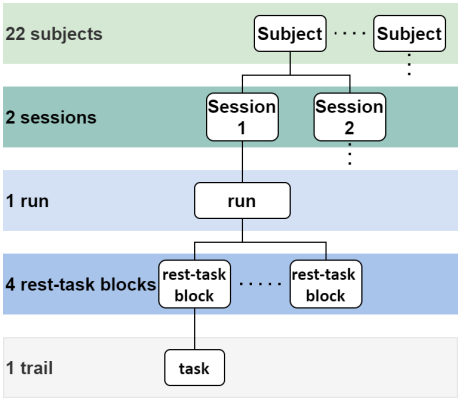

Fig. 2: Organization and structure of the collected data.

A. Data Collection

fNIRS data was collected from an experimental protocol in which participants completed a series of computerized cognitive tasks, which have been used in both clinical and cognitive psychological fields to invoke distinct types of WL [45], [46]. 22 participants were recruited from both graduate and undergraduate university students with median age equal to 22. Each of the 22 subjects participated in data collection for two sessions, collected 5 weeks apart, as shown in Fig. 2. Each session contains one run of a randomized block design with the four cognitive tasks interspersed with resting periods (25 seconds) in between each of the cognitive task blocks. Each block lasted for 60 seconds. In order to introduce jitter into the stimulus presentations within blocks, trials within each block were separated with a variable inner trial interval randomly selected from a gamma distribution on a per trial basis (k=2 seconds, θ=1.5 seconds).

Participants had normal or corrected to normal vision, and were seated 60cm away from a 26-inch computer monitor whose refresh rate was locked to 60Hz. fNIRS data was collected at a sampling rate of 10Hz using a Hitachi ETG-4000. As shown in Fig. 1, fNIRS optodes were arranged in a 3x11 array, with a 3cm optode separation distance, fitted into a cap, and then placed on the forehead of each participant symmetrically, covering the frontal cortex. The device uses near infrared light (695nm, 830nm) to measure both oxygenated and de-oxygenated hemoglobin levels in the blood at 52 separate measurement channels on each participant's brain. After the probes were placed, each fNIRS channel was calibrated using the tools provided in the ETG-4000 system to ensure that they were providing a satisfactory reading.

Cognitive Tasks: The cognitive tasks that participants completed were administered in a randomized block design format. They included an emotional working memory (ewm) task, which was a delayed recall task where participants were presented with an array of six letters on the screen, and were asked to memorize the letters in the array. During the delay period an image was displayed that was either intended to be neutral and produce no arousal, or an image intended to elicit high negative valence. The participants were then asked if a certain letter had appeared in the original array. They would indicate a "Yes" response by pressing the left arrow key on the keyboard, and a "No" response by pressing the right

arrow key. The audio n-back (anb) task involved participants holding a stream of continually adapting letter values [b, t, q, v] in their working memory while simultaneously attempting to recall whether or not a given letter was displayed two presentations prior. Importantly, information was presented through a speaker rather than on the screen, and we only administered 2-back tasks (with the number of items maintained, and continually being updated, set to 2). For a detailed description of the n-back see [47]. When the participants heard a letter, they would respond by pressing the left arrow key if their letter matched that of 2 presentations ago. If the letter did not match, they would press the right arrow key. The reaction time (rt) task displayed a fixation point in the center of the screen at the start of each trial. After a variable period of time for each trial (min=300ms, mean=500ms, max=700ms) a large "X" stimulus replaced the fixation point in the center of the screen. The participant's task was to respond as quickly as possible by pressing the left arrow key when the fixation point was replaced with the stimulus. The go-no-go (gng) task followed the protocol of Herrmann et al. [48], and included a red rectangle target stimulus that appeared in the center of the screen and a blue oval distractor stimulus. Participants were tasked with responding as quickly as possible when they were presented with the target stimulus by pressing the left arrow key on the keyboard, and with not responding when they were presented with the distractor stimulus. The stimulus appeared on screen for a variable amount of time (1 to 2 sec.); a variable inter-stimulus interval was presented between trials, during which a cross fixation point was displayed on the screen before the subsequent test began.

We selected the four cognitive benchmark tasks to elicit controlled levels of load on participants' WML and VPL resources. Both the anb and ewm tasks are carefully designed to engage participants' working memory resources, while the rt and gng tasks do not involve working memory. These tightly controlled tasks enabled us to build out our models using two different multi-labeling schemes, where we focus on each benchmark task's expected load on participants WML and VPL, as shown in Table I. More specifically, in the first labeling scheme, fNIRS data collected during these tasks are given multi-class label values of 0 (off), 1 (low) or 2 (high) across all two relevant WL sub-components, namely WML and VPL for modeling. We also explore a simpler binary classification, where we combine the labels of 1 (low) and 2 (high) into a single 'on' (1/on) class.

Task	WML-VPL
rt	0-1
gng	0-1
ewm	1-2
anb	2-0

(a) Multi-level multi-labeling scheme

Task	WML-VPL
rt	0-1
gng	0-1
ewm	1-1
anb	1-0

(b) Binary-level multi-labeling scheme

TABLE I: Labeling schemes of fNIRS data

B. Data Preprocessing

We employ a bandpass filter followed by Z-score normalization to remove noise artifacts from fNIRS data. More specifically, the density data, captured through the Hitachi ETG-4000, is first converted into the rate of change of oxy (ΔHbO) and de-oxy hemoglobin (ΔHb) values using modified Beer-Lambert Law [49]. These values are then band-pass filtered with low and high frequency values of 0.01Hz and 0.5Hz, respectively. Then, each channel of data for each participant was normalized by using Z-score normalization [50] independently for each run of experiments. We treat the fNIRS data like video data, preserving the relative locations of fNIRS channels on the head, which measure data over time. The layout of the 52 fNIRS channels is shown in Fig. 3. In order to preserve the channel layout without omitting any data, and have a suitable size matrix for our neural network encoders, we insert zeros at locations shown in Fig. 3. Thus, our data matrix is 6×22 at each time step.

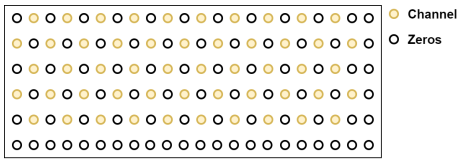

Fig. 3: Spatial arrangement of fNIRS channels.

C. Classification Model

As mentioned above, a Gated Recurrent Unit (GRU) [37] can capture dependencies among different time steps. In our proposed approach, we employ and modify a Bidirectional GRU (BiGRU) to apply it to fNIRS data. BiGRU consists of two GRU units, taking the input sequence in forward and backward directions. BiGRU provides better performance compared to unidirectional (regular) GRU [37]. The overall architecture of our classification model is shown in Fig. 4a. The sequence of data (denoted by M) is first encoded by a CNN, and then the extracted features are sent to a BiGRU. The attention mechanism from Seq2Seq [51], [52] is applied to provide a weighted sequential hidden feature. We consider two types of classifiers: multi-branch classifier and joint classifier. We introduce the self-supervision signal to these two classifier schemes jointly by using the self-supervised label augmentation (SLA) method [20]. All of these steps will be described in more detail below.

For the multi-label classification task, the data has been annotated with multiple labels. For our fNIRS data, there are two labels for every sample, each label indicating the level of WML or VPL as shown in Table I . The model needs to predict the activation level for the WML and VPL. In fully-supervised learning, one can formulate the multi-branch objective L_{MB} based on cross-entropy as:

$$\begin{aligned} L_{MB}(\mathbf{M}, y_{wml}, y_{vpl}; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = & L_{CE}(\sigma(f(\mathbf{M}; \boldsymbol{\theta}); \boldsymbol{\mu}), y_{wml}) \\ & + L_{CE}(\sigma(f(\mathbf{M}; \boldsymbol{\theta}); \boldsymbol{\nu}), y_{vpl}) \end{aligned} \tag{1}$$

where $f(\cdot; \boldsymbol{\theta})$ represents CNN-BiGRU attention model with parameters $\boldsymbol{\theta}$; $\mathbf{M}$ is the sequential input matrix; y_{wml} and

y_vpl are the congnitive level for WML and VPL respectively; and $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are the parameters of the two classifiers $\sigma(\cdot;\boldsymbol{\mu})$ and $\sigma(\cdot;\boldsymbol{\nu})$.

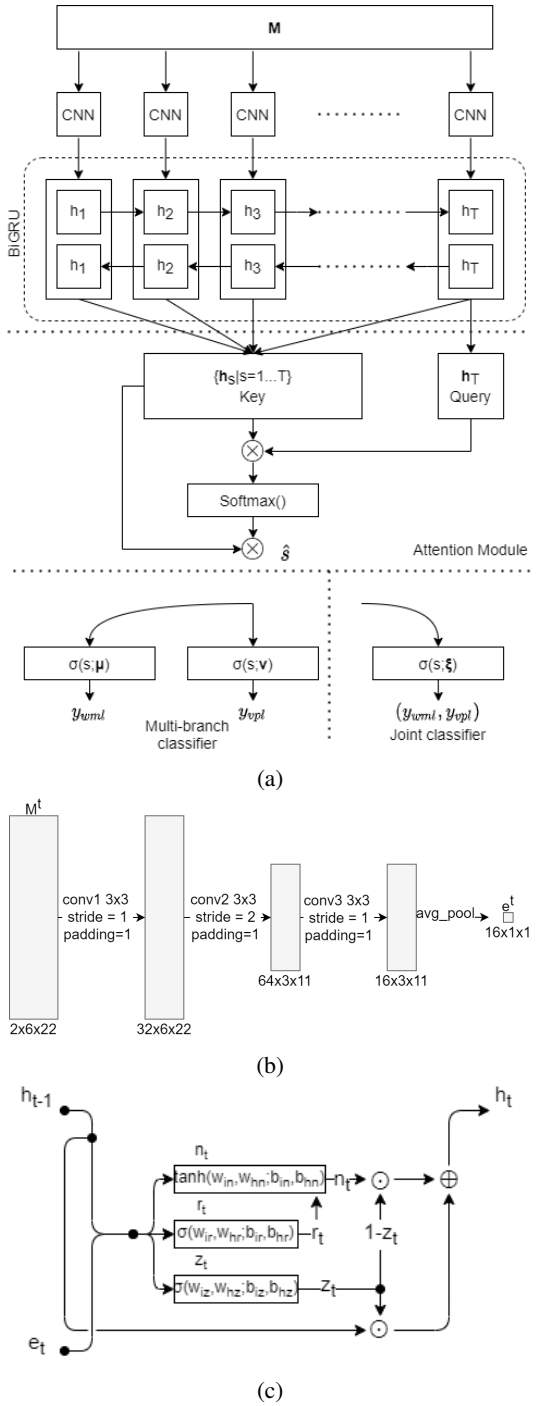

Fig. 4: **(a)** Architecture of our classification model, where $\mathbf{M}$ is the input sequential data. The upper part represents the CNN-BiGRU with attention model. The lower part includes two classifier schemes we considered: multi-branch classifier and joint classifier. **(b)** The structure of the CNN encoder. **(c)** Gated Recurrent Unit (GRU).

Unlike multi-branch classifier, the joint classifier uses a single classifier $\sigma(\cdot;\boldsymbol{\xi})$, which predicts a joint probability covering all the combinations of the multi-labeling scheme. In our case, Table I contains three different combinations for both binary and multi-class labeling schemes. In fully-supervised learning, one can describe the joint objective L_J based on cross-entropy as:

$$L_J(\mathbf{M},(y_{wml},y_{vpl});\boldsymbol{\theta},\boldsymbol{\xi}) = L_{CE}(\sigma(f(\mathbf{M};\boldsymbol{\theta});\boldsymbol{\xi}),(y_{wml},y_{vpl})) \quad (2)$$

1) CNN encoder: We propose a three-layer CNN structure to encode the spatial information at each time step t. Given the input data $\mathbf{M} = \{...,M^{t-1},M^t,M^{t+1},...\}$, where $M^t \in \mathbb{R}^{2\times H\times W}$ (where H is 6, W is 22, and 2 corresponds to Oxy and Deoxy channels), the encoder outputs a sequence of embedding vectors $\mathbf{e} = \{...,e_{t-1},e_t,e_{t+1},...\}$, where

$$e_t = CNN(M^t). \quad (3)$$

In other words, the output $e_t \in \mathbb{R}^E$ is the embedding vector of M^t, which is the fNIRS data at time step t. In this paper, we set E as 16. The detailed structure of the CNN encoder is shown in Fig. 4b. Every convolution layer has the same Conv→Relu→Batch Normalization structure. Conv1, conv2 and conv3 have 32, 64 and 16 filters, respectively.

2) BiGRU Attention Module: We use BiGRU to process the temporal information. As mentioned above, the BiGRU consists of a forward and backward GRU unit. The GRU unit is similar to LSTM, but is more concise and involves less parameters [37]. The structure of the GRU unit is shown in Fig. 4c. It can be represented as follows:

$$\begin{aligned} r_t &= \sigma(w_{ir}e_t + b_{ir} + w_{hr}h_{t-1} + b_{hr}) \\ z_t &= \sigma(w_{iz}e_t + b_{iz} + w_{hz}h_{t-1} + b_{hz}) \\ n_t &= tanh(w_{in}e_t + b_{in} + r_t \odot (w_{hn}h_{t-1} + b_{hn})) \\ h_t &= (1 - z_t) \odot n_t + z_t \odot h_{t-1}, \end{aligned} \quad (4)$$

where e_t denotes the embedding features from the CNN encoder and is the input of the GRU unit at time t; w and b are the weight and bias of the fully connected (FC) layers inside the GRU, respectively; σ is the sigmoid function and $\odot$ represents element-wise multiplication. All the previous information is saved in the hidden state h_{t-1}. Reset gate r_t controls how much h_{t-1} is ignored in new state n_t. Update gate z_t controls the weight of the new state and previous hidden state when outputting the current hidden state h_t. We use $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to represent the hidden state calculated from forward sequence and backward sequence, respectively. The hidden state of BiGRU can be represented as:

$$\mathbf{h}_t = \{\overrightarrow{h_t},\overleftarrow{h_t}\}. \quad (5)$$

Moreover, we apply a self-attention mechanism to weigh the sequential features. Self-attention allows calculating a weighted mapping between $\mathbf{h}_t$ and $\{\mathbf{h}_s|s=1,...,T\}$, where T is the length of input sequence, $\mathbf{h}_t$ and $\mathbf{h}_s$ are the `Query` and `Key` to guide the weighted `Value`. In self-attention, `Key` and `Value` share the same value. $\mathbf{h}_T$ is the last hidden state. $\mathbf{h}_s$ is the sequence of all hidden states at each time step. Self-attention can be written as:

$$\begin{aligned} \hat{s} &= \sum_{s=1}^{T} a_s\mathbf{h}_s, \quad \text{where} \\ a_s &= \beta\frac{exp(score(\mathbf{h}_T,\mathbf{h}_s))}{\sum_{s'=1}^{T} exp(score(\mathbf{h}_T,\mathbf{h}'_s))} \quad \text{and} \\ score(\mathbf{h}_T,\mathbf{h}_s) &= \mathbf{h}_T^T\mathbf{h}_s. \end{aligned} \quad (6)$$

$\hat{s}$ is the re-scaled attention output, which includes the semantic information from the input sequence, and it is the input of the classifier for the final classification. $\beta = \frac{1}{\sqrt{T}}$ is a scaling factor.

3) Self-Supervised Label Augmentation: In this section, we will first review the self-supervised label augmentation (SLA) [20], and then introduce our learning framework, which modifies and adopts SLA providing a self-supervision signal to multi-branch and joint classifier schemes.

SLA is an effective method to learn a single joint label with respect to primary- and pre-tasks. It uses a joint distribution to combine supervised and self-supervised signals together. More specifically, the supervised signal comes from the ground truth labels of inputs, and self-supervised signal comes from the transformations applied on the inputs. SLA lets the model learn a joint distribution of semantic labels and transformations, and finally expends the original semantic labels. For example, let's assume that a primary classification task has C classes, and a pre-task applies N transformations. Then, using the SLA method, the joint probability distribution on all possible combinations has $C \times N$ labels. Rotation, patches and color shift are common techniques to generate pre-task labels in self-supervision. Lee et al. [20] use rotation (4 transformations) and color permutation (6 transformations) to generate pre-task labels. In our case, this task is not as straightforward as applying these transformations directly to fNIRS data. Rotation and permutation of the data change the spatial arrangement of the channel locations, and was shown to degrade performance in an emotion classification task [41]. Instead, we apply a different transformation. As mentioned above, during fNIRS data collection, there is a controlled rest before every task. We sampled both the controlled rest data and task data, and concatenated them along the time axis in different orders as shown in Fig. 5. In other words, $\mathbf{M_{SLA}} \in \{(CR||TASK), (TASK||CR)\}$, where $(||)$ denotes the concatenation operation. Thus, the pre-task in our case is to classify the order of the input sequence (cr-first or task-first, thus 2 transformations).

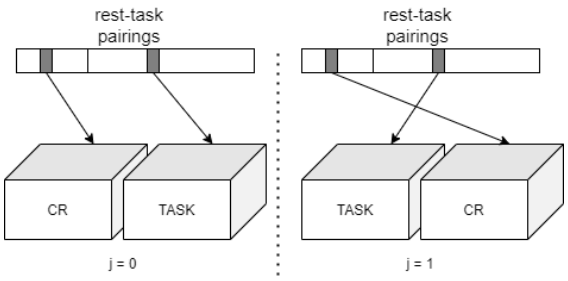

Fig. 5: Input augmented data $\mathbf{M_{SLA}}$ utilizing order based augmentation. j is the pre-task label.

One can modify Eq. (1) to multi-branch SLA objective with self-supervision, L_{MB_SLA}, as follows:

$$\begin{aligned} L_{MB\_SLA}(\mathbf{M_{SLA}}, y_{wml}, y_{vpl}; \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \\ L_{CE}(\sigma(f(\mathbf{M_{SLA}}; \boldsymbol{\theta}); \boldsymbol{\mu}), (y_{wml}, j)) \\ + L_{CE}(\sigma(f(\mathbf{M_{SLA}}; \boldsymbol{\theta}); \boldsymbol{\nu}), (y_{vpl}, j)), \end{aligned} \tag{7}$$

where

$$j = \begin{cases} 1, & r > 0.5 \\ 0, & otherwise \end{cases} \tag{8}$$

j represents the pre-task label, and r is a random value sampled from a uniform distribution ($r \sim unif(0,1)$). $j \in \{0,1\}$ where 0 and 1 represent the labels for $(CR||TASK)$ and $(TASK||CR)$, respectively.

Similarly, Eq. (2) can be rewritten to represent the joint learning scheme SLA objective with self-supervision, L_{J-SLA}, as:

$$\begin{aligned} L_{J-SLA}(\mathbf{M_{SLA}}, (y_{wml}, y_{vpl}); \boldsymbol{\theta}, \boldsymbol{\xi}) \\ = L_{CE}(\sigma(f(\mathbf{M_{SLA}}; \boldsymbol{\theta}); \boldsymbol{\xi}), (y_{wml}, y_{vpl}, j)). \end{aligned} \tag{9}$$

IV. Experiments

In the following, we refer to multi-branch classifier with full-supervision (Eq. (1)) as MB, joint classifier with full-supervision (Eq. (2)) as J, multi-branch classifier with SLA (Eq. (7)) as MB_SLA, and the joint classifier with SLA (Eq. (9)) as J_SLA. We performed "leave-one-participant-out" and 10-fold cross validation to evaluate the cross-subject performance. For 10-fold cross validation, 2 or 3 subjects were separated aside for testing in each fold. Since there are 22 subjects, 2 subjects were separated for testing in 8 folds, and 3 subjects were separated for testing for 2 folds. We have trained and tested on two label schemes shown in Table I. More specifically, for the multi-level case, we used three levels $y \in \{0,1,2\}$ (corresponding to off, low and high, respectively) to represent the WML and VPL. For the binary case, we used two levels $y \in \{0,1\}$ (corresponding to off and on, respectively). The comparison of multi-branch (MB) and joint classifiers (J) without SLA is presented in Sec. IV-B. The evaluation of these classifiers with SLA is presented in Sec. IV-C. When evaluating the performance of SLA, a single inference $P(y|\mathbf{M_{SLA}}, j)$ is run N times and the self-supervision scores are aggregated. In our experiments, we applied 2 order transformations as described above. Thus, $P(y|\mathbf{M_{SLA}}) = P(y|\mathbf{M_{SLA}}, j=0) + P(y|\mathbf{M_{SLA}}, j=1)$.

A. Setup

The duration of each task type varies in length with all tasks lasting longer than 60 seconds. When sampling the data, we use different sliding window step sizes in order to obtain balanced data such that different tasks have the same amount of data samples. The upper bound of the step size is set as 25 frames. Each sample has a duration of 50 frames (corresponding to 5 seconds). Each of the 22 participants joined two sessions. We obtained 6,641 samples from session 1 and 6,284 samples from session 2, resulting in a total of 12,925 samples.

The input of the fully-supervised models contains only task data, while self-supervised models use control rest and task data with different order transformations. Thus, the lengths of the inputs of the fully-supervised and self-supervised models are 50 frames and 100 frames, respectively.

We use cross-entropy as the loss function, and employ Adam optimizer with a learning rate of 1e-4 and weight decay of 2e-5. The embedding dimension of CNN encoder is set to be 16, and the hidden state dimension of BiGRU is 8. We

monitor the evaluation loss, and if the loss does not decrease for 5 epochs, training is stopped. Unlike the fNIRS work in [53], which uses different hyperparameters for different subjects/folds, we use the same set of hyperparameters (learning rate, number of layers and filters etc.) for different folds.

Initially, we performed the experiments by taking the same number of samples from each task. It should be noted that, depending on what type of load each task incurs (WML and/or VPL), the label balance is not always guaranteed when task balancing is applied. For instance, under the binary-level multi-labeling scheme (Table I(b)), WML has balanced label data while VPL does not (more tasks elicit VPL with more samples of '1'). Similarly, for multi-level multi-labeling scheme (Table I(a)), both WML and VPL do not have balanced data when the same number of samples are taken from each task (WML will have more samples of '0' and VPL will have more samples of '1'). After presenting the results in Sections IV-B and IV-C with this way of sampling, we also perform label balancing and present its results and a comparison with task-balanced data in Sec. IV-D.

B. Comparison between Fully-Supervised Multi-Branch and Joint Classifiers

While the MB classifier learns two independent distributions, ρ_{wml} and ρ_{vpl}, J learns a single joint distribution $\rho_{(wml,vpl)}$. For J, we calculate the accuracy (from the joint estimate) for the Working Memory Load (WML) as follows:

$$\frac{\sum_y^C (TP + TN | vpl = y)}{\sum_y^C (TP + TN + FP + FN | vpl = y)}. \quad (10)$$

Similarly, the accuracy of Visual Perception load (VPL) is calculated by:

$$\frac{\sum_y^C (TP + TN | wml = y)}{\sum_y^C (TP + TN + FP + FN | wml = y)}, \quad (11)$$

where C is the number of cognitive load levels. The results of multi-level and binary-level classification are summarized in Table II. As can be seen, with the binary-level multi-labeling scheme, MB and J have similar performance. MB has a slightly higher accuracy (0.8956) for WML, while J has a higher accuracy (0.8352) on VPL. With multi-level multi-labeling scheme, MB provided higher accuracy for both WML and VPL with average accuracy values of 0.757 and 0.767, respectively. The confusion matrices for this experiment are shown in Table III. As can be seen, MB and J have similar confusion matrices. For both WML and VPL, "off" is misclassified as "on" more often compared to "on" being classified as "off". With VPL, the classification accuracy for "off" is lower than "on", and this can be explained by the fact that there are more samples for "on" data, since we took equal number of samples from each task as explained above.

With the multi-level multi-labeling scheme, the task becomes more difficult, since it is not a binary decision anymore, and the levels of WML and VPL are also decided. From the confusion matrix of WML, we can see that, the most common error is "low" being misclassified as "high". As for the VPL, the most common error is "high" being miss-predicted as "off". Again, for WML there are not as many samples of "low" and "high" as that of "off". For VPL, there are not as many samples of "high" and "off" as that of "low" due to the task-balancing based sampling described above.

Classifier	WML	VPL
MB	0.8956	0.8129
Joint (J)	0.8935	0.8352
MB-SLA	0.8949	**0.9038**
J-SLA	**0.8966**	0.8837

(a) Binary-level Multi-label Scheme

Classifier	WML	VPL
MB	0.7575	0.7670
Joint (J)	0.7569	0.7569
MB-SLA	0.8065	0.7912
J-SLA	**0.8221**	**0.8221**

(b) Multi-level Multi-label Scheme

TABLE II: Average accuracy of leave-one-participant-out cross validation across 22 participants for binary and multi-level labeling schemes. MB and J represent multi-branch and joint classifiers with full-supervision, respectively, while MB-SLA and J-SLA refer to classifiers incorporating SLA.

		WML		VPL	
		P:off	P:on	P:off	P:on
MB	T:off	0.86	0.14	0.67	0.33
	T:on	0.06	0.94	0.14	0.86
J	T:off	0.83	0.17	0.68	0.32
	T:on	0.05	0.95	0.11	0.89

(a) Binary-level

		WML			VPL		
		P:off	P:low	P:high	P:off	P:low	P:high
MB	T:off	0.92	0.06	0.01	0.72	0.11	0.17
	T:low	0.11	0.46	0.43	0.01	0.93	0.05
	T:high	0.12	0.16	0.71	0.42	0.11	0.48
J	T:off	0.92	0.06	0.01	0.75	0.07	0.17
	T:low	0.12	0.42	0.47	0.01	0.92	0.06
	T:high	0.07	0.17	0.75	0.47	0.12	0.42

(b) Multi-level

TABLE III: Confusion matrix of leave-one-participant-out cross validation over 22 participants, 2 sessions each (44 data collection sessions total), with MB and J on binary- and multi-level multi-label schemes with fully supervision.

C. Comparison between Self-Supervised Multi-Branch and Joint Classifiers

We experimentally verified that our proposed MB-SLA and J-SLA both improve the performance for both multi-level and binary-level multi-labeling schemes. The results for the binary-level multi-labeling scheme are summarized in Table II(a). The self-supervised versions increase the accuracy for VPL (from 0.81 and 0.83 to 0.9 and 0.88). The results of multi-level multi-labeling scheme are shown in Table II(b). J-SLA provides the highest accuracy for both WML (0.82) and VPL (0.82), and increases the performance compared to fully supervised versions (from 0.7569 to 0.8221). The confusion matrices for the self-supervised case are shown in Table IV. The self-supervised models display similar confusion patterns as fully-supervised models, but with improvements on performance. The SLA method shows general improvements for all tasks, especially for the cases, where data for different classes were not balanced as explained above. For instance, for binary level, the data for WML is balanced while the data for VPL is not. The accuracy for "off" class (0.67) was lower than the "on" class (0.86) for VPL with MB, and it is improved to 0.79 and 0.94, respectively, with MB-SLA. Similarly, for multi-level,

the accuracy for "high" class (0.48), for VPL with MB, is improved to 0.58 with MB-SLA. The accuracy for "off" class (0.72), for VPL with MB, is improved to 0.85 with MB-SLA. Similar patterns are observed going from J to J-SLA.

		WML		VPL	
		P:off	P:on	P:off	P:on
MB-SLA	T:off	0.84	0.16	0.79	0.21
	T:on	0.05	0.95	0.06	0.94
J-SLA	T:off	0.87	0.13	0.72	0.28
	T:on	0.07	0.93	0.06	0.94

(a) Binary-level

		WML			VPL		
		P:off	P:low	P:high	P:off	P:low	P:high
MB-SLA	T:off	0.87	0.10	0.03	0.85	0.03	0.12
	T:low	0.14	0.63	0.23	0.01	0.86	0.12
	T:high	0.03	0.12	0.85	0.25	0.17	0.58
J-SLA	T:off	0.96	0.03	0.01	0.83	0.04	0.14
	T:low	0.18	0.54	0.28	0.01	0.96	0.03
	T:high	0.04	0.14	0.83	0.28	0.18	0.54

(b) Multi-level

TABLE IV: Confusion matrix of leave-one-participant-out cross validation over 22 participants, 2 sessions each (44 data collection sessions total), with MB-SLA and J-SLA on binary- and multi-level multi-labeling schemes.

		WML		VPL	
		P:off	P:on	P:off	P:on
TB	T:off	0.84	0.16	0.79	0.21
	T:on	0.05	0.95	0.06	0.94
LB	T:off	0.87	0.13	0.91	0.09
	T:on	0.04	0.96	0.13	0.87

(a) Binary-level

		WML			VPL		
		P:off	P:low	P:high	P:off	P:low	P:high
TB	T:off	0.87	0.10	0.03	0.85	0.03	0.12
	T:low	0.14	0.63	0.23	0.01	0.86	0.12
	T:high	0.03	0.12	0.85	0.25	0.17	0.58
LB	T:off	0.87	0.11	0.02	0.85	0.03	0.12
	T:low	0.09	0.68	0.23	0.02	0.87	0.12
	T:high	0.03	0.12	0.85	0.24	0.09	0.67

(b) Multi-level

TABLE V: Confusion matrix of MB-SLA model with task-balanced (TB) and label-balanced (LB) data for both binary and multi-class labeling schemes.

D. Comparison between Task-balanced and Label-balanced Data

As mentioned previously, for the above experiments, we balanced the data based on different tasks, i.e. we used the same number of data samples from each task. As seen in Table I, different tasks may incur different levels of cognitive load (for WML and VPL), and thus task balancing does not guarantee balancing of the individual labels for the WML and VPL. For instance, based on Table I (a), for the multi-class case, labels 0,1 and 2 are not balanced for either WML or VPL when task balancing is used.

Thus, we performed a new set of experiments by balancing the number of samples for different labels for binary (on-off) as well as 3-level (on, low, high) classification tasks. We present the results in this section, and compare them with the results of task-balancing. For binary, and multi-level multi-labeling, we used different sampling steps to reach a balanced distribution of labels. Since SLA models have provided better accuracy in the experiments presented in Sec. IV-C, we performed the new experiments in this section by using the MB-SLA and J-SLA models, and compared their performance on task-balanced and label-balanced data. We performed leave-one-participant-out cross validation across 22 participants, and used precision, recall, F1-score and accuracy as metrics to evaluate the performance.

For the binary-level labeling scheme, the labels for WML and VPL cannot be balanced simultaneously due to task-label relationship in Table I(b). So, the labels can be balanced for WML and VPL separately for comparison. As mentioned above, for the task-balanced data, the 'on (1)' and 'off (0)' labels were balanced for WML. Thus, in this part, we sampled the data to balance the labels for VPL, and presented all the results. For multi-level labeling, the labels are balanced for both WML and VPL. The performances obtained with task-balanced and label-balanced data are shown in Tables VI and VII for binary- and multi-level multi-label classification, respectively. The WML and VPL label ratios are listed as rounded, approximate numbers. As seen from Tables VI and VII, compared to the task-balanced data, higher performance is achieved with the label-balanced data in terms of precision, recall and F1 score for both binary- and multi-level classification of WML and VPL, and with 3.3% less data.

The confusion matrices obtained with task-balanced and label-balanced data and with MB-SLA model are shown in Table V. For binary-level classification of WML, the correct match ratio is increased, and the mismatch values are decreased. For VPL, the detection ratio of class 0 is increased, while the detection ratio of class 1 is slightly decreased. This can be explained by the imbalance in data for the task-balanced case, wherein there is much more data for class 1 (on) than class 0 (label ratio is 1:3). Thus, model can over-fit to class 1 providing higher accuracy. Label-balancing addresses this, and provides higher F1-score.

E. Comparison with other classifiers

We have also compared our proposed MB-SLA model with three types of commonly used machine learning classifiers, namely linear SVM [54], ANN [53] and CNN [53]. In terms of the number of parameters, our proposed MB-SLA model has around 30k trainable parameters, which is about 8x less than CNN-based model [53] and around 6x less than the ANN [53]. In this set of experiments, we used 10-fold cross-validation instead of leave-one-out cross validation [55]. After preprocessing, which includes bandpass filtering and z-score normalization, we use the extracted features to train SVM and ANN. To train the SVM, we used five features, namely mean, variance, skewness, kurtosis and slope, as noted in [54]. Before training SVM, since the large number of features would degrade the performance, we selected the channels with top 6 Fisher scores for the HbO and HbR signals. We also used all five types of features together by selecting the channels with top-1 and top-2 feature Fisher scores, which are denoted as all-top1 and all-top2, respectively, in Table VIII.

For training ANN, we used mean, variance, kurtosis, skewness, peak and slope [53] as features. While CNNs can

Model	Data	WML Ratio	VPL Ratio	WML				VPL			
				Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.
MB-SLA	Task-balanced	1:1	1:3	0.8989	0.8954	0.8947	0.8949	0.8757	0.8649	0.8701	**0.9038**
	Label-balanced	1:2	1:1	**0.9214**	**0.9145**	**0.9179**	**0.9361**	**0.8902**	0.8911	**0.8907**	0.8948
J-SLA	Task-balanced	1:1	1:3	0.8979	0.8969	0.8966	0.8967	0.8522	0.8305	0.8404	0.8837
	Label-balanced	1:2	1:1	0.9150	0.8986	0.9063	0.9279	0.8833	**0.8917**	0.8865	0.8896

TABLE VI: Leave-one-participant-out performance of binary-level multi-labeling scheme using SLA models. WML and VPL ratio represent the |'off'|:|'on'|. Pre, Rec, F1 and Acc. are precision, recall, F1-score and Accuracy, respectively.

Model	Data	WML Ratio	VPL Ratio	WML				VPL			
				Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.
MB-SLA	Task-balanced	2:1:1	1:2:1	0.7732	0.7836	0.7772	0.8065	0.7547	0.7653	0.7588	0.7912
	Label-balanced	1:1:1	1:1:1	**0.7984**	**0.7979**	**0.7972**	0.7994	**0.7980**	**0.7976**	**0.7968**	0.7992
J-SLA	Task-balanced	2:1:1	1:2:1	0.7871	0.7745	0.7748	**0.8221**	0.7871	0.7745	0.7748	**0.8221**
	Label-balanced	1:1:1	1:1:1	0.7796	0.7821	0.7791	0.7841	0.7796	0.7821	0.7791	0.7841

TABLE VII: Leave-one-participant-out performance of multi-level multi-labeling scheme using SLA models. WML and VPL ratios represent |'off'|:|'low'|:|'high'|. Pre, Rec. F1 and Acc. are precision, recall, F1-score and accuracy, respectively.

Linear SVM	WML				VPL			
FeatureTypes	Accu	Precision	Recall	F1	Accu	Precision	Recall	F1
mean	0.6832	0.6814	0.6856	0.6800	0.6852	0.6849	0.6875	0.6823
variance	0.3403	0.5244	0.3383	0.2590	0.3391	0.3336	0.3483	0.2603
skew	0.4159	0.4129	0.4167	0.4115	0.4161	0.4132	0.4169	0.4117
kurtosis	0.3372	0.3354	0.3349	0.3293	0.3368	0.3348	0.3344	0.3278
slope	0.3402	nan	0.3333	nan	0.3402	nan	0.3333	nan
all-top1	0.6649	0.6877	0.6655	0.6618	0.6649	0.6877	0.6654	0.6619
all-top2	0.6597	0.6736	0.6601	0.6490	0.6603	0.6744	0.6607	0.6496
	WML				VPL			
Model	Accu	Precision	Recall	F1	Accu	Precision	Recall	F1
ANN2-a	0.3424	0.3373	0.3400	0.2885	0.3408	0.3357	0.3374	0.2838
ANN2-b	0.3387	0.3354	0.3350	0.3224	0.3374	0.3340	0.3330	0.3141
ANN2-c	0.3391	0.3347	0.3370	0.3082	0.3415	0.3368	0.3384	0.3054
CNN1-a	0.7079	0.7090	0.7087	0.7089	0.7195	0.6871	0.6963	0.6906
CNN1-b	0.7232	0.7241	0.7278	0.7256	0.6862	0.7204	0.7298	0.7237
MB-SLA(Ours)	**0.7753**	**0.7741**	**0.7765**	**0.7742**	**0.7730**	**0.7722**	**0.7742**	**0.7741**

TABLE VIII: Multi-level classification results of linear SVM [54], ANN [53], CNN [53] and our proposed MB-SLA model with 10-fold cross validation. For linear SVM, "nan" in slope means that all samples of test splits were classified to a single level.

	WML				VPL			
Model	Accu	Precision	Recall	F1	Accu	Precision	Recall	F1
SVM mean feature	0.5934, 0.773	0.582, 0.7808	0.5963, 0.7749	0.5829, 0.7771	0.5947, 0.7757	0.5886, 0.7812	0.6012, 0.7738	0.5865, 0.7781
CNN1-b	0.6363, 0.8101	0.6373, 0.8109	0.6351, 0.8205	0.6371, 0.8141	0.6019, 0.7705	0.6371, 0.8037	0.6513, 0.8083	0.6373, 0.8101
MB-SLA	**0.6855, 0.8655**	**0.6747, 0.8735**	**0.6872, 0.8658**	**0.6771, 0.8713**	**0.6825, 0.8651**	**0.6759, 0.8635**	**0.6879, 0.8605**	**0.6783, 0.8699**

TABLE IX: The 95% (p=0.05) confidence interval of multi-level multi-label classification with 10-fold cross validation.

perform both feature extraction and classification, SVM and ANN focus on single-label classification. Thus, we trained two separate classifiers for WML and VPL. We used grid search to find the best hyperparameters (C of linear SVM, learning rate and batch size of ANN and CNN) for each fold of the 10-fold cross validation. We report the performance of traditional machine learning methods and our proposed model with 10-fold cross validation in Table VIII, where CNN1 and ANN2 are the best performing models reported in [53]. As shown in Table VIII, our proposed MB-SLA model provides the best performance in terms of all four metrics, namely accuracy, precision, recall and F1-score, and for both WML and VPL. We also calculated the 95% (p=0.05) confidence intervals under 10-fold cross-validation for SVM, CNN1 and our MB-SLA model (since ANN2 model performance was low on our dataset). The confidence intervals are shown in Table IX. Again, our proposed MB-SLA model has the highest interval values compared to linear SVM (using mean features) and CNN1-b. In addition, we performed permutation tests under 10-fold cross validation scheme, and trained 1,000 models with randomized labels. We have a total of 8 metrics to evaluate the performance, and only WML-precision ($p = 0.008$) and VPL-precision (p=0.001) have p-values larger than 0, indicating that our proposed model is functional.

F. Algorithm Transparency

Deep learning models are hard to interpret in general. We use a simple yet effective method to investigate the contributions of different channels by blocking some channels and recording the performance change. Yellow circles in Fig. 1 represent the mask locations, and for each mask position (0 to 19), the 4 surrounding channels are blocked. The performance change in terms of F1-scores, for each mask position, is presented in Fig. 6, which shows that the performance drop is more significant for mask positions 12, 13, 14, 17 and 18 (positions under the confidence low-bound of 95%), indicating that the channels surrounding these positions play more important roles than others. To visualize these results on the brain, we use the virtual spatial registration process of Tsuzuki et al. [56] to register the 52 channel locations onto a stereoscopic human brain, and obtain the Brodmann area (BA) mappings. This

mapping showed that the most influential region is BA 10, followed by BAs 47, 46 and 45. This makes sense since BA 10 is the frontopolar area, which is perhaps the most well known region being involved in cognitive load and executive function [57]. BAs 45, 46, 47 are often referred to as "Broca's complex" [58]. There is a strong link between Broca's areas and the phonological loop, where information is rehearsed in working memory, with studies of verbal working memory regularly implicating Broca's area as a part of the phonological loop, particularly in the articulatory rehearsal component [59]. This makes sense as the EWM and n-back tasks would have involved this verbal working memory rehearsal. The finding that our model gleaned maximum information from the frontopolar and Broca's complex is promising to see, as those regions would ideally be heavily involved in tasks that involved VPL and WML, which is what our model was trained to predict. We overlay the most informative mask positions on these BAs in Fig. 7.

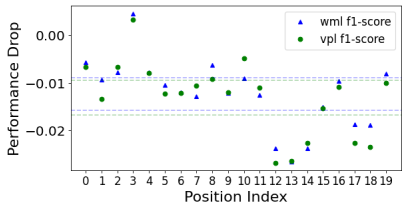

Fig. 6: Horizontal dash lines are the 95% confidence intervals. The most impactful positions are 12, 13, 14, 17 and 18.

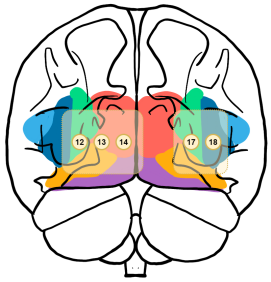

Fig. 7: The overlay of mask positions 12, 13, 14, 17 and 18 on the Brodmann areas 10, 45, 46 and 47, which are shown in red, dark blue, green and yellow respectively.

V. DISCUSSION

We have introduced a new method of modeling a large, cross-participants, cross-session set of high density functional near-infrared spectroscopy (fNIRS) data. We have proposed a new model to classify different levels of Working Memory Load (WML) and Visual Perceptual Load (VPL). In contrast to existing work, we go beyond classification of general workload (WL), and can further delineate WL into its specific sub-components, such as WML and VPL, and their different levels. Our proposed method is based on Bi-Directional Gated Recurrent Unit with an attention mechanism. We have also incorporated the self-supervision signal to the cognitive classification task by introducing a new transformation, which uses different ordering of the controlled rest and task data, and is more suitable for fNIRS data. We have shown that the spatio-temporal fNIRS data can be used by our model to make near real-time predictions of WML *plus* VPL. This is important as much machine learning work on fNIRS data to date has been done primarily on single-trial time segments ranging from roughly 40-60 seconds, and real-time adaptive systems need to select actions in shorter time periods than that. We have performed not only binary but also multi-level (off, low, high) classification of WML and VPL. With those two pieces of information, an adaptive system can choose not only whether or not a user needs support, but also what modality (auditory like through speakers or visually as on a monitor) to provide that support in. Future work should expand on this work, and increase the scale of the dataset by involving more participants and using more cognitive benchmark tasks in these studies. Cognitive benchmark tasks can include those where participants complete tasks on the computer that are more ecologically valid (like browsing the web, having a meeting over zoom, playing a video game with peers, etc.).

REFERENCES

[1] Kevin Mandrick, Vsevolod Peysakhovich, Florence Rémy, Evelyne Lepron, and Mickaël Causse, "Neural and psychophysiological correlates of human performance under stress and high mental workload," *Biological psychology*, vol. 121, pp. 62–73, 2016.
[2] Haleh Aghajani, Marc Garbey, and Ahmet Omurtag, "Measuring mental workload with eeg+ fnirs," *Frontiers in human neuroscience*, vol. 11, pp. 359, 2017.
[3] Christopher D Wickens, "Multiple resources and performance prediction," *Theoretical issues in ergonomics science*, vol. 3, no. 2, pp. 159–177, 2002.
[4] Ryan McKendrick, Raja Parasuraman, Rabia Murtza, Alice Formwalt, Wendy Baccus, Martin Paczynski, and Hasan Ayaz, "Into the wild: neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy," *Frontiers in human neuroscience*, vol. 10, pp. 216, 2016.
[5] Bin Xie and Gavriel Salvendy, "Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments," *Work & stress*, vol. 14, no. 1, pp. 74–99, 2000.
[6] Stephen H Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2009.
[7] Rebecca L Charles and Jim Nixon, "Measuring mental workload using physiological measures: a systematic review," *Applied ergonomics*, vol. 74, pp. 221–232, 2019.
[8] Ryan McKendrick, Bradley Feest, Amanda Harwood, and Brian Falcone, "Theories and methods for labelling cognitive workload: Classification and transfer learning," *Frontiers in human neuroscience*, vol. 13, pp. 295, 2019.
[9] Anne-Marie Brouwer, Thorsten O Zander, Jan BF Van Erp, Johannes E Korteling, and Adelbert W Bronkhorst, "Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls," *Front. in Neurosci.*, vol. 9, pp. 136, 2015.
[10] Sandra G Hart and Christopher D Wickens, "Workload assessment and prediction," in *Manprint*, pp. 257–296. Springer, 1990.
[11] Alan Gevins, Michael E Smith, Harrison Leong, Linda McEvoy, Susan Whitfield, Robert Du, and Georgia Rush, "Monitoring working memory load during computer-based tasks with eeg pattern recognition methods," *Human factors*, vol. 40, no. 1, pp. 79–91, 1998.
[12] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, no. 2, pp. 387–399, 2011.
[13] Etienne Combrisson and Karim Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of neuroscience methods*, vol. 250, pp. 126–136, 2015.
[14] Sandra G Hart and Lowell E Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.
[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int'l Conf. on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in *Proc. of the IEEE Int'l Conf. on computer vision*, 2015, pp. 1422–1430.

[17] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[18] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer, "S4l: Self-supervised semi-supervised learning," in *Proc. of the IEEE/CVF Int'l Conf. on Computer Vision*, 2019, pp. 1476–1485.

[19] Jiaze Sun, Binod Bhattarai, and Tae-Kyun Kim, "Matchgan: a self-supervised semi-supervised conditional generative adversarial network," in *Proceedings of the Asian Conf. on Computer Vision*, 2020.

[20] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin, "Self-supervised label augmentation via input transformations," in *Int'l Conf. on Machine Learning*. PMLR, 2020, pp. 5714–5724.

[21] Robert M Yerkes, John D Dodson, et al., "The relation of strength of stimulus to rapidity of habit-formation," *Punishment: Issues and experiments*, pp. 27–41, 1908.

[22] Angelika Dimoka, "How to conduct a functional magnetic resonance (fmri) study in social science research," *MIS quarterly*, pp. 811–840, 2012.

[23] Raja Parasuraman and Matthew Rizzo, *Neuroergonomics: The brain at work*, Oxford University Press, 2008.

[24] B Chance, Z Zhuang, Chu UnAh, C Alter, and L Lipton, "Cognition-activated low-frequency modulation of light absorption in human brain.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 8, pp. 3770–3774, 1993.

[25] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield, "Building predictive models of emotion with functional near-infrared spectroscopy," *Int'l Journal of Human-Computer Studies*, vol. 110, pp. 75–85, 2018.

[26] David A Boas, Clare E Elwell, Marco Ferrari, and Gentaro Taga, "Twenty years of functional near-infrared spectroscopy: introduction for the special issue," 2014.

[27] Erin T Solovey, Felix Putze, et al., "Improving hci with brain input: Review, trends, and outlook," *Foundations and Trends® in Human-Computer Interaction*, vol. 13, no. 4, pp. 298–379, 2021.

[28] Felix Putze, Sebastian Hesslinger, Chun-Yu Tse, YunYing Huang, Christian Herff, Cuntai Guan, and Tanja Schultz, "Hybrid fnirs-eeg based classification of auditory and visual perception processes," *Frontiers in neuroscience*, vol. 8, pp. 373, 2014.

[29] Daniel Afergan, Evan M Peck, Erin T Solovey, Andrew Jenkins, Samuel W Hincks, Eli T Brown, Remco Chang, and Robert JK Jacob, "Dynamic difficulty using brain metrics of workload," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Syst.*, 2014, pp. 3797–3806.

[30] Ziheng Wang, Ryan M Hope, Zuoguan Wang, Qiang Ji, and Wayne D Gray, "Cross-subject workload classification with a hierarchical bayes model," *NeuroImage*, vol. 59, no. 1, pp. 64–69, 2012.

[31] Norberto Eiji Nawa and Hiroshi Ando, "Classification of self-driven mental tasks from whole-brain activity patterns," *PloS one*, vol. 9, no. 5, pp. e97296, 2014.

[32] Leanne Hirshfield, Phil Bobko, Alex Barelka, Natalie Sommer, and Senem Velipasalar, "Toward interfaces that help users identify misinformation online: Using fnirs to measure suspicion," *Augmented Human Research*, vol. 4, no. 1, pp. 1, 2019.

[33] Aurélien Appriou, Andrzej Cichocki, and Fabien Lotte, "Towards robust neuroadaptive hci: exploring modern machine learning methods to estimate mental workload from eeg signals," in *Extended Abstr. of the CHI Conf. on Human Factors in Computing Syst.*, 2018, pp. 1–6.

[34] Yichuan Liu, Hasan Ayaz, and Patricia A Shewokis, "Multisubject "learning" for mental workload classification using concurrent eeg, fnirs, and physiological measures," *Frontiers in human neuroscience*, vol. 11, pp. 389, 2017.

[35] Christian Mühl, Camille Jeunet, and Fabien Lotte, "Eeg-based workload estimation across affective contexts," *Frontiers in neuroscience*, vol. 8, pp. 114, 2014.

[36] Alexander Grushin, Derek D Monner, James A Reggia, and Ajay Mishra, "Robust human action recognition via long short-term memory," in *Int'l Joint Conf. on Neural Networks*, 2013, pp. 1–8.

[37] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[38] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[39] Tomoyuki Hiroyasu, Kenya Hanawa, and Utako Yamamoto, "Gender classification of subjects from cerebral blood flow changes using deep learning," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 2014, pp. 229–233.

[40] Satoru Hiwa, Kenya Hanawa, Ryota Tamura, Keisuke Hachisuka, and Tomoyuki Hiroyasu, "Analyzing brain functions by subject classification of functional near-infrared spectroscopy data using convolutional neural networks analysis," *Computational intelligence and neuroscience*, 2016.

[41] Danushka Bandara, Leanne Hirshfield, and Senem Velipasalar, "Classification of affect using deep learning on brain blood flow data," *Journ. of Near Infrared Spectroscopy*, vol. 27, no. 3, pp. 206–219, 2019.

[42] Gauvain Huve, Kazuhiko Takahashi, and Masafumi Hashimoto, "Brain activity recognition with a wearable fnirs using neural networks," in *IEEE Int'l Conf. on mechatronics and automation*, 2017, pp. 1573–1578.

[43] Johannes Hennrich, Christian Herff, Dominic Heger, and Tanja Schultz, "Investigating deep learning for fnirs based bci," in *Int'l Conf. of the IEEE EMBC*, 2015, pp. 2844–2847.

[44] Thi Kieu Khanh Ho, Jeonghwan Gwak, Chang Min Park, and Jong-In Song, "Discrimination of mental workload levels from multi-channel fnirs using deep leaning-based approaches," *IEEE Access*, vol. 7, pp. 24392–24403, 2019.

[45] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies," *Human brain mapping*, vol. 25, no. 1, pp. 46–59, 2005.

[46] Nilli Lavie, "Perceptual load as a necessary condition for selective attention.," *Journal of Experimental Psychology: Human perception and performance*, vol. 21, no. 3, pp. 451, 1995.

[47] AM Owen, KM McMillan, AR Laird, and EN Bullmore, "back working memory paradigm: a meta–analysis of normative functional neuroimaging studies," *Hum Brain Mapp*, vol. 25, no. 1, pp. 46–59, 2005.

[48] Martin J Herrmann, Michael M Plichta, Ann-Christine Ehlis, and Andreas J Fallgatter, "Optical topography during a go–nogo task assessed with multi-channel near-infrared spectroscopy," *Behavioural brain research*, vol. 160, no. 1, pp. 135–140, 2005.

[49] Wesley B Baker, Ashwin B Parthasarathy, David R Busch, Rickson C Mesquita, Joel H Greenberg, and AG Yodh, "Modified beer-lambert law for blood flow," *Biomedical optics express*, vol. 5, no. 11, pp. 4053–4075, 2014.

[50] S Patro and Kishore Kumar Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," *Advances in neural information proc. systems*, vol. 27, pp. 3104–3112, 2014.

[52] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[53] Thanawin Trakoolwilaiwan, Bahareh Behboodi, Jaeseok Lee, Kyungsoo Kim, and Ji-Woong Choi, "Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution," *Neurophotonics*, vol. 5, no. 1, pp. 011008, 2017.

[54] Han-Jeong Hwang, Han Choi, Jeong-Youn Kim, Won-Du Chang, Do-Won Kim, Kiwoong Kim, Sungho Jo, and Chang-Hwan Im, "Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy," *Journal of biomedical optics*, vol. 21, no. 9, pp. 091303, 2016.

[55] Russell A Poldrack, Grace Huckins, and Gael Varoquaux, "Establishment of best practices for evidence for prediction: a review," *JAMA psychiatry*, vol. 77, no. 5, pp. 534–540, 2020.

[56] Daisuke Tsuzuki, Valer Jurcak, Archana K Singh, Masako Okamoto, Eiju Watanabe, and Ippeita Dan, "Virtual spatial registration of stand-alone fnirs data to mni space," *Neuroimage*, vol. 34, no. 4, pp. 1506–1518, 2007.

[57] Ke Peng, Sarah C Steele, Lino Becerra, and David Borsook, "Brodmann area 10: collating, integrating and high level processing of nociception and pain," *Progress in neurobiology*, vol. 161, pp. 1–22, 2018.

[58] Alfredo Ardila, Byron Bernal, and Monica Rosselli, "How localized are language brain areas? a review of brodmann areas involvement in oral language," *Archives of Clinical Neuropsychology*, vol. 31, no. 1, pp. 112–122, 2016.

[59] Corianne Rogalsky, William Matchin, and Gregory Hickok, "Broca's area, sentence comprehension, and working memory: an fmri study," *Frontiers in human neuroscience*, vol. 2, pp. 14, 2008.