K-means Clustering Machine Learning Approach Reveals Groups of Homogeneous Individuals with Unique Brain Activation, Task, and Performance Dynamics using fNIRS

Manob Jyoti Saikia

***Abstract*— Wearable functional near-infrared spectroscopy (fNIRS) for measuring brain function, in terms of hemodynamic responses, is pervading our everyday life and holds the potential to reliably classify cognitive load in a naturalistic environment. However, human's brain hemodynamic response, behavior, and cognitive and task performance vary, even within and across homogeneous individuals (with same training and skill sets), which limits the reliability of any predictive model for human. In the context of high-stakes tasks, such as in military and first-responder operations, the real-time monitoring of cognitive functions and relating it to the ongoing task, performance outcomes, and behavioral dynamics of the personnel and teams is invaluable. In this work, a portable wearable fNIRS system (WearLight) developed by the author was upgraded, and an experimental protocol was designed to image the prefrontal cortex (PFC) area of the brain of 25 healthy homogeneous participants in a naturalistic environment while participants performed n-back working memory (WM) tasks with four difficulty levels. The raw fNIRS signals were processed using a signal processing pipeline to derive the brain's hemodynamic responses. An unsupervised k-means machine learning (ML) clustering approach, utilizing the task-induced hemodynamic responses as input variables, suggested three unique participant groups. Task performance in terms of % correct, % missing, reaction time, inverse efficiency score (IES), and a proposed IES was extensively evaluated for each participant and the three groups. Results showed that, on average, brain hemodynamic response increased, whereas task performance degraded, with increasing WM load. However, the regression and correlation analysis of WM task, performance, and the brain's hemodynamic responses (TPH) revealed interesting hidden characteristics and the variation in the TPH relationship between groups. The proposed IES also served as a better scoring method that had distinct score ranges for different load levels as opposed to the overlapping scores of the traditional IES method. Results showed that the k-means clustering has the potential to find groups of individuals in an unsupervised manner using the brain's hemodynamic responses and to study the underlying relationship between the TPH in groups. Using the method presented in this paper, real-time monitoring of cognitive and task performance of soldiers, and preferentially forming small units to accomplish tasks based on the insights and goals may be helpful. The results showed that WearLight can image PFC, and this study also suggests future directions for the multi-modal body sensor network (BSN) combining advanced ML algorithms for real-time state classification, cognitive and physical performance prediction, and the mitigation of performance degradation in the high-stakes environment.**

***Index Terms*— Brain imaging, clustering, cognitive load, fNIRS, k-means, machine learning, military, performance, soldiers, unsupervised, working memory**

Manob Jyoti Saikia was with Soldier and Small Unit Ambulatory Virtual Environments Laboratory, Center for Applied Brain and Cognitive Sciences, Medford, MA 02155, USA. He is currently with Department of Electrical Engineering, University of North Florida, Jacksonville, FL 32224, USA (e-mail: manob.saikia@unf.edu).

I. Introduction

NEUROIMAGING techniques are widely used to study brain activation and have found that the neurons in the prefrontal cortex (PFC) area of the brain are associated with working memory (WM) related tasks [1], [2]. One of the well-established experimental methods in neuroscience and cognitive psychology involves the use of n-back WM tasks that can manipulate WM load levels [3]. In the n-back experiment, the WM load raises with the increasing number of items (n) to be memorized until the participant's WM capacity is reached. The increasing WM load degrades the task performance such as accuracy and reaction time, and also increases physiological arousal and heart rate [4]. Studies have shown the bilateral network activation, ventrolateral PFC (VLPFC) and dorsolateral PFC (DLPFC), and lateral and medial premotor cortices, frontal poles, dorsal cingulate, and medial and lateral posterior parietal cortices [5]. Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging modality applied for the measurement of cognitive load in naturalistic environments [6], [7]. The fNIRS measures neuronal activity by indirectly measuring changes in cerebral blood oxygenation, the blood oxygenation level-dependent (BOLD) functional image of the brain [8], similar to the functional Magnetic Resonance Imaging (fMRI), by noninvasively imaging temporal and spatial variation of the oxygenated (HbO_2) and deoxygenated (Hb) hemoglobin concentrations [9].

Functional near-infrared spectroscopy could be utilized to measure the cognitive workload of the operators in stressful critical jobs such as military operations, command and control,

air traffic control (ATC), and drone operations. Most fNIRS studies measure brain response differences in pre-selected participant groups [10], control vs experimental groups [11], [12], or under certain experimental conditions [7], [13]. These studies use statistical analysis to test predefined hypotheses and study the group differences. However, in the real-world dynamic high-stakes task situation where a group of homogeneous trained personnel is deployed, such as in the military and first-responder operations [14], a different approach is paramount.

Studying the relationship between task, performance and the hemodynamic response of the brain (TPH) is crucial, especially when fNIRS is intended to deploy in mission-critical environments. The performance of personnel involved in high-stakes tasks is not only dependent on the training they underwent, acquired skills, and physical and cognitive capability, but also it may depend on other human factors such as sleep quality, medication, lifestyle, emotion, stress, diet, physical exertion or as complex as the entire life experience [14]–[16]. Hence, the relationships between TPH could be subjective and can dynamically vary due to various unmeasurable factors.

In this work, an experimental protocol was developed as explained in Section II-B using n-back WM tasks to study the variation in the relationship between the WM load, task performance, and the brain's hemodynamic responses in the unknown sub-groups of individuals, which were automatically derived from a single homogeneous group implementing the unsupervised machine learning (ML) method. The use of a portable fNIRS system, WearLight [Fig. 1], helped to collect the brain's hemodynamic responses in the naturalistic environment while the participants performed the WM tasks. The protocol consists of four different n-back conditions (0 to 3-back) spread across 32 blocks of trials (8 blocks of each condition). The spreading was pseudo-randomized. A previously developed fNIRS system by the author was upgraded to collect the brain's signal in this experiment [Section II-A]. The fNIRS signal processing method is explained with a flowchart in Section II-C. After computing hemodynamic responses, an unsupervised k-means clustering approach presented in this paper formed unique participant groups using the brain responses [Section II-D]. The TPH relationship differences in the clusters and their characteristic were studied. Also, a new inverse efficiency score ($IES^{'}$), a performance scoring method, was introduced in Section II-E.

Using the presented method, evaluating TPH of each member in a homogeneous group in real-time and forming/reforming small units with similar or complementary exhibiting characteristics on the field may be helpful to achieve the overall best outcomes. The real-time TPH metrics may help commanders make better insights about the current situation and make decisions for the anticipated task demands. For example, with the TPH analysis, a commander might create specialized units for various tasks that are more or less cognitively and/or physically challenging (known as force management decision [17]). In addition, TPH metrics can aid leaders and trainers the useful information to create training protocols.

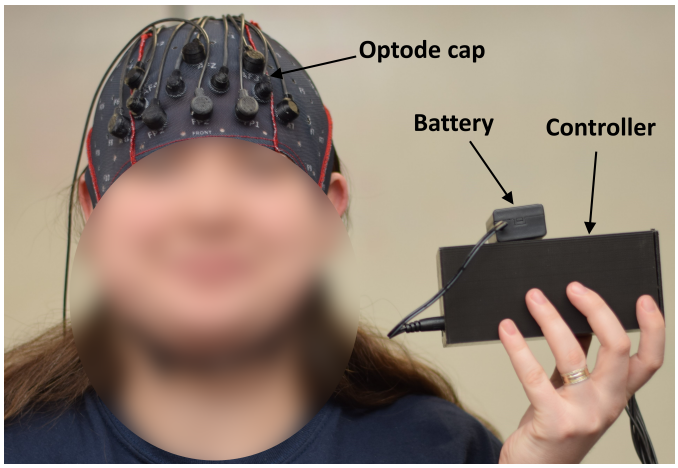

Fig. 1. WearLight fNIRS system. A user is wearing an fNIRS montage cap connected to the battery-operated controller for the brain imaging.

Results presented in Section III showed that on average hemodynamic response increased and task performance diminished with the WM load. And the k-means clustering to group participant from a homogeneous participant group was instrumental in studying the subtle differences in the relationship between experimental conditions, task performance, and the brain's hemodynamic responses in the suggested groups. The new IES also provided distinct score ranges for different cognitive load levels as opposed to the overlapping score ranges of the traditional IES method. Soldiers go through intensive training, however, the hemodynamic response, behavior, and cognitive and task performance may vary within and across soldiers on a given day depending on various factors [14], [15]. Using the method presented in this paper, real-time monitoring of the cognitive and task performance of soldiers, and dynamically forming small units to accomplish the goals of a mission could be helpful.

II. Materials and Methods

Functional near-infrared spectroscopy (fNIRS) is an optical brain imaging method based on measuring the temporal variation of oxy-hemoglobin (HbO_2) and deoxy-hemoglobin (Hb) blood concentration on the cortical surface of the brain [18], [19]. Since an fNIRS channel is formed with an NIR light source and a detector (commonly called optodes), by placing and mapping the spatial distribution of multiple of such optodes on the scalp similar to Fig. 2, it is possible to image a large cortical area of a brain. An fNIRS cap holds optodes, separates a detector from a corresponding source by a distance of about 25-45 mm, and comfortably places the optodes on the head as seen in Fig. 2 (A). The sources are sequentially turned ON and OFF at a high rate and the detectors measure the light reflected back from the cortical surface of the brain. The NIR light absorption by the HbO_2 and Hb on the cortical area is wavelength dependent and hence performing spectroscopic measurements at least at two wavelengths, fNIRS computes the ΔHbO_2 and ΔHb over time.

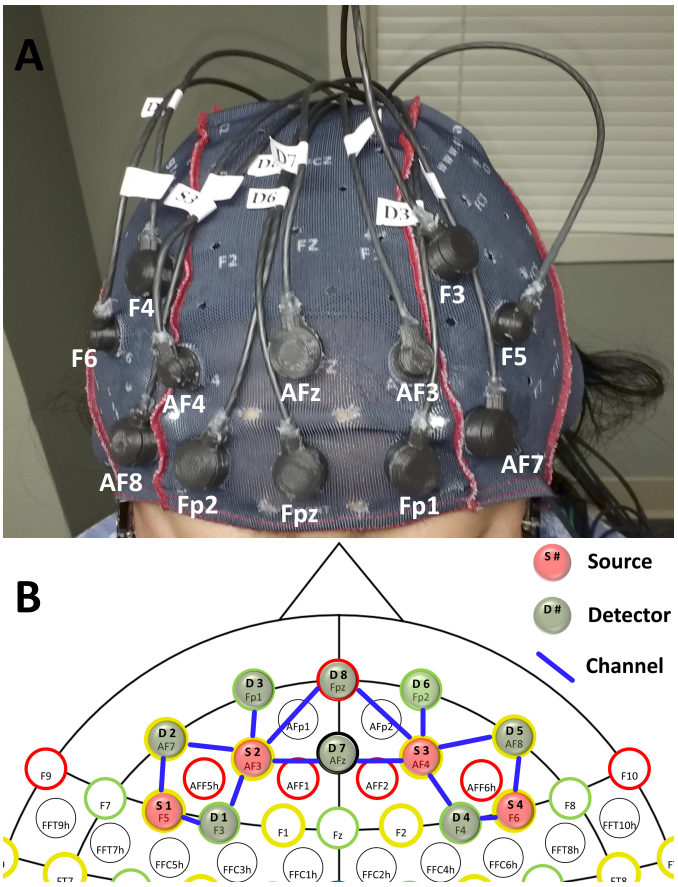

Fig. 2. **(A)** A participant wearing the fNIRS optode montage for the experimental study. **(B)** The fNIRS montage designed for the experiment. LEDs (red), detectors (green), and fNIRS channels (blue line) on a 10-20 standard EEG electrode system.

In this study, the brain activity, in terms of BOLD hemodynamic response, was continuously recorded while participants performed n-back WM tasks in an experimental protocol presented in Section II-B. In an n-back WM task, participants are sequentially presented with items (letters, numbers or patterns, etc.), and participants are engaged in remembering the previous n set of rapidly sequentially flashing items at any moment of time. The participants are asked to respond when the current item (stimulus) is the identical to the n^{th} item before the current item. The task difficulty level can be raised by increasing n, as the greater the n, the more items participants need to remember from the continuously shifting sequence of items. Thus, WM load increases with n. The participants respond using an input device such as a keypad, switch, or mouse. Both the participants' task responses and the fNIRS brain signal can be simultaneously recorded, synchronizing the data with the tasks. The participants' task performance can be measured by looking at the number of targets missed, wrong reactions, and reaction time along with the brain's hemodynamic signals due to the WM load.

A. Neuroimaging using WearLight

In this work, the brain signal recording was done using a laboratory-developed wearable continuous-wave fNIRS system. A previously developed WearLight fNIRS system [20] was upgraded for this study. A head montage was designed to accommodate four LED light sources and eight photodiode detectors to cover the prefrontal area of the brain. The LEDs emit 770 and 850 nm (peak wavelengths) NIR light. The maximum optical power was below 5 mW. The source-detector distances were about 35-45 mm. The source and detector landmark is shown in the 10-20 standard EEG electrode system in Fig. 2 (B). This montage provided 14 usable fNIRS channels (blue lines in Fig. 2 (B). Figure 2 (A) shows a participant wearing an fNIRS optode montage cap. The montage cap was interfaced with the control unit of the WearLight system. A computer, securely connected to the WearLight system via WiFi, controlled the fNIRS system and collected data. A previous graphical user interface (GUI) software [21] was customized for this study. The GUI software controled the WearLight system, displayed data in real time for signal quality checking, and saved data with a unique time-tagged file identifier. In addition, this software also recorded both TTL hardware and Software trigger to mark any event. The advantages and features of the WearLight fNIRS system in terms of hardware architecture and design, compatibility, signal quality, comfort and ease of customization, etc are presented in the previous papers [20]–[22]. Before the recording started, the optode-scalp interface was improved by carefully parting the hair underneath the optodes while assessing the signal quality on the GUI software. A computer with two displays, one for the participant and one for the experimenter, was used for the n-back task presentation. Both the n-back task data and the fNIRS data samples were time stamped, and there were additional software triggers in the GUI software; these features helped to align task performance data and the fNIRS data with the task in the post-processing phase. After a data collection session, fNIRS data and the task performance data files were securely stored in a file folder for post-processing using the data processing flowchart discussed in Section II-C.

B. Experimental paradigm

For this experiment, participants were asked to comfortably sit in front of a computer display. The display presented a stream of English alphabet letters, and participants were required to compare the currently presented letter with the letter that occurred n steps back. When the letter matched, participants responded by pressing the "0" button on the keypad with their dominant hand to identify the targeted stimulus. BCI2000 [23] controlled the task presentation and recorded the user response. The button pressing data, stimulus, and fNIRS data were time stammed in the experiment to help evaluate participants' performance, engagement level, and hemodynamic response with respect to the task. There were four different n-back task difficulty levels, 0, 1, 2, and 3-back as shown in Fig. 3 (A). In the 0-back task condition, participants were asked to respond to a single predefined target letter *"A"*. In the 1-back task condition, the target was the letter matching just the previous letter. In the 2-back and 3-back task conditions, the targets were the letters that were identical to the letter presented two and three letters before the current letter, respectively.

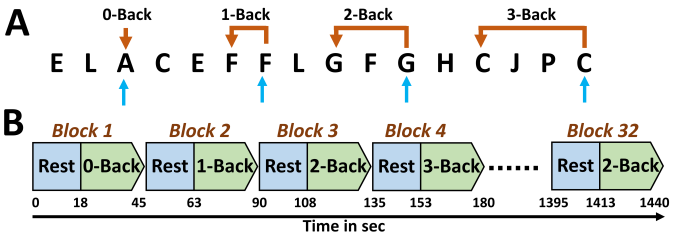

Fig. 3. Experimental paradigm. **(A)** Conceptual diagram of 0, 1, 2, and 3-back task, and **(B)** 32 task blocks, each block is either 0, 1, 2 or 3-back condition. The sequence of the four conditions was pseudo-randomized.

In total, there were 32 task blocks [Fig. 3 (B)]. Each block had a 0, 1, 2, or 3-back task condition, and the order of the four conditions was pseudo-randomized. The order of the 32 task condition blocks was [0 1 2 3 1 2 3 0 2 3 0 1 3 0 1 2 0 1 2 3 1 2 3 0 2 3 0 1 3 0 1 2]. Each block contained 3 target and 6 non-target letters. The sequence of the target and non-target letters was also randomized. In total, 288 letters appeared on the screen. Every letter was displayed for 500 milliseconds, and the screen was left blank for 2.5 seconds. Thus in 3 seconds, a new letter was presented resulting in each block lasting (9×3) 27 seconds. There were 18 seconds of *Rest* period between the task blocks, including 3 seconds for *Clue* to display information about which task condition (0, 1, 2, or 3-back) was about to start. In the *Rest* period, participants were instructed to relax so that hemodynamic activity could return to baseline. An instructional training session was conducted before the actual recording, where the task was explained and participants performed a short practice session. The entire data acquisition was for about 24 minutes as seen in the time-axis of Fig. 3 (B).

This study included 25 healthy participants (11 female and 14 male), who were right-handed, had a mean of 15 years of formal education, and ages 22-27 years (mean age = 23 years). The participants were the senior year students of STEM education and had fairly similar educational backgrounds. The participants were informed before the experimental study and gave written consent. In the recruitment process, participants completed a screening questionnaire. Criteria for inclusion included no neurological illness or no medical record of head trauma, at the time of the study no prescribed medication, fluent in English, and good eyesight. The study was in accordance with the approval of the Institutional Review Board (No. 1203762-2).

C. Data processing

The low-intensity optical measurement performed on a living body in a naturalistic setting is often affected by various artifacts, as contrast to performing experimental studies on phantoms in a controlled environment such as on a vibration-proof optical table [24]. The optodes used in this study were designed through an iterative design process to provide a suitable balance between comfortability and noise immunity [22]. However, occasional movement artifacts in the raw fNIRS signal were observed because participants, who performed the tasks in a naturalistic unconstrained manner, sometimes made abrupt movements. In addition, the combination of other

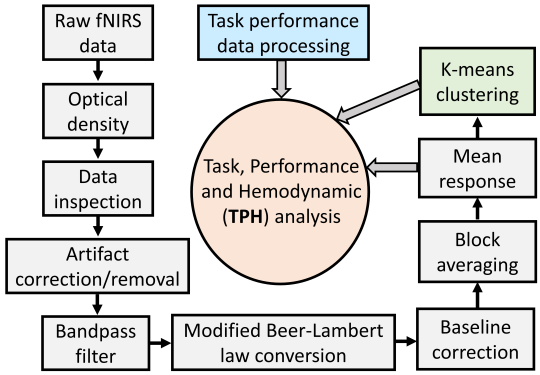

Fig. 4. Flow chart of data processing. Brain hemodynamic and task performance data were processed separately. After performing k-means clustering, the task, performance, and hemodynamic (TPH) analysis was performed.

physiological noise signals such as respiration, cardiac, and Mayer waves was present in the fNIRS signal that was filtered out later. Visually inspecting the quality of the fNIRS signal on GUI software and improving the optode-scalp interface (if required) prior to the data acquisition were emphasized for data quality control based on previous experiments.

After the data collection, fNIRS signal processing was performed using a custom-built MatLab script. The flow chart of the data processing is shown in Fig. 4. After converting the fNIRS measurements to optical density, the quality of the data was accessed. The contaminated time blocks were digitally marked and artifacts were removed. Artifacts were identified visually and through signal processing methods such as evaluating the signal amplitude above or below a threshold of 10-15 standard deviations from the mean and a threshold of 0.4 within a 6 second time period. After that, motion artifact correction was performed using principal component analysis (PCA) where a larger variation component was excluded to remove the motion artifact from the time series. Then a bandpass filter (0.01 Hz to 0.2 Hz) was applied to keep the low-frequency fNIRS signal. Using modified Beer-Lambert law (MBLL), the optical densities at wavelengths 770 and 850 nm, and their corresponding extinction coefficient values, the relative concentration changes ΔHbO_2 and ΔHb were computed. Since the fNIRS signal takes about 5 to 10 seconds to come to baseline level, with the stimulus duration of 27 seconds [Fig. 3 (B)], and combining 5 seconds of pre-stimulus and 13 seconds of post-stimulus duration, epochs of [-5 to 40] was extracted. The segmented epochs were coded with task block and task condition. Then for each fNIRS channel, a block average was performed to find the average hemodynamic response. The block averaging was computed by taking mean of the hemodynamic activity over the blocks that had the same task condition for each participant [25]. In addition, the channel mean response was computed by averaging all the fNIRS channels. Figure 5 presents the grand block averaging result, averaged over all the participants. On average, HbO_2 response increased with n, the task difficulty level. The response for the 0-back condition was minimal and hence excluded from the figures. The Hb response was in the

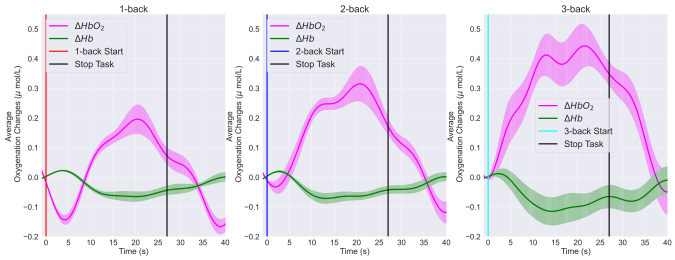

Fig. 5. Grand block averaged hemodynamic response in the 1, 2 and 3 back condition. $\boldsymbol{HbO_2}$ (solid magenta lines) and $\boldsymbol{Hb}$ (solid green lines) with $\pm$ standard error of means (SEM) bars. Vertical red, blue, cyan and black lines are the event markers.

opposite direction of the HbO_2 response, which was expected. The average hemodynamic response over the stimulus duration was used for the k-means clustering as explained in Section II-D. The k-means clustering, and task, performance, and hemodynamic response (TPH) analysis were performed using various Python packages.

D. K-means clustering

Clustering algorithms can be applied to find unique groups from multivariate data. Some popular algorithms are k-means clustering, fuzzy c-means clustering, subtractive clustering, mountain clustering, etc. Due to simplicity and applicability, k-means clustering is widely used to group multivariate data into k clusters. The k-means clustering method iteratively minimizes a cost function to form clusters that are as compact as possible and that are as separable as possible. The distances between the clusters and the centroids of the clusters are used to construct a cost function [26].

The k-means clustering method was utilized to find unique groups of participants where the participants in a group have a similar brain hemodynamic response trend in the four WM task conditions. This was an unsupervised machine learning approach as there was no additional instruction about the participants and what characteristics to look for. The algorithm was instructed to find groups from the brain's hemodynamic responses. The mean hemodynamic responses in the four task conditions (four variables) of each participant were the input for the k-means clustering algorithm. The within-cluster sum of squares (WCSS) was plotted and the elbow method was followed to find the optimal number of clusters (k = 3) as shown in Fig. 6 (A). The unique hemodynamic responses exhibited by the three groups are visible in the 2D plots in Fig. 6 (B-D) and the 3D plot in Fig. 6 (E). Since the response was minimal for the 0-back condition, the 0-back condition was not plotted. After obtaining the three clusters (participant groups), the mean hemodynamic responses of the three groups and all the participants in one group (All) were evaluated in the four n-back task stimulations to assess the overall group differences that are shown in Fig. 6 (F). On average, the hemodynamic responses of the three groups and all the participants (All) increased with the task difficulty (0 to 3-back). However, these increments have a unique trend that is further discussed in the Results section.

E. Task performance analysis

The task performance data were analyzed for each participant individually and for each group of participants to assess the relationships between task condition, task performance, and brain hemodynamic response. The task performance data were also used to determine if participants experienced any task difficulty at various levels, gauge participant engagement, and detect any anomalies, such as participant not performing the task; the data from the invalid participation could then be excluded.

The task accuracy was calculated by counting the total occurrence of wrong reactions, the ones when a participant incorrectly responded a non-targeted letter as a targeted letter by responding on the keypad (pressing the "0" key). The reaction time was the time taken by a participant to respond to a target letter. A missing response was when a participant was unable to respond when a target letter was presented. Missing responses were not included in the calculation of accuracy and reaction time; rather they were considered as an additional variable to compute the proposed IES explained in Section II-E.1. Task performance was thus analyzed by using five variables: accuracy, missing, reaction time, traditional IES, and a proposed IES. The task performance in the four stimulus conditions for the three participant groups was obtained from the k-means clustering, and all participants together were also analyzed.

1) Inverse efficiency score (IES): Inverse Efficiency Score (IES) combines accuracy and speed to measure task performance using a single consolidating variable [13], [27]. IES is calculated by dividing the reaction time (RT) by accuracy (% correct) [(1)]. It can be viewed as the average time spent on the correct responses. Since RT is measured in milliseconds, from (1), the unit of IES is milliseconds too. If two different task conditions cause the same average RT but differ in accuracy, then the IES of the condition with the higher accuracy will be less than the IES of the condition with the lower accuracy. The lower the IES, the higher the task performance. A trade-off between accuracy and speed was observed in the cognitive task. Therefore, IES offers a better summary of performance than use of a single isolated variable (RT or % correct) as a performance metric.

$$IES = \frac{RT}{100 - \%\ error} = \frac{RT}{\%\ correct} \tag{1}$$

$$IES^{'} = \frac{RT \times \%\ missing}{100 - \%\ error} = \frac{RT \times \%\ missing}{\%\ correct} \tag{2}$$

Traditionally, target missing is not considered as a separate variable, since the missed targets are counted as inaccurate responses [27]. For this work, the relationship between the target missing and task difficulty and performance was studied. It was found that greater the task difficulty, the higher the target missing. Because of this, target missing was incorporated as a separate variable in the measurement of task difficulty and performance, as shown in the construction of $IES^{'}$ [(2)]. The proposed $IES^{'}$ can be viewed as the average time spent on the correct responses with a penalty for missing (scaling factor).

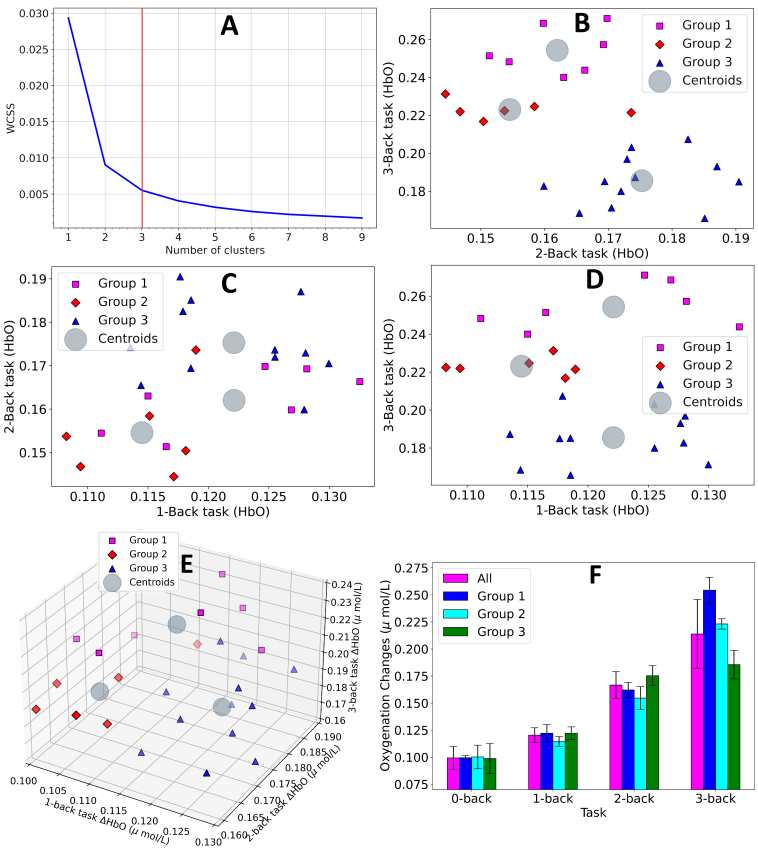

Fig. 6. **(A)** The Elbow method was used to find the optimum number of clusters (k = 3). Two dimensional scatter plots of the hemodynamic response of each group of participants (cluster): **(B)** 2-back vs 3-back, **(C)** 1-back vs 2-back, and **(D)** 1-back vs 3-back task, respectively. **(E)** The hemodynamic response of the three clusters in 3D plot. **(F)** Average hemodynamic response ($\mu \pm \sigma$) of the three groups and all the participants in All group.

III. Results

A. Task performance

Task performance in the participant groups was analyzed, and interesting differences between the groups were observed. As presented in Fig. 7, on average, the task performance of all the participants in the All group (magenta bars) degraded with the task difficulty (0 to 3-back). On average, the accuracy was 99.86%, 96.4%, 94.26%, and 90.43% for the 0, 1, 2, and 3-back tasks, respectively. The average reaction time was 465, 582, 685, and 806 ms, and the average target missing was 0.22%, 3.48%, 13.97%, and 32.76% in the 0, 1, 2, and 3-back tasks, respectively. From one-way ANOVA tests performed separately, accuracy, reaction time, and missing were found to be significantly different (each $p < 0.001$) for all participants across the four task conditions. Pairwise post hoc comparison (Tukey HSD test) showed that all the pairs of groups were also significantly different ($p < 0.001$).

The average accuracy, reaction time, and missing of the three groups followed the same trends as observed for the all groups with respect to task condition (magenta bars in in Fig. 7), however each group performed slightly different as seen in Fig. 7. The accuracy of Group 1 was consistently highest for all the conditions. Group 3 consistently had the second highest accuracy. However, in terms of reaction time, Group 1 was comparable with Group 3 for the 0, 1, and 2-back task conditions. In contrast, Group 2 consistently performed poorly in terms of accuracy, reaction time, and missing. Finally, Group 1 missed fewer targets in the 1 and 2-back tasks compared to Group 3. However, Group 1 missed more targets than Group 3 in the 3-back task.

B. Task, performance and hemodynamic (TPH)

The relationship between task-induced hemodynamic response and task performance for the three groups and for all participants together was analyzed. The average task performance and hemodynamic response due to the four task conditions are shown in the scatter plots for all participants in Fig. 8. Task performance, as measured by percent correct, reaction time, and percent missing, is on the x-axis and hemodynamic response is on the y-axis. Legend colors indicates participant group and the legend shape indicates task condition. These same plots in Fig. 8 also illustrate the relationship between task condition (legend shape), task performance (x-axis), and hemodynamic response (y-axis).

Linear regressions were performed to examine trends in the relationship between task performance (% correct, reaction time, and % missing) and the hemodynamic activity (ΔHbO_2) for the three groups and for all participants together. Linear regression lines with 95% confidence intervals are included in Fig. 8. Magenta, red and blue color lines represent Group

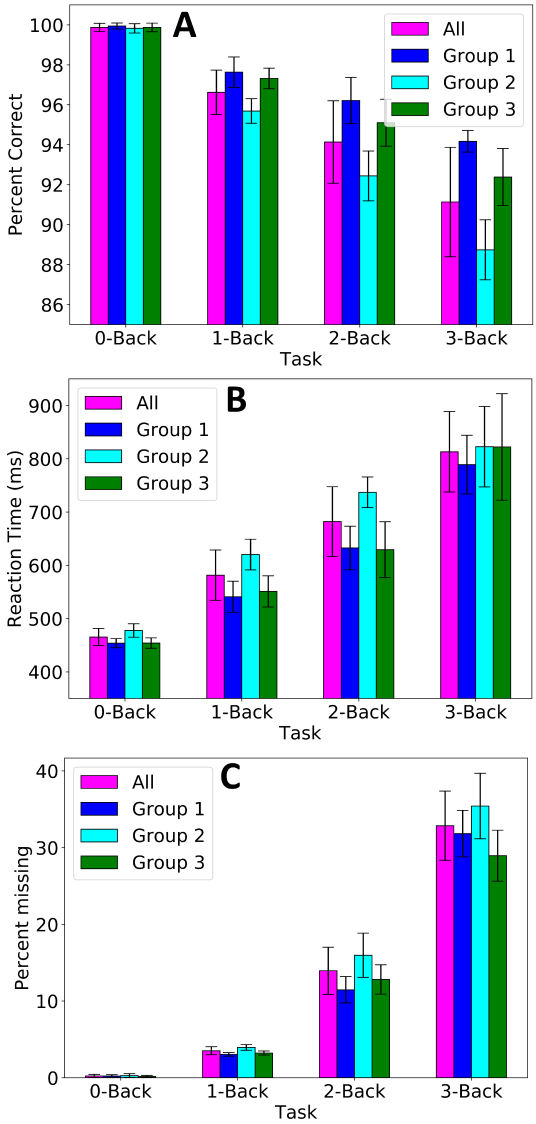

Fig. 7. Average task performance ($\mu \pm \sigma$) of the three groups and all the participants together (All) in the four task conditions (0-back to 3-back): **(A)** % correct, **(B)** reaction time, and **(C)** % missing, respectively.

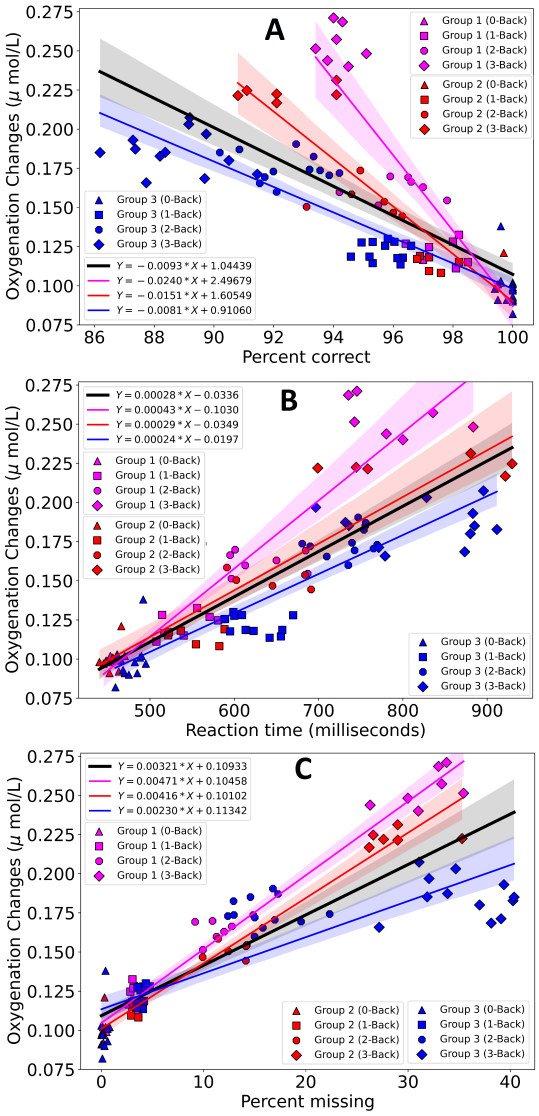

Fig. 8. Scatter plots of task performance vs hemodynamic response of all the participants: **(A)** % correct vs ΔHbO_2, **(B)** reaction time vs ΔHbO_2 , and **(C)** % missing vs ΔHbO_2 in 0-back (triangles), 1-back (squares), 2-back (circles), and 3-back (diamonds) task conditions in Group 1 (magenta), Group 2 (red), and Group 3 (blue), respectively. Four corresponding regression lines with 95% confidence intervals: magenta (Group 1), red (Group 2), blue (Group 3), and black (all the participants together) and their equations.

1, 2, and 3, respectively. The black regression lines, formed by fitting data from all participants on the same plot, are also plotted in Fig. 8 to compare all participants together with the individual groups of participants. The equations of the regression lines, also presented in Fig 8, differences in slopes and intercepts for all the lines indicate that the relationship between task performance and hemodynamic response is different for each group. For example, the sensitivities of the black, magenta, red, and blue color lines in Fig 8 (A) are -0.0093, -0.0240, -0.0151, and 0.0081 $\mu\ mol\ L^{-1}\ (\%\ correct)^{-1}$, respectively. The Pearson correlation coefficients of these regression lines are r = -0.720, -0.894, -0.909, and -0.907, respectively. Similarly in Fig 8 (B) and (C), the corresponding sensitivities of the black, magenta, red, and blue color lines are 0.00028, 0.00043, 0.00029, and 0.00024 $\mu\ mol\ L^{-1}\ (RT\ in\ ms)^{-1}$, and 0.00321, 0.00471, 0.00416, and 0.00230 $\mu\ mol\ L^{-1}\ (\%\ missing)^{-1}$, respectively. The corresponding Pearson correlation coefficients of the lines in Fig 8 (B) and (C) are r = 0.877, 0.986, 0.980, and 0.852, and r = 0.844, 0.938, 0.897, and 0.899, respectively.

In general, these plots reveal a linear relationship between task performance and task condition (0, 1, 2, and 3-back). As accuracy decreased, both reaction time and missing increased with n for all participants. This suggests that, with increasing n, participants experienced task difficulty and performance degradation. The increase in hemodynamic response (y-axis) with grater task difficulty [Fig. 8 (A-C)] is most likely due to the increased cognitive load necessary to complete more challenging tasks.

1) Inverse efficiency score (IES): Inverse efficiency score (IES) was calculated using a traditional method [(1)]. Figure 9 (A) presents inverse efficiency score vs hemodynamic response

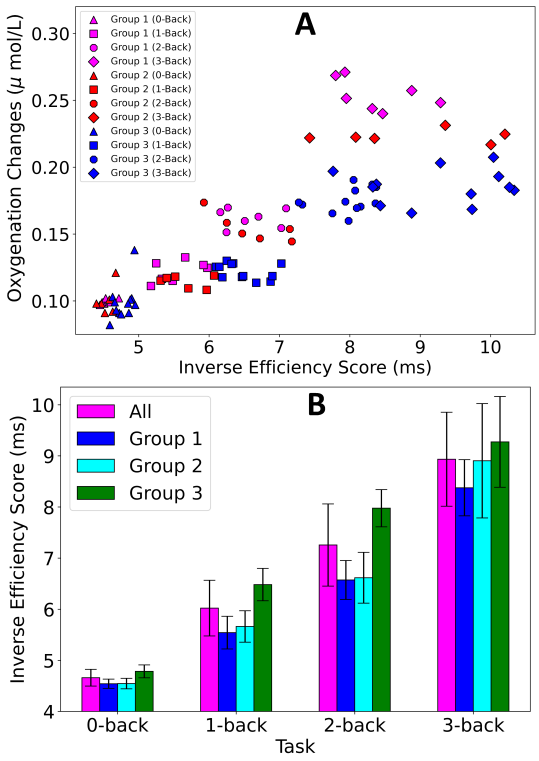

Fig. 9. **(A)** Scatter plot of Inverse Efficiency Score (IES) vs ΔHbO_2 for all participant groups in the 0-back (triangles), 1-back (squares), 2-back (circles), and 3-back (diamonds) task conditions in Group 1 (magenta), Group 2 (red), and Group 3 (blue), respectively. **(B)** Average IES ($\mu \pm \sigma$) of the three groups and all the participants in All group.

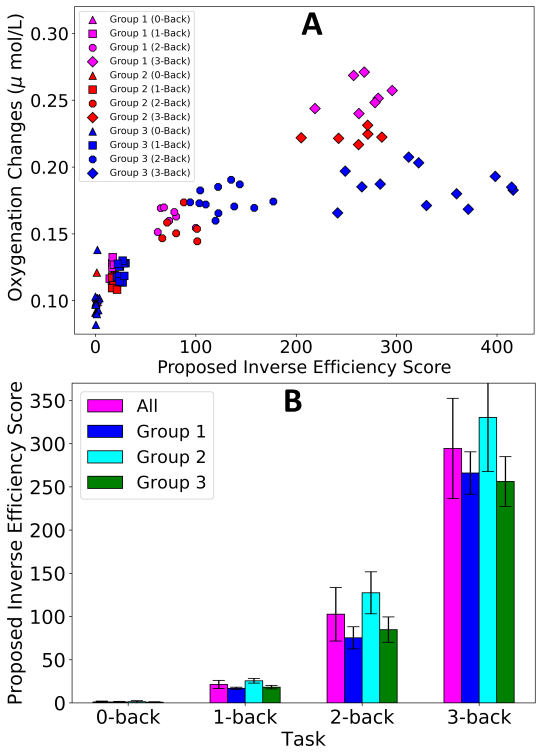

Fig. 10. **(A)** Scatter plot of proposed Inverse Efficiency Score (IES') vs ΔHbO_2 for all the participant groups in the 0-back (triangles), 1-back (squares), 2-back (circles), and 3-back (diamonds) task conditions in Group 1 (magenta), Group 2 (red), and Group 3 (blue), respectively. **(B)** Average proposed IES ($\mu \pm \sigma$) of the three groups and all the participants in All group.

for each participant. The legend shape and color are for the task condition and participant group, respectively. Figure 9 (B) presents the average IES of each group and all the participants. Both IES and hemodynamic response increased with task difficulty [Fig. 9 (A-B)].

However, there are important differences in the relationships of these variables in the group level. As shown in Fig. 9 (B), the average IES of Group 3 was the highest consistently in all four task conditions. However, on average, Group 3 had the highest hemodynamic response only in the 1 and 2-back conditions; in the 3-back condition, it decreased significantly as shown in Fig. 6 (F). In contrast, compared to Group 2 and Group 3, Group 1 had the lowest IES in all task conditions [Fig. 9 (B)] and had the highest hemodynamic response in the 3-back condition [Fig. 6 (F)]. Group 2 had highest IES standard deviation (δ_{IES}) [Fig. 9 (B)] but lowest hemodynamic response standard deviation δ_{HbO} [Fig. 6 (F)]. For Group 2, IES and hemodynamic response each seem to have a linear relationship with task difficulty. However, on average, IES and hemodynamic response and their corresponding standard deviations seem to have relationships that are more linear with task difficulty for Group 1.

2) Proposed inverse efficiency score (IES'): IES was computed using a new method [(2)] that includes target missing as a separate variable in IES construction to yield a consolidating scoring capable of detecting subtle performance differences in participants and clusters formed from relatively homogeneous participants (in our case STEM students). Figure 10 (A) and (B) present proposed IES vs hemodynamic response and average propose IES, respectively. Figure 9 (A) and Fig. 10 (A) present scatter plots of brain hemodynamic response (y-axis) and task performance in terms of IES (x-axis), but using two different methods. The comparison of the two scatter plots shows that the relationship between task performance across the four conditions and brain response of the three participant groups is more distinctly separable using the proposed IES [Fig. 10 (A)]. For example, the 3-back condition (diamonds), 2-back condition (circles), and 1-back condition (squares) have the proposed IES score ranges of [205 - 416], [62 - 177], and [14 - 30], respectively [Fig. 10 (A)]. These ranges are non-overlapping. Whereas the IES score ranges using the traditional method are [7.4 - 10.3], [5.9 - 8.4], and [5.2 - 7.0] for the same task conditions [Fig. 9 (A)]. These ranges are overlapping, and it could be harder to differentiate the task difficulty level using this scoring method for some participant. The difference between the two methods can also be visualized by comparing in Fig. 9 (B) and Fig. 10 (B), where the height of the bars changes distinctly with the task condition using the proposed IES method [Fig. 10 (B)]. Thus the new IES method seems to be more sensitive in the WM task for the homogeneous participants.

Correlation analysis was also performed to determine how the variables % correct (Acc), reaction time (RT), % missing (Miss), traditional Inverse Efficiency Score (IES), proposed IES (New IES), and hemodynamic response (HbO) are corre-

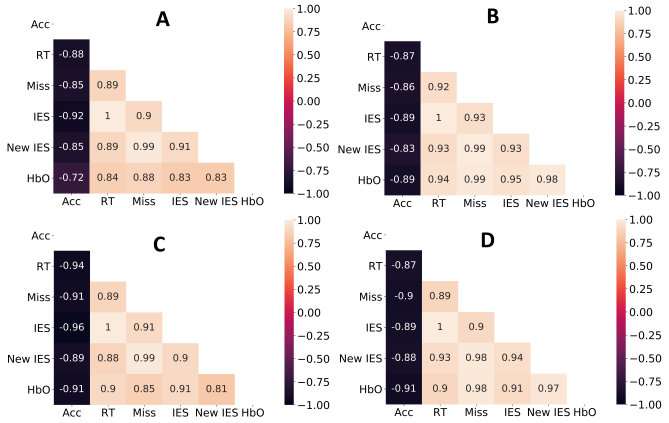

Fig. 11. Pearson correlation coefficients for the relationships between of percent correct (Acc), reaction time (RT), percent missing (Miss), traditional Inverse Efficiency Score (IES), proposed IES (New IES), and hemodynamic response (HbO): **(A)** All participants, **(B)** Group 1, **(C)** Group 2, and **(D)** Group 3.

lated with each other. As seen in Fig. 11 (A), the correlation coefficients of HbO with IES and HbO with New IES are equal, 0.83. However, IES is highly correlated (unity) with reaction time, which is also true for Groups 1 to 3. In contrast, the correlation coefficient of new IES with RT is not unity, which is desirable. This suggests that the influence of reaction time, along with the target missing and accuracy, to study the differences between participants and participant groups using the proposed new IES ($IES^{'}$) may be optimal.

IV. Discussion and Conclusion

Prefrontal cortex (PFC) activation during working memory tasks has been demonstrated by various studies. This study shows that WearLight fNIRS system can image hemodynamic response on PFC during the WM task and supports other studies. Most research has examined the relationships between hemodynamic responses, user states, and cognitive load conducted in laboratory environments using commercially available fNIRS systems. These experimental setups may yield little resemblance to real-world situations, limiting the general interpretation of the study outcomes. Also, most studies have focused on preselected participant groups (controlled vs experimental).

In this work, a laboratory-developed portable fNIRS system by the author was customized, and an experimental protocol was developed to conduct the research in a naturalistic environment. The participants in this study were a homogeneous group as they were Bachelor of Science in Engineering students with similar academic backgrounds. Their understanding of the task was also evaluated in an n-back task training and quiz session before the actual data acquisition session. This resulted in high overall task engagement and performance across all participants. However, differences in TPH were observed. A new IES method combining reaction time, missing, and accuracy demonstrated a better performance evaluation than the more traditional IES method. Results showed that target missing could be an indication of cognitive overloading, as well as a sign of higher task engagement. Also, increasing task difficulty demanded more cognitive effort, resulting in higher hemodynamic activation with (n) and diminishing task performance.

This work also describes methods to find unique groups of individuals by implementing the unsupervised machine learning approach on the brain's hemodynamic signal and to study the subtle TPH relationship differences between those groups. The TPH relationships could vary dynamically as they depend on various human factors. Evaluating TPH in real-time could help commanders make better insights, and forming small units based on the current status could lead to the overall best task outcomes.

Wearable fNIRS is a promising neuroimaging technology that can be applied in the naturalistic environment to monitor cognitive load. This could also lead to the development of predictive technology combining advanced ML algorithms that select an individual or a group of individuals with certain characteristics in terms of the relationship between task, performance, and cognitive and physical load. However, accessing the individual or team status merely from the brain's hemodynamic response may not be sufficient to predict performance outcomes. Other physiological measurements using body sensor network (BSN) and adopting a multi-modal approach could bridge the gap between state classification, cognitive and physical performance prediction, and the mitigation of performance degradation.

This work is a step towards the implementation of fNIRS in the high-stakes environment and describes a method to find subsets of trained personnel solely from the brain's fNIRS signal. This technique could be applied to form small units in real-time while the personnel are already engaged in mission-critical tasks. This work also contributed to current research in which a multi-modal approach combining different sensors and systems was implemented, and a BSN to collect data in both laboratory and field settings was created. These studies enable the measurement and prediction of a multitude of cognitive, behavioral, and physical parameters.

V. Author Disclaimer

The views expressed in this article are solely those of the author and do not reflect the official policies or positions of the Department of Army, the Department of Defense, or any other department or agency of the U.S. government.

References

[1] M. D'Esposito, B. R. Postle, D. Ballard, and J. Lease, "Maintenance versus manipulation of information held in working memory: An event-related fmri study," *Brain and Cognition*, vol. 41, pp. 66–86, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278262699910965

[2] B. Spitzer, D. Goltz, E. Wacker, R. Auksztulewicz, and F. Blankenburg, "Maintenance and manipulation of somatosensory information in ventrolateral prefrontal cortex," *Human Brain Mapping*, vol. 35, pp. 2412–2423, 2014, doi: 10.1002/hbm.22337. [Online]. Available: https://doi.org/10.1002/hbm.22337

[3] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." *Journal of Experimental Psychology*, vol. 55, pp. 352–358, 1958.

[4] R. W. Backs and K. A. Seljos, "Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task," *International Journal of Psychophysiology*, vol. 16, pp. 57–68, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167876094900426

[5] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies," *Human brain mapping*, vol. 25, pp. 46–59, 2005.

[6] R. da Silva Soares, A. Y. A. Oku, C. S. F. Barreto, and J. R. Sato, "Applying functional near-infrared spectroscopy and eye-tracking in a naturalistic educational environment to investigate physiological aspects that underlie the cognitive effort of children during mental rotation tests," *Frontiers in Human Neuroscience*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2022.889806

[7] S. Midha, H. A. Maior, M. L. Wilson, and S. Sharples, "Measuring mental workload variations in office work tasks using fnirs," *International Journal of Human-Computer Studies*, vol. 147, p. 102580, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581920301828

[8] J. Steinbrink, A. Villringer, F. Kempf, D. Haux, S. Boden, and H. Obrig, "Illuminating the bold signal: combined fmri–fnirs studies," *Magnetic Resonance Imaging*, vol. 24, pp. 495–505, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0730725X06000531

[9] M. Wolf, M. Ferrari, and V. Quaresima, "Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications," *Journal of biomedical optics*, vol. 12, p. 62104, 2007.

[10] Y. M. Han, M. C. Chan, M. M. Chan, M. K. Yeung, and A. S. Chan, "Effects of working memory load on frontal connectivity in children with autism spectrum disorder: a fnirs study," *Scientific Reports 2022 12:1*, vol. 12, pp. 1–14, 1 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-05432-3

[11] R. Li, G. Rui, C. Zhao, C. Wang, F. Fang, and Y. Zhang, "Functional network alterations in patients with amnestic mild cognitive impairment characterized using functional near-infrared spectroscopy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, pp. 123–132, 2020.

[12] V. Kumar, S. Nichenmetla, H. Chhabra, V. S. Sreeraj, N. P. Rao, M. Kesavan, S. Varambally, G. Venkatasubramanian, and B. N. Gangadhar, "Prefrontal cortex activation during working memory task in schizophrenia: A fnirs study," *Asian Journal of Psychiatry*, vol. 56, p. 102507, 2 2021.

[13] Y. Statsenko, T. Habuza, K. N.-V. Gorkom, N. Zaki, and T. M. Almansoori, "Applying the inverse efficiency score to visual–motor task for studying speed-accuracy performance while aging," *Frontiers in Aging Neuroscience*, vol. 12, p. 452, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2020.574401

[14] T. T. Brunyé, R. Brou, T. J. Doty, F. D. Gregory, E. K. Hussey, H. R. Lieberman, K. L. Loverro, E. S. Mezzacappa, W. H. Neumeier, D. J. Patton, J. W. Soares, T. P. Thomas, and A. B. Yu, "A review of us army research contributing to cognitive enhancement in military contexts," *Journal of Cognitive Enhancement*, vol. 4, pp. 453–468, 2020. [Online]. Available: https://doi.org/10.1007/s41465-020-00167-3

[15] G. E. Giles, S. Elkin-Frankston, T. T. Brunye, E. Navarro, J. F. Seay, K. L. McKenzie, S. A. Brown, J. L. Parham, T. N. Garlie, H. E. Choi-Rokas, K. B. Mitchell, K. Racicot, K. S. O'Fallon, J. W. Soares, W. R. Elmore, J. A. Cantelon, A. L. Gardony, T. J. Smith, J. P. Karl, J. M. Jayne, J. W. Ramsay, and M. D. Eddy, "Establishing a cognitive, health, physical, and social-emotional toolkit to predict soldier performance," *Medicine and Science in Sports and Exercise*, vol. 54, pp. 119–119, 9 2022.

[16] J. Vera, D. Janicijevic, S. Miras-Moreno, A. Pérez-Castilla, R. Jiménez, B. Redondo, and A. García-Ramos, "Intraocular pressure responses to a virtual reality shooting simulation in active-duty members of the spanish army: The influence of task complexity," *Physiology and behavior*, vol. 256, 11 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36070832/

[17] J. G. Wohl, "Force management decision requirements for air force tactical command and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, pp. 618–639, 1981.

[18] W. L. Chen, J. Wagner, N. Heugel, J. Sugar, Y. W. Lee, L. Conant, M. Malloy, J. Heffernan, B. Quirk, A. Zinos, S. A. Beardsley, R. Prost, and H. T. Whelan, "Functional near-infrared spectroscopy and its clinical application in the field of neuroscience: Advances and future directions," p. 724, 7 2020. [Online]. Available: www.frontiersin.org

[19] M. Althobaiti and I. Al-Naib, "Recent developments in instrumentation of functional near-infrared spectroscopy systems," *Applied Sciences*, vol. 10, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/18/6522

[20] M. Saikia, W. Besio, and K. Mankodiya, "Wearlight: Toward a wearable, configurable functional nir spectroscopy system for noninvasive neuroimaging," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, 2019.

[21] M. J. Saikia, G. Cay, J. V. Gyllinsky, and K. Mankodiya, "A configurable wireless optical brain monitor based on internet-of-things services." Institute of Electrical and Electronics Engineers Inc., 12 2018, pp. 42–48.

[22] M. Saikia and K. Mankodiya, "3d-printed human-centered design of fnirs optode for the portable neuroimaging," vol. 10870, 2019.

[23] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, pp. 1034–1043, 2004.

[24] M. J. Saikia, "A spectroscopic diffuse optical tomography system for the continuous 3d functional imaging of tissue -a phantom study," *IEEE Transactions on Instrumentation and Measurement*, p. 1, 2021.

[25] H. Santosa, F. Fishburn, X. Zhai, and T. J. Huppert, "Investigation of the sensitivity-specificity of canonical- and deconvolution-based linear models in evoked functional near-infrared spectroscopy," *Neurophotonics*, vol. 6, p. 1, 5 2019. [Online]. Available: https://www.spiedigitallibrary.org/journals/neurophotonics/volume-6/issue-02/025009/Investigation-of-the-sensitivity-specificity-of-canonical–and-deconvolution/10.1117/1.NPh.6.2.025009.full

[26] D. Pelleg and A. Moore, "Accelerating exact k-means algorithms with geometric reasoning," 1999, pp. 277–281.

[27] R. Bruyer and M. Brysbaert, "Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (ies) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)?" *Psychologica Belgica*, vol. 51, pp. 5–13, 2011.