Cognitive Workload Estimation in Conditionally Automated Vehicles Using Transformer Network Based on Physiological Signals

Ange Wang
Intelligent Transportation Thrust, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: awang324@connect.hkust-gz.edu.cn

Jiyao Wang
Intelligent Transportation Thrust, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: jwang297@connect.hkust-gz.edu.cn

Wenxin Shi
Robotics and Autonomous Systems Thrust, Systems Hub,
The Hong Kong University of Science and Technology (Guangzhou), China
Email: wshi559@connect.hkust-gz.edu.cn

Dengbo He*
Assistant Professor
Intelligent Transportation Thrust, Systems Hub
The Hong Kong University of Science and Technology (Guangzhou), China
HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China
Email: dengbohe@hkust-gz.edu.cn

Word Count: 6,696 words + 3 table (250 words per table) = 7,446 words

Submitted [April, 10, 2024]

ABSTRACT

Though driving automation is promised to improve driving safety, drivers are still required to be ready to take the control back in conditionally automated vehicles, which are defined as SAE L3 vehicles by the Society of Automotive Engineers. Thus, drivers' states can still affect driving safety in SAE L3 vehicles. Previous research found that high cognitive load may impair drivers' takeover performance. Thus, it is still necessary to estimate drivers' cognitive load in SAE L3 vehicles. However, existing driver cognitive load estimation algorithms mostly focused on vehicles with a lower level of driving automation (e.g., SAE L0), which may not apply to driver state estimation in SAE L3 vehicles, given that drivers' responsibilities are different, and several commonly used measures (e.g., driving performance) are unavailable when drivers are not continuously controlling the vehicle. Further, previous driver cognitive load estimation algorithms rarely considered the temporal information in the input features. Thus, we proposed a deep-learning algorithm to estimate driver cognitive load in SAE L3 vehicles, which integrated multiple physiological features (i.e., electrocardiogram, electrodermal activity, respiration) and considered the temporal correlation of the data using a Transformer-encoder-based Network. The performance of our algorithm was compared to baseline models on an open dataset. Results showed that our algorithm outperformed baseline models and achieved an accuracy of 94.4% using within-subject data partition (proportionally splitting data from the same subject into the training and testing datasets) and an accuracy of 89% using across-subjects data partition (dividing the training and testing datasets based on individual subjects).

Keywords: Cognitive Load, Conditional Automation, Deep Learning, Transformer Encoder, Physiological Signals

INTRODUCTION

The driving task is more than a visual-manual task but also a cognitive task. For instance, while driving, individuals are required to monitor their surroundings for traffic conditions (e.g., identifying potential hazards), select the appropriate lane, and keep safe following distance. At the same time, during driving, drivers may also engage in non-driving-related tasks (NDRTs), which also demand some level of cognitive recourse. The fluctuations in cognitive load in drivers has been found to be closely associated with driving safety and high levels of cognitive load among drivers can deteriorate driving performance (*1*). For example, an excessive cognitive load state may lead to delayed responses to emergency events (*2*), a diminished visual search scope (resulting in a visual tunnel effect) (*3*), and a reduced ability to anticipate hazards (*4*). Thus, monitoring drivers' cognitive load is critical in ensuring driving safety.

Although driving automation has brought new possibilities to improve safety (*5*) by removing human drivers from the vehicle control loop, at this stage, drivers are still expected to be responsible for driving safety in case of emergency. According to a taxonomy proposed by the American Society of Automotive Engineers (SAE) (*6*), the most advanced driving automation available in the market is still Level 3, in which the driving automation can execute driving maneuvers within specific traffic conditions, but the driver remains legally responsible for driving safety (*7*). Thus, the cognitive load of drivers may still affect driving safety in SAE Level 3 vehicles. For example, previous research found that high cognitive load is associated with increased reaction times when assuming vehicle control during assisted driving (*8*). While the physical workload of drivers may decrease with driving automation, it is conceivable that drivers' cognitive load may increase (*9*). Further, numerous studies indicate that drivers are inclined to engage more in non-driving-related tasks while using driving automation (*10*), potentially leading to higher cognitive load. On the other hand, in vehicles with SAE level 3 automation, by definition, drivers no longer need to monitor the automation all the time and they can engage in NDRTs such as video-streaming and internet browsing in vehicles. Hence, monitoring drivers' cognitive load is essential to driving safety in SAE Level-3 vehicles.

Many cognitive load estimation algorithms have been proposed in previous research, mainly in non-automated vehicles (i.e., SAE Level 0). For instance, in a driving simulator experiment, He et al. (*11*) investigated the classification accuracy of diverse machine learning models using different combinations of driver characteristics to classify three types of cognitive load imposed by a modified n-back task (*12*, *13*). They found that a random forest algorithm incorporating electrodermal (EDA) and heart rate (HR) can achieve a discriminative accuracy of 72.6% in evaluating driver cognitive load, outperforming models that solely relied on individual physiological measures. Another study by Solovey et al. (*14*) combined physiological data (i.e., HR and EDA) and driving data (vehicle speed, acceleration, etc.) using a naive Bayes model. They achieved an accuracy of 89% in differentiating high (n-back task) and low (no task) cognitive load using a real-world driving dataset. Other related driving studies that focused on cognitive estimation are summarized in **TABLE 1**.

TABLE 1 Driving studies that estimated drivers' cognitive load in recent five years.

Study	**Physiological Signal**	**Classification Problem**	**Data Partition**	**Automation Level**	**Model**	**Accuracy**
He et al. (*11*)	ECG, GSR	Three classes: no task, 1-back, 2-back	Within-subject	L0	KNN, SVM, FNN, RNN, RF	73%
Yang et al. (*15*)	EEG	Three classes: no task, visual-recall task, visual-auditory-recall task	Within-subject	L0	ARecNet	95%
Rahman et al. (*16*)	ECG	Two classes: no task, 1-back	Within-subject	L0	LR, SVM, LDA, FNN	87%

Meteier et al. (*17*)	ECG, EDA, RESP	Two classes: no task, counting back task	Within-subject	L3	MLP, RF, SVC	90% (three signals fusion)
Kumar et al. (*18*)	ECG, GSR, EEG	Two classes: cognitive secondary task, no task	Within-subject/Across-subjects	L0	RNN, SVM, FNN, KNN, DTC, NB, LDA	Within-subject: 88% Across-subjects: 86%
Angkan et al. (*19*)	ECG, EDA, eye -tracking data	Two classes: reading task, arithmetic tasks	Not Mentioned	L0	VGG Net	75%
Le et al. (*20*)	Hemodynamic data	Three classes: no task, 1-back, 2-back	Across-subjects	L0	BT	90%
Barua et al. (*21*)	ECG, EEG, EOG, GSR, RESP	Three classes: no task, 1-back, 2-back	Not mentioned	L0	RF	79%
Prabhakar et al. (*22*)	eye -tracking data	Two classes: no task, auditory n-back or visual n-back or auditory arithmetic task	Not mentioned	L0	NN	75%

Notes: ECG – Electrocardiogram; GSR - Galvanic Skin Response; EEG – Electroencephalogram; RESP – Respiration; EOG - Electrooculography; KNN - k-Nearest Neighbors; SVM - Support Vector Machine; FNN - Feedforward Neural Network; RNN - Recurrent Neural Network; BT - Bagged Trees; RF - Random Forest; LR - Logistic Regression; LDA - Linear Discriminant Analysis; MLP - Multilayer Perceptron; SVC - Support Vector Classifier; DTC - Decision Tree Classifier; NB - Naive Bayes; ARecNet - Attention-enabled Recognition Network; NN - Neural Network

However, it is worth mentioning that the aforementioned driver state monitoring algorithms primarily focused on the assessment of cognitive load in non-automated vehicles. These algorithms may not be applicable for driver state estimation in SAE Level 3 vehicles for several reasons. First, some features that have been commonly used in non-automated vehicles may not be available in SAE Level 3 vehicles. For example, as the drivers are relieved from vehicle controlling tasks and traffic monitoring tasks most of the time, the driving performance measures and the eye-tracking measures become invalid in SAE Level 3 vehicles. Second, even algorithms that solely used physiological measures may need to be scrutinized in SAE Level 3 vehicles. This is because regardless of how the cognitive load was imposed in previous studies, driving a non-automated vehicle still claims some level of visual-manual attentional resources, which may introduce unexpected fluctuation in physiological signals. Thus, it is necessary to develop algorithms to estimate cognitive load in SAE Level 3 vehicles. However, only a few studies have considered cognitive load estimation in vehicles with driving automation. To the best of our knowledge, the only one that can be identified is by Meteier et al.(*17*), which used traditional machine learning methods (i.e. SVC, RF, MLP) to recognize the driver's cognitive load in SAE Level 3 vehicles.

On the other hand, most of the previous driver cognitive detection methods ignored the temporal information in the features. This is especially a shortcoming when physiological measures are used, given that the temporal dependence of physiological data contains rich information (*23*, *24*). To date, the only study that considered the temporal information for driver cognitive load detection used the RNN-based models (*18*), which, by nature, follow the Markov property and assume each to be dependent only on the previously observed state. Thus, the RNN may not handle complex temporal interaction among multiple time windows (*25*) and its performance in extracting information in physiological signals is limited.

The transformer architecture (*26*) proposed in the naturalistic language processing domain provides a powerful tool to extract temporal information in temporal-dependent data, primarily attributed to the distinctive multi-head attention mechanism. In other words, transformers can simultaneously "attend" to and learn from the entire sequence (*27*). In recent years, the transformer architecture has been successfully adopted to estimate some driver states. For example, Gao et al.(*28*) estimated driver vigilance based on frequency-domain EEG features, and the transformer network achieved an accuracy of 78%. Another study by Wang et al. (*29*) introduced the gated linear unit in the attention and feed-forward networks sub-blocks, presenting an improved transformer architecture for one-dimensional feature vector classification. The results demonstrate that the enhanced transformer approach can achieve an accuracy of 86.97% in predicting driver fatigue state using EEG signals. Yang et al. (*30*) developed a Weakly Supervised Contrastive (W-SupCon) learning framework based on vision transformers, wherein distracted behaviors are quantified by computing the distances between the representations of distracted driving and a set of normal driving representations.

However, the transformer architecture has not been adopted for driver cognitive state estimation in vehicles with SAE Level 3 driving automation. In this paper, we propose a data fusion model to estimate cognitive load in SAE Level 3 vehicles, using physiological measures including EDA, ECG, and RESP. A multi-layer transformer model was used to capture signal variations and improve the accuracy of cognitive load estimation. Its performance was compared to other commonly used baseline models on a publicly available dataset by Meteier et al.(*31*).

DATASET AND PREPROCESSING

Data Source

In this study, we trained and tested our model on the dataset published by Meteier et al.(*31*). In the dataset, 90 drivers (mean age: 24.15; standard deviation of age: 5.95 years) drove an SAE Level 3 automated vehicle in a driving simulator. Half of the participants were required to perform a cognitive task called the oral digit span counting task, which involved verbally counting backward from 3,645 by a step of 2. This artificial task simulates a continuous cognitive task, resembling a conversation between individuals over the phone or between passengers in a vehicle. The post-drive questionnaire using the NASA-Task Load Index (NASA-TLX) indicated that the cognitive task effectively induced higher levels of cognitive load among 87 drivers. In the study, the physiological signals were recorded using the Biopac MP36 hardware at a sampling rate of 1,000 Hz. The EDA and ECG data were collected from participants using the SS57LA and SS2LB lead sets (Biopac) equipped with disposable Ag/AgCl pre-gelled electrodes (EL507 and EL503, Biopac). Additionally, respiratory activity was captured by identifying chest expansion and contraction using the SS5LB respiratory effort transducer (Biopac).

Data Preprocessing

In general, the preprocessing steps for ECG, EDA, and RESP signals include noise elimination, time-domain and frequency-domain feature extraction, and normalization. All procedures were carried out using the Neurokit library in Python (*32*). First, noise elimination was conducted to remove the noise in the data collection process. Specifically, all signals were down-sampled to 50 Hz to reduce the computation cost. For EDA, a low-pass filter with a cutoff frequency of 5Hz was applied; for ECG and RESP, band-pass filters with a cutoff frequency between 3Hz and 45Hz and between 0.1 Hz and 0.35 Hz were applied, respectively (*17*).

Next, the features from the denoised signals were extracted. For ECG, the R peaks were first identified using the local maxima in the QRS complexes, which represent the process of ventricular depolarization (the contraction of the ventricles) and consist of the Q wave, R wave, and S wave (*33*). followed by a calculation of HR and HR variability (HRV). As for HRV, we calculated low-frequency power (LF, from 0.04Hz to 0.15Hz) and high-frequency power (HF, from 0.15Hz to 0.4Hz) using a short-time Fourier Transform (*34*). As for EDA, we extracted its mean tonic level, which relates to the slower-acting components and background characteristics of the signal (i.e., the overall level, slow climbing, and slow declinations over time). As for RESP, the peak and valley of respiration were identified using the *findpeak* function in Python. Then, the average duration of each inhale (from a peak to its closest next valley) and exhale (from a valley to its closest next peak) were calculated. To satisfy the minimum requirement for accurate estimation of HRV, for all feature extraction, we adopted a time window of 60 seconds following previous research (*35–37*) to balance the feature extraction accuracy and timeliness of driver state estimation. A step size of 2 seconds (i.e., an overlap of 58 seconds) was adopted in our study so that the algorithm can estimate drivers' cognitive state every 2 seconds. Then, to eliminate the impact of individual differences in physiological signals, we employed the min-max normalization approach, where each feature value of a participant is subtracted by the minimum of that participant and then divided by the difference between the maximum and minimum values of that participant.

Feature Selection

Excessive features will increase computational complexity, leading to data redundancy (*38*). Thus, a feature selection procedure was adopted before model training. The Principal Component Analysis (PCA) was widely adopted in previous research. However, PCA primarily focuses on data variance and is an unsupervised method. Thus, the PCA may overlook the relationship between the selected features and the model output (*39*). To overcome this limitation, we adopted a LASSO method by Tibshirani (*40*) for variable selection. In the LASSO method, by introducing penalty terms in the model estimation and continuously shrinking the coefficients, unnecessary features in the model can be reduced, which can effectively avoid the issues of model overfitting (*40*). The definition of the LASSO method is as follows. For data $(x_i, y_i), i = 1,2,\ldots,n,$ where $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T$ represents the predictor variables and y_i represents the dependent variable, it is assumed that y_i is standardized, i.e., $\sum_i \frac{y_i}{n} = 0$ and $\sum_i \frac{{y_i}^2}{n} = 1$, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$. Then, the LASSO estimate can be defined as follows:

$$\hat{\beta}^{LASSO} = \arg\min\left\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2, subject\ to\ \sum_j\left|\beta_j\right| \leq \lambda\right\}, \quad (1)$$

where $\lambda \geq 0$ represents the penalty function employed to control the degree of compression; $\hat{\beta}_j^0$ denotes the ordinary least squares penalized estimation and p is the number of features. Let $\lambda_0 = \sum\left|\hat{\beta}_j^0\right|$, by selecting $\lambda \leq \lambda_0$, it is possible to induce a shrinkage effect on certain regression coefficients, causing them to approach zero.

MODEL TRAINING AND EVALUATION

The overall framework of our work is illustrated in **FIGURE 1**.

Problem Formulation

The recognition of driver cognitive load is a supervised classification task, wherein the workload levels are utilized as categorical labels. Then, the dataset can be described as follows:

$$D: \{T_m^l, y_m\}, m = 1,2,\ldots,N, \quad (2)$$

where T^l is the multimodal temporal sequences with the size of l; y is the cognitive workload of the driver, with two levels: low ($y = 0$), and high ($y = 1$); N is the total number of samples. For each sample,

$$T^l = [f_1, f_2, \ldots, f_l], \quad (3)$$

where f is the vector of features at each time step, in which f is consisted of extracted EDA, ECG, and RESP features. The research objective of the models is to generate the workload level $\hat{y}_m$ based on the temporal series information T_m^l, so that:

$$\hat{y}_m = \arg\max_{y_m \in \zeta} p\left(\left(y_m \middle| T_m^l\right)\right) \quad (4)$$

where $\zeta = \{0,1\}$ represents the label of cognitive workload levels.

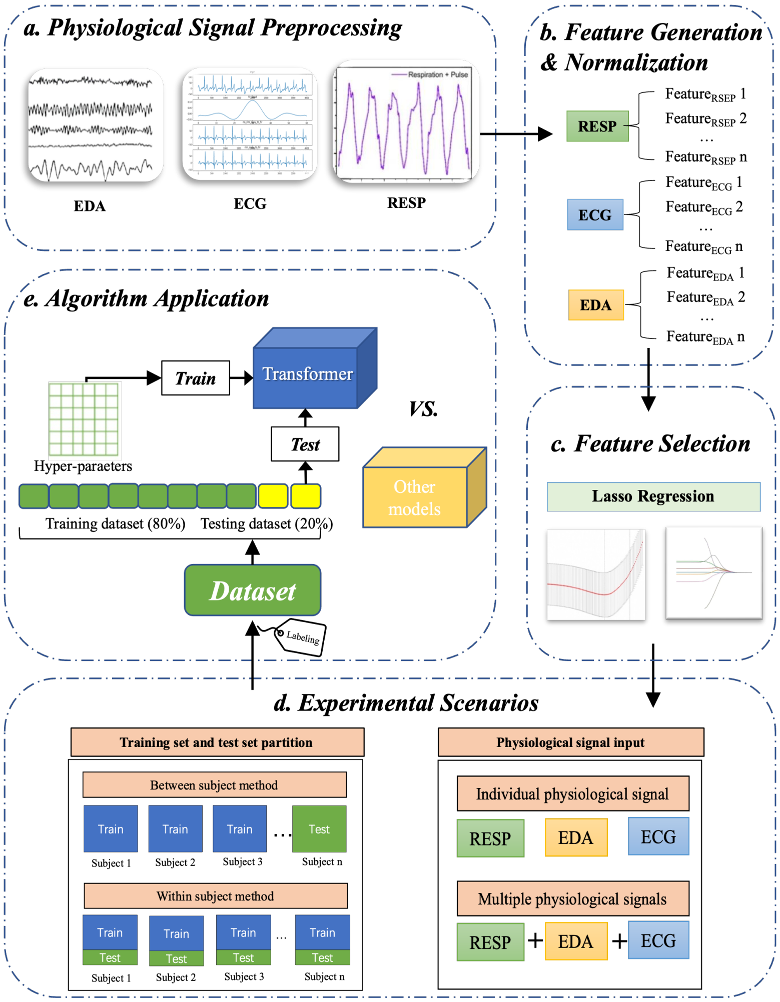


FIGURE 1 The framework of work

Transformer-encoder-based Network (TEN)

A typical Transformer architecture comprises four essential components: 1) positional encoding, 2) attention mechanism, 3) encoder block, and 4) decoder block. In this study, it is enough to extract information from the input sequence without decoding and generating an output sequence; hence, a Transformer-Encoder-based Network (TEN) was used. Firstly, the embedding layer maps the input features into an embedding dimension space, then the positional encoding layer encodes positional information within the sequence and the Transformer encoder layers encode and model the sequence. Finally, a fully connected layer is utilized to map the encoded sequence to the output dimension for binary classification prediction **(FIGURE 2a)**.

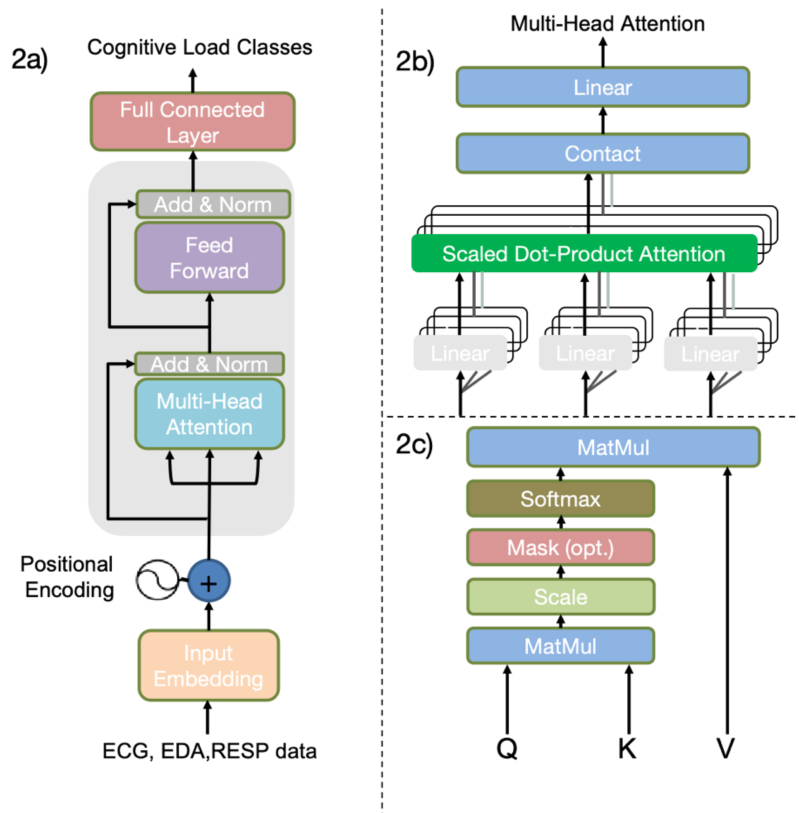


FIGURE 2 The Transformer-encoder-based Network: a) the framework of TEN; b) multi-head attention module in TEN; c) linear layer in TEN, where Q, K, and V are Query, Key, and Value from the from the attention mechanism of the Transformer.

Position Encoding

The Position Encoding component is used to incorporate positional information for each position within the input sequence. Due to the absence of recursion or convolution, the Transformer model itself does not directly consider the positional information within the sequence. Therefore, we embedded the position encoding in the input to keep the information of relative sample positions, leading to a refined representation. This is achieved through the linear transformation of the input sequence to a dimension of d^{model}, followed by the incorporation of Position Encoding (*PE)* originating from Equations (5) and (6).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d^{model}}}}\right), \tag{5}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d^{model}}}}\right), \tag{6}$$

where pos denotes the position of the sample (1 to d^{model}) and i means the index of the linearly transformed input sequence (1 to 7, total number of variables) at position pos, while $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$ represent the $2ith$ and $(2i+1)th$ components, respectively, of the encoding vector at position pos.

Multi-head Attention

The transformer forecasting model utilizes the multi-head attention mechanism to capture long-term correlations within time series data. The attention mechanism is represented by Equation (7), which is independently applied, then concatenated, and subsequently fed into a linear layer to generate sequences (**FIGURE 2b and 2c**).

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V, \tag{7}$$

Specifically, the weighted representation, $Attention\ (Q,K,V)$, is obtained through simultaneous calculation of matrices Q, K and V, where Q is the ‘‘query” (i.e., query matrix) that computes correlation scores (attention weights) with other positions to determine the level of attention each position pays to others); K is the ‘‘key” (i.e., key matrix), which is used to compute correlation scores (attention weights) of the query with other positions; and V is the ‘‘value” (i.e., value matrix), which carries information forwarded to the next layer, with each position retrieving information from the value matrix based on the attention weights. The multi-head attention scales values based on the relationship between key and query, which in essence, establishes connections between multiple instances of self-attention. Additionally, a residual connection between the input and output was incorporated to facilitate gradient propagation in the frame flow.

$$\text{Multi-Head}\ (Q,K,V) = contact(head_1, head_2, \ldots, head_h)W^0, \tag{8}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{9}$$

The given expression involves parameter matrices $W_i^{Q,K,V} \in \mathbb{R}^{d_{model} \times d_{q,k,v}}$ and $W^0 \in \mathbb{R}^{hd_v \times d_{model}}$. The utilization of multiple headers within the encoding layer enables simultaneous attention to diverse information sources in distinct representation subspaces at varying positions. This approach facilitates the extraction of contextually relevant features from different regions.

Baseline Models

In this work, we also compared the performance of the transformer network to other traditional classifiers and deep learning models that showed superior performance in previous studies. These baseline models include k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Light Gradient Boosting Machine (LightGBM), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM).

KNN

The KNN is a non-parametric supervised learning algorithm (*41*), in which the hyperparameter, the number of neighbors (N_neighbors), determines the count of training data samples that participate in predicting a given test data point.

SVM

The SVM is a supervised learning algorithm (*42*), in which the regularization parameter "C" was optimized to achieve a balance between model complexity and the tolerance of training errors, with a smaller C resulting in larger margins but higher tolerance on training errors. At the same time, the Kernel function in SVM transforms a non-linear decision surface to a linear equation in a space with a higher number of dimensions.

LightGBM

The LightGBM is a machine learning algorithm based on gradient boosting (*43*), which is known for its rapid training speed and high performance. The performance of LightGBM can be affected by the number of decision trees (N_estimators), the learning rate, the max depth of each trained tree (Max_depth), and the maximum number of leaves per tree (Num_leaves).

CNN

The CNN model is one of the most popular and widely used models in the field of deep learning in recent years. It adopts the approach of local connections and shared weights, directly extracting meaningful representations from raw data through alternating convolutional and pooling layers. This enables the automatic extraction of local features from the data and establishes dense and comprehensive feature vectors (*44*). In this study, the CNN consists of two convolutional layers followed by a fully connected (linear) layer. The first convolutional layer takes input from seven channels and employs 32 filters of size 3 with a stride of 1. Subsequently, the first max-pooling layer performs 1D max-pooling with a size of 2 and a stride of 2. The output of the first max-pooling layer (32 channels) serves as the input to the second convolutional layer, which utilizes 64 filters of size 3 with a stride of 1. Following this, the second max-pooling layer performs 1D max-pooling with a size of 2 and a stride of 2. The output from the last pooling layer is flattened into a 1D feature vector. This flattened feature vector is then fed into the fully connected (linear) layer, which maps it to the output dimension in **Equation** (2) for binary classification purposes.

LSTM

The LSTM network is an improved type of Recurrent Neural Network (RNN). By including the addition of extra forget gates (*45*, *46*), the improved LSTM can network address the issue of "vanishing gradients" during model training and enable the learning of long and short-term dependencies in time series data. In this study, the LSTM model comprises a single LSTM layer followed by a fully connected (linear) layer. The LSTM layer takes input sequences of dimensionality 7 (corresponding to the number of variables) and extracts the output at the last time step. Subsequently, this output is forwarded to the fully connected layer to obtain the ultimate classification outcome.

Training and testing sets partition

The data partition methods have a large impact on the model performance. Specifically, two common approaches are employed for dataset partitioning, namely, the between-subjects data partition and the within-subject data partition (*18*). The former involves dividing the training and testing datasets based on individual subjects (e.g., data from subjects 1 to 72 are used for training; while data from subjects 73 to 90 are used for testing). In this case, the testing dataset does not include any data from individuals in the training dataset. In contrast, the within-subject data partition proportionally splits data from the same subject into the training and testing datasets (e.g., taking 80% of data from each subject for training and the remaining 20% for testing). In practice, the labeled historical data of a new driver is unknown. Thus, the between-subjects data partition is practically more meaningful. However, the model performance usually drops dramatically from the within-subject data partition to the between-subjects data partition. To further validate the performance of our proposed transformer-based algorithm, we compared the model performance of the TEN and baseline models on both data partitions based on the average performance on a 10-fold cross-validation.

Evaluation Metric

To assess the performance of the transformer network, the following four metrics are used:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

$$Precision = \frac{TP}{TP + FP} \qquad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (13)$$

where true positive (*TP*) and true negative (*TN*) are the count of correctly predicted positive instances and negative instances, respectively. False positive (*FP*) and false negative (*FN*) are the count incorrectly predicted negative and positive instances, respectively.

RESULTS

In this segment, we compared the model performance on the selected dataset. All models were trained and tested on an Apple MacBook Pro (16 in., 2021) equipped with a 16-core GPU and 16 GB DDR4 RAM running at 3200 MHz. The deep learning framework, Torch 2.0.1 based on Python 3.9, was used.

LASSO Regression

In total, 56 distinct features were extracted from three physiological signals (i.e., 10 derived from EDA, 27 from ECG, and 19 from RESP). Through the utilization of the LASSO regression (optimal learning rate $\alpha = 0.01$), seven pertinent physiological features with non-zero coefficients were identified and are presented in **TABLE 2**.

TABLE 2 Variables incorporated into the model

Signal	Feature	Description
EDA	Mean filtered level	The mean value of filtered EDA signal
	Mean tonic level	The mean value of tonic EDA signal
ECG	Mean rate	The mean number of cardiac cycles per minute
	LF/HF	The ratio of low-frequency power (LF) to high-frequency power (HF)
RESP	Phase duration inspiration	The average inspiratory duration
	Phase duration expiration	The average expiratory duration
	Respiratory rate variability	The variation of respiratory durations

Optimal Hyper-Parameters and Performance Comparison

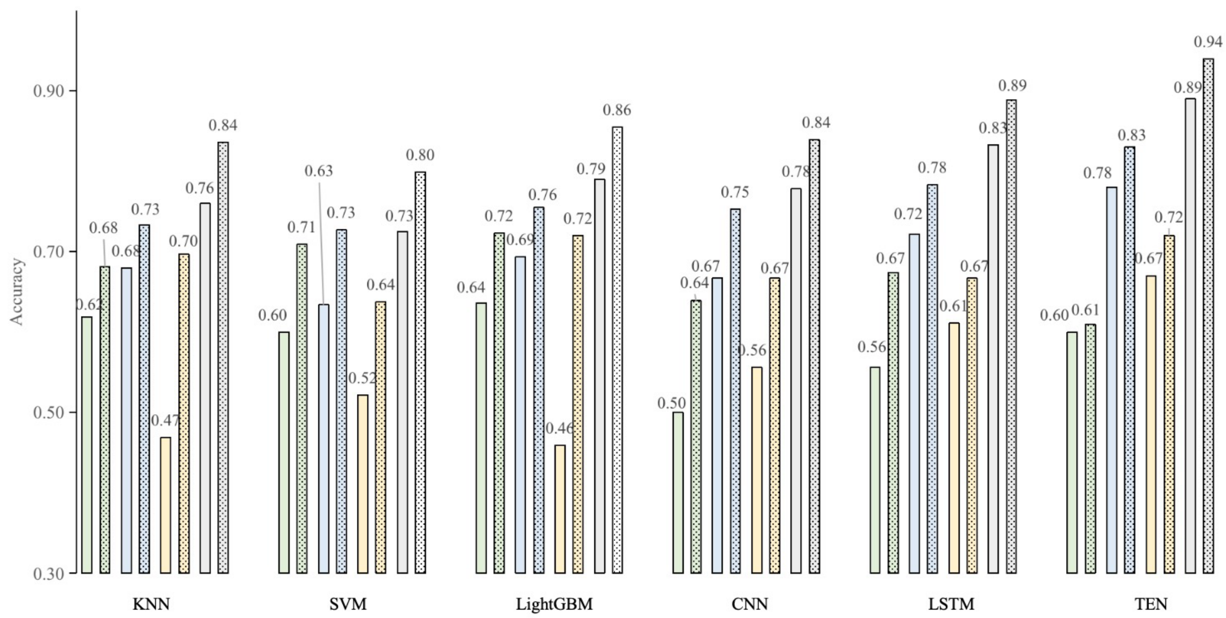
The candidate hyper-parameters of the models and the hyper-parameters leading to the best performance for all baseline models and the TEN are provided in **TABLE 3**. Then, **FIGURE 3** illustrates the average of the evaluation metrics (including accuracy, precision, recall, and F1 score) on the testing dataset with the within-subject and across-subjects cross-validations.

TABLE 3 Search ranges of hyperparameters

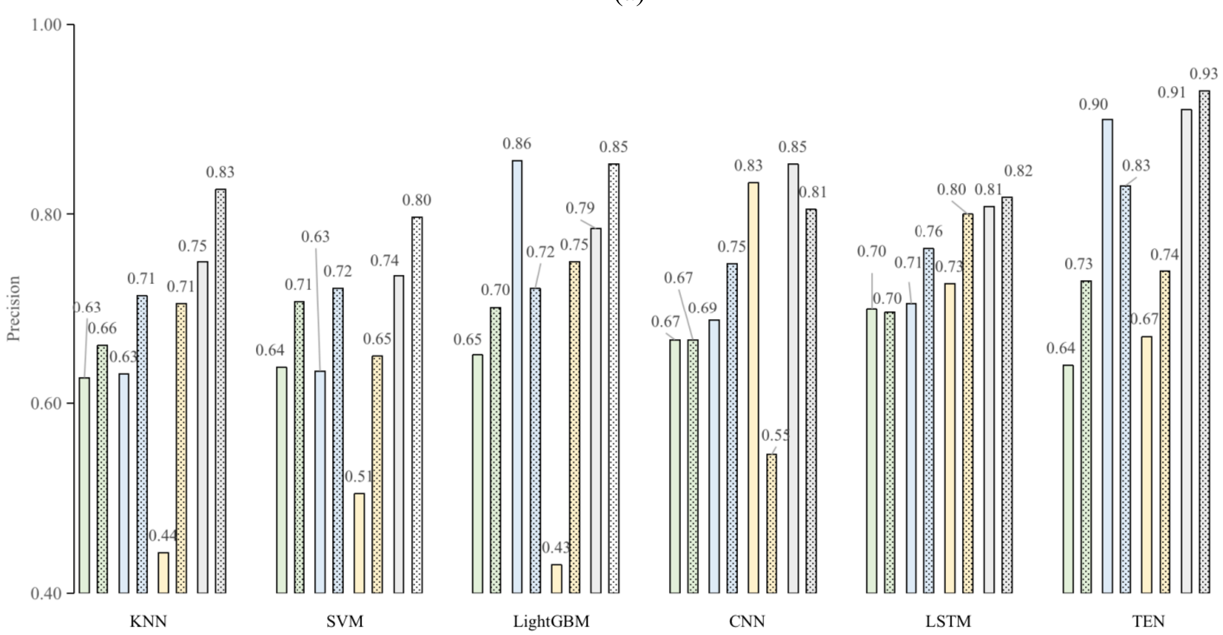
Model	Hyperparameter	Range
KNN	N_neighbors	**3**, 5, 7, 9
SVM	C	**0.1**, 1, 10
	Kernel	**linear**, RBF, polynomial
LightGBM	N_estimators	50, **100**, 200
	Learning rate	**0.01**, 0.1, 0.2
	Max_depth	-1, **10**, 20
	Num_leaves	**31**, 63, 127
CNN	Batch size	16, 32, **64,** 128, 256

	Learning rate	0.0001, **0.001**, 0.01, 0.1
LSTM	Number of layers	**1**, 2, 3, 4
	Number of hidden units	32, **64**, 128
	Batch size	16, **32**, 64, 128, 256
	Learning rate	0.0001, **0.001**, 0.01, 0.1
TEN	Dropout rate	0.1, **0.2**, 0.3, 0.4
	Batch size	16, **32**, 64, 128, 256
	Learning rate	0.0001, **0.001**, 0.01, 0.1
	Number of heads	2, 4, **8**, 16
	Number of hidden units	32, **64**, 128

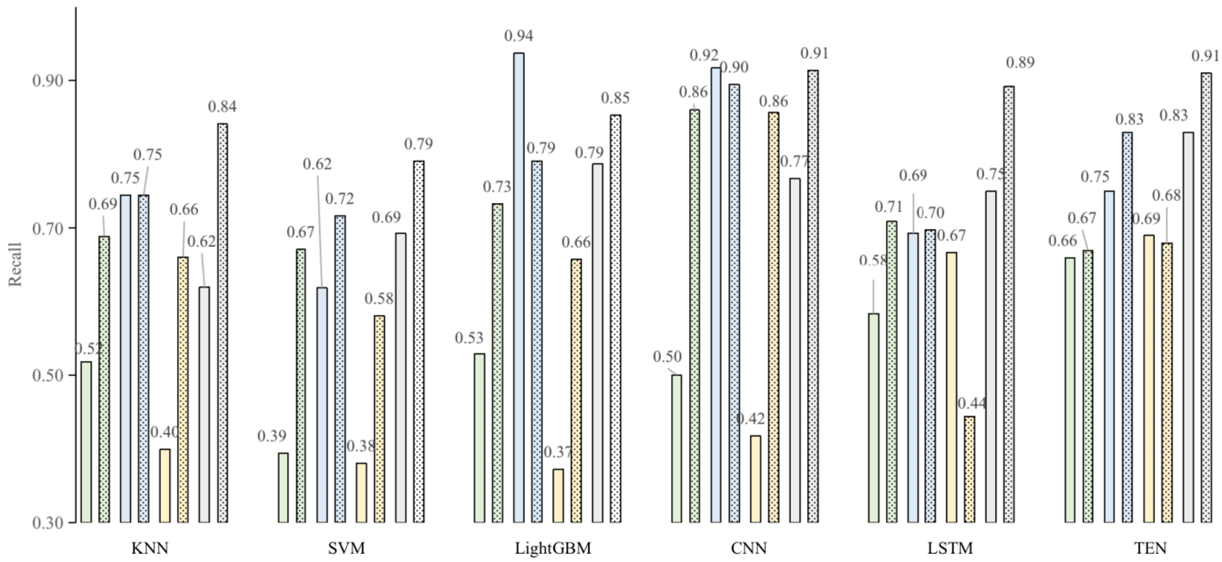
Notes: Bold values denote the optimal hyperparameters.



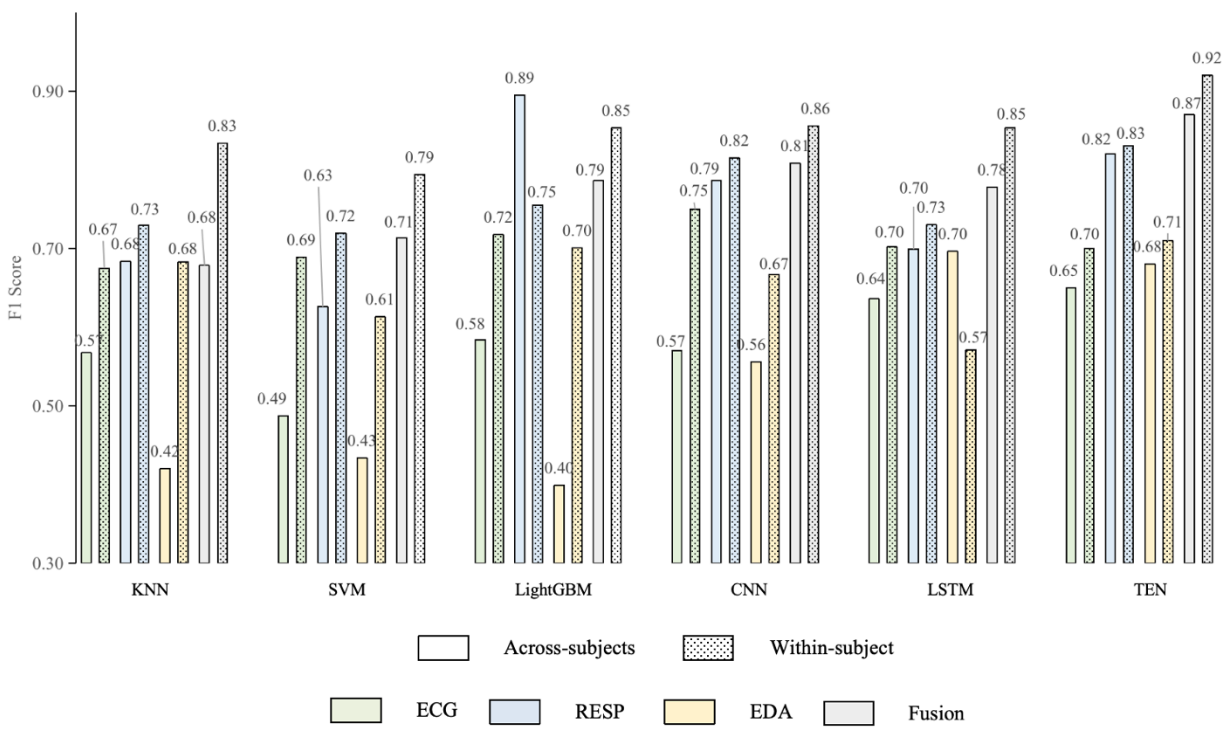
(a)



(b)



(c)



(d)

FIGURE 3 Comparison of model performance across baseline models and TEN using different physiological signals on two data partition methods: (a) Accuracy; (b) Precision; (c) Recall; (d) F1 score.

DISCUSSION

In this study, a Transformer-Encoder-based Network (TEN) was proposed to estimate drivers' cognitive load in SAE Level 3 vehicles based on physiological measures only. The performance of the model was compared to 5 baseline models (KNN, SVM, LightGBM, CNN, and LSTM) that reached satisfying performance in previous studies.

First of all, in traditional classifiers, LightGBM exhibited the best performance, yielding 79% and 86% accuracies for across-subjects and within-subject data partition. In contrast, the TEN proposed in our study outperforms all baseline models in terms of accuracy by at least 5% regardless of the data partition methods. The TEN also exceeded the ML models that have been adopted for cognitive load detection in SAE Level 3 vehicles by 4% in terms of accuracy (i.e., 94% with TEN versus 90% in (*17*), which used support vector classifier on the same dataset, with within-subject data partition and using the same physiological measures as ours). Given that the LSTM also reached 83% and 89% accuracy with across-subjects and within-subject data partition, it is likely that the high cognitive load may have led to some characteristics of the physiological signals in the temporal domain, which can hardly be captured by traditional machine learning models (e.g., RF) that used discrete samples. The temporal characteristics may also not fully satisfy the Markov process and can hardly be modeled using LSTM, which may explain better performance yielded by TEN.

Further, by further scrutinizing the precision and recall, it can be observed that the high accuracy of the TEN can be attributed to its superior performance in precision compared to other baseline models. In other words, the judgement that a driver is experiencing a high cognitive load is more likely to be true if such judgement is from TEN compared to that by other baseline models. Practically, it seems that the TEN is conservative in terms of detecting high cognitive load and whenever it marks a high cognitive load state, following countermeasures should better be taken to help alleviate cognitive load in drivers.

Agreeing with previous studies (*11*, *18*), we also found that the model performance would drop with across-subjects data partition compared to that with within-subject partition. This is not surprising, given the heterogeneity of the physiological signals across drivers (*47*, *48*). However, it should be noted that, the TEN showed good robustness to handle this heterogeneity. Specifically, from within-subject data partition to across-subjects data partition, the drop of the accuracy was the smallest (1% to 5%) with TEN as compared to other baseline models (i.e., KNN: from 5% to 23%, SVM: from 7% to 12%, LightGBM: from 7% to 26%, CNN: from 6% to 14%, and LSTM: from 6% to 11%). Again, it is possible that the TEN has learned some common temporal features of the high cognitive among drivers, which can hardly be captured by other models. If possible, future studies should try to extract these temporal features (e.g., through SHapley Additive exPlanations approach (*49*)), which may facilitate future optimization of the cognitive estimation algorithms and even help understand the physiological mechanism behind high cognitive load in drivers. Future work should also validate the effectiveness of these features and the generalizability of the models by training and testing our proposed model on multiple datasets when other datasets regarding drivers' cognitive levels in conditional automated vehicles become available.

Finally, again, as expected, the fusion of the algorithms can greatly boost the high cognitive load detection accuracies of the ML models. Previous research has repeatedly pointed out the importance of data fusion for the detection of drivers' states (e.g., (*50*, *51*) in fatigue detection, (*52*) in drowsiness detection, and (*15*) in cognitive load detection). Our research further indicates that the data fusion not just improves the driver state estimation accuracy, but also improves the generalizability of the model. Specifically, with single physiological measures, from within-subject data partition to across-subjects data partition, the average accuracy drop was 8.5% for ECG, 6.8% for RESP, and 13.8% for EDA. In contrast, when the fused measures were used, the average drop was 6.5%. These results suggest that fusing more measures can also be an effective approach to increase the robustness of the models to individual heterogeneity.

CONCLUSIONS

As one of the first studies that aim to estimate cognitive load among drivers in SAE Level 3 vehicles, using an open dataset, our study demonstrates the superior performance of TEN as compared to selected baseline models (SVM, KNN, LightGBM, CNN, and LSTM in our study and RF in study (*17*)) when only ECG, RESP, and EDA features were used. To evaluate the capability of the models in handling individual heterogeneity, all models were tested based on both within-subject data partition and across-subjects data partition. The results show that the TEN can reach an accuracy of 94% with within-subject data partition and 89% with across-subjects data partition when all ECG, RESP, and EDA features were used. However, it should be noted that the data used for this study was still collected in a driving simulator study and only one artificial cognitive task (i.e., oral digit span counting task) was used in the study. Future studies should check the model performance in detecting cognitive load imposed by other real driving (e.g., complex navigation information) or non-driving-related tasks (e.g., phone call task) in on-road studies. Further, the physiological measures in the dataset were collected through research-level devices, which is intrusive and may impede the application of the driver state detection methods. Future research should investigate the robustness of our algorithms when only signals collected from consumer-grade devices are available (e.g., study (*13*)).

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 52202425) and partially by the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), and the Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001).

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: **Ange Wang**: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Jiyao Wang**: Validation, Formal analysis, coding assistance. **Wenxin Shi**: Validation, Formal analysis. **Dengbo He**: Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Coughlin, J. F., B. Reimer, and B. Mehler. Monitoring, Managing, and Motivating Driver Safety and Well-Being. *IEEE Pervasive Computing*, Vol. 10, No. 3, 2011, pp. 14–21. https://doi.org/10.1109/MPRV.2011.54.
2. Harbluk, J. L., 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, 2007, pp. 372–379. https://doi.org/10.1016/j.aap.2006.08.013.
3. Recarte, M. A., 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, 2000, pp. 31–43. https://doi.org/10.1037/1076-898X.6.1.31.
4. Muhrer, E., and M. Vollrath. The Effect of Visual and Cognitive Distraction on Driver's Anticipation in a Simulated Car Following Scenario. *Transportation Research Part F: Psychology and Behaviour*, Vol. 14, No. 6, 2011, pp. 555–566. https://doi.org/10.1016/j.trf.2011.06.003.
5. Fernandes, P., and U. Nunes. Platooning With IVC-Enabled Autonomous Vehicles: Strategies to Mitigate Communication Delays, Improve Safety and Traffic Flow. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 1, 2012, pp. 91–106. https://doi.org/10.1109/TITS.2011.2179936.

6. J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International J3016_201806 (2018), , 2018.
7. Committee, O.-R. A. D. (ORAD). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. SAE International, 2014.
8. Du, N., F. Zhou, E. M. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang. Predicting Driver Takeover Performance in Conditionally Automated Driving. *Accident Analysis and Prevention*, Vol. 148, 2020. https://doi.org/10.1016/j.aap.2020.105748.
9. Stapel, J., F. A. Mullakkal-Babu, and R. Happee. Automated Driving Reduces Perceived Workload, but Monitoring Causes Higher Cognitive Load than Manual Driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, 2019, pp. 590–605. https://doi.org/10.1016/j.trf.2018.11.006.
10. He, D., and B. Donmez. Influence of Driving Experience on Distraction Engagement in Automated Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2673, No. 9, 2019, pp. 142–151. https://doi.org/10.1177/0361198119843476.
11. He, D., Z. Wang, E. B. Khalil, B. Donmez, G. Qiao, and S. Kumar. Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2676, No. 10, 2022, pp. 670–681. https://doi.org/10.1177/03611981221090937.
12. Mehler, B. Establishing the Sensitivity of Physiological Measures of Cognitive Load in the Driving Environment. 2012. https://doi.org/10.13140/2.1.2471.5521.
13. 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, 2019, pp. 362–371. https://doi.org/10.1109/THMS.2019.2917194.
14. Solovey, E. T., M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler. Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies. Presented at the CHI ’14: CHI Conference on Human Factors in Computing Systems, Toronto Ontario Canada, 2014.
15. Yang, H., J. Wu, Z. Hu, and C. Lv. Real-Time Driver Cognitive Workload Recognition: Attention-Enabled Learning with Multimodal Information Fusion. *IEEE Transactions on Industrial Electronics*, 2023, pp. 1–11. https://doi.org/10.1109/TIE.2023.3288182.
16. Rahman, H., M. Ahmed, S. Barua, and S. Begum. Non-Contact-Based Driver’s Cognitive Load Classification Using Physiological and Vehicular Parameters. *BIOMEDICAL SIGNAL PROCESSING AND CONTROL*, Vol. 55, 2020. https://doi.org/10.1016/j.bspc.2019.101634.
17. Meteier, Q., M. Capallera, S. Ruffieux, L. Angelini, O. Abou Khaled, E. Mugellini, M. Widmer, and A. Sonderegger. Classification of Drivers’ Workload Using Physiological Signals in Conditional Automation. *Frontiers in Psychology*, Vol. 12, 2021. https://doi.org/10.3389/fpsyg.2021.596038.
18. S. Kumar, D. He, G. Qiao, and B. Donmez. Classification of Driver Cognitive Load Based on Physiological Data: Exploring Recurrent Neural Networks. Presented at the 2022 International Conference on Advanced Robotics and Mechatronics (ICARM), Guilin, China, 2022.
19. Angkan, P., B. Behinaein, Z. Mahmud, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad. Multimodal Brain–Computer Interface for In-Vehicle Driver Cognitive Load Measurement: Dataset and Baselines. *IEEE Transactions on Intelligent Transportation Systems*, 2024, pp. 1–16. https://doi.org/10.1109/TITS.2023.3345846.

20. Le, A. S., H. Aoki, F. Murase, and K. Ishida. A Novel Method for Classifying Driver Mental Workload Under Naturalistic Conditions With Information From Near-Infrared Spectroscopy. *Frontiers in Human Neuroscience*, Vol. 12, 2018. https://doi.org/10.3389/fnhum.2018.00431.
21. Barua, S., M. U. Ahmed, and S. Begum. Towards Intelligent Data Analytics: A Case Study in Driver Cognitive Load Classification. *Brain Sciences*, Vol. 10, No. 8, 2020, p. 526. https://doi.org/10.3390/brainsci10080526.
22. Prabhakar, G., A. Mukhopadhyay, L. Murthy, M. Modiksha, D. Sachin, and P. Biswas. Cognitive Load Estimation Using Ocular Parameters in Automotive. *Transportation Engineering*, Vol. 2, 2020, p. 100008. https://doi.org/10.1016/j.treng.2020.100008.
23. Qiu, S., H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino. Multi-Sensor Information Fusion Based on Machine Learning for Real Applications in Human Activity Recognition: State-of-the-Art and Research Challenges. *Information Fusion*, Vol. 80, 2022, pp. 241–265. https://doi.org/10.1016/j.inffus.2021.11.006.
24. Porges, S. W., and E. A. Byrne. Research Methods for Measurement of Heart Rate and Respiration. *Biological Psychology*, Vol. 34, No. 2, 1992, pp. 93–130. https://doi.org/10.1016/0301-0511(92)90012-J.
25. Xiao, S., J. Yan, M. Farajtabar, L. Song, X. Yang, and H. Zha. Learning Time Series Associated Event Sequences With Recurrent Point Process Networks. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 10, 2019, pp. 3124–3136. https://doi.org/10.1109/TNNLS.2018.2889776.
26. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. No. 30, 2017.
27. Koya, S. R., and T. Roy. Temporal Fusion Transformers for Streamflow Prediction: Value of Combining Attention with Recurrence.
28. Gao, D., B. Du, X. Tao, and J. Lu. Driver Vigilance Detection from EEG Signals Using Transformer Networks. Presented at the GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022.
29. Wang, J., Y. Xu, J. Tian, H. Li, W. Jiao, Y. Sun, and G. Li. Driving Fatigue Detection with Three Non-Hair-Bearing EEG Channels and Modified Transformer Model. *Entropy*, Vol. 24, No. 12, 2022, p. 1715. https://doi.org/10.3390/e24121715.
30. Yang, H., H. Liu, Z. Hu, A.-T. Nguyen, T.-M. Guerra, and C. Lv. Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 2023, pp. 1–12. https://doi.org/10.1109/TITS.2023.3316203.
31. Meteier, Q., M. Capallera, E. de Salis, L. Angelini, S. Carrino, M. Widmer, O. Abou Khaled, E. Mugellini, and A. Sonderegger. A Dataset on the Physiological State and Behavior of Drivers in Conditionally Automated Driving. *Data in Brief*, Vol. 47, 2023. https://doi.org/10.1016/j.dib.2023.109027.
32. Makowski, D., T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing. *Behavior Research Methods*, Vol. 53, No. 4, 2021, pp. 1689–1696. https://doi.org/10.3758/s13428-020-01516-y.
33. Hamilton, P. Open Source ECG Analysis. Presented at the Computers in Cardiology, 2002.
34. Elsenbruch, S., Z. Wang, W. C. Orr, and J. D. Z. Chen. Time-Frequency Analysis of Heart Rate Variability Using Short-Time Fourier Analysis. *Physiological Measurement*, Vol. 21, No. 2, 2000, p. 229. https://doi.org/10.1088/0967-3334/21/2/303.

35. Patel, M., S. K. L. Lal, D. Kavanagh, and P. Rossiter. Applying Neural Network Analysis on Heart Rate Variability Data to Assess Driver Fatigue. *Expert Systems with Applications*, Vol. 38, No. 6, 2011, pp. 7235–7242. https://doi.org/10.1016/j.eswa.2010.12.028.
36. Begum, S., S. Barua, R. Filla, and M. U. Ahmed. Classification of Physiological Signals for Wheel Loader Operators Using Multi-Scale Entropy Analysis and Case-Based Reasoning. *Expert Systems with Applications*, Vol. 41, No. 2, 2014, pp. 295–305. https://doi.org/10.1016/j.eswa.2013.05.068.
37. Laborde, S., E. Mosley, and J. F. Thayer. Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research–Recommendations for Experiment Planning, Data Analysis, and Data Reporting. *Frontiers in psychology*, Vol. 8, 2017, p. 213.
38. Zheng, L., X.-Q. Qiao, T. Ni, W. Yang, and Y.-N. Li. Driver Cognitive Loads Based on Multi-Dimensional Information Feature Analysis. *Zhongguo Gonglu Xuebao/China Journal of Highway and Transport*, Vol. 34, No. 4, 2021, pp. 240–250. https://doi.org/10.19721/j.cnki.1001-7372.2021.04.021.
39. McVean, G. A Genealogical Interpretation of Principal Components Analysis. *PLOS Genetics*, Vol. 5, No. 10, 2009, p. e1000686. https://doi.org/10.1371/journal.pgen.1000686.
40. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, 1996, pp. 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
41. Zhang, M.-L., and Z.-H. Zhou. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recognition*, Vol. 40, No. 7, 2007, pp. 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019.
42. Joachims, T. *Making Large-Scale SVM Learning Practical*. Publication 1998,28. Technical Report, 1998.
43. Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. No. 30, 2017.
44. Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, 1998, pp. 2278–2324. https://doi.org/10.1109/5.726791.
45. Hochreiter, S., and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, Vol. 9, No. 8, 1997, pp. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
46. Graves, A., A. Mohamed, and G. Hinton. Speech Recognition with Deep Recurrent Neural Networks. Presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.
47. Zhang, J., Z. Yin, P. Chen, and S. Nichele. Emotion Recognition Using Multi-Modal Data and Machine Learning Techniques: A Tutorial and Review. *Information Fusion*, Vol. 59, 2020, pp. 103–126. https://doi.org/10.1016/j.inffus.2020.01.011.
48. Yang, L., H. Yang, B.-B. Hu, Y. Wang, and C. Lv. A Robust Driver Emotion Recognition Method Based on High-Purity Feature Separation. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 12, 2023, pp. 15092–15104. https://doi.org/10.1109/TITS.2023.3304128.
49. Shapley, L. S. Stochastic Games*. *Proceedings of the National Academy of Sciences*, Vol. 39, No. 10, 1953, pp. 1095–1100. https://doi.org/10.1073/pnas.39.10.1095.
50. Du, G., L. Zhang, K. Su, X. Wang, S. Teng, and P. X. Liu. A Multimodal Fusion Fatigue Driving Detection Method Based on Heart Rate and PERCLOS. *IEEE Transactions on*

Intelligent Transportation Systems, Vol. 23, No. 11, 2022, pp. 21810–21820. https://doi.org/10.1109/TITS.2022.3176973.

51. Yu, X., C.-H. Chen, and H. Yang. Air Traffic Controllers' Mental Fatigue Recognition: A Multi-Sensor Information Fusion-Based Deep Learning Approach. *Advanced Engineering Informatics*, Vol. 57, 2023, p. 102123. https://doi.org/10.1016/j.aei.2023.102123.
52. Yang, L., H. Yang, H. Wei, Z. Hu, and C. Lv. Video-Based Driver Drowsiness Detection With Optimised Utilization of Key Facial Features. *IEEE Transactions on Intelligent Transportation Systems*, 2024, pp. 1–13. https://doi.org/10.1109/TITS.2023.3346054.