Preprint version; final version available at http://ieeexplore.ieee.org

# Predicting Secondary Task Performance: A Directly Actionable Metric for Cognitive Overload Detection

Pierluigi Vito Amadori, *Member, IEEE*, Tobias Fischer, *Member, IEEE*, Ruohan Wang, *Member, IEEE*, Yiannis Demiris, *Senior Member, IEEE*

***Abstract*—In this paper, we address cognitive overload detection from unobtrusive physiological signals for users in dual-tasking scenarios. Anticipating cognitive overload is a pivotal challenge in interactive cognitive systems and could lead to safer shared-control between users and assistance systems. Our framework builds on the assumption that decision mistakes on the cognitive secondary task of dual-tasking users correspond to cognitive overload events, wherein the cognitive resources required to perform the task exceed the ones available to the users. We propose DecNet, an end-to-end sequence-to-sequence deep learning model that infers in real-time the likelihood of user mistakes on the secondary task, i.e., the practical impact of cognitive overload, from eye-gaze and head-pose data. We train and test DecNet on a dataset collected in a simulated driving setup from a cohort of 20 users on two dual-tasking decision-making scenarios, with either visual or auditory decision stimuli. DecNet anticipates cognitive overload events in both scenarios and can perform in time-constrained scenarios, anticipating cognitive overload events up to 2s before they occur. We show that DecNet's performance gap between audio and visual scenarios is consistent with user perceived difficulty. This suggests that single modality stimulation induces higher cognitive load on users, hindering their decision-making abilities.**

***Index Terms*—Cognitive Workload, User Monitoring, Decision Anticipation, Simulated Driving.**

## I. Introduction

COGNITIVE load modeling has received significant research interests in recent years thanks to its wide range of applications spanning from human-robot interaction [1], to human-computer interaction [2], and intelligent vehicles [3], [4], [5]. Accurate inference of a user's cognitive state in real-time could lead to disruptive benefits towards optimized interface designs and adaptive user interfaces [4], [6], more effective and situational-aware robots [7], [8], as well as safer and smarter vehicles [9], [10], [11]. Modeling and inferring human cognitive states is also an inherently multidisciplinary task, where numerous fields, such as psychology [12], neuroscience [13], engineering and artificial intelligence [14], [15], overlap.

Despite the evident applications in human-robot interaction and intelligent vehicles, cognitive state inference still has not reached a level suitable for real-world applications, compared to other machine learning domains, e.g., computer vision [17], [18], [19] and natural language processing [20], [21]. Also, cognitive load inference does not provide a directly actionable feedback signal for real-world assistance systems, as humans may still perform well under stress [22]. Because of this, in this paper, we propose a paradigm shift from cognitive load inference towards cognitive overload detection. We identify as cognitive overload the instances in which the amount of cognitive resources required to perform a task exceed the ones currently available to a user, therefore leading to severe decrease in both task performance and safety.

Driving is a popular scenario for cognitive state modeling [10], [23], [15], as it is a highly cognitive demanding task that requires drivers to be constantly aware of the surrounding environment, while continuously making decisions and taking actions [24]. Also, the possible causes of cognitive overload and distraction in drivers are numerous [18], such as visual, e.g., eyes off-road due to the use of mobile phones or in-vehicle information systems, or auditory, e.g., mind off-road due to holding hand-free cellphone conversations or even e-mail systems. While visual distractions can have a clear and observable effect, e.g., the driver is not looking at the road, auditory/cognitive distractions have more subtle effects, e.g., driving performance degrade and hazard perception is hindered [18]. Thus, the design of systems that can infer cognitive distraction is critical to improve safety, albeit particularly challenging [25].

In this paper, we focus our attention on a dual-task human-in-the-loop simulated virtual reality (VR) driving scenario. Here, human participants are tasked to drive and avoid obstacles (primary task), while performing a cognitively demanding "n-back task" (secondary task) [26]. Leveraging on the known relationship between task performance and cognitive overload [27], we assume that mistakes on the secondary task correspond to cognitive overload instances in the participant. Differently from previous cognitive load inference methods [15], [25], the proposed approach provides a directly actionable and unambiguous feedback signal for assistance systems as it anticipates the practical effects of cognitive overload. Given the focus on cognitive overload detection in simulated VR driving, we use the terms user and driver interchangeably throughout the paper.

We investigate whether we can predict the practical impacts of cognitive overload events and distraction on secondary task decision making from unobtrusive physiological signals from the driver, namely eye gaze and head pose. To do this, we exploit the widespread availability of affordable and unobtrusive sensors and improvements in algorithms [28], [29] to collect

Manuscript received May 17, 2021; revised August 13, 2021; accepted September 5, 2021. This work was supported in part by UK DSTL/EPSRC Grant EP/P008461/1, and a Royal Academy of Engineering Chair in Emerging Technologies to Yiannis Demiris.

All authors are with the Personal Robotics Lab, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BT, U.K. (e-mail: {p.amadori, r.wang16, y.demiris}@imperial.ac.uk; info@tobiasfischer.info)

Research presented in this paper is a continuation of Amadori et al. [16].

physiological data from drivers. Our work may be interpreted as an extension to conventional cognitive load classification [4], wherein we demonstrate that predicting the correctness of cognitive state-dependent decisions is feasible.

The contributions of the paper are:

1) We present an end-to-end long short-term memory (LSTM; [30])-based model, namely DecNet, for anticipating cognitive overload events in humans. The proposed approach can reliably infer in real-time the likelihood of a user's mistake on an imminent secondary task decision;
2) We collect a dataset containing physiological and behavioral data from a cohort of twenty participants in a realistic driver-in-the-loop virtual reality simulation. Participants were instructed to drive along the road while avoiding obstacles (primary task) and to make cognitive-based decisions (secondary task) in two separate scenarios with visual and auditory decision stimuli, respectively;
3) We analyze DecNet's performance on these scenarios and investigate the effects that combined visual-auditory and visual-visual stimuli have on cognitive stimulation and decision in the driver;
4) We demonstrate that DecNet estimates that the task difficulty in the visual-visual scenario is higher than that of the visual-auditory scenario, which is in line with the task's perceived difficulty obtained from questionnaires, as well as several models of multitasking [31], [32].

The rest of the paper is organized as follows: Section II provides a detailed overview of related works. Section III formalizes the problem of decision anticipation and introduces the proposed model, DecNet, to solve it. Section IV explains in detail both the experimental protocol and the data collection/preprocessing procedure. Section V provides an in-depth presentation of the experimental setup used to evaluate and test DecNet performance on the collected dataset. Section VI analyzes the results and performance achieved by DecNet against both classic methods and comparable recurrent neural network architectures and investigates the impact of multimodal stimuli on driver performance. Finally, Section VII summarizes the contributions of the paper and its main limitations, and outlines future research directions.

## II. Related Works

Our focus on cognitive overload detection during simulated driving is closely related to cognitive load classification, human-machine interaction, and the role of secondary tasks during driving. This section overviews related literature.

### A. Cognitive Load Classification

Cognitive load classification is inherently a very complex task, as different cognitive load levels experienced by humans are not directly measurable [33]. When addressing cognitive load classification, sophisticated feature engineering is often required to improve data quality and extract useful features from raw sensor signals, e.g., [4], [10], [23], [25]. These studies have proven statistical correlations between cognitive load and physiological signals, although they can vary significantly among experiment participants.

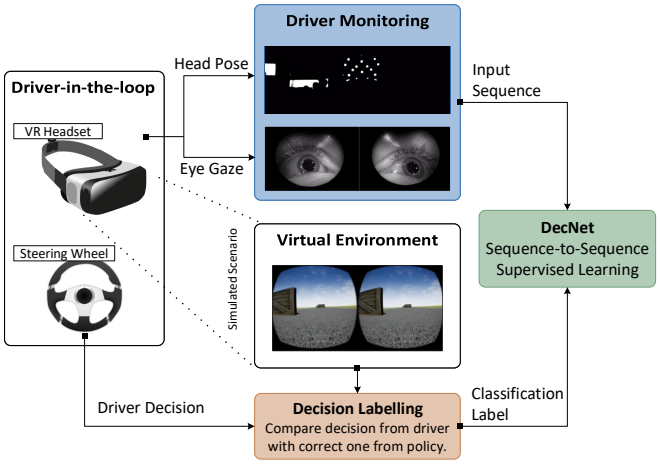



Fig. 1. Block diagram of DecNet framework for decision correctness anticipation. We frame the problem as sequence-to-sequence supervised learning. We use a Virtual Reality (VR) headset with integrated head pose and eye gaze tracking to monitor the user (blue). The simulated environment prompts the user to make decisions according to a specific policy. The correctness of the decisions identify the labels (orange), while head pose and eye gaze data represent the inputs of DecNet. The final goal of DecNet (green) is to anticipate the correctness likelihood of future secondary task decisions, which is indicative of events of cognitive overload on users.

Personalized models can tackle the problem, as seen in [34], [35], however such models may become impractical as data collection and model training are required for every new user. Toward this end, [15] introduced a novel end-to-end framework for real-time cognitive load classification, where the network is capable of learning useful feature representations directly from data.

### B. Gaze Patterns in Cognitive Human-Machine Interaction

Gaze patterns have been widely used in cognitive human-machine interaction. For example, [36] used gaze patterns to infer a user's level of domain knowledge in the domain of genomics, while [37] focused on knowledgeability prediction using a noninvasive eye-tracking method on mobile devices with Support Vector Machines (SVMs).

Gaze patterns also allow for the detection of cognitive-behavioral patterns [38] and internal thought (directing attention away from a primary visual task) [39] in intelligent user interfaces. Interestingly, it was also shown that a human's gaze is a requirement for perspective-taking in human-robot interactions, which allows a robot to infer the world's characteristics from the human's viewpoint [40], [41]. In a similar manner, studies have shown that there are clear correlations between gaze patterns and cognitive load [33], [42].

### C. Multitasking in Driving

Multitasking scenarios have been extensively employed by assistive and intelligent vehicles research communities [1], [43], [44], [45], [46]. In [43], authors have investigated the impact of secondary tasks on driving performance and showed that they lead to clear safety-related issues, such as off-road glances and unplanned lane deviations. Recently, [45], investigated how engaged are drivers on secondary tasks in vehicles with

different degrees of automation. In [46], authors investigated the role of secondary tasks on highly automated vehicles and studied how drivers regulate their resources to complete primary and secondary tasks and how they react during take-over requests. In [47], various secondary tasks were classified based on the EEG dynamics. While very good accuracy was achieved, EEG is intrusive and subject-dependent, whereas our method generalizes across subjects and is based on easy-to-access signals.

Our work is also related to that of Ersal *et al.* [48], who proposed a radial-basis neural network to predict the actions that a driver would have taken if there had not been a secondary task present. In [49], based on the finding that secondary tasks impact the driver's driving abilities, the take-over readiness of drivers was modeled by explicitly taking the secondary task into account. Also, Engström *et al.* [50] introduced a framework that predicts the effects of cognitive load on driver performance, and argued that secondary tasks hinder driving tasks that rely on cognitive control, while automatic performance is unaffected.

### D. Final Remarks

In the sections above, we have shown how the proposed study builds on past literature both for its experimental assumptions and in its model design. We use cognitive secondary tasks decisions as a proxy for cognitive overload instances, as numerous studies have shown that secondary tasks have direct effects on driver behavior, safety and cognitive states. Also, our choice for head pose and eye gaze as input signals to DecNet is supported by findings in cognitive human-machine interaction studies, which have shown that gaze patterns strongly correlate with human cognitive states.

Differently from cognitive load inference literature, we propose a paradigm shift from cognitive load classification towards cognitive overload detection via secondary task decision correctness anticipation. The proposed paradigm shift offers a directly actionable feedback metric that assistive systems can use to intervene and prevent the practical impacts of cognitive overload instances in users. Also, we propose a novel generalized user-agnostic model that can operate in real-time and that employs a sequence-to-sequence learning paradigm to encourage feature extraction from early observations.

## III. Proposed Solution

We frame cognitive overload detection as a supervised classification problem where the correctness of secondary task decisions is used as a label, as shown in Fig. 2. Specifically, we consider a dataset

$$\mathcal{D} : \left\{ \left( \mathbf{X}_j, y_j \right) \right\}_{j=1}^{N}, \tag{1}$$

where $\mathbf{X}_j$ is a temporal sequence of size $N_{\text{steps}}$, $y_j$ represents the corresponding label, and $N$ identifies the total number of samples. We omit time dependency on $\mathbf{X}_j$ to simplify notation. In our study, $\mathbf{X}_j$ denotes a sequence of physiological signals, and $y_j$ the correctness of the secondary task decision, as a proxy to cognitive overload events. Our supervised classification problem uses binary labels, which are derived by evaluating whether the secondary task decision performed by the driver is correct ($y_j = 1$) or wrong ($y_j = 0$). Please see Section IV for a detailed presentation on experimental procedure and data collection.

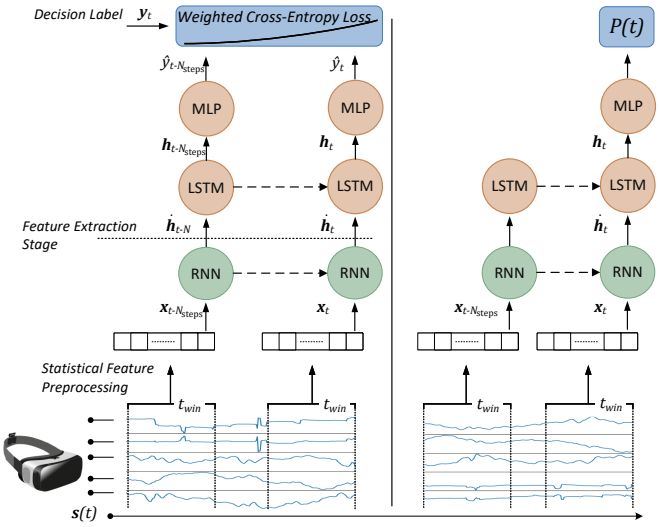



Fig. 2. Overview of DecNet for secondary task decision correctness anticipation during training (left) and inference (right). Sequential sensor readings from the driver $\mathbf{s}(t)$ are collected by a sliding window of length $t_{win}$ to extract the sequence input to DecNet, $\mathbf{X}_j$. The RNN stage first processes the input into a sequence of features/hidden states $\dot{\mathbf{h}}_t$ via Eq. (3), which is then used as input to the LSTM stage according to Eq. (4). Finally, the hidden states of the LSTM are projected to decision correctness likelihood via Eq. (9).

Since we framed cognitive overload detection as a classification problem, the final goal is to identify a model $p_\theta$ that minimizes the cross-entropy loss $L$ as

$$L = \sum_{j=1}^{N} - \log p_\theta(y_j | \mathbf{X}_j). \tag{2}$$

In this paper, we parametrize $p_\theta$ with DecNet, which comprises of a cascade of two sequential models: a recurrent neural network (RNN) and a long short-term memory network (LSTM) [30].

### A. Decision Anticipation Network (DecNet)

DecNet is a two-stage end-to-end sequential model that jointly learns to extract the most relevant features via an RNN module and to exploit them via an LSTM network in order to infer cognitive overload events by anticipating the correctness likelihood of an imminent decision, as shown in Fig. 2. In other words, the hidden states of the RNN, see Eq. (3), are used as input to the LSTM module. Finally, we project the hidden states of the LSTM stage with a multilayer perceptron (MLP) with a Rectified Linear Unit(ReLU) nonlinearity, followed by a softmax layer to predict the decision correctness probability.

Given a sequence of observations $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_{\text{steps}}})$ where $\mathbf{x}_t \in \mathbb{R}^{N_x}, \forall t$, the initial RNN stage operates as

$$\dot{\mathbf{h}}_t = \tanh(\mathbf{W}^{rnn}\mathbf{x}_t + \mathbf{H}^{rnn}\dot{\mathbf{h}}_{t-1} + \mathbf{b}^{rnn}), \tag{3}$$

where $\dot{\mathbf{h}}_t \in \mathbb{R}^{N_{rnn}}$ identifies the hidden state/feature vector at the time step $t$ and $\tanh(\cdot)$ represents the hyperbolic tangent

function. The output of the RNN stage is a sequence of hidden states/feature vectors $(\dot{\mathbf{h}}_1, \dot{\mathbf{h}}_2, ..., \dot{\mathbf{h}}_{N_{\text{steps}}})$, which is used as input to the LSTM network for cognitive overload detection via secondary task decision correctness anticipation. The parameters to be learned at this stage are $\mathbf{W}^{rnn} \in \mathbb{R}^{N_{rnn} \times N_x}$, $\mathbf{H}^{rnn} \in \mathbb{R}^{N_{rnn} \times N_{rnn}}$ and $\mathbf{b}^{rnn} \in \mathbb{R}^{N_{rnn}}$.

The LSTM network stage operates on the sequence of feature vectors and outputs a sequence of hidden states $\mathbf{h}_t \in \mathbb{R}^{N_{lstm}}$, as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}^i \mathbf{h}_{t-1} + \mathbf{I}^i \dot{\mathbf{h}}_t + \mathbf{b}_i) \tag{4}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}^f \mathbf{h}_{t-1} + \mathbf{I}^f \dot{\mathbf{h}}_t + \mathbf{b}_f) \tag{5}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}^o \mathbf{h}_{t-1} + \mathbf{I}^o \dot{\mathbf{h}}_t + \mathbf{b}_o) \tag{6}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}^c \mathbf{h}_{t-1} + \mathbf{I}^c \dot{\mathbf{h}}_t + \mathbf{b}_c) \tag{7}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{8}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, and $\mathbf{c}_t$ identify input gate, forget gate, output gate, and memory cell, respectively. The parameters to be learned are $\mathbf{W}^* \in \mathbb{R}^{N_{lstm} \times N_{lstm}}$, $\mathbf{I}^* \in \mathbb{R}^{N_{lstm} \times N_{rnn}}$ and $\mathbf{b}^* \in \mathbb{R}^{N_{lstm}}$, where $*$ is used to represent $\{i, f, o\}$.

Finally, after computing the hidden states of the LSTM stage, DecNet performs a probability projection via a fully connected layer followed by a $\mathrm{softmax}$ activation, as

$$\hat{\mathbf{y}}_t = \mathrm{softmax}(\mathbf{W}^y \mathbf{h}_t + \mathbf{b}^y), \tag{9}$$

where $\mathbf{W}^y \in \mathbb{R}^{2 \times N_{lstm}}$ and $\mathbf{b}^y \in \mathbb{R}^2$ are the parameters to be learned for the projection stage.

### B. Model Training

When training DecNet, we employ a sequence-to-sequence learning paradigm, similar to [15], [9], [16]. For the training process, we adopt a label smoothing technique [51] by replacing the binary labels of decision correctness, i.e., correct ($y_j = 1$) and wrong ($y_j = 0$), with soft labels, i.e., correct ($y_j = 0.9$) and wrong ($y_j = 0.1$), as they have shown to be a particularly effective strategy to improve learning stabilization and model generalization. We assume a weighted cross-entropy loss across each input sequence $\mathbf{X}_j$:

$$loss = \sum_{j=1}^{N} \sum_{t=1}^{N_{\text{steps}}} -e^{(t-N_{\text{steps}})} \log p(y_j | \mathbf{x}_{1 \to t}), \tag{10}$$

where $\mathbf{x}_{1 \to t} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$ identifies the sub-sequence of observations until timestamp $t$.

Eq. (10) defines a weighted loss across the entire sequence and builds on the assumption that longer sequences contain extra information for correct inference. It may also be interpreted as a form of auxiliary loss similar to [52], whereby the network is encouraged to extract relevant features from early observations, increase the gradient signal that gets propagated back, and provide additional regularization.

The exponential weights in the loss, i.e., $e^{(t-N_{\text{steps}})}$, reduce the impact of earlier decisions over performance, as they are made when less context is available for error anticipation [9], and incentivize the role of latter decisions. Also, this loss was shown to have positive effects as a regularizer to prevent early overfitting [15].

## IV. Experiments

In this section, we introduce the experimental protocol adopted, the simulated scenarios, the dataset collection procedure and data pre-processing. During each recording session, we collected both physiological signals and behavioral data from the driver in two separate simulated scenarios, i.e., one where decision stimuli are provided via audio cues and one where they are visually presented. For physiological signals, we collect gaze and head pose, due to their unobtrusive nature and their known correlation with human cognitive states [15]. While driving, participants were given a series of instructions to follow to complete a cognitive secondary task, and their decisions were recorded as behavioral data.

### A. Participants

Twenty participants (mean age 26.4, standard deviation 3.3) with normal or corrected to normal vision consented to participate in our experiments. Before beginning, each participant was introduced to sensors and experimental protocol. Participants were given a chance to do a test drive with the simulator. This allowed them to familiarize themselves with the driving task and the simulated environment, and to reduce learning factors on data collection. During each participant's drive, an observer monitored physiological signals and behavioral data integrity. This study has been approved by the Ministry of Defence Research Ethics Committee (MoDREC).

### B. Setup

We setup a realistic dual-tasking driver-in-the-loop virtual reality (VR) simulation for the experiment (see Fig. 3). The setup included: a physical simulator, a VR headset, and a screen for monitoring purposes. The VR headset has integrated eye gaze and head pose tracking, and requires an infra-red camera mounted above the steering wheel to operate. We developed and designed the simulated driving environments using the Unreal Engine (https://www.unrealengine.com/). The use of a simulated environment allows to have complete control on both on the environment, i.e., driving maneuvers and speed, and the tasks that participants experience during the experiment, i.e., the frequency and the number of decisions, in addition to providing a safe environment to the participants.

### C. Experimental Protocol

During each drive, participants were instructed to jointly perform two tasks. The primary task was to drive along a straight highway and avoid stationary rectangular obstacles. The secondary task is an "*n-back*" [26] based task which required participants to perform a cognitive-based decision when the simulator prompted them to do so. Since the cognitive load is inherently not a measurable metric, this task has often been used in the literature as a proxy to modulate different levels of cognitive load on the driver [23], [25], [24]. The paradigm builds on the core assumption that a participant's cognitive load while performing a task is strongly correlated with the working memory required to perform such a task [26]. The task


Fig. 3. Driver-in-the-loop simulation. Top: The two simulated scenarios. For the audio stimuli scenario, obstacles are simple boxes, while in the visual stimuli case, numbers are displayed directly on the obstacles. Bottom: The participant wears a VR headset with integrated eye gaze tracker and head pose estimation. The screen displays the scene observed by the participant and sensor readings in real-time for monitoring during the trial.

allows to easily to modulate different levels of cognitive load by increasing/decreasing the "n", and it also has been shown to be an effective tool to predict individual fluid intelligence and higher cognitive functions, especially when used to induce higher levels of load, such as 3-back [53].

We designed the simulator to prompt the secondary task numbers to the participants at regular intervals. Participants were instructed that each number corresponded to a specific category and that their task was to iteratively remember the category of the number they were presented three steps before. The numbers spanned from 1 to 12 and corresponding categories were as follows: category $a$ corresponded to the set of numbers $\{1, 2, 3\}$, category $b$ corresponded to the set of numbers $\{4, 5, 6\}$ and so on. Participants were instructed that four buttons on the steering wheel were dedicated to the task, with each button corresponding to a different category.

To illustrate the experiment condition, let us consider a participant in the audio stimuli scenario, presented with the sequence of numbers $3, 5, 8, 7, 6, 9, 10$. Given a 3-back task the participant would not be required to perform any decision until prompted the number 7. In fact, when provided with the number 7, the participant is storing 3 numbers in their memory, i.e., $3, 5, 8$ and is therefore required to "make a decision" based on the number they were presented 3-steps before, i.e., 3, and press the button that corresponds to the category $a$. From this moment on, every time the participant is presented with a new number, the participant is asked to decide/remember to which category the last number in their memory buffer corresponded to.

Secondary tasks can have disruptive effects on the primary task performance [54], [55]. To avoid this, when describing the secondary task, participants were informed that, although it was important for them to correctly perform the secondary task, their main focus must always be to safely perform the primary task, i.e., driving and avoiding obstacles. We enforced this, by reminding the participants before each drive that driving safety was of utmost importance. Their compliance was reflected in the fact that not a single crash was recorded amongst all participants and all driving scenarios.

The experimental protocol required each participant to drive the simulator on two separate scenarios of 180s. In each scenario, a different modality of stimuli for the secondary task was in place: one auditory and one visual. For the visual stimuli, numbers were displayed on the obstacles, as shown in Fig. 3 middle. In the scenario with the auditory stimuli, numbers were announced to the participants at every obstacle, as shown in Fig. 3 top. The two scenarios are designed to induce a constant level of cognitive load by prompting constant decisions in the participant. The 3-back task was also chosen for the secondary task to be challenging and cognitively demanding to perform. In fact, since we build on the assumption that decision mistakes from drivers are indicative of cognitive overload occurrences, the secondary task needed to be complex enough to induce events of cognitive overload in the participants.

Collecting the data on these two scenarios opens to the possibility to investigate two separate cognitively demanding cases: one where multiple modalities are stimulated and one where a single modality is engaged. Also, we can investigate whether a single modality engaging task could lead to sensory overload, as numerous works have shown that concurrent tasks using the same modality lead to performance decrease [31], [32]; and whether cognitive load and its effects on driver decisions can be distributed with different stimuli.

### D. Simulated Environment

The two simulated environments, as shown in Fig. 3 top and middle, both assumed a daylight scenario in good weather conditions. This ensured that obstacles were clearly visible to the participants and also limited the possible sources of distractions during the experiment. We designed the road to be 10m wide, enough to be divided into three 3.3m wide virtual lanes, and the obstacles to be 3.5m wide in order to entirely block a virtual lane and to ensure that drivers had to steer to avoid them. Before each drive, obstacles are randomly placed along one of three lanes and 100 meters apart. We implemented a proportional–integral–derivative controller on the simulator to have a consistent cruise speed of 120 km/h. This helps ensure a consistent cognitive load throughout the experiment, an equal number of decisions for all participants and a consistent driving scenario. More specifically, participants were prompted to make a decision approximately every 3 seconds and for a total of 55 decisions per drive in both scenarios. We also designed the obstacles so that they did not exert any effects on the vehicle upon contact, while sending a notification to the monitoring researcher. This allows reducing the potential loss of focus on the primary and secondary tasks from the participant caused by a crash with the obstacle, while still tracking their performance. All participants successfully

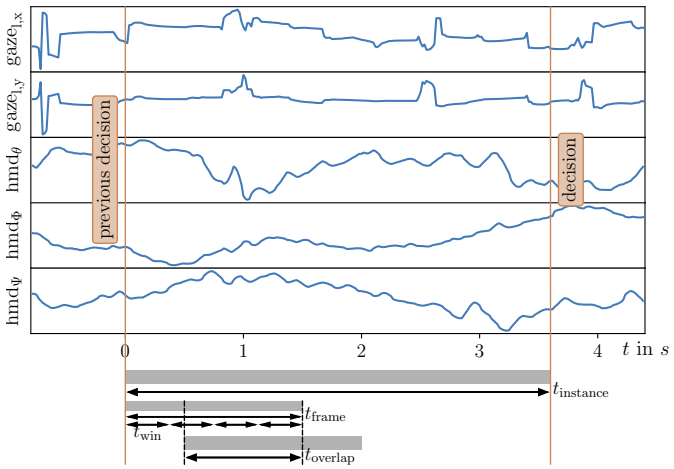



Fig. 4. Sequence generation. We frame the sequence of observations that precede a secondary decision as a "*decision instance*" of variable length $t_{\text{instance}}$. Within each decision instance, we extract inputs of length $t_{\text{frame}}$ with a sliding windows approach with fixed overlap $t_{\text{overlap}}$. For sequential models, each element of the sequence was computed over windows of duration $t_{\text{win}}$.

complied with instructions on safety, as not a single crash was recorded amongst all participants and all driving scenarios. Finally, to ensure that no memory effect could occur between trials, numbers used for the secondary task were automatically regenerated for each drive.

The location of the obstacles on the lanes is implemented according to a custom-defined discrete distribution, which allows us to reduce the probability of cases where a virtual lane is free of obstacles for extended periods. Assuming $c_i$ as the distance between the current obstacle location and the previous obstacle location in lane $i$. We then define the obstacle placement probability distribution in lane $i$ as

$$p(i) = \frac{e^{c_i/\text{IntervalSize}}}{\sum_i e^{c_i/\text{IntervalSize}}}, \tag{11}$$

where IntervalSize represents the distance between two adjacent obstacles. For instance, consider a case where the $i$-th lane has not been blocked for the past 5 obstacles, the custom distribution ensures that the probability for that lane to be blocked is $e^5$ times higher than the one of the most recently blocked lane.

### E. Dataset Collection

During each drive, we collect: 1) instantaneous two-dimensional gaze locations for left and right eye at 60 Hz, 2) three-dimensional head pose at 60 Hz, and 3) driver decisions. At time $t$, the integrated eye-tracker in the VR headset provides two-dimensional vectors with the gaze position on the screens, as seen through the headset lenses for both eyes as follows

$$\mathbf{e}^l(t) = [e^l_x(t), e^l_y(t)], \quad \mathbf{e}^r(t) = [e^r_x(t), e^r_y(t)], \tag{12}$$

where superscripts $l$ and $r$ differentiate between left and right eye, respectively, and subscripts specify the axis of the data. Vectors are normalized in the range $[-1, 1]$ along both x-axis and y-axis, so that the center is $(0, 0)$, bottom-left is $(-1, -1)$ and top-right is $(1, 1)$. Head pose information is directly inferred from position and rotation of the headset, with position and rotation during calibration being considered as reference. The headset position is specified in Cartesian coordinates, and the rotation is described in Euler angles (roll-pitch-yaw notation):

$$\mathbf{h}(t) = [h_x(t), h_y(t), h_z(t), h_\phi(t), h_\theta(t), h_\psi(t)], \tag{13}$$

where subscripts $\phi$, $\theta$ and $\psi$ specify yaw, pitch and roll data, respectively. All the physiological data was collected in a vector $\hat{\mathbf{s}}(t)$ as follows:

$$\hat{\mathbf{s}}(t) = \left[e^r_x, e^r_y, e^l_x, e^l_y, h_x, h_y, h_z, h_\phi, h_\theta, h_\psi\right], \tag{14}$$

where time dependency on the single elements of $\hat{\mathbf{s}}(t)$ has been omitted to simplify notation.

The secondary task decisions were recorded in a vector $\mathbf{u}$:

$$\mathbf{u}(t_d) = [n_s,\ id,\ rt], \tag{15}$$

where $t_d$ identifies the time at which occurred the decision, $n_s$ the number that was provided as stimulus, $id$ the category chosen by the participant and $rt$ the reaction-time of the participant.

### F. Pre-processing and Dataset Split

After data collection, each gaze pattern data sample was processed to provide distance $\delta$ from the previous sample on both axes and the absolute distance from the center of the field of view ($d_{\text{fov}}$). This procedure ensures that the network does not learn to associate mistakes and correct decisions with specific gaze locations, but on the dynamics of the eye movements. For the head pose, we keep the absolute values of position and rotation, as head movements are characterized by slower shifts than eye gaze. This led to a sample for each time step with 11 raw features, as follows:

$$\mathbf{s}(t) = \left[\delta^r_x, \delta^r_y, \delta^l_x, \delta^l_y, d_{\text{fov}}, h_x, h_y, h_z, h_\phi, h_\theta, h_\psi\right]. \tag{16}$$

We process the dataset for classification by splitting each participant's data into decision instances of variable length $t_{\text{instance}}$. Secondary task decision instances are bounded by the timestamp at which the driver made a decision $t_d$ and the timestamp immediately after the previous decision $t_{d-1}$, as shown in Fig. 4.

Dataset splitting into train, validation and test set is only performed after processing the data into a sequence of decision instances. This procedure ensures that data from decisions in the train set and the validation/test sets are entirely separated and not correlated. In other words, we always perform training, validation and testing on separate decisions.

We pre-process the data within each decision instance $j$ with a sliding window approach. We extract the input sequences for our models from sequences of raw sensor data of length $t_{\text{frame}}$. The window of raw data is processed into a $N_{\text{steps}}$-long sequence $\mathbf{X}_j(t)$ of fixed-size feature vectors $\mathbf{x}_{j,t}$:

$$\mathbf{X}_j(t) = [\mathbf{x}_{j,t-N_{\text{steps}}}, ..., \mathbf{x}_{j,t}]. \tag{17}$$

We compute each feature vector $\mathbf{x}_{j,t}$ from a sliding window of length $t_{\text{win}}$, as follows

$$\mathbf{x}_{j,t} = f[\mathbf{s}(t - t_{\text{win}}), ..., \mathbf{s}(t)], \tag{18}$$

where the operator $f(\cdot)$ computes mean, standard deviation, median, 25th and 75th percentiles, maximum, minimum and range of its argument. We chose this set of features to capture central tendencies, variability, and extremes of each physiological signal. For non-sequential models, i.e., logistic regression and SVM, we directly compute the aforementioned statistical features over data windows of length $t_{\text{frame}}$.

We normalize the features to have zero-mean and unit variance, and we uniformly sample the input sequences $t_{\text{frame}}$ via a sliding window approach with overlap $t_{\text{overlap}} = \xi \cdot t_{\text{frame}}$. The parameter $\xi$ represents the overlap ratio, which is fixed to 95% for all the models considered in the paper.

As we frame the problem as supervised learning, we need to identify the binary labels for the cognitive overload events. Our main assumption is that mistakes on the secondary task are representative of cognitive overload events, therefore we assign labels according to the following policy:

$$y_j = \begin{cases} 0 \text{ if } id \neq \text{category}(n_s) \\ 1 \text{ if } id = \text{category}(n_s), \end{cases} \tag{19}$$

where category$(\cdot)$ is the operator that extracts the category of the number that was given as stimulus. In other words, if the participant could not recall the correct category to the number stored in their memory, the data corresponding to that decision is assigned to an event of cognitive overload, i.e., $y_j = 0$. On the other hand, correct decision corresponded to a level of cognitive workload that the participant could sustain.

## V. Experimental Setup

In this section, we present the experimental setup we assumed to address the following research questions:

1) Do gaze patterns and head movements correlate with driver secondary task decision-making processes?
2) Can these correlations be exploited to anticipate the likelihood of making a mistake on the secondary task, i.e., a cognitive overload event?
3) How far in advance can we anticipate a cognitive overload event so that a closed-loop assistance system can intervene?
4) What is the impact of different stimuli on cognitive stimulation and decision on the driver?

To answer these questions, we evaluate the performance of DecNet on the collected dataset and compare it with various models.

### A. Classification Scenarios

For critical safety applications, we focus on the model's ability to anticipate the likelihood of future secondary task decision mistakes of the driver (wrong decision classification), as they relate to cognitive overload events which might lead to dangerous maneuvers. However, it is not advisable for a model to be unable to robustly infer the likelihood of future correct decisions (correct decision classification). If the assistance system takes over too often, even when it would not have been necessary, it may cause discomfort and distrust on the driver.

Consequently, we evaluate DecNet performance on three separate classification scenarios: correct decision, wrong decision, and normalized decision classification. Correct and wrong decision classification scenarios focus on evaluating whether the model can anticipate correct or wrong decisions, respectively. Instead, the normalized classification scenario evaluates DecNet's ability to anticipate the general correctness of the next decision. In this scenario, we first compute performance metrics for both correct and wrong decisions. Then, their average is weighted according to the support, i.e., the number of true instances for each decision label.

### B. Evaluation Setup

We evaluate classification performance in terms of precision, recall, and $F_1$-score:

$$P = \frac{tp}{tp + fp}, \quad R = \frac{tp}{tp + fn}, \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{20}$$

where $tp$, $fp$ and $fn$ identify true positives, false positives and false negatives, respectively.

While we focus on the classification performance of DecNet in terms of precision, recall, and $F_1$-score, it is important to stress that the output of the proposed model is the correctness likelihood of the next secondary/cognitive task decision. We map the likelihood value to a binary class, i.e., correct or wrong decision, via a classification threshold $\eta$. In other words, if the correctness likelihood is above the threshold $\eta$, we classify the next decision as correct, while if the value falls below, we classify it as a wrong decision.

Classification performance are computed in an offline test setting, where we use 80% of data for training, 10% of data for validation, and the remaining 10% for testing. When comparing model performances, we report the mean and standard deviation of each metric for all algorithms using 5-fold cross-validation.

All models were implemented in Python on Keras (https://keras.io/) and trained with Adam optimizer [56]. For the training of the networks, we set the learning rate to 0.0001, the total number of epochs to 100 and we performed early stopping on the validation set. Model training and testing were deployed on an Intel Core i7-6800K 3.40GHz CPU and NVIDIA GeForce GTX 1080 8GB GPU.

## VI. Results

In this section, we compare DecNet performance with our previous sequence-to-sequence model (Seq2Seq; [16]) and with two sequential neural network model baselines, i.e., standard recurrent neural network (RNN) and long short-term memory (LSTM). In addition to these, we also compare against non-sequential baselines, i.e., Logistic Regression (LogReg) and Support Vector Machines (SVMs). The above baseline models are in line with cognitive load classification literature [15], [25], [57], [2], which have focused on sequential modeling, e.g., as LSTMs and RNNs [15], Hidden Markov Models (HMM) [25], Logistic Regression and SVM [57] and Naive Bayes classifier [2].

TABLE I
AUDIO TASK: CLASSIFICATION PERFORMANCE.

| Method | $t_{\text{frame}} = 0.5s$ $F_1$-Score | $t_{\text{frame}} = 1s$ $F_1$-Score | $t_{\text{frame}} = 1.5s$ $F_1$-Score |
|---|---|---|---|
| Normalized Decision Classification | | | |
| LogReg | 0.60±0.02 | 0.60±0.03 | 0.62±0.02 |
| SVM | 0.70±0.02 | 0.69±0.03 | 0.68±0.02 |
| RNN | 0.75±0.03 | 0.70±0.02 | 0.69±0.03 |
| LSTM | 0.73±0.03 | 0.71±0.03 | 0.70±0.04 |
| Seq2Seq | 0.74±0.01 | 0.75±0.02 | 0.73±0.03 |
| DecNet | 0.75±0.01 | 0.75±0.02 | 0.78±0.01 |
| Wrong Decision Classification | | | |
| LogReg | 0.50±0.03 | 0.50±0.03 | 0.51±0.03 |
| SVM | 0.59±0.03 | 0.57±0.03 | 0.54±0.03 |
| RNN | 0.60±0.03 | 0.61±0.03 | 0.60±0.03 |
| LSTM | 0.62±0.02 | 0.62±0.02 | 0.61±0.04 |
| Seq2Seq | 0.64±0.02 | 0.64±0.02 | 0.63±0.03 |
| DecNet | 0.65±0.01 | 0.67±0.02 | 0.68±0.02 |
| Correct Decision Classification | | | |
| LogReg | 0.64±0.02 | 0.64±0.03 | 0.68±0.03 |
| SVM | 0.75±0.02 | 0.75±0.02 | 0.75±0.02 |
| RNN | 0.75±0.04 | 0.74±0.02 | 0.73±0.02 |
| LSTM | 0.77±0.04 | 0.75±0.03 | 0.75±0.03 |
| Seq2Seq | 0.83±0.02 | 0.83±0.02 | 0.77±0.02 |
| DecNet | 0.84±0.01 | 0.82±0.01 | 0.84±0.01 |

TABLE II
VISUAL TASK: CLASSIFICATION PERFORMANCE.

| Method | $t_{\text{frame}} = 0.5s$ $F_1$-Score | $t_{\text{frame}} = 1s$ $F_1$-Score | $t_{\text{frame}} = 1.5s$ $F_1$-Score |
|---|---|---|---|
| Normalized Decision Classification | | | |
| LogReg | 0.56±0.02 | 0.56±0.03 | 0.56±0.01 |
| SVM | 0.61±0.02 | 0.61±0.02 | 0.61±0.02 |
| RNN | 0.64±0.02 | 0.63±0.02 | 0.62±0.04 |
| LSTM | 0.64±0.04 | 0.65±0.03 | 0.63±0.01 |
| Seq2Seq | 0.65±0.02 | 0.64±0.02 | 0.65±0.01 |
| DecNet | 0.66±0.02 | 0.66±0.02 | 0.66±0.03 |
| Wrong Decision Classification | | | |
| LogReg | 0.49±0.03 | 0.47±0.04 | 0.47±0.02 |
| SVM | 0.53±0.01 | 0.53±0.03 | 0.53±0.03 |
| RNN | 0.61±0.02 | 0.61±0.02 | 0.60±0.03 |
| LSTM | 0.61±0.02 | 0.61±0.01 | 0.60±0.02 |
| Seq2Seq | 0.62±0.01 | 0.61±0.01 | 0.60±0.05 |
| DecNet | 0.61±0.01 | 0.62±0.01 | 0.63±0.02 |
| Correct Decision Classification | | | |
| LogReg | 0.61±0.02 | 0.62±0.02 | 0.62±0.02 |
| SVM | 0.66±0.03 | 0.66±0.02 | 0.67±0.02 |
| RNN | 0.75±0.01 | 0.75±0.01 | 0.75±0.01 |
| LSTM | 0.75±0.01 | 0.75±0.01 | 0.76±0.01 |
| Seq2Seq | 0.76±0.02 | 0.75±0.01 | 0.75±0.01 |
| DecNet | 0.76±0.03 | 0.76±0.01 | 0.75±0.01 |

TABLE III
DECNET: CLASSIFICATION PERFORMANCE AGAINST FEATURES.

| Scenario | $F_1$ Gaze | $F_1$ Head | $F_1$ Comb |
|---|---|---|---|
| Audio Norm. | 0.67±0.01 | 0.75±0.01 | 0.78±0.01 |
| Audio Wrong | 0.54±0.02 | 0.61±0.02 | 0.68±0.02 |
| Audio Correct | 0.81±0.01 | 0.81±0.02 | 0.84±0.01 |
| Visual Norm. | 0.63±0.01 | 0.63±0.01 | 0.66±0.03 |
| Visual Wrong | 0.59±0.02 | 0.56±0.03 | 0.63±0.02 |
| Visual Correct | 0.75±0.01 | 0.75±0.02 | 0.75±0.01 |

### A. Classification Performance

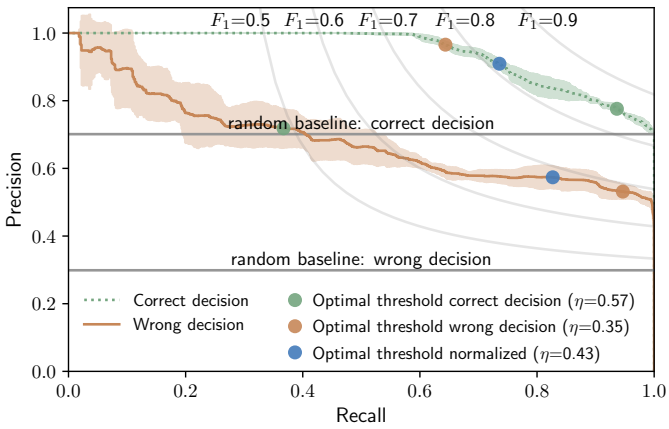
In this section, we list the classification performance as a function of the length of the input $t_{\text{frame}}$ used for training and testing for both the audio and the visual stimuli experiment.

In Table I and Table II, we report $F_1$-scores under the three classification settings, i.e., normalized classification, correct and wrong decision discovery, for the audio and the visual stimuli experiment, respectively. For each classification scenario, we compute the classification threshold $\eta$ according to the best $F_1$ performance achieved in the validation set.

DecNet outperforms all other models on all classification tasks. The performance gap with other models becomes more accentuated when longer inputs are provided to the models. On frame length $t_{\text{frame}} = 1.5s$ for the audio stimuli experiment, DecNet shows an overall $F_1$-score performance improvement of 8%, more specifically of $\sim 0.05$, $\sim 0.05$, and $\sim 0.07$, for the normalized, wrong and correct decision classification, respectively. A similar behavior appears on the visual task on the wrong classification scenario, where DecNet has 5% performance increase over the second best performing model, Seq2Seq. In general, we can see that sequence models better capture the secondary task decision-making process than simpler non-sequential models, with DecNet providing a performance increase especially on the wrong decision classification scenario.

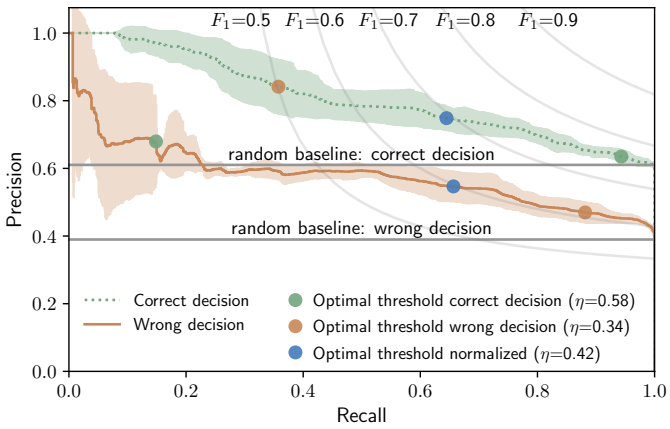
In Table III, we evaluate $F_1$-scores for DecNet when $t_{\text{frame}} = 1.5s$ for the three classification scenarios as a function of the physiological signals used for training and testing. Here, $F_1$ Gaze identifies the performance of DecNet when only gaze information is used, $F_1$ Head when only head pose information is used and finally $F_1$ Comb lists the performance when both gaze and head pose information are combined. As we can see, both streams of information contribute to the performance of DecNet Comb, with head pose data being able to achieve better performance when considered alone. The performance of DecNet Gaze on the audio task, however, are not surprising, as they are comparable to previous studies on knowledgeability anticipation from gaze information alone [37]. On the visual task, instead, we can see that gaze data is more relevant to the final $F_1$-Score performance of DecNet, as it identifies the main resource used by the participants to capture the information required to complete the task.

In Fig. 5, we collect the precision-recall curves for DecNet when $t_{\text{frame}} = 1.5s$ for the audio-stimuli and visual-stimuli tasks. The plots show the precision/recall performance curves for both correct and wrong decision classification as a function of the decision threshold in the audio, Fig. 5a, and visual task scenario, Fig. 5b. We can see that in both scenarios and for both classes the precision-recall curves are well above the random baseline, despite the complexity of the task. The plots indicate that DecNet can effectively separate the two classes of correct, i.e., high load, and wrong decisions, i.e., cognitive overload, on the secondary task as performance as a function




(a) Audio Stimuli Experiment

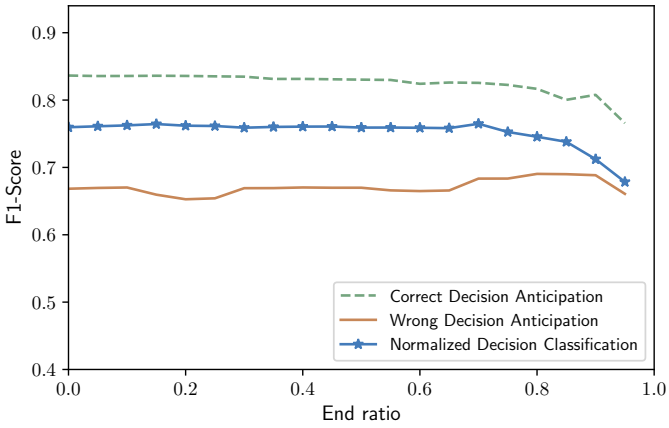
| Scenario | $P$ | $R$ | $F_1-$Score |
|---|---|---|---|
| Norm. Dec. Class. | 0.81 | 0.77 | 0.79 |
| Wrong Dec. Class. | 0.58 | 0.83 | 0.68 |
| Correct Dec. Class. | 0.91 | 0.74 | 0.82 |



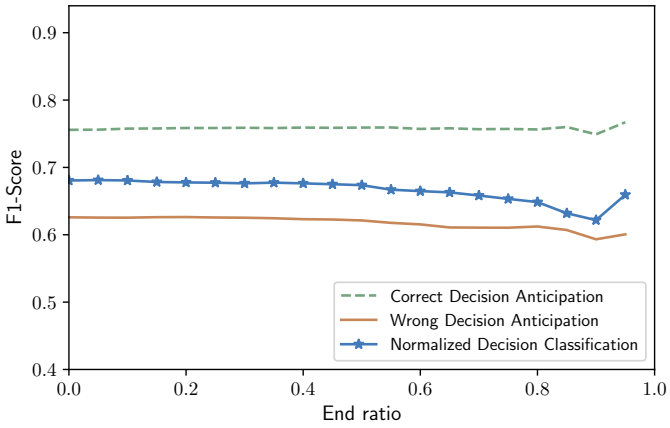

(b) Visual Stimuli Experiment

| Scenario | $P$ | $R$ | $F_1-$Score |
|---|---|---|---|
| Norm. Dec. Class. | 0.67 | 0.65 | 0.66 |
| Wrong Dec. Class. | 0.55 | 0.66 | 0.60 |
| Correct Dec. Class. | 0.75 | 0.65 | 0.70 |

Fig. 5. Precision-recall curves for the wrong decision (solid orange line) and correct decision (dotted green line) classes. The left plot represents the audio task, while the right plot shows the visual task. All classification curves are clearly above the random baseline. Note that the random baselines are different depending on the task and decision class, as the decision classes are (slightly) imbalanced in our dataset. Shaded areas represent the standard deviation across 10 runs. Green and orange dots depict the best decision threshold in terms of $F_1$ score for correct and wrong decisions respectively, while the blue dots depict the best performance when balancing correct and wrong decisions. The table below each plot shows the performance in terms of precision, recall and $F_1$-score for the three classification scenarios when using the normalized classification threshold. All thresholds have been selected based on the validation set.




(a) Audio Stimuli Experiment




(b) Visual Stimuli Experiment

Fig. 6. Performance as a function of the time available to anticipate the correctness of an incoming decision. The end ratio indicates how early DecNet is required to anticipate the next decision. For instance, assuming a decision instance of $t_{\text{instance}} = 3.5s$, an input of $t_{\text{frame}} = 1s$ and an end ratio of ER $= 0.8$, DecNet would produce a prediction ER $\cdot\ (t_{\text{instance}} - t_{\text{frame}}) = 2s$ before such decision takes place.

of the classification threshold $\eta$ are consistent.

### *B. Decision Correctness Anticipation Performance*

The main goal of DecNet is to provide an actionable metric for assistance systems to be able to intervene if the occurrence of a mistake is detected. In this section, we evaluate how well and how far in advance DecNet is able to anticipate the correctness of an imminent decision. Without loss of generality, we assume that the need for a decision has already been detected by the assistive system, as incoming decision detection is beyond this paper's scope.

In Fig. 6, we show the $F_1$-score performance of DecNet as a function of time available to the classifier before providing the correctness likelihood of the next decision. We assume a $t_{\text{frame}} = 1s$ input. Performances for correctness anticipation for the audio task are fairly stable for all the scenarios considered.

However, it is interesting to notice that DecNet appears to be more able to identify incoming mistakes in the time frame between $2s$ and $3s$ before the decision. This could suggest that the features of an incoming mistake from the driver are more robust and relevant during the moments that precede a decision. On the other hand, the features that predict a correct incoming decision might be more prominent after the model has had more time to observe the driver since the past decision has passed, i.e., when the driver has had time to switch their attention from the past decision to the next one.

The final goal of DecNet resides in its ability to be implemented and operate in real-time to provide timely assistance

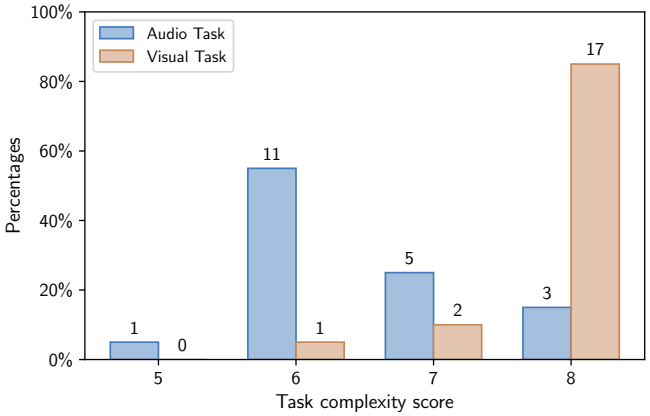



Fig. 7. Participant perceived difficulty of the task. Participants were asked to rate the perceived complexity of each of the experiments with a value between 0, i.e., low complexity, to 8, i.e, highest complexity.

to the user. To evaluate this, we have computed the inference time, given an input sequence of 1.5s. The total inference time of DecNet is 8ms, which corresponds to an inference rate of $\sim$ 125Hz. Since the data from the eye and head pose tracker is captured at 60Hz, we conclude that the proposed DecNet is capable of operating in real-time.

### C. Effect of Stimuli Modality on Performance

Performance in Tables I and II and Fig. 5 indicate that DecNet can better anticipate the correctness of incoming decisions when participants were provided audio stimuli, in comparison to decisions prompted by visual cues. To investigate this, we asked each participant to rate the secondary task's perceived complexity with a value ranging between 0 to 8, with 0 identifying a low demanding task and 8 a highly demanding one.

Responses of the participants are collected in Fig. 7, where we can see that 85% of the participants considered the visual task to be of higher complexity than the audio stimuli task. This shows that DecNet performance on decision anticipation and the human perceived complexity match in both experimental scenarios, and suggests that single modality stimulation exerts higher levels of cognitive load on drivers, directly affecting their ability to make correct decisions. Overall, participants agreed that visual scenario is more demanding than the audio. In fact, in the audio task there are no confounding variables, as numbers are announced every time an obstacle is passed, while on the visual task, participants had to read the numbers displayed on the obstacles. However, there was no evident difference in their driving performance as no crashes were recorded. It is also interesting to notice that drivers assigned different levels of complexity for each scenario, showing that drivers' perception of cognitive load can differ also on the same task and suggesting that they divide energies between the two tasks using different strategies. This highlights the ambiguity of a cognitive load-based metric, and further stresses the benefits of a metric based on secondary task decision mistakes, which inherently occur when drivers are overloaded.

Our results confirm the findings of numerous studies in the multitasking theory literature, such as [31], [32]. The multitasking model in [31] assumes that auditory and visual perception use different resources, therefore if two joint tasks use different modalities their performance are expected to improve, and worsen if they require the same modality. Similarly, in the working memory theory by [32] it was shown that participants experience significant performance disruption when two or more concurrent tasks operate on the same visual modality, which is consistent with our results on the visual stimuli scenario.

## VII. Conclusions

In this paper, we introduced DecNet, an end-to-end multi-stage recurrent deep model that anticipates the correctness of an imminent decision from a driver as a proxy to cognitive overload instances. We collected a dataset from a cohort of participants on two separate decision-making scenarios: one where decision stimuli are presented visually and one where they are auditory. We investigated the ability of the proposed model to anticipate the secondary task decisions on both scenarios from non-obtrusive physiological signals only, namely eye gaze and head pose.

Our results showed that DecNet is high performing in the task of decision correctness anticipation, achieving 81% precision and 77% recall on the auditory stimuli task, and 67% precision and 65% recall on the visual stimuli task. The proposed model outperforms comparable models on all the scenarios considered. We tested the real-time capabilities of DecNet and proved that the proposed model can reliably infer the correctness likelihood of a decision up to 2s before such a decision takes place.

We have also investigated the effects that different stimuli modalities have on cognitive overload events of the driver, i.e., their decision accuracy on the secondary task, and therefore more generally on their level of cognitive load. Our analyses indicate that when a single modality is overloaded, as for the visual stimuli task, both drivers and DecNet tend to be less reliable performance-wise. This suggests that, in case of take-over, it would be preferable to use a different modality than the one currently used to perform such a task. While DecNet is capable of running online in real-time, given its inference rate of 125Hz, all shown classification performance refer to offline testing. Therefore, it would be interesting to investigate how DecNet performs in a closed-loop setting, where a human driver is interacting with the simulated environment.

Our study has shown that unobtrusive physiological signals are strongly correlated with cognitive overload events in the driver and that DecNet can exploit such correlations to anticipate these events. The proposed model DecNet shows solid and reliable performance, however it represents an initial step towards real-time cognitive overload estimation and it would be interesting to investigate the application of more advanced machine learning models to this problem. Also, the proposed study focused on simulated scenarios without external distractions, and it would be therefore worth exploring how well we can generalize our methods to more complex scenarios, such as real-world driving. In our study, we induced a high level of cognitive load in the driver by means of a "3-back" task, which is inherently artificial. Although common in driver

cognitive states studies due to their ease of implementation, these tasks still represent a proxy for real-life driving tasks and may represent a limitation of the applicability of the proposed methods to natural driving scenarios.

## REFERENCES


[1] T. Carlson and Y. Demiris, "Collaborative control for a robotic wheelchair: evaluation of performance, attention, and workload," *IEEE Trans. on Sys., Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 876–888, 2012.

[2] E. Haapalainen, S. Kim *et al.*, "Psycho-physiological measures for assessing cognitive load," in *Intern. Conf. on Ubiquitous Comput.*, 2010, pp. 301–310.

[3] M. L. Reyes and J. D. Lee, "Effects of cognitive load presence and duration on driver eye movements and event detection performance," *Transp. Res. Part F: Traffic Psychol. and Behav.*, vol. 11, no. 6, pp. 391–402, 2008.

[4] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. on Intell. Transp. Sys.*, vol. 6, no. 2, pp. 156–166, 2005.

[5] R. Bose, H. Wang *et al.*, "Regression-based continuous driving fatigue estimation: Toward practical implementation," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 2, pp. 323–331, June 2019.

[6] D. Grimes, D. S. Tan *et al.*, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *SIGCHI Conf. on Hum. Factors in Comput. Sys.*, 2008, pp. 835–844.

[7] M. Gombolay, A. Bair *et al.*, "Computational design of mixed-initiative human–robot teaming that considers human factors: Situational awareness, workload, and workflow preferences," *Int. J. of Robot. Res.*, vol. 36, no. 5-7, pp. 597–617, 2017.

[8] A. Steinfeld, T. Fong *et al.*, "Common metrics for human-robot interaction," in *ACM SIGCHI/SIGART Conf. Human-Robot Interaction*, 2006, pp. 33–40.

[9] A. Jain, A. Singh *et al.*, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *IEEE Intern. Conf. on Robot. and Automat.*, 2016, pp. 3118–3125.

[10] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Trans. on Intell. Transp. Sys.*, vol. 8, no. 2, pp. 340–350, 2007.

[11] C. P. Lam, A. Y. Yang *et al.*, "Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems," in *IEEE Int. Conf. on Intell. Robots and Syst.*, Sept 2015, pp. 5776–5783.

[12] D. L. Strayer and W. A. Johnston, "Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone," *Psychol. Sci.*, vol. 12, no. 6, pp. 462–466, 2001.

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

[14] M. Wollmer, C. Blaschke *et al.*, "Online driver distraction detection using long short-term memory," *IEEE Trans. on Intell. Transp. Sys.*, vol. 12, no. 2, pp. 574–582, 2011.

[15] R. Wang, P. V. Amadori, and Y. Demiris, "Real-time workload classification during driving using hypernetworks," in *IEEE Intern. Conf. on Intell. Robots and Sys.*, 2018, pp. 3060–3065.

[16] P. V. Amadori, T. Fischer *et al.*, "Decision anticipation for driving assistance systems," in *IEEE Intern. Conf. on Intell. Transp. Sys.*, 2020.

[17] M. Petit, T. Fischer, and Y. Demiris, "Lifelong augmentation of multimodal streaming autobiographical memories," *IEEE Trans. Cogn. Devel. Syst.*, vol. 8, no. 3, pp. 201–213, Sept. 2016.

[18] Y. Xing, C. Lv *et al.*, "Driver activity recognition for intelligent vehicles: A deep learning approach," *IEEE Trans. on Veh. Tech.*, vol. 68, no. 6, pp. 5379–5390, 2019.

[19] T. Billah, S. M. Rahman *et al.*, "Recognizing distractions for assistive driving by tracking body parts," *IEEE Trans. on Circuits and Sys. for Video Tech.*, vol. 29, no. 4, pp. 1048–1062, 2018.

[20] M. Petit and Y. Demiris, "Hierarchical action learning by instruction through interactive grounding of body parts and proto-actions," in *IEEE Int. Conf. on Robot. and Automat.*, 2016, pp. 3375–3382.

[21] A. Taniguchi, T. Taniguchi, and T. Inamura, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Trans. Cogn. Devel. Syst.*, vol. 8, no. 4, pp. 285–297, Dec. 2016.

[22] M. A. Recarte and L. M. Nunes, "Mental workload while driving: Effects on visual search, discrimination, and decision making." *J. of Exp. Psychol.: Applied*, vol. 9, no. 2, pp. 119–137, 2003.

[23] E. T. Solovey, M. Zec *et al.*, "Classifying driver workload using physiological and driving performance data: Two field studies," in *SIGCHI Conf. on Hum. Factors in Comput. Sys.*, 2014, pp. 4057–4066.

[24] B. Mehler and B. Reimer, "How demanding is "just driving"? A cognitive workload-psychophysiological reference evaluation," in *Intern. Driving Symp. on Hum. Factors in Driver Assessment, Training and Vehicle Design*, 2019, pp. 363–369.

[25] L. Fridman, B. Reimer *et al.*, "Cognitive load estimation in the wild," in *CHI Conf. on Hum. Factors in Comput. Sys.*, 2018, pp. 652:1–652:9.

[26] B. Reimer, C. Gulash *et al.*, "The MIT AgeLab n-back: a multi-modal android application implementation," in *Intern. Conf. on Automot. User Interfaces and Interactive Veh. Applications*, 2014.

[27] G. Hossain and M. Yeasin, "Analysis of cognitive dissonance and overload through ability-demand gap models," *IEEE Trans. Cogn. Devel. Syst.*, vol. 9, no. 2, pp. 170–182, June 2015.

[28] X. Zhang, Y. Sugano *et al.*, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 41, no. 1, pp. 162–175, 2017.

[29] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Eur. Conf. on Comput. Vision*, 2018, pp. 339–357.

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

[31] C. D. Wickens, "Multiple resources and mental workload," *Hum. Factors*, vol. 50, no. 3, pp. 449–455, 2008.

[32] A. Baddeley, "Working memory: Theories, models, and controversies," *Annual Review of Psychol.*, vol. 63, pp. 1–29, 2012.

[33] H. K. Wong, J. Epps, and S. Chen, "A comparison of methods for mitigating within-task luminance change for eyewear-based cognitive load measurement," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 4, pp. 681–694, Dec. 2018.

[34] E. Ferreira, D. Ferreira *et al.*, "Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults," in *IEEE Symp. on Comput. Intell., Cogn. Alg., Mind, and Brain*, 2014, pp. 39–48.

[35] T. Georgiou and Y. Demiris, "Adaptive user modelling in car racing games using behavioural and physiological data," *User Model. and User-Adapted Interact.*, vol. 27, no. 2, pp. 267–311, 2017.

[36] M. J. Cole, J. Gwizdka *et al.*, "Inferring user knowledge level from eye movement patterns," *Information Processing & Management*, vol. 49, no. 5, pp. 1075–1091, 2013.

[37] O. Celiktutan and Y. Demiris, "Inferring human knowledgeability from eye gaze in mobile learning environments," in *Eur. Conf. on Comput. Vision Workshops*, 2018, pp. 193–209.

[38] R. Bednarik, S. Eivazi, and H. Vrzakova, "A computational approach for prediction of problem-solving behavior using support vector machines and eye-tracking data," in *Eye Gaze in Intelligent User Interfaces*, 2013, pp. 111–134.

[39] M. X. Huang, J. Li *et al.*, "Moment-to-moment detection of internal thought during video viewing from eye vergence behavior," in *ACM Intern. Conf. on Multimedia*, 2019, p. 2254–2262.

[40] T. Fischer and Y. Demiris, "Markerless perspective taking for humanoid robots in unconstrained environments," in *IEEE Intern. Conf. on Robot. and Automat.*, 2016, pp. 3309–3316.

[41] T. Fischer and Y. Demiris, "Computational modelling of embodied visual perspective-taking," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 4, pp. 723–732, 2020.

[42] A. Dasgupta, S. M. Bhattacharya, and A. Routray, "A system for noncontact estimation of cognitive load using saccadic parameters based on a serio-parallel computing framework," *IEEE Trans. Cogn. Devel. Syst.*, vol. 11, no. 3, pp. 450–459, 2019.

[43] M. Blanco, W. J. Biever *et al.*, "The impact of secondary task cognitive processing demand on driving performance," *Accident Anal. & Prevention*, vol. 38, no. 5, pp. 895–906, 2006.

[44] N. Merat, A. H. Jamson *et al.*, "Highly automated driving, secondary task performance, and driver state," *Hum. Factors*, vol. 54, no. 5, pp. 762–771, 2012.

[45] F. Naujoks, C. Purucker, and A. Neukum, "Secondary task engagement and vehicle automation–comparing the effects of different automation levels in an on-road experiment," *Transp. Res. part F: Traffic Psychol. and Behav.*, vol. 38, pp. 67–82, 2016.

[46] B. Wandtner, N. Schömig, and G. Schmidt, "Secondary task engagement and disengagement in the context of highly automated driving," *Transp. Res. part F: Traffic Psychol. and Behav.*, vol. 58, pp. 253–263, 2018.

[47] V. Alizadeh and O. Dehzangi, "The impact of secondary tasks on drivers during naturalistic driving: Analysis of EEG dynamics," in *IEEE Intern. Conf. on Intell. Transp. Sys.*, 2016, pp. 2493–2499.
[48] T. Ersal, H. J. Fuller *et al.*, "Model-based analysis and classification of driver distraction under secondary tasks," *IEEE Trans. on Intell. Transp. Sys.*, vol. 11, no. 3, pp. 692–701, 2010.
[49] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intell. Transp. Sys. Mag.*, vol. 9, no. 4, pp. 10–22, 2017.
[50] J. Engström, G. Markkula *et al.*, "Effects of cognitive load on driving performance: The cognitive control hypothesis," *Hum. factors*, vol. 59, no. 5, pp. 734–764, 2017.
[51] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" *Neural Information Processing Systems*, 2019.
[52] C. Szegedy, W. Liu *et al.*, "Going deeper with convolutions," in *IEEE Conf. on Comput. Vision and Pattern Recognition*, 2015, pp. 1–9.
[53] S. M. Jaeggi, M. Buschkuehl *et al.*, "The concurrent validity of the N-back task as a working memory measure," *Memory*, vol. 18, no. 4, pp. 394–412, 2010.
[54] R. C. Williges and W. W. Wierwille, "Behavioral measures of aircrew mental workload," *Human Factors*, vol. 21, no. 5, pp. 549–574, 1979.
[55] X. Wu and Z. Li, "Secondary task method for workload measurement in alarm monitoring and identification tasks," in *Int. Conf. on Cross-Cultural Design*, 2013, pp. 346–354.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[57] F. Tango and M. Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Trans. on Intell. Transp. Sys.*, vol. 14, no. 2, pp. 894–905, 2013.

**Pierluigi Vito Amadori** (S'14, M'17) received the M.Sc. degree (Hons.) in Telecommunications Engineering from the University of Rome La Sapienza, Rome, Italy, in 2013 and the Ph.D. degree in Electronic Engineering from the Department of Electrical & Electronic Engineering, University College London, London, U.K., in 2017.

He currently holds a position as a Postdoctoral Research Associate at the Personal Robotics Laboratory at Imperial College London, London, U.K. His main research interests include driver monitoring, user modeling and driving assistance systems.

**Tobias Fischer** (M'16) received the B.Sc. degree from Ilmenau University of Technology, Germany, in 2013, the M.Sc. degree in Artificial Intelligence from the University of Edinburgh, U.K., in 2014, and the Ph.D. degree from the Personal Robotics Lab, Imperial College London, London, U.K, in 2018.

His research interests include both computer vision and human vision, visual attention and computational cognition. He is interested in applying this knowledge to cognitive robotics.

Dr. Fischer was a recipient of the Queen Mary Award for the Best U.K. Robotics PhD Thesis in 2018 and the Eryl Cadwaladr Davies prize for the best departmental thesis in 2017-2018.

**Ruohan Wang** (S'16) received the B.Sc. degree (Hons.) from National University of Singapore, Singapore, in 2012, and the Ph.D. degree from the Personal Robotics Lab, Imperial College London, London, U.K, in 2021.

He is currently a research scientist with Institute of Infocomm Research, A*STAR, Singapore. His main research interests include machine learning and its application in assistive robotics. Ruohan was a recipient of National Science Scholarship, Singapore.

**Yiannis Demiris** (SM'03) received the B.Sc. (Hons.) degree in artificial intelligence and computer science and the Ph.D. degree in intelligent robotics from the Department of Artificial Intelligence, University of Edinburgh, Edinburgh, U.K., in 1994 and 1999, respectively.

He is a Professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is the Royal Academy of Engineering Chair in Emerging Technologies, and the Head of the Personal Robotics Laboratory. His current research interests include human-robot interaction, machine learning, user modeling, and assistive robotics. He has published more than 200 journal and peer-reviewed conference papers in the above areas.

Prof. Demiris was a recipient of the Rector's Award for Teaching Excellence in 2012 and the FoE Award for Excellence in Engineering Education in 2012. He is a Fellow of IET, BCS, and Royal Statistical Society.