Topics in Cognitive Science 00 (2023) 1–42
© 2023 Cognitive Science Society LLC.
ISSN: 1756-8765 online
DOI: 10.1111/tops.12669

This article is part of the topic "The Emerging Cognitive Science of Human-Autonomy Teams," Christopher W. Myers, Nancy J. Cooke, Jamie Gorman, and Nathan McNeese (Topic Editors).

Estimating Systemic Cognitive States from a Mixture of Physiological and Brain Signals

Matthias Scheutz,[a] Shuchin Aeron,[b] Ayca Aygun,[a] J.P. de Ruiter,[a,c] Sergio Fantini,[d] Cristianne Fernandez,[d] Zachary Haga,[a] Thuan Nguyen,[a] Boyang Lyu[b]

[a]*Department of Computer Science, Tufts University*
[b]*Department of Electrical and Computer Engineering, Tufts University*
[c]*Department of Psychology, Tufts University*
[d]*Department of Biomedical Engineering, Tufts University*

Received 26 April 2021; received in revised form 16 May 2023; accepted 16 May 2023

Abstract

As human–machine teams are being considered for a variety of mixed-initiative tasks, detecting and being responsive to human cognitive states, in particular *systematic cognitive states*, is among the most critical capabilities for artificial systems to ensure smooth interactions with humans and high overall team performance. Various human physiological parameters, such as heart rate, respiration rate, blood pressure, and skin conductance, as well as brain activity inferred from functional near-infrared spectroscopy or electroencephalogram, have been linked to different systemic cognitive states, such as workload, distraction, or mind–wandering among others. Whether these multimodal signals are indeed sufficient to isolate such cognitive states across individuals performing tasks or whether additional contextual information (e.g., about the task state or the task environment) is required for making appropriate inferences remains an important open problem.

In this paper, we introduce an experimental and machine learning framework for investigating these questions and focus specifically on using physiological and neurophysiological measurements to learn classifiers associated with systemic cognitive states like cognitive load, distraction, sense of urgency, mind wandering, and interference. Specifically, we describe a multitasking interactive experimental

Correspondence should be sent to Matthias Scheutz, Department of Computer Science, Tufts University. 420 Joyce Cummings Cente, 177 College. Avenue, Medford, MA, 02155 USA. E-mail: matthias.scheutz@tufts.edu

setting used to obtain a comprehensive multimodal data set which provided the foundation for a first evaluation of various standard state-of-the-art machine learning techniques with respect to their effectiveness in inferring systemic cognitive states. While the classification success of these standard methods based on just the physiological and neurophysiological signals across subjects was modest, which is to be expected given the complexity of the classification problem and the possibility that higher accuracy rates might not in general be achievable, the results nevertheless can serve as a baseline for evaluating future efforts to improve classification, especially methods that take contextual aspects such as task and environmental states into account.

Keywords: Physiological and brain sensing; Classifying cognitive states; Cognitive state inferences; Domain generalization; Machine learning

1. Introduction

Recent advances in robotics and autonomous systems point to a future where humans and machines will jointly perform tasks, ranging from collaborative manufacturing with industrial co-robots, to the many harvesting scenarios in agriculture, search and rescue operations after natural disasters, deep space missions, and many more. Imagine a joint ground-air search and rescue mission in an urban environment after an earthquake where a team of first responders is tasked to conduct a search for wounded people in collapsed buildings. The mission is supported by an autonomous system *S* consisting of various networked unmanned ground vehicles (UGVs) and unmanned air vehicles (UAVs) that can monitor a variety of important *systemic cognitive states* of their human team members in real-time while performing their own assigned tasks (e.g., see Scheutz, DeLoach, & Adams, 2017 for the description of a computational framework). As two human searchers are trying to deploy communication devices with the help of UGVs, first their *sense of urgency* and subsequently their *cognitive workload* are both increasing as the process turns out to be more complicated and takes longer than expected, while the third member's *vigilance* is dropping as she is watching out for inclement weather activity. The search leader, in the meantime, is becoming increasingly *distracted* due to difficulties with her communication device. *S* notices a lack of team cohesion due to cognitive state changes in the individuals and takes immediate action. First, *S* tasks two UAVs to explore the areas down the road, knowing that this task will have to be done next. *S* then tasks the closest UGV to provide a verbal update on the UAV mission to the *mind wandering* team member, quickly restoring *vigilance and alertness*, and proposes that the member helps the two other struggling teammates, which lowers their *workload* and prevents *urgency* from increasing further. As the UAVs report additional areas with potentially trapped human survivors to the southeast and northwest, *S* relays that information to the search leader through the closest UGV, which subsequently refocuses the leader's attention on the areas still to be searched.

The autonomous system *S* in the above scenario was able to intervene and proactively support the team by being aware of human cognitive states (indicated in italics) and then using its explicit task knowledge (e.g., the need to perform surveillance operations) to make autonomous decisions to act in the interest of task goals and interact with humans to improve team coherence. One way in which *S* could obtain the necessary information about its teammates' systemic cognitive states is through monitoring their physiological signals which often carry important information about human performance and possibly workload or interference. Understanding the extent to which such systemic cognitive states can be inferred from a mixture of physiological and neurophysiological signals is thus of great interest for understanding the different effects these states can have on human task performance and thus also on team effectiveness. Moreover, being able to detect such cognitive states, in particular, ones that lower performance, can form the basis of interventions to mitigate states that lower and move toward states that improve task performance (e.g., refocusing attention after distraction, removing lower priority tasks to reduce workload, engaging with regular activities to prevent mind wandering, etc.).

While there has been increasing interest in developing experimental paradigms to develop multimodal data sets which can form the basis for developing detection algorithms (e.g., learning appropriate classifiers), there is currently no available multimodal data set that comprises a comprehensive set of physiological (e.g., heart and respiration rate, arterial saturation and blood pressure, skin conductance) and neurophysiological data (e.g., functional near-infrared spectroscopy (fNIRS), electroencephalogram (EEG), and eye gaze) paired with behavioral measures (e.g., communication events, as well as task-based actions, such as braking in a driving task or performing detection response tasks). Yet, such a comprehensive data set is needed for developing a comprehensive understanding of which combination of signals (if any) can be used for developing (reasonably) accurate inference methods of various systemic cognitive states. While we would not expect there to be a perfect alignment with any subset of signals and systemic cognitive states, the verdict is still out on whether there exists a *sufficiently systematic correlation between the measured signals and systemic cognitive states that can be utilized by machine learning methods to develop corresponding classifiers that work across individuals*.

The goal of this paper is thus two-fold: (1) We present data from an experimental paradigm aimed specifically at developing a comprehensive multi-modal data set for studying systematic cognitive states; and (2) we use the data set for a first evaluation of standard machine learning methods using various types of physiological signals, including fNIRS, EEG, and eye gaze (pupil diameter) to determine the extent to which they are able to make reasonably accurate inferences about human cognitive states from a subset of the multimodal signals across subjects. Note that while traditionally machine learning methods train and test the learned model on data from the same individual, we are tackling the more challenging setting where a model is trained on several participants but needs to generalize to new individuals which we addressed by applying an advanced technique called *domain generalization* to improve the generalization capability of the learned models. The results of these efforts not only demonstrate the potential and limitations of domain generalization methods, but more importantly can serve as a baseline for future methods that include additional constraining factors such as task context and observable events in the task environment to push the classification accuracy to what can be achieved at best without additional individual adaptations of the models.

Table 1
Examples of prior work using five sensing modalities to measure any of the five cognitive states we are investigating.

Cognitive state	fNIRS	EEG	Card.v.	Skin c.	Eye gaze
Cognitive load	Causse, Chua, Peysakhovich, Del Campo, and Matton (2017)	Berka et al. (2007)	Stuiver and Mulder (2014)	Mehler, Reimer, Coughlin, and Dusek (2009)	Palinko, Kun, Shyrokov, and Heeman (2010)
Distraction	Ozawa and Hiraki (2017)	Wang, Jung, and Lin (2015)	Beckers, Schreiner, Bertrand, Mehler, and Reimer (2017)	Rajendra and Dehzangi (2017)	Liang and Lee (2008)
Sense of urgency	Holtzer et al. (2017)	Cheng (2017)	Liu, Lu, Huang, and Fu (2017b)	Kurniawan, Maslov, and Pechenizkiy (2013)	Liu, Hsieh, Lo, and Hwang (2017a)
Mind wandering	Durantin, Dehais, and Delorme (2015)	Baldwin et al. (2017)	Keller, Ruthruff, and Keller (2017)	Blanchard, Bixler, Joyce, and D'Mello (2014a)	Grandchamp, Braboszcz, and Delorme (2014)
Interference	León-Carrion et al. (2008)	Cooper et al. (2015)	Canabarro, Garcia, Satler, and Tavares (2017)	Collet, Petit, Priez, and Dittmar (2005)	Chatham, Frank, and Munakata (2009)

Abbreviations: Card. v., cardiovascular activity; EEG, electroencephalogram; fNIRS, functional near-infrared spectroscopy; Skin c., skin conductance.

2. Human cognitive states

It is well-known that various human cognitive states can significantly influence individual task performance and thus are likely to affect team behavior as well. Among these task-relevant states are *cognitive load* (Cooper, Medeiros-Ward, & Strayer, 2013), *distraction* (Strayer et al., 2015), *sense of urgency* (Ordonez & Benson, 1997), *mind wandering* (Kane et al., 2007), *vigilance* (McIntire, McKinley, Goodyear, & Nelson, 2014), and *interference* (Appelbaum, Boehler, Davis, Won, & Woldorff, 2014). Various methods have been proposed in the literature to measure these states (see Table 1), including using complementary brain sensing techniques involving EEG and fNIRS, which can also be combined to investigate neurovascular coupling (i.e., the relationship between neuronal activation and associated blood flow changes Tong et al., 2005; Dutta, Jacob, Chowdhury, Das, & Nitsche, 2015). While EEG is directly sensitive to neuronal activity, fNIRS is sensitive to hemodynamic changes associated with brain activity as well as systemic physiological changes. To take into account potential systemic contributions to the fNIRS signal, it is thus important to also monitor heart rate, arterial saturation, respiration, and arterial pressure (e.g., Fantini, Aggarwal, Chen, Franceschini, & Ehrenberg, 2003; Kainerstorfer, Sassaroli, Tgavalekos, & Fantini, 2015). These systemic measurements serve a dual purpose: first, they can help isolate brain-specific components of the fNIRS signal; second, they provide complementary information on systemic physiological states which can be further enhanced by including eye gaze and skin conductance information.

There is a large number of prior studies attempting to characterize systemic cognitive states in terms of multimodal physiological signals. Here, we can only provide a brief excerpt with respect to the systemic cognitive states we are investigating in this work.

2.1. *Cognitive workload*

Khedher et al. collected both EEG and human gaze data from 15 students in a virtual learning environment for the classification of *cognitive workload* in two distinct groups: students who could complete the tasks successfully and students who could not (Khedher, Jraidi, & Frasson, 2019). This study reported *k-Nearest Neighbor* as the best classifier over other classification techniques. Another study used the fusion of EEG and fNIRS to assess cognitive workload by building independent classifiers for each sensor (Coffey, Brouwer, & van Erp, 2012). Then, the classification results were combined to calculate the final decision. However, the results of the fusion method did not show notable enhancement over just using EEG alone.

2.2. *Distraction*

Some other studies explored the fusion of different physiological signal modalities in predicting *distraction*. Engstrom et al. combined electrocardiogram (ECG), gaze position, and vehicle measurements, such as lane position and steering wheel, to determine the level of distraction of the participants (Engström, Johansson, & Östlund, 2005). The results of the fusion technique indicated better performance over using single physiological modalities in assessing distraction. Craye et al. introduced a driving simulation environment that includes multiple sensor modalities to extract different physiological features, such as depth map, heart rate, steering wheel, and pedal positions. The authors utilized *hidden Markov models* to fuse the extracted features along with the contextual information to estimate the driver's fatigue and distraction levels. Their results showed an accurate prediction of fatigue and distraction with the combination of various physiological signal types (Debie et al., 2019).

2.3. *Sense of urgency*

Relatively, few studies have investigated the effects of the fusion of multiple signal types on predicting the *sense of urgency*. In Khalaf et al. (2020), authors recorded various physiological signals, such as ECG, continuous blood pressure, respiration, impedance cardiogram, and facial electromyography (EMG) to assess the participant's challenge and threat states who completed three mental arithmetic tasks. Another study used blood volume pressure, galvanic skin response, and skin temperature to determine the anxiety levels of the participants (Šalkevicius, Damaševičius, Maskeliunas, & Laukienė, 2019).

2.4. Mind wandering

Mind wandering is another cognitive state which has been investigated by leveraging different types of physiological signal modalities to determine. Blanchard et al. combined skin conductance and skin temperature to assess mind wandering by training different supervised classification models (Blanchard, Bixler, Joyce, & D'Mello, 2014b). Bixler et al. used the fusion of eye gaze, skin conductance, and skin temperature along with contextual information such as task difficulty and time on task to determine automatic multimodal detection of mind wandering (Bixler, Blanchard, Garrison, & D'Mello, 2015). Grandchamp et al. examined the variations in the gaze position, the frequency of blinking, and the pupil size caused by mind wandering which was generated by a monotonous breathcounting task where the participants were asked to fix their eyes to a point, keep counting their breath, and report when they lose counting (Grandchamp et al., 2014).

2.5. Cognitive interference

There are also a few efforts on assessing *cognitive interference* based on different physiological biomarkers. Robertson et al. simultaneously recorded EEG and fNIRS during a multisource interference task. Their results indicate that the combination of EEG and fNIRS improves the performance of assessing cognitive interference (Robertson, Thomas, Prato, Johansson, & Nittby, 2014). Nigbur et al. collected EEG and electrooculogram to investigate the influence of EEG theta activity on multiple sources of cognitive interference. Their results demonstrated the sensitivity of the theta power to the recruitment of executive control in interference circumstances (Nigbur, Ivanova, & Stürmer, 2011). In González-Villar, Samartin-Veiga, Arias, and Carrillo-de-la Pe na (2017), authors calculated the slope of the power spectrum density (PSD) taken from EEG and considered it as an indicator of neural noise and investigated the variations in neural noise during cognitive interference in fibromyalgia patients. Their results demonstrated that neural noise increases during cognitive interference.

2.6. Experimental conditions for inducing five systemtic cognitive states

In order to assess which combinations of physiological and brain measures are best suited for cognitive state inferences and, for future work, which context-based aspects might be needed for accurate classifications, we developed an experimental paradigm that allowed us to instantiate different systemic cognitive states naturally as part of a driving task and collect a comprehensive multimodal data set that could be used to train machine learning models to classify these states. The experimental task thus needed to allow for the instrumentation of participants with a full suite of physiological and brain sensors (see Fig. 1) without having too much of an impact on their task performance. Moreover, the tasks needed to allow for controlled variations to induce *cognitive load*, *distraction*, *sense of urgency*, *mind wandering*, and *interference*. We settled on a seated driving task in a driving simulator where participants only needed to use their right hand for steering and their right foot for operating the gas and brake pedals. There was minimal head motion involved and we were able to reduce

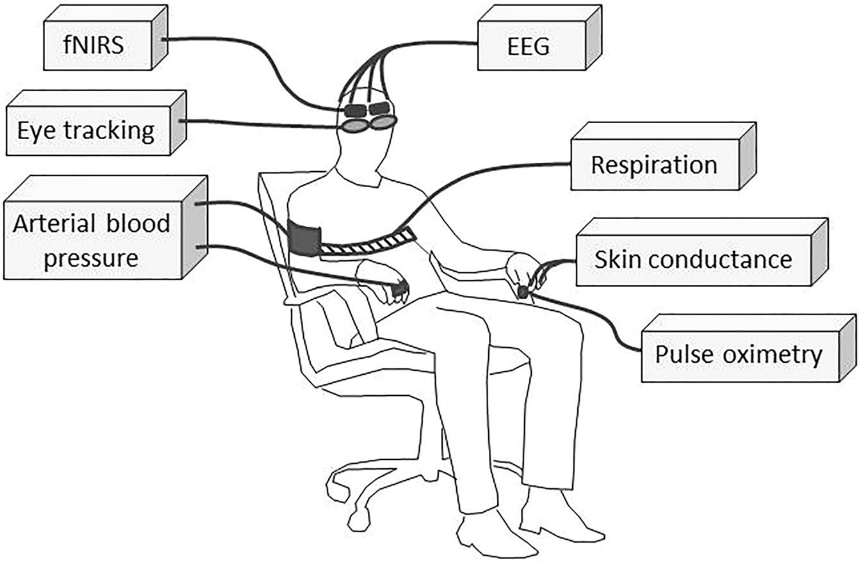


Fig. 1. Schematic of the experimental setup for the collection of various physiological signals during the driving simulator protocol, including functional near-infrared spectroscopy (fNIRS) and electroencephalogram (EEG).

motion artifacts in EEG and fNIRS. During driving, participants had to respond to various environmental situations, such as avoiding crashing into a braking automobile, or situations that restricted participants' reactions (e.g., being surrounded by cars while traveling under an overpass). In addition to varying the amount of traffic or the need for braking to increase or decrease the task difficulty, we added two secondary tasks in some experimental conditions that participants had to perform during driving while avoiding accidents. The first was the tactile *detection response task (DRT)* for Standardization Road Vehicles—Transport Information and Systems (2016) task which participants had to perform continuously. This is a validated method for assessing cognitive workload while driving if used right (see also Stojmenova & Sodnik, 2018). The second was a communication task where drivers had to respond to different types of questions.

Our motivation for choosing this particular combination of tasks was based on considerations of ecological validity. The situation in which someone is driving while traffic conditions may require braking, and a conversation with a passenger takes place simultaneously is one that occurs frequently in real life. The DRT task roughly corresponds with the driver wearing a smartwatch that vibrates when someone calls them, and that the driver subsequently has to cancel.

The multitask setting then allows for the definition of specific conditions that participants encountered periodically throughout the experiment and which would best induce the five cognitive states we discussed in Table 2: *cognitive load*, *distraction*, *urgency*, *mind wandering*, and *interference*.

Table 2
Experimental conditions for inducing five systemic cognitive states

Cognitive state	Events or conditions
Cognitive load	DRT, braking, communication
Distraction	DRT, car proximity, on-ramp areas
Sense of urgency	communication, car proximity
Mind wandering	baseline
Interference	DRT, brake pedal

Abbreviation: DRT, detection response task.

2.6.1. *Cognitive load*

Cognitive load is expected to increase with increasing task demands over a longer period of time, that is, when there is not only the visual driving task but also the tactile perception of the DRT as well as the auditory awareness and engagement of answering the questions. High cognitive load during the experiment is expected when participants have regular communication events paired with sudden braking events while also performing the DRT.

2.6.2. *Distraction*

Distractions are events or tasks that interrupt the continuity of driving performance and require instantaneous adaptation and reaction to a changed driving environment. Distraction episodes are defined as periods of time where peripheral task demands such as the DRT could interrupt the focus or attention of the participant (e.g., during a braking event).

2.6.3. *Sense of urgency*

The operational definition of urgency is when there are environmental objects and task demands that insist on a timely manner of reaction such as when one needs to answer a question (performing a communication task) quickly or sudden changes to the driving environment that require immediate braking.

2.6.4. *Mind wandering*

During the first 3 minutes of the experiment, peripheral traffic and road conditions were at a minimal level and no task demand arises (i.e., no communications, no braking required, and no DRT) other than driving in the right lane on a straight highway. This portion of the experiment was used as an introduction to the experiment and used to extract a baseline measurement of task performance. This time period of the experiment was also defined as a period that would evoke the mind wandering mental condition.

2.6.5. *Interference*

These are cognitive resource interferences which could affect the performance of the participant, for example, having to perform a physical brake action to avoid a collision while also having to perform another independent physical action to respond to the DRT simultaneously.

Note that the above systemic cognitive states can overlap as the same external event (e.g., braking event) can contribute to multiple different systemic states (e.g., sense of urgency, distraction, and interference). Moreover, whether the external events (i.e., the experimental manipulations) actually cause a particular systemic cognitive state also depends on the individual, how they allocate cognitive resources, and how well they are able to cope with the task demands. For example, whether a DRT event causes distraction will depend on an individual's focus of attention and ability to multitask. Similarly, whether a braking event causes interference with the DRT will depend on the individual's ability to focus on and perform simultaneous physical tasks (omissions of DRT responses in some individuals, e.g., are indicative of interference when braking is required at the same time). Consequently, being able to generalize cognitive states across multiple instances of the same type of event not only within a single subject, but also across subjects is a challenging problem that in general will require some form of calibration that is beyond the work reported here. Rather, we will tackle the simpler, but still difficult problem of trying to detect the same systemic cognitive state type across multiple instances within the same subject.

2.7. *Comparison with other work using driving tasks*

Neubauer et al. introduced an autonomous simulated driving platform to infer human cognitive states by leveraging stochastic filtering which is then used to determine the decision for engaging or disengaging the driving assistant. They used several physiological signal types, such as electro-dermal activity (EDA), electroencephalography (EEG), heart rate, and heart rate variability (Bixler et al., 2015). Although this paper provided a way to explore the effects of different signal modalities on estimating human cognitive states, it did not explore some important physiological markers, including the morphological characteristics of fNIRS, blood pressure, and respiration signals which have the potential for an accurate estimation of human cognitive states. In our study, we acquired an extended number of signal types, including EEG, fNIRS, human gaze, arterial blood pressure (ABP), skin conductance, and respiration to further investigate their capability on predicting cognitive states.

Another research study explores drivers' stress levels by utilizing multiple signal types recorded from 22 participants, including EDA, ECG, and EEG in a driving simulation environment (Mühlbacher-Karrer et al., 2017). Despite this paper presented a method for the assessment of drivers' stress levels based on cellular neural networks (CNNs) with the help of multiple sensor modalities, their data set has an inadequate number of participants which might cause their model to overfit.

Zahabi et al. investigated the effectiveness of using video-based methods to learn advanced driver-assistance systems which would be used to reduce high crash rates associated with degradations in older people's cognitive and physical abilities (Zahabi, Razak, Shortz, Mehta, & Manser, 2020). The authors used fNIRS and EEG collected from 20 older participants who have an average age of 63.1 years and leveraged them to measure the degradation of participants' cognitive capabilities. Similarly, this data set includes a lower number of participants within a specific age range. We generated our data set with a sufficient number of participants which allows us to develop machine–learning models that generalize well.

Huang et al. conducted a study to assess drivers' mental workload in a simulated driving platform. The authors recorded multiple signal types such as EEG, ECG, and EDA to predict drivers' cognitive load in a NASA-TLX setting by leveraging state-of-the-art machine learning methodologies, including XGBoost, CNN, long short-term memory (LSTM), and the fusion of CNN and LSTM (Huang, Liu, & Peng, 2022). In our study, we used an advanced methodology called domain generalization to enhance the generalization performance.

Brouwer et al. examined the physiological impacts of participants' behaviors as a response to real driving in an adaptive cruise control (ACC) system (Brouwer et al., 2017). The authors recorded heart rate and EEG signals from 15 participants and specifically focused on heart rate and blink responses to the participants. Even though this paper examines the variations in heart rate and blinks as a response to ACC behavior, the data set also provides an insufficient number of participants. Moreover, the stress level of the participants was generated based only on acceleration and deceleration events. In our experimental setup, the stress levels of the participants were generated by different combinations of secondary tasks added to the primary driving task. The secondary tasks, such as braking, communication, and tactile stimulation and their various combinations, are essential for investigating different types of cognitive states.

Overall, none of the above experiments include interactive components with other humans such as the communication events included in our paradigm. Most other studies also used a much smaller number of subjects, did not utilize multitasking paradigms to cause changes in systemic cognitive states, and did not collect a similarly comprehensive set of relevant physiological and neurophysiological data together with behavioral and event data that can also be used for determining the extent to which context information is necessary for inferring the requisite systemic cognitive states.

3. Methods

3.1. Participants

One hundred and thirteen participants from the local community were recruited to participate in a single session study that lasted approximately 120 min. Thirty-three participants were excluded: 14 due to technical issues and 19 due to simulator sickness or other discomfort. In our final data set of 82 participants, the average age was 20 years old[1] (standard deviation of 3 years), 46.8% identified as female and the remainder identified as male, all were right-handed, had normal or corrected to normal vision, had a valid driver's license, and drove at least 1 day a week on average. Participants were compensated \$20 ($n = 18$) or 2 h of research credits for an introductory Psychology course ($n = 62$). The research protocol was approved by the Institutional Review Board of Tufts University and in accordance with the Declaration of Helsinki.

3.2. Equipment and measurements

This study utilized a medium fidelity partial-cab driving simulator. Software and hardware were provided by RTI (Ann Arbor, MI). The simulator displayed the environment via five 45-inch liquid crystal displays which created a 180-degree field of view of the forward road scene. The partial-cab had a working steering wheel, brake pedal, accelerator pedal, and automatic gear shifter. A straight four-lane (two lanes in each direction) highway was simulated. The highway environment was lined with trees, had clear weather, and took place during the day. Traffic was light, with cars in the left lane passing the driver roughly every 30 s. The posted speed limit was 65 mph. The simulator generated images and recorded the driving data at 60 Hz. This included recordings of driver behavior from cameras and microphones. Audio of the driving environment was presented to participants through noise–canceling earbuds (Bose QuietComfort 20).

The tactile DRT was implemented using a cylindrical vibrotactile motor (14 mm in diameter and 4.5 mm thick) attached to the participants' right collar bone/shoulder. A response button was attached to their right index fingertip with hook and loop tape. Participants were instructed to respond to tactile stimuli that occurred randomly every 6–10 s. The motor vibrated for 1 s or until the button was pressed; whichever came first.

During the entire experimental session, various physiological signals were collected. A summary image of these measurements is shown in Fig. 1. fNIRS was measured by a NIRScout (NIRx Medical Technology, Berlin, Germany) device which consisted of light emitting diode source pairs (at wavelengths of 760 and 850 nm) fiber bundle coupled photo-diode detectors (see Fig. 2). These optical data were collected at 7.81 Hz. These data were complemented with a suite of physiological measurements; respiration, skin conductance, ABP, and peripheral oxygen saturation.[2] An RSP100C (BIOPAC Systems, Goleta, CA) respiration belt was attached around the participant's chest. On the participant's left hand, skin conductance was measured with an EDA100C (BIOPAC Systems) EDA sensor as well as a NIBP100D (BIOPAC Systems) beat-to-beat finger plethysmography system. An OXY100E (BIOPAC Systems) finger clip pulse oximeter was attached to the left thumb. This suite of data streams was collected at 20 samples per second.

EEG was collected at 500 Hz using an 8-channel Enobio (Neuroelectrics, Cambridge, MA, USA) system. 3.14 cm^2 silver/silver chloride electrodes were placed at the international 10-10 system locations FC1, FC2, FC5, FC6, CP1, CP2, CP5, and CP6.

A Pupil Core (Pupil Labs, Berlin, Germany) was used to collect eye movements, pupil diameter, and blink rate during this experiment. This eye tracker contained dual 200 Hz eye cameras and a 120 Hz world camera.

Lab streaming layer was used to synchronize and aggregate the time series data across different data acquisition devices and programs via a dedicated, high-bandwidth computer network (Asus GT-AC5300 router). Fig. 3 shows a system schematic of this instrument synchronization through LSL.

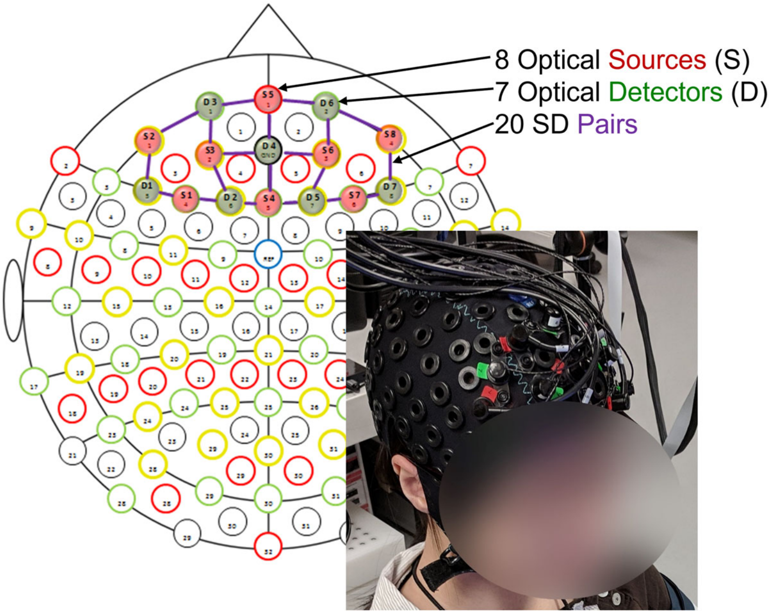


Fig. 2. Arrangement of functional near-infrared spectroscopy (fNIRS) optodes over the participant's prefrontal cortex, and an image of the fabric cap used to apply the array.

3.3. *Experimental procedure*

Participants completed surveys on driving history and demographics. Next, they were brought to the driving simulator and were set up with the physiological monitoring equipment. The experiment consisted of two driving scenarios: one with the DRT and one without the DRT. The order in which these scenarios were presented was counterbalanced over participants. The DRT was setup and introduced to half of the participants at this time; the other half of the participants were setup and introduced to the DRT during the brake before the second half of the experiment. Participants were then introduced to the driving simulator. They were instructed to stay in the right lane for the entire drive and to maintain a comfortable and appropriate speed while keeping in mind the posted speed limit of 65 mph.

Each scenario was 37.4 km long and took approximately 25 min to complete. The beginning of the drive consisted of 5.4 km (approximately 3 minutes) of just driving to allow the driver to acclimate to the simulation. After this section of the drive, the DRT began in one of the two drives (counterbalanced across participants). For the remainder of the scenarios (regardless of the presence or absence of the DRT), participants periodically engaged in six braking events and in four lure braking events. Braking events consisted of a vehicle appearing 200 m in front of the driver. Participants approached this lead vehicle until it was 75 m ahead and then followed this lead vehicle at a fixed distance of 75 m for 20 s. At that point, the lead vehicle rapidly decelerated for 5 seconds while its brake lights activated. After a braking event, the lead vehicle rapidly accelerated away from the driver. Lure braking events were

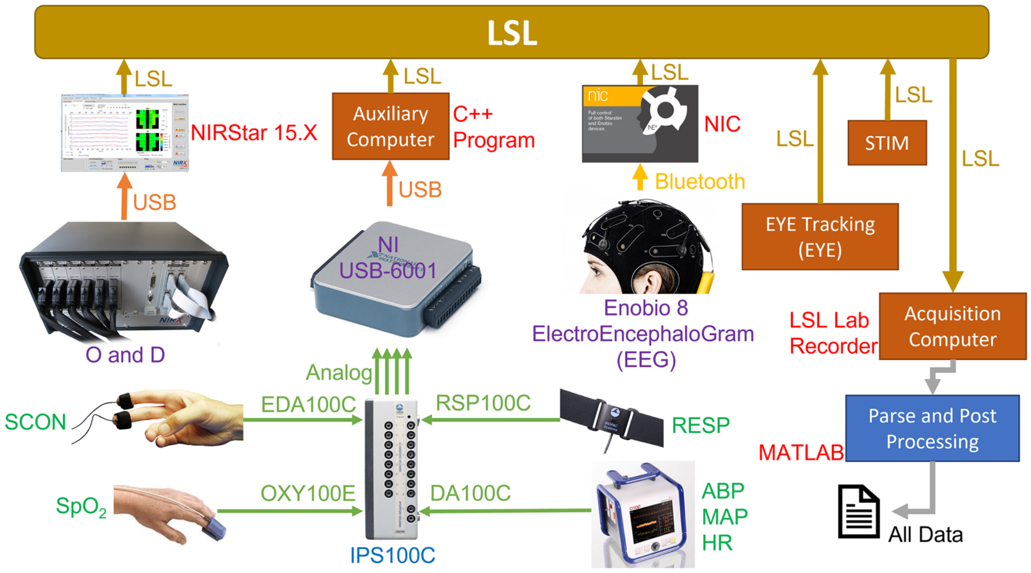


Fig. 3. Schematic of various data collection devices and coregistration.
Abbreviations: ABP, arterial blood pressure; D, de-oxyhemoglobin; HR, heart rate; LSL, lab streaming layer; MATLAB, matrix laboratory; O, oxyhemoglobin; RESP, respiration; SCON, skin conductance; SpO2, peripheral oxygen saturation; STIM, stimulus system including driving simulator and detection response task; USB, universal serial bus.

similar to real braking events; however, after 20 s, the lead vehicle accelerated away from the driver and did not brake. The braking and lure braking events were spaced out throughout the drive so that they were approximately 1–3 minutes apart. The events were presented in different orders across participants to minimize any possible impact of order effects, and the scenarios were otherwise identical in terms of the number and type of events.

During each scenario, participants responded to a series of basic fact questions about themselves (e.g., “Are you right-handed?”). Twenty questions were asked during each drive (for a total of 40 different questions), occurring roughly every 30–60 s. In addition, participants were allowed to rest briefly between the two scenarios. After the second scenario, participants filled out a final questionnaire that asked about aspects of the drive and contained the simulator sickness questionnaire.

3.4. Performance and behavioral label generation

Participant performance was measured by driving performance and reaction times to the various tasks (i.e., braking events, communication events, and DRT responses). These performance measures were used as behavioral labels. Considering the individual differences across participants, we assumed that event-based labels may not fully reflect the actual cognitive states for all participants. Thus, 13 additional labels were created based on the participant’s behavior and performance during communication events, braking events, the participant's control of the car's steering, the vehicle's heading angle and position. For the DRT signal, an additional label missed DRT was added due to the distraction of the participant.

Driving performance was assessed by the position offset from the middle of the lane, the vehicle's heading offset, and the change in the steering position. Data were consolidated into 1-s windows of data through the duration of the trial. Each driving performance category was partitioned into their best 15% and worst 15% of their windowed data.

Task performance was measured by the participants reaction times to the DRT, the communication events, and the braking events. A unique label was made if the participant did not respond to the DRT stimulus. The slow DRT label was the longest 25%, or 1st quartile, of the reaction times to the DRT and the fast DRT label was the shortest 25%, or 3rd quartile, of that participant's reaction times. The labels for the communication event were slow communication which was the longest 25% of the total individual's response times and fast communication was the shortest 25% of their reaction times.

3.5. Data preprocessing

As physiological data are known to be noisy, we deployed various data pre-processing techniques to remove noise and other artifacts that would negatively impact the training of machine learning models for inferring cognitive states.

3.5.1. Functional near-infrared spectroscopy and arterial blood pressure

Preprocessing of fNIRS began with the elimination of motion artifacts and drifts and the removal of channels with large noise. Raw asynchronous continuous-wave intensity measurements from each source-detector pair and each wavelength from the fNIRS instrument were first interpolated (using p-chip interpolation, matrix laboratory [MATLAB]) onto a continuous time axis with a sampling frequency of 20 Hz. Linear peace-wise detrending was then done by finding regions in which the variance of the signal is above the 75th and below the 25th percentiles, and using these as breakpoints in the detrend in order to remove any drifts or jumps in the signal that was not physiological.

Cleaned intensity measurements for each source-detector pair were then used to calculate changes in the concentration of oxy-hemoglobin (ΔHbO) and changes in the concentration of deoxy-hemoglobin (ΔHb) or changes in the concentration of total-hemoglobin (ΔHbT) using modified Beer–Lambert law (Blaney, Sassaroli, Pham, Krishnamurthy, & Fantini, 2019). A wavelength-dependent differential path-length factor based on absolute absorption coefficient (μ_a) and absolute reduced scattering coefficient (μ_s') values previously reported on healthy participants (mean age, 28 ± 4 years) taken on the forehead (Hallacoglu et al., 2012). Next, the noise in temporal ΔHbO (ΔHbO(t)), temporal ΔHb (ΔHb(t)), and temporal ΔHbT (ΔHbT(t)) were calculated above physiologically relevant frequencies to identify any channels with high instrumental noise for exclusion. Each signal was first high-pass filtered above 1.7 Hz (i.e., above heart rate). The windowed variance of the high-pass filtered signal was calculated, and the condition was set that if the median variance was above a threshold, the channel was neglected in further analysis. A threshold of 1 μM was used as this threshold for all signals.

ABP was measured using finger plethysmography (CNSystems CNAP Monitor 500, Graz, Austria). ABP was measured beat-to-beat on the subject's left index or middle finger, and it represents instantaneous ABP values, which thus provide systolic maxima and diastolic minima.[3] We interpolated ABP signals using p-chip interpolation, MATLAB onto the same 20 Hz time axis as hemodynamic signals. Recalibration of the arterial blood pressure (ABP) signal was a common issue, which caused an ABP reading that is of no use. Each ABP time trace was automatically searched for these segments, and in sections in which recalibration occurred, the segment was excluded from further analysis. We collected and processed the ABP signals for the further evaluation of their effects on an assessment of different cognitive states. However, this study does not include an analysis of the performance of ABP signals on cognitive state estimation.

3.5.2. *Electroencephalogram (EEG) Power spectral density (PSD)*

The PSD of EEG is one of the most widely used features of EEG signals (Qin, Zheng, & Chen, 2019; Hossain, Yaacob, & Nordin, 2021; Hamzah, Norhazman, Zaini, & Sani, 2016; Al-Nafjan, Hosny, Al-Wabil, & Al-Ohali, 2017). Specifically, PSD measures the power distribution of a given signal for each frequency in a given time-frequency transform (Stoica et al., 2005). From the raw EEG data, we extracted the EEG-PSD features using the five standard EEG frequency bands: δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and γ (30–100 Hz). Even though PSD is one of the most common extracted features for EEG signals, there is no common consensus on how to select the time window for the periodogram function. Specifically, if one selects a too narrow time window, the frequency analysis might be inaccurate, leading to a poor frequency resolution. On the other hand, a wide time window might give a better frequency resolution but also leads to a poor time resolution. Following the seminal work in Wang, Nie, and Lu (2014b), Garg et al. (2021), Zheng, Zhu, Peng, and Lu (2014), Zheng and Lu (2015), we decide to use a periodogram having a 1 s nonoverlapping rectangular window to estimate the PSD using the MATLAB Signal Processing Toolbox. Particularly, the periodogram PSD estimator produces the average spectral power over each frequency via discrete Fourier transform (Al-Nafjan et al., 2017; Hamzah et al., 2016). The spectral power is then integrated over each EEG frequency band to produce the EEG-PSD features. Finally, from eight EEG channels, using a time window of 1 s and five frequency bands, a 1 s frame of EEG-PSD data corresponds to a data matrix of size 40×1.

3.5.3. *Eye gaze (pupil diameter)*

Eye gaze is a good indicator of cognitive processes. Eye gaze data were recorded using a Pupil Core (Pupil Labs) eye tracker to obtain the pupillometry signal with a sampling rate of 400 Hz that contains a 120 Hz world camera and a 200 Hz binocular camera. Each pupil diameter sample contains two parameters: the left eye diameter and the right eye diameter. By averaging the pupil diameter of two eyes separately over a time window of 1 s, a 1 s frame of pupil diameter data corresponds to a data matrix of size 2×1.

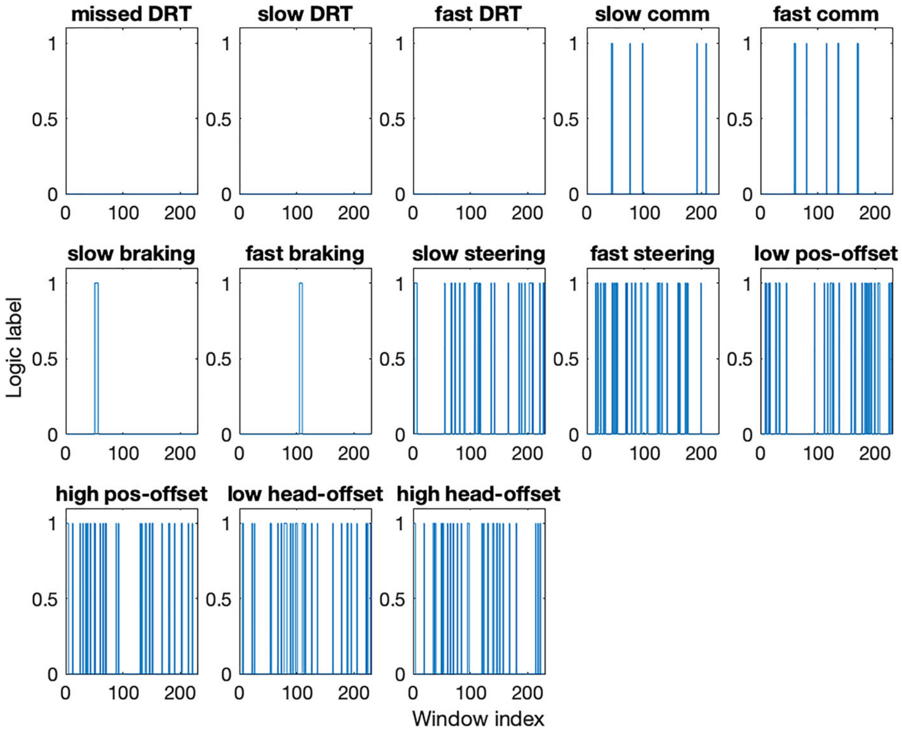


Fig. 4. Visualization of 13 behavioral markers with a window size of 6 s. The vertical axis is the binary label for each marker and the horizontal axis is the index of the window. Noting that the label imbalance happens in all the markers. In some markers, the imbalance is more severe than in others.

3.5.4. Data balancing

As previously discussed at the end of Section 3.4, 13 different logical markers were generated based on the participants' behaviors, such as DRT response time, communication event, braking event, steering event, the position of the car, and the heading error as the initial labels. These labels were automatically generated from the observed events and classified into two categories: fast/high and slow/low identified by a threshold. For example, the Slow Steering marker denoted the bottom 15% of the change in the steering wheel angle, while the Fast Steering marker took the top 15% of the change in the steering wheel angle, Heading Offset Low represented the lowest 15% of the participant's heading error, while Heading Offset High is the highest 15% of the participant's heading error. Fig. 4 visualizes these 13 behavioral markers.

If two behavioral markers are generated from the same participant's behavior but belong to different categories, then they are called extreme behavioral markers. For instance, since High Heading Offset and Low Heading Offset are both generated from Heading Offset behavior but belong to two different categories, that is, the highest 15% and the lowest 15% of the participant's heading error, they are a pair of extreme behavioral markers. Finally, from 13 behavioral markers, we form six extreme marker pairs: Slow DRT - Fast DRT, Slow

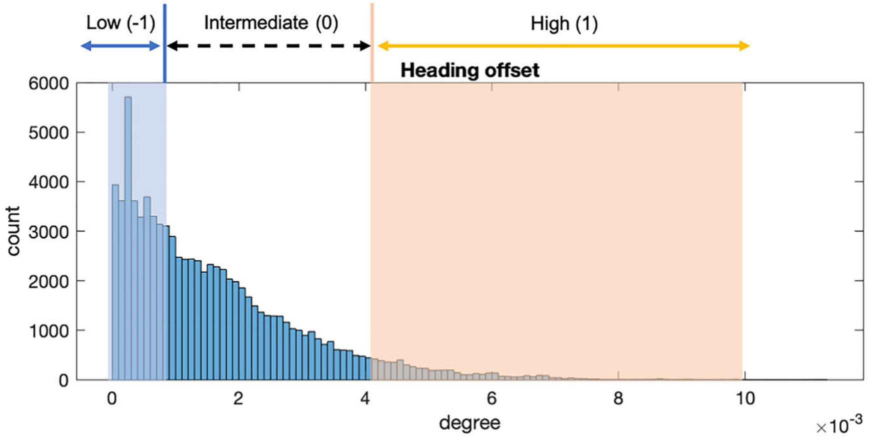


Fig. 5. Combining the Heading Offset Low class with the Heading Offset High class and removing the intermediate class leads to a balanced data set. The horizontal axis denotes the Heading Offset Error in degrees, while the vertical axis denotes the number of samples. Heading Offset Low (labeled by "–1") represents the lowest 15% of the participant's heading error, while Heading Offset High (labeled by "1") is the highest 15% of the participant's heading error. After removing the intermediate class, there is an equal chance to classify any data point into Heading Offset Low or Heading Offset High.

Communication - Fast Communication, Slow Braking - Fast Braking, Slow Steering - Fast Steering, Low Position Offset - High Position Offset, and Low Heading Offset - High Heading Offset.

Since these markers are constructed based on a biased threshold (e.g., 15% top and bottom of Heading Offset), the resulting labels are heavily imbalanced, making it difficult to quantify the classification performance. To address this problem, we combined two extreme behavioral markers from the same category and remove the intermediate labels to form balanced datasets. Taking the Heading Offset markers as an example, we first took the union of an extreme behavioral marker pair, that is, the union of High Heading Offset and Low Heading Offset labels, making it with three classes: Heading Offset Low (denote by "–1"), Heading Offset High (denote by "1"), and intermediate class (neither Heading Offset Low nor Heading Offset High, denote by "0"). Since both Heading Offset Low and Heading Offset High are based on the same threshold of 15%, these two classes must have the same number of samples. Therefore, removing the intermediate class will lead to a binary classification problem with balanced labels. In other words, after removing the intermediate class, any data point can only be assigned to two classes: Heading Offset Low ("–1") or Heading Offset High ("1") with an equal chance of 50%. Fig. 5 illustrates our data balancing process using Heading Offset Low and Heading Offset High markers.

3.5.5. *Motivation of domain generalization*

Based on the balanced data, the goal is to develop a predictor/classifier such that given the measured data and behavioral states from several participants, that is, whether they are fast or slow when communicating, braking, steering, and so on in a particular experimental linked to the various cognitive states, one can predict the corresponding behavioral states for the new-coming participants. The key challenge is, in practice, the measured sensor data from new-coming participants may not share the same data distribution as that of the training participants (Duan et al., 2020; Han & Jeong, 2021; Raza, Rathee, Zhou, Cecotti, & Prasad, 2019; Wu, Xu, & Lu, 2020; Zhao, Yan, & Lu, 2021), which violates the basic assumption in most traditional machine learning algorithms, requiring that the training and testing data are independently and identically distributed. This distribution shift phenomenon has been observed not only for fNIRS signals (Lyu et al., 2021) but also for EEG signals (Raza et al., 2019); Wu et al., 2020)). Recent works have shown that the performance of a predictor/classifier trained on the data from one group of participants usually degrades when testing on the data from another group of new participants. For instance, an fNIRS-based cognitive load estimator may not generalize well across different participants (Lyu et al., 2021), and a well-trained drowsiness-driving classifier based on EEG data performs badly when applied to the new participants (Cui, Xu, & Wu, 2019). Lots of work has been proposed for addressing the distribution shift problem, here we mainly focus on domain generalization (DG) (Blanchard, Lee, & Scott, 2011) methods, which aim to find the models that can generalize well on the new (unseen) participants. Since DG does not require accessing unseen (test) data during the training time, it is considered a realistic but challenging problem (Wang, Lan, Liu, Ouyang, & Qin, 2021). In the next section, we introduce some well-known DG methods and employ them to our problem to overcome the challenge of distribution shift.

3.5.6. Notations and problem formulation

The data (fNIRS/EEG-PSD) are first segmented by a nonoverlapping sliding window of size w. Since EEG-PSD has a higher temporal resolution together with a shorter time-response than fNIRS, we decide to use $w = 3$ s for fNIRS data and $w = 1$ s for EEG-PSD. For the pupil-diameter signals, we decide to select the same window $w = 1$ as EEG-PSD. The data set $\mathbf{X}$, therefore, is a collection of data segments and their labels, that is, $\mathbf{X} = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{X}_i$ denotes i^{th} segment, y_i denotes its corresponding label, and N denotes the number of segments. Each segment $\boldsymbol{X}_i$ corresponds to a tensor of size $(c) \times (f \times w)$, where c, f, and w represent the number of channels, the sampling frequency, and the size of sliding window, respectively. For instance, if the data are EEG-PSD, $c = 40$ (five bands with eight channels per band) and $f = 1$ Hz, then each data segment $\boldsymbol{X}_i$ with window size $w = 1$ s corresponds to a data matrix size 40×1. If the data are fNIRS-ΔHbO, $c = 20$ and $f = 20$, then each data segment $\boldsymbol{X}_i$ with window size $w = 3$ s corresponds to a data matrix size 20×60. If the data are pupil diameter, $c = 2$ (diameters of the left eye and the right eye) and $f = 1$ (the average value of pupil diameter in the same window), then each data segment $\boldsymbol{X}_i$ with window size $w = 1$ s corresponds to a data matrix size 2×1. The label y_i is one of the six extreme behavioral marker pairs: Slow DRT - Fast DRT, Slow Communication - Fast Communication, Slow Braking - Fast Braking, Slow Steering - Fast Steering, Low Position Offset - High Position Offset, Low Heading Offset - High Heading Offset. We follow the procedure described in Section 3.5 to form a balanced data set, that is, the label $y_i \in \{-1, 1\}$, where –1 and 1 denote the Slow/Low and Fast/High events, respectively.

3.6. Machine learning models for fNIRS, EEG, and eye gaze

Even though fNIRS and EEG are both linked to the same neural activity, these signals are very different but complementary to temporal and spatial resolution. Therefore, we use different models to deal with fNIRS data and EEG-PSD data.[4] Next, we describe how the learning models for fNIRS and EEG-PSD are separately selected.

3.6.1. Learning models for fNIRS

Motivated by the state-of-the-art time-series classification learning models in Karim, Majumdar, Darabi, and Harford (2019), Fawaz et al. (2020), we decide to use the multivariate long short-term memory fully convolutional network (MLSTM-FCN) (Karim et al., 2019), and InceptionTime (Fawaz et al., 2020) as two candidate models for fNIRS.

- *Multivariate long short-term memory fully convolutional network (Karim et al.*, 2019). The MLSTM-FCN model consists of two branches: an LSTM block and a fully convolutional network (FCN) block. Two blocks are operated in parallel, where one can be treated as an augmentation of the other. The FCN (Wang, Yan, & Oates, 2017) block is composed of three temporal convolutional blocks where each is followed by batch normalization and a rectified linear unit (ReLU) (Agarap, 2018) activation function. Squeeze-and-excitation blocks are added behind the first two convolutional blocks for input feature maps recalibration, while a global average pooling layer is added to the end of the last convolutional block. The output from the FCN block and long short-term memory (LSTM) block is concatenated and fed to a linear classifier for the final classification task. We keep all model parameters as same as the settings in the original paper (Karim et al., 2019) but set the number of LSTM cells as eight without applying a grid search.
- *InceptionTime (Fawaz et al.*, 2020). The InceptionTime model consists of an ensemble of five Inception networks that are initialized randomly to better stabilize the model. Each Inception network cooperates ResNet (He, Zhang, Ren, & Sun, 2015) modules with the inception modules where filters with various lengths are applied simultaneously to the input time series (Ruiz, Flynn, Large, Middlehurst, & Bagnall, 2020) for diverse feature extraction. The usage of the bottleneck layers (He et al., 2015) further reduces the model complexity and speeds up the training process. Again, we keep all the parameters of the model the same as the setting in the original paper (Fawaz et al., 2020).

3.6.2. Learning models for electroencephalogram (EEG) power spectral density (PSD)

For EEG-PSD data, we use a multi-layer perceptron (MLP) having two fully connected (FC) layers with a ReLU activation function (Agarap, 2018) followed by a linear layer as the learning model. To prevent the neural networks from overfitting, a dropout layer is added after the ReLU layer (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Indeed, the first two FC layers are aimed at extracting meaningful features, while the last linear layer acts as a classifier. It is worth noting that the MLP architecture is extensively used

Table 3
Multi-layer perceptron (MLP) architecture for EEG-PSD data

Layer	Operation	Output size
Input	–	(N, 40)
The first FC layer	Linear(40, 40) + ReLU + Dropout(0.25)	(N, 40)
The second FC layer	Linear(40, 32) + ReLU + Dropout(0.25)	(N, 32)
The last linear layer	Linear(32, 2)	(N, 2)

in literature for learning from EEG-PSD data (Arsalan, Majid, Butt, & Anwar, 2019); Katmah et al., 2021; Kuremoto, Baba, Obayashi, Mabu, & Kobayashi, 2015; Lin, Wang, Wu, Jeng, & Chen, 2007). For convenience, the learning model for EEG-PSD data is called power spectral network (PSD-NET). Details of the MLP structure for PSD-NET can be found in Table 3.

3.6.3. Learning models for eye gaze

Motivated by the state-of-the-art time-series classification learning models in Karim et al. (2019), Fawaz et al. (2020), we decide to use the MLSTM-FCN (Karim et al., 2019), and InceptionTime (Fawaz et al., 2020) as two candidate models for eye gaze (pupil diameter). We use the same model settings for MLSTM-FCN and InceptionTime as described for fNIRS.

3.6.4. Baseline algorithm

Based on these learning models, the Empirical Risk Minimization (ERM) algorithm serves as the baseline learning algorithm for fNIRS, EEG-PSD, as well as pupil diameter signals. In particular, ERM aims for minimizing the empirical risk (classification error) from all trained participants without employing any DG techniques.

3.6.5. Domain generalization algorithms

To address the distribution shift problem, three different DG methods: Maximum Mean Discrepancy-Adversarial Autoencoder (MMD-AAE) (Li, Pan, Wang, & Kot, 2018b), Meta-Learning Domain Generalization (MLDG) (Li, Yang, Song, & Hospedales, 2018a), and Correlation Alignment (CORAL) (Sun, Feng, & Saenko, 2017) are employed. Here, we utilize the implementation in DomainBed (Gulrajani & Lopez-Paz, 2020) and adapt the feature extractor in PSD-NET for each DG method.

- *Maximum Mean Discrepancy-Adversarial Autoencoder (Li et al.*, 2018b). MMD-AAE is an adversarial training-based DG method. MMD-AAE uses an adversarial autoencoder to align the distributions in the representation space of different domains via minimizing their MMD and matching the learned representation distribution to a prior distribution in an adversarial manner.
- *Meta-Learning Domain Generalization (Li et al.*, 2018a). MLDG is a meta-learning DG method that separates multiple seen domains into meta-train and meta-test domains for reducing the distribution shift and performing optimization which leads to an improvement in learning performance.

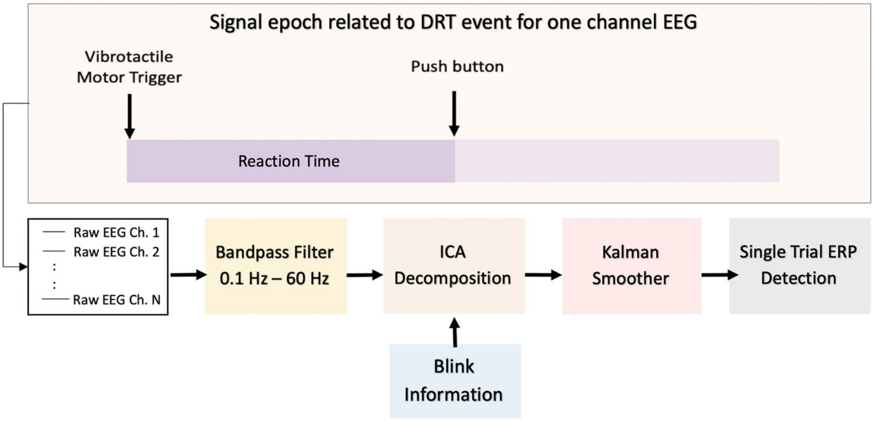

Signal epoch related to DRT event for one channel EEG
Vibrotactile Motor Trigger
Push button
Reaction Time
Raw EEG Ch. 1
Raw EEG Ch. 2
Raw EEG Ch. N
Bandpass Filter 0.1 Hz – 60 Hz
ICA Decomposition
Kalman Smoother
Single Trial ERP Detection
Blink Information

Fig. 6. Overview of single trial event-related potential (ERP) extraction.
Abbreviations: DRT, detection response task; EEG, electroencephalogram; ICA, independent component analysis.

- *Correlation Alignment (Sun et al.*, 2017).CORAL is a DG method that is based on the idea of matching the mean and covariance of feature distributions from different domains to perform domain alignment.

3.7. *Event-related potential estimation*

We also investigate single-trial event-related potentials (ERPs) during a DRT event which is shown in Fig. 6. The prediction of ERPs from EEG signals is significant to assess the cognitive states of an individual. The most common method to extract ERPs from EEG is to take the grand average of EEG channels from multiple trials with the aim of eliminating the sensor-based and ERP-independent neuronal activity noise. Although the averaging procedure is practical to determine the main morphological characteristics of event related potentials (ERPs), it does not provide a way to assess human's responses to specific types of stimulations which differ across trials (Cecotti & Ries, 2017). An ERP response is characterized by different brain waves that occur following the onset of the stimulus, such as N1, N2, and P3. N1 is assumed to appear between 90 and 200 ms after the onset of the stimulus (Sur & Sinha, 2009). Although there have been research works propose that N1 is correlated with selective attention (Thornton, Harmer, & Lavoie, 2007) or emotional stimulus (Hu et al., 2017), the early potentials are usually associated with physical and sensory stimulation (Golob et al., 2009). N2, which occurs between 180 and 325 ms following the stimulation, is related to the recognition and characterization processes of the brain (Patel & Azzam, 2005). P3 is evoked between 300 and 400 ms after the onset of the stimulation and is correlated with selective attention (i.e., higher attention generates higher P3 amplitudes (Sur & Sinha, 2009). All the mentioned ERP components are associated with attentional interest and mental workload (Ghani, Signal, Niazi, & Taylor, 2020). The ERP generation techniques, which aim to assess the cognitive workload, are investigated into two categories in terms of the task type that the participants accomplish: dual-task and single-task (Ghani et al., 2020). In this study, we utilize the dual-task technique by considering the driving task and the DRT event (pushing the button after the tactile stimulation) as the primary and the secondary tasks, respectively.

First, signal epochs of all EEG channels related to the DRT events are taken. The onset of the DRT event is assumed to be the stimulation of the vibrotactile motor fixed to the participants' right collar bone/shoulder (depicted at the top of Fig. 6). Then, a $6th - order$ Butterworth Bandpass Filter between $0.1Hz - 60Hz$ was applied to the EEG signal epochs to remove the out-of-band noise. Next, the mixture of signal epochs is decomposed into its statistically independent components via independent component analysis (ICA). The ICA component which is related to blink artifacts is removed by using blink information taken from gaze recording. To do this, ICA components are compared with the blink information, the component which includes instantaneous spikes on the amplitude at the same time as the blinks are determined, and the blink artifact-related ICA component is removed manually. A Kalman Smoother is utilized to smooth the blink artifact-removed EEG channels (Kalman, 1960). Finally, the ERPs are extracted from cleaned EEG epochs.

3.7.1. Independent component analysis

There are several motion artifacts induced by body movements and recording devices which contaminate EEG signals, such as eye movements, blinks, respiratory exertion, muscle, and cardiac activity (Louis et al., 2016). Among those, eye movements and blinks are considered as fundamental sources of motion-corrupted EEG signals (Joyce, Gorodnitsky, & Kutas, 2004). ICA is an effective tool to decompose a mixture of linear signal streams into its hidden components. ICA is a generative model and is implemented with the presumption that the latent components are statistically independent and non-Gaussian, and the number of components is the same as the number of input signal streams. In this study, ICA is expressed as follows:

$$\mathbf{s} = A \times \mathbf{c}, \tag{1}$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_M)$, $\mathbf{c} = (c_1, c_2, \ldots, c_M)$, and $A = [a_{ij}]$ for $i, j = 1, 2, \ldots, M$ represent the vector of the linear mixture of EEG channels, the vector of statistically independent hidden components, and the unknown mixing matrix, respectively. Here, A and $\mathbf{c}$ are unknown and the aim is to find the best predictor of M independent components, $\hat{\mathbf{c}}$, from M observations, $\mathbf{s}$, by estimating A^{-1} which is the inverse of mixing matrix A. Then, A^{-1} is used to obtain the latent components with the following expression:

$$\hat{\mathbf{c}} = A^{-1} \times \mathbf{s}. \tag{2}$$

Finally, the mixture of linear EEG signals is reconstructed with $\mathbf{s} = A \times \hat{\mathbf{c}}$. In this study, $M = 8$ which represents the number of EEG channels. To calculate the unmixing matrix A^{-1}, we have utilized fastICA algorithm that explores a linear combination of non-Gaussian components while increasing the statistical independence within the components as much as possible (Hyvärinen & Oja, 2000). Fig. 7 depicts one channel raw EEG signal and the blink information taken from pupil data. The amplitude of zero and one of the blink signals repre-

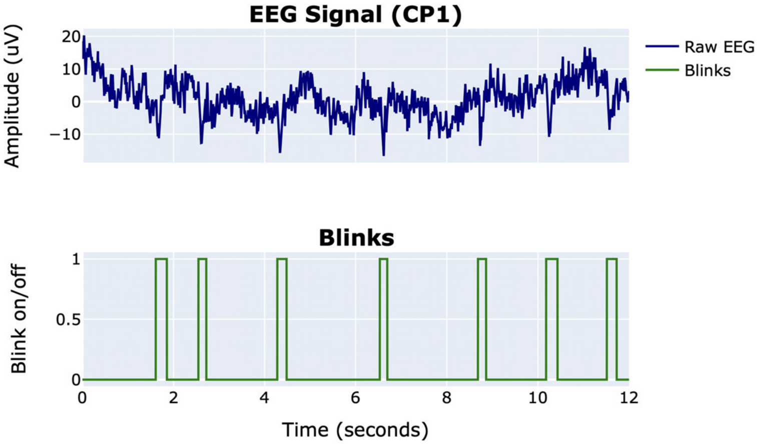


Fig. 7. Raw electroencephalogram (EEG) (Channel CP1) with blink information.

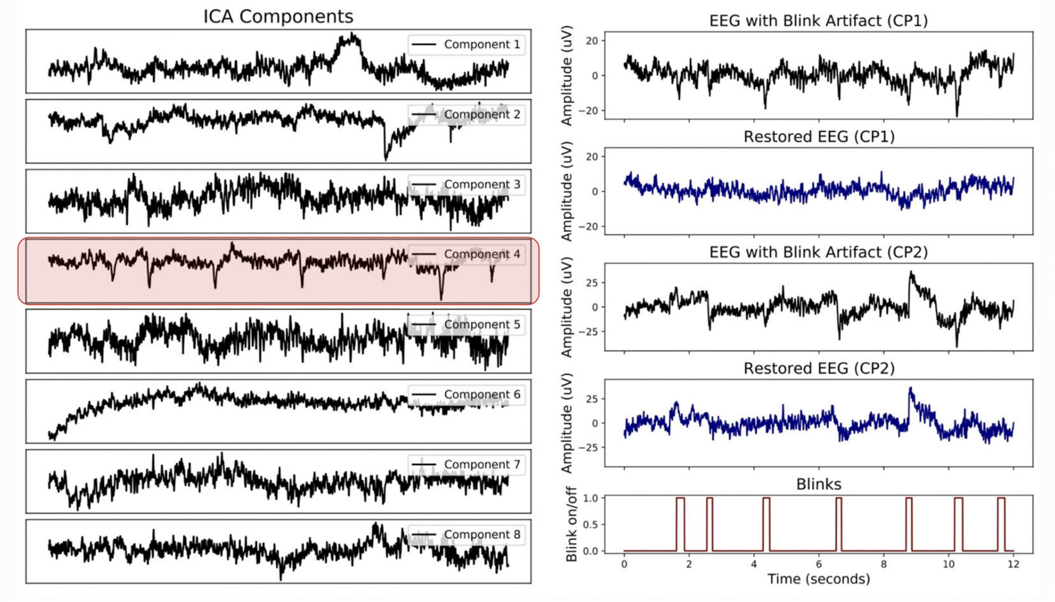


Fig. 8. Blink artifact removal: Independent component analysis (ICA) components where Component 4 represents the blink artifact (left) and different electroencephalogram (EEG) channels with and without blink artifact (right).

sent the eye opening and the eye closure, respectively. The blinks are observed on EEG signal as spikes with higher amplitude. To extract the blink artifacts, fastICA is applied to preprocessed EEG. Fig. 8 (left) shows the decomposed ICA components taken from one participant. Here, the component 4, which is related to the blink artifact, is removed to reconstruct the EEG channels. Fig. 8 (right) depicts two EEG channels (CP1 and CP2) before and after ICA removal. It can be seen that the spikes corresponding to blink artifacts do not appear in reconstructed EEG channels.

3.7.2. *Kalman smoother*

This is a common technique to predict the state of dynamic linear structures in the presence of noise Kalman (1960). It is a backward algorithm used to improve the estimation of previous states based on subsequent observations. In this study, we used the Python library called "tsmoothie" (Cerliani, 2021) to smooth the EEG signals.

3.8. *Eye gaze*

Human eye gaze is another benchmark to assess cognitive states which has the capability of exposing clues about mental conditions of a person, such as visual attention, situational awareness, cognitive workload, fatigue, emotional arousal, stress, comprehension, and immersion. However, a careful examination of gaze parameters is needed for an accurate prediction of cognitive states as multiple cognitive states may be linked to the same gaze parameter. For example, mean fixation duration is inversely correlated with the mental load during flight simulation (Holmqvist et al., 2011), while there is an explicit relationship between fixation duration and visual attention (Skaramagkas et al., 2021). A contextual information in addition to human gaze can be leveraged to interpret different conditions which influence the cognitive status of a human.

There are several gaze parameters, such as fixation, blink, saccadic movements, and pupil diameter. Fixation represents the preservation of eye gaze to a specific point (Skaramagkas et al., 2021). Fixation count, which is the number of fixations on a specific object, is inversely correlated with search efficiency (Bjørneseth, Renganayagalu, Dunlop, Hornecker, & Komandur, 2012). Another study suggests that a higher fixation count is related to a greater cognitive workload (Schmutz, Roth, Seckler, & Opwis, 2010). Longer fixation duration is associated with task difficulty and hardship in information selection (Wang, Yang, Liu, Cao, & Ma, 2014a). Blinks are the spontaneous opening and shutting movements of eyelids which are related to the mental exertion of an individual (Shojaeizadeh, Djamasbi, Paffenroth, & Trapp, 2019). One study indicates that blink frequency has a negative correlation with visual attention (Sakai et al., 2017). Another work associates decreased impulsive eye blink rates with the level of stress (Merkies, Ready, Farkas, & Hodder, 2019). Saccades are described as instantaneous eye movements between fixations (Shojaeizadeh et al., 2019). The amount of microsaccades has an opposite correlation with concentration level, while variation in saccadic activations is associated with mental fatigue as a result of time-on-task (Di Stasi et al., 2013; Buettner, Baumgartl, & Sauter, 2019). Pupil dilation, which represents the variations in pupil size, is another metric to evaluate the mental status of a human. Pupil dilation is directly related to the locus coeruleus activity which is effective in controlling physiological arousal and cognition (Eckstein, Guerra-Carrillo, Singley, & Bunge, 2017; Varazzani, San-Galli, Gilardeau, & Bouret, 2015).

In this study, we examine the variations in fixation counts during DRT events to which participants had to respond by pushing the button. For each DRT event, a 1 s time frame is

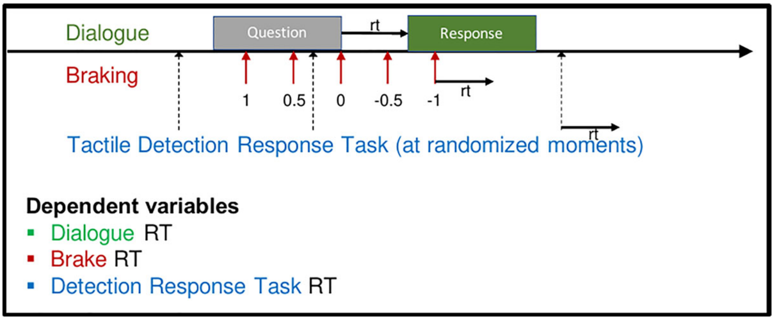


Fig. 9. Relationship between communication and braking event times (units in seconds) for a detection response task experiment session.
Abbreviation: RT, response time.

picked where the onset of the time frame is determined as the stimulation of the vibrotactile motor. Then, the 1 s time frame is divided into 100 ms time windows which are concatenated by 50 ms, and the fixation counts are calculated for every time window. Finally, the fixation counts are averaged over all 1 s time epochs of each participant.

We also explore the change in pupil diameter within the first 3 minutes of the experiment considering that the participants focus on the task at the beginning of the driving and they might gradually lose their attention as a result of mind wandering. We applied three-step preprocessing to denoise the pupillometry signal. First, we used amplitude thresholding to remove the signal partitions lower than 0.8 mm and greater than 10 mm by considering that the values lower than 0.8 mm are potential blink artifacts (Saeedpour-Parizi, Hassan, & Shea, 2020) and the pupil dilation is measurable up to 10 mm (Wildemeersch, Peeters, Saldien, Vercauteren, & Hans, 2018). Second, we applied linear interpolation to fix the extracted parts (Saeedpour-Parizi et al., 2020). Third, we utilized a fifth-order Butterworth low-pass filter with a cutoff frequency of 10 Hz to cancel baseline wander (Smallwood et al., 2011). Finally, we applied a moving average with a window size of 10 s.

4. Analyses and results

4.1. Communication and driving behavior

To test the effect of braking events on communication performance, and vice versa, we varied the relative timing of the communication events and braking events (see Fig. 9 showing the condition with DRT). The time difference between the end of the question and the beginning of the braking event was varied from –1 to +1 s, in steps of 0.5 s. This offset is termed stimulus onset asynchrony (SOA). A positive SOA value means that the braking event occurs before the end of the presented question, and a negative value means that the braking event occurs after the end of the question.

The dependent measure of relevance for the communication is the floor transfer offset (FTO), which is the time in seconds between the end of the question posed to the participant and the beginning of their articulation of a response. This value can be negative, in which case the response temporally overlapped with the question. The results regarding the FTO (which can be interpreted as a communication response time) were surprising: the participants were significantly faster in the condition with the DRT ($F(1, 717) = 7.35$, $p < 0.01$). They appeared faster in the SOA conditions (in which they simultaneously had to respond to a traffic event by pressing the brake pedal) than in the baseline condition without braking, but this difference was not significant in a mixed model with participant and communication event as random factors ($t(38) = 1.321$, $p = .19$).

In a substantial number of communication trials (15%), the participants used *filled pauses*, like "uh" or "uhm" at the beginning of their response. As using a filled pause signals an upcoming delay (a short one for "uh" and a longer one for "uhm," (Clark & Tree, 2002), this could indicate that they were delaying their response due to higher cognitive load. Therefore, we tested whether the proportion of filled pauses in the verbal responses of the participants was sensitive to the SOA condition and the presence/absence of DRT events. In a logistic regression using a mixed model with participants as a random factor, we found no significant effect of either factor on the proportion of filled pauses.

Please note that in the previous analysis of the FTO related to the presence of DRT and SOA conditions, we used the FTO as measured from the beginning of the articulation, so if the participant started with saying "uh(m)," the FTO values were measured with respect to the beginning of the articulation of "uh(m)." We repeated the same analysis with FTO values measured from the end of the "uh(m)" token (if there was an initial "uh(m)" token), as well as from the end of the pause following the "uh(m)" token (if there was an initial "uh(m)" token) and the pattern of result, in terms of what was significant and what was not, was identical.

Fig. 10 shows the communication response times in the different SOA and DRT conditions. There were no significant interactions between the factors SOA and DRT at the $\alpha = 0.05$ level.

For the braking response time, there was no significant difference for the presence or absence of the DRT or for the presence or absence of the communication event ($F(1, 708) < 1$), suggesting that there was no resource interference between the two tasks (see Fig. 11 for an overview of these braking data). There was also no significant difference between braking events which involved a simultaneous communication event and braking events that did not ($F(1, 876) = 2.01$, $p = .16$), lending further evidence to hypothesis that DRT did not interfere with the other tasks, that is, braking and DRT responses were not sharing cognitive resources as the presence or absence of one of them did not affect the reaction times of the other.

There was no significant difference in DRT response times for the participants who had the scenario with DRTs first, or second ($t(351) = 0.939$, $p = .35$). To look at the effect of time (fatigue) on DRT, we computed the Pearson correlation coefficient between the time of the DRT event in the scenario and the DRT response time. This correlation was 0.06, in the expected direction (longer response times after longer time) but was not significant ($t(381) = 1.093$, $p = .275$).

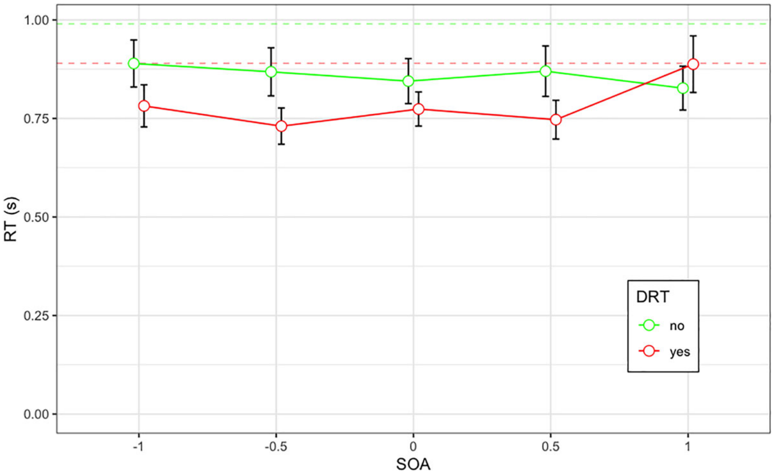


Fig. 10. Communication response time (RT) for different stimulus onset asynchrony (SOA) (in seconds) of detection response task (DRT) and non-DRT periods. Solid lines represent results during braking, while dashed lines represent the baseline period.

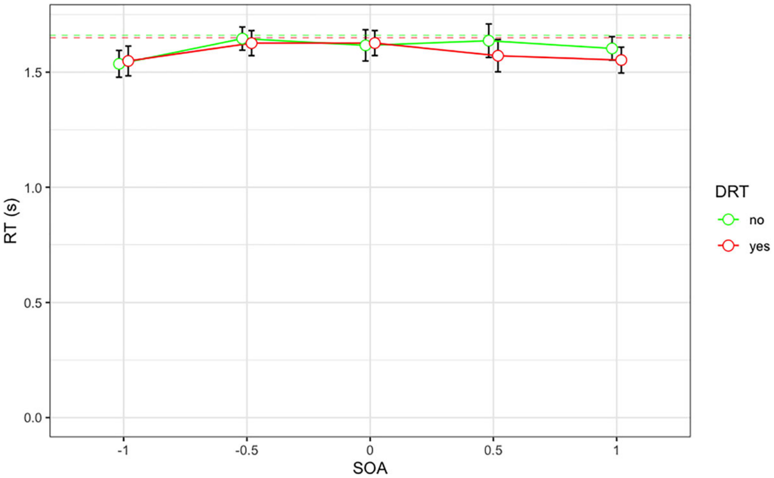
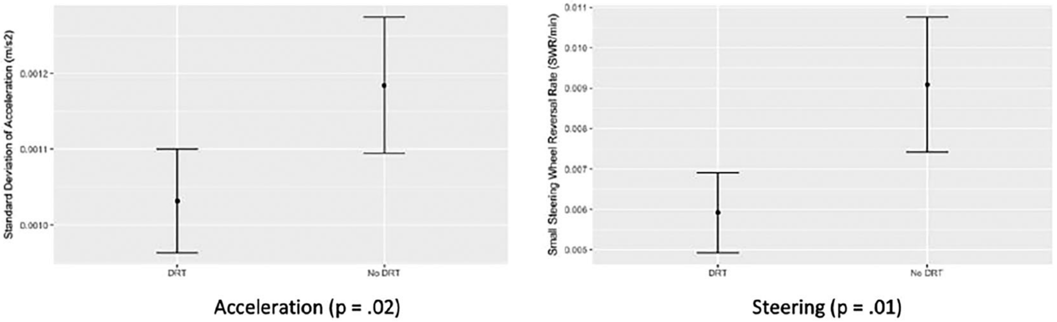


Fig. 11. Braking response time (RT) for different stimulus onset asynchrony (SOA) (in seconds) of detection response task (DRT) and non-DRT periods. Solid lines represent results during braking, while dashed lines represent the baseline period.

Surprisingly, the presence of the DRT task revealed a positive effect on the driving quality, as operationalized by the standard deviation of the acceleration, as well as the steering wheel reversal rate (see Fig. 12).

The DRT reaction times themselves were differentially sensitive to multitasking load caused by the communication events. We divided the DRT events into four categories. They were labeled as *in-question* if they occurred during the auditory presentation of the



Acceleration (p = .02)

Steering (p = .01)

Fig. 12. Standard deviation of acceleration and steering wheel reversal rate.

Table 4
Mean detection response task (DRT) response times (ms) in relation to communication events

In-question	In-transition	In-answer	Outside-communication
1236	1165	1046	1126

communication question. They were labeled as *in-transition* if they occurred right after the presentation of the communication question, but before the participant answered. They were labeled as *in-answer* if they occurred during the time the participant answered the question. Finally, they were labeled as *outside-communication* if they occurred outside any communication-related event. The resulting average DRT reaction times are in Table 4. The only significant difference in a post-hoc (Tukey) analysis was the difference between in-question and in-answer ($p < .01$). So, if the participant was listening to the question, their DRT response was substantially slower than when they were already articulating their answer. We know from recent work on communication processing that participants in conversation have to plan their response already while they are listening to the current turn (Magyari, De Ruiter, & Levinson, 2017; Levinson & Torreira, 2015). Our result suggests that simultaneous listening and response planning require attentional resources that slow down the DRT responses. However, once the articulation of the verbal response has been launched, there is no slowdown anymore. This suggests that the articulation of the planned response appears to be an automatic process that does not require extra attentional resources (see, e.g., (Levelt, Richardson, & La Heij, 1985)).

4.2. *Single-trial ERP extraction*

Fig. 13 shows the examples of ERPs taken from the CP1 channels of DRT sessions of two different participants. During these DRT events, the participants perform dual-task by responding to the tactile stimulation and maintaining the driving task simultaneously which causes a higher cognitive workload. The results show a precise estimation of N1, N2, and P3 components of ERPs for different DRT events taken from these two participants. It is seen

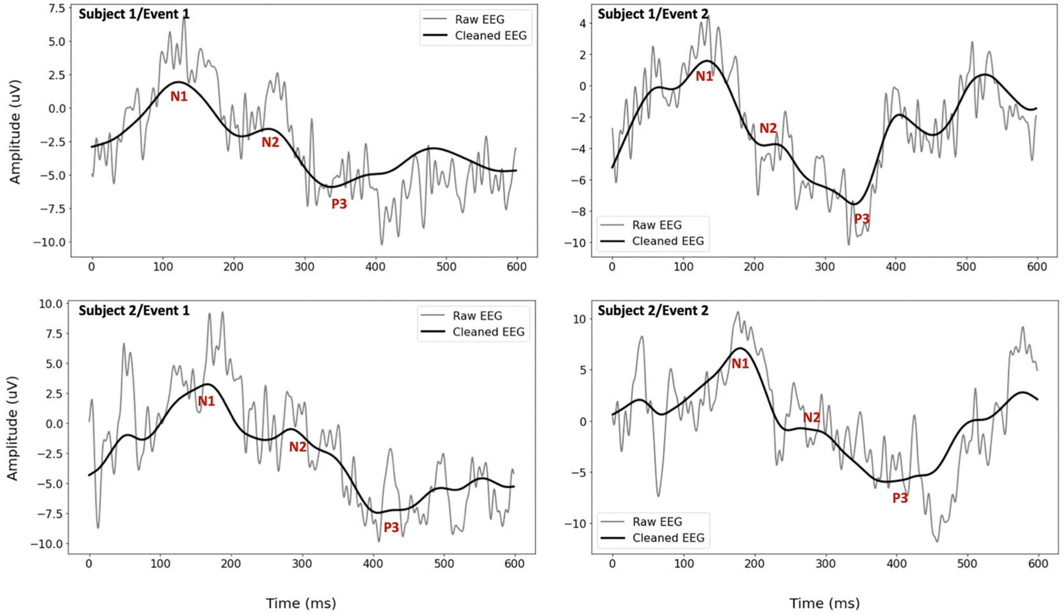


Fig. 13. Examples of event related potential (ERP) responses to detection response task (DRT) events for two events of two different participants.

from the figure that the ERP components are noticeable at specific time points after the onset of the secondary task which intensifies the cognitive workload of the participants. Yet, despite the accurate detection of ERPs for some DRT events, there are other cases where ERPs are not detected at similar points during experimental runs (even within the same subject) even though participants received the tactile stimulation and pushed the button. There are multiple possible explanations for this failure, the immediate one being that more discriminating computational methods for extracting ERPs might be needed. Alternatively, the participants' DRT responses might not always manifest themselves in ERPs (e.g., due to the variations in the cognitive context), in which case ERPs might just be of limited utility for inferring the kinds of systemic cognitive states.

4.3. Eye gaze

Fig. 14 shows the average fixation counts over the responded events of DRT sessions for four different participants. The results indicate that the average number of fixations increases after the stimulation. The DRT event occurs every 6 –10 s during the driving simulation where the participants were instructed to respond to the tactile stimuli. Both performing the driving task and responding to the tactile stimuli increase cognitive workload which results in increased fixation counts. Although there are many cases with an extended number of fixations during DRT events, there are also some reversed cases where a decreased number of fixations are observed, pointing to the limitations of using a single modality for estimating

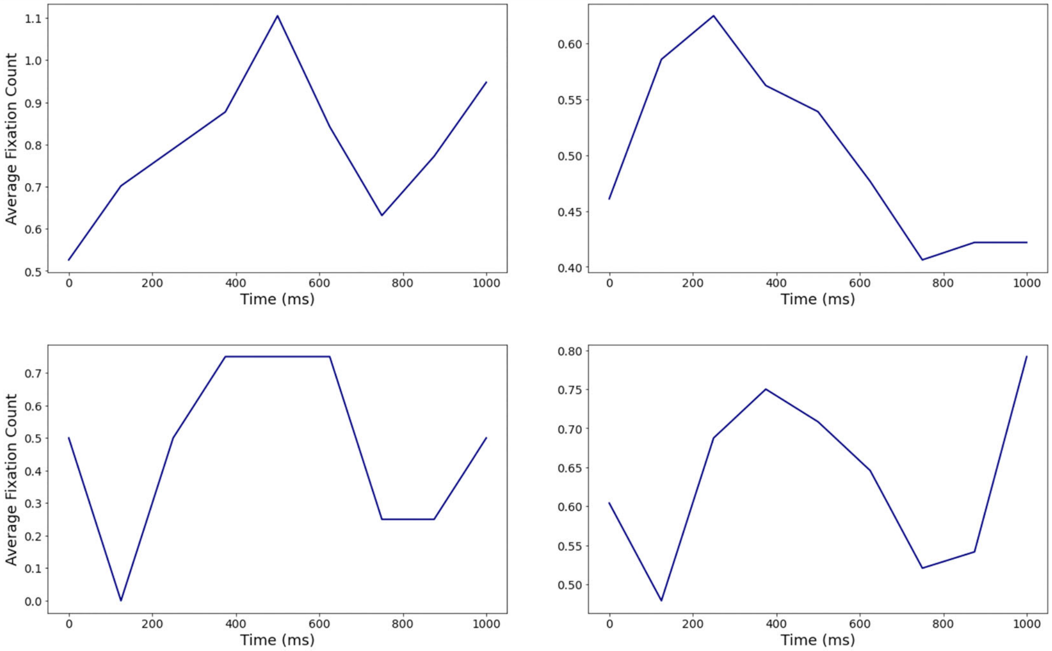


Fig. 14. Average fixation counts over the responded events of detection response task (DRT) sessions for different participants.

cognitive states. This is in part the case because human cognitive states are modulated by a variety of factors that the single modality might not be able to pick up. It is thus important to further investigate additional human gaze parameters (e.g., pupil dilation) to get better and more comprehensive estimates of cognitive states.

Fig. 15 depicts the variations in pupil diameter within the first 3 minutes of the experiment which were taken from two sessions of three different participants. The starting points represent the onset of the experiments. The results indicate that the pupil diameter gradually decreases after the experiment is initiated. The underlying reason is that the participants focus on the driving task at the beginning of the experiment and after that, they lose their attention progressively which is a demonstration of mind wandering state (consistent with the findings of Grandchamp et al., 2014) and as discussed in Section 3.8).

4.4. Learning behavioral states from EEG, fNIRS, and pupil diameter

In this section, our goal is to train learning models that are capable of classifying the behavioral states. As previously discussed, assessing behavioral states acts as an intermediate step to infer the cognitive states. Specifically, we randomly selected 10 participants from a total of 89 participants and use the fNIRS, EEG-PSD, and pupil diameter signals collected from these 10 participants to form the training and test data set. We follow the leave-one-subject-out protocol (Dou, Coelho de Castro, Kamnitsas, & Glocker, 2019) to conduct our experiment, that

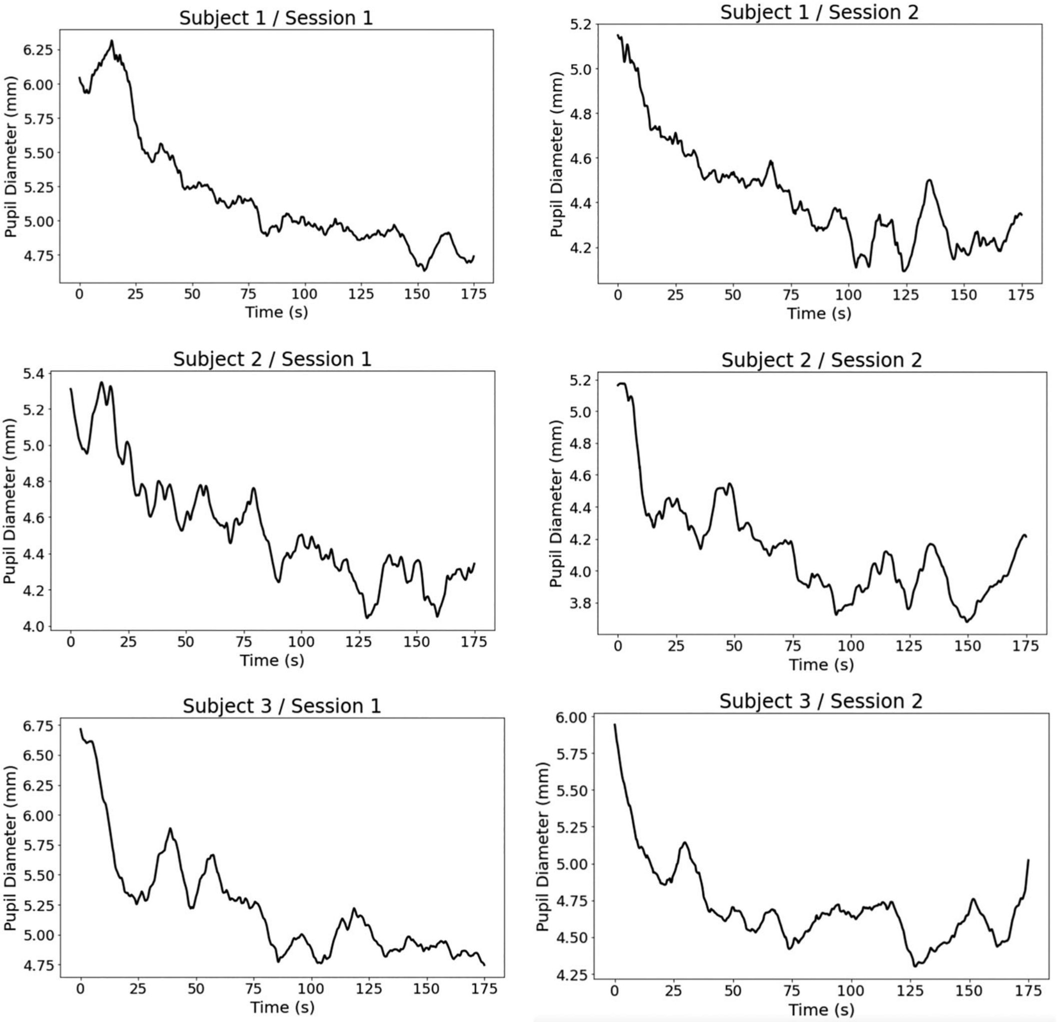


Fig. 15. Pupil diameter variation within the first 3 minutes of the experiments taken from three different participants.

is, using the data collected from nine participants for training and the data collected from the rest (one participant) for testing. The training data are split into a training set and a validation set with a proportion of 80% and 20%. As previously discussed in Section 3.5, the label is alternatively selected from one of the six extreme behavioral marker pairs: Slow DRT - Fast DRT, Slow Communication - Fast Communication, Slow Braking - Fast Braking, Slow Steering - Fast Steering, Low Position Offset - High Position Offset, Low Heading Offset - High Heading Offset, where labels "–1" and "1" denote the Slow/Low and Fast/High behavioral marker, respectively. Instead of jointly predicting these labels, we consider a simpler problem of predicting them separately. The joint prediction problem will be preserved as our future work.

Table 5
Numerical results for baseline models

	DRT		Non-DRT	
Data type	InceptionTime	MLSTM-FCN	InceptionTime	MLSTM-FCN
ΔHb	49.2 ± 1.9	52.0 ± 0.9	51.9 ± 1.0	51.1 ± 1.5
ΔHbO	51.5 ± 2.0	50.6 ± 1.7	**56.1** ± 0.8	55.3 ± 3.0
ΔHbT	52.9 ± 2.3	53.1 ± 1.1	53.3 ± 1.1	52.6 ± 1.1
Data type	InceptionTime	MLSTM-FCN	InceptionTime	MLSTM-FCN
Pupil diameter	53.7 ± 1.2	52.6 ± 1.9	**57.2** ± 1.0	54.1 ± 2.4
Data type	PSD-NET		PSD-NET	
EEG-PSD	59.1 ± 0.2		**60.1** ± 0.4	

Abbreviations: DRT, detection response task; EEG, electroencephalogram; ΔHb, changes in the concentration of deoxyhemoglobin; ΔHbO, changes in the concentration of oxy-hemoglobin; ΔHbT, changes in the concentration of total-hemoglobin; MLSTM-FCN, multivariate long short-term memory fully convolutional network; PSD-NET, power spectral network; PSD, power spectral density.

Hyperparameters tuning is performed for all DG methods, we apply a grid search over a range of [0.001,1] for all the hyper-parameters with a $\log_{10}$ scale and choose parameters and the corresponding model that produce the lowest validation loss. The hyperparameter tuning procedure is repeated for every test participant, following Gulrajani and Lopez-Paz, 2020). All models working on the EEG-PSD data are trained for 200 epochs using the Adam optimizer (Kingma & Ba, 2014) with the learning rate 5×10^{-4}. The batch size is set to be 32 for ERM and 144 (nine participants, 16 samples from each participant) for MMD-AAE, MLDG, and CORAL. For the models applied to fNIRS data, we set the batch size as 144 with the learning rate as 10^{-4} for Adam optimizer (Kingma & Ba, 2014). For the models applied to pupil diameter data, we set the batch size as 144 with the learning rate as 5×10^{-5} for Adam optimizer (Kingma & Ba, 2014). The whole experiment is repeated three times and the average accuracy and standard deviation values are reported. The above procedures are applied for all six extreme behavioral marker pairs separately. The details of these extreme behavioral maker pairs are described in Section 3.5. For the final results, we only report the accuracy where Fast/Slow steering is used as the label due to its superior performance in comparison with other behavioral markers.

The results for baseline models, that is, the models without employing DG techniques for EEG-PSD, fNIRS, and pupil diameter data, can be viewed in Table 5. As seen, the performance of two chosen algorithms for fNIRS data is close to a random predictor. Particularly, the highest accuracy for fNIRS data is 56.1% achieved by InceptionTime for Non-DRT sessions using ΔHbO data. On the other hand, the best accuracy of the baseline algorithm for EEG-PSD and pupil diameter is 60.1% and 57.2% both achieved in Non-DRT sessions, respectively.

Next, the learning performances of three DG methods are shown in Tables 6 and 7. The accuracy provided by DG methods achieves comparable performances as the baseline for

Table 6
Numerical results for three domain generalization (DG) algorithms with electroencephalogram (EEG) power spectral density (PSD) data

DG algorithms	DRT	Non-DRT
ERM	59.1 ± 0.2	60.1 ± 0.4
MLDG	59.4 ± 0.2	60.6 ± 0.4
MMD-AAE	58.5 ± 0.3	**61.8** ± 0.8
CORAL	59.1 ± 0.4	60.9 ± 0.8

Abbreviations: CORAL, correlation alignment; DRT, detection response task; ERM, empirical risk minimization; MLDG, meta-learning domain generalization; MMD-AAE, maximum mean discrepancy-adversarial autoencoder.

Table 7
Numerical results for three domain generalization (DG) algorithms with pupil diameter data

	DRT		Non-DRT	
DG algorithms	InceptionTime	MLSTM-FCN	InceptionTime	MLSTM-FCN
ERM	53.7 ± 1.2	52.6 ± 1.9	57.2 ± 1.0	54.1 ± 2.4
MLDG	54.5 ± 1.4	53.9 ± 2.8	59.3 ± 1.7	55.5 ± 2.9
MMD-AAE	53.9 ± 1.5	54.2 ± 2.3	57.2 ± 2.0	56.0 ± 0.9
CORAL	56.2 ± 1.1	54.0 ± 3.1	**61.5** ± 2.8	56.5 ± 1.2

Abbreviations: CORAL, correlation alignment; DRT, detection response task; ERM, empirical risk minimization; MLDG, meta-learning domain generalization; MMD-AAE, maximum mean discrepancy-adversarial autoencoder.

DRT sessions and generally surpasses the baseline algorithm at least 0.5% up to 4.3% for Non-DRT sessions. The highest improvements belong to EEG-PSD and pupil diameter data. More specifically, the highest accuracy for EEG-PSD data using a baseline algorithm is 60.1% for Non-DRT sessions, while the best accuracy of DG method is 61.8%. On the other hand, the highest accuracy for pupil diameter data using a baseline algorithm is 57.2% for Non-DRT sessions, while the best accuracy of DG method is 61.5%. As seen, applying the DG algorithms consistently improves the accuracy of learning models compared to the traditional ERM method.

Even though we employed state-of-the-art DG methods, the classification performance of behavioral states is modest. This is not unexpected given that assessing these behavioral states across multiple instances within subjects and across subjects is a hard problem. It points to potential limitations of current DG methods which are mainly designed for computer vision data sets and may not be feasible to apply directly to physiological data sets without appropriate modifications. Yet, given that new DG methods are being proposed all the time, it will be important to evaluate their performance on the data set and also consider the inclusion of additional signals and context-based information in an effort to improve classifier performance (if it is possible).

5. Discussion and future work

The overarching question for our experimental framework that motivated all of our machine learning efforts was whether it is possible to achieve a sufficiently high classification accuracy of systemic cognitive states across subjects using state-of-the-art machine learning models. This is important not only for monitoring and potentially aiding individual humans, but also for improving performance in mixed-initiative teams where humans and autonomous artificial agents work together in the pursuit of common goals. Current autonomous systems, however, are unaware of human cognitive states, they have no notion of team capabilities, tasks, and goals, and they lack the ability to interact with humans and adapt their behaviors on team dynamics. Genuine artificial teammates, instead, will need to have the ability to assess human physiological and cognitive states, to understand human goals and intentions as these dynamically shift based on varying task demands, and to anticipate errors and changes in plans so they can proactively intervene in order to preserve team coherence and performance.

Hence, an important step along the way toward genuine artificial teammates would be the demonstration of a successful method that is able to classify systemic cognitive states across individuals (and ideally also across tasks) or to provide a conceptual "impossibility argument" for why such inferences from physiological and neurophysiological data are not possible (e.g., appealing to noise in the data, large variations of the signals within an individual, distributional signal drifts during task performance, etc.). To the best of our knowledge, neither position has been convincingly made in the literature so far.

The classification results of behavioral states we obtained using standard machine learning methods (which can then be linked to cognitive states based on the experimental setting) showed that current DG methods yield only modest classification accuracy (and neither do ERPs in EEGs, although eye gaze showed promising results for some). In a way, the failure to obtain accurate classifiers is not unexpected because the classification problem across multisubject multimodal (noisy) time series data is known to be hard and it is quite possible that there is just not enough common information in those signals to generate consistent cognitive state abstractions and that at the very least additional constraining context information will ultimately be needed for machine learning model to be able to cope with individual variation, context-based shifts, and signal drifts. This last point is implicitly supported by the lack of published methods that demonstrate sufficiently accurate systemic cognitive state classification based on multimodal physiological and neurophysiological signals across multiple subjects, despite significant efforts by the community and partial successes for specific cognitive states within individuals.

It is, however, important to mention that while a general method that works across subjects for all of the considered systemic cognitive states might be impossible, more specific methods targeting individual cognitive states could still yield high classification results. A case in point is our recent success at achieving high classification performance for one of the five systemic cognitive states: *cognitive workload*. Using the different experimental/behavioral conditions to define three different levels of cognitive workload: only driving (level 0), driving and communicating (level 1), and driving with braking events and communication (level 2). We performed statistical analyses of various physiological signals, including eye gaze, EEG, and

ABP, and utilized several machine learning methodologies, including *k-Nearest Neighbor*, *Naive Bayes*, *Random Forest*, *Support-Vector Machines*, and *Neural Network-based models* to infer the three workload levels. The results revealed that direct cognitive workload classification on eye gaze information alone (without predicting behavioral states), specifically *percentage change in pupil size*, was able to achieve an accuracy of 80.45 ∓ 3.15 using *Support-Vector Machines* while combining eye gaze with EEG was able to reach an accuracy of 77.08 ∓ 3.22 using a *Neural Network-based model* (see Aygun, Nguyen, Haga, Aeron, & Scheutz, 2022 for details).

This is but one example of how the experimental framework introduced in this paper and the resultant data set (with the accompanying data pre-processing methods) can form the basis for additional analytical and modeling work that we expect to provide further results for which sensory modalities are essential for cognitive state inference (and should thus be collected in other experiments in the future) to demonstrate generalization not only across subjects, as we have pursued here, but more generally also across different tasks.

6. Conclusion

It is currently still an open research question to what extent statistical machine learning methods are able to classify systemic cognitive states based on multimodal physiological and neurophysiological signals across subjects and ideally across tasks. Yet, being able to detect and track such cognitive states would not only allow for the development of adaptive technologies that would benefit individuals by taking their cognitive state into account, but also human–machine teams where the effectiveness of interactions and cooperation critically depend on individual human performance modulated by systemic human cognitive states.

In this paper, we introduced a multimodal experimental paradigm that was specifically designed to collect a comprehensive suite of empirical data from human performance paired with task and event-based context as the basis for investigating methods and machine learning models for cognitive state inference that can help answer this question. We applied state-of-the-art DG methods to our data and obtained modest classification results for behavioral states which are linked to systemic cognitive states. The failure to obtain more accurate classifiers points to the need for novel classification methods that likely will need to incorporate additional task-based context in order to improve classification performance. In that sense, our results can serve as a baseline for evaluating future machine learning models on our data set, but they could also be taken as a hint that general methods for inferring systemic cognitive states across subjects (and across tasks) might just not be attainable, despite our best efforts.

Acknowledgments

The authors would like to acknowledge funding provided by AFOSR grant # FA9550-18-1-0465 and also thank Chris Myers as well as the anonymous reviewers for their feedback that led to a much improved paper.

Notes

1 Since the participants are mostly recruited from the college with the average age of 20 years old and, the results in our experiment may not be suitable to apply on other subjects with higher age and different education.

2 Even though these physiological signals have been collected in our experiment and are available for analysis, some, for example, ABP and skin conductance, are not further investigated in this paper. However, we still included them in the description to provide a complete list of the available signals in our data set.

3 Here, the ABP signal is beat-to-beat ABP, not continuous ABP; however, for convenience, we just use ABP to denote beat-to-beat ABP in the rest of this manuscript.

4 There is some evidence that combining fNIRS and EEG signals could provide better results than processing them separately (Aghajani & Omurtag, 2016; Omurtag, Aghajani, & Keles, 2017; Putze et al., 2014), which we will leave this approach for our future work.

References

Agarap, A. F. (2018). Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*.

Aghajani, H., & Omurtag, A. (2016). Assessment of mental workload by EEG+ fNIRS. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3773–3776). IEEE.

Al-Nafjan, A., Hosny, M., Al-Wabil, A., & Al-Ohali, Y. (2017). Classification of human emotions from electroencephalogram (EEG) signal using deep neural network. *International Journal of Advanced Computer Science and Applications*, *8*(9), 419–425.

Appelbaum, L., Boehler, C. N., Davis, L., Won, R. J., & Woldorff, M. (2014). The dynamics of proactive and reactive cognitive control processes in the human brain. *Journal of Cognitive Neuroscience*, *26*, 1021–1038.

Arsalan, A., Majid, M., Butt, A. R., & Anwar, S. M. (2019). Classification of perceived mental stress using a commercially available EEG headband. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2257–2264.

Aygun, A., Nguyen, T., Haga, Z., Aeron, S., & Scheutz, M. (2022). Investigating methods for cognitive workload estimation for assistive robots. *Sensors*, *22*(*18*), 337–348.

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and quantifying mind wandering during simulated driving. *Frontiers in Human Neuroscience*, *11*, pp 406.

Beckers, N., Schreiner, S., Bertrand, P., Mehler, B., & Reimer, B. (2017). Comparing the demands of destination entry using Google glass and the Samsung Galaxy S4 during simulated driving. *Applied Ergonomics*, *58*, 25–34.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*(5), B231–B244.

Bixler, R., Blanchard, N., Garrison, L., & D'Mello, S. (2015). Automatic detection of mind wandering during reading using gaze and physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 299–306).

Bjørneseth, F. B., Renganayagalu, S. K., Dunlop, M. D., Hornecker, E., & Komandur, S. (2012). Towards an experimental design framework for evaluation of dynamic workload and situational awareness in safety critical maritime settings. In *The 26th BCS Conference on Human Computer Interaction 26* (pp. 309–314).

Blanchard, G., Lee, G., & Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, *24*, 2178–2186.

Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. (2014a). Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems* (pp. 55–60). Cambridge: Springer.

Blanchard, N., Bixler, R., Joyce, T., & D'Mello, S. (2014b). Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems* (pp. 55–60). Springer.

Blaney, G., Sassaroli, A., Pham, T., Krishnamurthy, N., & Fantini, S. (2019). Multi-distance frequency-domain optical measurements of coherent cerebral hemodynamics. *Photonics*, *6*(3) (pp. 83).

Brouwer, A.-M., Snelting, A., Jaswa, M., Flascher, O., Krol, L., & Zander, T. (2017). Physiological effects of adaptive cruise control behaviour in real driving. In *Proceedings of the 2017 ACM Workshop on An Application-oriented Approach to BCI Out of the Laboratory* (pp. 15–19).

Buettner, R., Baumgartl, H., & Sauter, D. (2019). Microsaccades as a predictor of a user's level of concentration. In *Information systems and neuroscience* (pp. 173–177). Springer International Publishing.

Canabarro, S. L. S., Garcia, A., Satler, C., & Tavares, M. C. H. (2017). Interaction between neural and cardiac systems during the execution of the Stroop task by young adults: Electroencephalographic activity and heart rate variability. *Neuroscience*, *4*, 28–51.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports*, *7*(1), 5222.

Cecotti, H., & Ries, A. J. (2017). Best practice for single-trial detection of event-related potentials: Application to brain–computer interfaces. *International Journal of Psychophysiology*, *111*, 156–169.

Cerliani, M. (2021). Tsmoothie. https://github.com/cerlymarco/tsmoothie.

Chatham, C. H., Frank, M. J., & Munakata, Y. (2009). Pupillometric and behavioral markers of a developmental shift in the temporal dynamics of cognitive control. *Proceedings of the National Academy of Sciences*, *106*(14), 5529–5533.

Cheng, S.-Y. (2017). Evaluation of effect on cognition response to time pressure by using EEG. In *Advances in human factors and ergonomics in healthcare and medical devices: Proceedings of the AHFE 2017 International Conferences on Human Factors and Ergonomics in Healthcare and Medical Devices, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA, 8*, (pp. 45–52). Springer International Publishing.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Coffey, E. B., Brouwer, A.-M., & van Erp, J. B. (2012). Measuring workload using a combination of electroencephalography and near infrared spectroscopy. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1822–1826). Los Angeles, CA: Sage Publications.

Collet, C., Petit, C., Priez, A., & Dittmar, A. (2005). Stroop color–word test, arousal, electrodermal activity and performance in a critical driving situation. *Biological Psychology*, *69*(2), 195–203.

Cooper, J. M., Medeiros-Ward, N., & Strayer, D. L. (2013). The impact of eye movements and cognitive workload on lateral position variability in driving. *Human Factors*, *55*(5), 1001–1014.

Cooper, P. S., Wong, A. S. W., Fulham, W. R., Thienel, R., Mansfield, E., Michie, P. T., & Karayanidis, F. (2015). Theta frontoparietal connectivity associated with proactive and reactive cognitive control processes. *NeuroImage*, *108*, 354–363.

Cui, Y., Xu, Y., & Wu, D. (2019). EEG-based driver drowsiness estimation using feature weighted episodic training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(11), 2263–2273.

Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M., & Abbass, H. A. (2019). Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics*, *51*(3), 1542–1555.

Di Stasi, L. L., McCamy, M. B., Catena, A., Macknik, S. L., Canas, J. J., & Martinez-Conde, S. (2013). Microsaccade and drift dynamics reflect mental fatigue. *European Journal of Neuroscience*, *38*(3), 2389–2398.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, *32*, 6450–6461.

Duan, T., Shaikh, M. A., Chauhan, M., Chu, J., Srihari, R. K., Pathak, A., & Srihari, S. N. (2020). Meta learn on constrained transfer learning for low resource cross subject EEG classification. *IEEE Access*, *8*, 224791–224802.

Durantin, G., Dehais, F., & Delorme, A. (2015). Characterization of mind wandering using fNIRS. *Frontiers in Systems Neuroscience*, *9*, 45.

Dutta, A., Jacob, A., Chowdhury, S. R., Das, A., & Nitsche, M. A. (2015). EEG-NIRS based assessment of neurovascular coupling during anodal transcranial direct current stimulation - A stroke case series. *Journal of Medical Systems*, *39*(4), 36.

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 97–120.

Fantini, S., Aggarwal, P., Chen, K., Franceschini, M. A., & Ehrenberg, B. L. (2003). Near-infrared spectroscopy and polysomnography during all-night sleep in human subjects. In *Proceedings of SPIE*, volume 5068, pp. 155–162.

Fawaz, H. I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962.

Garg, Nikhil, et al. (2022). Decoding the neural signatures of valence and arousal from portable EEG headset. *Frontiers in Human Neuroscience*, *16*, p. 808.

Ghani, U., Signal, N., Niazi, I. K., & Taylor, D. (2020). ERP based measures of cognitive workload: A review. *Neuroscience & Biobehavioral Reviews*, *118*, 18–26.

Golob, E., Ringman, J., Irimajiri, R., Bright, S., Schaffer, B., Medina, L., & Starr, A. (2009). Cortical event-related potentials in preclinical familial Alzheimer disease. *Neurology*, *73*(20), 1649–1655.

González-Villar, A. J., Samartin-Veiga, N., Arias, M., & Carrillo-de-la Pe na, M. T. (2017). Increased neural noise and impaired brain synchronization in fibromyalgia patients during cognitive interference. *Scientific Reports*, *7*(1), 1–8.

Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Oculometric variations during mind wandering. *Frontiers in Psychology*, *5*, pp. 31.

Gulrajani, I., & Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Hallacoglu, B., Sassaroli, A., Fantini, S., Wysocki, M., Guerrero-Berroa, E., Beeri, M. S., Haroutunian, V., Shaul, M., Rosenberg, I. H., & Troen, A. (2012). Absolute measurement of cerebral optical coefficients, hemoglobin concentration and oxygen saturation in old and young adults with near-infrared spectroscopy. *Journal of Biomedical Optics*, *17*(8), 081406.

Hamzah, N., Norhazman, H., Zaini, N., & Sani, M. (2016). Classification of EEG signals based on different motor movement using multi-layer perceptron artificial neural network. *Journal of Biological Sciences*, *16*(7), 265–271.

Han, D.-K., & Jeong, J.-H. (2021). Domain generalization for session-independent brain–computer interface. In *2021 9th International Winter Conference on Brain–Computer Interface (BCI)* (pp. 1–5). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. corr abs/1512.03385 (2015).

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Holtzer, R., Schoen, C., Demetriou, E., Mahoney, J. R., Izzetoglu, M., Wang, C., & Verghese, J. (2017). Stress and gender effects on prefrontal cortex oxygenation levels assessed during single and dual-task walking conditions. *European Journal of Neuroscience*, *45*(5), 660–670.

Hossain, M. F., Yaacob, H., & Nordin, A. (2021). Development of unified neuro-affective classification tool (UNACT). In *IOP Conference Series: Materials Science and Engineering*, volume 1077 of *1* (pp. 012031). IOP Publishing.

Hu, B., Rao, J., Li, X., Cao, T., Li, J., Majoe, D., & Gutknecht, J. (2017). Emotion regulating attentional control abnormalities in major depressive disorder: An event-related potential study. *Scientific Reports*, *7*(1), 1–21.

Huang, J., Liu, Y., & Peng, X. (2022). Recognition of driver's mental workload based on physiological signals, a comparative study. *Biomedical Signal Processing and Control*, *71*, 103094.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430.

ISO for Standardization Road Vehicles—Transport Information, I. O., & Systems, C. (2016). Detection-response task (DRT) for assessing attentional effects of cognitive load in driving.

Joyce, C. A., Gorodnitsky, I. F., & Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, *41*(2), 313–325.

Kainerstorfer, J. M., Sassaroli, A., Tgavalekos, K. T., & Fantini, S. (2015). Cerebral autoregulation in the microvasculature measured with near-infrared spectroscopy. *Journal of Cerebral Blood Flow & Metabolism*, *35*(6), 959–966.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kane, M. J., Brown, L. H., McVay, J. C., PJ, P. J. S., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science*, *18*, 614–621.

Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate LSTM-FCNS for time series classification. *Neural Networks*, *116*, 237–245.

Katmah, R., Al-Shargie, F., Tariq, U., Babiloni, F., Al-Mughairbi, F., & Al-Nashash, H. (2021). A review on mental stress assessment methods using EEG signals. *Sensors*, *21*(15), 5043.

Keller, J., Ruthruff, E., & Keller, P. (2017). Mindfulness and divergent thinking: The value of heart rate variability as an objective manipulation check. *Universal Journal of Psychology*, *5*(3), 95–104.

Khalaf, A., Nabian, M., Fan, M., Yin, Y., Wormwood, J., Siegel, E., Quigley, K. S., Barrett, L. F., Akcakaya, M., Chou, C.-A., & Ostadabbas, S. (2020). Analysis of multimodal physiological signals within and between individuals to predict psychological challenge vs. threat. *Expert Systems with Applications*, *140*, 112890.

Khedher, A. B., Jraidi, I., & Frasson, C. (2019). Predicting learners' performance using EEG and eye tracking features. In *The Thirty-Second International Flairs Conference*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuremoto, T., Baba, Y., Obayashi, M., Mabu, S., & Kobayashi, K. (2015). To extraction the feature of EEG signals for mental task recognition. In *2015 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* (pp. 264–269).

Kurniawan, H., Maslov, A. V., & Pechenizkiy, M. (2013). Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (pp. 209–214).

León-Carrion, J., Damas-López, J., Martín-Rodríguez, J. F., Domínguez-Roldán, J. M., Murillo-Cabezas, F., Barroso y Martin, J. M., & Domínguez-Morales, M. R. (2008). The hemodynamics of cognitive control: The level of concentration of oxygenated hemoglobin in the superior prefrontal cortex varies as a function of performance in a modified Stroop task. *Behavioural Brain Research*, *193*(2), 248–256.

Levelt, W. J., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, *24*(2), 133–164.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731.

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2018a). Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, H., Pan, S. J., Wang, S., & Kot, A. C. (2018b). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5400–5409).

Liang, Y., & Lee, J. D. (2008). Driver cognitive distraction detection using eye movements. In *Passive eye monitoring* (pp. 285–300). Berlin, Heidelberg: Springer International Publishing.

Lin, Y.-P., Wang, C.-H., Wu, T.-L., Jeng, S.-K., & Chen, J.-H. (2007). Multilayer perceptron for EEG signal classification during listening to emotional music. In *TENCON 2007-2007 IEEE Region 10 Conference* (pp. 1–3). IEEE.

Liu, C.-W., Hsieh, A.-Y., Lo, S.-K., & Hwang, Y. (2017a). What consumers see when time is running out: Consumers' browsing behaviors on online shopping websites when under time pressure. *Computers in Human Behavior*, *70*, 391–397.

Liu, W., Lu, Y., Huang, D., & Fu, S. (2017b). An analysis of pilot's workload evaluation based on time pressure and effort. In *Engineering psychology and cognitive ergonomics: Performance, emotion and situation awareness* (pp. 32–41). Springer International Publishing.

Louis, E. K. S., Frey, L. C., Britton, J. W., Hopp, J. L., Korb, P., Koubeissi, M. Z., Lievens, W. E., & Pestana-Knight, E. M. (2016). The normal EEG. *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants [Internet].*

Lyu, B., Pham, T., Blaney, G., Haga, Z., Sassaroli, A., Fantini, S., & Aeron, S. (2021). Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS. *Journal of Biomedical Optics*, *26*(2), 022908.

Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology*, *8*, 211.

McIntire, L. K., McKinley, R. A., Goodyear, C., & Nelson, J. (2014). A comparison of the effects of transcranial direct current stimulation and caffeine on vigilance and cognitive performance during extended wakefulness. *Brain Stimulation*, *7*, 499–507.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*, 6–12.

Merkies, K., Ready, C., Farkas, L., & Hodder, A. (2019). Eye blink rates and eyelid twitches as a non-invasive measure of stress in the domestic horse. *Animals*, *9*(8), 562.

Mühlbacher-Karrer, S., Mosa, A. H., Faller, L.-M., Ali, M., Hamid, R., Zangl, H., & Kyamakya, K. (2017). A driver state detection system–Combining a capacitive hand detection sensor with physiological sensors. *IEEE Transactions on Instrumentation and Measurement*, *66*(4), 624–636.

Nigbur, R., Ivanova, G., & Stürmer, B. (2011). Theta power as a marker for cognitive interference. *Clinical Neurophysiology*, *122*(11), 2185–2194.

Omurtag, A., Aghajani, H., & Keles, H. O. (2017). Decoding human mental states by whole-head EEG+ fNIRS during category fluency task performance. *Journal of Neural Engineering*, *14*(6), 066003.

Ordonez, L., & Benson, L. (1997). Decisions under time pressure: How time constraint affects risky decision making. *Organizational Behavior and Human Decision Processes*, *71*(2), pp. 121–140.

Ozawa, S., & Hiraki, K. (2017). Distraction decreases prefrontal oxygenation: A NIRS study. *Brain and Cognition*, *113*, 155–163.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141–144).

Patel, S. H., & Azzam, P. N. (2005). Characterization of n200 and p300: Selected studies of the event-related potential. *International Journal of Medical Sciences*, *2*(4), 147.

Putze, F., Hesslinger, S., Tse, C.-Y., Huang, Y., Herff, C., Guan, C., & Schultz, T. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Frontiers in Neuroscience*, *8*, 373.

Qin, X., Zheng, Y., & Chen, B. (2019). Extract EEG features by combining power spectral density and correntropy spectral density. In *2019 Chinese Automation Congress (CAC)* (pp. 2455–2459). IEEE.

Rajendra, V., & Dehzangi, O. (2017). Detection of distraction under naturalistic driving using galvanic skin responses. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (pp. 157–160).

Raza, H., Rathee, D., Zhou, S.-M., Cecotti, H., & Prasad, G. (2019). Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain–computer interface. *Neurocomputing*, *343*, 154–166.

Robertson, J. A., Thomas, A. W., Prato, F. S., Johansson, M., & Nittby, H. (2014). Simultaneous fMRI and EEG during the multi-source interference task. *PLoS One*, *9*(12), e114599.

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2020). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), pp. 1–49.

Saeedpour-Parizi, M. R., Hassan, S. E., & Shea, J. B. (2020). Pupil diameter as a biomarker of effort in goal-directed gait. *Experimental Brain Research*, *238*(11), 2615–2623.

Sakai, T., Tamaki, H., Ota, Y., Egusa, R., Inagaki, S., Kusunoki, F., Sugimoto, M., Mizoguchi, H. (2017). EDA-based estimation of visual attention by observation of eye blink frequency. *International Journal on Smart Sensing and Intelligent Systems*, *10*(2), 296–307.

Šalkevicius, J., Damaševičius, R., Maskeliunas, R., & Laukienė, I. (2019). Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, *8*(9), 1039.

Scheutz, M., DeLoach, S., & Adams, J. (2017). A framework for developing and using shared mental models in human–agent teams. *Journal of Cognitive Engineering and Decision Making*, *11*(3), 203–224.

Schmutz, P., Roth, S. P., Seckler, M., & Opwis, K. (2010). Designing product listing pages—Effects on sales and users' cognitive workload. *International Journal of Human–Computer Studies*, *68*(7), 423–431.

Shojaeizadeh, M., Djamasbi, S., Paffenroth, R. C., & Trapp, A. C. (2019). Detecting task demand via an eye tracking machine learning system. *Decision Support Systems*, *116*, 91–101.

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E. E., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, pp 260–277.

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., Carlson, J. M., & Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS One*, *6*(3), e18298.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stoica, P., &, Moses, R. L. (2005). *Spectral analysis of signals*. Upper Saddle River, NJ: Pearson Prentice Hall.

Stojmenova, K., & Sodnik, J. (2018). Detection-response task—Uses and limitations. *Sensors*, *2*, 594.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors*, *57*(8), 1300–1324.

Stuiver, A., & Mulder, B. (2014). Cardiovascular state changes in simulated work environments. *Frontiers in Neuroscience*, *8*, pp. 399.

Sun, B., Feng, J., & Saenko, K. (2017). Correlation alignment for unsupervised domain adaptation. In *Domain adaptation in Computer Vision Applications* (pp. 153–171). Springer International Publishing.

Sur, S., & Sinha, V. K. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal*, *18*(1), 70.

Thornton, A. R. D., Harmer, M., & Lavoie, B. A. (2007). Selective attention increases the temporal precision of the auditory n100 event-related potential. *Hearing Research*, *230*(1–2), 73–79.

Tong, Y., Rooney, E. J., Bergethon, P. R., Martin, J. M., Sassaroli, A., Ehrenberg, B. L., Van Toi, V., Aggarwal, P., Ambady, N., & Fantini, S. (2005). Studying brain function with near-infrared spectroscopy concurrently with electroencephalography. In B. Chance, R. R. Alfano, B. J. Tromberg, M. Tamura, & E. M. Sevick-Muraca (Eds.), *Proceedings of the SPIE* (pp. 444).

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *Journal of Neuroscience*, *35*(20), 7866–7877.

Wang, J., Lan, C., Liu, C., Ouyang, Y., & Qin, T. (2021). Generalizing to unseen domains: A survey on domain generalization. In Z.-H. Zhou (Ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI-21* (pp. 4627–4635). International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014a). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, *62*, 1–10.

Wang, X.-W., Nie, D., & Lu, B.-L. (2014b). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, *129*, 94–106.

Wang, Y. K., Jung, T. P., & Lin, C. T. (2015). EEG-based attention tracking during distracted driving. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*, 1085–1094.

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1578–1585). IEEE.

Wildemeersch, D., Peeters, N., Saldien, V., Vercauteren, M., & Hans, G. (2018). Pain assessment by pupil dilation reflex in response to noxious stimulation in anaesthetized adults. *Acta Anaesthesiologica Scandinavica*, *62*(8), 1050–1056.

Wu, D., Xu, Y., & Lu, B.-L. (2020). Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(1), 4–19.

Zahabi, M., Razak, A. M. A., Shortz, A. E., Mehta, R. K., & Manser, M. (2020). Evaluating advanced driver-assistance system trainings using driver performance, attention allocation, and neural efficiency measures. *Applied Ergonomics*, *84*, 103036.

Zhao, L.-M., Yan, X., & Lu, B.-L. (2021). Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Zheng, W.-L., & Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175.

Zheng, W.-L., Zhu, J.-Y., Peng, Y., & Lu, B.-L. (2014). EEG-based emotion classification using deep belief networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6).