Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures

Dengbo He
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: dengbo.he@mail.utoronto.ca

Ziquan Wang
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: ziquan.wang@mail.utoronto.ca

Elias B. Khalil
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: khalil@mie.utoronto.ca

Birsen Donmez, Corresponding Author
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: donmez@mie.utoronto.ca

Guangkai Qiao
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: qiaogk@sjtu.edu.cn

Shekhar Kumar
Department of Mechanical and Industrial Engineering
University of Toronto, Toronto, Ontario, Canada M5S 3G8
Email: shekhar.kumar@mail.utoronto.ca

Submitted to TRB: August 1, 2021
Revised based on TRB reviews, and submitted to TRR: October 26, 2021
Final files submitted to TRR: March 6, 2022

Funding Statement: The funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Hitachi Solutions, Ltd.

Data accessibility: Data sharing is not applicable to this article as no new data were created in this study.

ABSTRACT

In-vehicle infotainment systems can increase cognitive load and impair driving performance. These effects can be alleviated through interfaces that can assess cognitive load and adapt accordingly. Eye-tracking and physiological measures that are sensitive to cognitive load, such as pupil diameter, gaze dispersion, heart rate (HR), and galvanic skin response (GSR), can enable cognitive load estimation. The advancement in cost-effective and non-intrusive sensors in wearable devices provides an opportunity to enhance driver state detection by fusing eye-tracking and physiological measures. As a preliminary investigation of the added benefits of utilizing physiological data along with eye-tracking data in driver cognitive load detection, this paper explores the performance of several machine learning models in classifying three levels of cognitive load imposed on 33 drivers in a driving simulator study: no external load, lower difficulty 1-back task, and higher difficulty 2-back task. We built five machine learning models, including k-nearest neighbor, support vector machine, feedforward neural network, recurrent neural network, and random forest (RF) on (1) eye-tracking data only, (2) HR and GSR, (3) eye-tracking and HR, (4) eye-tracking and GSR, and (5) eye-tracking, HR, and GSR. Although physiological data provided 1% - 15% lower classification accuracies compared to eye-tracking data, adding physiological data to eye-tracking data increased model accuracies, with an RF classifier achieving 97.8% accuracy. GSR led to a larger boost in accuracy (29.3%) over HR (17.9%), with the combination of the two boosting accuracy by 34.5%. Overall, utilizing both physiological and eye-tracking measures shows promise for driver state detection applications.

Keywords: Cognitive Load Estimation, Machine Learning, Heart Rate, Galvanic Skin Response, Eye Measures

INTRODUCTION

Factors such as road environment (i.e., high traffic conditions), bad weather, and the usage of in-vehicle technologies (e.g., cellphone and infotainment systems) can increase the cognitive load experienced by drivers. Both simulator and on-road studies have shown that high cognitive load can impair driving performance and visual scanning behaviors (1, 2). Real-time assessment of cognitive load can enable vehicle manufacturers to provide preventative warnings and develop adaptive interfaces that can support drivers, for example, by actively limiting functionality on menu interfaces (3) and automatically filtering information when high levels of cognitive load is detected (4). Automated vehicle systems can also utilize cognitive load estimates to intelligently transfer vehicle control to the driver (5).

As summarized in Table 1, a variety of measures were found to be responsive to varying levels of external cognitive load experienced by drivers, including: 1) eye-tracking measures, such as pupil diameter (2, 6), blink rate (7), and standard deviation (SD) of horizontal gaze position (7, 8), 2) physiological measures, such as heart rate (HR) (8-10), galvanic skin response (GSR) (8, 9), and Electroencephalography (EEG) (11, 12); 3) driving performance measures, such as vehicle speed (8); 4) and subjective measures, such as NASA-Task Load Index (NASA-TLX) (1).

TABLE 1 Example Cognitive Load Measurements

Measure	Trend with Increased Cognitive Taskload
Eye-tracking	
Pupil diameter	↑ (2, 6)
Blink	Rate ↑ (7)
Gaze position	Periphery/mirror/instrument check rate ↓ (1)
	SD of horizontal position ↓ (7, 8)
	SD of vertical position ↓ (7)
Physiological	
HR	HR ↑ (8-10)
	HR variability ↓ (10)
GSR	↑ (8, 9)
EEG	Power of alpha band ↓ (11, 12)
	P300 latency ↑ (13)
Respiration	Rate ↑ (9)
Performance-based	
Vehicle speed	Average ↑ (9) ↓ (8)
	SD ↑ (9) ↓ (8)
Steering wheel	Reversal rate ↑ (8)
Subjective	
NASA-TLX	↑ (1)

It is widely acknowledged that no single measure alone can provide sufficient information to estimate cognitive load (9, 11). Indeed, multiple measures have been combined in previous research to estimate the cognitive load experienced by drivers. For example, Solovey et al. (14) reached 89% accuracy in classifying 2 levels of cognitive load (no-task vs. an auditory recall 2-back task) using driving performance, GSR, and HR data collected in an on-road study. Liang, Reyes (15) reached 81.1% accuracy in identifying 2 levels of cognitive load (no-task vs. an auditory stock ticker task) using driving performance and eye-tracking data collected in a simulator study. In general, driving performance measures used in these earlier studies (e.g., speed and lane position) are highly sensitive to traffic conditions and may require additional driving context-assessment to improve their utility in driver state detection (16). This need for additional information can be a barrier for the use of driving performance measures in driver state detection.

The fusion of eye-tracking and physiological measures seems to be more promising for real-time assessment of driver cognitive load, yet research is lacking in this area. Eye-tracking measures have been adopted in a number of production cars for detecting visual distraction (e.g., 17) and drowsiness (e.g., 18).

They have not yet been adopted for cognitive load detection, although pupil diameter (e.g., 2, 6), blink rate (e.g., 7), and gaze dispersion (e.g., 7, 8) are known to be sensitive to cognitive load variation. Physiological measures, such as HR (e.g., 8, 9, 10) and GSR (e.g., 8, 9), also react well to variations in cognitive demand and can now be collected through cost-effective and non-intrusive sensors, e.g., in wearable devices such as the Apple Watch (19) and FitBit (20). Thus, combining eye-tracking with HR and GSR data is now a feasible solution for in-car applications, yet, it is unknown what level of performance enhancement this combination may provide in driver cognitive load detection.

Using a dataset collected in a driving simulator study, this paper investigates the benefits of fusing eye-tracking and physiological data for driver cognitive load classification. It is hypothesized that increasing the number of features in the dataset by fusing different measure types would improve classification performance. In the simulator study, three levels of cognitive load (no external load, 1-back task, and 2-back task) were imposed by an audio-verbal cognitive task, the modified n-back task (21), while participants drove through an urban environment. A variety of machine learning methods used in earlier studies were explored on this three-class driver state estimation problem, including k-nearest neighbor (KNN, e.g., 14), support vector machine (SVM, e.g., 22)**,** feedforward neural network (FNN, e.g., 14), recurrent neural network (RNN, e.g., 23), and random forest (RF, e.g., 24). The models were built and compared using the following measures to investigate the benefits of fusing eye-tracking and different physiological data (in particular, HR and GSR):

- eye-tracking data only, including eye closure (i.e., fraction of the iris covered by the upper and lower eye lid), pupil diameter, and gaze rotation angle (i.e., the orientation of the eye gaze with respect to the world coordinate system);
- physiological data only, including HR and GSR;
- eye-tracking data and HR combined;
- eye-tracking data and GSR combined;
- eye-tracking data, HR, and GSR combined.

DATA SOURCE

The data utilized in this paper was collected from 33 participants in a driving simulator study originally reported in (21, 25), which investigated the effects of different levels of external cognitive demand on drivers' physiological, eye-tracking, and driving performance. In a within subject design, participants completed three counterbalanced conditions (in three drives total): no external task, and two difficulty levels of an external cognitive task (i.e., a secondary task). Eye-tracking measures, including the level of eye closure, pupil diameter, and gaze rotation angle, as well as physiological measures, including Electrocardiography (ECG) and GSR, were collected. Below we provide an overview of the experimental methods but a more detailed description of the methods can be found in (21). He et al. (26) also utilized physiological measures from this dataset for a preliminary machine learning application, but only with relatively simple machine learning models and without using eye-tracking data.

Participants

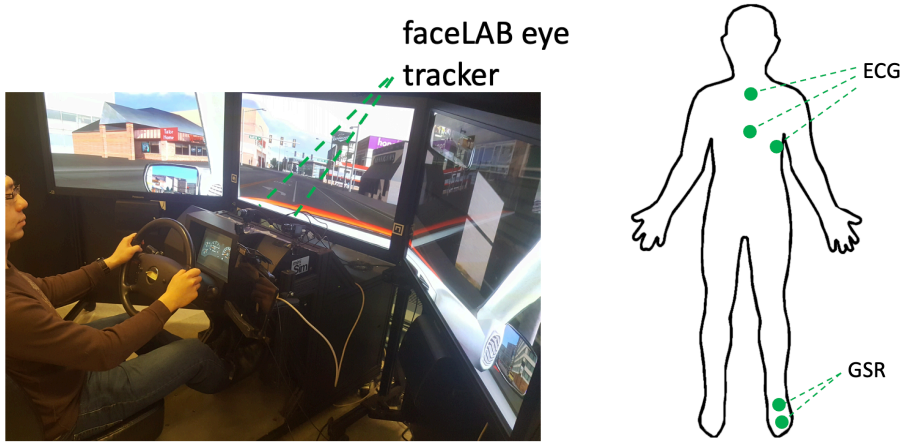
Thirty-three drivers (18 males and 15 females), recruited through campus and online posts, completed this driving simulator study. Participants were required to drive at least several times per month, to hold a full driver's license (G license in Ontario, Canada or equivalent) for at least 3 years, and to be under 35 years old (average age: 27.6; SD: 4.45). The compensation was C$12 per hour, and the participants were told that they could receive a bonus of up to C$14 based on their secondary task performance as an incentive for engaging in the secondary task.

Apparatus

The study was conducted on a NADS miniSim^{TM} driving simulator (Figure 1a). This fixed-based simulator has three 42-inch screens, creating a 130^{o} horizontal and 24^{o} vertical field at a 48-inch viewing distance. The center screen displays the left and center parts of the windshield; the right screen displays

the rest of the windshield, the rear-view mirror, and the right-side window and mirror; while the left screen displays the left-side window and mirror.

The eye-tracking information was collected at 60 Hz by using the faceLAB 5.0, a dashboard mounted eye-tracker by Seeing Machines. ECG was collected with three solid gel foam electrodes placed on participants' chest; and GSR was collected with one solid gel foam electrode beneath the bare left foot and the other under the heel (Figure 1b). Both ECG and GSR sensors were from Becker Meditec and the data was collected at 240 Hz using the D-Lab software developed by Ergoneers.



(a) **(b)**

Figure 1 Apparatus: (a) driving Simulator – NADS miniSim and faceLAB 5.0 eye-tracking system; (b) placement of ECG and GSR Sensors

Experimental Tasks

The secondary task used in this study was a modified version of an auditory-verbal n-back task widely used in driving research (e.g., 8, 9), and was validated to impose graded levels of cognitive load on drivers (21, 25). The modification was performed to minimize physiological signal interference due to speech. In each n-back task, participants listened to a pre-recorded series of 10 letters, separated by approximately 2.5-second intervals, for an overall duration of approximately 25 seconds. For the 1-back task (lower cognitive load), participants were asked to silently count the number of times two identical letters appeared back-to-back (e.g., PP). For the 2-back task (higher cognitive load), participants were asked to silently count the number of times two identical letters appeared in pairs separated by one letter in between (e.g., DTD). Participants were asked to verbally provide their answer at the end of each n-back task. Figure 2 offers examples of auditory input provided to participants with the target instances highlighted with different colors.

(a) (b)

Figure 2 Visualization of the modified n-back task: (a) example 1-back task, the correct answer is "1"; (b) example 2-back task, the correct answer is "2"

The driving scenarios were designed to involve mainly operational driving, without tactical decisions (e.g., navigation or passing a vehicle). Participants were asked to follow a lead vehicle at a speed of 40 mph (around 64.4 km/h) and a comfortable headway on a 4-lane urban road. In addition to training drives, each participant completed three drives with different levels of the modified n-back task: baseline with no task, lower cognitive load with 1-back task, and higher cognitive load with 2-back task.

The order of these three taskload levels was counterbalanced across participants. For machine learning models presented in this paper, data from four n-back tasks (a series of 10 letters for each n-back task) per drive were utilized. Participants had completed another two n-back tasks in each drive, but these corresponded to lead vehicle braking event response, which could affect physiological measures and was deemed to be outside the scope of the current analysis. In each n-back drive, the participants spent 100 seconds performing the four n-back tasks. This 100-second period for each n-back drive (i.e., 1-back and 2-back) and a corresponding 100-second period for the no-task drive were used in model building, leading to 300 seconds of data per participant being used in the machine learning models.

Procedures

After verifying their eligibility, participants were asked to sign a consent form. All participants completed a practice drive that was identical to the route used in the three experimental drives. Participants were then provided written and oral instructions on the modified n-back task and practiced it without driving to ensure that they fully understood the secondary task. The eye-tracking system was then calibrated and the physiological sensors were placed on participants. Then, participants completed another practice drive while performing the secondary task. Participants went on to complete the three experimental drives (no-task, 1-back, and 2-back).

DATA PROCESSING & MODEL TRAINING

A three-class classification problem was pursued in our analysis: no-task vs. 1-back task vs. 2-back task. We fitted KNN, SVM, FNN, RNN and RF models to different combinations of eye-tracking and physiological data. Overall, five different datasets were created as described in the section below.

Signal Processing and Feature Extraction

Table 2 summarizes our signal processing steps. In total, two physiological features were generated (HR and GSR) along with six eye-tracking features, including eye closure (EC) raw data, blink duration (BD), blink frequency (BF), pupil diameter (PD), eyeball rotation speed (eyeRS), and percentage of time at least 75% of iris is covered by eyelids (PERCLOS). These features were selected based on previous studies, which showed relationships between BF, PD, gaze dispersion and cognitive load, as summarized in Table 1. In place of SD of gaze position and periphery/mirror/instrument check rate reported in Table 1, we used eyeRS, which captures gaze dispersion and can be calculated independently of the driving scene (e.g., extracted through video of driver's face only). Although no clear relationship has been found between BD and cognitive load (e.g., 27) and PERCLOS is primarily a measure of drowsiness (e.g., 18), we opted to include these features, as they can be readily available from eye closure data. All eye-tracking features were calculated for each eye and were then averaged across the two eyes.

The measures that were originally sampled at a rate higher than 60 Hz (i.e., physiological measures) were down-sampled to 60 Hz in order have equal number of data points across all features. All features, except EC, GSR and PD, were calculated within a moving window. With a step size of 1/60 sec (i.e., 60 Hz), a window size of 10 sec was used for the eye-tracking features and a window size of 5 sec was used for HR. The ECG data is noisy and thus is commonly converted to inter-beat interval (ibi) data after R-peak detection (e.g., 14) and a running window procedure is necessary for this conversion. The 5-second window was adopted for HR based on our preliminary work on the same data investigating cognitive load detection (26). A longer time window was deemed necessary for eye-tracking measures given that, for example, the average blink frequency was recorded to be around 10 times/minute in (28) and 24 times/min in our dataset. Thus, a 10-second window size was chosen to provide a long enough period for reliable eye-tracking data extraction, but not too long compared to the entire data extraction period for each level of cognitive load (100 seconds), although earlier research used a longer time window for PERCLOS (30 seconds in (29) and 1 minute in (30)).

TABLE 2 Processing of Eye-Tracking and Physiological Measures to Obtain Machine Learning Features (i.e., Inputs)

Type	Measure	Features	Processing Steps
Eye-tracking Sampling frequency: 60 Hz	Eye closure (left and right): The fraction of the iris covered by the upper and lower eye lids (0: fully open to 1: fully closed)	4 features at 60 Hz: **EC**: eye closure raw data **BD**: blink duration (ms) **BF**: blink frequency (number/sec) **PERCLOS**: % of time at least 75% of iris is covered by upper and lower eyelids	1) Identify frames with eye closure over 75% and frames with eye fully closed (for left and right eye separately) 2) For each eye, calculate average **BD** and **PERCLOS** within a window size of 10 sec and step size of 1/60 sec (i.e., 60 Hz) 3) For each eye, calculate the inter-blink intervals (intervals between two eye closures) and convert them into **BF** within a window size of 10 sec and step size of 1/60 sec 4) Average each feature across the two eyes
	Pupil diameter (left and right)	1 feature at 60 Hz: **PD**: pupil diameter (mm)	1) Calculate average **PD** across the two eyes
	Gaze rotation angle (left and right): The orientation change of eyeball with respect to the world coordinate system, horizontally and vertically (rad)	1 feature at 60 Hz: **eyeRS**: eyeball rotation speed (rad/sec)	1) For each eye, calculate the eyeball rotation angle for each data point, i.e., the root-sum square of the horizontal and the vertical gaze rotation angles 2) Calculate average **eyeRS** (sum of eyeball rotation angles over the 10 sec time window / 10 sec) with a window size of 10 sec and a step size of 1/60 sec 3) Average the feature across the two eyes
Physiological Sampling frequency: 240 Hz	ECG	1 feature at 60 Hz: **HR** (/min)	1) Remove polynomial trend of raw ECG data using the *polyval* function in MATLAB and identify R-peaks of de-trended ECG using the *findpeaks* function (31) 2) Calculate the average inter-beat interval (ibi) with a window size of 5 sec with a step size of 1/60 sec 3) Convert ibi to **HR**
	GSR	1 feature at 60 Hz: **GSR** (μSiemens)	1) Calculate the average **GSR** every 1/60 sec (i.e., 60 Hz)

All features were extracted at 60 Hz leading to 594,000 rows of data (sampling frequency of 60 Hz * 25-seconds of data for each n-back task * 4 n-back tasks in each drive * 3 drives per participant * 33 participants). We built five datasets to train and evaluate our machine learning models:

- **eyeSet**: with eye-tracking features only;
- **physioSet**: with HR and GSR;
- **eyeHRset**: with eye-tracking features and HR;
- **eyeGSRset**: with eye-tracking features and GSR;
- **eyePhysioSet**: with all features.

Data Partition

As shown in Figure 3, a within-driver data partition approach was adopted, aiming to represent all participants in the training, validation, and test datasets. This data partition is selected over a between-drivers data partition method (which allocates some participants to the test dataset and the remaining to the training dataset), as we do not have a large enough sample to capture individual differences across participants that would be required for a between-drivers data partition. Hyperparameters were tuned using a 10-fold cross-validation on 90% of the data from each cognitive load level from each participant. The last 10% of the data from each level of cognitive load from each participant was used as the test dataset. Ten different splits (see Figure 3) were generated for the 10-fold cross-validation, each for one fold, with the training conducted on 81% of the data and the validation on 9%. A random split of training and test datasets was not appropriate for this data given its temporal nature and the resulting correlation over time.

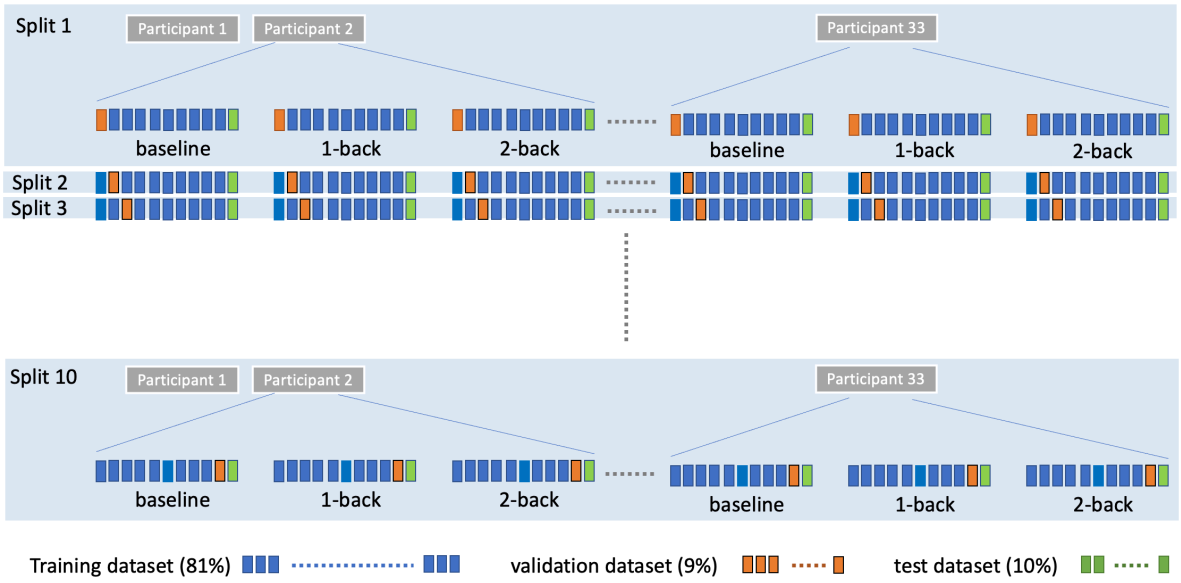


Figure 3 Visualization of data splits for cross validation and testing. Note that the boxes are arranged to fall on a timeline for each cognitive load level and represent 10 sec of data (with 100 sec for each level of cognitive load and 300 sec in total for each participant). Each blue and orange box represents 9% (i.e., 540 consecutive samples) and each green box represents 10% (i.e., 600 consecutive samples) of the total samples for each cognitive load level for each participant.

Data Preparation

Previous work has shown that individual differences among drivers influence the accuracy of driver state classification when eye-tracking and physiological data is used (32, 33). To minimize this

effect, each participant's data was normalized with respect to their no-task responses in the training dataset as in **Equation 1**:

$$X_{score} = \frac{X_{raw} - \bar{X}_{notask}}{S_{notask}} \quad (1)$$

where, X_{score} is the normalized feature value, X_{raw} is the raw feature value, and $\bar{X}_{notask}$ and S_{notask} are the mean and standard deviation of that feature for the no-task condition in the training dataset.

Further, each feature was also standardized using the scale of the features from the training dataset. Data standardization can control for scale differences across features and has been shown to improve the overall model accuracy in models such as KNN (34) and neural networks (35). Feature standardization was performed as in **Equation 2**:

$$X_{scaled} = \frac{X_{score} - \bar{X}_{training}}{S_{training}} \quad (2)$$

where, X_{scaled} is the standardized feature score, X_{score} is the normalized feature score in **Equation 1**, and $\bar{X}_{training}$ and $S_{training}$ are the mean and the standard deviation of the feature in the training dataset.

Model Training

All machine learning models were built in Python. Modules from the Scikit-Learn library (36) were used to train and test SVM, FNN, KNN, and RF, whereas Keras (37) was used for RNN. The specific functions used are documented in Table 3. Hyperparameters were tuned through a grid-search approach using cross-validation (i.e., iterating over combinations of parameters and selecting those that resulted in the highest average accuracy on validation datasets). All models were trained on an Apple MacBook Pro (16-inch, 2019) with 2.6GHz 6-Core Intel i7 CPU and 16 GB 2667 MHz DDR4 RAM. Graphical Processing Units were not used in the training of the neural network models.

Table 3 summarizes the candidate hyperparameters we tested and the best ones for each dataset and for each machine learning model. For KNN, the number of neighbours specifies the number of training data samples that can vote for the prediction of a given test data point. The weight function dictates how the voting samples are weighted. The distance metric dictates how the distances between the unknown sample and the voting samples are calculated. For SVM, a Radial Basis Function (RBF) kernel was used, as it yielded the best performance in most of the previous attempts in classifying levels of cognitive load with SVM (e.g., 15, 22, 26). Two additional hyperparameters were tuned for SVM. The regularization parameter defines "*how far the influence of a single training example reaches*" and the kernel coefficient defines "*how far the influence of a single training example reaches*" (36). For FNN, the activation function defines the (non-linear) output of the neuron in the network given a set of inputs, the learning rate controls how quickly the model is adapted to the problem and the L2 regularization rate decides how much the model is regularized to reduce the likelihood of overfitting.

Traditional RNNs may suffer from the exploding or vanishing gradient problems, i.e., if the input sequence is too long, the RNN model might be unstable (38, 39). A Long Short Term Memory (LSTM) architecture solves this problem by adding three gates to the network (40, 41), and has been found to perform well in time-sequence classification (42). For this reason, we used LSTM in our RNN architecture. Further, the use of a sliding window has been shown to improve the performance of neural networks for time-series prediction (43). Thus, we explored several combinations of moving window size and step size. Our RNN consisted of two LSTM layers, each followed by a dropout layer, which prevents the model from overfitting by randomly setting the input units to 0. The dropout rate defines the frequency of the input units being ignored (i.e., set to 0). Then, 1 to 5 layers of fully-connected layers were used, each followed by a dropout layer as well. The last fully-connected layer outputs the classification of the estimated cognitive load into one of the three classes. For RF, if bootstrap is true, the whole dataset is used to build each tree; otherwise, bootstrap samples are used when building trees, which means a subset of the samples was used for the training of each tree.

TABLE 3: Overview of model fitting and hyperparameters explored. The best combinations of hyperparameters for each model are indicated with the following superscripts: [e]eyeSet, [p]physioSet, [eh]eyeHRset, [eg]eyeGSRset, [ep]eyePhysioSet.

Model	Functions and candidate hyperparameters
KNN	Function: *KNeighborsClassifier* from Scikit-Learn library Number of neighbours: 1, 2[e, p, eh, ep], 3, 4[eg], 5, 6, 7, 8, 9, 10 Weight function: uniform, distance [e, p, eh, eg, ep] Distance metric: Euclidean, Manhattan [e, p, eh, eg, ep]
SVM	Function: *svm.SVC* from Scikit-Learn library Kernel (fixed): RBF [e, p, eh, eg, ep] Regularization parameter: 50.0[eh], 100.0[e, p, eg, ep]h Kernel coefficient:1[e, p, eh, eg, ep], 0.1
FNN	Function: *MLPClassifier* from Scikit-Learn library Architecture (number of neurons in each hidden layer): 2-8, 8-32, 16-64, 2-4-2, 8-16-8, 32-64-32, 2-4-8-4, 8-16-32-16, 16-32-64-32[e, p, eh, eg, ep] Minibatch size (fixed): 50[e, p, eh, eg, ep] Activation function: *tanh* [e, p, eh, eg, ep], *ReLU* Learning rate: constant at 0.001 [eh, eg], adaptive (initialized at 0.001) [e, p, ep] Regularization rate for L2 penalty: 0.001, 0.01[e, p, eh, eg, ep]
RNN	Function: *Sequential* from Keras library Batch size: 32, 64[e, p, eh, eg, ep], 128 Sliding window (window-overlap): 6-3, 8-4[e, p, eh, eg, ep], 10-5, 20-10 Learning rate: 0.01, 0.001[e, p, ep], 0.0001 [eh, eg] Architecture: 1st LSTM layer: • Dimensionality of the output space: 32-512 • Activation function (fixed): *tanh* [e, p, eh, eg, ep] • Whether to return the last state in addition to the output: True [e, p, eh, eg, ep] Dropout layer: • Drop rate: 0-0.3 2nd LSTM layer: • Dimensionality of output space: 32-512 • Activation function (fixed): *tanh* [e, p, eh, eg, ep] • Whether to return the last state in addition to the output: False [e, p, eh, eg, ep] Dropout layer: • Input units to drop: optimal value between 0-0.3 1st to nth (n: 1[p, eh], 2[eg], 3[ep], 4[e], 5) dense layer, each followed with one dropout layer: • Dimensionality of output space: 32-512 • Activation function: *ReLU*, *tanh*, *Sigmoid* • Drop rate: 0-0.3 Last dense layer: • Activation function (fixed): *Softmax* [e, p, eh, eg, ep]
RF	Function: *RandomForestClassifier* from Scikit-Learn library Number of trees in the forest: 5, 10, 20, 30, 50, 100 [e, p, eh, eg, ep] Function to measure the quality of a split (fixed): Gini impurity [e, p, eh, eg, ep] Bootstrap or not: True [e, p, eh, eg, ep], False

RESULTS

Figure 4 shows the classification accuracy on the test dataset for each machine learning model for the different combinations of features. Figure 5 provides confusion matrices on the test dataset. The average accuracies on cross-validation were comparable to the test accuracies and thus are not reported; this indicates that overfitting or underfitting are unlikely to have occurred.

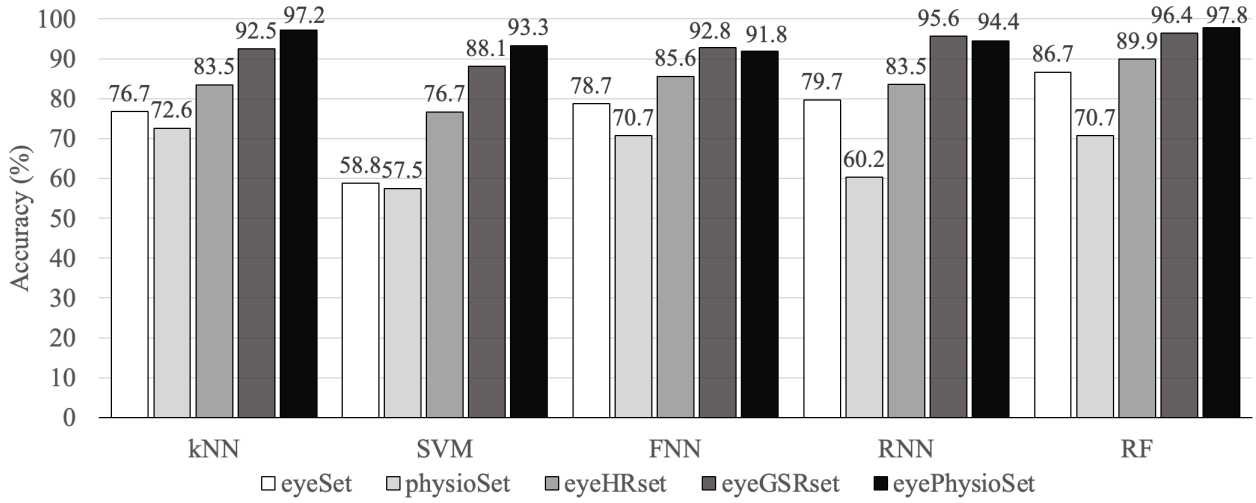


Figure 4 Classification accuracies in identifying three levels of cognitive load on test dataset

It can be observed that RF generated the highest prediction accuracy (97.8%) when all eye-tracking and physiological features were used (eyePhysioSet data). Although using physiological features alone (physioSet) generated worse prediction accuracy (1-15%) compared to using eye-tracking features alone (eyeSet), adding physiological features in addition to eye-tracking features increased the model prediction accuracies. Among the physiological features, GSR seems to have provided more predictive power compared to HR. When comparisons are made across machine learning models, it can be seen that the discrepancies between different models decreased with the expansion of the feature set. The confusion matrices indicate that different models may be good at identifying different levels of cognitive load, even if the models may be comparable in terms of their overall accuracy. For example, with all features utilized, RF was better at differentiating no task from 1-back task (i.e., lower level of cognitive load) compared to KNN, but KNN was better at differntiating 1-back task from 2-back task (i.e., higher level of cognitive load).

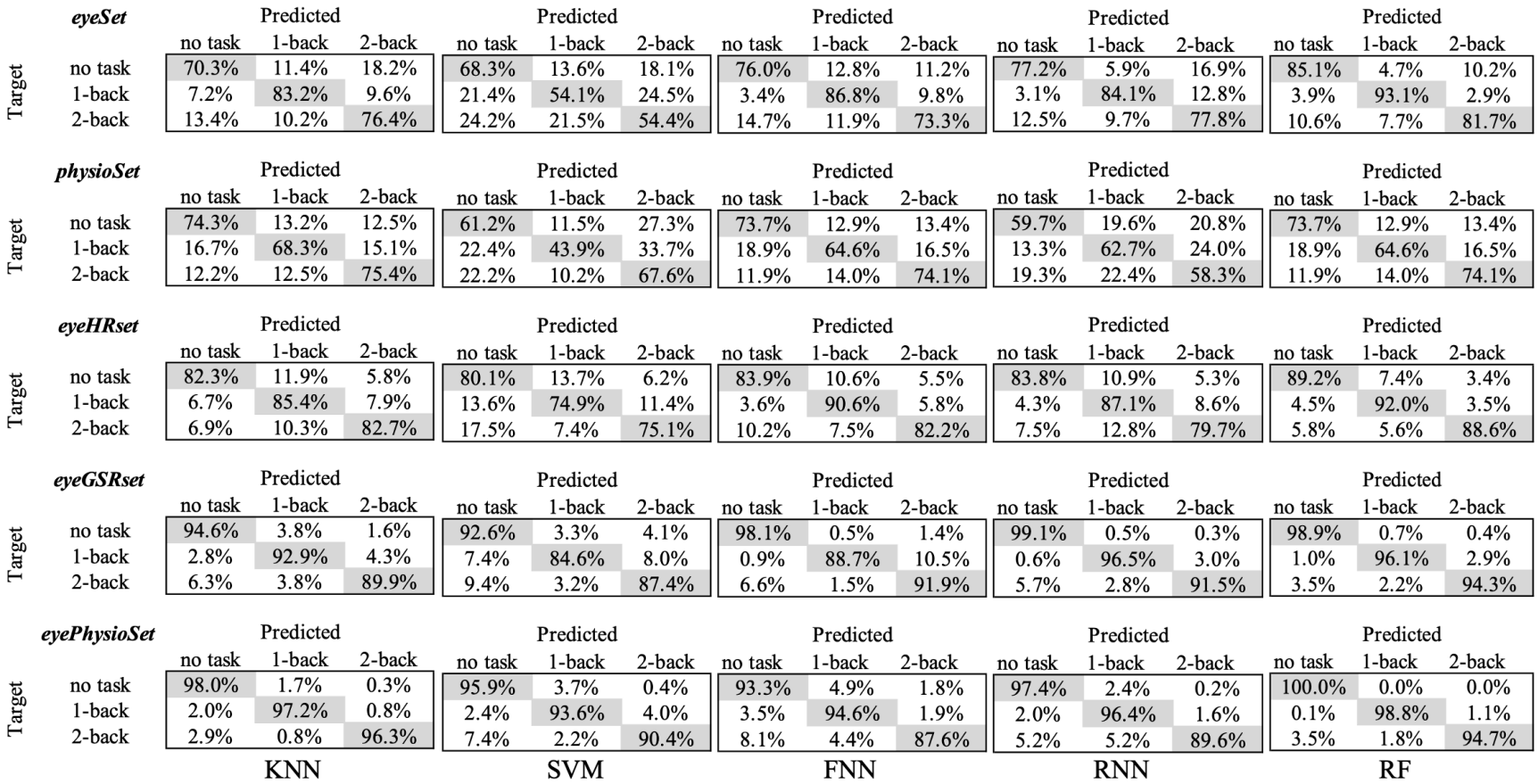


Figure 5 Confusion matrices for classification performance on the test dataset

DISCUSSION

This paper revealed the potential for improving cognitive load estimation through combining eye-tracking and physiological measures. Eye-tracking measures are being adopted in production vehicles for distraction (e.g., 17) and drowsiness (e.g., 18) detection, but also hold promise for cognitive load detection (e.g., 2, 6, 7). Further, physiological measures are also promising for cognitive load detection: earlier studies that used physiological predictors to classify drivers' cognitive load reported classification accuracies of 85-96% (14, 22, 26, 44). However, not all physiological measures are suitable for driver state estimation due to the intrusiveness of associated sensors (e.g., Electroencephalography). Further, although the fusion of eye-tracking and physiological measures seems to be more promising for real-time assessment of driver cognitive load, research is lacking in this area.

In this paper, two physiological measures that are available in consumer-grade wearable devices, i.e., HR and GSR, were fused with eye-tracking measures, leading to 97.8% accuracy with a random forest (RF) model in classifying three levels of cognitive load (no task, lower difficulty 1-back task, and higher difficulty 2-back task). This result is promising when compared with the accuracies reached in previous research that combined driving performance with eye-tracking measures (e.g., 81.1% in (15)), and also with the accuracies reached in previous research that combined driving performance measures with physiological measures (e.g., 89% in (14)), especially considering that our 3-class classification problem is more challenging than the 2-class problems tackled in these earlier studies. GSR contributed more to the performance of the models compared to HR: when HR was added to eye-tracking data, the model accuracies increased from 3.2% (with RF) to 17.9% (with SVM), whereas with HR, the increases ranged from 9.7% (with RF) to 29.3% (with SVM). Although adding both HR and GSR to eye-tracking data yielded the highest accuracies for most models, the benefit of adding HR on top of GSR was relatively small, with changes in model accuracies ranging from -1.2% (with RNN) to 5.2% (with SVM). Collecting and processing more features may come with monetary and computational costs, and if a choice is to be made, GSR may be preferred over HR.

Our findings reveal that, with the increased number of features, the advantage of using a specific machine learning model becomes less obvious. With only eye-tracking measures, RF yielded the highest accuracy (86.7%) and SVM the lowest (58.8%). When both eye-tracking and physiological measures were used, all models reached over 91% accuracy. It is possible that with more features, the classification problem became easy enough for most models to handle. At the same time, we also note that KNN, which yielded the second highest accuracy (97.2%) on the combined eye-tracking/physiological dataset, took the shortest to train with 7.6 seconds, and RF, which resulted in the highest accuracy (97.8%), took slightly longer with 61.6 seconds. The training time for FNN, RNN and SVM were two orders of magnitude longer (all over 5,000 seconds) than that of KNN and RF. Thus, even with the computation cost of model training considered, RF and KNN are preferred over other models for the cognitive load detection problem explored in our study.

Although overall accuracies across models were comparable when all features were utilized, the confusion matrices reveal that different models may be good at identifying different levels of cognitive load. For example, when all features were utilized, RF reached the overall highest accuracy, but KNN performed better than RF in differentiating 2-back from 1-back. Thus, the choice of models may not be based solely on overall accuracies, but also on the specific purpose of the in-vehicle applications, for example, which level of cognitive load is most critical to differentiate from other levels for an alert to be issued.

Individual differences among drivers have been shown to impact the accuracy of driver state classification based on physiological data (32), and thus we used a normalization strategy, which would require the system to have prior data from each driver, or learn from the driver over time. This is a reasonable expectation but prior data may not always be available for each driver. Future research should utilize a larger sample with a more diverse set of drivers to test the generalization of our models by training the models based on a group of participants and predicting the cognitive load of a different group of participants (i.e., between-drivers data partition). It should however be noted that if model

training/testing is based on a between-drivers data partition, individual differences can have a significant impact on model performance.

Further, although we utilized a validated secondary task to impose cognitive workload on our participants, the task is artificial and there is a need to study more the tasks that drivers normally perform in their vehicles (e.g., talking on a cellphone). Additionally, in this paper, we focused on physiological measures (i.e., HR and GSR) that can be collected through wearable devices, but our data came from a research-grade system utilized in a driving simulator study. Further, the eye-tracking measures were collected using an eye-tracker built for the laboratory environment. There are bound to be additional signal noise issues when these measures are collected through wearable devices and cameras, and in a real vehicle. Although our study provided evidence that HR and GSR have the potential to be fused with eye-tracking data to improve driver cognitive load estimation, more research is needed to develop signal processing algorithms for relevant data collected through consumer-grade wearable devices in motion and through in-vehicle eye-tracking systems under varying lighting conditions. As in (45-47), there is indeed significant research activity to improve these devices and accompanying algorithms. Future work needs to be conducted to train the models based on larger sample size collected from real-world driving environment, instead of from driving simulator experiments, to improve the feasibility of the models. Finally, the time window sizes used for feature extraction may affect the performance of the models. Future research should explore different time window sizes appropriate for this application.

In summary, physiological and eye-tracking features that can be collected through in-vehicle or wearable devices combined with the algorithms developed in our paper have the potential to support less intrusive driver state detection. Given that our approach excluded driving performance measures, it can also inform driver state detection for automated vehicles where driving performance data may not be indicative of driver state.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: D. He, B. Donmez; analysis and interpretation of results: D. He, Z. Wang, E. B. Khalil, B. Donmez, and G. Qiao; draft manuscript preparation: D. He, E. B. Khalil, B. Donmez, Z. Wang, and S. Kumar. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Harbluk JL, Noy YI, Trbovich PL, Eizenman M. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. Accident Analysis & Prevention. 2007;39(2):372-9.
2. Recarte MA, Nunes LM. Effects of verbal and spatial-imagery tasks on eye fixations while driving. Journal of Experimental Psychology: Applied. 2000;6(1):31.
3. Rogers S, Fiechter C-N, Thompson C, editors. Adaptive user interfaces for automotive environments. Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat No 00TH8511); 2000; Dearborn, MI, USA: IEEE.
4. Strayer DL, Cooper JM, Turrill J, Coleman JR, Hopman RJ. The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale. 2017;71(2):93.
5. Johns M, Sibi S, Ju W, editors. Effect of cognitive load in autonomous vehicles on driver performance during transfer of control. Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications; 2014; New York, NY, USA: Association for Computing Machinery.
6. Recarte MA, Nunes LM. Mental workload while driving: Effects on visual search, discrimination, and decision making. Journal of Experimental Psychology: Applied. 2003;9(2):119.

7. Liang Y, Lee JD. Combining cognitive and visual distraction: Less than the sum of its parts. Accident Analysis & Prevention. 2010;42(3):881-90.
8. Mehler B, Reimer B, Coughlin JF. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. Human Factors: Journal of Human Factors and Ergonomics Society. 2012;54(3):396-412.
9. Mehler B, Reimer B, Coughlin JF, Dusek JA. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transportation Research Record. 2009;2138(1):6-12.
10. Brookhuis KA, de Vries G, De Waard D. The effects of mobile telephoning on driving performance. Accident Analysis & Prevention. 1991;23(4):309-16.
11. Ryu K, Myung R. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. International Journal of Industrial Ergonomics. 2005;35(11):991-1009.
12. Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neuroscience and Biobehavioral Reviews. 2014;44:58-75.
13. Strayer DL, Turrill J, Cooper JM, Coleman JR, Medeiros-Ward N, Biondi F. Assessing cognitive distraction in the automobile. Human Factors: Journal of Human Factors and Ergonomics Society. 2015;57(8):1300-24.
14. Solovey ET, Zec M, Garcia Perez EA, Reimer B, Mehler B, editors. Classifying driver workload using physiological and driving performance data. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14; 2014; Toronto, ON, Canada.
15. Liang Y, Reyes ML, Lee JD. Real-time detection of driver cognitive distraction using support vector machines. IEEE Transactions on Intelligent Transportation Systems. 2007;8:340-50.
16. Miller S. Literature Review Workload Measures. Iowa City, United States: National Advanced Driving Simulator; 2001. Report No.: N01-006.
17. Cadillac. Super Cruise - Hands Free Driving | Cadilac Ownership 2021 [Available from: https://www.cadillac.com/world-of-cadillac/innovation/super-cruise.
18. Khan MQ, Lee S. A comprehensive survey of driving monitoring and assistance systems. Sensors. 2019;19(11):2574.
19. APPLE. The Future of Health is on Your Wrist [Available from: https://www.apple.com/ca/watch/.
20. Fitbit. Understand Your Stress So You Can Manage It [Available from: https://www.fitbit.com/global/us/technology/stress.
21. He D, Donmez B, Liu CC, Plataniotis KN. High cognitive load assessment in drivers through wireless Electroencephalography and the validation of a modified n-back task. IEEE Transactions on Human-Machine Systems. 2019;49(4):362-71.
22. Wang YK, Jung TP, Lin CT. EEG-based attention tracking during distracted driving. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2015;23(6):1085-94.
23. Shimizu T, Shima K, Mukaeda T, Muraji S, Matsuo J, Horiue M, editors. Real-time evaluation of driver cognitive loads based on multivariate biosignal analysis. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2020: IEEE.
24. Barua S, Ahmed MU, Begum S. Towards intelligent data analytics: A case study in driver cognitive load classification. Brain Sciences. 2020;10(8):526.
25. He D, Liu CC, Donmez B, Plataniotis KN, editors. Assessing high cognitive load in drivers through Electroencephalography. Proceeding of Transportation Research Board 96th Annual Meeting; 2017; Washington D.C.

26. He D, Risteska M, Donmez B, Chen K. Driver mental workload classification through the use of physiological data. In: Eslambolchilar P, Komninos A, Dunlop M, editors. Digital Signal Processing and Machine Learning for Interactive Systems Developers. Association for Computing Machinery: ACM Books; 2021.
27. Tsai Y-F, Viirre E, Strychacz C, Chase B, Jung T-P. Task performance and eye activity: predicting behavior relating to cognitive workload. Aviation, space, and environmental medicine. 2007;78(5):B176-B85.
28. De Padova V, Barbato G, Conte F, Ficca G. Diurnal variation of spontaneous eye blink rate in the elderly and its relationships with sleepiness and arousal. Neuroscience Letters. 2009;463(1):40-3.
29. Cardone D, Filippini C, Mancini L, Pomante A, Tritto M, Nocco S, et al., editors. Driver drowsiness evaluation by means of thermal infrared imaging: Preliminary results. Infrared Sensors, Devices, and Applications XI; 2021: International Society for Optics and Photonics.
30. Rodríguez-Ibáñez N, García-González MA, Fernández-Chimeno M, Ramos-Castro J, editors. Drowsiness detection by thoracic effort signal analysis in real driving environments. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2011: IEEE.
31. MathWorks. R Wave Detection in the ECG [Available from: https://www.mathworks.com/help/wavelet/ug/r-wave-detection-in-the-ecg.html.
32. Lin C-T, Wu R-C, Liang S-F, Chao W-H, Chen Y-J, Jung T-P. EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Transactions on Circuits and Systems I: Regular Papers. 2005;52(12):2726-38.
33. Li K, Jin L, Jiang Y, Xian H, Gao L. Effects of driver behavior style differences and individual differences on driver sleepiness detection. Advances in Mechanical Engineering. 2015;7(4):1687814015578354.
34. Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4(2):1883.
35. Shanker M, Hu MY, Hung MS. Effect of data standardization on neural network training. Omega. 1996;24(4):385-97.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
37. Chollet F. Keras: The python deep learning library. Astrophysics Source Code Library. 2018.
38. Pascanu R, Mikolov T, Bengio Y, editors. On the difficulty of training recurrent neural networks. International Conference on Machine Learning; 2013; Atlanta, USA.
39. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:150600019. 2015.
40. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems. 2016;28(10):2222-32.
41. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-80.
42. Wang J-H, Liu T-W, Luo X, Wang L, editors. An LSTM approach to short text sentiment classification with word embeddings. Proceedings of the 30th Conference on Computational linguistics and Speech Processing (ROCLING 2018); 2018; Hsinchu, Taiwan, China.
43. Frank RJ, Davey N, Hunt SP. Time series prediction and neural networks. Journal of Intelligent and Robotic Systems. 2001;31(1-3):91-103.
44. Kohlmorgen J, Dornhege G, Braun ML, Blankertz B, Müller K-R, Curio G, et al. Improving human performance in a real operating environment through real-time mental workload detection. In: Dornhege G, Millan JdR, Hinterberger T, McFarland DJ, Mller K-R, editors. Toward Brain-Computer Interfacing. Cambridge, Massachusetts: MIT Press; 2007. p. 409-22.
45. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review. JMIR mHealth and uHealth. 2020;8(9):e18694.

46. Zhu L, Du D, editors. Improved Heart Rate Tracking Using Multiple Wrist-type Photoplethysmography during Physical Activities. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018: IEEE.
47. Binaee K, Sinnott C, Capurro KJ, MacNeilage P, Lescroart MD, editors. Pupil Tracking Under Direct Sunlight. ACM Symposium on Eye Tracking Research and Applications; 2021.