Classification of Driver Cognitive Load based on Physiological Data: Exploring Recurrent Neural Networks

Shekhar Kumar, Dengbo He, Guangkai Qiao, and Birsen Donmez

Abstract— **In-vehicle systems can lead to high cognitive load that impairs driving performance. Interfaces that can detect and adapt to cognitive load accordingly may alleviate these effects. Previous research explored machine learning models to classify drivers' cognitive load based on physiological signals but most conducted training and testing on data from the same participants (i.e., within-driver partitioning), which raises generalizability and practical feasibility concerns. In this paper, we explored the performance of widely-used models by training and testing them on data from different subjects (i.e., across-drivers partitioning), and further compared them with a more recent model that is effective for time-series data, the recurrent neural network (RNN). A driving simulator dataset was used to classify 2 levels of cognitive load (external cognitive secondary task vs. no task). All models performed better with within-driver partitioning. RNN outperformed other models with mean accuracies of 88.1% and 85.6% with within-driver and across-drivers partitioning, respectively.**

I. Introduction

Both simulator and on-road studies have found that high cognitive load can negatively impact driving performance and visual scanning behaviors of drivers [1, 2]. The cognitive load experienced by drivers can be due to a number of sources, both from outside (e.g., heavy traffic) and inside the vehicle (e.g., engagement with in-vehicle technologies). The assessment of cognitive load in real time can potentially improve driving safety through preventative warnings and adaptive interfaces that can, for example, limit functionality [3] or filter information [4] in infotainment systems. The intelligent transfer of vehicle control to drivers can also be made possible when high cognitive load is detected in automated vehicles [5]. However, while visual demands can be detected through technologies such as gaze tracking [6], cognitive load is more challenging to capture.

The variation in various physiological measures have been found to be associated with external cognitive load experienced by drivers, such as galvanic skin response (GSR), and Electroencephalography (EEG), Electrocardiography (ECG) [7, 8]. Physiological measures have become particularly promising for real-time assessment of external cognitive load experienced by drivers, as devices that can collect such measures are being incorporated to consumer products and have become less intrusive to the wearer [9, 10].

Previous studies have investigated the use of physiological measures alone and in combination with other predictors to classify drivers' cognitive load (see Table I for an overview). These studies explored a variety of machine learning methods including support vector machine (SVM), artificial neural network (ANN), k-nearest neighbor (KNN), decision tree classifier (DTC), Naïve Bayes (NB), and linear discriminant analysis (LDA) with some level of success. The majority of the models explored in these previous studies were trained and tested on the same group of participants (i.e., within-driver partition), which raises generalizability and practical feasibility concerns. The generalizability of these models has to be assessed via across-drivers partitioning, or in other words, training should be performed on one set of drivers while testing is performed on another set. Further, within-driver partitioning assumes that in practice a new model would have to be trained for each new user of a system that utilizes these models. This would not be practically feasible. To explore the generalizability of the machine learning models used in previous research in estimating driver cognitive load through physiological data, this paper utilized widely-used machine learning models (i.e., SVM, ANN, KNN, DTC, NB, LDA) and compared their accuracy when they are trained and tested on the same participants (i.e., within-driver partitioning) vs. different participants (i.e., across-drivers partitioning).

Further, a more recent machine learning method, i.e., Recurrent Neural Networks (RNN), was also explored. RNNs differ from earlier machine learning methods, in that they are trained on time sequence (i.e., temporal) data but have not yet been explored for driver cognitive load classification. Although they have not been explored for driver cognitive load classification, RNNs have been used to classify general working memory load; Kuanar et al. [11] achieved a high level of accuracy (92.5%) for classifying four difficulty levels of a memory task using a 64-channel EEG. RNNs are promising for driver cognitive load classification, given that driving data is highly temporal. Thus, this paper also compared the driver cognitive load classification performance of widely-used machine learning models to that of RNN.

II. Data Source

In a driving simulator study, we collected physiological, eye-tracking, and driving performance data from 33 participants. This dataset was initially reported in [17, 18], which investigated how different levels of external cognitive demand (as imposed by an external cognitive task) are correlated with driving performance and driver's physiology. The experiment adopted a within-subject design, i.e., each participant completed three drives, each with a unique experimental condition: a baseline drive without external task load, and two drives with two variations of external cognitive task (i.e., a secondary task) with different difficulties. The order of the three drives were counterbalanced. Physiological

Research supported by Natural Sciences and Engineering Research Council of Canada (NSERC). Shkhar Kumar Guangkai Qiao and Birsen Donmez (corresponding author) are with the Univeristy of Toronto, Toronto, ON M5S 3G8 Canada (email: shekhar.kumar@mail.utoronto.ca; qiaogk@sjtu.edu.cn; donmez@mie.utoronto.ca). Dengbo He is now with the Systems Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He is also with the Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong SAR, China. He was with the University of Toronto, Toronto, ON, Canada (e-mail: dengbohe@ust.hk).

TABLE I: PREVIOUS RESEARCH ON DRIVER COGNITIVE LOAD CLASSIFICATION THROUGH PHYSIOLOGICAL SIGNALS

Study	Features	Classification Problem	Models	Performance
He et. al [12] Simulator study with 33 participants	Power spectra from 2-channel EEG; Heart Rate (HR); GSR	***3 classes***: No-task vs. two levels of an auditory-verbal n-back task (1-back and 2-back)	SVM, ANN, KNN, DTC, NB, and LDA within-driver partition (25% of each participant data allocated to testing)	Highest accuracy obtained with SVM: 79.4%
Kohlmorgen et. al [13] On-road study with 17 participants	Power spectra from 32-channel EEG	***2 classes (problem 1)***: No-task vs. listening to an audio book in the presence of another audio stream (news) ***2 classes (problem 2)***: No-task vs. silently counting down by 27 and responding verbally when prompted	LDA within-driver partition (1/11th of each participant data allocated to testing)	***Problem1:*** 95.6% max. and ~49% min. accuracy across participants ***Problem 2:*** 91.8% max. and ~55% min. accuracy across participants
Liang et. al [14] Simulator study with 10 participants	Eye movement; driving performance (steering wheel position, steering error, lane position)	***2 classes***: No-task vs. identifying the direction of value changes for two target stocks in an auditory stock ticker task	SVM within-driver partition (randomly selected testing data from all participants)	81.1% accuracy
Solovey et. al [15] On-road study with 99 participants	HR; GSR; driving performance (speed, steering wheel position, acceleration)	***2 classes***: No-task vs. an auditory recall n-back task (2-back)	ANN, KNN, DTC, and NB across-drivers partition (10% of participants allocated to testing)	Highest accuracy obtained with NB: 89%
Wang et. al [16] Simulator study with 10 participants	Power spectra from 32-channel EEG	***2 classes***: No-task vs. verifying the correctness of arithmetic equations presented visually	SVM within-driver partition (test data obtained from all participants)	84.6% accuracy

measures that were collected included EEG, ECG, and GSR. A detailed description of the experiment design can be found in [18]. We only provide a brief summarization of the experiment design in this paper. Some preliminary machine learning models were implemented based on this dataset in He et al. [12] (as presented in Table I), but without RNN and only through the use of within-driver partitioning for pre-processing of the data.

A. Participants

Eighteen males and fifteen females (33 total), recruited through campus and online posts, completed the study. Participants were screened based on their years of licensure (over 3 years), driving frequency in the past year (at least several times a month), type of license they held (full driver's license in Ontario, Canada) and age (less than or equal to 35). The mean age of the participants was 27.6 with a standard deviation (SD) of 4.45.

The participants were told that they would receive a compensation at a rate of C$12/hour, and a bonus of up to C$14 based on their performance in the secondary task to encourage them to engage in the task as they drove. However, all participants received C$50 regardless of their performance.

B. Apparatus

The study was conducted on a fixed-based NADS miniSim™ driving simulator (Figure 1a). It has three 42-inch screens that creates a 130-degree horizontal and 24-degree vertical field when viewed at a distance of 48-inch (around 1.22 meters). The left and center parts of the windshield are displayed in the center screen; the rest of the windshield, the rear-view mirror, and the right-side window and mirror are displayed in the right screen, and the left-side window and mirror in the left screen.

A commercial-grade EEG wireless headband (i.e., Muse™ by Interaxon, shown in Figure 1b) collected EEG data with dry sensors at a sampling frequency of 220 Hz. Data was collected from Fp1 and Fp2 positions (two EEG positions on the forehead) which were found to be sensitive to changes in drivers' cognitive load [18]. The software, MuseLab, was used to record and analyze the EEG signals. Similar to He et. al [18], the EEG signals were transferred into five power bands at 10 Hz.

The sensors by Becker Meditec were used to collect ECG and GSR signals, both at 240 Hz sampling frequency. ECG and GSR data were recorded using the D-Lab software by Ergoneers. Three solid gel foam electrodes were used to collect the ECG signal. The electrodes were placed on participants' chest (as shown in Figure 1c). Two solid gel foam electrodes were placed beneath the bare left foot to collect the GSR signal – one in the middle and the other under the heel (Figure 1c).

C. Experimental Tasks

The participants followed a lead vehicle on a 4-lane urban road and were asked to drive at a speed of 40 mph (around 64.4 km/h).

An auditory-verbal n-back task was used as the secondary task to impose different levels of cognitive load on drivers. In each of the three experimental drives, participants were assigned one level of the secondary task: baseline (no task), 1-back task (lower cognitive load), and 2-back task (higher cognitive load).

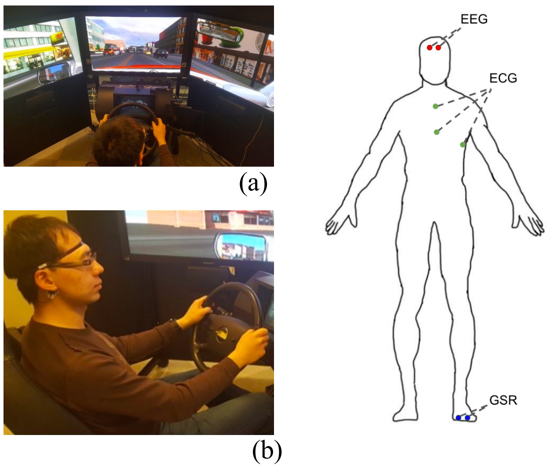


Figure 1. Apparatus: (a) driving Simulator; (b) EEG Measurement – Muse by Interaxon; (c) placement of EEG, ECG, and GSR Sensors

To minimize interference with the EEG signal, the n-back task was modified as described in [17, 18]. In this modified n-back task, participants listened to a series of 10 letters from a recording. The letters were separated by 2.5-second intervals, leading to a total duration of 25 seconds for each n-back task. There were 4 n-back tasks within each drive, leading to a 100 second total. A corresponding 100-second section was extracted from the baseline drive for data analysis. Thus, a total of 300-second data from each participant was used for this paper.

For the n-back task, participants needed to listen to the letters and count silently how many times specific patterns appeared. The pattern for the 1-back task was two identical letters appearing back to back (e.g., DD), for the 2-back task it was two identical letters appearing back to back but separated by one letter in between (e.g., GTG). After each n-back task, participants were required to say how many times they heard the pattern. In Figure 2, examples of 1-back (highlighted in red) and 2-back patterns (highlighted in green) are provided.

Stimuli (auditory):	C F F B H B C A C G
1-back response (vocal):	1
2-back response (vocal):	2

Figure 2. Visualization of the modified n-back task. For 1-back task, the correct answer is "1"; for 2-back task, the correct answer is "2".

D. Procedures

Consent was obtained after the verification of participant eligibility. After consent, participants completed a practice drive with no secondary task, followed by written and oral instructions on the modified n-back task and a practice of the task without driving. The participants then finished another practice drive with EEG, ECG and GSR sensors on. They were also encouraged to practice secondary task as much as possible in this drive. Participants then finished the three experimental drives in counterbalanced order. The routes used in the practice drives and the experimental drives were identical.

III. Data Processing & Model Training

A. Signal Processing and Feature Extraction

Table II summarizes our signal processing steps. In total, 12 attributes were generated: 10 for EEG, 1 for ECG, and 1 for GSR. The total dataset had 12 features and 99000 data points (i.e., 10 Hz sampling frequency of physiological data * 25-seconds of data per n-back task * 4 n-back tasks per drive * 3 drives per participant * 33 participants). Earlier analysis reported on this dataset did not reveal significant differences in driving performance between 1-back and 2-back task conditions, although there were differences in physiological measures [18]. Thus, a binary classification problem was pursued in this paper: no-task vs. n-back tasks (either 1-back or 2-back). Binary classification was also adopted in all studies reported in Table I except one. We fitted RNN, SVM, ANN, KNN, DTC, NB, and LDA.

TABLE II: Signal Procesing and Feature Extraction

Physiological Signal	*Processing Steps*	*Output*
EEG	- Remove 60 Hz utility frequency - Apply Fast Fourier Transformation (FFT) - Integrate to obtain the power of delta (1-4 Hz), theta (4-8 Hz), alpha (7.5-13 Hz), beta (13-30 Hz), gamma (30-44 Hz) bands for every 0.1 second of data	***10 attributes***: the power of 5 bands (delta, theta, alpha, beta, gamma) from Fp1 and Fp2 positions at 10 Hz
ECG	- De-trench raw ECG data - Calculate the average inter-beat interval for every 5 seconds and convert to average heart rate (HR)	***1 attribute***: average HR updated every 5 seconds
GSR	- Calculate the average GSR for every 0.1 second interval	***1 attribute***: average GSR at 10 Hz

B. Data Partition: Training and Test Sets

Similar to the approaches used in previous studies, i.e., He et. al [12], a within-driver data partition aimed to represent all participants both in the training and the testing datasets was used. In order to test model robustness, each model was trained and tested five times on five different splits of the data, where the test dataset for each split consisted of 20% of data from each of the three cognitive load conditions (no-task, 1-back task, and 2-back task), with the remaining 80% used for training. While in the across-drivers partition, about 20% of the participants (n=6 or 7) were allocated to the test and the remaining to the training dataset. This split was performed five times and each model was fitted five times in order to assess their robustness. The chronological order of participant data was first randomized before the across-drivers partitioning.

C. Data Preparation

Previous work has identified negative effects of individual differences on the accuracy of driver state classification when using physiological measures [19][20]. Thus, each participant's data was normalized using their data from the baseline condition in the training dataset to minimize this effect:

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

where, X_{score} represents the normalized feature (an attribute from Table II) value; X_{raw} refers to the raw feature value; and $\bar{X}_{notask}$ and S_{notask} are the mean and SD of that feature calculated from the baseline condition in the training dataset. Next, each feature was further standardized as follow:

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

where, X_{scaled} represents the standardized feature score; X_{score} refers to the previously computed normalized feature score; and $\bar{X}_{training}$ and $S_{training}$ are the mean and the SD of the features calculated based on all three task conditions in training dataset.

For the RNN model specifically, two additional transformations were then applied. First, a sliding window with a size of 5 and overlap of 4, was applied to smooth the data. Second, a tanh activation function (as shown below) was applied in order to scale each attribute to the interval of [-1,1], where, x refers to the input of the activation function (i.e., X_{scaled} for our data).

$$tanh\, x = \frac{sinh\, x}{cosh\, x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \quad (3)$$

D. Model Training

The machine learning models explored in this study were built in Python, with the details of the hyperparameters presented in Table III.

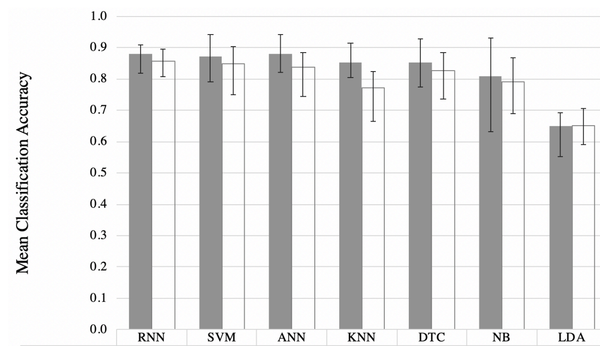
TABLE III: OVERVIEW OF MODEL DETAILS

Model	Details
RNN	Function: Sequential in Keras Layers: 2-layer LSTM Sequence length: 20 Sequence step: 2
SVM	Function: svm.SVC in Scikit-Learn Regularization parameter: 1.0 Kernel: Radial basis function Polynomial: 3rd degree
ANN	Function: MLPClassifier in Scikit-Learn Activation function: Rectified linear unit Number of hidden layers: 100 Learning rate: Constant
KNN	Function: KNeighborsClassifier in Scikit-Learn Number of neighbours: 3 Leaf size: 30
DTC	Function: DecisionTreeClassifier in Scikit-Learn Max depth: 7 Split criterion: Gini
NB	Function: GaussianNB in Scikit-Learn Variance smoothing: 1e-9
LDA	Function: LinearDiscriminantAnalysis in Scikit-Learn Solver: Singular value decomposition Shrinkage: None

The SVM, ANN, KNN, DTC, NB, and LDA were built using the Scikit-Learn library [21]; the RNN models were built using Keras [22]. As cross validation was not feasible due to the time-series nature of the training data, RNN hyperparameters were selected as those which achieved the highest level of accuracy on the test set; whereas the hyperparameters of other models were tuned using 5-fold cross-validation through a grid-search approach (i.e., out of all possible combinations of hyperparameters, select the ones that resulted in the highest accuracy). To avoid gradient exploding problem and the vanishing problem of RNN [23, 24], a Long Short Term Memory (LSTM) architecture was adopted by adding three gates to the RNN network [25, 26], which has been found to work well for classification based on time-sequence data [27].

IV. RESULTS

Figure 3 presents the accuracy of the various models for both the within-driver and the across-drivers partitioning of the training and test datasets. Accuracy for within-driver partitioning was larger than across-drivers partitioning, likely because models trained on all participants' data are more sensitive to individual differences, and therefore more prone to overfitting. RNN reached the highest classification accuracy compared to other models for both the within-driver and the across-drivers partitioning. Further, it had the narrowest range of accuracy values obtained for the five splits (see error bars in Figure 3) suggesting a more robust performance.



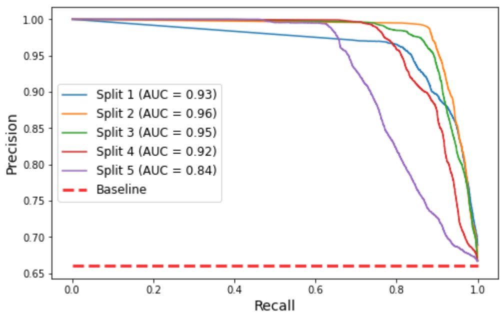
	RNN	SVM	ANN	KNN	DTC	NB	LDA
Within-Driver	0.881	0.872	0.879	0.854	0.852	0.808	0.649
Across-Drivers	0.856	0.849	0.837	0.772	0.828	0.790	0.652

Figure 3. Mean classification accuracy (error bars represent the range of accuracies across five model fits for each method and partitioning).

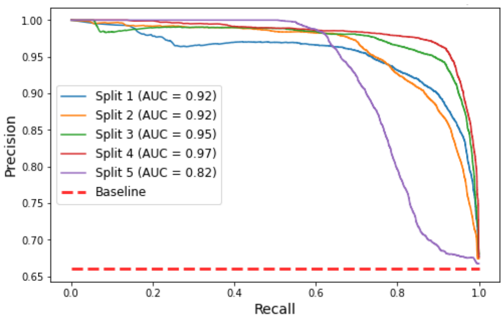
Although accuracy is a useful tool in benchmarking model performance, relying exclusively on accuracy may be misleading given that it treats false positives and false negatives equally and does not provide a detailed account of the two. Furthermore, given that the experiment was balanced across three conditions, the binary classification perspective led to a data imbalance: 1/3 of the data came from the no-task condition and 2/3 from an n-back task condition (either 1-back or 2-back). Given this imbalance, receiver operating characteristic (ROC) curves were not appropriate for model performance assessment because of their tendency to overestimate performance on imbalanced datasets [28]. Therefore, precision-recall curves, which are sensitive to data imbalances can expose performance gaps that are not visible in ROC curves [29], were generated (Figure 4). It was found that RNNs achieved high levels of precision, without significant performance reductions in recall.

V. DISCUSSION

Driver cognitive load classification in real-time can enable the implementation of tailored warnings and adaptive interfaces (e.g., information content filtering). Further, real-time assessment of cognitive load can also improve automated vehicle design by making it possible to intelligently transfer control to the driver from the vehicle.



(a)



(b)

Figure 4. PR curves for RNN with (a) within-driver partitioning, (b) across-drivers partitioning

Given that most of the relevant previous research trained and tested their models on the same participants (i.e., within-driver partitioning), we also compared each our model's performance when training and testing was performed this way and when training and testing was performed on data from distinct participants (i.e., across-drivers training). All models performed better for within-driver partitioning by an average accuracy difference of 3.0% (standard deviation: 2.6%). This difference may or may not be considered practically significant based on the application but can inform designers of vehicle systems that may utilize machine learning models to classify driver cognitive load.

The RNN model with a 2-layer LSTM architecture achieved a mean accuracy of 0.88 for within-driver partitioning and 0.86 for across-drivers partitioning. These accuracies are comparable to the model accuracies obtained in earlier research, which are presented in Table I. LDA had the lowest mean accuracy levels 0.65 for both within-driver and across-drivers partitioning. Although the remaining models (e.g., SVM, ANN, DTC) also had relatively large mean accuracy levels, their performance was more variable (i.e., a larger range of accuracies) compared to that of RNN. Thus, RNN was more robust.

A potential reason for RNN's superior performance is that RNN considers temporal information, while other widely-used machine learning tested in this paper did not (i.e., they had no memory of a signal's previous values) and thus are highly vulnerable to individual differences across drivers. Although RNN outperformed the other methods explored, it is a "black-box" method and not as interpretable as more traditional methods such as SVM or DTC [30]. In general, explainable and interpretable algorithms may find higher levels of acceptance from vehicle manufacturers and system designers who may be concerned about interpretability of the methods. Future research should explore other methods, such as ensemble algorithms [31], which may provide comparable accuracy levels to RNN but not at the cost of interpretability.

As stated earlier, He et al. [12] also utilized the same dataset used in this paper to explore a number of machine learning models through within-driver partitioning. SVM resulted in the best performance with an accuracy of 0.79 for classifying three levels of cognitive load (i.e., no-task, 1-back, 2-back). Not surprisingly, the two-class problem pursued in the current paper is easier than the three-class problem and SVM for within-driver partitioning led to an average accuracy level of 0.87.

A limitation of our study is the small sample size used to train and test our models. We had 33 participants -- a sample size higher than most relevant research (see Table I) but is still limited – and focused only on younger drivers. Future research should investigate these models with a larger and more representative sample (e.g., with different age groups and driving experience). Further, although using a validated secondary task to impose cognitive load can benefit model fitting, future work is needed to verify the effectiveness of the models in identifying high cognitive load imposed by the tasks that are common in drivers' own vehicles (e.g., voice control).

VI. Conclusion

In this paper, we explored the performance of several more traditional machine learning models (SVM, ANN, KNN, DTC, NB, LDA) as well as an emergent one particularly appropriate for temporal data, the RNN, in detecting whether drivers are engaged in a cognitive secondary task. All models we explored in this study performed better with within-driver partitioning compared to across-drivers partitioning. RNN reached the highest accuracy with mean accuracies of 88.1% and 85.6% with within-driver and across-drivers partitioning, respectively. Future research with larger sample size and based on real in-vehicle tasks is needed to validate the machine learning models explored in our study.

Acknowledgment

The funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada.

References

[1] J. L. Harbluk, Y. I. Noy, P. L. Trbovich, and M. Eizenman, "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance," *Accident Analysis & Prevention,* vol. 39, no. 2, pp. 372-379, 2007.

[2] M. A. Recarte and L. M. Nunes, "Effects of verbal and spatial-imagery tasks on eye fixations while driving," *Journal of Experimental Psychology: Applied,* vol. 6, no. 1, p. 31-43, 2000.

[3] S. Rogers, C.-N. Fiechter, and C. Thompson, "Adaptive user interfaces for automotive environments," *IEEE Intelligent Vehicles Symposium*, Dearborn, MI, USA, pp. 662-667, 2000.

[4] D. L. Strayer, J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman, "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,* vol. 71, no. 2, pp. 93-110, 2017.

[5] M. Johns, S. Sibi, and W. Ju, "Effect of cognitive load in autonomous vehicles on driver performance during transfer of control," The *6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Seattle, WA, USA, pp. 1-4, 2014.

[6] B. Donmez, L. N. Boyle, and J. D. Lee, "Mitigating driver distraction with retrospective and concurrent feedback," *Accident Analysis & Prevention,* vol. 40, no. 2, pp. 776-786, 2008.

[7] B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek, "Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers," *Transportation Research Record,* vol. 2138, no. 1, pp. 6-12, 2009.

[8] B. Mehler, B. Reimer, and J. F. Coughlin, "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,* vol. 54, no. 3, pp. 396-412, 2012.

[9] P. Bashivan, I. Rish, and S. Heisig, "Mental state recognition via wearable EEG," *arXiv preprint arXiv:1602.00985,* 2015.

[10] J. Gaskin, J. Jenkins, T. Meservy, J. Steffen, and K. Payne, "Using wearable devices for non-invasive, inexpensive physiological data collection," The *50th Hawaii International Conference on System Sciences*, Hawaii, USA, pp. 597-605, 2017.

[11] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive analysis of working memory load from EEG, by a deep recurrent neural network," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 2576-2580, 2018.

[12] D. He, M. Risteska, B. Donmez, and K. Chen, "Driver mental workload classification through the use of physiological data," *Digital Signal Processing and Machine Learning for Interactive Systems Developers*, P. Eslambolchilar, A. Komninos, and M. Dunlop Eds. Association for Computing Machinery, pp.409-429, 2021.

[13] J. Kohlmorgen *et al.*, "Improving human performance in a real operating environment through real-time mental workload detection," *Toward Brain-Computer Interfacing*, G. Dornhege, J. d. R. Millan, T. Hinterberger, D. J. McFarland, and K.-R. Mller Eds. Cambridge, Massachusetts: MIT Press, pp. 409-422, 2007.

[14] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Transactions on Intelligent Transportation Systems,* vol. 8, pp. 340-350, 2007.

[15] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data," The *32nd Annual ACM Conference on Human Factors in Computing Systems*, Toronto, Canada, pp. 4057-4066, 2014.

[16] Y. K. Wang, T. P. Jung, and C. T. Lin, "EEG-based attention tracking during distracted driving," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 23, no. 6, pp. 1085-1094, 2015.

[17] D. He, C. C. Liu, B. Donmez, and K. N. Plataniotis, "Assessing high cognitive load in drivers through electroencephalography," *Transportation Research Board 96th Annual Meeting*, Washington D.C., USA, no. 17-02615, 2017.

[18] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, "High cognitive load assessment in drivers through wireless Electroencephalography and the validation of a modified n-back task," *IEEE Transactions on Human-Machine Systems,* vol. 49, no. 4, pp. 362-371, 2019.

[19] C. T. Lin, R. C. Wu, S. F. Liang, W. H. Chao, Y. J. Chen, and T. P. Jung, "EEG-based drowsiness estimation for safety driving using independent component analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 52, pp. 2726-2738, 2005.

[20] K. Li, L. Jin, Y. Jiang, H. Xian, and L. Gao, "Effects of driver behavior style differences and individual differences on driver sleepiness detection," *Advances in Mechanical Engineering,* vol. 7, no. 4: 1687814015578354, 2015.

[21] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[22] F. Chollet, "Keras: The python deep learning library," *Astrophysics Source Code Library,* 2018.

[23] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *International Conference on Machine Learning*, Atlanta, USA, pp. 1310-1318, 2013.

[24] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019,* 2015.

[25] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 28, no. 10, pp. 2222-2232, 2016.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[27] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," The *2018 Conference on Computational Linguistics and Speech Processing*, Hsinchu, Taiwan, pp. 214-223, 2018.

[28] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," The *23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, pp. 233-240, 2006.

[29] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One,* vol. 10, no. 3: e0118432, 2015.

[30] D. J. Livingstone, D. T. Manallack, and I. V. Tetko, "Data modelling with neural networks: Advantages and limitations," *Journal of Computer-Aided Molecular Design,* vol. 11, no. 2, pp. 135-142, 1997.

[31] K. W. De Bock and D. Van den Poel, "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models," *Expert Systems with Applications,* vol. 39, no. 8, pp. 6816-6826, 2012.