Augmented Human Research (2020) 5:15
https://doi.org/10.1007/s41133-020-00034-y

ORIGINAL PAPER

Identification of Potential Task Shedding Events Using Brain Activity Data

Danushka Bandara[1] · Trevor Grant[2] · Leanne Hirshfield[2] · Senem Velipasalar[1]

Received: 12 November 2019 / Revised: 10 March 2020 / Accepted: 12 March 2020 / Published online: 30 March 2020

Abstract
In Human–Machine Teaming environments, it is important to identify potential performance drops due to cognitive overload. If identified correctly, they can help improve the performance of the human–machine system by offloading some tasks to less cognitively overloaded users. This can help prevent user error that can result in critical failures. Also, it can improve productivity by keeping the human operators at an optimal performance state. This paper explores a new method for identifying user cognitive load by a three-class classification using brain activity data and by applying a convolutional neural network and long short-term memory model. The data collected from a set of cognitive benchmark experiments were used to train the model, which was then tested on two separate datasets consisting of more ecologically valid task environments. We experimented with various models built with different benchmark tasks to explore which benchmark tasks were better suited for the prediction of task shedding events in these compound tasks that are more representative of real-world scenarios. We also show that this method can be extended across-tasks and across-subject pools.

Keywords Human–Machine Teaming · fNIRS · Brain data · Task shedding · Convolutional neural networks · LSTM · Classification

Introduction

As computing devices become more ubiquitous, the need for greater human–machine symbiosis becomes an important factor. This concept, of human and computer agents working together to accomplish a goal, or a set of goals, is referred to as Human–Machine Teaming (HMT). In any teaming environment, whether it is a team of all human agents, a team of all machines, or a combination of the two, it is important that resources within the team are allocated as efficiently as possible such that the team may achieve its goal while simultaneously putting the least amount of strain possible on any of the team members. The finite resources of an HMT, such as the processing power of a machine, or the limited cognitive capacity of a human agent, could be viewed as potential bottlenecks within an HMT system, where the team's ability to accomplish a task may falter. Although the processing power of a machine may have been the primary factor that stopped HMTs from achieving optimal performance in the past, processing power is now an easily obtainable resource. As such, recent efforts to improve the performance of HMTs have shifted focus to improving the communication modalities between humans and machine agents. As human agents have limited cognitive capacity within a time-sensitive task environment, ideal task performance is dependent on optimizing humans' information processing capabilities, which are affected by the complex interplay between their perceptual processing load, mental workload, and emotional state. Task performance within an HMT is also dependent on the ability of machine agents to detect and interpret the signs of potential overload on the part of the human agent, and to

✉ Danushka Bandara
dsbandar@syr.edu

Trevor Grant
Trevor.Grant@colorado.edu

Leanne Hirshfield
Leanne.Hirshfield@colorado.edu

Senem Velipasalar
svelipas@syr.edu

[1] Syracuse University, Syracuse, USA

[2] Institute of Cognitive Science, University of Colorado Boulder, Boulder, USA

have the ability to take meaningful action to assuage, or at the very least reduce, the load placed on the human agent.

For a machine agent within an HMT to properly make predictions about the current state of the system, the machine must not only possess knowledge relevant to the task that needs to be completed, but also the amount and type of *mental workload* that is currently being placed on the human agent(s) within the HMT. The machine must also be able to discern if this amount of workload is significant enough to induce degradation in the human agent's or team's performance that might limit the HMT's ability to complete the current task. *Mental workload*, in this context, is the brain's finite amount of processing capacity to allocate to a given task. As theoretical and experimental work by Wickens' Multiple Resource Theory (MRT) [45] has shown, there are different types of cognitive resources that the brain can allocate simultaneously, and the overload of one type of cognitive resource does not necessarily lead to the overload of another [45]. When a person is required to perform multiple tasks that require the same type of mental resource, resource may become overloaded, and as a result, person's performance at the given task will degrade. If the overload is adequately high, the individual may eschew the task altogether, an event known as 'task shedding' [48]. A common area where this concept expressed is in the field of piloted aircraft and unmanned air vehicles (UAVs). In those cases, the autonomous system and pilot cooperate to achieve objectives [38]. For this type of cooperation to work, the machine needs to be able to sense and intervene when the human's performance starts dropping due to increases in task load. [5, 37]. Parasuraman and Hancock have shown that task shedding can be triggered by high workload and low certainty [29]. Although researchers have previously attempted to model this task shedding behavior using simple and basic tasks [26, 27, 30, 33] as well as more complex real-world scenarios [6, 23], more accurate predictions about when task shedding events are likely to occur are needed if these models are to be implemented in real time.

In this paper, we introduce a novel method by which one can use information about task performance and taskload metrics to tabulate a measure that we call the 'Task Shedding Index (TSX),' which is a combination of taskload and performance. The TSX can be used as an indicator of the potential of the user for task shedding. Further, we use this new proposed measure to predict task shedding instances using across-task, across-subject machine learning methods. The goal is to introduce an agnostic framework, not tied to any specific task that a particular HMT would try to complete. To achieve this goal, we created a model, using psychophysiological data recorded from a functional near-infrared spectroscopy (fNIRS) device, to detect when task load on a human participant was high, and task shedding instances were therefore likely to occur. To isolate specific types of mental workload (working memory, visuospatial attention), we trained models on multiple cognitive benchmark tasks used widely in the fields of cognitive psychology and cognitive neuroscience and tested those models on more ecologically valid tasks. We show that such a system can provide a reliable prediction of such events such that an autonomous agent with access to this type of physiological data would be able to predict when moments of mental overload might lead to performance decrements or task shedding events, and as a result would be able to take over for, or provide assistance to, the human agent within the HMT.

Background

Human–Machine Teaming and Task Shedding

The importance of finding a way for an autonomous agent in an HMT to detect when the human agent may be subjected to events of higher cognitive workload, and thus task shedding events, has been shown in past systems design research [34]. Past work in the fields of human factors and cognitive engineering has provided evidence that when a human agent's performance is supplemented with a machine agent's ability to assist on tasks, the reported workload of the human agent decreases. This decrease in perceived workload coincides with an increase in the human agent's self-confidence and trust in the HMT as well as an increase in the overall performance of the HMT [14]. Despite these advances, there are still many issues that need to be addressed to ensure HMTs can be optimized to task performance [8]. With these issues, highlighted by past research, in mind we use predictive modeling on multiple cognitive resources to detect when increases in mental workload are likely to lead to performance decrements.

Using Psychophysiological Sensors to Measure Workload

Task shedding detection through brain activity requires sensors that are robust to noise, portable and noninvasive. The fNIRS device works well for this application since the device can be set up quickly and can target specific areas of the brain that are implicated in cognitive resources that are prone to becoming overloaded when engaged in cognitively demanding tasks [15, 42]. The fNIRS device works by using multiple pairs of optodes that are placed on the scalp and pulse infrared light (690 nm and 830 nm) through the skull and into the brain. The reflected light intensity that is received by the detector is dependent on

the amount of oxygenated and deoxygenated hemoglobin in the incident area over which the optodes are placed [11]. Since oxygen is consumed during the metabolic processes involved in brain activity, the concentration of hemoglobin is correlated with increased brain activity. fNIRS has in the past been used for classification of workload levels [20]. More specifically, the fNIRS' ability to measure brain activity in the frontal cortex of the brain gives it a unique ability to predict workload levels, as greater activation in the frontal cortex is associated with higher levels of mental workload [19, 36].

The ability to measure mental states, and thus workload, has already been well documented in the human factors literature [35]. Other work has found great success in being able to predict mental workload in real-world computer environments using fNIRS [31, 41]. As advances in both fNIRS technology and portability increase, fields such as brain–computer interfacing (BCI) have argued that the fNIRS' increased spatial resolution would make fNIRS an invaluable tool for collecting real-time psychophysiological data and building systems that incorporate and adapt to that data in real time [10]. The portability of the fNIRS system, combined with its ability to measure workload and communicate those measurements to a machine agent, could provide a solution to make strides toward correcting documented issues in intelligent system designs [24].

Machine Learning Classifiers on fNIRS Data

Researchers have used traditional machine learning classifiers such as support vector machines [12], artificial neural networks [7], hidden Markov models [49] as well as other statistical methods [4, 39] to preprocess and classify fNIRS data. However, more recent work has demonstrated the ability to use deep learning algorithms [18] to capture the characteristics of the fNIRS signal. One category of deep learning algorithms, convolutional neural networks (CNNs) [25], is typically used in the image processing domain because of their ability to capture the spatial structure of image data. They are well suited for fNIRS analysis due to the same reasons. Our previous work [3] showed that the oxygenated and deoxygenated data provided by fNIRS can be used similar to the RGB channels of an image when fed into a CNN-based classifier. We also used a long short-term memory (LSTM) network, to capture the time-series behavior of the fNIRS data. LSTMs are a version of recurrent neural networks, which can capture long- and short-term dependencies in the data [21]. LSTMs have become popular for machine learning on electroencephalography data [1, 13]. They have also recently been used in fNIRS analysis [40]. In [3], the above-described model was used in across-subject classification by dividing the subject pool into folds. In this paper, we use the same LSTM model on across-subject, across-task, three-label classification tasks.

Methods

Experimental Protocol

fNIRS data were collected from 45 participants (14 females, 31 males, mean age = 26, min age = 21, max age = 36) who were selected from the undergraduate and graduate student population at a university in the Northeast United States. The data were collected using a Hitachi ETG-4000 fNIRS device at a sampling rate of 10 Hz. The optodes were arranged into an fNIRS cap with a 3×11 probe configuration and were placed on the participant's forehead area in a symmetrical manner (Fig.1). Using this configuration, the fNIRS device can capture information about the oxygenated and deoxygenated hemoglobin levels in the frontal cortex of the participants. The fNIRS was calibrated to ensure that all probes were recording proper readings and adjusted to account for ambient light. After setting up the fNIRS device on the participant's head, a Patriot Polhemus 3D digitizer device was used to measure the location of each source/detector to account for variances in head size and shape. All participants gave informed consent under the restrictions and guidelines of the University's Institutional Review Board.

From the set of 45 participants, a subset of these participants ($n = 25$, 7 females, 18 males) completed the cognitive benchmark tasks described in Sect. 3.2, as well as the triage cyber analyst task described in Sect. 3.3.1. Both the cognitive benchmark tasks and the triage task used a variable interstimulus interval (ISI) between the offset of a trial and the onset of a new trial, during which a cross-fixation point in the center of the screen was displayed on the screen. The length of the ISI was an exponential distribution (mean = 4 s, min = 2 s, max = 8 s). The order of tasks was as follows,

1. Consent
2. fNIRS sensor set up
3. Session 1 (cognitive benchmark tasks, a-d order randomized)
   (a) N-back, NASA-TLX
   (b) Visual search, NASA-TLX
   (c) Posner cueing paradigm, NASA-TLX
   (d) Words task, NASA-TLX
   (e) Simple reaction time task, NASA-TLX
   (f) Reverse Go/No-Go, NASA-TLX
4. 15-min break
5. Session 2

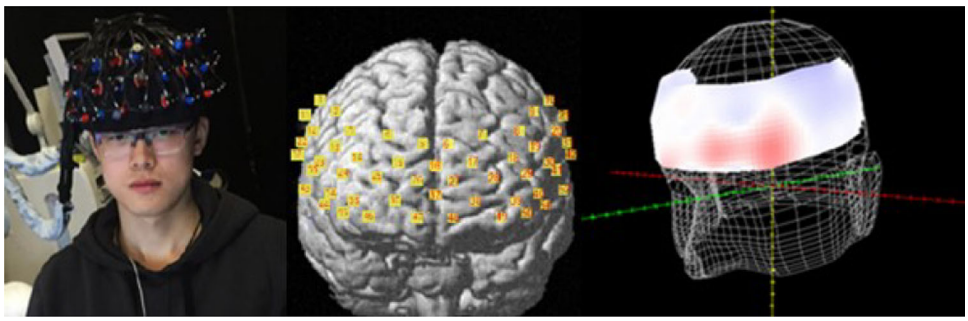

Fig. 1 The 3 × 11 fNIRS probe configuration

(a) Triage task training
(b) Triage task
(c) NASA-TLX

The remaining 20 participants performed the Multi-Attribute Task Battery (MATB) testbed described in Sect. 3.3.2. The order of MATB tasks was as follows,

1. Consent
2. fNIRS sensor set up
3. Session 1
   (a) 1-min MATB low difficulty, NASA-TLX
   (b) 1-min MATB medium difficulty, NASA-TLX
   (c) 1-min MATB high difficulty, NASA-TLX
4. 1-min break
5. Session 2
   (a) 1-min MATB high difficulty, NASA-TLX
   (b) 1-min MATB low difficulty, NASA-TLX
   (c) 1-min MATB medium difficulty, NASA-TLX
6. 1-min break
7. Session 3
   (a) 1-min MATB medium difficulty, NASA-TLX
   (b) 1-min MATB high difficulty, NASA-TLX
   (c) 1-min MATB low difficulty, NASA-TLX

Since an adjustment in screen size is correlated with certain physiological responses [32], all tasks were displayed to the participants on a 22-inch monitor with a screen resolution of 1280 × 1024 pixels. Participants were seated in a stationary chair so that the distance between their eyes and the monitor was 65cm. The participants used the computer mouse as the only form of response to all tasks (by clicking buttons or making selections). This was done to minimize noise in the data due to subject movement during the experiment. All participants would begin each fNIRS session with a 30-s session of controlled rest during which the participant fixated on a plain black plus symbol in the center of a white display. Participants would then perform ten trials of a reaction time task before they began the rest of the cognitive benchmark tasks. To mitigate the fatigue effects involved in cognitive testing [43], the benchmark task order was randomized between subjects.

Stimulus Materials: Training Data

Visual-Lexical Processing, Adaptive Words

This adaptive words task was developed to induce workload on participant's visual-lexical processing resources. In this task, the words for the numerical values of the digits one through eight were displayed vertically for a variable amount of time in the center of the screen. The participant's goal was to determine whether the word that was displayed on the screen corresponded to either an odd or even numerical value (Fig. 2).

Visual Search Task, Visual Search

The visual search task was designed to cause cognitive load on people's visual processing resources and was modeled after the task design developed by Wang, Cavanagh and Green [44]. A circular array of nine letters consisting of a distractor (backward Ns) and a target (normal facing Ns) was displayed to the participant for a variable amount of time. The participant's task was to determine whether or not the target was displayed within the array (Fig. 3).

Response Inhibition, Go/No-Go

The response inhibition task was the Go/No-Go task, which involved one target stimuli (a large blue circle) and one distractor stimuli (a large blue square). The development of stimulus materials was guided by Huettel, Mack, and McCarthy [22]. The participant was tasked with responding to the stimuli if the target was presented, and not responding to the stimuli when the distractor was presented (Fig. 4).

Working Memory, N-Back Task

The N-Back task (Fig. 5) was designed to cause cognitive load on people's working memory resources by requiring

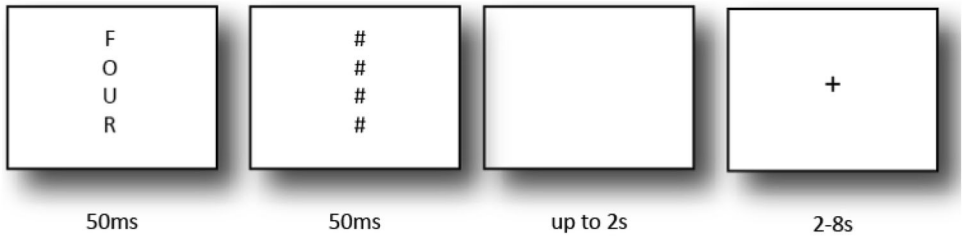

Fig. 2 Adaptive Words Task Presentation

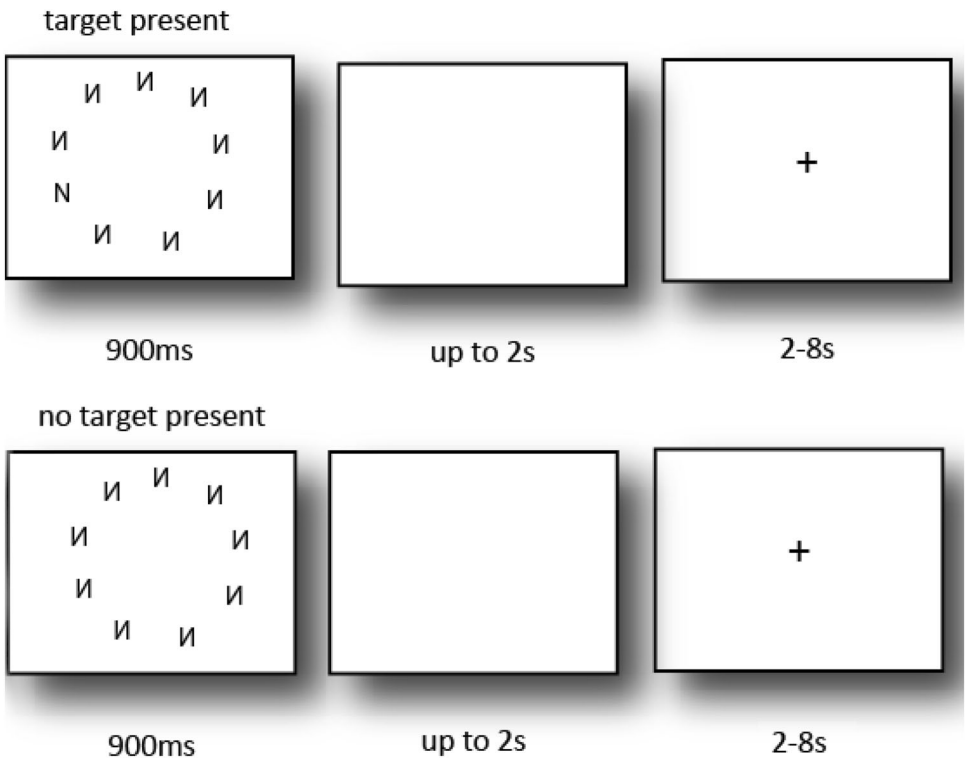

Fig. 3 Visual search task presentation

participants to hold a stream of characters in their working memory and responding when a new character that is presented to them matches one of the characters they are currently holding. The task development was based on Harvey et al. [17]. The task presented participants with a series of letters, a single letter at a time, for a duration of 500ms each. The letters would appear in the center of the screen with a plain white background. Only the letters B, D, G, T, V along with their lower-case variants, b, d, g, t, v, were used. Before each block, participants were given an 'n' value of either one, two, three or four. The participant's goal was to determine whether the current letter presented to them matched the letter that was 'n' presentations behind the current letter that was displayed (case insensitive). For example, in Fig. 5, if an 'n' of two had been given to the participant, then the correct response would be 'yes.' If the participant was given an 'n' value of one, however, the correct response would be 'no.'

Stimulus Materials: Test Data

Triage Task

The triage analyst task acts as an ecologically valid representation of a cyber-security network analyst's position and is based on the work of Greenlee et al. [16]. The task involved the participant viewing what is at first an empty table in the center of the screen. The table headings were 'Source IP,' 'Source Port,' 'Destination IP,' 'Destination Port.' Participants were informed before the task beginning that they did not require working knowledge of the terminology involved to complete the task. The table would then be populated with incoming 'transmissions' on the 'network' the participant was monitoring. Starting from the top of the table, new 'transmissions' would fill the table until a maximum of five 'transmissions' was on the screen. After five 'transmissions' were shown on the screen, the bottom transmission would be removed from the screen to make room for a new incoming transmission

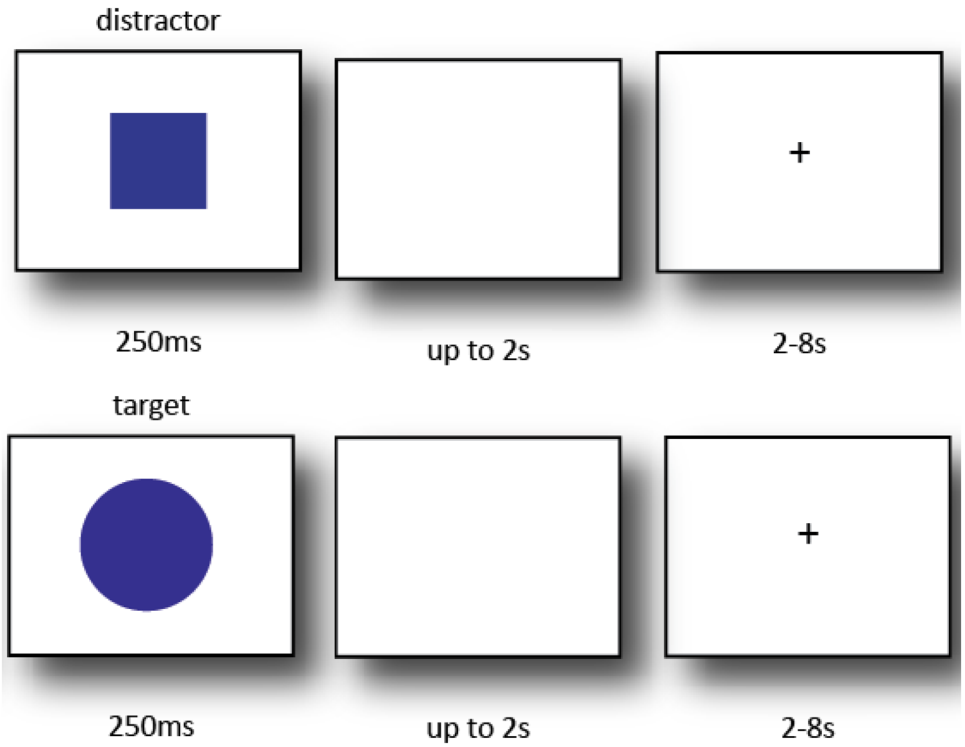

Fig. 4 The Go/No-Go task presentation

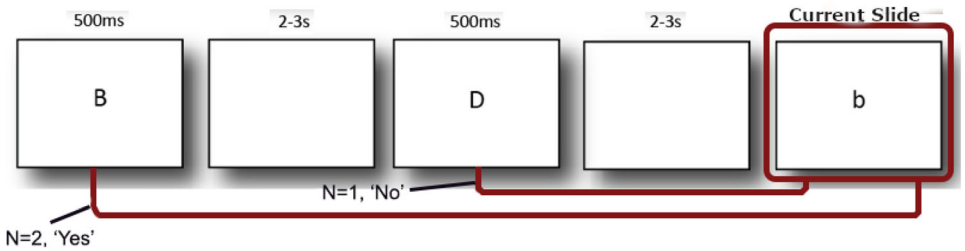

Fig. 5 N-back task presentation

at the top, bumping the rest of the table down one slot. The participant was tasked with detecting 'intrusions' on the network. These 'intrusions' were defined as either two different 'transmissions' on the table having the same destination information (both 'Destination IP' and 'Destination Port') or two different 'transmissions' on the table having the same source information (both 'Source IP' and 'Source Port'). The participant was only asked to identify whether the newest (topmost) 'transmission' was or was not an 'intrusion.' Figure 6 shows an example of intrusion. The triage testbed tracked participant response times as well as their performance (logging correct, incorrect, or no response events) throughout the task.

Multi-Attribute Task Battery

The Multi-Attribute Task Battery (MATB) condition used in our stimulus materials was closely based on the version of the task implemented by Comstock and Arnegard [9]. This implementation of the task used in our experimental

Fig. 6 Source intrusion

Source IP	Source Port	Destination IP	Destination Port
103.17.22.62	82	198.176.21.9	14
56.254.13.15	11	33.98.47.72	12
226.12.22.132	63	108.71.226.62	77
103.17.22.62	82	251.102.18.3	65
42.113.56.5	44	56.225.11.89	43

scenario as a task which involved a high level of multi-tasking was needed in order to produce a situation in which task shedding would be likely. The difficultly of the task necessitated the need for high achieving users to not become overloaded or overly stressed, which forced the user to prioritize their actions based on the most time-sensitive or important needs of the task at the time. The procedure's final implementation used the air force's updated version of the Multi-Attribute Task Battery (AF_MATB) [28], and a level of difficulty that required a high amount of multi-tasking and mental effort was chosen for the recorded experimental trials.

As a result of the selected difficulty level, the task was nearly impossible to complete perfectly. Piloting the task revealed that in order to maintain an adequate performance score participants needed to remain engaged during the entire duration of the task. Similar to the original MATB, AF_MATB variation of the task contains six different windows, all of which provide the information needed for the participant to complete four different subtasks (see Fig. 7). The tasks, System Monitoring, Communications, Resource Management and Tracking, which all require different inputs from the user in order to perform successfully on the task. The 'tracking' subtask was disabled during this experiment to reduce the physical motion of the participant activating the motor cortex in the brain as well as reducing motion artifacts from the fNIRS data. The other two windows, which contain Scheduling and Pump Status information, are resources that the user can use to improve performance during the task. The first requirement in the system monitoring subtask (top-left pane) is to keep track of the two lights at the top of the window and keep them at their original status by toggling them on/off using the buttons below them. The second requirement in the system monitoring subtask requires the user to be aware of the four scales and press the corresponding button if any of the scales deviates from the center by more than one tick mark. The communication subtask (bottom left pane) involves the subject listening for verbal requests to change the frequency of specific radio calls. The verbal requests include a call-sign, and if the call-sign is not the call-sign of the subject, the request is to be ignored. The resource management subtask (bottom center pane) requires the subject to keep the fuel levels in tanks A and B within 500 units of the initial level of 2500 units each. The pumps connected to the tanks can be used to pump fuel from the lower supply tanks to tanks A and B. The pumps can be turned on/off by clicking on the particular pump. However, these pumps can malfunction for periods of time during the experiment. If a pump is malfunctioning, it cannot be turned on. The AF_MATB maintains of record of participant performance throughout the whole task, and after the task is completed a report of each subject's performance data on the subtasks within the MATB is exported. In our pilot studies, we used a set of five pilot subjects to complete the MATB task with varying number of task frequencies (10 configurations). Then, we chose three of the 10 MATB difficulty levels as low/high/medium based on the average scores obtained by the pilot study participants for each task configuration. The three configurations were chosen by dividing the 10 configurations based on average score obtained by participants and breaking the 10 configurations into three clusters of low/medium/high difficulty levels. Then the three configurations, which have the

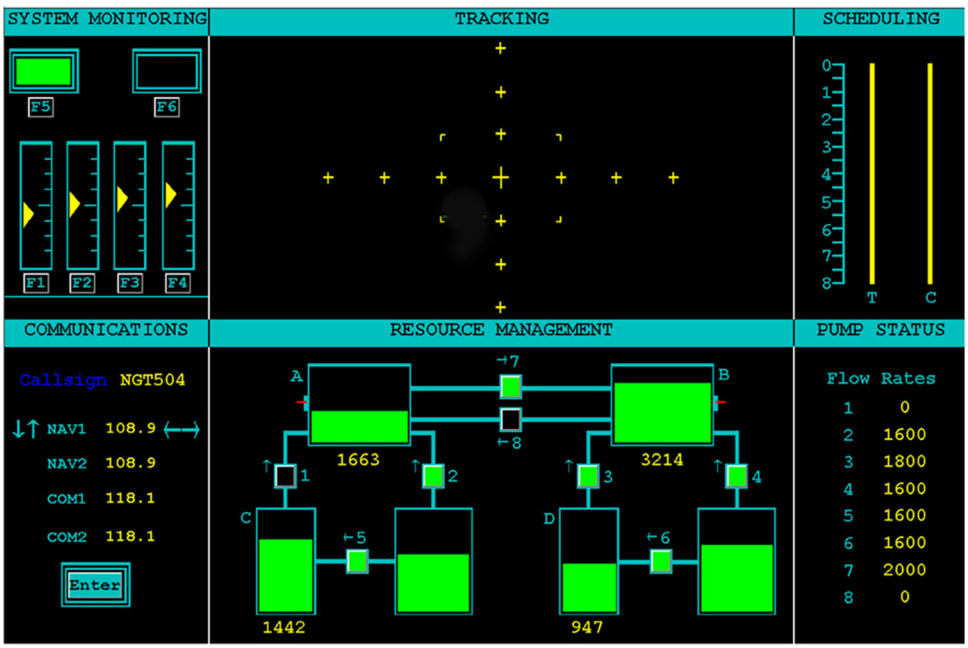

Fig. 7 The Multi-Attribute Task Battery

lowest standard deviation among the clusters, were chosen as MATB low/medium/high configurations for the experiment.

Data Analysis

fNIRS Data

The fNIRS provides 104 channels (52 channels of oxygenated hemoglobin and 52 channels of deoxygenated hemoglobin) of data at 10 Hz. These data were band-pass filtered with a window of .01/.5Hz to remove cardiac, respiratory and high-frequency unwanted noise [2]. The time-series data from fNIRS were divided into 5-s blocks, and the average value of the fNIRS data was obtained for each of the 104 data channels for those 5-s blocks.

Label Calculation

In this experiment, we relied on objective performance data to calculate our labels. The labels were calculated based on data logged by the software that presented the task itself. This way, we were able to match the exact times the stimuli were presented and the duration of the stimuli using the log files from the task software. The log files for the benchmark and triage tasks were generated by capturing information about when an event was triggered via the stimulus software, as well as information about when the participant responded to the event. Every event logged by the software was tagged with a UNIX timestamp. The participant's reaction times were then calculated, in milliseconds, for each trial included in the logfile. The logfile also indicated whether or not the participant responded correctly to the trial. The MATB logfile collected data at 10 Hz. The datapoints collected include the number of accurate and inaccurate user responses within each time segment and the difference between the current tank levels from the target tank level in the resource management task.

Using these performance and taskload metrics (further discussed in Sect. 4.2.1), we introduce a measure called 'Task Shedding Index (TSX),' which is a combination of taskload and performance. The TSX is used as an indicator of the potential of the user for task shedding.

Task Shedding Index (TSX)

As described in Sect. 2.1, when users are put in a high-workload scenario, and the workload of the task is over a certain threshold, they tend to shed that task or switch to another task on which they may be able to perform better [46]. At this point, the performance on the prior task would degrade, and therefore, be an apt point of intervention on the part of the machine agent in the HMT [5]. To account for this, a hybrid workload and performance model is needed to predict task shedding tendencies. Using the distance between normalized task load and normalized performance (detailed next) as our ground truth in this analysis, the taskload (T) and performance (P) of each subject are normalized and the Task Shedding Index (TSX) is defined as the subtraction of normalized P from normalized T (Eq. 1).

$$\begin{aligned}&\text{Task Shedding Index}\,(TSX)\\&\quad=\text{Normalize}(T)-\text{Normalize}(P)\end{aligned}\tag{1}$$

The ranges of the T, P and TSX for our dataset are shown in Fig. 8.

Task load was obtained by adding together the number of tasks that were presented to the user during those 5 s. If certain tasks started or ended during each 5-s chunk of time, they were apportioned according to the amount of time they were present during the time frame.

In Fig. 9, Task A would contribute 50% to the current period's task load. Task B would contribute 100%, and task C would contribute 25%. Therefore, the task load would be calculated according to Eq. (2),

$$\text{Apportioned TaskLoad}\,(T)=0.5+1+0.25=1.75\tag{2}$$

Performance is apportioned similarly as in Eq. (3),

$$\begin{aligned}&\text{Apportioned Perf.}\,(P)=0.5*\text{TaskA}_{\text{Perf}}\\&\quad+1*\text{TaskB}_{\text{Perf}}+0.25*\text{TaskC}_{\text{Perf}}\end{aligned}\tag{3}$$

To calculate performance in the benchmark and triage tasks, the onset, duration and accuracy of user response were saved to a file during the experiment by the testbed.

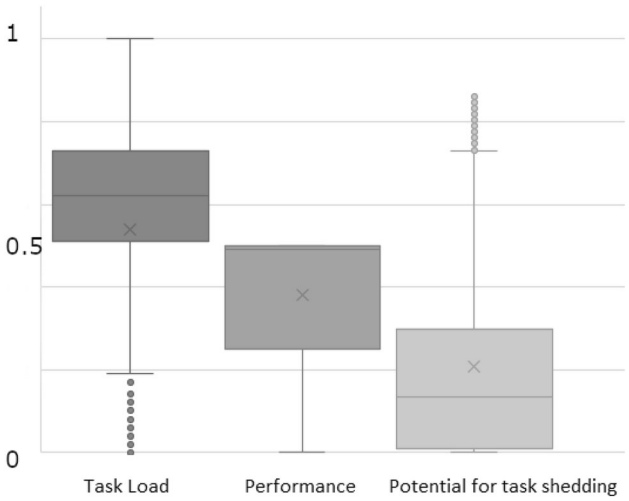

Fig. 8 Data ranges for task load, performance and Task Shedding Index

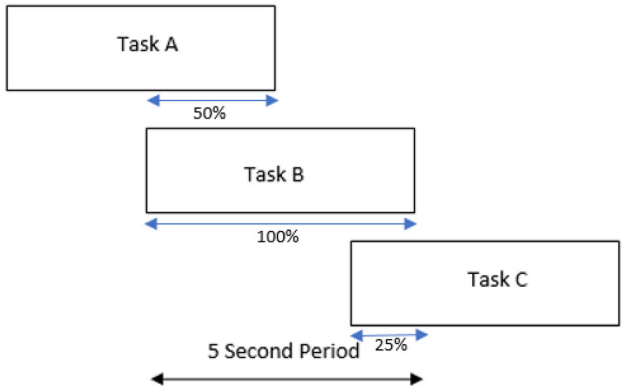

Fig. 9 Apportioning tasks to the time segments

Calculating user performance for benchmark and triage tasks was done using Eq. (4) for each 5 s.

$$\text{Triage}_{\text{perf.}} = \frac{(\text{Correct Responses}) - (\text{Incorrect Responses})}{\text{TaskLoad}(T)} \quad (4)$$

Since MATB is a multi-attribute task, the performance calculation for MATB was done by taking the compound performance of all the subtasks. The MATB contains the following subtasks,

- Time point
- Number of accurate light toggle responses (L)
- Number of accurate gauge responses (G)
- Number of accurate communications responses (C)
- Number of inaccurate light toggle responses (l)
- Number of inaccurate gauge responses (g)
- Number of inaccurate communications responses(c)
- Difference of Tank A value from desired target value (AD)
- Difference of Tank B value from desired target value (BD)

Equation (5) was used to calculate the user performance for the MATB task.

$$MATB_{perf.} = \frac{(L+G+C) - (l+g+c) + FuelTankA_{perf.} + FuelTankB_{perf.}}{TaskLoad(T)}, \quad (5)$$

where Fuel Tank A performance is calculated by using the following equation,

$$\begin{cases} 0 & ABS(\Delta AD) > 100 \\ 1 & ABS(\Delta AD) \leq 100 \end{cases}$$

The same method was used to calculate Fuel Tank B performance.

Discretization of Ground Truth Labels

The TSX labels follow a positively skewed distribution. We are interested in breaking down the distribution into three levels of TSX, namely 'low,' 'moderate' and 'high.' This categorization has been chosen since it allows us to look at the 'moderate' level of TSX as the desired level, the low level of TSX as when the user is 'idling' and the 'high' level of TSX as when the user is overloaded. This enables different strategies for dealing with each of these cases. The TSX labels were discretized using Equal Frequency Discretization over all the datasets. The Equal Frequency Discretization algorithm sorts all values of continuous variable in ascending order and divides the range into three intervals so that every interval contains the same number of sorted values. The resulting ranges from Equal Frequency Discretization for the TSX labels were:

- Low TSX: Less than 0.12
- Moderate TSX: 0.12 to 0.323
- High TSX: 0.323 upward

The resulting discretized label distribution for each of the datasets is given in Table 1.

Table 1 Ground truth label distribution for each of the datasets

	Low TSX labels No./ Percentage (%)	Moderate TSX labels No./ Percentage(%)	High TSX labels No./ Percentage (%)
Adaptive Words	130/36%	115/32%	120/33%
Visual Search	298/45%	205/31%	157/24%
Go/No-Go	240/37%	253/39%	153/24%
nback	210/31%	277/42%	180/27%
Triage	2050/34%	1802/30%	2106/35%
MATB	402/23%	678/40%	614/36%

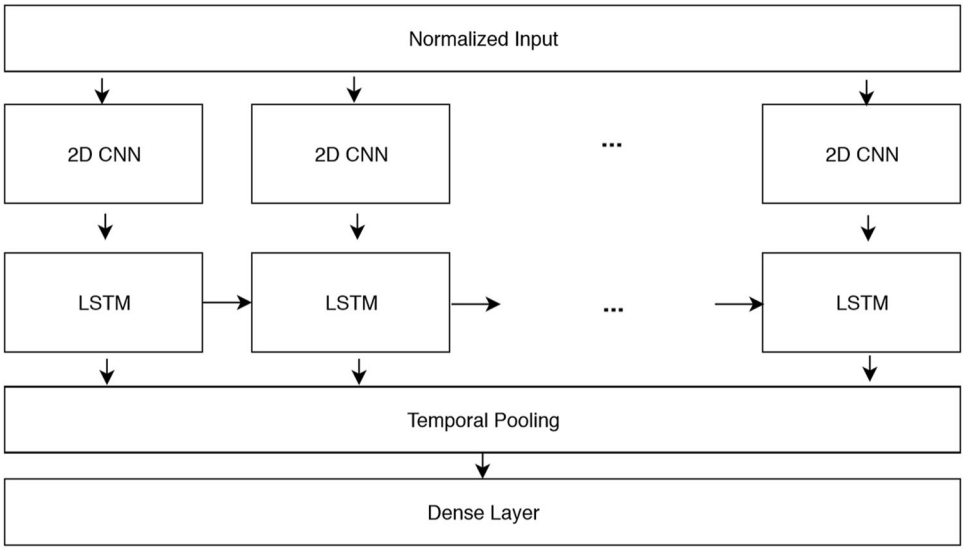

Fig. 10 CNN+LSTM classification model

Classification Model

We trained a classification model that we developed in our earlier work to classify fNIRS data using a combination of a convolutional neural network (CNN) and long short-term memory (LSTM) as shown in Fig. 10. Since CNNs are well suited for capturing the spatial nature of fNIRS data, and LSTMs are good at capturing the temporal behavior of fNIRS data, this model provided performance improvement over traditional machine learning methods [3]. We evaluated the accuracy of the classifier for different LSTM time step sizes and found that the optimal time step size for LSTM was 3 time steps (15 s).

Model Evaluation

Data from the benchmark tasks, detailed in Sect. 3.2, were used to train four separate models. Each model was trained using the respective benchmark task and was tested against both the triage task and MATB task. It is notable that the four models were trained on benchmark data of 25 subjects and tested on triage data for the same subjects during a separate triage session, whereas the MATB test data were from a different subject pool of 20 subjects (e.g., model transfer to new, previously unseen subjects with the MATB classifications). Each benchmark task involves a different cognitive resource, with each resource capable of being independently overloaded. We hypothesize that the different models trained on these different benchmark tasks will perform differently based on the type of task shedding that it is predicting. For instance, a model trained on the n-back task would perform better for a task that involved the utilization of short-term memory, such as remembering a string of digits for a variable period, whereas the visual search might perform better on a task that involved finding a salient stimulus in a cluttered visual environment. In addition to the models trained on individual benchmark tasks, a combined ensemble model was also trained on all benchmark tasks using a voting classifier, which takes the predicted probabilities for each class from each model and averages them. The predicted labels are calculated as follows:

$$\text{Adaptive words class probabilities} = [p_{low1}, p_{medium1}, p_{high1}] \tag{6}$$

$$\text{Visual search class probabilities} = [p_{low2}, p_{medium2}, p_{high2}] \tag{7}$$

$$\text{Go no go class probabilities} = [p_{low3}, p_{medium3}, p_{high3}] \tag{8}$$

$$\text{Nback class probabilities} = [p_{low4}, p_{medium4}, p_{high4}] \tag{9}$$

$$\text{Ensemble class probabilities} = [Avg(p_{low}), Avg(p_{medium}), Avg(p_{high})] \tag{10}$$

$$\text{Ensemble voting prediction} = argmax[Ensemble\ class\ probabilities] \tag{11}$$

We tested the models on both a sequential task (triage) and a concurrent task (MATB) [47]. The report on our model performance in these different tasks is detailed in Sect. 5.

Results

We were interested in the overall classification performance of our models as well as the performance on each of the label types because each of the 'low,' 'moderate' and 'high' TSX labels can be used to keep the system in an optimally productive state (Fig. 13). Overall, the five trained models had better accuracy on the MATB task than the triage task based on a Students' t test ($p < 0.007$), where accuracy is defined by,

$$\text{Accuracy} = (\text{Correctly Classified Instances}/\text{Total Instances})$$

Accuracy results from the tests are described in Table 2.

As shown in Table 2, the overall accuracy of the models was around 60%, with the best overall accuracy results obtained by the ensemble model, with 61% accuracy for triage data and 63% accuracy for MATB data. This is promising considering that random guessing would result in 33% accuracy on a balanced 3-class problem. Next, we look at the confusion matrices and the precision, recall and f1 scores of the models to better understand model performance in the context of task shedding events. Below we present the confusion matrices from the testing done on triage data and MATB data using the ensemble model (Figs. 11, 12). In these figures, 'low' TSX is represented by 1, 'moderate' TSX by 2 and 'high' TSX by 3.

Table 2 Accuracy results of the models on the test sets

	Triage accuracy (%)	MATB accuracy (%)
Adaptive words	60	63
Visual search	60	60
Go/No-Go	59	60
nback	61	61
Ensemble	61	63

Fig. 11 Confusion matrix for ensemble model tested on triage task. 'Low' TSX is represented by 1, 'moderate' TSX by 2 and 'high' TSX by 3

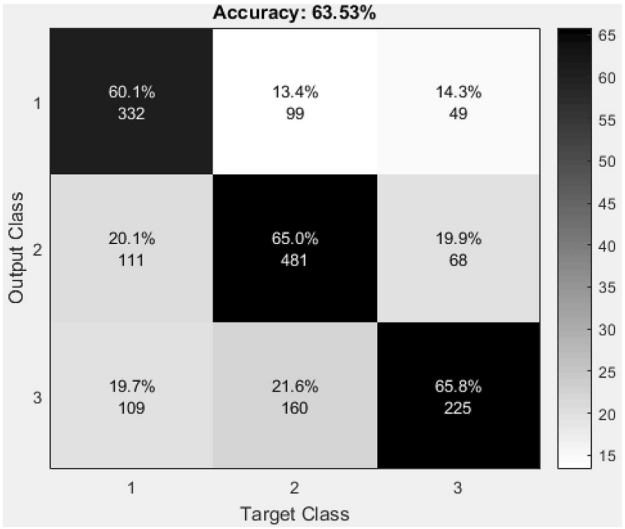

Fig. 12 Confusion matrix for ensemble model tested on MATB task. 'Low' TSX is represented by 1, 'moderate' TSX by 2 and 'high' TSX by 3

The results indicated in the confusion matrices (Figs. 11, 12) show that the ensemble model performed well when identifying instances of 'low' TSX for the triage task. However, the model had difficulty in correctly identifying the 'moderate' and 'high' TSX instances within the triage task. On the MATB test, the ensemble model performed evenly well when predicting TSX values. This difference in model performance could be due to the sequential nature of the triage task, which has the user performing one single task multiple times in a row, constantly engaging one subset of cognitive resources, not having to switch to any other type of cognitive processing. As a result of this sequential presentation of the stimulus, the same brain region may have been continuously activated throughout the completion of the entire task and the physiological response may have been so gradual that the model was unable to predict when sharp changes in TSX occurred. This may account for differences seen between 'moderate' and 'high' TSX labels being not as easy to detect as the difference between 'low' and 'moderate' labels. On the contrary, the MATB, as a concurrent task, activates multiple brain regions due to task presentation happening all at once, with the user having to switch between using multiple cognitive resources to complete each task within the battery. Though it is a possibility that the differences in accuracy between the model are due to the nature of the tasks, more ecologically valid datasets would be required to

Table 3 Precision, recall and f1 score of the 'low' label for models on the test sets

Model	Triage			MATB		
	Precision	Recall	f1 score	Precision	Recall	f1 score
Adaptive words	0.56	0.94	0.70	0.66	0.66	0.66
Visual search	0.6	0.85	0.70	0.71	0.59	0.64
Go/No-Go	0.60	0.89	0.71	0.69	0.55	0.61
nback	0.58	0.90	0.70	0.68	0.62	0.65
Ensemble	0.60	0.90	0.72	0.69	0.66	0.67

Table 4 Precision, recall and f1 score of the 'Medium' label for models on the test sets

Model	Triage			MATB		
	Precision	Recall	f1 score	Precision	Recall	f1 score
Adaptive words	0.66	0.53	0.59	0.70	0.68	0.69
Visual search	0.62	0.42	0.50	0.66	0.60	0.63
Go/No-Go	0.71	0.44	0.54	0.72	0.61	0.66
nback	0.64	0.43	0.51	0.70	0.59	0.64
Ensemble	0.68	0.45	0.54	0.73	0.60	0.66

Table 5 Precision, recall and f1 score of the 'high' label for models on the test sets

Model	Triage			MATB		
	Precision	Recall	f1 score	Precision	Recall	f1 score
Adaptive words	0.60	0.66	0.63	0.50	0.74	0.6
Visual search	0.58	0.60	0.59	0.43	0.61	0.5
Go/No-Go	0.50	0.67	0.57	0.37	0.74	0.49
nback	0.41	0.58	0.48	0.41	0.68	0.51
Ensemble	0.58	0.61	0.59	0.45	0.56	0.49

test out whether or not the presentation of the task, sequential or concurrent, shows a similar effect on the model accuracy.

As mentioned above, we were also interested in the per-label performance of the models. Precision and recall values were calculated for each of the TSX labels. Precision is a measure of how many of the 'true' classifications are relevant in each class. Recall or sensitivity refers to the true prediction of each class when it is actually true. F1 score (harmonic mean of precision and recall) is a measure of the accuracy of each of the tests. The F1 score is more useful than accuracy when the cost of false positives and false negatives is variable. Precision, recall and f1 score values for each of the five separate models are presented in Tables 3, 4 and 5. Each table is organized by its TSX value category.

The precision for the 'low' TSX labels was higher for the MATB task classification than the triage task (from Students' t test resulting in $p < 0.00001$). Precision for 'moderate' TSX labels was also higher for the MATB task than triage task ($p < 0.01$). Precision for 'high' TSX labels was higher for the triage task than MATB task ($p < 0.0003$). Recall for the 'low' TSX labels was higher for the triage task compared to MATB ($p < 0.00005$). Recall for 'moderate' TSX labels was higher for MATB task than triage task ($p < 0.00002$), and recall for 'high' TSX labels was also higher for MATB task than triage task ($p < 0.22$). Interestingly, even though the MATB data were collected from a different subject pool than the one that the models were trained on, neither triage nor MATB test did significantly better over all the labels. Precision and recall values for the model could stand to be improved, perhaps through the use of more training data. However, for a general model such as the one proposed here, establishing criteria for what the acceptable values are for both precision and recall scores may prove difficult as these two values tend to be task-dependent. Recall, in the case of the two ecologically valid examples that were trained on for this experiment would be an important factor for 'high' TSX label as a result of the false-negative cost being too high (Fig. 13). If the task shedding event is not detected, and the human agent is currently overloaded, the machine agent has no way of knowing that it should interfere and assist the human agent within the HMT with accomplishing

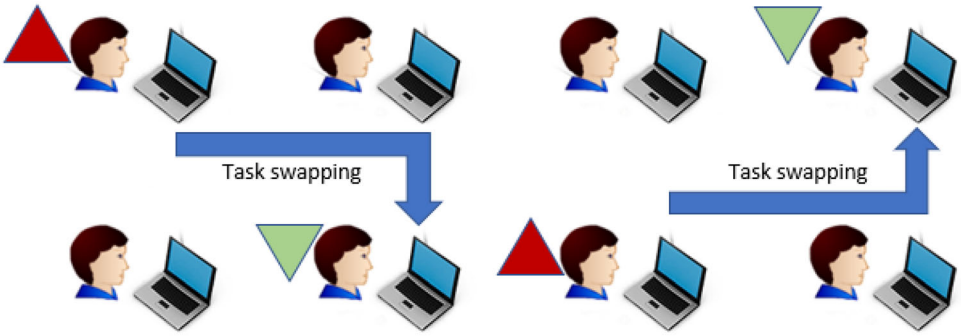

Fig. 13 An HMT system swapping tasks from overloaded (high TSX) users to underloaded (low TSX) users

the current goal. This is a moment in which task shedding will become likely due to the high mental workload of the human agent within the HMT. Precision, when considered with the ecologically valid tasks that the model was tested on, is also important to consider because a high false-positive rate may lead to user behavior being interrupted unnecessarily, increasing both the human agent's frustration with the system and negative affect. Frequent and unnecessary interventions by the machine agent might carry a performance hit for the HMT both by interrupting human agents in the HMT and the overhead involved in task switching. Models using this type of data to provide feedback for an autonomous system within an HMT environment will need to, therefore, strike a good balance of precision and recall when it comes to the predicting 'high' TSX values.

Good prediction performance for the low TSX label is also important because, when a user is in low TSX state, the HMT system can load some tasks from overloaded users to the 'low' TSX user (Fig. 13), thereby keeping the overall productivity rate of the HMT high. However, false positives, in this case, could be harmful because the system might load more tasks to somebody who is not at 'low' value of TSX. False negatives, however, may not prove to be as harmful as their counterpart because the system will then simply ignore an idling user. Though this would cause a productivity decline within the system, it would not cause any catastrophic failures. Therefore, precision should be prioritized over recall when it comes to 'low' TSX values.

The 'moderate' TSX is the 'ideal' state in HMT systems, and this is the state in which the human user is productive without being at risk for being overloaded. False positives for this label mean that no action will be taken by the system. If the user is actually at a 'low' TSX, then this would not cause any major issue, but if the user is actually overloaded, or in the 'high' TSX category, this may lead to task shedding by the user and is not desirable. False negatives in case of 'moderate' TSX label depend on the predicted state. If the user is actually at 'moderate' TSX and the system predicts 'low' TSX, the system might overload the user by assigning more tasks to him. On the other hand, if the user is at 'moderate' TSX and system predicts 'high' TSX, it might unload the user putting him into idle state. Because of these aspects, precision and recall values should be considered carefully depending on both the nature of the task and the tolerance of potential performance decrements within the HMT environment. Also of note is that, as hypothesized, the different models trained on different benchmark tasks performed differently. And the ensemble model combining all models performed better than any of the other models. This was expected because each cognitive benchmark task on which the models were trained elicits a distinct type of cognitive load, with a signature neural correlate that is recorded up in the fNIRS data. Though further analysis is needed, and better labeled ecologically valid tasks could be used to generate even more accurate models, this could be helpful for training models used by autonomous agents within HMTs designed to accomplish tasks that rely heavily on different types of cognitive processing by the human agent. For example, a model trained on the adaptive words task could be more suitable to capture verbal working memory load and therefore might perform better on a task in which a human agent performs a task that requires the use of working memory and visual-lexical processing, such as remembering a set of instructions, than a model trained on the Go/No-Go task, which measures one's level of response inhibition.

Conclusion

One of the major challenges today in fNIRS research is the difficulty of obtaining large datasets to run analysis on. In this work, we demonstrate that across-task machine learning is possible on fNIRS data with promising performance. This would indicate that researchers would be able to combine data from multiple experiments to develop models that generalize well. As mentioned above, we were able to generalize not only across-tasks but also across-subject pools. These have important implications for using

psychophysiological data from fNIRS in real-world applications and environments.

In a multi-tasking situation similar to MATB, which an air force pilot faces, any operator performance dips can cause catastrophic incidents. Therefore, our method would be useful in such scenarios to prevent user error due to cognitive overload. Other such scenarios can include air traffic controller interfaces, or stock–broker interfaces, where the cost of performance degradation due to cognitive overload is very high. Also, once the operator TSX state is detected, it can be used to improve the overall productivity of the HMT, by loading tasks to users who are in idle state. In this way, the same method we introduced here can be used as a productivity tool in collaborative workplaces. For example, if a group of users is doing some data entry task, the HMT can monitor users who are overloaded and swap the tasks to the underloaded users, thereby improving overall productivity. Therefore, this method can be used in both critical and noncritical systems to improve system behavior.

In the above discussion, we focused on multi-user environments as an application area for this work. However, this method could be adapted to a single user as well. For example, a user doing multi-tasking could occasionally have some parts of their visual cognitive faculties overloaded. In this scenario, the system can offload some of the visual tasks from the user and present tasks that occupy different cognitive faculties. Such a system could improve the performance of a single user. This idea could also be extended to multi-user, multi-tasking environments, where during task swapping, the system could check for users who are more suitable to accept the task type based on their current cognitive faculties being used.

In this article, we have introduced a measure combining both task load and performance to be able to detect task shedding events within HMT environments. We have been able to classify the Task Shedding Index into three levels: 'low,' 'moderate' and 'high.' We have considered two test cases which are the triage (Cyber Analyst) task and the MATB task in our testing. The spatial and temporal nature of fNIRS enabled us to use a CNN+LSTM model designed to capture both the spatial and temporal nature of brain activity. As stated in previous sections, this method of classification shows great promise in both the domains of HCI and human factors, though it also has more direct practical implications for Human–Machine Teaming and multi-tasking environments where various cognitive resources of a user are occupied at different times. Specifically, our across-task performance shows promise for training models on simpler tasks and being able to generalize to compound tasks. In conclusion, our study demonstrates that we can obtain a generalizable classifier that performs well across multiple subject pools as well as across multiple tasks, thus enabling adaptive Human–Machine Teaming across diverse real-world settings.

Acknowledgements We thank NSF for supporting this research through NSF Award #1816732.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Alhagry S, Fahmy AA, El-Khoribi RA (2017) Emotion recognition based on eeg using lstm recurrent neural network. Emotion 8(10):355
2. Ayaz H, Shewokis PA, Curtin A, Izzetoglu M, Izzetoglu K, Onaral B (2011) Using mazesuite and functional near infrared spectroscopy to study learning in spatial navigation. J Vis Exp JoVE 56:e3443
3. Bandara D, Hirshfield L, Velipasalar S (2019) Classification of affect using deep learning on brain blood flow data. J Near Infra Red Spectrosc 27(3):206–219
4. Bandara D, Velipasalar S, Bratt S, Hirshfield L (2018) Building predictive models of emotion with functional near-infrared spectroscopy. Int J Hum Comput Stud 110:75–85
5. Bliss JP, Harden JW, Dischinger Jr HC (2013) Task shedding and control performance as a function of perceived automation reliability and time pressure. In: Proceedings of the human factors and ergonomics society annual meeting, vol 57. SAGE Publications, Los Angeles, CA, pp 635–639
6. Brumby DP, Salvucci DD, Howes A (2009) Focus on driving: how cognitive constraints shape the adaptation of strategy when dialing while driving. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1629–1638
7. Chan J, Power S, Chau T (2012) Investigating the need for modelling temporal dependencies in a brain-computer interface with real-time feedback based on near infrared spectra. J Near Infrared Spectrosc 20(1):107–116
8. Chen JY, Barnes MJ (2014) Human-agent teaming for multirobot control: a review of human factors issues. IEEE Trans Hum Mach Syst 44(1):13–29
9. Comstock JR Jr, Arnegard RJ (1992) The multi-attribute task battery for human operator workload and strategic behavior research. NASA, Washington
10. Coyle S, Ward T, Markham C, McDarby G (2004) On the suitability of near-infrared (nir) systems for next-generation brain-computer interfaces. Physiol Meas 25(4):815
11. Cui X, Bray S, Bryant DM, Glover GH, Reiss AL (2011) A quantitative comparison of nirs and fmri across multiple cognitive tasks. Neuroimage 54(4):2808–2821
12. Cui X, Bray S, Reiss AL (2010) Speeded near infrared spectroscopy (nirs) response detection. PLoS ONE 5(11):e15474
13. Davidson PR, Jones RD, Peiris MT (2007) Eeg-based lapse detection with high temporal resolution. IEEE Trans Biomed Eng 54(5):832–839
14. de Visser E, Parasuraman R (2011) Adaptive aiding of human–robot teaming: effects of imperfect automation on performance, trust, and workload. J Cognit Eng Decis Mak 5(2):209–231
15. Ferrari M, Quaresima V (2012) A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. Neuroimage 63(2):921–935

16. Greenlee ET, Funke GJ, Warm JS, Sawyer BD, Finomore VS, Mancuso VF, Funke ME, Matthews G (2016) Stress and workload profiles of network analysis: not all tasks are created equal. In: Ahram TZ, Nicholson D (eds) Advances in human factors in cybersecurity. Springer, Berlin, pp 153–166
17. Harvey PO, Fossati P, Pochon JB, Levy R, LeBastard G, Lehéricy S, Allilaire JF, Dubois B (2005) Cognitive control and brain resources in major depression: an fmri study using the n-back task. Neuroimage 26(3):860–869
18. Hennrich J, Herff C, Heger D, Schultz T (2015) Investigating deep learning for FNIRS based BCI. In: EMBC, pp 2844–2847
19. Herff C, Heger D, Fortmann O, Hennrich J, Putze F, Schultz T (2014) Mental workload during n-back task-quantified in the prefrontal cortex using fnirs. Front Hum Neurosci 7:935
20. Hirshfield LM, Solovey ET, Girouard A, Kebinger J, Jacob RJ, Sassaroli A, Fantini S (2009) Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 2185–2194
21. Hochreiter S, Schmidhuber J (1997) Lstm can solve hard long time lag problems. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems. MIT Press, Cmbridge, pp 473–479
22. Huettel SA, Mack PB, McCarthy G (2002) Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nat Neurosci 5(5):485
23. Janssen CP, Brumby DP (2010) Strategic adaptation to performance objectives in a dual-task setting. Cognit Sci 34(8):1548–1560
24. Kontogiannis T, Kossiavelou Z (1999) Stress and team performance: principles and challenges for intelligent decision aids. Saf Sci 33(3):103–128
25. LeCun Y, Bengio Y et al (1995) Convolutional networks for images, speech, and time series. In: Arbib MA (ed) The handbook of brain theory and neural networks, vol 3361, 10th edn. MIT Press, Cambridge
26. Liu Y (1996) Queueing network modeling of elementary mental processes. Psychol Rev 103(1):116
27. Meyer DE, Kieras DE (1997) A computational theory of executive cognitive processes and multiple-task performance: part i. Basic mechanisms. Psychol Rev 104(1):3
28. Miller WD Jr (2010) The us air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior. Technical report, Consortium Research and Fellows Program, Arlington VA
29. Parasuraman R, Hancock PA (2001) Adaptive control of mental workload. In: Hancock PA, Desmond PA (eds) Human factors in transportation. Stress, workload, and fatigue. Lawrence Erlbaum Associates Publishers, pp 305–320
30. Pashler HE, Sutherland S (1998) The psychology of attention, vol 15. MIT press, Cambridge
31. Peck EM, Afergan D, Yuksel BF, Lalooses F, Jacob RJ (2014) Using FNIRS to measure mental workload in the real world. In: Gilleade K (ed) Advances in physiological computing. Springer, Berlin, pp 117–139
32. Reeves B, Lang A, Kim EY, Tatar D (1999) The effects of screen size and message content on attention and arousal. Med Psychol 1(1):49–67
33. Salvucci DD, Taatgen NA (2010) The multitasking mind. Oxford University Press, Oxford
34. Sirevaag EJ, Kramer AF, Reisweber M, Wickens CD, Strayer DL, Grenell JF (1993) Assessment of pilot performance and mental workload in rotary wing aircraft. Ergonomics 36(9):1121–1140
35. Smith ME, Gevins A, Brown H, Karnik A, Du R (2001) Monitoring task loading with multivariate eeg measures during complex forms of human–computer interaction. Hum Factors 43(3):366–380
36. Solovey ET, Zec M, Garcia Perez EA, Reimer B, Mehler B (2014) Classifying driver workload using physiological and driving performance data: two field studies. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp. 4057–4066
37. Strenzke R, Schulte A (2011) Modeling the human operator's cognitive process to enable assistant system decisions. GAPRec 2011:38
38. Strenzke R, Uhrmann J, Benzler A, Maiwald F, Rauschert A, Schulte A (2011) Managing cockpit crew excess task load in military manned–unmanned teaming missions by dual-mode cognitive automation approaches. In: AIAA guidance, navigation, and control conference, p 6237
39. Tai K, Chau T (2009) Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface. J Neuroeng Rehabilit 6(1):39
40. Tamaki T, Hiwa S, Hachisuka K, Okuno E, Hiroyasu T (2016) Region-of-interest estimation using convolutional neural network and long short-term memory for functional near-infrared spectroscopy data. Front Neuroinform 12:10
41. Treacy Solovey E, Afergan D, Peck EM, Hincks SW, Jacob RJ (2015) Designing implicit interfaces for physiological computing: guidelines and lessons learned using fnirs. ACM Trans Comput Hum Interaction (TOCHI) 21(6):35
42. Villringer A, Chance B (1997) Non-invasive optical spectroscopy and imaging of human brain function. Trends Neurosci 20(10):435–442
43. Van der Linden D, Frese M, Meijman TF (2003) Mental fatigue and the control of cognitive processes: effects on perseveration and planning. Acta Psychol 113(1):45–65
44. Wang Q, Cavanagh P, Green M (1994) Familiarity and pop-out in visual search. Percept Psychophys 56(5):495–500
45. Wickens CD (1991) Processing resources and attention. Mult Task Perform 1991:3–34
46. Wickens CD, Gutzwiller RS, Santamaria A (2015) Discrete task switching in overload: a meta-analyses and a model. Int J Hum Comput Stud 79:79–84
47. Wickens CD, McCarley JS (2007) Applied attention theory. CRC Press, London
48. Wickens CD, Santamaria A, Sebok A (2013) A computational model of task overload management and task switching. In: Proceedings of the human factors and ergonomics society annual meeting, vol 57. SAGE Publications Sage CA, Los Angeles, CA, pp 763–767
49. Zimmermann R, Marchal-Crespo L, Edelmann J, Lambercy O, Fluet MC, Riener R, Wolf M, Gassert R (2013) Detection of motor execution using a hybrid fnirs-biosignal bci: a feasibility study. J Neuroeng Rehabilit 10(1):4