Driver Cognitive Load Estimation in Conditional Driving With Aligned Attention-Enabled Multimodal Fusion

Ange Wang[a,b], Haohan Yang[c], Jiyao Wang[a], Hai Yang[a,b], Dengbo He[a,b,d*]

[a]*Intelligent Transportation Thrust, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China*

[b]*Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China*

[c]*School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore*

[d]*HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen*

Abstract

Despite the promise of autonomous driving in enhancing road safety, drivers in conditionally automated vehicles are still responsible for driving safety and thus driver state still matters to driving safety. Though driving automation can reduce taskload of drivers, they may still experience high cognitive load, which can impair takeover performance. However, existing cognitive load estimation algorithms were primarily designed for non-automated vehicles, which may not be applicable in conditionally automated vehicles, due to the differences in driver responsibilities and the availability of certain metrics (e.g., driving performance measures are absent when drivers are not controlling the vehicle). Further, existing driver cognitive load algorithms rarely considered the integration of both spatial and temporal information in the input features. Therefore, we proposed an aligned-attention transformer network that integrates the multi-stream transformer network with alignment attention to estimate the cognitive load of drivers in conditionally automated vehicles. The algorithm fuses physiological measures that can potentially be measured non-intrusively in vehicles, i.e., electrocardiogram, electrodermal activity, and respiration signals. To validate the efficacy of the algorithm, we supplement a European dataset with a self-collected Chinese dataset, in which 42 drivers engaged in various cognitive tasks (i.e., memory, calculation, and spatial tasks). The results showed that our algorithm outperformed state-of-the-art driver cognitive estimation algorithms on both within-subject and across-subjects data partitions. Further, ablation tests validated the robustness of our algorithm and the effectiveness of the network modules. This research can guide the design of driver state monitoring systems in both non-automated vehicles and conditionally automated vehicles.

Keywords

Cognitive load, Physiological signal, Aligned attention, Transformer Network, Automated driving

1. Introduction

Human error is recognized as one of the dominating factors in road accidents (Singh, 2015). It has been commonly acknowledged that, compared to driving tasks that are visually and manuallly demanding (e.g., speed controlling), the cognitive demanding tasks (e,g., hazard perception and driving strategy selection) can be more safety-critical, and thus drivers' performance in these tasks has been widely adopted as key metrics differentiating novice and experienced drivers (Jackson et al., 2009). In addition to driving-related tasks, the introduction of infotainment functions in the smart cabin (e.g., video-streaming and internet browsing) and the prevalence of the bring-in smart devices (e.g., smartphones) may also increase the task load of drivers.

Cognitive load is typically defined as "*the information processing capacity or cognitive resources required to meet the actual demands*" (Babiloni, 2019). The high cognitive load in driving has been found to be closely related to driving safety. For example, a high cognitive load may lead to delayed responses to emergency events (Harbluk et al., 2007), visual tunnel effect (Recarte & Nunes, 2000), decreased ability to anticipate hazards (Muhrer & Vollrath, 2011), and increased reaction times (Du et al., 2020) in traditional vehicles without advanced driving automation systems (ADASs). The introduction of driving automation in recent years may be a solution, as a lower overall task load has been observed in vehicles equipped with ADASs (Huang et al., 2024). However, some other studies found that ADAS can increase drivers' cognitive load due to the additional responsibility to monitor automation (Stapel et al., 2019). For example, Melnicuk et al., (2021) found that high cognitive load can impair takeover performance when driving with ADAS. Thus, estimating drivers' high cognitive states or adaptively supporting drivers when they are experiencing high cognitive loads (He et al., 2019) are critical to driving

safety when both driving with and without ADAS – both require accurate estimation of drivers' cognitive states.

However, most of the previous cognitive load estimation algorithms were developed for non-automated vehicles, which may not apply to vehicles with ADAS, especially SAE Level 3 vehicles (*J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, 2018). This is because drivers no longer need to control and monitor the vehicle constantly in SAE Level-3 vehicles. As a result, the driving performance and eye-tracking measures that were commonly used in non-automated vehicles (Sun et al., 2021; Wang et al., 2022) become unavailable in SAE Level 3 vehicles, and drivers may exhibit different physiological and eye-tracking patterns when performing different tasks (Wang et al., 2024). Especially, eye movement behaviour is highly context-dependent. Different driving environments (such as urban roads versus highways) can lead to very different visual behaviour patterns (Du et al., 2024). This makes eye-tracking a less reliable indicator of cognitive load in practical applications, especially in multitasking driving scenarios. Thus, evaluating and developing new models for driver state estimation is critical to SAE Level-3 vehicles, especially as drivers were more inclined to engage in non-driving-related tasks (NDRTs) when driving with ADAS (He & Donmez, 2019).

To the best of our knowledge, we can only identify three studies that focused on high cognitive load estimation in vehicles with driving automation (Meteier et al., 2021; Shi et al., 2024; Wang et al., 2024), and most of the existing driver cognitive estimation approaches (including the ones for non-automated vehicles and vehicles with driving automation) have two major limitations:

- Most previous models relied on manual feature extraction of physiological features (Harbluk et al., 2007; He et al., 2022; Kumar et al., 2022; Meteier et al., 2021; Wang et al.,

2024), for example, heart rate (HR) extracted from electrocardiogram (ECG) (He et al., 2022), and respiratory rate extracted from respiratory signals (RESP) (Qu et al., 2023). While satisfactory accuracy has been achieved in previous research, the extraction of these handcrafted low-level features is computational costing and may lead to loss of information in the raw signals. To the best of our knowledge, only one driving study utilized the raw physiological signals directly for driver cognitive load estimation (Shi et al., 2024), but only an ECG measure was utilized.

- Most driver cognitive load estimation studies used classical machine learning models that ignored the temporal-spatial dependency of physiological signals. Only a few studies used Recurrent Neural Networks (RNNs) (Kumar et al., 2022) and Long Short-Term Memory (LSTM) networks (Ansari et al., 2022; Yang et al., 2023) that could consider temporal information for driver cognitive load estimation. However, RNN or LSTM may still neglect the long-distance temporal dependencies present in time series and previous studies did not fuse multiple physiological features, ignoring the spatial dependency among signals.

To address the aforementioned challenges, in this study, we proposed an Attention-Aligned Transformer algorithm for Cognitive (CogFormer) load estimation in vehicles with driving automation. The CogFormer adopted a multi-stream transformer architecture to fuse ECG, Electrodermal Activity (EDA), and RESP signals. Specifically, separate transformer networks were designed for the three physiological signals, followed by a synthesization network with the aligned-attention mechanism at a decision level. We then trained and validated our model on both the European SAE Level-3 dataset (Meteier et al., 2023) with 88 participants and on a self-collected dataset, the China HIS@HKUST(GZ) Cognitive Load Driving (CAM-CLD), which involved multiple types of cognitive tasks and 42 participants in a simulated SAE Level-3 driving experiment.

2. Related work

2.1. Physiological signals-based driver cognitive load recognition

Physiological measures are associated with both the central and autonomic nervous systems and have been found to be associated with drivers' cognitive states. These measures include heart activity (HR and heart rate variability (HRV)), EDA (tonic and phasic skin conductance), RESP (respiratory rate and amplitude), and brain activity measured through EEG. Among these measures, the EEG signals are directly related to brain activities and have been widely used for driver state estimation. For example, an EEG-based Bayesian neural network has been successfully used to determine driver fatigue (Chai et al., 2017). At the same time, ECG signals were also widely utilized, as it is relatively easy to collect. For example, the ECG signal alone (Tjolleng et al., 2017) or in combination with Galvanic Skin Response (GSR) (Kumar et al., 2022; Wei et al., 2023) has been used for cognitive detection in previous research. However, it should be noted that, to date, we can only identify one study that utilized high-level features for drivers' cognitive state estimation (Shi et al., 2024) and most previous research used handcrafted physiological signals and manually adjusted hyper-parameters in the models, which is time-consuming and may lead to loss of information in the raw signals.

2.2. Learning-based driver cognitive load recognition methods

Previous cognitive load estimation algorithms mainly targeted drivers in non-automated vehicles (i.e., SAE Level 0). These methods can be categorized into two major classes, i.e., traditional machine learning classifiers (e.g., LightGBM algorithm, K-Nearest Neighbors (KNN) and Random Forest (RF)) and deep learning algorithms. The former is mostly based on handcrafted features and rarely considers temporal information in signals. The deep learning approaches have gained increasing attention in recent years due to their capability to handle complex patterns in signals. For example, the Convolutional Neural

Network (CNN) has been used to detect high workload among driver (Gao et al., 2019; Hajinoroozi et al., 2016; Kose et al., 2019). The LSTM and its variants were found to exhibit superior performance in predicting driver's cognitive load based on physiological signals (Ansari et al., 2022; Yang et al., 2023). Wang et al. (2018) proposed an end-to-end framework for real-time cognitive load classification based on a mixed Hyper Long Short-Term Memory network (m-HyperLSTM), a variant of hypernetwork. More recent work has even explored the combination of CNN and LSTM for cognitive load estimation (Yu et al., 2023; Zhang et al., 2021). However, previous studies only considered data fusion at the feature level and failed to fully leverage the complementarity of multimodal data (Yang et al., 2024); in contrast, Yang et al. (2023) developed an attention recognition network with the cross-attention mechanism that can fuse the features at the decision level, which can automatically select useful features from multimodal time-series data. Other driving-related cognitive load estimation studies using physiological signals studies in the past five years are summarized in Table 1.

Table 1. Driver cognitive load estimation methods with physiological signals in the past five years.

Reference	Physiological Signal	Input	Cognitive Task	Data Partition	Automation Level	Models	Accuracies
Wang et al. (2024)	ECG	RD	2 classes: n-back task and math task	Across subjects	L0	CogDG-ECG	Dataset 1: 61.21%; Dataset 2: 69.33%; Dataset 3: 74.06%
Angkan et al. (2024)	ECG, EEG, Eye-tracking data	FEM and RD	3 classes: reading and arithmetic tasks	Across subjects (Leave one subject out)	L0	ResNet	64.53%
He et al. (2022)	ECG, GSR	FEM	3 classes: n-back task	Within-subject	L0	KNN, SVM, FNN, RNN, RF	72.6%
Yang et al. (2023)	EEG	FEM	3 classes: memory and recall tasks	Within-subject	L0	ARecNet	95%
Rahman et al. (2020)	ECG	FEM	2 classes: n-back task	Within-subject	L0	LR, SVM, LDA, FNN	87%
Meteier et al. (2021)	ECG, EDA, RESP	FEM	2 classes: math task	Within-subject	L3	MLP, RF, SVC	90%

Kumar et al. (2022)	ECG, GSR, EEG	FEM	2 classes: n-back task	Within-subject/Across-subjects	L0	RNN, SVM, FNN, KNN, DTC, NB, LDA	Within-subject: 88.1%; Across-subjects: 85.6%
Barua et al. (2020)	ECG, EEG, EOG, GSR, RESP	FEM	3 classes: n-back task	Not mentioned	L0	RF	79%
Prabhakar et al. (2020)	Eye -tracking data	RD	2 classes: n-back or math tasks	Not mentioned	L0	NN	75%
Wang et al. (2024)	ECG, EDA,RESP	FEM	Two classes: math task	Within-subject	L3	Transformer	94.4%
Shi et al. (2024)	ECG	Spectrogram	Two classes: n-back	Within-subject	L3	Lightweight Neural Network	92%

Notes: FEM - Features Extracted Manually; RD - Raw Data; KNN - k-Nearest Neighbors; SVM - Support Vector Machine; FNN - Feedforward Neural Network; RF - Random Forest; LR - Logistic Regression; LDA - Linear Discriminant Analysis; MLP - Multilayer Perceptron; SVC - Support Vector Classifier; DTC - Decision Tree Classifier; NB - Naive Bayes; NN - Neural Network; CogDG-ECG - Cognitive Domain Generalization with Electrocardiogram; XGM - Extreme Gradient Boosting.

3. Attention-Aligned Transformer for Cognitive (CogFormer) Estimation

3.1 Problem Formulation

The recognition of driver cognitive load is approached as a supervised classification task, wherein the workload levels are utilized as categorical labels, as shown in Figure 1. Then, the dataset can be described as follows:

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

where T^l is the multiple temporal sequences with the size of l; y is the cognitive workload of the driver, including low load (baseline, $y = 0$), and higher load ($y = 1, 2, 3$); N is the total number of samples. For each sample,

$$T^l = [f_1, f_2, \dots, f_l], \quad (2)$$

where f is the vector of physiological signals at each time step, where f consists of ECG, RESP, and EDA signals, represented by $f = [\alpha, \beta, \gamma]$, therefore from the perspective of signal type, T^l can be represented by $T^l = [\alpha^l, \beta^l, \gamma^l]$. Here, $f \in \mathbb{R}^{42}$ and $T^l \in \mathbb{R}^{42 \times l}$. The research objective of the model is to generate the workload level $\hat{y}_m$ based on the temporal series information T_m^l, so that:

$$\hat{y}_m = \arg\max_{y_m \in \zeta} p(y_m | T_m^l), \quad (3)$$

where $\zeta = \{0,1,2,3\}$ represents the label of cognitive workload levels.

3.2 Model Construction

As shown in Figure 1, the developed CogFormer contains three parallel transformer-encoder-based modules , the aligned and self-attention mechanism, three max-pooling layers, and a classifier layer.

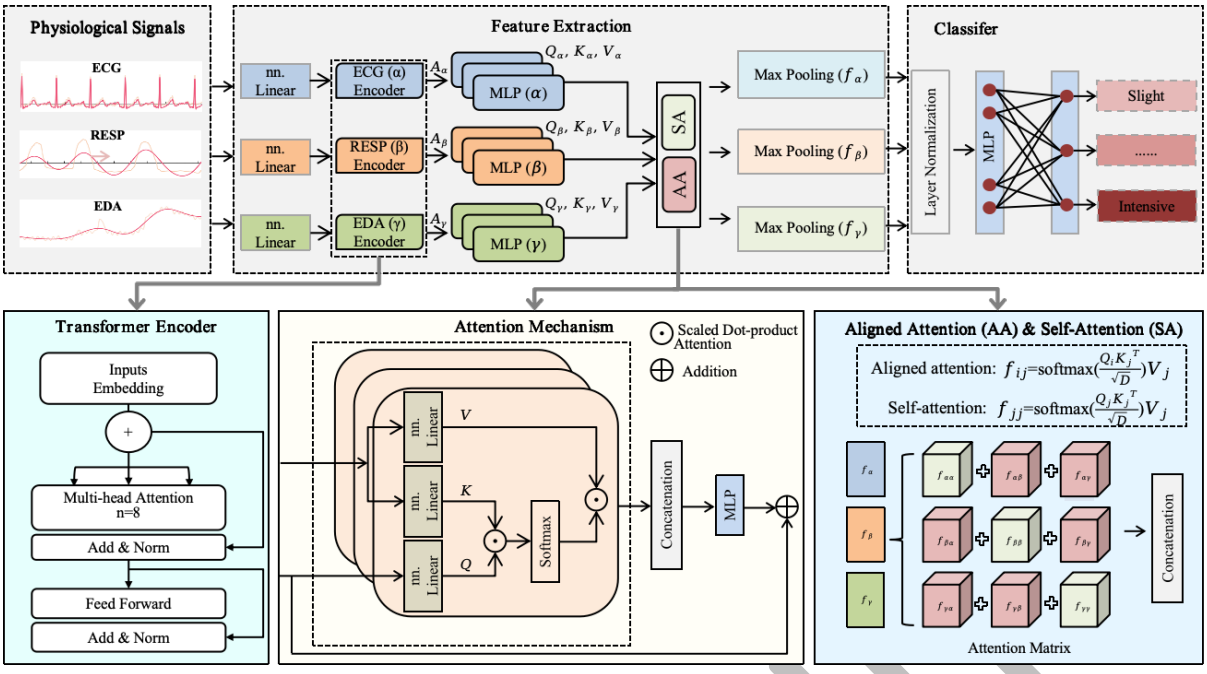


Figure 1. Overview of the proposed CogFormer for driver cognitive load detection.

- Input preprocessing

The initial phase involves the preprocessing of three physiological signals, i.e., ECG data $X^{\alpha} \in \mathbb{R}^{l \times F_{\alpha}}$, RESP data $X^{\beta} \in \mathbb{R}^{l \times F_{\beta}}$, and EDA data $X^{\gamma} \in \mathbb{R}^{l \times F_{\gamma}}$, Here, F_s represents feature dimensions for each physiological signal. Linear transformations were applied to all signal inputs to adjust the input vector dimensions to be compatible with subsequent computations:

$$H^s = X^t W^s + b^s, s = \{\alpha, \beta, \gamma\}, \quad (4)$$

where X^t denotes the transpose of X, W^s and b^s are the weight matrices and bias vectors, respectively, transforming each physiological data into a unified feature representation, here $W^s \in \mathbb{R}^{F_s \times D}$, D is the hidden layer dimension.

- Multi-Head Self-Attention

Our model employs both self-attention and aligned attention within its alignment attention mechanism. Specifically, 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 (5), which is independently applied, then concatenated, and subsequently fed into a linear layer to generate sequences:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (5)$$

where, the weighted representation, $Attention\ (Q, K, V)$, is obtained through simultaneous calculation of matrices Q, K and V, and 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; 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.

For each head in the multi-head attention mechanism, the queries Q, keys K , and values V are computed:

$$Q_i^s = H^s W_{Qi},\ K_i^s = H^s W_{Ki},\ V_i^s = H^s W_{Vi}, \quad (6)$$

where W_{Qi}, W_{Ki} and $W_{Vi} \in \mathbb{R}^{D\times n}$ are the weight matrices for the *i*-th head. *n* is the projection dimension for each attention head.

The outputs of multiple heads are concatenated and linearly transformed:

$$MultiHead^s = Concat\ (Head_1^s, Head_2^s, \ldots, Head_h^s)W_O, \quad (7)$$

where $W_O \in \mathbb{R}^{h\cdot d_v\times D}$ is the output linear transformation matrix.

- Attention design

The aligned attention mechanism integrates multiple physiological signals, i.e., ECG, RESP, and EDA signals, by utilizing self-attention and aligned attention. Self-attention captures dependencies within each signal type, while aligned attention aligns and integrates information across different signals. This approach allows the model to weigh and combine the most relevant features from each signal, creating a cohesive and comprehensive representation.

The ability of the mechanism to fuse diverse data sources can allow the model to understand and predict physiological patterns by capturing complex interdependencies. Self-attention and aligned attention are expressed by equations (8) and (9), respectively.

$$f_{ii} = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_k}}\right) V_i, \tag{8}$$

$$f_{ij} = softmax\left(\frac{Q_i {K_j}^T}{\sqrt{d_k}}\right) V_j, i, j \in s \; and \; i \neq j. \tag{9}$$

- Classifier layer

The CogFormer calculates the probability of the subject experiencing each level of cognitive laod by applying a nonlinear projection through a classifier layer.

$$\hat{y} = softmax(O), \tag{10}$$

the classifier output $\hat{y}$ is obtained by applying the softmax function to the result of the output layer O. $\hat{y}$ denotes the predicted probabilities for each category.

- Model training

Training is performed using the cross-entropy loss function:

$$Loss = -\sum_{n=1}^{N} \sum_{c=1}^{C} y_{nc} \, log(\hat{y}_{nc}), \tag{11}$$

where the term y_{nc} represents an indicator of the true label for the n^{th} sample in category c. $\hat{y}_{nc}$ denotes the probability predicted by the model that the n^{th} sample belongs to category *c.*

4. Datasets

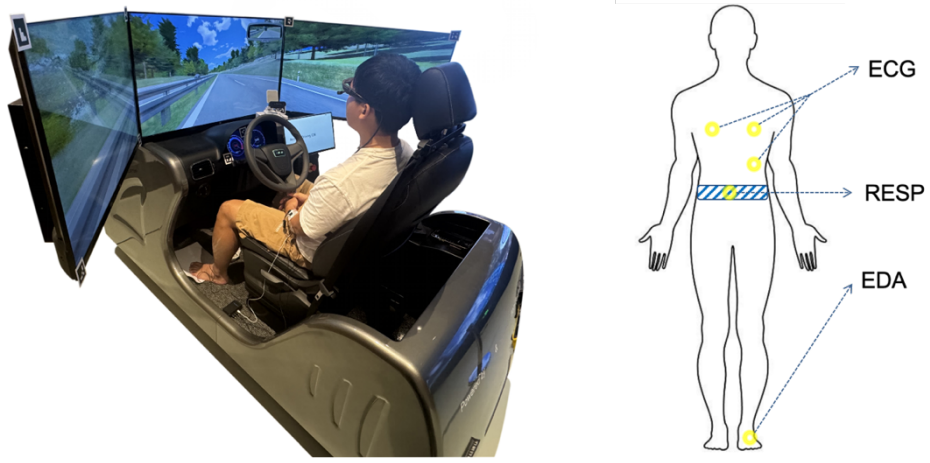
To the best of our knowledge, we can only identify one open dataset that focused on drivers' cognitive load in SAE Level-3 vehicles (i.e., (Meteier et al., 2023)). However, only one verbal-cognitive NDRT was used in this dataset, i.e., the arithmetic counting down task, which may not evaluate the generalizability of the model. Thus, to better validate our proposed model, we constructed the CAM-CLD dataset, which focuses on drivers' cognitive states in

simulated SAE Level-3 vehicles. The details of the experiment for CAM-CLD are provided below.

4.1 CAM-CLD Dataset

4.1.1 Equipment

The experiment was conducted in a fixed-base driving simulator (Figure 2a), which has three 43-inch displays, showing a horizontal view angle of 150° and a vertical viewing angle of 47°. Participants could engage and disengage the ADAS by pressing the virtual buttons on the 15-inch screen next to the steering wheel. They could also disengage the ADAS and take over control of the vehicle by turning the steering wheel and pressing the brake. The driving data was logged at a frequency of 60 Hz by the simulation software, i.e., the Silab 7.1 by WIVW GmbH. The ECG, RESP, and EDA were collected using the sensors by Ergoneers GmbH at a sampling frequency of 100 Hz (see Figure 2(b)). It should be noted that the eye-tracking data was also collected by Dikablis 3 in our study, but the data was not used in our model. All data was synchronized in the Human Research Tool (HRT) software by Info Instrument.



(a) (b)

Figure 2. The equipment (a) of the driving platform and (b) physiological sensor placements.

4.1.2 Participants

Based on the sample size estimation with MorePower (Campbell & Thompson, 2012), at least 24 participants are required to ensure a statistical power of 80%. Thus, 42 drivers were recruited for this experiment (25 males and 17 females), with an average age of 35.28 years (Standard deviation [SD]: 9.10, min: 23, max: 53). All participants were required to have a valid driver's license for a minimum of one year, and were compensated at 70 RMB per hour. Additionally, a performance-based bonus of up to 30 RMB was offered as an incentive for cognitive NDRTs provided in the experiment. This study was approved by the Hong Kong University of Science and Technology (HREP-2023-0199).

4.1.3 Driving Tasks

Each participant completed one drive on a dual carriageway with six lanes with 21 straight sections, each stretching around 7 km. The traffic is free-flowing, with a vehicle density of 6 passenger car units per km per lane. All participants were required to engage the driving automation when possible in the experiment and only take over the control of the vehicle when they felt necessary or prompted by the takeover request (TOR). The TOR initiated in 3 types of events, i.e., ramp, traffic accident and fog area. The driving automation can work under the speed of 110 km/h and the speed limit of the carriageway was 60 km/h.

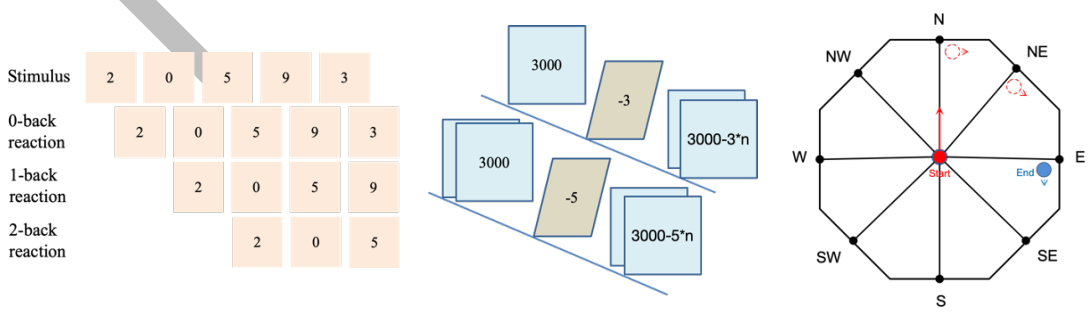
4.1.4 Cognitive Tasks

Three types of cognitive tasks were adopted in the experiment, requiring three different types of cognitive resources with different levels of difficulty, i.e., the n-back task for working memory (3 levels), the math calculation task for mental calculation (2 levels), and the cognitive spatial task for spatial resources (1 level). Notably, the baseline task consisted solely of driving; other scenarios combined driving tasks with NDRTs.

- **The n-back task (Jaeggi et al., 2010)**, has been widely used to manipulate cognitive load in various contexts, including driving (Mehler, 2012). In the n-back task, a series of stimuli,

such as numbers or letters, are presented, with a sustained pause between each stimulus. During this pause, participants are required to recall the stimuli presented *n* positions earlier (refer to Figure 3a for an example). The levels within the n-back task, denoted as *n*, indicate the difficulty and complexity of the task, with a larger *n* corresponding to higher task difficulty. In our study, 0-, 1-, and 2-back tasks were used.

- **The math task (Meteier et al., 2021)** involves counting backward from 3,000 in decrements of 3 (non-integer) or 5 (integer) (refer to Figure 3b for an example) in our study. This task resembles continuous communication activities, such as phone conversations, and can induce high levels of cognitive load over an extended duration. The difficulty level of the task can be controlled by the initial number (i.e., 3000 in our study) and the decrements (i.e., 3 and 5 in our study).
- **The cognitive spatial task** simulates spatial processing in navigation (Liang & Lee, 2010), requiring participants to listen to an audio clip describing a person's route, and then orally identify the main direction (e.g., East, North, Southwest) faced by this person in the end. As shown in Figure 3c, the map has eight stations connected to the center. Participants were told to go to one of the stations from the center, and then go *n* steps clockwise or counterclockwise. For example, when being asked "Where is this person when she goes to the north station and moves two stations clockwise?", the answer is east. This task simulates the process of directional information when using navigation systems.



(a) (b) (c)

Figure 3. Cognitive tasks, including (a) n-back task, (c) math task, and (d) cognitive spatial task.

4.1.5 Procedures

Participants were reminded to maintain regular sleep habits, avoid alcohol, and not consume caffeine 24 hours before the study. Upon arrival, participants provided written consent, acknowledging their voluntary involvement in the study. This was followed by a 30-minute orientation session regarding the experimental procedures, operation of the vehicles, and cognitive tasks. Then, the participant underwent a practice driving session, in which the driver experienced one TOR. Afterward, the participants were equipped with physiological sensors and eye-tracking devices, and then the formal experiment began, after each TOR, the driver was required to pull over and complete subjective questionnaires, including the NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988) and Karolinska Sleepiness Scale (KSS) (Kaida et al., 2006). The analysis of the TOR is out of the scope of this study, and we only used the physiological signals associated with NDRTs and baseline before the start of the TOR. Table 4a shows the entire experimental process, while Table 4b presents the different stages of each scenario. Each experiment took around 3 hours in total, and leading to the number of samples shown in Table 2.

Table 2. The number of samples of different datasets.

Dataset	t_w=1s	t_w=3s	t_w=5s
MADT-D: Math task	52800	17600	10560
CAM-CLD: Spacial task	32375	10790	6475
CAM-CLD: Math task	47490	15830	9498
CAM-CLD: n-back task	66315	22105	13263

Note: t_w is the time windows length.

4.1.6 Experiment Design

A 3 (Takeover Scenarios) by 7 (NDRTs) within-subject design was used. The 3 scenarios were counterbalanced using a Latin squared design. Within each scenario, the 7 NDRT types were counterbalanced using a Latin squared design, leading to 7 possible orders. This combination of scenario and NDRT resulted in a total of 21 unique experimental conditions.

Each combination (in total 21) of experimental conditions had two participants. Specifically, after completing all NDRTs within one Takeover Scenario, participants took a 10-minute break before proceeding to the next 7 takeover scenarios.

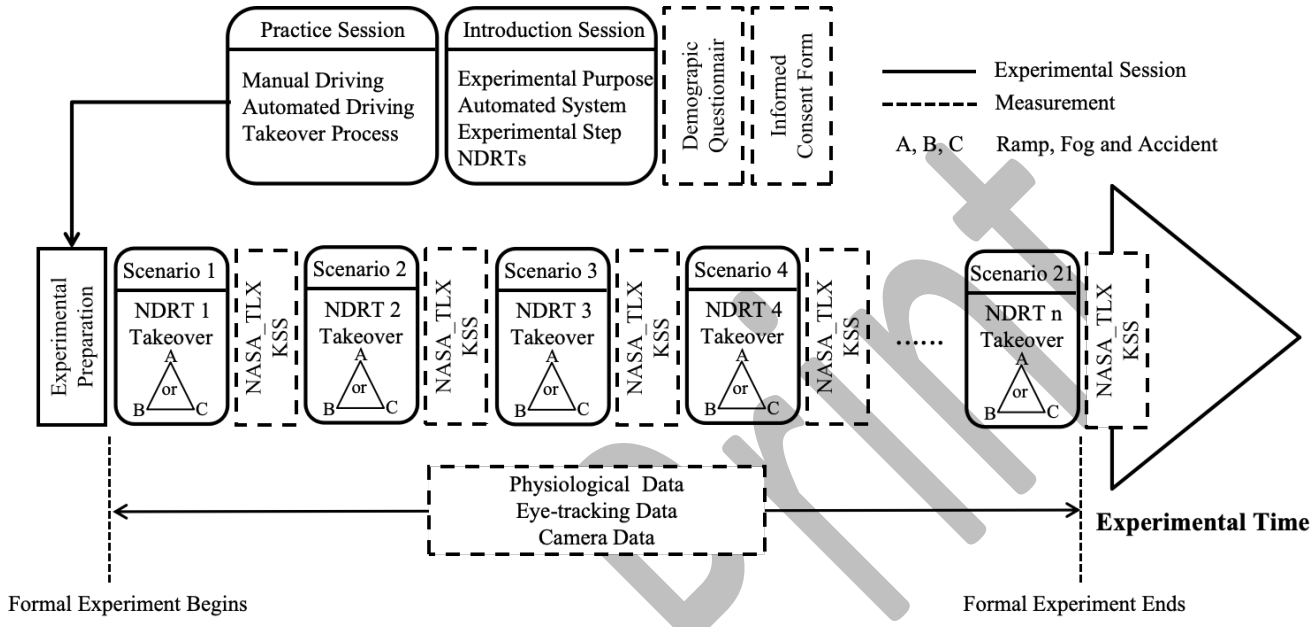


Figure 4a. Experimental process.

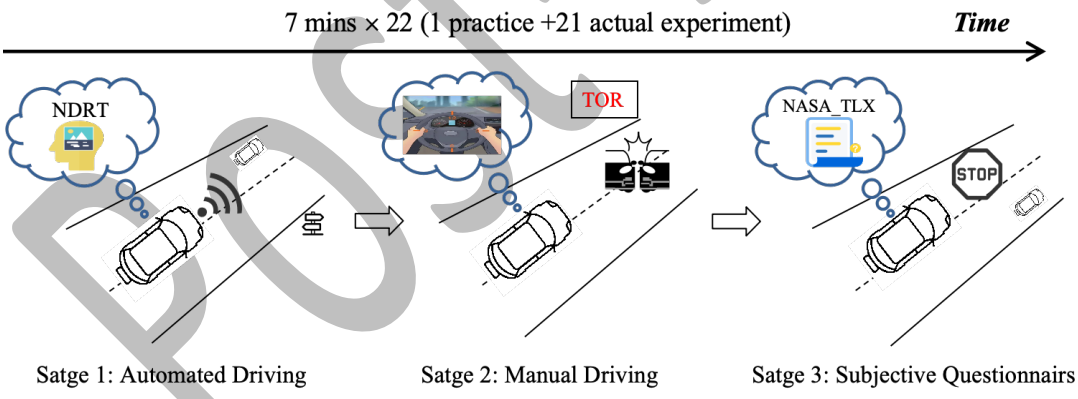


Figure 4b. Scenario procedure.

4.1.7 Validation

The data tested did not follow a normal distribution, therefore we used the Friedman test to compare the imposed cognitive load levels across different task difficulty levels within a type of cognitive task. If a significant effect was observed, a Wilcoxon signed-rank test was used for post-hoc comparisons. As shown in Figure 5, our cognitive load tasks imposed

significant and distinguishable load levels in drivers. Therefore, we always labeled the baseline as low cognitive load; for the math task, we labeled MT2 as medium cognitive load and MT1 as high cognitive load; for the n-back task, we labeled 0-B as medium-low cognitive load; 1-B as medium-high cognitive load and 2-B as high cognitive load.

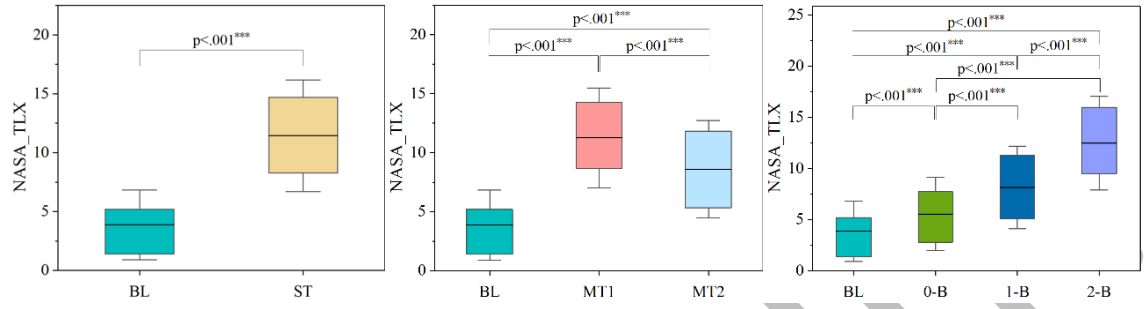


Figure 5. Significance test results of self-reported NASA-TLX. (a) Cognitive spatial (ST) task, (b) Math task, and (c) N-back task. Note: MT1 is math task 1 (decrement of 3), MT2 is math task 2 (decrement of 5), BL is the baseline (driving only), 0-B is 0-back task, 1-B is 1-back task and 2-B is the 2-back task, *** means $p < .05$.

In addition, we examined whether it is necessary to identify the specific sources of cognitive load in our model. Since the purpose of detecting a driver's cognitive load in conditional automated driving is to help drivers maintain a moderate level of cognitive load to prepare for potential takeover events, we compared the effects of different cognitive load sources (including road monitoring only and performing cognitive load tasks) on takeover performance, when the drivers self-reported to have similar cognitive load levels. Specifically, we developed a mixed linear model (accounting for repeated measures and individual differences) based on our collected dataset. Samples with varying NASA TLX scores, regardless of the sources of the high cognitive load, were divided into three quartiles (i.e., low, medium, high). Notably, no high-cognitive samples originated from road monitoring alone. Thus, our model was based on low and medium cognitive load levels. As shown in Tables 3

and 4, the metrics for assessing takeover performance include speed, acceleration, steering wheel angle, and lateral deviation. The results indicate that the source of cognitive load had no significant effect on takeover quality ($p > .05$), suggesting that distinguishing between cognitive load sources is not essential.

Table 3. The impact of cognitive source on takeover quality at a low cognitive level.

Dependent Variable	Independent Variable	F value	***p*-value**
V	Source of cognitive load	0.04	**0.838**
AX		2.25	**0.135**
AY		0.14	**0.712**
SW		0.7	**0.403**
Mean LD		0.64	**0.425**
Max LD		1.04	**0.308**

Note: in the table, speed - V, longitudinal acceleration – AX, lateral acceleration - AY, steering wheel angle - SW, maximum lateral deviation - Max LD, mean lateral deviation - Mean LD,

Table 4. The impact of cognitive source on takeover quality at a medium cognitive level.

Dependent Variable	Independent Variable	F value	***p*-value**
V	Source of cognitive load	0.47	**0.493**
AX		1.6	**0.206**
AY		0.01	**0.949**
SW		0.02	**0.879**
Mean LD		0.2	**0.652**
Max LD		0.1	**0.757**

4.2 Datasets

Apart from our own dataset, we also tested our model on the mathematical and autonomous driving tasks dataset (MADT-D) published by the University of Applied Sciences and Arts of Western Switzerland (Meteier et al., 2023). The MADT-D consists of driving data from 90 participants (with two participants deemed invalid). In the experiment, half of the participants were instructed to perform a cognitive task known as the oral digit span counting task, requiring them to verbally count backward from 3,645 in decrements of 2 (labeled as high cognitive load); while the rest half conducted a driving task only (labeled as low cognitive

load). To ensure that both load levels appear in the leave-one-out test, we combined one participant performing the NDRT with one participant not performing the NDRT to form a new subject. Consequently, there are a total of 44 new participants. Similarly, in the experiment, the physiological data, i.e., EDA, RESP, and ECG, were collected using sensors by BioPac at a frequency of 1,000Hz in the dataset. The differences in subjective ratings of cognitive load between tasks were validated in Meteier et al., (2021). We extracted data spanning 10 minutes from the cognitive load phase in the MADT-D dataset. The attributes of these datasets are summarized in Table 5. These datasets provide a comprehensive view of cognitive workload across different age groups, cultural backgrounds, and driving experiences, making the findings more broadly applicable and reducing the potential for model bias.

Table 5. Attributes of the datasets.

Aspect	**CAM-CLD Dataset**	**MADT-D Dataset**
Age Range and Mean Age	Diverse age range (23 to 53 years), Mean age: 35.28 years	Narrower, younger demographic, Mean age: 24.15 years
Cultural Background	Asian backgrounds	Primarily European participants
Professional Background	Wide range of occupations	Student-based sample
Driving Experience	Participants with at least one year of licensed driving experience	Predominantly student population with less varied driving experience
Generalization Benefit	Captures varied cognitive workload across life stages and real-world driving expertise	Provides insights into cognitive workload among younger, less experienced drivers

4.3 Signal Preprocessing

Given that we used the raw data as inputs in the model, only noise elimination was conducted to enhance the quality of the data. Specifically, all signals underwent down-sampling to a frequency of 100 Hz (from 1000Hz in MADT-D) to optimize computational efficiency and ensure the consistency between two datasets. For EDA, a low-pass filter with a cutoff frequency of 5 Hz was employed; while for ECG and RESP, band-pass filters were applied within the frequency ranges of 3 Hz to 45 Hz and 0.1 Hz to 0.35 Hz, respectively (Meteier et al., 2021). The preprocessing was executed using Python 3.8.

5 Baseline Models and Evaluation Metrics

We compared our model to selected learning-based approaches from prior studies for cognitive load estimation in driving:

- **MTS-CNN**: This model represents a variant of the CNN-based architecture designed to discern cognitive workload levels by amalgamating multivariate temporal features captured across all convolutional layers. The model parameters were set following Xie et al., (2020).
- **DecNet**: As a variant of the LSTM-based network, this model extracts features through an RNN module and infers cognitive workload levels via an LSTM network. The structure of the DecNet model in this study aligns with that outlined in (Amadori et al., 2022).
- **CNN-LSTM**: Combining feature extractions from the convolutional layer and sequence predictions in LSTM layers, the CNN-LSTM model synergistically integrates the strengths of both CNN and LSTM architectures for workload-level inference. The parameters of the CNN-LSTM model in this investigation were adopted from Huang et al., (2022).
- **m-HyperLSTM**: An adaptation of HyperNetworks, the m-HyperLSTM incorporates multimodal information fused at the feature level (Wang et al., 2018). Consequently, a single HyperLSTM-based module is employed in the m-HyperLSTM network. The

maximum epoch was set to 30 to ensure convergence of the validation loss, with 128 and 64 hidden units selected for the main LSTM and HyperLSTM, respectively. Other parameters were the same as those in the ARecNet illustrated below.

- **ARecNet**: This model captures and enhances time-series feature representations through HyperLSTM-based modules and a cross-attention mechanism (Yang et al., 2023). In our study, this model was trained using the Adam optimizer with a learning rate of 0.001, and batch size of 64. The training was capped at a maximum of 20 epochs, with the model achieving the lowest validation loss adopted. The optimal number of hidden units was identified as 64 for the main LSTM and 32 for the HyperLSTM.
- **CogFormer**: In CogFormer, we selected the hyperparameters using Grid Search. The search ranges for hyperparameters are listed in following Table 6. The data for each signal was encoded using its own Transformer Encoder Layer, configured with a hidden size of 128, 8 attention heads, and 1 layer. The encoded outputs were concatenated with an aligned attention mechanism and passed through a fully connected layer to produce the final output. The model was trained using the Adam optimizer with a learning rate of 0.001, a dropout rate of 0.1, and a batch size of 64. The training was capped at 20 epochs with early stopping based on the lowest validation loss.

Table 6. Search Ranges of Hyperparameters

Hyperparameter	**Description**	**Searched Values**
Learning Rate	Controls the step size during model weight updates.	0.01, **0.001**, 0.0001
Batch Size	Number of samples per gradient update.	16, 32, **64**, 128
Hidden Dimension	Size of internal representations within the model layers.	64, **128**, 256

Dropout Rate	Probability of dropping units to prevent overfitting.	**0.1**, 0.3, 0.5
The Number of Attention Heads	Parallel attention mechanisms in multi-head attention.	2, 4, **8**, 16

The model was assessed under two data partition protocols (i.e., within-subject data partition with 5-fold cross validation and across-subjects data partition with leave-one-subject-out validation (Esterman et al., 2010)). For the within-subject data partition, 80% of the data from each participant was used for training, and the remaining 20% of data from each participant was used for testing. As for the across-subjects data partition, data from 41 participants (in CAM-CLD dataset) and 43 participants (in MADT-D dataset) were used for training and the rest 1 participant (both in two datasets) was used for validation.

The average accuracy, F1 score, and Area Under the Curve (AUC) on the validation datasets were used for model evaluation. The AUC is defined as the area under the receiver operating characteristic curve and the accuracy and F1 score are formulated as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN},$$

$$F1\ score = \frac{2TP}{2TP + FP + FN},$$

where $TP,\ FP,\ FN\ and\ TN$ denote true positives, true negatives, false positives, and false negatives, respectively.

6 Results and Discussions

6.1 Model Accuracies and Discussions

As shown in Table 7 and Table 8, the model comparison encompassed different data partitioning protocols (i.e., within-subject and across-subjects), various cognitive load tasks, and classification categories, showcasing the model performance across different time horizons (i.e., the length of physiological signal time series, represented by t_w). Figure 6 further shows

the confusion matrices for CogFormer under the within-subject and across-subjects partition protocols. In the figure, the bottom row and the rightmost column are the averages of each column and each row, respectively.

It can be found that under the within-subject data partition protocol, m-HyperLSTM consistently outperformed MTS-CNN and DecNet. Under the across-subjects data partition protocol, m-HyperLSTM consistently outperformed other LSTM-based models (including CNN-LSTM), indicating that its adaptive module could capture feature representations more effectively than traditional LSTM architectures. Additionally, ARecNet surpassed other LSTM and CNN-based models, indicating that the cross-attention mechanism can enhance the learning of useful features in multimodal time series data.

Nevertheless, our proposed CogFormer consistently achieved the highest accuracy in all models. Furthermore, we observed that for some tasks, increasing the time horizon did not necessarily lead to an increase in recognition accuracy. It is also noteworthy that although satisfactory recognition results were obtained under the within-subject protocol, the across-subjects protocol led to a significant decrease in accuracy for all models. Such a decrement in model accuracy has been widely documented in previous research (e.g., 64.53% recognition accuracy in differentiating driver's cognitive load in different driving scenarios, using eye movement, EEG, and ECG data in (Angkan et al., 2024)), indicating the challenges faced when handling unknown subject's data.

Finally, according to the confusion metrics, in the math task, low cognitive load and high cognitive load were frequently misclassified as medium load. For instance, in Figure 6a (c) at t_w = 1 s, CogFormer misclassified 1.2% of low cognitive load (baseline) and 3.0% of medium cognitive load (MT2) as high cognitive load (MT1). It can be observed that the CogFormer was more likely to confuse medium and high cognitive loads (3.0% and 3.3%). Additionally, in the n-back task shown in Figure 6 (d), low (baseline) and higher (1-back task) cognitive

loads were often considered medium (0-back task) loads. Similar patterns were also observed in the results of the cross-subject partitioning method shown in Figure 6b. Thus, specific algorithms may need to be selected if a certain level of cognitive load is of interest.

Table 7. Results on within-subject data partition protocol.

Model	$t_w = 1\ s$			$t_w = 3\ s$			$t_w = 5\ s$		
	Accuracy (%)	F1-score	AUC	Accuracy (%)	F1-score	AUC	Accuracy (%)	F1-scores	AUC
MADT-D: Math task (two classes)									
MTS-CNN$^{\Delta}$	87.71 ± 4.86	0.876 ± 0.044	0.877 ± 0.046	88.83 ± 4.14	0.884 ± 0.042	0.882 ± 0.049	88.29 ± 4.03	0.887 ± 0.042	0.885 ± 0.046
DecNet$^{\Delta}$	86.76 ± 3.29	0.866 ± 0.035	0.863 ± 0.033	87.27 ± 4.47	0.877 ± 0.046	0.876 ± 0.041	87.45 ± 4.34	0.876 ± 0.043	0.872 ± 0.049
CNN-LSTM$^{\Delta}$	87.51 ± 4.14	0.871 ± 0.043	0.879 ± 0.042	87.87 ± 5.63	0.876 ± 0.053	0.874 ± 0.048	88.38 ± 3.59	0.888 ± 0.033	0.881 ± 0.036
m-HyperLSTM$^{\Delta}$	89.68 ± 5.26	0.898 ± 0.055	0.896 ± 0.053	89.32 ± 5.19	0.894 ± 0.053	0.881 ± 0.051	89.14 ± 4.72	0.892 ± 0.045	0.893 ± 0.046
ARecNet$^{\Phi}$	92.46 ± 4.22	0.921 ± 0.045	0.919 ± 0.040	91.93 ± 3.12	0.913 ± 0.031	0.911 ± 0.033	92.56 ± 3.38	0.921 ± 0.031	0.922 ± 0.032
CogFormer (ours)$^{\Phi}$	**95.28** ± 3.98	**0.952** ± 0.039	**0.955** ± 0.037	**93.79** ± 4.25	**0.937** ± 0.047	**0.936** ± 0.044	**94.03** ± 3.81	**0.946** ± 0.039	**0.947** ± 0.035
CAM-CLD: Spacial task (two classes)									
MTS-CNN$^{\Delta}$	86.08 ± 2.85	0.863 ± 0.026	0.865 ± 0.027	87.34 ± 2.39	0.874 ± 0.022	0.879 ± 0.021	88.57 ± 1.97	0.882 ± 0.022	0.885 ± 0.021
DecNet$^{\Delta}$	86.91 ± 2.43	0.867 ± 0.028	0.863 ± 0.023	86.23 ± 1.47	0.867 ± 0.017	0.866 ± 0.016	88.06 ± 2.44	0.888 ± 0.024	0.882 ± 0.025
CNN-LSTM$^{\Delta}$	89.46 ± 2.32	0.899 ± 0.027	0.898 ± 0.026	88.41 ± 2.52	0.887 ± 0.027	0.884 ± 0.023	89.46 ± 2.37	0.892 ± 0.026	0.894 ± 0.023
m-HyperLSTM$^{\Delta}$	88.64 ± 3.76	0.884 ± 0.034	0.889 ± 0.036	88.96 ± 1.29	0.886 ± 0.016	0.884 ± 0.014	89.32 ± 3.45	0.898 ± 0.032	0.893 ± 0.036
ARecNet$^{\Phi}$	91.92 ± 2.46	0.912 ± 0.027	0.909 ± 0.028	90.09 ± 1.33	0.897 ± 0.015	0.890 ± 0.013	91.49 ± 2.96	0.916 ± 0.029	0.918 ± 0.027
CogFormer (ours)$^{\Phi}$	**93.05** ± 2.39	**0.928** ± 0.026	**0.931** ± 0.023	**91.68** ± 1.68	**0.916** ± 0.013	**0.921** ± 0.018	**93.63** ± 2.12	**0.933** ± 0.023	**0.931** ± 0.022
CAM-CLD: Math task (three classes)									
MTS-CNN$^{\Delta}$	85.16 ± 3.18	0.854 ± 0.035	0.852 ± 0.036	86.28 ± 3.13	0.867 ± 0.034	0.864 ± 0.032	86.99 ± 2.35	0.868 ± 0.021	0.865 ± 0.026
DecNet$^{\Delta}$	85.59 ± 2.75	0.858 ± 0.025	0.853 ± 0.026	86.64 ± 2.92	0.863 ± 0.027	0.863 ± 0.029	87.37 ± 2.51	0.870 ± 0.025	0.872 ± 0.024
CNN-LSTM$^{\Delta}$	86.86 ± 2.27	0.864 ± 0.022	0.862 ± 0.021	87.59 ± 2.23	0.879 ± 0.021	0.877 ± 0.023	87.98 ± 3.12	0.879 ± 0.034	0.872 ± 0.037
m-HyperLSTM$^{\Delta}$	87.99 ± 2.36	0.872 ± 0.025	0.876 ± 0.023	87.87 ± 2.13	0.875 ± 0.023	0.879 ± 0.019	88.25 ± 2.55	0.888 ± 0.022	0.885 ± 0.027
ARecNet$^{\Phi}$	88.28 ± 2.16	0.889 ± 0.020	0.884 ± 0.022	88.17 ± 3.37	0.883 ± 0.039	0.882 ± 0.032	89.23 ± 2.13	0.894 ± 0.025	0.901 ± 0.022
CogFormer (ours)$^{\Phi}$	**91.13** ± 1.31	**0.909** ± 0.012	**0.910** ± 0.015	**90.64** ± 2.43	**0.902** ± 0.023	**0.904** ± 0.022	**91.92** ± 2.36	**0.919** ± 0.022	**0.921** ± 0.026
CAM-CLD: n-back task (four classes)									
MTS-CNN$^{\Delta}$	84.35 ± 3.08	0.841 ± 0.032	0.842 ± 0.031	84.93 ± 2.55	0.846 ± 0.026	0.845 ± 0.023	85.33 ± 2.14	0.856 ± 0.023	0.852 ± 0.021
DecNet$^{\Delta}$	84.26 ± 2.09	0.848 ± 0.021	0.845 ± 0.019	85.01 ± 2.58	0.855 ± 0.025	0.851 ± 0.026	85.43 ± 2.33	0.851 ± 0.024	0.856 ± 0.024
CNN-LSTM$^{\Delta}$	85.88 ± 2.94	0.859 ± 0.029	0.850 ± 0.030	85.12 ± 2.64	0.858 ± 0.024	0.852 ± 0.025	85.36 ± 2.19	0.858 ± 0.023	0.851 ± 0.021
m-HyperLSTM$^{\Delta}$	85.85 ± 2.57	0.854 ± 0.025	0.857 ± 0.026	84.47 ± 2.26	0.845 ± 0.023	0.841 ± 0.021	86.17 ± 2.21	0.864 ± 0.024	0.867 ± 0.026
ARecNet$^{\Phi}$	86.17 ± 1.52	0.862 ± 0.016	0.869 ± 0.014	85.13 ± 1.54	0.858 ± 0.016	0.852 ± 0.018	87.42 ± 1.74	0.872 ± 0.017	0.870 ± 0.016
CogFormer (ours)$^{\Phi}$	**88.72** ± 1.78	**0.890** ± 0.019	**0.892** ± 0.018	**87.67** ± 1.44	**0.878** ± 0.014	**0.872** ± 0.015	**88.14** ± 1.88	**0.882** ± 0.018	**0.885** ± 0.017

Notes: In this table and the following tables, Δ means feature-level fusion model, Φ means decision-level fusion model and bold texts are the best model.

Table 8. Results on across-subjects data partition protocol.

Model	$t_w = 1\ s$			$t_w = 3\ s$			$t_w = 5\ s$		
	Accuracy (%)	F1-scores	AUC	Accuracy (%)	F1-scores	AUC	Accuracy (%)	F1-scores	AUC
MADT-D: Math task (two classes)									
MTS-CNNΔ	53.63 ± 5.29	0.547 ± 0.069	0.534 ± 0.045	51.43 ± 4.53	0.527 ± 0.043	0.516 ± 0.040	51.93 ± 4.77	0.524 ± 0.061	0.512 ± 0.057
DecNetΔ	52.79 ± 2.96	0.536 ± 0.035	0.514 ± 0.021	50.89 ± 2.23	0.518 ± 0.029	0.496 ± 0.015	50.69 ± 2.11	0.514 ± 0.027	0.491 ± 0.012
CNN-LSTMΔ	54.23 ± 5.98	0.557 ± 0.046	0.558 ± 0.055	52.43 ± 5.12	0.536 ± 0.057	0.539 ± 0.058	51.93 ± 5.22	0.539 ± 0.042	0.536 ± 0.057
m-HyperLSTMΔ	56.70 ± 6.80	0.579 ± 0.068	0.545 ± 0.060	57.22 ± 7.10	0.411 ± 0.021	0.549 ± 0.033	55.46 ± 6.57	0.629 ± 0.084	0.555 ± 0.066
ARecNetΦ	59.09 ± 5.99	0.582 ± 0.051	0.598 ± 0.057	58.19 ± 4.89	0.588 ± 0.047	0.587 ± 0.056	60.59 ± 4.14	0.605 ± 0.037	0.616 ± 0.042
CogFormer (ours)Φ	**64.93** ± 5.39	**0.669** ± 0.074	**0.677** ± 0.051	**63.06** ± 5.90	**0.628** ± 0.046	**0.662** ± 0.075	**65.71** ± 3.12	**0.656** ± 0.040	**0.683** ± 0.091
CAM-CLD: Spacial task (two classes)									
MTS-CNNΔ	55.33 ± 6.19	0.562 ± 0.078	0.550 ± 0.054	56.03 ± 6.09	0.567 ± 0.075	0.549 ± 0.052	55.83 ± 6.05	0.563 ± 0.063	0.547 ± 0.054
DecNetΔ	54.69 ± 3.51	0.556 ± 0.044	0.528 ± 0.039	54.49 ± 3.56	0.554 ± 0.037	0.531 ± 0.036	54.29 ± 3.53	0.551 ± 0.041	0.530 ± 0.031
CNN-LSTMΔ	56.53 ± 6.72	0.576 ± 0.055	0.576 ± 0.054	56.23 ± 6.78	0.574 ± 0.051	0.572 ± 0.045	55.93 ± 5.74	0.572 ± 0.053	0.568 ± 0.046
m-HyperLSTMΔ	56.61 ± 3.22	0.604 ± 0.042	0.610 ± 0.036	56.74 ± 3.24	0.580 ± 0.069	0.591 ± 0.049	56.94 ± 3.72	0.585 ± 0.033	0.603 ± 0.031
ARecNetΦ	60.28 ± 3.53	0.633 ± 0.033	0.665 ± 0.030	59.29 ± 3.37	0.615 ± 0.028	0.601 ± 0.021	61.27 ±3.50	0.628 ± 0.030	0.638 ± 0.037
CogFormer (ours)Φ	**65.90** ± 4.36	**0.658** ± 0.064	**0.681** ± 0.045	**65.10** ± 7.61	**0.638** ± 0.051	**0.672** ± 0.057	**66.59** ± 3.64	**0.631** ± 0.035	**0.708** ± 0.036
CAM-CLD: Math task (three classes)									
MTS-CNNΔ	53.83 ± 6.95	0.543 ± 0.060	0.529 ± 0.073	54.03 ± 7.12	0.537 ± 0.062	0.522 ± 0.069	54.18 ± 5.30	0.540 ± 0.058	0.526 ± 0.065
DecNetΔ	53.19 ± 3.16	0.539 ± 0.034	0.515 ± 0.038	52.49 ± 2.98	0.533 ± 0.031	0.513 ± 0.036	53.69 ± 5.05	0.541 ± 0.050	0.518 ± 0.059
CNN-LSTMΔ	55.03 ± 7.42	0.561 ± 0.073	0.556 ± 0.074	54.23 ± 5.88	0.557 ± 0.069	0.553 ± 0.072	54.53 ± 6.90	0.559 ± 0.071	0.550 ± 0.070
m-HyperLSTMΔ	56.95 ± 5.74	0.570 ± 0.052	0.629 ± 0.060	56.37 ± 2.66	0.560 ± 0.032	0.619 ± 0.036	58.81 ± 5.34	0.577 ± 0.059	0.601 ± 0.054
ARecNetΦ	57.78 ± 4.23	0.618 ± 0.051	0.580 ± 0.049	57.53 ± 4.90	0.620 ± 0.053	0.575 ± 0.050	59.34 ± 4.58	0.614 ± 0.045	0.622 ± 0.048
CogFormer (ours)Φ	**62.92** ± 4.11	**0.629** ± 0.046	**0.650** ± 0.055	**63.14** ± 5.95	**0.630** ± 0.085	**0.652** ± 0.049	**63.21** ± 8.78	**0.633** ± 0.090	**0.661** ± 0.064
CAM-CLD: n-back task (four classes)									
MTS-CNNΔ	52.18 ± 7.15	0.528 ± 0.080	0.504 ± 0.091	51.33 ± 6.02	0.522 ± 0.078	0.509 ± 0.089	52.53 ± 6.85	0.530 ± 0.072	0.505 ± 0.069
DecNetΔ	51.74 ± 7.95	0.523 ± 0.082	0.492 ± 0.076	50.89 ± 7.12	0.520 ± 0.070	0.498 ± 0.078	51.19 ± 7.53	0.521 ± 0.075	0.490 ± 0.075
CNN-LSTMΔ	53.53 ± 8.32	0.546 ± 0.091	0.531 ± 0.092	52.78 ± 6.38	0.540 ± 0.063	0.536 ± 0.069	53.23 ± 8.82	0.543 ± 0.089	0.533 ± 0.090
m-HyperLSTMΔ	54.47 ± 5.54	0.576 ± 0.047	0.537 ± 0.050	54.11 ± 4.53	0.579 ± 0.049	0.534 ± 0.046	54.85 ± 4.98	0.568 ± 0.051	0.530 ± 0.044
ARecNetΦ	55.84 ± 5.07	0.587 ± 0.051	0.582 ± 0.045	56.97 ± 5.15	0.592 ± 0.051	0.562 ± 0.056	57.43 ± 4.02	0.603 ± 0.037	0.597 ± 0.039
CogFormer (ours)Φ	**60.12** ± 6.77	**0.602** ± 0.068	**0.617** ± 0.039	**61.10** ± 6.12	**0.614** ± 0.061	**0.609** ± 0.063	**61.21** ± 6.74	**0.615** ± 0.053	**0.628** ± 0.059

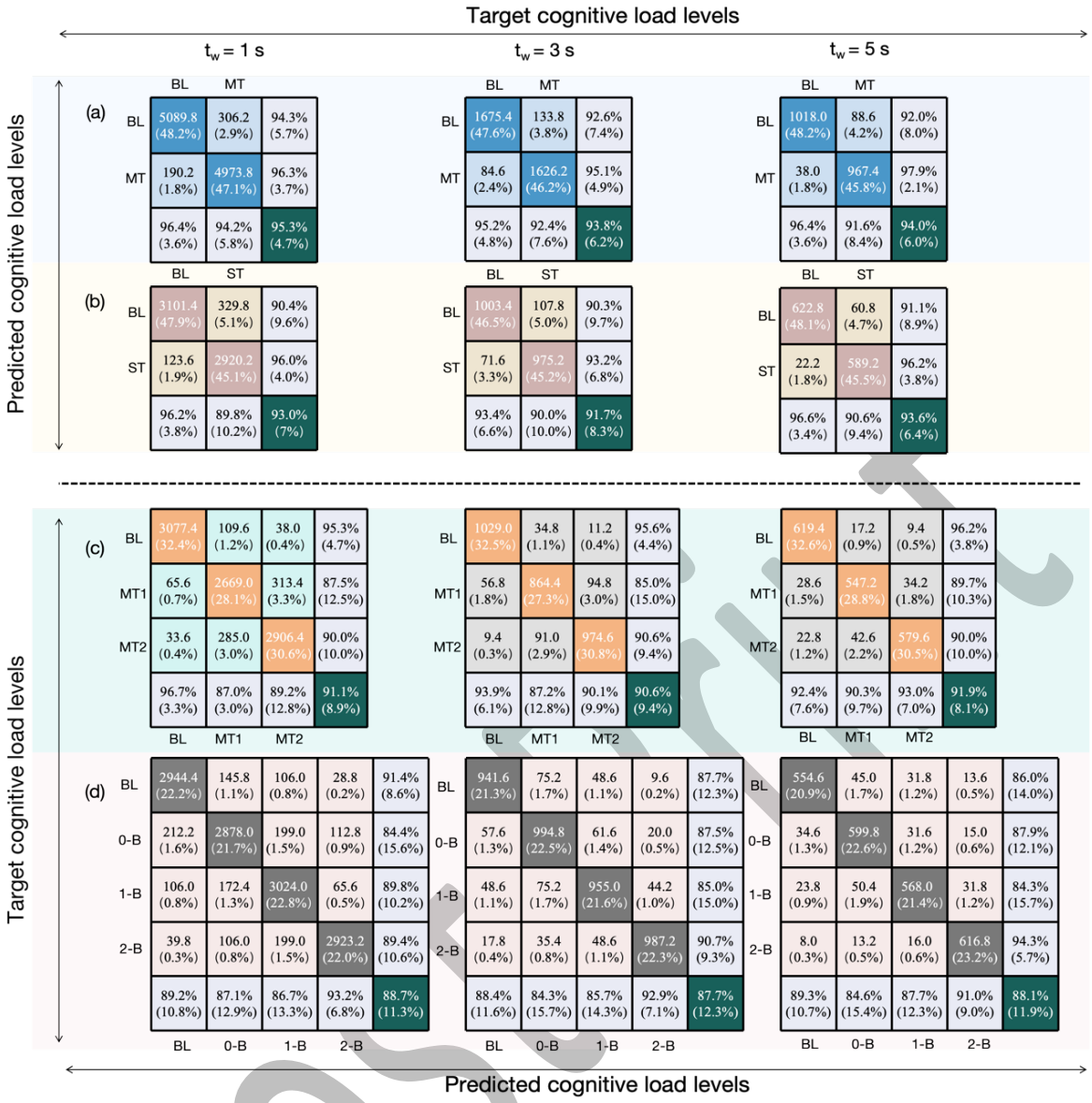


Figure 6a. Confusion matrices under within-subject data partition protocol with different time horizons: (a) MADT-D: Math task, (b) CAM-CLD: Cognitive spatial task, (c) CAM-CLD: Math task, and (d) CAM-CLD: N-back task.

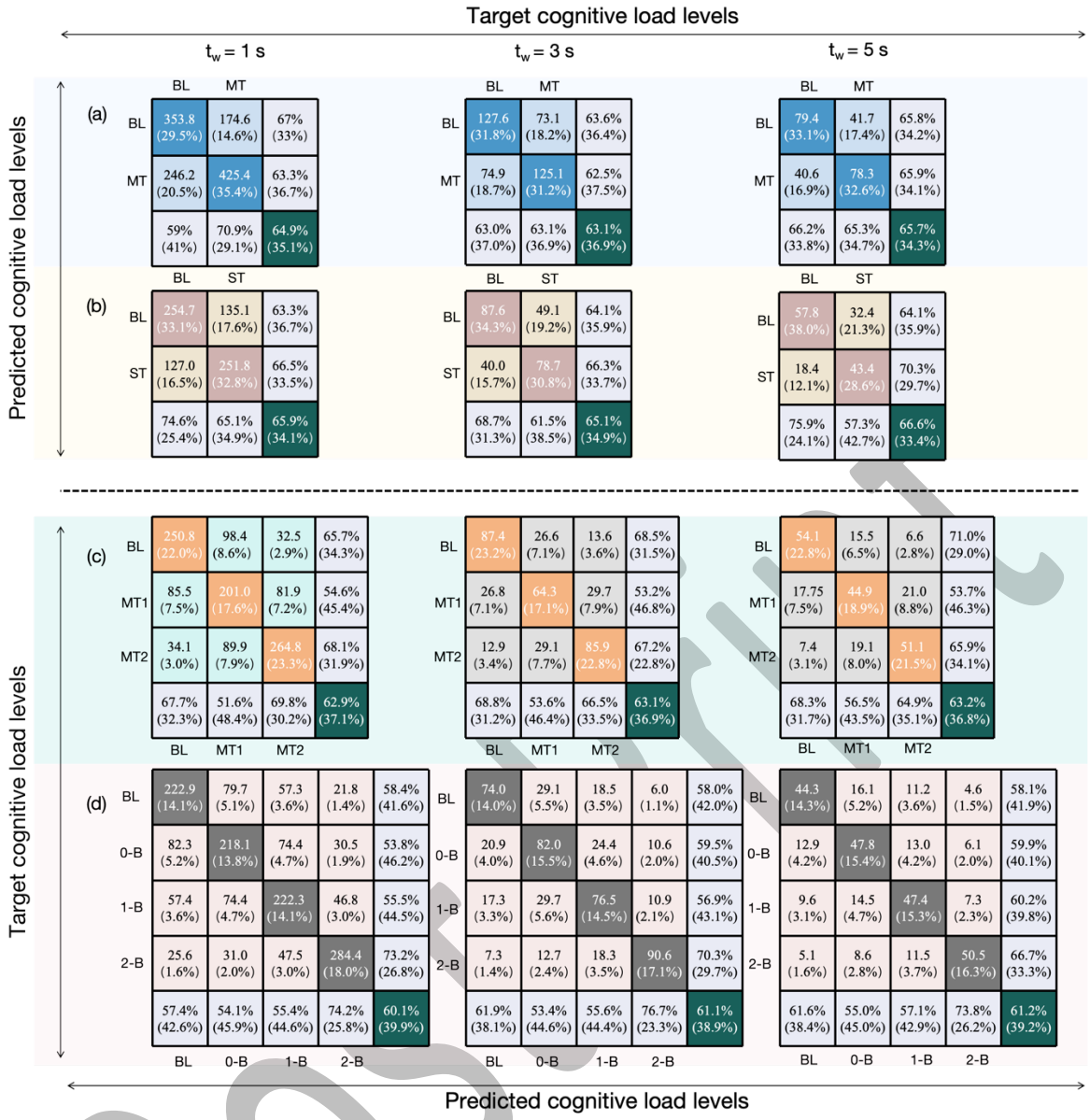


Figure 6b. Confusion matrices under across-subjects data partition protocol with different time horizons: (a) MADT-D: Math task, (b) CAM-CLD: Cognitive spatial task, (c) CAM-CLD: Math task, and (d) CAM-CLD: N-back task.

6.2. Ablation Experiments and Discussions

6.2.1 Input Ablation

Figure 7 describes the recognition accuracy and standard deviation of CogFormer across various cognitive load tasks under different combinations of physiological signals and historical horizons. The results clearly indicate that CogFormer, when integrating all physiological signals, significantly outperformed the models with part of the physiological

signals. This finding is in line with previous studies that indicated more channels of signals would boost the model performance (He et al., 2022).

At the same time, different tasks may have different preferences for signals. Specifically, in MADT-D, the accuracy of CogFormer without ECG signals (i.e., RESP+EDA) is lower than that of CogFormer including ECG data (i.e., ECG+EDA, ECG+RESP). In contrast, in CAM-CLD, including EDA always boosted the accuracy. This discrepancy in best signal combinations in different datasets demonstrates the challenges for determining the informative physiological signals for driver state estimation. Specifically, the best feature combination can hardly be determined before testing, as it may depend on the quality of signals and other unknown factors. Thus, a self-attention mechanism that can automatically select informative signals or features instead of handcrafted features should be preferred for real-world applications.

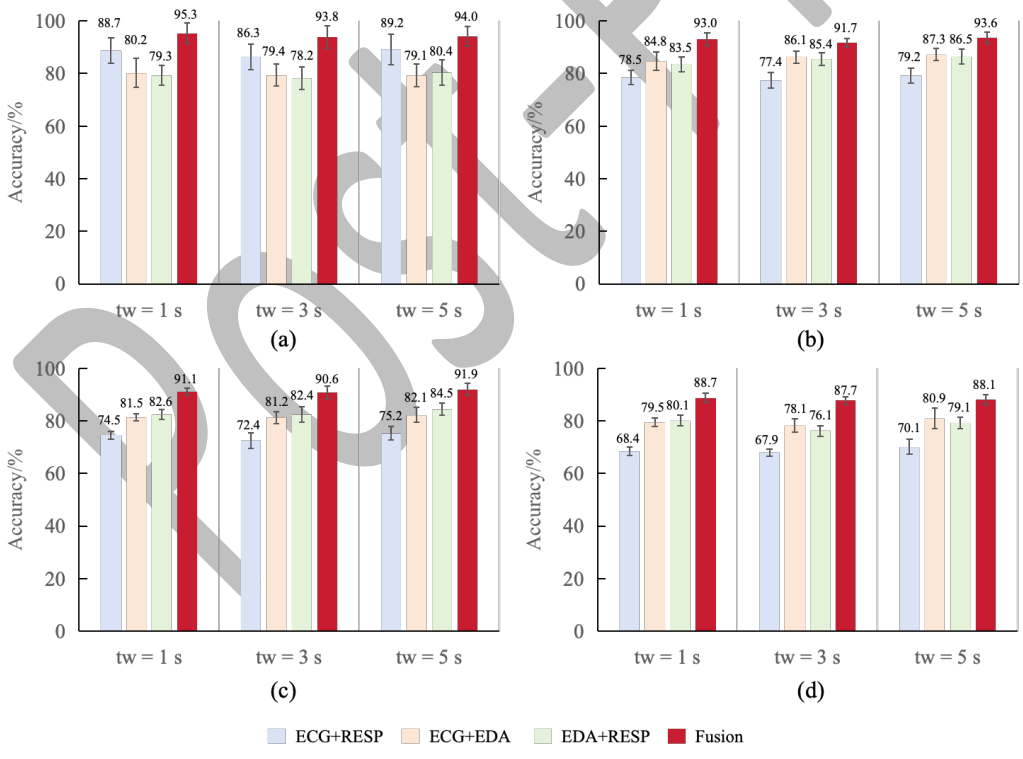


Figure 7. Accuracies of the CogFormer with different combinations of physiological signals and time horizons: (a) MADT-D: Math task, (b) CAM-CLD: Cognitive Spatial task, (c) CAM-CLD: Math task, and (d) CAM-CLD: N-back task.

6.2.2 Model Ablation

To identify the crucial components within CogFormer, specifically the Multi-stream Encoding and Aligned Attention modules, the model ablation tests were conducted. The accuracies of all model variants are summarized in Table 9.

Table 9. Ablation study with different time horizons on various cognitive tasks.

	N1	N2	N3
MADT-D: Math task			
$t_w = 1s$	91.41 (↓4.23%)	91.85 (↓**3.73%**)	**95.28**
$t_w = 3s$	90.47 (↓**3.67%**)	90.64 (↓3.92%)	**93.79**
$t_w = 5s$	91.58 (↓**2.68%**)	91.06 (↓3.26%)	**94.03**
CAM-CLD: Spatial task			
$t_w = 1s$	88.51 (↓5.13%)	89.23 (↓**4.28%**)	**93.05**
$t_w = 3s$	88.03 (↓4.15%)	88.34 (↓**3.78%**)	**91.68**
$t_w = 5s$	90.61 (↓**3.33%**)	90.50 (↓3.46%)	**93.63**
CAM-CLD: Math task			
$t_w = 1s$	87.70 (↓3.91%)	87.87 (↓**3.71%**)	**91.13**
$t_w = 3s$	85.99 (↓5.41%)	86.99 (↓**4.20%**)	**90.64**
$t_w = 5s$	88.10 (↓4.34%)	88.21 (↓**4.21%**)	**91.92**
CAM-CLD: N-back task			
$t_w = 1s$	82.96 (↓6.94%)	84.18 (↓**5.39%**)	**88.72**
$t_w = 3s$	82.75 (↓5.68%)	83.66 (↓**4.79%**)	**87.67**
$t_w = 5s$	84.91 (↓**3.80%**)	84.93 (↓4.23%)	**88.14**

Notes: Network 1 (N1) represents the model without the Multi-stream Encoding and aligned attention modules. Network 2 (N2) includes Multi-stream Encoding but lacks the aligned attention module. Network 3 (N3) contains both the Multi-stream Encoding and aligned attention modules. The values in parentheses indicate the decrease in accuracy compared to CogFormer under different module ablation scenarios; the bold font represents the case with the smallest decrease in accuracy among the two variants.

It can be observed that the Aligned Attention module consistently improved performance across all tests, underscoring its enhanced ability to capture informative features compared to conventional transformer models. Additionally, the variation with the Multi-stream Encoding module outperformed the variation without it in scenarios with short time horizons; however, with an increase in the time horizon, the benefits started to diminish, indicating that the Multi-stream Encoding module can extract useful multimodal information in long sequences; but with short time horizon, the benefit of it may have been shadowed by increased model complexity.

Additionally, as showen in Figure 8, we conducted paired t-tests to assess the performance difference between the network without the Aligned Attention module and Multi-stream Encoding (N1) and the CogFormer (N3). The statistical results are shown below, where a single asterisk (*) and double asterisks (**) denote p-values below 0.05 and 0.01, respectively. Notably, for time window lengths of 1 second and 3 seconds, N3 significantly outperformed N1 across three datasets. For time windows of 3 seconds and 5 seconds, N3 significantly outperformed N1 on one dataset.

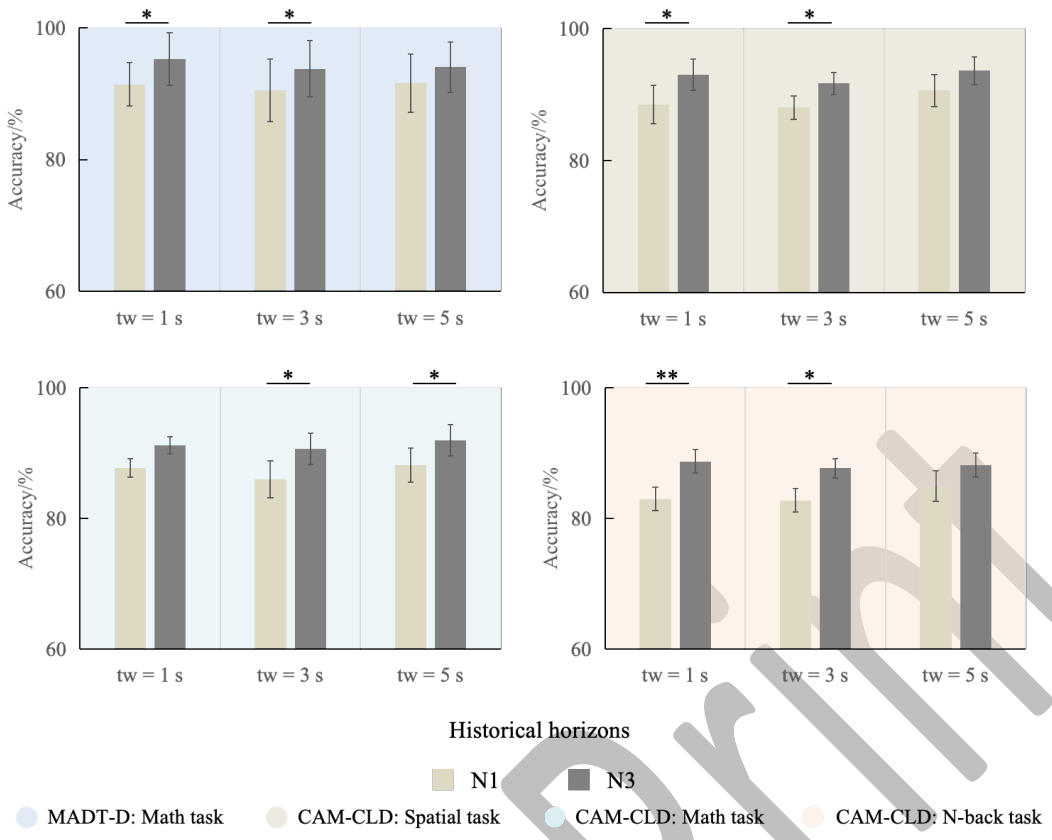


Figure 8. Statistical test results between N1 and N3.

6.2.3 Robustness Testing

In practical applications, time series signals are inevitably influenced by missing values and Gaussian White Noise (GWN), particularly when sourced from multiple information channels [19]. Consequently, monitoring cognitive load levels necessitates network robustness to information distortion (Yang et al., 2023; Zhu et al., 2024). Thus, robustness testing was conducted. Following Yang et al., (2023), we intentionally added 10% or 20% missing multimodal physiological information ($\mathcal{M}$), 30% missing ECG signals (α), RESP (β)), or EDA (γ), GWN with standard deviations σ = 0.1 or 0.2, or mixed information distortion comprising 20% missing multimodal physiological data and GWN σ = 0.2 in the test dataset. It should be noted that the missing values were substituted with the values from previous time steps. Table 10 shows the estimation results of CogFormer with a time horizon of 5 seconds under various distortion conditions.

Table 10. Comparison of estimation accuracy of the decision-fusion models with varied information distortions when t_w=5s.

Model	Normal (%)	Missing 10% ($\mathcal{M}$)	Missing 20% ($\mathcal{M}$)	Missing 30% (α)	Missing 30% (β)	Missing 30% (γ)	GWN σ=0.1	GWN σ=0.2	Mixed distortion
MADT-D: Math task									
CogFormer	94.03	92.80 (↓1.33%)	91.36 (↓2.92%)	92.39 (↓1.78%)	92.70 (↓1.43%)	92.74 (↓1.39%)	93.61 (↓0.45%)	92.30 (↓1.87%)	90.80 (↓**3.56%**)
ARecNet	92.56	90.81 (↓1.93%)	88.75 (↓4.29%)	90.20 (↓2.62%)	90.35 (↓2.45%)	91.18 (↓1.51%)	91.85 (↓0.77%)	90.00 (↓2.84%)	87.14 (↓6.22%)
CAM-CLD: Spacial task									
CogFormer	93.63	91.87 (↓1.92%)	91.51 (↓2.32%)	91.86 (↓1.93%)	92.39 (↓1.34%)	91.60 (↓2.21%)	93.27 (↓0.39%)	91.18 (↓2.69%)	89.61 (↓**4.49%**)
ARecNet	91.49	89.36 (↓2.38%)	87.98 (↓3.99%)	89.52 (↓2.20%)	89.46 (↓1.15%)	88.94 (↓2.87%)	90.82 (↓0.74%)	88.33 (↓3.58%)	86.87 (↓5.32%)
CAM-CLD: Math task									
CogFormer	91.92	90.29 (↓1.81%)	88.81 (↓3.50%)	90.18 (↓1.93%)	90.79 (↓1.24%)	89.77 (↓2.40%)	91.29 (↓0.69%)	89.76 (↓2.41%)	86.39 (↓6.40%)
ARecNet	89.23	87.71 (↓1.73%)	86.21 (↓3.50%)	86.78 (↓1.67%)	87.31 (↓2.20%)	86.77 (↓2.84%)	88.88 (↓0.39%)	86.47 (↓2.04%)	83.99 (↓**6.24%**)
CAM-CLD: n-back task									
CogFormer	88.14	85.50 (↓3.09%)	83.71 (↓5.29%)	85.83 (↓2.69%)	86.12 (↓2.35%)	86.00 (↓2.49%)	87.22 (↓1.05%)	84.80 (↓3.94%)	81.73 (↓**7.84%**)
ARecNet	87.42	83.89 (↓4.21%)	82.67 (↓5.75%)	85.25 (↓2.55%)	85.89 (↓1.78%)	85.22 (↓2.58%)	86.09 (↓1.54%)	83.41 (↓4.81%)	80.20 (↓9.00%)

Notes: The values in parentheses represent the decrease in accuracy compared to the normal situation; the bold font indicates the model with the smallest decrease in accuracy under mixed missing data.

As can be observed from Table 8, under mixed distortion conditions, for most tasks (other than the math task in the CAM-CLD dataset), the accuracy drop of CogFormer was smaller than that of ARecNet. In other words, our proposed model shows greater robustness compared

to the best baseline model, potentially attributing to the parallel transformer architecture with a coherent attention mechanism in the model.

7. Limitations

It should still be noted that our model focuses on cognitive load estimation without differentiating the sources of the cognitive load. Though integrating the outputs of other computer vision models (Craye & Karray, 2015; Doniec et al., 2020) and our models may resolve this issue, future research may better address this limitation by considering more types of information (such as drivers' visual behaviours and traffic context information) in end-to-end models. Further, our model is still not good at handling across-subjects scenarios. Future research should explore how to improve the generalizability of the models to make them practically more feasible.

8. Overall Discussion and Conclusions

Being different from previous approaches that utilized traditional machine learning models, we developed a decision-level multi-physiological information fusion architecture to extract temporal (i.e., chronological feature inputs at multiple time steps) and spatial information (i.e., parallel features being input into the model at a single time step) from multiple physiological signals. Experimental results demonstrate that the proposed CogFormer surpassed other baseline models in terms of estimation accuracy and robustness. In addition, based on the model ablation study and robustness test, we find:

- The preferred feature combinations for driver state estimation may depend on the type of targeted tasks, data collection quality, and driving context. Thus, the performance of models based on handcrafted features and manual feature selection may not be guaranteed in real-world applications.

- A longer time horizon, though can provide richer information, may not necessarily increase the model performance, potentially because the models may not be able to capture the complex features in the data. Thus, the degree of matching between the models and the characteristics of data should be considered when designing driver state-monitoring algorithms.
- All models, including our proposed model, were highly susceptible to individual differences – the performance of all models dropped significantly when across-subjects data partition was applied, indicating that the models, even with the attention mechanism, may still not be able to capture the individual-invariant features of high cognitive load states. Future research should design specific algorithms and structures to handle this issue (e.g., (Wang et al., 2024))
- Though our algorithm was not designed specifically to handle the noise and data distortion, our model showed better robustness compared to the best baseline model. The noise and missing data are common in real-world applications. Thus, future research should consider a more specific algorithm design (e.g., (Kieu et al., 2015; Middleton, 1999; Yang et al., 2025; Yang et al., 2023)) and validate the proposed model based on real-world datasets. Further, we focused on driver cognitive estimation out of takeover events. Future research may consider driver cognitive estimation during the takeover events to guide the adaptive strategies in the control transferring process.

CRediT authorship contribution statement

Ange WANG: Methodology, Software, Validation, Formal analysis, Writing – original draft. Haohan YANG: Formal analysis, Conceptualization, Writing – review & editing. Jiyao WANG: Formal analysis, Writing – review & editing. Hai YANG: Writing – review & editing, Supervision. Dengbo HE: Methodology, Writing – review & editing, Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province of China (2024A1515010392), and partially by the National Natural Science Foundation of China (grant no. 52202425), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011) and the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

References

Amadori, P. V., Fischer, T., Wang, R., & Demiris, Y. (2022). Predicting Secondary Task Performance: A Directly Actionable Metric for Cognitive Overload Detection. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(4), 1474–1485. IEEE Transactions on Cognitive and Developmental Systems. https://doi.org/10.1109/TCDS.2021.3114162

Angkan, P., Behinaein, B., Mahmud, Z., Bhatti, A., Rodenburg, D., Hungler, P., & Etemad, A. (2024). Multimodal Brain–Computer Interface for In-Vehicle Driver Cognitive Load Measurement: Dataset and Baselines. *IEEE Transactions on Intelligent Transportation Systems*, 1–16. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2023.3345846

Ansari, S., Naghdy, F., Du, H., & Pahnwar, Y. (2022). Driver Mental Fatigue Detection Based on Head Posture Using New Modified reLU-BiLSTM Deep Neural Network. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, *23*(8), 10957–10969. https://doi.org/10.1109/TITS.2021.3098309

Babiloni, F. (2019). Mental Workload Monitoring: New Perspectives from Neuroscience. In L. Longo & M. C. Leva (Eds.), *Human Mental Workload: Models and Applications* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-32423-0_1

Barua, S., Ahmed, M. U., & Begum, S. (2020). Towards Intelligent Data Analytics: A Case Study in Driver Cognitive Load Classification. *Brain Sciences*, *10*(8), Article 8. https://doi.org/10.3390/brainsci10080526

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Chai, R., Naik, G. R., Nguyen, T. N., Ling, S. H., Tran, Y., Craig, A., & Nguyen, H. T. (2017). Driver Fatigue Classification With Independent Component by Entropy Rate Bound Minimization Analysis in an EEG-Based System. *IEEE Journal of Biomedical and Health Informatics*, *21*(3), 715–724. https://doi.org/10.1109/JBHI.2016.2532354

Craye, C., & Karray, F. (2015). *Driver distraction detection and recognition using RGB-D sensor* (arXiv:1502.00250). arXiv. https://doi.org/10.48550/arXiv.1502.00250

Dengbo He, Kanaan, D., & Donmez, B. (n.d.). A Taxonomy of Strategies For Supporting Time-Sharing With Non-Driving Tasks in Automated Driving—Dengbo He, Dina Kanaan, Birsen Donmez, 2019. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*, 2088–2092. https://doi.org/10.1177/1071181319631

Doniec, R. J., Sieciński, S., Duraj, K. M., Piaseczna, N. J., Mocny-Pachońska, K., & Tkacz, E. J. (2020). Recognition of Drivers' Activity Based on 1D Convolutional Neural Network. *Electronics*, *9*(12), Article 12. https://doi.org/10.3390/electronics9122002

Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation Research Part C: Emerging Technologies*, *112*, 78–87. https://doi.org/10.1016/j.trc.2020.01.006

Du, N., Zhou, F., Tilbury, D. M., Robert, L. P., & Jessie Yang, X. (2024). Behavioral and physiological responses to takeovers in different scenarios during conditionally automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *101*, 320–331. https://doi.org/10.1016/j.trf.2024.01.008

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*(2), 572–576. https://doi.org/10.1016/j.neuroimage.2009.10.092

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., & Zuo, S. (2019). EEG-Based Spatio–Temporal Convolutional Neural Network for Driver Fatigue Evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2755–2763. https://doi.org/10.1109/TNNLS.2018.2886414

Hajinoroozi, M., Mao, Z., Jung, T., Lin, C., & Huang, Y. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, *47*, 549–555. Scopus. https://doi.org/10.1016/j.image.2016.05.018

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention*, *39*(2), 372–379. https://doi.org/10.1016/j.aap.2006.08.013

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

He, D., & Donmez, B. (2019). Influence of Driving Experience on Distraction Engagement in Automated Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(9), 142–151. https://doi.org/10.1177/0361198119843476

He, D., Wang, Z., Khalil, E. B., Donmez, B., Qiao, G., & Kumar, S. (2022). Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(10), 670–681. https://doi.org/10.1177/03611981221090937

Huang, Chunxi, Xie, Weiyin, Huang, Qihao, Zhu, Yan, Cui, Dixiao, & He, Dengbo. (2024). Effect of Advanced Driver Assistance Systems on Fatigue Levels of Heavy Truck Drivers in Prolonged Driving Tasks. *Journal of Tongji University (Natural Science)*, *52*(6), 846–855.

Huang, J., Liu, Y., & Peng, X. (2022). Recognition of driver's mental workload based on physiological signals, a comparative study. *Biomedical Signal Processing and Control*, *71*, 103094. https://doi.org/10.1016/j.bspc.2021.103094

J3016—Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. (2018). SAE International J3016_201806 (2018).

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics*, *52*(2), 154–164. https://doi.org/10.1080/00140130802030714

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412. https://doi.org/10.1080/09658211003702171

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., & Fukasawa, K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology*, *117*(7), 1574–1581. https://doi.org/10.1016/j.clinph.2006.03.011

Kieu, L.-M., Bhaskar, A., & Chung, E. (2015). A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data. *Transportation Research Part C: Emerging Technologies*, *58*, 193–207. https://doi.org/10.1016/j.trc.2015.03.033

Kose, N., Kopuklu, O., Unnervik, A., & Rigoll, G. (2019). Real-Time Driver State Monitoring Using a CNN Based Spatio-Temporal Approach. In *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019* (pp. 3236–3242). https://doi.org/10.1109/ITSC.2019.8917460

Kumar, S., He, D., Qiao, G., & Donmez, B. (2022). Classification of Driver Cognitive Load based on Physiological Data: Exploring Recurrent Neural Networks. *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*, 19–24. https://doi.org/10.1109/ICARM54641.2022.9959588

Liang, Y., & Lee, J. D. (2010). Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis & Prevention*, *42*(3), 881–890. https://doi.org/10.1016/j.aap.2009.05.001

Mehler, B. (2012). *Establishing the Sensitivity of Physiological Measures of Cognitive Load in the Driving Environment*. https://doi.org/10.13140/2.1.2471.5521

Melnicuk, V., Thompson, S., Jennings, P., & Birrell, S. (2021). Effect of cognitive load on drivers' State and task performance during automated driving: Introducing a novel method for determining stabilisation time following take-over of control. *Accident Analysis and Prevention*, *151*, 105967. Scopus. https://doi.org/10.1016/j.aap.2020.105967

Meteier, Q., Capallera, M., de Salis, E., Angelini, L., Carrino, S., Widmer, M., Abou Khaled, O., Mugellini, E., & Sonderegger, A. (2023). A dataset on the physiological state and behavior of drivers in conditionally automated driving. *Data in Brief*, *47*. Scopus. https://doi.org/10.1016/j.dib.2023.109027

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., Widmer, M., & Sonderegger, A. (2021). Classification of Drivers' Workload Using Physiological Signals in Conditional Automation. *Frontiers in Psychology*, *12*. Scopus. https://doi.org/10.3389/fpsyg.2021.596038

Middleton, D. (1999). Non-Gaussian noise models in signal processing for telecommunications: New methods an results for class A and class B noise models. *IEEE Transactions on Information Theory*, *45*(4), 1129–1149. IEEE Transactions on Information Theory. https://doi.org/10.1109/18.761256

Muhrer, E., & Vollrath, M. (2011). The effect of visual and cognitive distraction on driver's anticipation in a simulated car following scenario. *Transportation Research Part F: Psychology and Behaviour*, *14*(6), 555–566. https://doi.org/10.1016/j.trf.2011.06.003

Prabhakar, G., Mukhopadhyay, A., Murthy, L., Modiksha, M., Sachin, D., & Biswas, P. (2020). Cognitive load estimation using ocular parameters in automotive. *Transportation Engineering*, *2*, 100008. https://doi.org/10.1016/j.treng.2020.100008

Qu, Y., Hu, H., Liu, J., Zhang, Z., Li, Y., & Ge, X. (2023). Driver State Monitoring Technology for Conditionally Automated Vehicles: Review and Future Prospects. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–20. https://doi.org/10.1109/TIM.2023.3301060

Rahman, H., Ahmed, M. U., Barua, S., & Begum, S. (2020). Non-contact-based driver's cognitive load classification using physiological and vehicular parameters. *Biomedical Signal Processing and Control*, *55*, 101634. https://doi.org/10.1016/j.bspc.2019.101634

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*(1), 31–43. https://doi.org/10.1037/1076-898X.6.1.31

Shi, W., Wang, Z., Wang, A., & He, D. (2024). Classification of Driver Cognitive Load in Conditionally Automated Driving: Utilizing Electrocardiogram-Based Spectrogram with Lightweight Neural Network. *Transportation Research Record*. https://doi.org/10.1177/03611981241252797

Singh, S. (2015). Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. *Traffic Safety Facts - Crash Stats*, Article DOT HS 812 115. https://trid.trb.org/view.aspx?id=1346216&source=post_page----------

Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving.

Transportation Research Part F: Traffic Psychology and Behaviour, *60*, 590–605. https://doi.org/10.1016/j.trf.2018.11.006

Sun, W., Aguirre, M., Jin, J. (Judy), Feng, F., Rajab, S., Saigusa, S., Dsa, J., & Bao, S. (2021). Online distraction detection for naturalistic driving dataset using kinematic motion models and a multiple model algorithm. *Transportation Research Part C: Emerging Technologies*, *130*, 103317. https://doi.org/10.1016/j.trc.2021.103317

Tjolleng, A., Jung, K., Hong, W., Lee, W., Lee, B., You, H., Son, J., & Park, S. (2017). Classification of a Driver's cognitive workload levels using artificial neural network on ECG signals. *Applied Ergonomics*, *59*, 326–332. https://doi.org/10.1016/j.apergo.2016.09.013

Wang, A., Huang, C., Wang, J., & He, D. (2024). The association between physiological and eye-tracking metrics and cognitive load in drivers: A *meta*-analysis. *Transportation Research Part F: Traffic Psychology and Behaviour*, *104*, 474–487. https://doi.org/10.1016/j.trf.2024.06.014

Wang, A., Wang, J., Shi, W., & He, D. (2024). Cognitive Workload Estimation in Conditionally Automated Vehicles Using Transformer Networks Based on Physiological Signals. *Transportation Research Record*. https://doi.org/10.1177/03611981241250023

Wang, J., Wang, A., Hu, H., Wu, K., & He, D. (2024). Multi-Source Domain Generalization for ECG-Based Cognitive Load Estimation: Adversarial Invariant and Plausible Uncertainty Learning. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1631–1635. https://doi.org/10.1109/ICASSP48485.2024.10447676

Wang, R., Amadori, P. V., & Demiris, Y. (2018). Real-Time Workload Classification during Driving using HyperNetworks. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3060–3065. https://doi.org/10.1109/IROS.2018.8594305

Wang, X., Xu, R., Zhang, S., Zhuang, Y., & Wang, Y. (2022). Driver distraction detection based on vehicle dynamics using naturalistic driving data. *Transportation Research Part C: Emerging Technologies*, *136*, 103561. https://doi.org/10.1016/j.trc.2022.103561

Wei, W., Fu, X., Zhong, S., & Ge, H. (2023). Driver's mental workload classification using physiological, traffic flow and environmental factors. *Transportation Research Part F: Traffic Psychology and Behaviour*, *94*, 151–169. Scopus. https://doi.org/10.1016/j.trf.2023.02.004

Xie, Y., Murphey, Y. L., & Kochhar, D. S. (2020). Personalized Driver Workload Estimation Using Deep Neural Network Learning From Physiological and Vehicle Signals. *IEEE Transactions on Intelligent Vehicles*, *5*(3), 439–448. https://doi.org/10.1109/TIV.2019.2960946

Yang, H., Liu, H., Hu, Z., Nguyen, A.-T., Guerra, T.-M., & Lv, C. (2023). Quantitative Identification of Driver Distraction: A Weakly Supervised Contrastive Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1–12. https://doi.org/10.1109/TITS.2023.3316203

Yang, H., Wu, J., Hu, Z., & Lv, C. (2023). Real-Time Driver Cognitive Workload Recognition: Attention-Enabled Learning with Multimodal Information Fusion. *IEEE Transactions on Industrial Electronics*, 1–11. https://doi.org/10.1109/TIE.2023.3288182

Yang, H., Zhou, Y., Wu, J., Liu, H., Yang, L., & Lv, C. (2025). Human-Guided Continual Learning for Personalized Decision-Making of Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, 1–0. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2024.3524609

Yang, L., Yang, H., Hu, B.-B., Wang, Y., & Lv, C. (2023). A Robust Driver Emotion Recognition Method Based on High-Purity Feature Separation. *IEEE Transactions on Intelligent Transportation Systems*, *24*(12), 15092–15104. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2023.3304128

Yang, L., Yang, H., Wei, H., Hu, Z., & Lv, C. (2024). Video-Based Driver Drowsiness Detection With Optimised Utilization of Key Facial Features. *IEEE Transactions on Intelligent Transportation Systems*, 1–13. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2023.3346054

Yu, X., Chen, C.-H., & Yang, H. (2023). Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach. *Advanced Engineering Informatics*, *57*, 102123. https://doi.org/10.1016/j.aei.2023.102123

Zhang, S., Zhang, Z., Chen, Z., Lin, S., & Xie, Z. (2021). A novel method of mental fatigue detection based on CNN and LSTM. *International Journal of Computational Science and Engineering*, *24*(3), 290–300. https://doi.org/10.1504/IJCSE.2021.115656

Zhu, J., Lv, C., Ma, Y., Yang, H., & Zhang, Y. (2024). Quantitative Estimation of Driver Cognitive Workload: A Dual-Stage Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1–13. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2024.3451144