Accident Analysis and Prevention 121 (2018) 118–128

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Adapting artificial neural networks to a specific driver enhances detection and prediction of drowsiness

Charlotte Jacobé de Naurois[a,b,*], Christophe Bourdin[a], Clément Bougard[b], Jean-Louis Vercher[a]

[a] Aix Marseille Univ, CNRS, ISM, Marseille, France
[b] Groupe PSA, Centre Technique de Vélizy, Vélizy-Villacoublay, Cedex, France

ARTICLE INFO

Keywords:
Monitoring
ANN
Adaptive learning
Inter-individual variability
Drowsiness

ABSTRACT

Monitoring car drivers for drowsiness is crucial but challenging. The high inter-individual variability observed in measurements raises questions about the accuracy of the drowsiness detection process. In this study, we sought to enhance the performance of machine learning models (Artificial Neural Networks: ANNs) by training a model with a group of drivers and then adapting it to a new individual. Twenty-one participants drove a car simulator for 110 min in a monotonous environment. We measured physiological and behavioral indicators and recorded driving behavior. These measurements, in addition to driving time and personal information, served as the ANN inputs. Two ANN-based models were used, one to detect the level of drowsiness every minute, and the other to predict, every minute, how long it would take the driver to reach a specific drowsiness level (moderately drowsy). The ANNs were trained with 20 participants and subsequently adapted using the earliest part of the data recorded from a 21st participant. Then the adapted ANNs were tested with the remaining data from this 21st participant. The same procedure was run for all 21 participants. Varying amounts of data were used to adapt the ANNs, from 1 to 30 min, Model performance was enhanced for each participant. The overall drowsiness monitoring performance of the models was enhanced by roughly 40% for prediction and 80% for detection.

1. Introduction

Driving while drowsy is a safety issue, and a major cause of accidents. Numerous fundamental and applicative studies focus on detection of drowsiness as a way to improve accident prevention. However, simply detecting drowsiness is not enough: once the driver is drowsy, it is probably already too late to prevent the accident. The key challenge is to predict how and when drowsiness will occur, how often it will occur and who might become drowsy under which conditions. Prediction refers here to the timely identification of when a given event will occur within a given range of future states, in our case a given level of drowsiness. Watson and Zhou (2016) detected the occurrence of micro-sleep episodes with 96% accuracy and were able to predict the next micro-sleep between 15 s and 5 min in advance, although obviously not the time of occurrence of the first micro-sleep. A recent study (Jacobé de Naurois et al., 2017) showed that an Artificial Neural Network (ANN) can not only detect the level of drowsiness but can also predict, in advance, the time at which this impaired driver's state will occur.

Various sources and types of information can be used to estimate the operator's functional state. For car driving, measurements must be easily recordable, not invasive, and reliable. The literature contains a variety of sources of information (Dong et al., 2011), mainly based on ocular and eyelid movements (Chen and Ji, 2012; Liu et al., 2009). For instance, PERCLOS (PERcentage of eye CLOSure, the percentage of time, generally during one minute, when eyes are closed more than 80%) indicates how long on average the eyes are closed. Physiological measurements are also often used to assess the driver's state through the central and the neuro-vegetative systems, offering the advantage of being continuously available, objective and fairly direct indicators of the functional state. The most commonly used physiological signal is the electroencephalogram (EEG). However, EEG recording during driving is rather intrusive and constraining (despite continuous technological advances), which can be a real disadvantage. Electrocardiogram (EKG) and respiration measurements are also often used. Yet it remains difficult to define a direct relationship between physiological features and a given cognitive state, since these physiological features vary with other states like stress, emotions, workload, physical effort and fatigue, or with the context.

Finally, driving behavior and performance, such as the standard deviation of car position relative to lane midline (also termed standard deviation of lane position (SDLP)) or steering wheel movements,

* Corresponding author at: Aix Marseille Univ, CNRS, ISM, Marseille, France.
E-mail addresses: charlotte.jacobe-de-naurois@etu.univ-amu.fr (C. Jacobé de Naurois), jean-louis.vercher@univ-amu.fr (J.-L. Vercher).

https://doi.org/10.1016/j.aap.2018.08.017
Received 5 April 2018; Received in revised form 18 July 2018; Accepted 14 August 2018
Available online 20 September 2018

(Arnedt et al., 2001; De Valck et al., 2003; Liu et al., 2009; Philip et al., 2004) are also common measures used to detect the driver's state. However, here again, driving performance and activity are not specific indicators of drowsiness.

To deal with the above limitations, recent research has sought to improve prediction through complex approaches combining multivariate, heterogeneous information via data fusion (Dong et al., 2011; Samiee et al., 2014). Findings from these studies show that this hybrid approach can provide better accuracy (Awais et al., 2014).

However, current models need to deal with yet another challenge to their prediction power. It is now widely recognized that neurobehavioral and cognitive performance vary considerably from one individual to another (Van Dongen et al., 2004aa; b). In car driving tasks, according to (Ingre et al., 2006), there is extensive inter-individual variability in driving behavior and eye behavior. Under similar conditions, individuals' patterns of drowsiness evolution over time can differ, and for a given self-declared drowsiness level, markers such as eye blink duration also vary considerably. Van Dongen et al. (2003) showed that individuals probably also differ in their vulnerability to sleep deprivation, and that this is partially predictable from individual cognitive performance without deprivation, i.e. from the individual cognitive profile. In driving simulator studies, drowsiness is often observed to develop in differing ways (Thiffault and Bergeron, 2003). Situational and personality factors, sleeping habits and driving history help explain why some people fall asleep at the wheel while others do not. This confirms the need to consider drivers' traits or profiles to calibrate systems for the detection and prediction of drowsiness (Jacobé de Naurois et al., 2017).

Such large inter-individual variability makes creating algorithms that will perform well for all individuals a challenge. As most studies use machine-learning algorithms, the difficulty is finding a general model trained with a limited number of drivers which can then be applied to the majority of individual drivers (Karrer et al., 2004). One of the main issues with machine learning is uncertainty about the generalization of a given model to a new participant. To ascertain whether an algorithm generalizes well, the dataset is segregated into either two (training and testing) or three (training, validating and testing) datasets (in most cases, the segregation is randomly performed on the full set of recorded data). Thus, it is impossible to be sure that the algorithm will perform well for another participant whose data is unknown to the model.

This problem can be approached in different ways. One is to train the model with as large a population as possible: the more data, the better the model. However, this method is based on the assumption that for each new individual, the model has previously encountered a similar individual. This makes it difficult to determine the number of participants required to deal with the large inter-individual variability. Furthermore, the level of similarity between two individuals is hard to quantify. This method would thus be extremely time-consuming, not only in terms of training the model but also in terms of data collection. A second solution is to have a specific model for each driver, but this obviously involves collecting and labeling sufficient data from each driver as well as training the specific model with these data, another time-consuming option. A third way is to use methods such as transfer learning or adaptive learning, which combine the advantages of the two preceding methods by permitting capitalization on a group of individuals and personalization for each new individual. In particular, these methods are applied on Brain Computer Interface systems (Wang et al., 2015). To detect driver drowsiness, studies applied such techniques on EEG signals (Wei et al., 2015; Wu et al., 2015, 2016) and found that transfer learning applied to EEG significantly enhances model performance. Our aim here was to test a similar method based on adaptive learning but using non-intrusive measurements including eyelid movements, head movements, EKG, respiration rate, driving activity and performance, as in our previous study (Jacobé de Naurois et al., 2017).

The goal of the present study is to enhance the performance of machine-learning models both in detecting the level of driver drowsiness and in predicting when a given impaired state will be reached, by first training a model and then adapting it to each new individual. The model uses Artificial Neural Networks. We hypothesize that training an ANN with a group of individuals and then personalizing the ANN for a new individual (whose data were not encountered by the model during training) will improve the performance of the model for this specific individual. We also assess the amount of data required to enhance the generalization performance of the model.

2. Materials and methods

The participants and the protocol, including data collection and preprocessing, were the same as used for our previous study (Jacobé de Naurois et al., 2017). Data modeling methods were specifically developed for the present study.

2.1. Participants

Twenty-one participants were included in the study (mean age 24.09 ± 3.41 years; 11 men, 10 women). Inclusion criteria were: valid driver's license for at least 6 months, no visual correction needed to drive, not susceptible to simulator sickness, as assessed by the Motion Sickness Susceptibility Questionnaire, Short-form (MSSQ-Short, Golding, 1998), and an Epworth scale score (assessing susceptibility to drowsiness) below 14 (Johns, 1991) (for more detail, see (Jacobé de Naurois et al., 2017)). The following participant information was collected: Epworth scale score (assessing susceptibility to drowsiness (Johns, 1991)), quality of the previous night's sleep (on a scale from 1 to 10), caffeine consumption (never, rarely, one or two cups per day, more than two cups per day), driving frequency (occasionally, several times a month/a week/a day), distance (kilometers) driven per year and score on the Horne and Östberg morning/evening questionnaire (Horne and Ostberg, 1975).

2.2. Protocol

The participants drove for 100 to 110 min in a static driving simulator in an air-conditioned room with temperature control set at 24 °Celsius. They drove just after lunchtime, a time considered as risky in terms of drowsiness (Horne and Reyner, 1999). The road and traffic were generated with SCANeR Studio®. A webcam located on top of the central screen of the simulator video-recorded the participants during the session to establish the ground truth (see below). The (static) simulator, provided by Oktal® and powered with SCANeR Studio® software, is made of a real car seat, 3 video screens (24″ in format 16/9 each, forming a tryptic), a steering wheel, pedals and a small screen (10″) for the dashboard, located just behind the wheel. The driving environment was displayed at a resolution of 1280 × 1024 pixels onto the three forward screens providing a 210° horizontal forward field of view. A rear screen provided a 60° rear field of view, corresponding to the normal use of the central rearview and two side mirrors. A stereo sound system provided simulated engine, road, and traffic sounds. An example of the field of view is presented on the figure below, which has been added to Fig. 1. The simulated car had an automatic gearbox, so the driver had only access to the steering wheel, gas and brake pedals. At the beginning of the session, the participants drove along a highway for roughly 90 min, then turned off the highway and drove for around 5 min to reach a city. Finally, they drove in an urban environment for roughly 5 min. There was no traffic during most of the highway stretch. The very monotonous environment (without event or traffic) was selected in order to induce drowsiness. Somewhere 2/3 of the way along, 22 cars appeared from the right of the highway, disappearing a few kilometers later (Fig. 2). This sudden addition of traffic was intended to change the driver's level of drowsiness. Rossi et al. (2011)

Fig. 1. The static driving simulator by Oktal®. A represents the road scene displayed on the 3 video screens. B represents the dashboard. C is the hardware faceLAB®. D are two of the three electrodes used for ECG, E is the respiratory belt. F are the electrodes for EDA (not used in this study because of important signal loss).

demonstrated that a driver is more susceptible to sleepiness in a simulator with a monotonous scenario.

2.3. Data collection and processing (inputs)

During the driving, data on driving performance, eyelid and head movements, as well as physiological data, were recorded using the following hardware and software: SCANeR Studio® for driving performance at 10 Hz, faceLAB® for sensorimotor signals at 60 Hz, and EKG, pulse plethysmography (PPG), respiration with the Biopac® MP150 system and Acqknowledge® software at 1000 Hz. Even if EEG is a gold standard, it is a method quite intrusive, so its use seems difficult in a context of industrialization in real cars. Several indicators were extracted from each source of information. The different indicators are summarized in Table 1. The variables were recorded at a frequency of 1/60 Hz (one by minute) because it is the lowest frequency common to all sources of information, including the ground truth. The participant information and the driving time (time elapsed since the beginning of the session) were both included as input based on the results of the previous study (Jacobé de Naurois et al., 2017).

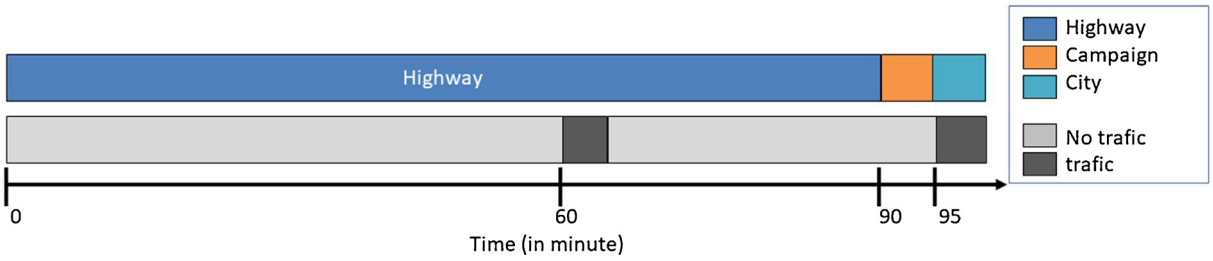

Fig. 2. Diagram of the scenario with different types of road and the associated traffic.

Table 1
All the variables (grouped in columns by source of information) computed for each participant, averaged for each minute of driving time, and used as inputs for the ANNs.

Physiological data	Behavioral data	Car data
HR: Heart Rate (average and standard deviation) (beat/min)	Blink duration (average and standard deviation)	Lateral distance from the closest lane and the center of the car in m (average and standard deviation)
Svlf: HR signal Very Low Frequency Power (0.0-0.04 Hz)	Blink frequency (average and standard deviation) (per minute)	Time to lane crossing (average and standard deviation)
Slf: HR signal Low Frequency Power (0.04-0.15 Hz)	PERCLOS (average and standard deviation) (% of eye-closure time)	Steering angle (average and standard deviation)
Shf: HR signal High Frequency Power (0.15-0.4 Hz)	Head position x (average and standard deviation)	Steering angle velocity (average and standard deviation)
Svhf: HR signal Very High Frequency Power (0.4-3.0 Hz)	Head position y (mean and standard deviation)	Steering entropy (computed from steering angle)
Sympathetic ratio (Slf / (Svlf + Slf + Shf))	Head position z (average and standard deviation)	Number of direction changes (0-crossings) per minute (computed from steering angle)
Vagal ratio (Shf / (Svlf + Slf + Shf))	Head rotation x (average and standard deviation)	Accelerator pedal angle (average and standard deviation)
Sympathetic-vagal ratio (Slf / Shf)	Head rotation y (average and standard deviation)	Lateral shift of the vehicle center relative to the lane center (average and standard deviation)
Respiration Rate (average and standard deviation) (per minute)	Head rotation z (average and standard deviation)	Vehicle speed (km/h) (average and standard deviation)
	Saccade frequency (mean and standard deviation) (per minute)	Number of runs-off-road per minute

2.4. Model (ANN)

Driver's drowsiness was modeled with an artificial neural network created with the neural network toolbox (Beale et al., 1992) of Matlab R2013a. A feedforward neural network with one hidden layer optimized with the Levenberg-Marquardt algorithm (Levenberg, 1944) was used. The number of neural units in the hidden layer varied between 1 and 25 and was optimized via a grid-search method (applied by steps of 2). This model was trained with a subset (n-1) of participants and adapted to a further participant according to the method described below (Section 2.7).

2.5. Detection of the real level of drowsiness (ground truth)

The real (ground truth) level of drowsiness was determined based on a method proposed by Wierwille and Ellsworth (1994). Every minute of driving, two raters evaluated the driver's state as ranging between 0 (alert state) and 4 (extremely drowsy) (Table 2). The mean of both raters was used as the drowsiness level. Inter-rater reliability was computed with Pearson's linear correlation ($R = 0.71$ and $p = 0.00$). Even if this subjective rating by a third (informed) person is validated by the "consistency and reliability in the rating produced" as stated by Wierwille and Ellsworth (1994), it is difficult to relate the ratings to the real state of drowsiness, thus one cannot conclude because the real drowsiness is a hidden measure, not directly accessible.

2.6. Outputs of models

As in the previous study, the goal here was both to detect the current level of drowsiness and to continuously predict when the driver's state would reach a given threshold. We therefore used two ANNs, one for detection, the other for prediction, as follows. After appropriate training, the first ANN detected a level of drowsiness with an output in the range 0–4 by steps of 0.5. If detected drowsiness was lower than 1.5, the second ANN predicted (in minutes) when it would reach 1.5: this time was its output, otherwise 0. The threshold was set at 1.5, meaning that at a given time, one of the two raters evaluated the state of the participant as moderately drowsy (level 2) while the other rated it as slightly drowsy (level 1). The impaired state was defined according to the level of drowsiness and not driving performance or event detection performance because there is no direct and reciprocal relation between driving performance and driver's level of drowsiness. The driving performance is not necessarily impaired when the driver shows signs of sleepiness (Philip et al., 2005).

Table 2
Trained observer rating based on a scale by Wierwille and Ellsworth (1994) cited by Rost et al. (2015).

Level	Drowsiness State	Video image indicators
0	not drowsy	Normal fast eye blinks, often reasonably regular; Apparent focus on driving with occasional fast sideways glances; Normal facial tone; Occasional head, arm and body movements.
1	slightly drowsy	Increase in duration of eye blinks; Possible increase in rate of eye blinks; Increase in duration and frequency of sideway glances; Appearance of "glazed eye" look; Appearance of abrupt irregular movements – rubbing face/eyes, moving restlessly on the chair; Abnormally large body movements following drowsiness episodes; Occasional yawning.
2	moderately drowsy	Occasional disruption of eye focus; Significant increase in eye blink duration; Disappearance of eye blink patterns observed during alert state; Reduction on degree of eye opening; Occasional disappearance of facial tone; Episodes without any body movements.
3	very drowsy	Discernable episodes of almost complete eye closure, eyes never fully open; Significant disruption of eye focus; Periods without body movements (longer than for level 2) and facial tone followed by abrupt large body movements.
4	extremely drowsy	Significant increase in duration of eye closure; Longer duration of episodes of no body movement followed by large isolated "correction" movements.

2.7. Methods for adaptive learning with the ANN

The present study set out to test whether the ANN can efficiently be adapted to each specific driver. The following methodology was used to create the adaptive ANN (hereafter termed Ad-ANN). Fig. 3 presents the overall process, in two phases: a classic training phase (steps 1, 2a, 3, 4), similar to our previous study (Jacobé de Naurois et al., 2017), and an adaptation phase (steps 2b, 5 and 6) aimed at improving the performance of the system for a particular participant.

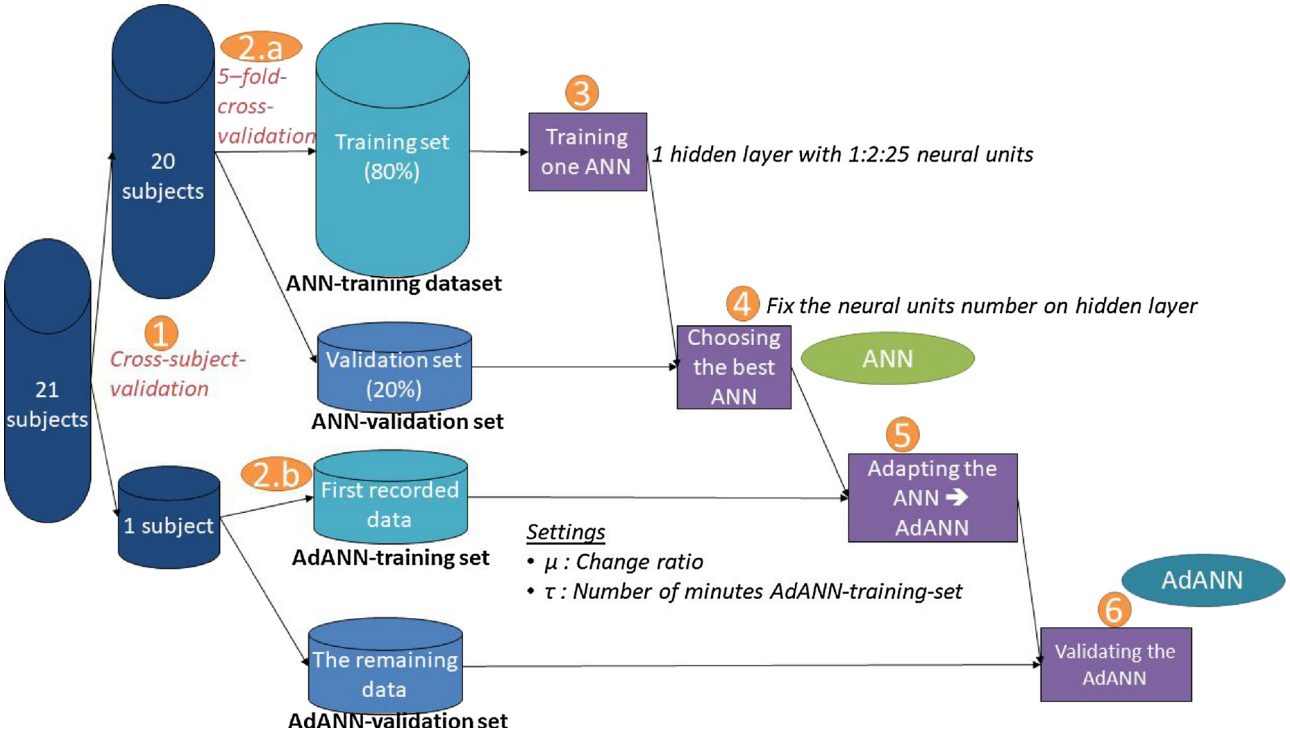

Fig. 3. The training method used to adapt the ANNs. Cylinders represent different datasets. The small numbered circles represent the step of the process defined in part 2.7 Methods for adaptive learning with the ANN. Rectangles represent steps in the process.

Step 1: a cross-subject validation (also known as a one-subject-left-strategy) was computed, testing a different participant at each validation cycle: one dataset was composed using data from 20 (n-1) participants to train the ANN and the other dataset was composed using data from the 21 st participant. This process was repeated 21 times, each time using a different participant's data as test set.

Step 2: the datasets were again divided into sub-datasets. The dataset composed of 20 participants was divided in two sub-datasets (step 2a) for a 5-fold cross-validation with random distribution: 80% for training the ANN and 20% for testing and validating the ANN (hereafter ANN-training dataset and ANN-validation dataset). The dataset from the remaining single participant/driver was also divided into two sub-datasets (step 2b): one composed of the first data segment recorded during the driving task, i.e. the first minutes of driving (hereafter Ad-ANN-training dataset) and the other composed of the remaining data (hereafter Ad-ANN-validation dataset). Between 1 and 30 min of recorded data were used in the Ad-ANN-training dataset, i.e. between 1 and 30 lines of the dataset. This time variable, hereafter τ, constituted the first parameter for the adaptation. Thus, if the first parameter was equal to 5 and the participant drove for 110 min, the Ad-ANN-training dataset contained 5 lines corresponding to the first 5 min, while the Ad-ANN-validation dataset contained 105 lines covering the 6th to the 110th minutes. This second step yielded four datasets: ANN-training dataset, ANN-validation dataset, Ad-ANN-training dataset, and Ad-ANN-validation dataset. The ANN-training dataset was used to train a general ANN on 20 drivers. The ANN-validation dataset was used to validate and choose the general ANN defining the number of neural units in the hidden layer. The Ad-ANN-training dataset was used to adapt the ANN (Ad-ANN), i.e. to personalize the ANN for each driver. Finally, the Ad-ANN-validation dataset was used to assess the performance of the adaptive ANN, i.e. its performance with data not previously encountered by the model.

Step 3: after dividing the datasets, the training dataset (ANN-training dataset) composed of 20 participants was used to train feed-forward neural networks. This ANN is the general ANN (see Jacobé de Naurois et al., 2017 for more details).

Step 4: the ANN-validation-dataset was used to choose the best-performing neural networks (best number of neural units on hidden layer). The ANN with the lowest root mean square error (RMSE) was kept.

Step 5: the chosen ANNs were adapted using the AdANN-training dataset containing the first segment of recorded data from the 21 st driver. This step restarts the training process, this time based on the AdANN-training dataset (hereafter, Ad-ANN). For the adaptation, different change ratios μ were tested from 10^{-5} to 10^{3}. Thus, μ is the second parameter of adaptation, changing the weight of the Ad-ANN.

Finally (Step 6), the remaining dataset (Ad-ANN-validation dataset) was used to choose both best parameters (amount of data on Ad-ANN-training dataset τ and μ).

2.8. Evaluation and performance

For each step described above, the performance function was the root mean square error (RMSE) between the network outputs and the target outputs. This performance metric was used to compare the error on each dataset: before the adaptation, using the general ANN and after the adaptation, using the Ad-ANN. The lower this metric, the better the model. After the adaptation, the RMSEs of the ANN-training dataset and the ANN-validation dataset were expected to increase while the RMSEs of the AdANN-training dataset and the AdANN-validation dataset were expected to decrease, because the model would now be adapted specifically to the data of the 21 st driver. Both RMSE mean and standard deviation (SD) before (step 4) and after (step 6) the adaptation were compared to test this hypothesis, assuming that SD would also decrease after the adaptation. RMSE was also analyzed with respect to the amount of data used in the Ad-ANN-training dataset and in the AD-ANN-validation dataset.

2.9. Statistical analysis

In order to compare pre- and post-adaptation RMSE, a two-sample F-test for equal variances was performed before a two-sample matched

t-test for means. Bonferroni corrections were performed for each value of τ to compare RMSE variation as a function of τ.

2.10. Subject-specific performance assessment of AdANN

To assess the subject-specific performance of AdANN, the same learning process was repeated. This time, the ANN was trained with only 19 participants and adapted with the 20th participant (A). Finally, the performance of the ANN and the Ad-ANN were evaluated on the 21 st participant (B). These two drivers (A and B) changed with each iteration. The ANN was adapted using the first segment of recorded data from driver A and the resulting AdANN was tested on driver B. To test subject-specific performance, RMSE on driver A and driver B before adaptation (general ANN) and after adaptation (Ad-ANN) were compared. Our hypothesis was that RMSE would decrease for driver A (since the model was specifically adapted to this driver) but would increase for driver B.

3. Results

One participant had sign of simulators sickness, so the session was stopped immediately, and the participant was excluded from the study. Some participants had road departure at different levels of drowsiness. Before analyzing the performance of models, it is important to note that all participants did not reach the same level of drowsiness at the same moment. Some participants reached the level "extremely drowsy" (level 4), while others only reached "moderately drowsy" (level 2) and this at different temporalities. Only one participant was a little particular because he reached at the maximum this level (moderately drowsy) after one hour of driving, which was a very long delay as compared to the others. We first (Section 3.1) present the results with both best parameters, i.e.: those yielding the lowest RMSE (τ: amount of data in AdANN-training dataset and μ: change ratio during adaptation) with the different sources of information ("all", "physiological", "behavioral" and "car"). Next (Section 3.2), we focus on the best source of information to analyze RMSE variation as a function of τ, i.e. the amount of data used to adapt the ANN for each participant. Finally (Section 3.3), we examine the subject-specific performance of the procedure, presenting results on the adaptation performed with data from one driver and tested with data from another driver, again for the best source of information. Each section gives results on both detection and prediction modeling.

3.1. Best mu and best tau on AdANN-validation dataset

In this section, the lowest mean RMSE for the cross-subject validation is presented with both best τ and μ parameters, for each different source of information and for both detection and prediction ANNs.

3.1.1. Detection of level of drowsiness

The lowest average RMSEs pre- and post-adaptation for each source of information, and in each case for both best parameters, are presented in Fig. 4. For all sources of information, mean and SD are significantly lower after adaptation than before adaptation. For the "all" category ($t = 4.294$; $p < 0.001$ and $F = 25.826$; $p < 0.001$ for RMSE mean and SD respectively), best performance is for $\tau = 20$ min and $\mu = 50$. For the "physiological" category ($t = 2.335$; $p = 0.027$ and $F = 8.620$; $p < 0.001$ for the mean and SD respectively), post-adaptation best performance is for $\tau = 18$ min and $\mu = 10^{-5}$. For the" car" category ($t = 3.231$; $p = 0.003$ and $F = 24.420$; $p < 0.001$ for the mean and SD respectively), best performance is for $\tau = 19$ min and $\mu = 5$. Finally, for the "behavioral" category ($t = 3.231$; $p = 0.003$ and $F = 28.652$; $p < 0.001$ for the mean and SD respectively), best performance is for $\tau = 30$ min and $\mu = 5$. RMSE is lowest when the "all" source of information is used. Furthermore, it is only after adaptation that the average RMSE (for all participants) is lower than 1 level of drowsiness, i.e. model error is more or less one level of drowsiness. Model adaptation did not improve performance for some participants: 2 participants for the "all" source of information, 8 for "physiological", 1 for "behavioral" and 5 for "car", but it is worth noting that adaptation never resulted in higher error.

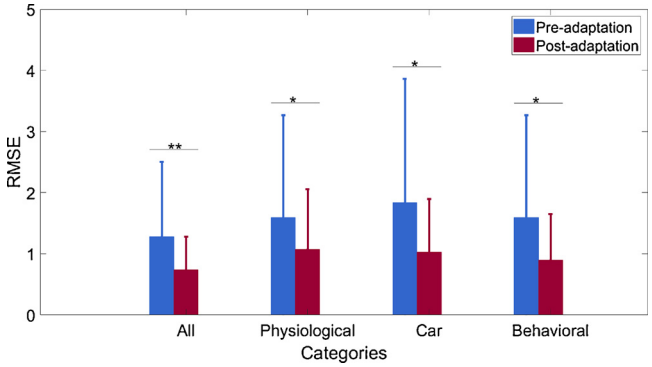

Fig. 4. Root mean square error (RMSE) and standard error of the detected level of drowsiness, based on the AdANN-validation dataset, for different sources of information, before and after adaptation. Stars represent the level of significance (NS: $p > .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

3.1.2. Prediction of time of occurrence of impaired driver state

The lowest average RMSEs pre- and post-adaptation for both parameters (τ and μ) are presented in Fig. 5. In terms of prediction of driver impairment, once again, for all sources of information, mean and SD are significantly lower after adaptation. For the "all" source of information category ($t = 3.349$; $p = 0.003$ and $F = 179.079$; $p < 0.001$ for the mean and SD respectively), best performance is with $\tau = 30$ min and $\mu = 10$. For the "physiological" category ($t = 4.262$; $p = 0.004$ and $F = 14.051$; $p < 0.001$ for the mean and SD respectively), best performance is with $\tau = 29$ min and $\mu = 0.5$. For the "car" category ($t = 3.938$; $p = 0.001$ and $F = 1235.924$; $p < 0.001$ for the mean and SD respectively), best performance is with $\tau = 30$ min and $\mu = 10$. Finally, for the "behavioral" category ($t = 3.938$; $p < 0.001$ and $F = 37.474$; $p < 0.001$ for the mean and SD respectively), best performance is with $\tau = 28$ min and $\mu = 10^{-1}$. The lowest RMSE is achieved with "behavioral" data, but this is not statically different from other sources of information. Moreover, it is only after adaptation that the mean of RMSE is lower than 6 min, i.e. the model error is more or less 6 min. As with detection, however, adaptation did not improve the prediction performance of the model for some participants (9 participants for the "all" source of information, 3 for "physiological", 4 for "behavioral" and 4 for "car"). Again, model performance never

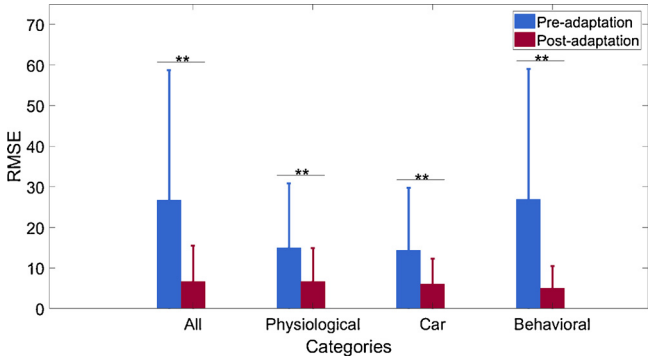

Fig. 5. Mean of RMSE and SD of the predicted time of occurrence of drowsiness level 1.5, based on AdANN-validation dataset, for different sources of information, before and after adaptation. Stars represent the level of significance of the difference in means (NS: $p > .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

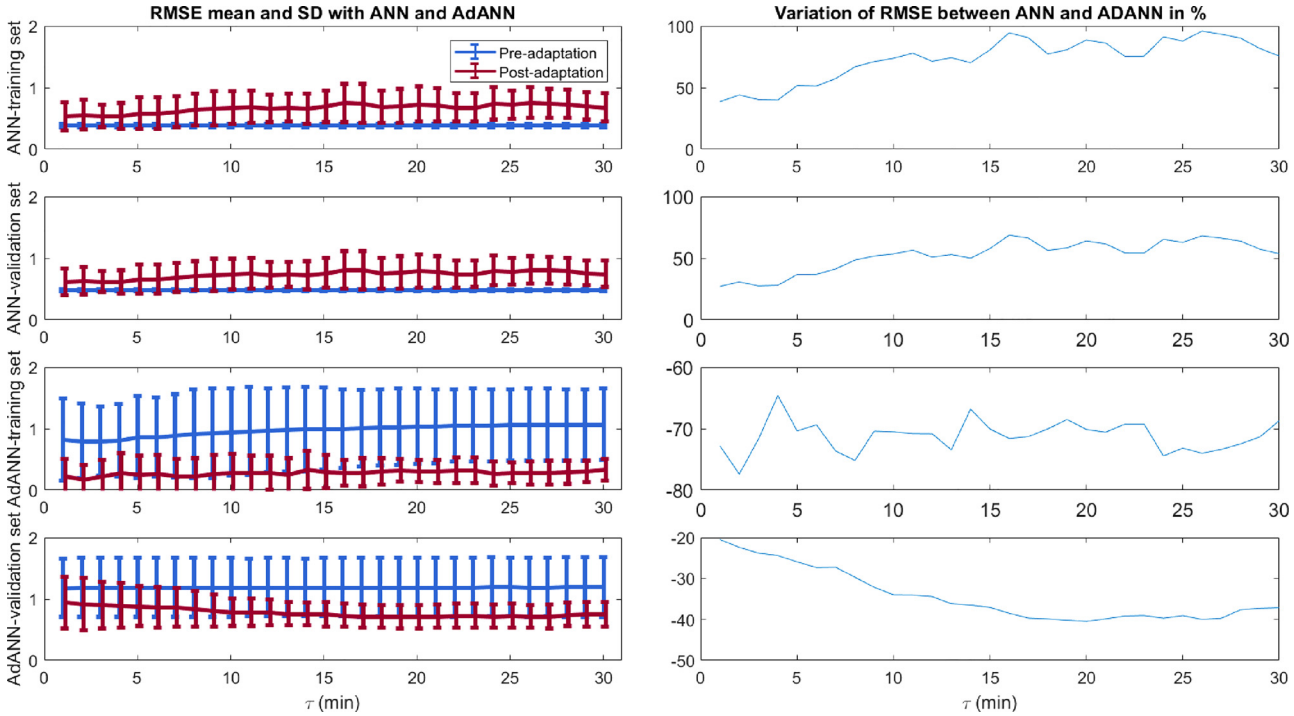

Fig. 6. Detection: Mean and SD of RMSE for the different datasets (ANN-training set, ANN-validation set, AdANN-training set, AdANN-validation set) as a function of amount of data (τ, in minutes) used to adapt the ADANN-training dataset.

decreased after adaptation.

3.2. Effect of varying the amount of data (τ) used for the adaptation

Detailed results on the best source of information with variations of τ, the amount of data used in the adaptation dataset (AdANN-training dataset), for both detection and prediction are outlined below. The "all" dataset yields best detection of drowsiness and the "behavioral" dataset best prediction of drowsiness. For each τ chosen (amount of data), the best μ was chosen. Both pre- and post-adaptation variations in RMSE are presented for the ANN-training dataset, the ANN-validation dataset, the AdANN-training dataset and the AdANN-validation dataset.

3.2.1. Detection of level of drowsiness

Fig. 6 shows the mean and the SD of RMSE for the cross-subject validation using the different datasets as a function of τ (amount of data in the AdANN-training dataset used to adapt the ANN). First, for both the ANN-training and the ANN-validation datasets, RMSE mean and SD are significantly higher after adaptation, whatever the amount of data used (for the mean, t-values between -2.787 and -3.968; $p < 0.01$ for the first value of τ and t-values between -4.440 and -5.487; $p < 0.001$for $\tau = 7$ to 30, for the SD, $p < 0.001$; F-values between 0.002 and 0.018 in all cases). For the AdANN-training dataset, RMSE mean and SD are significantly lower after adaptation for each τ (for the mean t-values between 3.592 and 5.896; $p < = 0.001$ for all cases, for the SD F-values between 3.440 and 11.713; $p < 0.001$ for all τ except 2, 3 and 6: $F = 6.691$; $p = 0.001$, $F = 4.658$; $p = 0.008$ and $F = 4.469$; $p = 0.002$). However, for the AdANN-validation dataset, RMSE is significantly lower when three minutes are used to adapt the ANN (t-values between 2.111 and 2.572; $p < 0.05$ for $\tau = 3$ to 7, t-values between 2.906 and 3.618; $p < 0.01$ for $\tau = 8$ to 12 and t-values between 3.869 and 4.292; $p < 0.001$ from the 13th minute to the end). SD is significantly smaller since $\tau = 8$ to 30 (for $\tau = 8$, $F = 2.670$; $p = 0.033$, for 9, $F = 3.176$; $p < 0.05$, for $\tau = 10$, $F = 3.176$; $p = 0.013$, for $\tau = 11$, $F = 4.543$; $p < 0.01$ and from $\tau = 12$ to 30, F-values between 5.525 and 7.369; $p < 0.001$). Moreover, the mean RMSE and SD of the AdANN-validation dataset decreases as a function of the amount of data used in the AdANN-training dataset, affording more accurate performance. In addition, there is a roughly 40% improvement in the performance of the AdANN-validation dataset after adaptation.

Comparison of RMSE between the different τ for the Ad-ANN-validation dataset (Fig. 7) shows that compared to $\tau = 1$, RMSE mean is significantly higher for $\tau = 2$–30 (t-values between 3.945 and 4.517). Compared to $\tau = 2$, RMSE mean is significantly higher for $\tau = 6$–27 (t-values between 3.774 and 4.228). Thus, more than 2 time-samples are required to significantly improve the performance of the ANNs.

3.2.2. Prediction of time of occurrence of impaired driver state

Fig. 8 shows the mean and the SD of RMSE for the cross-subject validation for the different datasets as a function of the amount of data

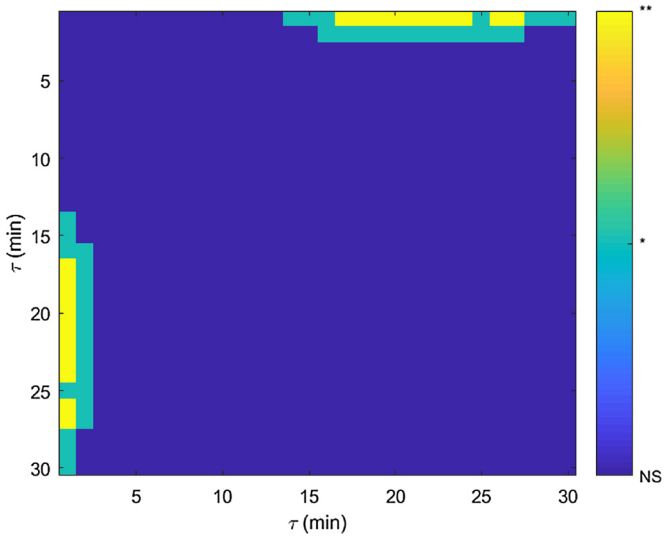

Fig. 7. P values for RMSE means compared between the different τ for the Ad-ANN-validation dataset after adaptation.

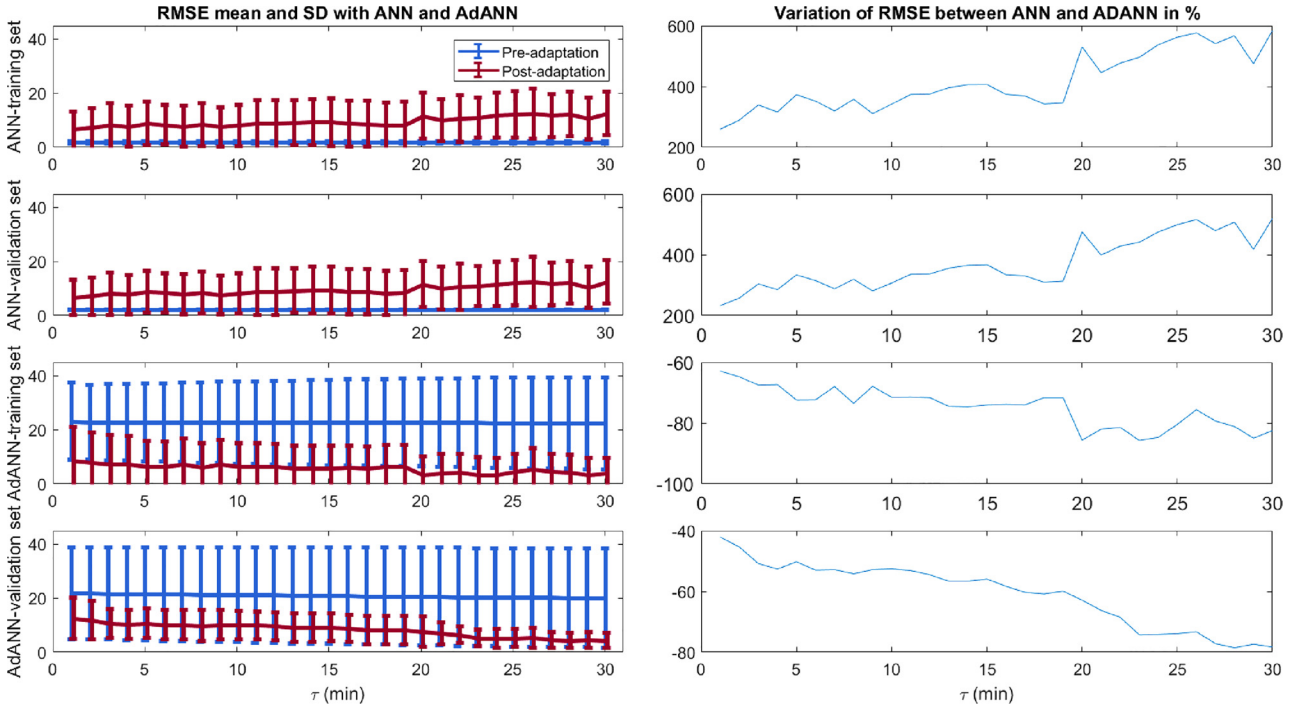

Fig. 8. Prediction: Mean and SD of RMSE for the different datasets (ANN-training set, ANN-validation set, AdANN-training set, AdANN-validation set) as a function of amount of data (τ, in minutes) used to adapt the AdANN-training dataset.

in the AdANN-training dataset used to adapt the ANN. For both the ANN-training and the ANN-validation datasets, RMSE increases significantly after adaptation (for $\tau = 1–20$, t-values between -5.239 and -3.340; $p < 0.01$ and for $\tau = 1–21$ to 30, t-values between -6.021 and -4.657; $p < 0.001$). SD is always significantly higher for both ANN datasets (F-values between 0.000 and 0.002; $p < 0.001$). For the AdANN-training dataset, the RMSE mean is always significantly lower after adaptation (t-values between 2.272 and 5.030; $p < 0.001$) but the SD is only lower for $\tau = 8–30$ (for $\tau = 8–11$, F-values between 2.642 and 3.300; $p < 0.05$, for $\tau = 12–19$, F-values between 3.300 and 3.245; $p < 0.01$, for $\tau = 20–30$, F-values between 4.797 and 9255; $p < 0.01$). For the AdANN-validation dataset, mean RSME is significantly lower after adaptation (for $\tau = 1$ min, $t = 3.535$; $p = 0.001$; t-values between 2.272 and 5.031; $p = 0.001$ for $\tau = 2–30$), while SD is significantly lower after adaptation for $\tau = 8–30$, (for $\tau = 8–11$, F-values between 2.642 and 3.300; $p < 0.05$, for τ 12–19, F-values between 3.378 and 3.245; $p < 0.01$, for τ to 20–30, F-values between 4.797 and 9.255; $p < 0.001$). For the AdANN-validation dataset, there is a small decrease followed by a stabilization of RMSE as a function of τ.

This can be confirmed by determining the p-value, comparing each τ with each other τ (Fig. 9). Compared to $\tau = 1$, RMSE mean is significantly lower for $\tau = 21–30$ (t-values between 4.377 and 7.010). Compared to $\tau = 2$, RMSE mean is significantly lower for $\tau = 21–30$ (t-values between 4.658 and 6.102). Compared to $\tau = 3$, RMSE mean is significantly lower for $\tau = 23$ and 27–30 (t-values between 3.933 and 4.361). Compared to $\tau = 4$, RMSE mean is significantly lower for $\tau = 27$, 28 and 30 (t-values between 3.885 and 4.016). Compared to $\tau = 5$, RMSE mean is significantly lower for $\tau = 23–30$ (t-values between 3 and 4.444). Compared to $\tau = 6$, RMSE mean is significantly lower for $\tau = 28$ and 30 ($t = 3.932$ and 3.913, respectively). Compared to $\tau = 7$, RMSE mean is significantly lower for $\tau = 28$ and 30 ($t = 3.955$ and $t = 3.936$, respectively). Compared to $\tau = 8–30$, RMSE is not significantly different; thus RMSE can be considered as stable at this level. Moreover, there is an almost 80% improvement in performance in the AdANN-validation dataset after adaptation.

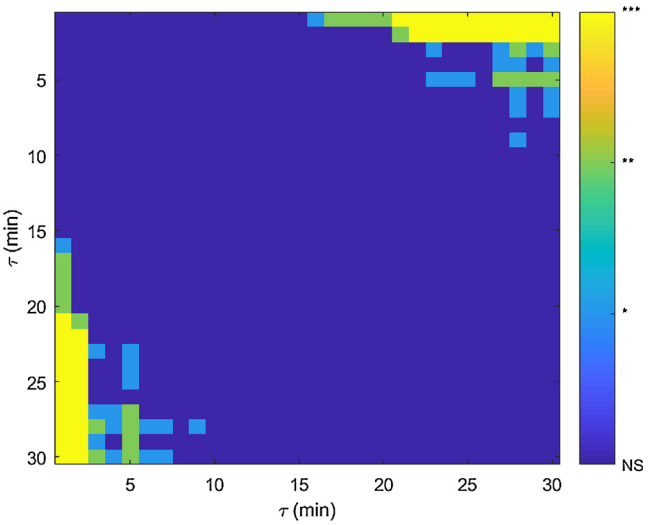

Fig. 9. P-value for RMSE mean comparing each τ with each other τ for the AD-ANN-validation dataset after adaptation.

3.3. Subject-specific performance of ANN adapted to one driver and tested with another

In this session, the ANN was trained with data from 19 participants and adapted with a 20th called A. Then the AdANN was tested on the 21 st participant, called B. Applied to all participants, both for the detection and the prediction of drowsiness, this procedure was performed 21 × 20 times (with a different A and B each time). Our objective was to determine whether the adaptation is specific to one participant or can improve model performance for another participant whose data is unknown to the ANN. To limit calculation time, the best parameters (number of neural units in the hidden layer, μ and τ values used for adaptation and best sources of information) defined in the first section of results (see 3.1) were chosen.

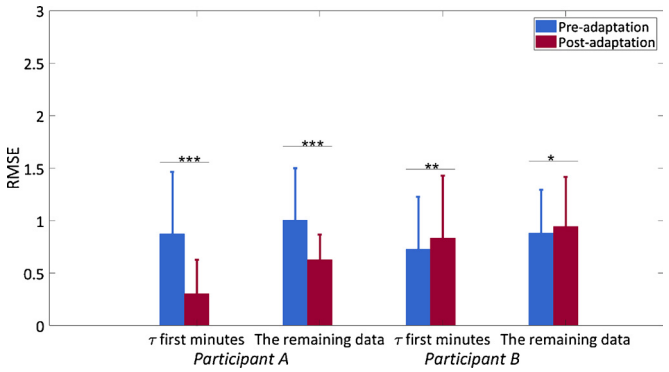

Fig. 10. Detection of level of drowsiness: RMSE mean and SD for participants used for the adaptation (A) and other participants (B) never previously encountered by the ANN, before and after adaptation with the τ first data recorded on participant A. Best τ is used here. Stars represent the level of significance (NS: $p > .0.05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

3.3.1. Detection of level of drowsiness

On average (Fig. 10), for the participant used for the adaptation (A), RMSE mean and SD are significantly lower after adaptation (for the τ first segment of the data, $t = 10.696$; $p < 0.001$, $F = 14.213$; $p =$ and for the remaining data, $t = 11.636$; $p < 0.01$, $F = 10.496$ $p =$ for mean and SD respectively). For the other tested participant (B), RMSE mean and SD are significantly higher after adaptation (for the τ first segment of the data, $t = -2.737$; $p = 0.006$, $F = 0.482$; $p < 0.001$ and for the remaining data, $t = -2.028$; $p = 0.043$, $F = 0.670$; $p < 0.001$ for mean and SD respectively).

3.3.2. Prediction of time of occurrence of impaired driver state

On average (Fig. 11), participant A's RMSE mean and SD are significantly lower after adaptation (for the τ first segment of the data, $t = 10.999$; $p < 0.001$, $F = 23.454$; $p < 0.001$ and for the remaining data, $t = 8.431$; $p < 0.001$, $F = 80.829$; $p < 0.001$ for mean and SD respectively). The other tested participant B's RMSE mean and SD are significantly higher after adaptation (for the τ first segment of the data, $t = -2.729$; $p = 0.007$, $F = 0.482$; $p < 0.001$ and for the remaining data, $t = -2.507$; $p = 0.012$, $F = 0.348$; $p < 0.001$ for mean and SD respectively). In the case of this B, an outlier participant was removed from the analysis because post-adaptation RMSE is too high (around 10^5), although adaptation also increases RMSE for this outlier participant.

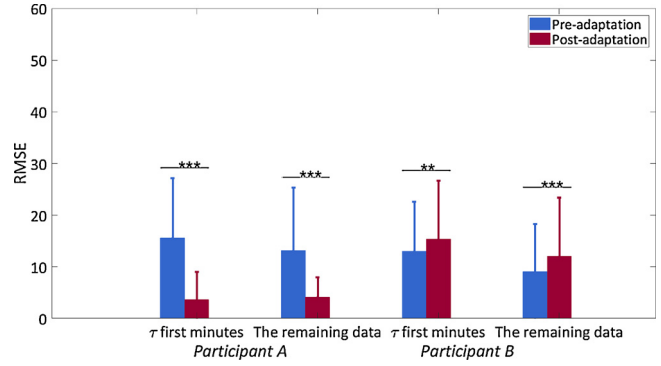

Fig. 11. Prediction of impaired driver state: RMSE mean and SD for participants used for the adaptation (A) and other participants never previously encountered (B), before and after adaptation with the τ first data recorded on participant A. Stars represent the level of significance (NS: $p > .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

4. Discussion

To maintain performance and safety, driving a car requires physiological and cognitive resources. While much attention has been given to detecting driver drowsiness and, more recently, to predicting a degraded state (Jacobé de Naurois et al., 2017), predicting the time of occurrence of a given state of drowsiness remains a challenge. A major issue is the inter-individual variability found even in a very monotonous situation. Seeking a methodology for driver-specific detection and prediction of drowsiness, the objective of this study was to test whether a model trained with a limited set of driver data could be easily and efficiently adapted to a further driver, using only a limited set of data from this driver.

The inter-rater reliability about the ground truth was high as the two independent ratings were quite close enough, the trend was the same for both raters, excepted that in a few cases one rater detected a change a little bit in advance (in a range between 1 to 5 min). Inter-rater correlation showed a similar value of R than in Weirwille and Ellsworth. (1994). Nonetheless, a potential bias in detection might be suspected from the fact that the ground truth is based on subjective evaluations from video recordings of the participant's motor behavior, which could be thought to explain the superior performance of the behavioral dataset. However, it is worth noting that these features are consensually described in the literature as the most objective and pertinent indicators of drowsiness. It is therefore difficult to conclude on whether the high performance of a model trained with behavioral data is due to the way ground truth is set or to the greater relevance of this particular set of data.

Secondly, not all types of data were found to have the same effect on adaptation. The model performs best on prediction using behavioral data alone, while the best performance on detection is achieved using the full dataset (behavioral, physiological, car data, as well as personal information and driving time). In this latter case, however, there is no significant difference between the different combinations of dataset. Similar to our previous study (Jacobé de Naurois et al., 2017). Daza et al. (2014) obtained better results with features extracted from eyelid movements (such as PERCLOS) than with features extracted from car data. Taken together, these results suggest a predominant role of oculomotor data compared to other types of data. Furthermore, when using different features, the neural network has to learn the dependencies between these different kinds of information. While this poses no problem in terms of detection, it is more of an issue for temporal prediction. When the aim is prediction, adding different types of information could be counterproductive due to the added complexity.

Thirdly, our innovative study shows that adapting an ANN to a participant whose data were not included in the initial ANN training, using only a few minutes of recorded data, can significantly improve the model's performance for this participant. This is consistent with the findings of another study in detection of drowsiness with transfer learning and EEG signals (Wu et al., 2015). In other words, the performance of the ANN, for each new participant, is enhanced after adaptation. For detection of drowsiness, the performance of the model is improved by about 40%, and for prediction, improvement reaches 80%. The difference in performance improvement between the two models can be explained by the fact that the outputs have different scales, or granularities: drowsiness is coded with 9 levels, between 0 and 4 by steps of 0.5, while time before an impairment state is between 0 and 60, by steps of 1 min. The improvement after adaptation corresponds to ½ level for detection of drowsiness and 15 min for prediction. The improvement in performance is significant above a threshold of data input used to adapt the model: two minutes (for prediction) or three minutes (for detection). In addition, increasing data up to 15 min further improves performance, though using an even larger amount of data (from 16 to 30 min) does not significantly improve performance further. An important result is that performance is improved using only a limited amount of data for model adaptation. This can probably be

explained by the fact that our ANNs are simple, not deep, networks. It is worth noting that the adapted ANN performs much better on the participant used for the adaptation. Its performance is worse when applied to data used for the initial (general) training, and even worse when applied to a previously unencountered participant, whose data were not used to adapt the ANN. Thus, adaptation with data from a particular participant makes the ANN very subject-specific, and less efficient for others. To our knowledge, this particular angle has never before been explored in the context of adaptation, at least concerning driver drowsiness.

However, while adapting the model to a previously unencountered participant gives better results than testing this new participant with a model trained on data from others, does adaptation really improve ANN performance compared to its performance with the participants used for the training? To answer this question, we need to compare the present results (with adaptation) to results from the previous study (Jacobé de Naurois et al., 2017), which used random 10-fold cross-validation with the same protocol.

In this previous study, the average RMSE (between different sources of information) was 0.49 (1/2 level) for detection of drowsiness and 2.33 min for temporal prediction of drowsiness. In the present study, the average RMSE after adaptation to a new participant is 0.93 (less than 1 level) for detection of drowsiness and 6.15 min for prediction. Therefore, adapting the ANNs (with a limited amount of data) is not as efficient as training the ANN with a larger set of data from this same participant, among others. However, our results show that, although less efficient, the adaptation method is also more parsimonious, and allows the model to be rapidly tuned to a new driver.

The difference in performance between the two methods may also be due to the high inter-individual variability of the effects of drowsiness on driving performance and on differences in physiological signals (Liu et al., 2009; Van Dongen et al., 2004aa; b). This may increase the gap between what is learned during the training phase and what is observed during the testing phase, due to overfitting. In machine learning, overfitting occurs when a greater difference in error is observed for the test dataset than for the validation dataset. One solution may be to train one ANN for each participant. However, this would obviously require a large amount of data from each driver, and thus quite a long driving time to collect this data, before the ANN was operational. Yet an ANN previously trained with a larger set of data only needs a limited dataset to be quickly (though less accurately) adapted (customized) to a new driver. It can be speculated that extending the adaptation, i.e. collecting more data from the new driver, might further improve performance and customization. This was not possible in our study, since using more data for adaptation means less data for the test phase (this is why here we limited the range of τ to only 1 to 30 min, and reserved the remaining data, from $\tau + 1$ to 110 min, as the test set).

Other studies on transfer learning or adaptive learning, for example through a BCI (Brain Computer Interface), either randomly selected data from one participant as their dataset for transfer or adaptation, or used a dataset from another session (Wei et al., 2015; Wu et al., 2015, 2016). To our knowledge, the present study is the first to use data initially recorded within the session to adapt the ANN for drowsiness detection. Actually, using randomly selected lines of data in a real car would probably give more information about variability; however, it would significantly increase the time required for the adaptation since it would involve more data collection.

Finally, while the present study addressed inter-individual variability, not intra-individual variability, the latter is also a highly relevant issue. Obviously, people do not behave and react exactly in the same way on different days, nor do they present the same tendency to fall asleep during driving. For a given individual, a large range of factors may influence the risk of becoming drowsy, such as quality and duration of sleep, daytime activity, health, medication, drug and/or alcohol use, etc. Although one individual's performance was found to be stable after sleep deprivation repeated on different days (Van Dongen et al., 2004aa), a recent study on a car driving simulator (Nilsson et al., 2017) found intra-individual differences when comparing several day-time and night-time sessions. Here, with the dataset collected and the modeling approach used, it is impossible to decide whether adaptation is only necessary once or whether an adaptation phase is needed every time a given individual starts driving on a new day. To address this issue, the same participant would have to be tested at least a second time, on a different day, as proposed by Zhang et al. (2017). This would involve a further study addressing intra-individual variability and assessing the optimal frequency of adaptation renewal.

Obviously, the number of participants used here to create the generalized model and test its customization was limited. We deliberately chose to study only a specific population (young people) at a particular time of day (post-lunch, attention dip) and in a static driving simulator. Different drowsiness dynamics are usually observed in different conditions (Anund et al., 2018; Fors et al., 2018). The transfer of knowledge from a younger population to an older one, from one-time slot to another, from a non-professional driver to a professional driver (Anund et al., 2018), or from a driving simulator to real road driving, are situations where transfer learning and adaptive learning could be further tested. Such an approach would allow the implementation of a general model (for instance in a new car), its progressive adaptation to a new driver, and longer term, its continuous adaptation to this driver to take account of changes in his/her behavior, mood, health, etc.

5. Conclusion

In this study, ANNs were adapted for a new participant from very limited data and used either to detect this driver's drowsiness level or to predict the onset of an impaired driving state. The ANNs were trained on a group of individuals and subsequently adapted for each specific participant to make allowance for high inter-individual variability. Our results with this new method using adaptive learning confirm that the ANN can rapidly be made subject-specific. This adaptation to a specific driver's data provides a promising first response to the challenge of high inter-individual variability, although other issues like intra-individual variability and/or driving on different days, time-on-day or different road conditions remain to be addressed. Moreover, other individual information could also be tested and added in the models such as for instance quality of rest/hours since last drive, drive task demand, lack of demands/monotony, personality trait as other methods of transfer learning… In the future, the model, trained beforehand with a limited set of drivers and used for a detection and prediction of drowsiness will probably need to be adapted to each new driver, and also updated (for a given driver) at a frequency still to be determined, in order to maintain the model in phase with potential evolutions of the driver's characteristics and behavior.

Conflict of interest

None.

Acknowledgments

This research was funded by a PhD grant from PSA Group via the OpenLab agreement with Aix-Marseille University and CNRS entitled "Automotive Motion Lab". We thank Marjorie Sweetko for correcting and improving the English manuscript, all the participants in this study, L. Marrou for helping to evaluate video recordings, P. Vars, V. Honnet, and M. Hing for their help with SCANeR® Software Developments.

References

Anund, A., Ahlström, C., Fors, C., Åkerstedt, T., 2018.). Are professional drivers less sleepy than non-professional drivers? Scand. J. Work Environ. Health 44 (1), 88–95. https://doi.org/10.5271/sjweh.3677.

Arnedt, J.T., Wilde, G.J., Munt, P.W., MacLean, A.W., 2001. How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task? Accid. Anal. Prev. 33 (3), 337–344.

Awais, M., Badruddin, N., Drieberg, M., 2014. A non-invasive approach to detect drowsiness in a monotonous driving environment. TENCON 2014 - 2014 IEEE Region 10 Conference 1–4. https://doi.org/10.1109/TENCON.2014.7022356.

Beale, M., Hagan, M.T., Demuth, H.B., 1992. Neural network toolbox. Neural Network Toolbox, The Math Works 5. pp. 25.

Chen, J., Ji, Q., 2012. Drowsy driver posture, facial, and eye monitoring methods. In: Eskandarian, A. (Ed.), Handbook of Intelligent Vehicles. Springer, London, pp. 913–940.

Daza, I.G., Bergasa, L.M., Bronte, S., Yebes, J.J., Almazán, J., Arroyo, R., 2014. Fusion of optimized indicators from Advanced Driver Assistance Systems (ADAS) for driver drowsiness detection. Sensors 14 (1), 1106–1131.

De Valck, E., De Groot, E., Cluydts, R., 2003. Effects of slow-release caffeine and a nap on driving simulator performance after partial sleep deprivation. Percept. Mot. Skills 96 (1), 67–78.

Dong, Y., Hu, Z., Uchimura, K., Murayama, N., 2011. Driver inattention monitoring system for intelligent vehicles: a review. IEEE Trans. Intell. Transp. Syst. 12 (2), 596–614.

Fors, C., Ahlstrom, C., Anund, A., 2018. A comparison of driver sleepiness in the simulator and on the real road. J. Transp. Saf. Secur. 10 (1–2), 72–87. https://doi.org/10.1080/19439962.2016.1228092.

Golding, J.F., 1998. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. Brain Res. Bull. 47 (5), 507–516.

Horne, J.A., Ostberg, O., 1975. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. Int. J. Chronobiol. 4 (2), 97–110.

Horne, J., Reyner, L., 1999. Vehicle accidents related to sleep: a review. Occup. Environ. Med. 56 (5), 289–294.

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, Gör, 2006. Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. J. Sleep Res. 15 (1), 47–53.

Jacobé de Naurois, C., Bourdin, C., Stratulat, A., Diaz, E., Vercher, J.-L., 2017. Detection and prediction of driver drowsiness using artificial neural network models. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2017.11.038.

Johns, M.W., 1991. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep 14 (6), 540–545.

Karrer, K., Vöhringer-Kuhnt, T., Baumgarten, T., Briest, S., 2004. The role of individual differences in driver fatigue prediction. Third International Conference on Traffic and Transport Psychology 5–9 Citeseer.

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. 2 (2), 164–168.

Liu, C.C., Hosking, S.G., Lenné, M.G., 2009. Predicting driver drowsiness using vehicle measures: recent insights and future challenges. J. Safety Res. 40 (4), 239–245.

Nilsson, E., Ahlström, C., Barua, S., Fors, C., Lindén, P., Svanberg, B., et al., 2017. Vehicle Driver Monitoring: Sleepiness and Cognitive Load. Statens väg-och transportforskningsinstitut.

Philip, P., Taillard, J., Sagaspe, P., Valtat, C., Sanchez-Ortuno, M., Moore, N., et al., 2004. Age, performance and sleep deprivation. J. Sleep Res. 13 (2), 105–110.

Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Åkerstedt, T., et al., 2005. Fatigue, sleepiness, and performance in simulated versus real driving conditions. Sleep 28 (12), 1511.

Rossi, R., Gastaldi, M., Gecchele, G., 2011. Analysis of driver task-related fatigue using driving simulator experiments. Proc. Soc. Behav. Sci. 20, 666–675.

Rost, M., Zilberg, E., Xu, Z.M., Feng, Y., Burton, D., Lal, S., 2015. Comparing contribution of algorithm based physiological indicators for characterisation of driver drowsiness. J. Med. Bioeng. 4 (5).

Samiee, S., Azadi, S., Kazemi, R., Nahvi, A., Eichberger, A., 2014. Data fusion to develop a driver drowsiness detection system with robustness to signal loss. Sensors 14 (9), 17832 14248220.

Thiffault, P., Bergeron, J., 2003. Monotony of road environment and driver fatigue: a simulator study. Accid. Anal. Prev. 35 (3), 381–391.

Van Dongen, H.P.A., Rogers, N.L., Dinges, D.F., 2003. Sleep debt: theoretical and empirical issues. Sleep Biol. Rhythms 1 (1), 5–13. https://doi.org/10.1046/j.1446-9235.2003.00006.x.

Van Dongen, H.P.A., Baynard, M.D., Maislin, G., Dinges, D.F., 2004a. Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. Sleep 27 (3), 423–433.

Van Dongen, H.P.A., Maislin, G., Dinges, D.F., 2004b. Dealing with inter-individual differences in the temporal dynamics of fatigue and performance: importance and techniques. Aviat. Space Environ. Med. 75 (3), A147–A154.

Wang, P., Lu, J., Zhang, B., Tang, Z., 2015. A review on transfer learning for brain-computer interface classification. 5th International Conference on Information Science and Technology (ICIST), 2015 315–322.

Watson, A., Zhou, G., 2016. Microsleep prediction using an EKG capable heart rate monitor. 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) 328–329. https://doi.org/10.1109/CHASE.2016.30.

Wei, C.-S., Lin, Y.-P., Wang, Y.-T., Jung, T.-P., Bigdely-Shamlo, N., Lin, C.-T., 2015. Selective transfer learning for EEG-based drowsiness detection. 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC) 3229–3232.

Wierwille, W.W., Ellsworth, L.A., 1994. Evaluation of driver drowsiness by trained raters. Accid. Anal. Prev. 26 (5), 571–581.

Wu, D., Chuang, C.-H., Lin, C.-T., 2015. Online driver's drowsiness estimation using domain adaptation with model fusion. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) 904–910.

Wu, D., Lawhern, V.J., Gordon, S., Lance, B.J., Lin, C.-T., 2016. Offline EEG-based driver drowsiness estimation using enhanced batch-mode active learning (EBMAL) for regression. Proc. IEEE Int'L Conf. on Systems, Man and Cybernetics.

Zhang, J., Wang, Y., Li, S., 2017. Cross-subject mental workload classification using kernel spectral regression and transfer learning techniques. Cogn. Technol. Work. 1–19.