SLEEPJ, 2025, **48**, 1–14
doi:10.1093/sleep/zsaf149
Advance Access publication 3 June 2025
Original Article

Original Article

Personalized alertness prediction using video-based ocular and facial features

Manivannan Subramaniyan[1,2], Francisco G. Vital-Lopez[1,2], Tracy J. Doty[3], Ian Anlap[4], William D. S. Killgore[4] and Jaques Reifman[1,*]

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research & Development, Medical Research and Development Command, Fort Detrick, MD, USA,
[2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA,
[3]Behavioral Biology Branch, Center for Military Psychiatry and Neuroscience Research, Walter Reed Army Institute of Research, Silver Spring, MD, USA and
[4]Department of Psychiatry, University of Arizona College of Medicine, Tucson, AZ, USA

[*]Corresponding author. Jaques Reifman, Senior Research Scientist and Director, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research & Development, Medical Research and Development Command, ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012, USA. Email: jaques.reifman.civ@health.mil.

Abstract

Study Objectives: Alertness impairment is generally assessed by the psychomotor vigilance test (PVT). However, performing a PVT in the real world is not practical because it is time-consuming and interrupts everyday activities. Here, we aimed to replace the PVT with passively recorded facial videos and use these measurements to make personalized alertness-impairment predictions.

Methods: We retrospectively analyzed data from a 62-hour total sleep deprivation (TSD) challenge involving 26 healthy young adults (14 men), where every 3 hours they performed a 5-minute PVT followed by a 3-minute video recording of the face. We then extracted ocular and facial features from the first 1 minute of the videos, used the features to train linear mixed-effects models that predicted PVT mean reaction times, and used the predicted PVT to customize the unified model of performance (UMP) and make personalized alertness-impairment predictions for each participant.

Results: For the mixed-effects models, the average root mean square error (RMSE) between the measured and predicted PVT data was 39 ms (standard deviation, 9 ms). For the personalized UMP predictions based on PVT predicted from the videos, the average RMSE between the measured PVT data and the model-predicted alertness impairment was 36 ms (standard error, 5 ms), which is nearly indistinguishable from the within-participant variability of 30 ms for PVT mean reaction time under rested conditions.

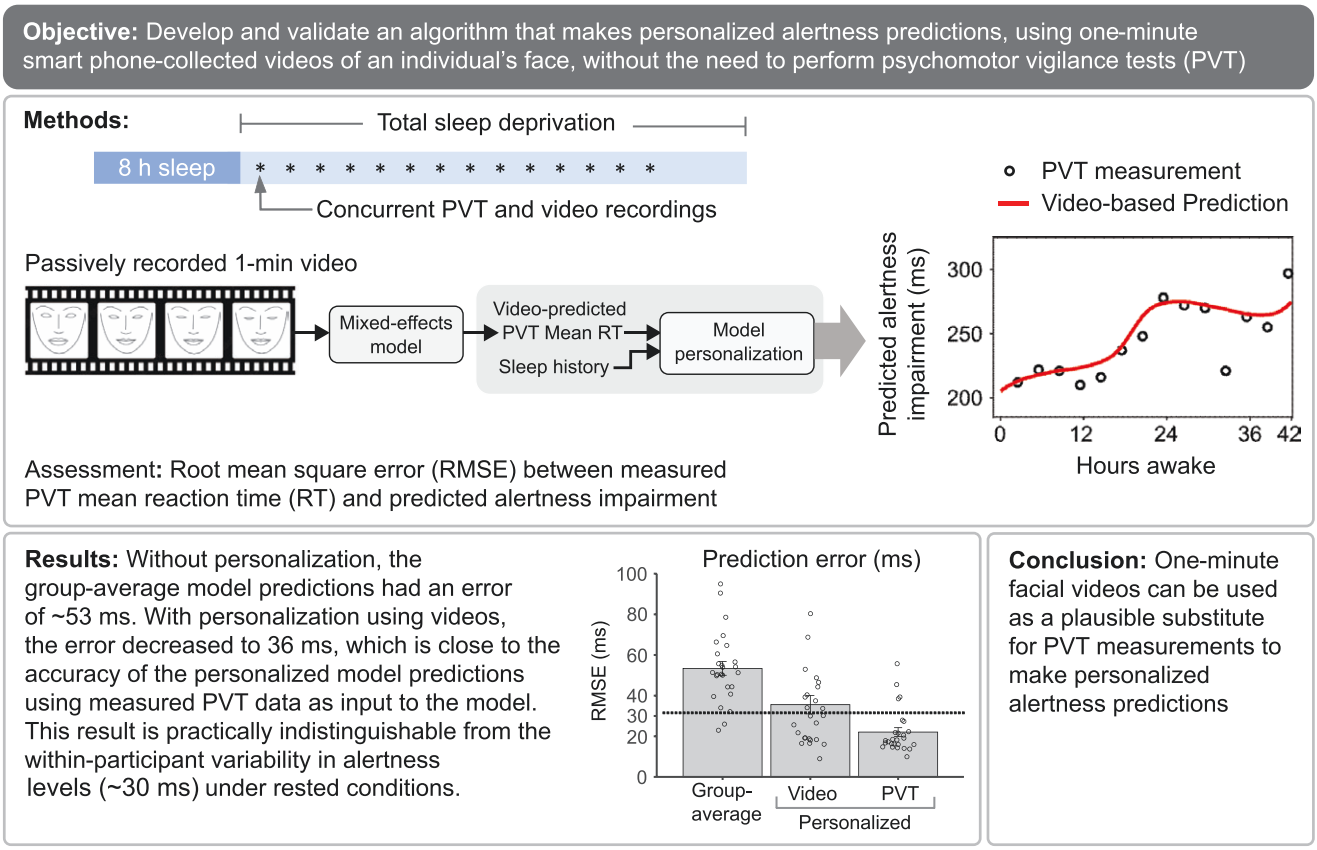
Conclusions: As a proof of principle, we developed a practical approach for predicting an individual's alertness impairment using passively recorded facial videos.

Clinical Trial Information: Title: "Real-Time Caffeine Optimization During Total Sleep Deprivation." Registration number: NCT04399083. Website: https://clinicaltrials.gov/study/NCT04399083.

Key words: alertness; eye blinks; mathematical model; psychomotor vigilance test; sleep loss; video recordings

Submitted for publication: November 27, 2024; **Revised:** May 13, 2025

Graphical Abstract



Statement of Significance

Alertness impairment, which compromises work and safety in civilian and military settings, is often assessed using the psychomotor vigilance test (PVT). However, this test, which is also used to customize alertness-prediction models, is impractical for everyday use. Here, obviating the need to perform PVTs, we developed and assessed a new approach to customize predictive models and make personalized alertness predictions using passively recorded 1-minute facial videos. This new capability offers the potential to popularize the use of personalized predictive models of alertness, which have been handicapped by the need to perform PVTs. Future efforts should focus on validating our results with videos collected outside a laboratory environment, as was done in this study.

The psychomotor vigilance test (PVT), which measures the reaction time (RT) to a visual stimulus, is the "gold standard" neurobehavioral test for assessing alertness impairment following sleep loss [1]. The PVT is a simple test, takes 5–10 minutes to complete, and is a well-validated, sensitive assay applicable to different sleep-loss challenges [2]. However, performing a PVT outside of a laboratory setting is often not practical because it interrupts everyday activities and is time-consuming, as evidenced by the small number of test results collected in shift-worker studies [3, 4]. In addition, PVT results depend on the individual's level of effort, further underscoring the practical challenges of assays that require active participation [5–7]. An ideal neurobehavioral test would assess alertness-impairment passively and unobtrusively.

As an alternative to the PVT, a number of studies have investigated the use of ocular, oculomotor, and facial features for building mathematical models that assess alertness impairment in adults. For example, Abe et al. [8] built a Bayesian model using ocular features (eye blink duration and average eye-opening degree) and oculomotor features (saccade and microsaccade properties) extracted from facial videos recorded during a 38-hour total sleep deprivation (TSD) challenge. From the videos, the model estimated the probability of each of the three levels of attention vigilance, as quantified by PVT mean RT. Other less-challenging sleep-loss studies, involving one night of 4 hours of sleep restriction [9–11] or one night of TSD [12], built logistic regression models to estimate the probability of each of two levels of fatigue [9–11] or of driving performance impairment [9, 12]. As model predictors, Akerstedt et al. [9] used variability in the vehicle lateral position during driving tasks, blink duration, and the ratio of blink amplitude to peak eyelid-closing velocity (i.e. the eyelid-closing amplitude-velocity ratio); Wilkinson et al. [11] used blink duration, inter-event duration (i.e. the interval between when the eyelid-closing and -reopening velocities reach their respective peak values), eyelid-closing amplitude-velocity ratio, and the percentage of time the eyes are fully closed; Shiferaw et al. [12] used blink rate, blink duration, fixation rate, saccade amplitude, driving duration, and binary sleep-deprivation status (i.e. yes or no); and Puspasari et al. [10] used blink duration, amplitude

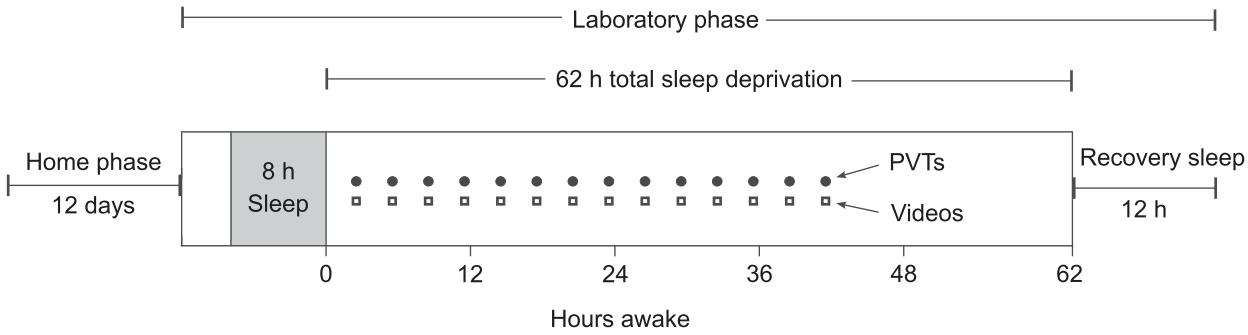


Figure 1. Study design. The study consisted of a home phase and a laboratory phase. During the 12 days of the home phase, participants maintained their habitual sleep and caffeine schedules. The laboratory phase started with an 8-hour sleep opportunity, was immediately followed by a 62-hour total sleep deprivation (TSD) challenge, and ended with a 12-hour recovery sleep. During the first 42 hours of TSD, participants performed 5-minute psychomotor vigilance tests (PVTs) every 3 hours (filled circles). Five minutes after each PVT, we obtained a 3-minute video recording (open squares) of each participant facing a mobile phone camera.

and peak velocity of saccades, the percentage of time the eyes are more than 80% closed (PERCLOS), and microsleep (blinks with a duration greater than 500 ms) frequency. A few studies also built data-driven classifiers using other machine learning approaches, including artificial neural networks [13], support vector machines [14, 15], or a collection of different machine learning models [15, 16]. These models used ocular features alone [14] or a combination of ocular and facial features [13, 15, 16] to classify individuals into one of two [14–16] or three [13] levels of alertness impairment induced by night driving or 24–30 hours of TSD.

Although these studies developed the ability to estimate alertness impairment based on facial and ocular features, all of their models–except those of Massoz et al. [14]–classified alertness impairment into discrete categories rather than quantifying it on a continuous scale. Such a quantification would help us detect declining trends in vigilance and administer countermeasures early on rather than waiting until an individual is classified as impaired. Importantly, except for a few studies [8, 13–16], others [9–12] did not assess model performance on an independent dataset, and the one study that validated model performance through cross-validation and predicted alertness impairment on a continuous scale showed a relatively large error (106 ms) in the prediction of PVT mean RT based on ocular features [14]. This error is more than three times as large as the within-participant variability of ~30 ms in alertness impairment under rested conditions [17], indicating the need for more accurate models.

Here, we aimed to develop personalized alertness-prediction models based on facial and ocular features, rather than on PVT measurements as in our previous modeling efforts [18–21]. To this end, we used laboratory data from a recent 62-hour TSD study involving 26 participants, where we collected PVT data every 3 hours (14 tests in total) and concurrently collected 3-minute facial videos from each participant. Using a 5-fold nested cross-validation (CV) procedure, we built (i.e. trained) personalized models to quantitatively predict continuous alertness-impairment levels and assessed their performance on independent data not used for model training.

Methods

Study design

To develop personalized models that predict alertness levels based on an individual's facial and ocular features, we retrospectively analyzed data from a recent sleep-deprivation study conducted at the Social, Cognitive, and Affective Neuroscience (SCAN) Lab at the University of Arizona College of Medicine (Tucson, AZ) [22]. The study included healthy men ($n = 14$) and non-pregnant, non-lactating healthy women ($n = 12$), with ages ranging from 18 to 36 years [mean = 21.6 years, standard deviation (SD) = 3.9 years] and no history of sleep disorders or physical and mental health problems (Table S1). The study consisted of a 12-day home phase, where participants maintained their habitual sleep and caffeine-consumption schedules, followed by a laboratory phase (Figure 1). The laboratory phase started with 8 hours of time in bed (23:00 to 07:00), was immediately followed by a 62-hour TSD challenge during which participants performed 5-minute PVTs every 3 hours, and ended with a 12-hour recovery sleep. Five minutes after the end of each PVT, we collected a 3-minute video recording of each participant's face as they sat facing the camera of a mobile phone (Samsung Galaxy S20). We set up the camera so that the participant's face covered approximately one-third of the height of the frame (frame height, 1920 pixels; frame width, 1080 pixels) and recorded the videos at 240 frames per second.

After the first 42 hours of TSD, participants consumed caffeine as part of the experimental design of the original study [22]. Therefore, to exclude caffeine effects, we only analyzed data collected during the first 42 hours of the TSD challenge, which consisted of 14 PVT sessions and their associated videos (Figure 1). The protocol was reviewed and approved by the University of Arizona College of Medicine Institutional Review Board and the US Army Office of Human Research Oversight.

Personalized alertness predictions

Based on the two-process model postulated by Borbély and Achermann [23], we previously developed and validated the unified model of performance (UMP) [18–21] that, given sleep history and caffeine history (optional), predicts alertness impairment as measured by the PVT mean RT. The UMP predicts the temporal patterns of alertness impairment P (Table 1, Equation 1) as a function of the circadian process, which depends on the time of day (Equation 2), and the homeostatic process, which depends on the time awake (Equation 3). The model also considers the amount of sleep debt in the prior sleep history and its effect on recovery sleep (Equations 4–6) [19]. To personalize the UMP to a particular individual [18], the model recursively adjusts the values of its five most sensitive parameters (Table 1) after each PVT measurement so that, progressively, the model predictions reflect the participant's response to sleep loss measured by the PVT. At the start of this process, the model assumes that the participant is

Table 1. Governing equations and parameter values of the unified model of performance (UMP)

Performance impairment (*P*):

$$P(t) = S(t) + \kappa C(t), \quad (1)$$

where *S* and *C* denote the homeostatic and circadian processes of the two-process model at time *t*, respectively, and κ represents the circadian amplitude.

Circadian Process (*C*):

$$C(t) = \sum_{i=1}^{5} a_i \sin\left[i\frac{2\pi}{\tau}(t+\phi)\right], \quad (2)$$

where a_i, $i = 1, \ldots, 5$, represents the amplitude of the five harmonics ($a_1 = 0.97$, $a_2 = 0.22$, $a_3 = 0.07$, $a_4 = 0.03$, and $a_5 = 0.001$), and τ and ϕ, respectively, denote the period and phase of the circadian oscillator (~24 h).

Homeostatic Process (*S*):

$$\frac{dS(t)}{dt} = \begin{cases} [U - S(t)]/\tau_w & \text{during wakefulness} \\ [L(t) - S(t)]/\tau_s & \text{during sleep} \end{cases} \quad (3)$$

where *L* and *U* denote the lower and upper asymptotes of process *S*, respectively, τ_w [mean (standard error) 23.0 (3.2) h] and τ_s [4.0 (1.0) h] denote the time constants of the increasing and decreasing sleep pressure during wakefulness and sleep, respectively. $S(0) = S_0$ and $L(0) = L_0$ correspond to the initial state values for *S* and *L*, respectively.

Lower Asymptote (*L*) of Process *S*:

$$L(t) = U \times Debt(t), \quad (4)$$

where *Debt* denotes the sleep debt.

Sleep Debt (*Debt*):

$$\frac{dDebt(t)}{dt} = [Loss(t) - Debt(t)]/\tau_{LA}, \quad (5)$$

$$Loss(t) = \begin{cases} 1 & \text{during wakefulness} \\ -2 & \text{during sleep,} \end{cases} \quad (6)$$

where τ_{LA} [7.0 (2.6) d] denotes the time constant of the exponential decay of the effect of sleep history on performance impairment.

Personalized predictions:
Customization of the model to capture an individual's sleep-loss phenotype requires that we update the value of five model parameters (U, κ, ϕ,S_0, and L_0) after each PVT measurement. The initial values (and standard errors) for these parameters, which correspond to the group-average model [20], are: $U = 497$ (31) ms, $\kappa = 75$ (7) ms, $\phi = 2.5$(0.2) h, $S_0 = 176$ (15) ms, and $L_0 = 140$ (14) ms.

an "average" individual and uses group-average parameter values obtained by fitting the model to the group-average PVT data in the study by Belenky et al. [24]. Then, after each PVT, the model uses a Bayesian learning approach to balance the weight of each PVT measurement against that of the group-average model. As the number of PVT measurements increases, the importance given to the measurements increases, leading to a personalized model that represents the individual's sleep-loss phenotype [18]. In a recent study, we showed that 12 PVTs measured during 36 hours of TSD are sufficient to personalize the UMP [25]. (Note that the UMP can also account for the effect of caffeine; however, we excluded this for simplicity because our study does not involve caffeine use.).

In addition to the PVT, changes in an individual's facial and ocular features have been shown to be indicative of alertness levels, as reported in multiple studies listed in Table 2. Therefore, in place of performing a PVT to personalize the UMP, we sought to identify facial and ocular features predictive of PVT results that we could use instead of the measured PVT as input to the UMP to personalize alertness predictions (Figure 2B). To this end, we investigated several facial and ocular features extracted from the video recordings that could be used to predict the measured PVT.

Feature extraction

From each video frame, we used the open-source Python library *dlib* [48] to extract 68 facial landmark points (Figure 3A). As a first step towards extracting ocular features, we computed "eye aspect ratio" values from the facial landmarks, as described previously [49]. Briefly, for each video frame, we computed the vertical height h_e between the upper and lower eyelids (Figure 3B), which indicated the level of eye-opening. To minimize artifactual changes in h_e due to a participant's movement relative to the camera, we normalized it by dividing it by the horizontal eye width w_e (Figure 3B and Supplementary Materials). To obtain a single eye aspect ratio h_e/w_e, we computed the ratio for each eye and averaged them. Similar to the eye aspect ratio, we computed a mouth aspect ratio by dividing the width w_m of the mouth by its height h_m (Figure 3C and Supplementary Materials). The time series of eye aspect ratios over a recording indicated the dynamics of eyelid movement, where sharp downward peaks (troughs) indicated blinks (Figure 3D). We extracted several ocular features based on these blinks (Figure S1). Table 2 shows the definition of the 15 ocular and facial features we extracted from the time series of landmarks.

Prediction of PVT based on ocular and facial features

To predict PVT data from ocular and facial features, we used linear mixed-effects models that represented the measured PVT data as a linear function of the features. These models consisted of fixed- and random-effects components, where the fixed component ("global model") captured the average linear relationship between the features and the PVT data, and the random component captured the within- and between-participant variabilities (Supplementary Materials). To this end, we split the study data into training and testing datasets, where we fitted (trained) the models using participants

Table 2. Definition of facial and ocular features investigated in the study

Feature number	Feature name (units)	Feature definition	Reference
1	Baseline eye-opening level (unitless)	Mean of the baseline eye aspect ratio (horizontal dashed traces in Figure S1; Supplementary Materials)	[13, 26][a]
2	Blink amplitude*,† (unitless)	Vertical distance between the baseline eye aspect ratio and the trough of a blink event (vertical dashed line in Figure S1B)	[9, 13, 27]
3	Blink rate (min^{-1})	Number of blinks per minute	[12, 13, 15, 27–31]
4	PERCLOS (%)	Percentage of time the eyes were more than 80% closed (Supplementary Materials)	[32–37]
5	Eyelid-closing velocity† (s^{-1})‡	Magnitude of the peak instantaneous velocity during the eyelid-closing phase (Figure S1B)	[9, 13]
6	Eyelid-reopening velocity† (s^{-1}) ‡	Magnitude of the peak instantaneous velocity during the eyelid-reopening phase (Figure S1B)	[9]
7	Eyelid-velocity ratio† (unitless)	Ratio between eyelid-closing velocity and eyelid-reopening velocity	§
8	Eyelid-closing duration† (s)	Interval between the time when the eyes were 33% closed during the eyelid-closing phase and the time when the eyelid-closing phase ended (thick dark horizontal lines in Figure S1B)	[14, 38]
9	Eyelid-reopening duration† (s)	Interval between the time when the eyelid-reopening phase began and the time when the eyes were 33% closed during the eyelid-reopening phase (thick light horizontal lines in Figure S1B)	[14, 38]
10	Blink duration† (s)	Interval between the time when the eyes were 33% closed during the eyelid-closing phase and the time when the eyes were 33% closed during the eyelid-reopening phase (Figure S1B)	[8–15, 27, 38–46]
11	Eyelid-closing amplitude-velocity ratio† (s)	Ratio between blink amplitude and eyelid-closing velocity	[9, 11, 13, 34, 42, 43, 46]
12	Eyelid-reopening amplitude-velocity ratio† (s)	Ratio between blink amplitude and eyelid-reopening velocity	[11, 34, 40, 42, 43]
13	Head-movement velocity (a.u./s)	Mean instantaneous velocity of the head	[15, 16][a], [47]
14	Variance of head-movement velocity (a.u./s^2)	Variance of the instantaneous velocity of the head	[47][a]
15	Baseline mouth aspect ratio (unitless)	Median value of the time series of the mouth aspect ratio	[26][a]

[a]Studies reporting metrics that essentially captured the characteristics of the feature reported here.
*The blink amplitude served as a reference to define the level of eye closure. For example, we assumed that the eyes were 80% closed when the eye aspect ratio decreased from baseline to 80% of the blink amplitude.
†Metrics averaged across the blinks detected in a video recording.
‡Units for velocity, s^{-1}, derived from the change in normalized eye-opening height (unitless) per second.
§ Feature not previously studied.
a.u., arbitrary units.

from the training datasets and assessed model predictions using participants from the testing datasets. We developed these models using a 5-fold nested CV procedure (Steps 1–9, Figure S2), in which we trained the global models using different subsets of participants and extracted features. To obtain the feature subsets, in each of the five outer CV folds of the CV procedure, we used a forward selection procedure, where at each cycle we added the next most informative feature to the model and at the end selected the feature subset with the best model performance (Steps 2–6, Figure S2). The data used to *train* a global mixed-effects model to obtain the video-predicted PVT mean RT for a specific participant did not include any data (measured PVT or extracted ocular/facial features) from the participant we wished to predict. Specifically, to train a global model for a participant, we used measured PVT data and video-extracted features collected from approximately 20 other participants in the training dataset (boxes labeled "Train" in Figure S2), explicitly excluding the specific participant.

To independently assess model performance for the participants in the testing dataset not used for model training, we computed the root mean square error (RMSE) between the model predictions (i.e. the video-predicted PVT mean RT) and the measured PVT data (Steps 11a and 12a, Figure S2). To reduce the effect of outliers in the selection of the training and testing datasets, we repeated the CV procedure 100 times (Step 10, Figure S2), each time using a different subset of participants (and hence different features to construct 100 distinct global models for each specific participant). Finally, to obtain the video-predicted PVT mean RT for a specific participant, we used video-extracted features from this participant as input to each of 100 distinct global models and averaged the predicted results (Steps 11b and 12b, Figure S2). This average, obtained without using any information from the specific participant during model training, served as the video-predicted PVT mean RT for that participant for the time the video was recorded.

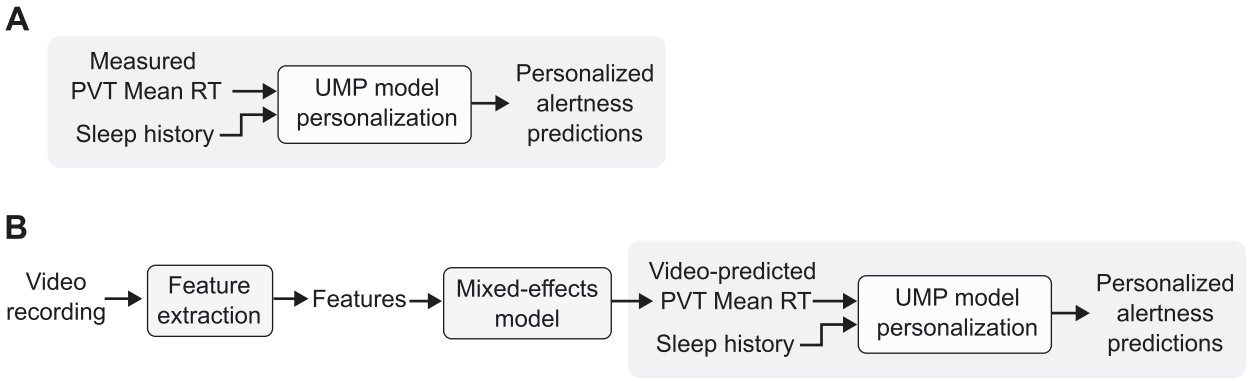


Figure 2. Inputs and outputs of the personalized unified model of performance (UMP) and the process of using ocular and facial video-extracted features to make personalized alertness predictions. (A) The UMP is personalized to an individual based on the individual's sleep history and mean reaction times (RTs) measured by a psychomotor vigilance test (PVT). (B) Instead of using the measured PVT mean RTs, we propose to use predicted PVT mean RTs from videos to personalize the UMP. To this end, we built linear mixed-effects models that predicted PVT mean RTs from facial and ocular features extracted from video recordings.

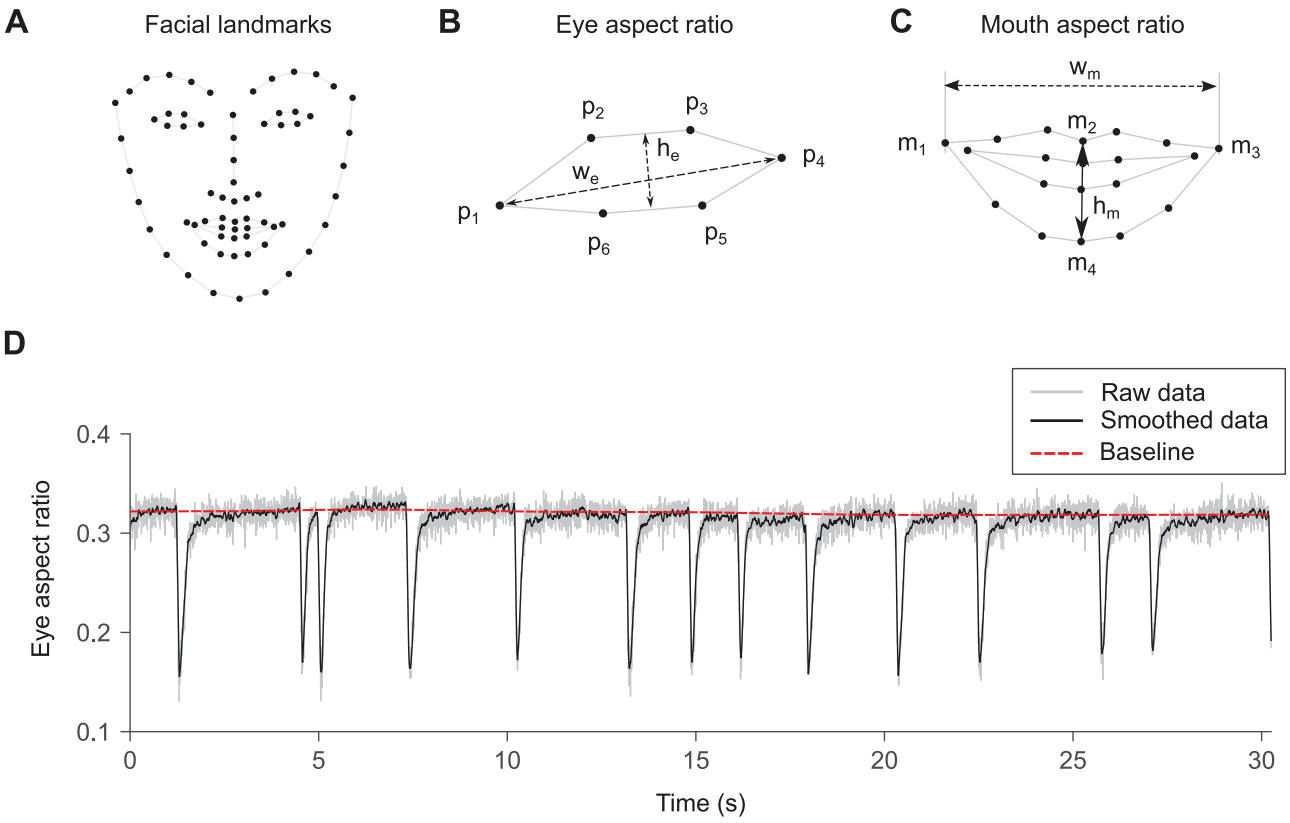


Figure 3. Feature extraction from facial videos. (A) Facial landmarks (68 points, filled circles) from a video frame of a representative participant. (B) Eye landmarks (corresponding to the left eye in panel A) from which we computed the eye aspect ratio as the ratio between the vertical eye-opening height h_e and the horizontal eye width w_e. (C) Mouth landmarks (from panel A) from which we computed the mouth aspect ratio as the ratio between the width of the mouth w_m (the distance between points m_1 and m_3) and the height of the mouth h_m (the distance between points m_2 and m_4). (D) Eye aspect ratio time-series data (light trace, raw data; dark trace, smoothed data) averaged across the left and right eyes for a representative participant. The dashed trace indicates the baseline of the eye aspect ratio, and the downward peaks (troughs) indicate blinks, from which we extracted several ocular features.

Video-length analysis

To assess the effect of video duration on the accuracy of the video-predicted PVT mean RT, we analyzed frames of video lengths ranging from 10 to 180 seconds of each recording. Then, from these frames, we extracted features, predicted PVT data, and assessed model performance.

Video-based personalization of the UMP

To personalize the UMP for each participant using video recordings, we followed the procedure discussed above (*Personalized alertness predictions*), except that we replaced each PVT measurement with the average of the predicted PVTs over the 100 mixed-effects model predictions for each video recording (Step 12b, Figure S2). Accordingly, after each prediction corresponding to one of the

14 video recordings of a given participant, we adjusted the UMP model parameters by providing as inputs to the UMP the averaged video-predicted PVT data along with the sleep history (Figure 2B). After adjusting the UMP parameters using all of the 14 predicted PVTs, we obtained the personalized UMP, which we used to predict alertness impairment for the corresponding participant for a given time of day. To obtain a measure of variability around the predictions of the personalized UMP for each participant, we repeated the model-personalization procedure 100 times, each time using the predicted PVT data from one of the 100 mixed-effects models of the CV procedure. Then, from the 100 video-based personalized UMP predictions, we computed a percentile-based 95% confidence interval.

Assessment of video-based personalization of the UMP

To assess the video-based personalized UMP's ability to learn an individual's response to sleep loss, for each participant we computed the RMSE between the measured PVT data and the time-matched UMP predictions (i.e. UMP predictions calculated at the time points when PVT measurements occurred). We also computed the RMSE between the predictions of the personalized PVT-based UMP (i.e. using the actual PVT measurements to individualize the model) and the measured PVT data. Finally, to assess the benefit of the personalized video-based UMP over a model with no individualization, i.e. a group-average UMP [25], we computed the RMSE between the measured PVT data and the predictions of the group-average model, where the only input to the model was sleep history.

Statistical analyses

We compared the prediction accuracy of the three types of UMP models (group-average UMP, personalized PVT-based UMP, and personalized video-based UMP) using repeated-measures analysis of variance (ANOVA) in MATLAB with Tukey-Kramer post hoc tests. Using the same statistical analysis, we also compared the accuracy of video-predicted PVT mean reaction times obtained from mixed-effects models built using 1-, 2-, or 3-minute videos. To assess whether individual facial and ocular features were predictive of the measured PVT data, we performed a univariate analysis by separately fitting a linear mixed-effects model to each feature and computing two complementary R^2 metrics: the marginal R_m^2 and the conditional R_c^2 [50]. R_c^2 captured the proportion of the variance explained by a participant-specific model (i.e. the full model, which includes both the fixed and the random effects), whereas R_m^2 captured the proportion of the variance explained by the global model (i.e. the fixed-effects component of the model) and provided a measure of the ability of the model to predict PVT data for participants not used for model fitting. Therefore, we used the R_m^2 values associated with the individual features to rank them and used a cutoff value of 0.1 to identify features that should be considered informative in predicting PVT data. The reported p values indicated the statistical significance ($p < .05$) of the slope of the univariate mixed-effects model fits not being equal to zero. To assess the extent of agreement between the measured and predicted data, we used the 14 data points for each participant to compute the concordance correlation coefficient (CCC) [51], using the *epiR* statistical package in R [52]. For each participant, we computed two CCC values: one for the measured PVT versus the video-predicted PVT and the other for the measured PVT versus the personalized video-based UMP alertness (i.e. PVT) prediction.

Results

Video-length determination

Our analysis of the effect of video duration on the ability to predict PVT data indicated that videos ≥1 minute captured at least one blink in each of the 14 video recordings of each participant, with no significant group differences in RMSE between the measured and video-predicted PVT data across 1-, 2-, or 3-minute videos [repeated-measures ANOVA, $F(2,50) = 0.05$, $p = .95$, for differences in RMSEs ≤ 0.20 ms over the 100 predictions for each of the 26 participants]. Therefore, we report results based on videos of 1-minute duration.

Feature assessment

To assess whether extracted features were predictive of PVT data, we separately fitted linear mixed-effects models to each of the 17 features (the 15 ocular and facial features in Table 2 plus age and sex). We observed large participant-to-participant variations in the range of feature values and PVT data. Figure 4 shows the models for three of the 17 features, including the global model based on all participants (thick line) and individual models (thin lines), where we captured individual differences by including participant-specific intercepts, and Table 3 shows the statistical significance of the fitted slopes and the proportion of variance explained by the models, as measured by the marginal R_m^2 and the conditional R_c^2. The top three features ranked by R_m^2 were baseline eye-opening level, PERCLOS, and eyelid-closing velocity. For these features, R_m^2 ranged from 0.13 to 0.18, indicating that the global model only captured a small proportion of the total variance. The corresponding R_c^2 values were considerably larger (0.54–0.58), indicating that the between-participant variability (see arrows in the bottom panel of Figure 4) contributed substantially towards the total variance. Analysis of the model slopes indicated that the PVT data decreased with baseline eye-opening level and eyelid-closing velocity, but increased with PERCLOS ($p < .001$). For the remaining 14 features, R_m^2 was less than 0.10, suggesting that these features are likely to be less predictive of PVT data, and R_c^2 ranged from 0.45 to 0.61, indicating large between-participant variability (Figure S3). Analysis of the slope of the model for these features indicated that there was no significant linear relationship between the PVT data and five features, i.e. age, sex, variance of head-movement velocity, blink rate, and eyelid-reopening amplitude-velocity ratio ($p > .06$). The remaining nine features showed a significant linear relationship with the PVT data ($p < .05$).

PVT data prediction from videos

Model selection

To develop multivariate mixed-effects models to predict PVT data (Figure 2B), in each of the five outer CV folds of the repeated (100 times) nested CV procedure, we used a forward selection procedure. Table 3 (last column) shows the frequencies at which the procedure selected each feature. The top three most frequently selected features appeared in 29%–51% of the 5 × 100 models, while the remaining features appeared in less than 29% of the models. The median number of features in the models (Step 6, Figure S2) was 2 (interquartile range, 1–3), indicating model parsimony.

Model evaluation

To evaluate the performance of the models on data not used for model training, we computed the RMSE between the video-predicted and the measured PVT data in the test datasets (Step 8,

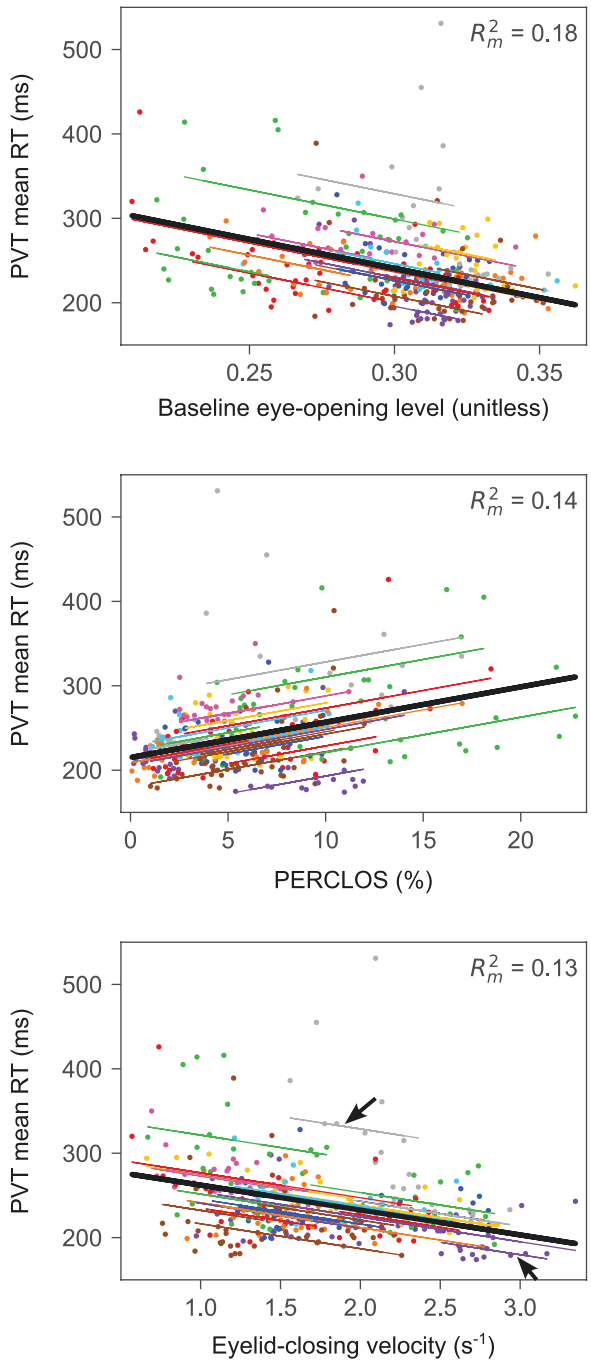


Figure 4. Linear association between psychomotor vigilance test (PVT) mean reaction time (RT) and individual ocular features, as determined by univariate linear mixed-effects model fit. The panels show model fitting for the top three features ranked by the marginal R^2, R^2_m. In each panel, the thick line indicates the global model fit (i.e. the fixed-effects component of the mixed-effects model), and the thin lines indicate participant-specific model fits for the 26 participants using the fixed and random effects. The arrows in the bottom panel mark two participants with very different ranges of predictor and response variable values. The fitted slopes were significant for each of the three features ($p < .001$). PERCLOS, percentage of time the eyes were more than 80% closed.

Figure S2). The average RMSE across the repetitions of the 5-fold CV (Step 12a, Figure S2) was 39 ms (SD, 9 ms). The average R^2_m and R^2_c across the repetitions of the 5-fold CV (Step 12a, Figure S2) were 0.18 (SD, 0.06) and 0.55 (SD, 0.05), respectively, indicating that the fixed-effects component captured a moderate amount of the total variance and that a large proportion of the variance came from participant-to-participant variations. In addition, for each participant, we computed the CCC between the video-predicted PVT mean RT data (closed circles in Figure 5) and the measured PVT mean RT data (open circles in Figure 5). We observed considerable between-participant variability, with generally low CCC values, which ranged from -0.19 to 0.50 (Figures 6A and Figure S4).

Personalization of the UMP

To assess the utility of using facial and ocular features to develop personalized predictive models of alertness, we compared the performance of the UMP developed using three different inputs to the model: (1) sleep history only (group-average UMP), (2) sleep history and measured PVT data (personalized PVT-based UMP), and (3) sleep history and video-predicted PVT data (personalized video-based UMP). Figure 5 shows the predictions of each of the three models along with the 95% confidence intervals for the predictions of the video-based personalized UMP (shaded areas), the measured PVT data (open circles), and the video-predicted PVT data (closed circles). The 26 participants in the figure were sorted from the most resilient to the most vulnerable to sleep loss (Supplementary Materials). Some of the most extreme participants in terms of sleep-loss phenotype (e.g. #1, #25, and #26) showed the largest differences between the two personalized models, although some participants with a less extreme sleep-loss phenotype (e.g. #9 and #11) also showed large differences. For other participants (e.g. #3, #4, #7, and #12), the two personalized models yielded very similar performance, which occurred when the video-predicted PVT data matched the measured data. In contrast, for the majority of the participants, the group-average UMP did not perform as well as either of the two personalized models.

To obtain an overall assessment of the models, we averaged the RMSEs across the 26 participants for each of the three UMP models (Figure 7). A repeated-measures ANOVA yielded a significant group difference in the averaged RMSE values [$F(2,50) = 44.8$, $p < .001$]. Tukey-Kramer post hoc tests indicated that the personalized video-based UMP model yielded an average RMSE that was significantly smaller than that of the group-average UMP {36 ms [standard error (SE), 5 ms] vs. 53 ms (SE, 3 ms); $p < .001$} and significantly larger than that of the personalized PVT-based UMP model [36 ms (SE, 5 ms) vs. 22 ms (SE, 2 ms); $p < .001$]. Importantly, for nearly half of the participants ($n = 12$), the personalized video-based UMP models yielded RMSEs that were lower than the conservatively estimated within-participant variability of 30 ms for PVT mean RT [17] (dotted horizontal line in Figure 7).

We also computed the CCC values to assess the extent of agreement between the measured PVT mean RT data (open circles in Figure 5) and the predicted PVT data obtained using the personalized video-based UMP predictions (dashed lines in Figure 5). Figures 6B and Figure S5 show the results, which indicated a considerably better agreement with the measured data than the video-predicted results in Figure 6A, with the CCC achieving a maximum of 0.85 for participant #12.

Discussion

Objective measurements of alertness level are based on neurobehavioral tests, such as the PVT, which are time-consuming and dependent on the individual's level of effort. Here, based on a 42-hour TSD challenge, we developed and validated models, which obviated the need for such tests and predicted personalized

Table 3. Statistical results of the univariate linear mixed-effects model to predict psychomotor vigilance test mean reaction time as a function of each of the 15 ocular and facial features extracted from 1-minute video recordings, age, and sex as well as the frequencies at which these features were selected to build the best models in the 5-fold nested cross-validation rocedure

Feature number	Feature name	R^2_m †	R^2_c#	Slope‡		Selection frequency (rank order)
				Trend	*P*-value	
1	Baseline eye-opening level	0.18	0.58	Neg	<.001***	.32 (2)
2	PERCLOS	0.14	0.54	Pos	<.001***	.51 (1)
3	Eyelid-closing velocity	0.13	0.54	Neg	<.001***	.29 (3)
4	Baseline mouth aspect ratio	0.09	0.61	Neg	.008**	.00 (17)
5	Eyelid-closing amplitude-velocity ratio	0.08	0.53	Pos	<.001***	.27 (4)
6	Eyelid-velocity ratio	0.06	0.52	Neg	<.001***	.23 (5)
7	Blink duration	0.06	0.53	Pos	<.001***	.05 (16)
8	Blink amplitude	0.06	0.49	Neg	<.001***	.07 (11)
9	Eyelid-reopening velocity	0.04	0.49	Neg	.002**	.05 (15)
10	Eyelid-closing duration	0.03	0.50	Pos	<.001***	.10 (7)
11	Eyelid-reopening duration	0.02	0.51	Pos	.001**	.06 (13)
12	Sex	0.01	0.48	Pos	.406	.07 (10)
13	Head-movement velocity	0.01	0.48	Pos	.028*	.08 (8)
14	Blink rate	0.01	0.45	Pos	.184	.11 (6)
15	Eyelid-reopening amplitude-velocity ratio	0.01	0.48	Pos	.112	.06 (14)
16	Variance of head-movement velocity	0.01	0.47	Pos	.080	.08 (9)
17	Age	0.00	0.48	Pos	.621	.06 (12)

Neg, negative slope; PERCLOS, percentage of time the eyes were more than 80% closed; Pos, positive slope.
‡Slope, slope of the mixed-effects model.
†R^2_m, proportion of the variance explained by a global model fit (i.e. a line fit with an intercept and a slope common to all participants).
#R^2_c, proportion of the variance explained by a participant-specific model fit (i.e. a line fit with an intercept that is participant specific and a slope that is common to all participants).
***$p < .001$, **$p < .01$, *$p < .05$.

alertness impairment on a continuous scale based on 1-min video recordings. Based on independent test data not used for model training, we found that the average error of 36 ms between the measured PVT data and the individualized video-based UMP predictions was nearly indistinguishable from the within-participant variability in alertness impairment (30 ms) under rested conditions. As a proof of principle, this finding suggests that for TSD we can substitute facial videos for PVTs to obtain adequate personalized alertness predictions.

Personalized alertness prediction based on video recordings involves two steps: predicting PVT data using features extracted from videos and using these data to personalize the UMP. In the first step, the 5-fold repeated (100 times) CV procedure results in 5 × 100 different mixed-effects models, each with its own feature subsets and model parameters. Hence, in real-world use, to obtain personalized alertness predictions based on a new video recording, one should extract features from the recording, use the 500 models to predict 500 PVT results, average the results, and then use the average PVT to personalize the UMP.

When building the multivariate mixed-effects models to predict PVT data from video recordings, we minimized model overfitting through cross-validation and by using a forward feature selection that only added informative features to the model. The results indicated that 75% of the models only required three or fewer features, indicating model simplicity and generalizability. Among the features used in the models, PERCLOS was the most frequently selected (used in 51% of the 500 models), an expected result given that in vigilance tests this feature is one of the most sensitive indicators of alertness [32, 53]. In agreement with previous studies that used PVT data as a measure of fatigue [33–35], PERCLOS correlated with alertness impairment, indicating its suitability for passively detecting fatigue. In a future effort, this feature-selection process could be simplified by using a tree-based method, such as XGBoost [54, 55].

Baseline eye-opening level, which measures the extent of "droopy" or "hanging eyelids," was the second most frequently selected feature (used in 32% of the models). The relationship between this feature and alertness impairment has not been well studied. A previous investigation [26] in which observers rated photographs of rested and sleep-deprived individuals with respect to facial cues of fatigue found that the extent of hanging eyelids was greater (i.e. the eyelids were more droopy) in the sleep-deprived group. In agreement with these results, we found that a relatively lower value of this feature (indicating more hanging of the upper eyelids) was associated with greater alertness impairment. Another investigation [13], involving on-road driving, assessed the correlation between sleepiness and mean eye-opening level (equivalent to the baseline eye-opening level in our study) and found that, in agreement with our results, the mean eye-opening level decreased with sleepiness. The upper eyelid is kept in the open position by the tonic activity of the levator palpebrae muscle, whose tone is affected by alertness level [56], potentially explaining the decreased baseline eye-opening level under low-alertness conditions.

Eyelid-closing velocity was the third most frequently selected feature (used in 29% of the models). This feature was evaluated as a fatigue indicator only by two previous studies [9, 13], which found that it decreased with sleep deprivation time [9] and sleepiness level

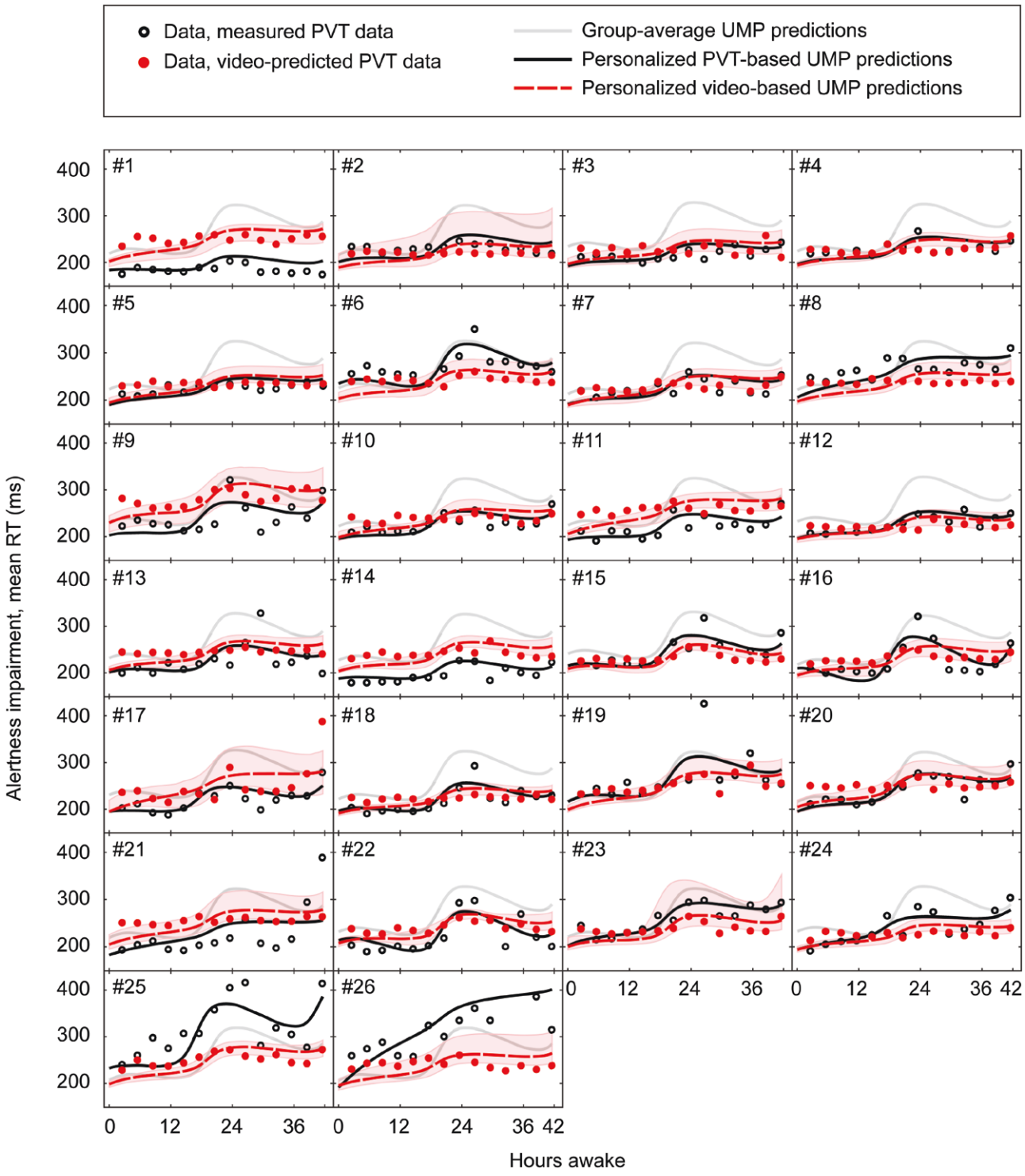


Figure 5. Personalization of the unified model of performance (UMP). Each panel shows the data (circles) and UMP predictions (lines) associated with each of the 26 participants. The open circles correspond to the measured psychomotor vigilance test (PVT) mean reaction times (RTs), and the filled circles correspond to the PVT mean RT predicted from the facial video features. The light continuous traces correspond to the predictions of the group-average UMP model. The dark continuous lines correspond to the predictions of the UMP personalized by sleep history and measured PVT data (personalized PVT-based UMP). The dashed lines correspond to the predictions of the UMP personalized by sleep history and video-predicted PVT data (personalized video-based UMP). The shaded regions indicate 95% confidence intervals for the predictions of the personalized video-based UMP model. For participant #26, the two measured PVT data points that were above 440 ms are not shown.

[13] during driving tasks. Supporting and extending their findings to non-driving scenarios and PVT-based vigilance evaluations, we found that eyelid-closing velocity decreased with alertness impairment. For all remaining significant features (Table 3), the direction of their change with alertness impairment was consistent with that found by previous studies (see references in Table 2).

Using mixed-effects models based on the most informative features, we predicted PVT mean RT and used these predictions

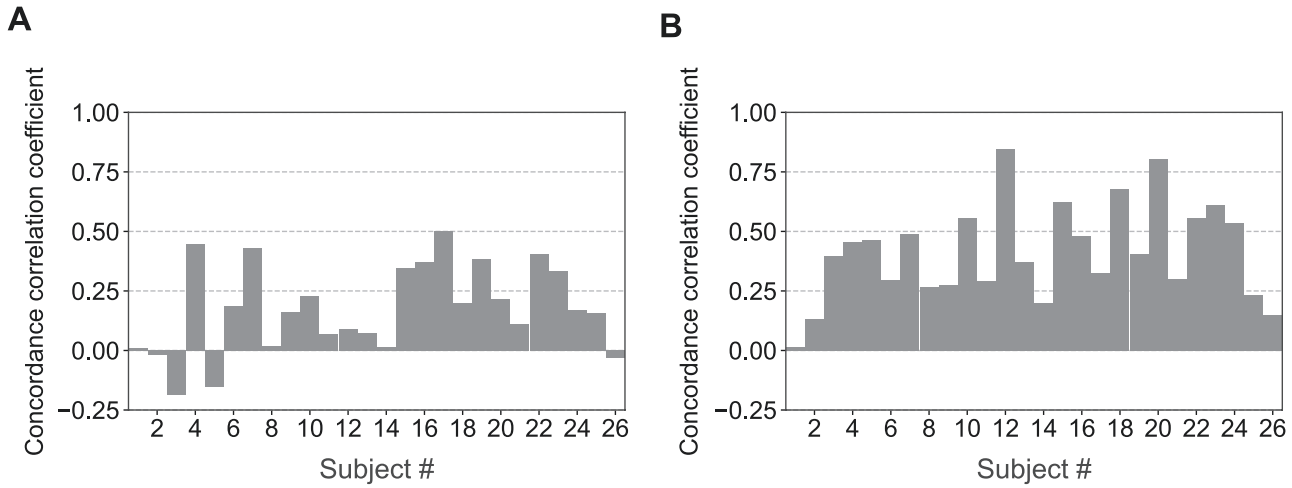


Figure 6. Distribution of concordance correlation coefficients for (A) the measured psychomotor vigilance test (PVT) versus the video-predicted PVT and (B) the measured PVT versus the personalized video-based UMP (unified model of performance) predictions. The participant population is indicated in sequential numbers from #1 to #26.

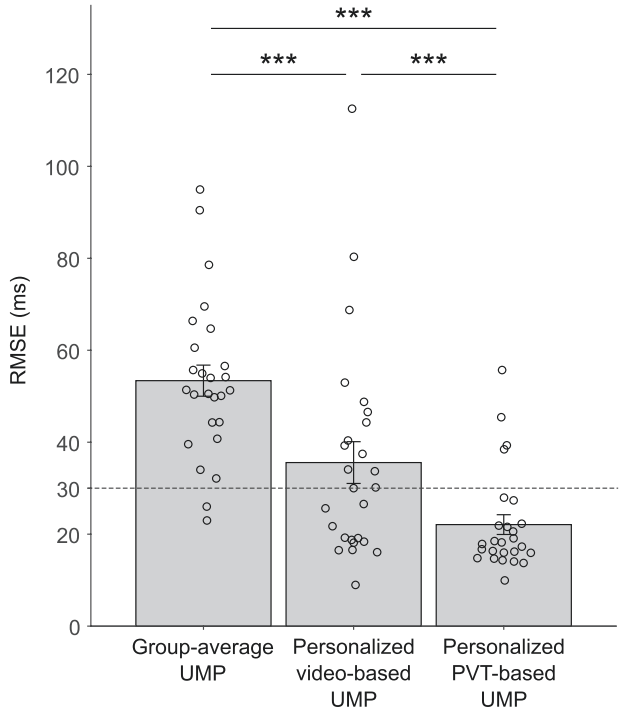


Figure 7. Comparison of the performance of the unified model of performance (UMP) based on three different inputs to the model: (A) sleep history only (group-average UMP), (B) sleep history and measured PVT data (personalized PVT-based UMP), and (C) sleep history and video-predicted PVT data (personalized video-based UMP). Each open circle represents the root mean square error (RMSE) between the measured PVT data and the predictions of the corresponding UMP model for each of the 26 participants. Error bars denote one standard error of the mean. The horizontal dotted line marks a conservative estimate of the within-participant variability (30 ms) in the measured PVT data. $^{***}p < .001$, repeated-measures ANOVA with Tukey-Kramer post hoc tests.

to personalize the UMP for each participant (Figure 2B). The average error of 36 ms for these personalized UMP predictions was substantially lower than that of the group-average model (53 ms), indicating the benefit of personalized prediction of alertness. In a previous study, a support vector machine regression model also predicted alertness-impairment levels based on video-extracted ocular features [14]. However, this model yielded a substantially higher discrepancy between measured and predicted PVT mean RT (106 ms), highlighting the advantage of our modeling approach. Although the average error of the personalized video-based predictions across the participants was relatively low, for half of the participants the error was >30 ms, partially due to the large range in feature values and PVT mean RT data. For the other half of the participants, the error was <30 ms, indicating good model performance for these participants. Nevertheless, our models could be further improved. As a potential next step to enhance PVT data prediction from video recordings, we could take advantage of recent developments in deep neural network architectures (e.g. Transformers) [57], which often have the potential to predict alertness impairment based on entire video recordings without the need to define and extract features.

To assess the agreement between the measured and predicted PVT data, we also computed the CCC values for both the video-predicted PVT and the personalized video-based UMP predictions versus the measured PVT mean RT (Figure 6, A and B, respectively). For nearly all of the participants, the CCC was considerably higher when we used the UMP. We attribute this improvement to two factors: (1) the structure of the UMP itself, which accounts for the homeostatic and circadian processes and thereby allows the predictions to more closely follow the physiology of sleep regulation, and (2) that the UMP predictions are based on a series of PVT estimates from the videos, as opposed to a single estimate. These results support the notion that the absolute error of an estimated quantity (in our case, the PVT based on a single video) should not be considered in isolation, but rather within the context of its use (in our case, as an input to the UMP to predict future PVT values) [58]. We also assessed whether the agreement between the measured and predicted PVT data was reflected in the personalized UMP model parameters. As expected, there was a significant positive correlation between the RMSE of the measured and video-predicted PVTs and the deviations in key UMP model parameters (S_0, L_0, and U) personalized with measured versus video-predicted PVTs (Pearson correlation coefficient $r = 0.24–0.94$, $p < .001$), with lower RMSE values associated with smaller absolute deviations between the corresponding UMP model parameters.

Although we used 1-minute video recordings for alertness predictions, shorter recordings, ideally a few seconds in duration, would be more practical for real-world implementation. However, for participants with a low blink rate, videos with a few seconds in duration may be devoid of the sufficient number of blinks required to compute informative features (PERCLOS and eyelid-closing velocity). As an alternative, within a time window of ≤1 hour, one could record a collection of videos of short durations during which individuals face the camera and use the collection of recordings to compute the required features.

Limitations

This study has limitations. First, we developed our models using data from a homogeneous population of healthy relatively young adults. Therefore, it is unclear whether our models are applicable to a heterogeneous, older population. In this regard, a recent study [39] found that sleep deprivation (29 hours of TSD) affected ocular metrics (PERCLOS, blink duration, and the frequency of long eye closures [>500 ms]) only in young adults (mean age, 24 years) but not in older adults (mean age, 57 years), warranting further research to develop age-appropriate alertness-prediction models. Second, we developed our models using data collected during a TSD challenge. Hence, additional studies involving different sleep-loss challenges, including chronic sleep restriction, are necessary to assess the generalizability of our findings. Finally, we obtained videos from participants who sat consistently facing the camera. In the real world, individuals may not be able to always face the camera as they may be engaged in various activities. Hence, it is unclear if our models will apply in those settings. However, when continuous video monitoring is possible, one could build predictive models based on features extracted from those time periods when individuals are facing the camera.

Conclusions

In this proof-of-concept study, we created a more practical method for predicting an individual's alertness level using passively collected video recordings of their face. The method involved two steps: prediction of PVT mean RT based on mobile phone video recordings and personalized predictions of alertness using the predicted PVT data. Assessment of model performance indicated that the personalized video-based predictions were adequately accurate with errors lower than the within-participant variability in alertness impairment under rested conditions for half of the participants, significantly better than the group-average model predictions, but not as accurate as the personalized model predictions based on actual PVT data. Future efforts should focus on investigating the use of artificial intelligence-based Transformer models [57] to capture facial features in video recordings predictive of alertness impairment.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Funding

This work was sponsored by the Military Operational Medicine Program Area Directorate of the U.S. Army Medical Research and Development Command (USAMRDC), Fort Detrick, MD. The Henry M. Jackson Foundation was supported by the USAMRDC under Contract No. W81XWH20C0031.

Disclosure Statements

Financial disclosure: none. *Nonfinancial disclosure:* none.

Author Contributions

Manivannan Subramaniyan: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, and Writing – review & editing. Francisco G. Vital-Lopez: Conceptualization, Methodology, and Writing – reviewing & editing. Tracy J. Doty: Conceptualization and Writing – review & editing. Ian Anlap: Investigation, Data curation, and Writing – reviewing & editing. William D. S. Killgore: Conceptualization, Investigation, Data curation, and Writing – reviewing & editing. Jaques Reifman: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, and Project administration.

Data Availability

The raw video recordings used in this article cannot be shared publicly because the protocol and informed consent documents do not support data sharing outside of entities defined in the original documentation. All other data will be made available following a written request to the corresponding author, along with a summary of the planned research.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Defense Health Agency, the U.S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The investigators have adhered to the policies for the protection of human participants as prescribed in AR 70–25. This paper has been approved for public release with unlimited distribution.

References

1. Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behav Res Methods Instrum Comput.* 1985;**17**(6):652–655. doi:10.3758/bf03200977
2. Basner M, Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep.* 2011;**34**(5):581–591. doi:10.1093/sleep/34.5.581
3. Song YM, Choi SJ, Park SH, Lee SJ, Joo EY, Kim JK. A real-time, personalized sleep intervention using mathematical modeling and wearable devices. *Sleep.* 2023;**46**(9):zsad179. doi:10.1093/sleep/zsad179
4. Knock SA, Magee M, Stone JE, *et al.* Prediction of shiftworker alertness, sleep, and circadian phase using a model of arousal dynamics constrained by shift schedules and light exposure. *Sleep.* 2021;**44**(11):zsab146. doi:10.1093/sleep/zsab146
5. Brewer GA, Lau KKH, Wingert KM, Ball BH, Blais C. Examining depletion theories under conditions of within-task transfer. *J Exp Psychol Gen.* 2017;**146**(7):988–1008. doi:10.1037/xge0000290
6. Massar SAA, Lim J, Sasmita K, Chee MWL. Rewards boost sustained attention through higher effort: a value-based decision

making approach. *Biol Psychol.* 2016;**120**:21–27. doi:10.1016/j.biopsycho.2016.07.019
7. Robison MK, Unsworth N, Brewer GA. Examining the effects of goal-setting, feedback, and incentives on sustained attention. *J Exp Psychol Hum Percept Perform.* 2021;**47**(6):869–891. doi:10.1037/xhp0000926
8. Abe T, Mishima K, Kitamura S, *et al*. Tracking intermediate performance of vigilant attention using multiple eye metrics. *Sleep.* 2020;**43**(3):zsz219. doi:10.1093/sleep/zsz219
9. Akerstedt T, Ingre M, Kecklund G, *et al*. Reaction of sleepiness indicators to partial sleep deprivation, time of day and time on task in a driving simulator - the DROWSI project. *J Sleep Res.* 2010;**19**(2):298–309. doi:10.1111/j.1365-2869.2009.00796.x
10. Puspasari MA, Iridiastadi H, Sutalaksana IZ, Sjafruddin A. Ocular indicators as fatigue detection instruments for Indonesian drivers. *Indust Eng Manage Syst.* 2019;**18**:748–760. doi:10.7232/iems.2019.18.4.748
11. Wilkinson VE, Jackson ML, Westlake J, *et al*. The accuracy of eyelid movement parameters for drowsiness detection. *J Clin Sleep Med.* 2013;**9**(12):1315–1324. doi:10.5664/jcsm.3278
12. Shiferaw BA, Downey LA, Westlake J, *et al*. Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Sci Rep.* 2018;**8**(1):2220. doi:10.1038/s41598-018-20588-7
13. Friedrichs F, Yang B. Camera-based drowsiness reference for driver state classification under real driving conditions. IEEE Intelligent Vehicles Symposium, La Jolla, CA, USA, 2010.
14. Massoz Q, Langohr T, François C, Verly JG. The ULg multimodality drowsiness database (called DROZY) and examples of use. IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016.
15. Daley MS, Gever D, Posada-Quintero HF, Kong Y, Chon K, Bolkhovsky JB. Machine learning models for the classification of sleep deprivation induced performance impairment during a psychomotor vigilance task using indices of eye and face tracking. *Front Artif Intell.* 2020;**3**:17. doi:10.3389/frai.2020.00017
16. Kong Y, Posada-Quintero HF, Daley MS, Chon KH, Bolkhovsky J. Facial features and head movements obtained with a webcam correlate with performance deterioration during prolonged wakefulness. *Atten Percept Psychophys.* 2021;**83**(1):525–540. doi:10.3758/s13414-020-02199-5
17. Rupp TL, Wesensten NJ, Balkin TJ. Trait-like vulnerability to total and partial sleep loss. *Sleep.* 2012;**35**(8):1163–1172. doi:10.5665/sleep.2010
18. Liu J, Ramakrishnan S, Laxminarayan S, Balkin TJ, Reifman J. Real-time individualization of the unified model of performance. *J Sleep Res.* 2017;**26**(6):820–831. doi:10.1111/jsr.12535
19. Rajdev P, Thorsley D, Rajaraman S, *et al*. A unified mathematical model to quantify performance impairment for both chronic sleep restriction and total sleep deprivation. *J Theor Biol.* 2013;**331**:66–77. doi:10.1016/j.jtbi.2013.04.013
20. Ramakrishnan S, Wesensten NJ, Kamimori GH, Moon JE, Balkin TJ, Reifman J. A unified model of performance for predicting the effects of sleep and caffeine. *Sleep.* 2016;**39**(10):1827–1841. doi:10.5665/sleep.6164
21. Reifman J, Ramakrishnan S, Liu J, *et al*. *2B-Alert* App: a mobile application for real-time individualized prediction of alertness. *J Sleep Res.* 2019;**28**(2):e12725. doi:10.1111/jsr.12725
22. Vital-Lopez FG, Doty TJ, Anlap I, Killgore WDS, Reifman J. *2B-Alert* App 2.0: personalized caffeine recommendations for optimal alertness. *Sleep.* 2023;**46**(7):zsad080. doi:10.1093/sleep/zsad080
23. Borbély AA, Achermann P. Sleep homeostasis and models of sleep regulation. *J Biol Rhythms.* 1999;**14**(6):559–570.
24. Belenky G, Wesensten NJ, Thorne DR, *et al*. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *J Sleep Res.* 2003;**12**(1):1–12. doi:10.1046/j.1365-2869.2003.00337.x
25. Priezjev NV, Vital-Lopez FG, Reifman J. Assessment of the unified model of performance: accuracy of group-average and individualised alertness predictions. *J Sleep Res.* 2023;**32**(2):e13626. doi:10.1111/jsr.13626
26. Sundelin T, Lekander M, Kecklund G, Van Someren EJ, Olsson A, Axelsson J. Cues of fatigue: effects of sleep deprivation on facial appearance. *Sleep.* 2013;**36**(9):1355–1360. doi:10.5665/sleep.2964
27. Morris TL, Miller JC. Electrooculographic and performance indices of fatigue during simulated flight. *Biol Psychol.* 1996;**42**(3):343–360. doi:10.1016/0301-0511(95)05166-x
28. Stern JA, Boyer D, Schroeder D. Blink rate: a possible measure of fatigue. *Hum Factors.* 1994;**36**(2):285–297. doi:10.1177/001872089403600209
29. Abe T, Nonomura T, Komada Y, *et al*. Detecting deteriorated vigilance using percentage of eyelid closure time during behavioral maintenance of wakefulness tests. *Int J Psychophysiol.* 2011;**82**(3):269–274. doi:10.1016/j.ijpsycho.2011.09.012
30. Barbato G, Ficca G, Beatrice M, Casiello M, Muscettola G, Rinaldi F. Effects of sleep deprivation on spontaneous eye blink rate and alpha EEG power. *Biol Psychiatry.* 1995;**38**(5):340–341. doi:10.1016/0006-3223(95)00098-2
31. Crevits L, Simons B, Wildenbeest J. Effect of sleep deprivation on saccades and eyelid blinking. *Eur Neurol.* 2003;**50**(3):176–180. doi:10.1159/000073060
32. Abe T. PERCLOS-based technologies for detecting drowsiness: current evidence and future directions. *Sleep Adv.* 2023;**4**(1):zpad006. doi:10.1093/sleepadvances/zpad006
33. Chua EC, Tan WQ, Yeo SC, *et al*. Heart rate variability can be used to estimate sleepiness-related decrements in psychomotor vigilance during total sleep deprivation. *Sleep.* 2012;**35**(3):325–334. doi:10.5665/sleep.1688
34. Ftouni S, Rahman SA, Crowley KE, Anderson C, Rajaratnam SM, Lockley SW. Temporal dynamics of ocular indicators of sleepiness across sleep restriction. *J Biol Rhythms.* 2013;**28**(6):412–424. doi:10.1177/0748730413512257
35. Jackson ML, Kennedy GA, Clarke C, *et al*. The utility of automated measures of ocular metrics for detecting driver drowsiness during extended wakefulness. *Accid Anal Prev.* 2016;**87**:127–133. doi:10.1016/j.aap.2015.11.033
36. Wierwille WW, Ellsworth LA. Evaluation of driver drowsiness by trained raters. *Accid Anal Prev.* 1994;**26**(5):571–581. doi:10.1016/0001-4575(94)90019-1
37. Wierwille WW, Wreggit SS, Kirn CL, Ellsworth LA, Fairbanks RJ. *Research on Vehicle-Based Driver Status/Performance Monitoring; Development, Validation, and Refinement of Algorithms For Detection of Driver Drowsiness. Final Report*. Washington, DC: United States Department of Transportation National Highway Traffic Safety Administration; 1994.
38. Caffier PP, Erdmann U, Ullsperger P. Experimental evaluation of eye-blink parameters as a drowsiness measure. *Eur J Appl Physiol.* 2003;**89**(3-4):319–325. doi:10.1007/s00421-003-0807-5
39. Cai AWT, Manousakis JE, Singh B, *et al*. On-road driving impairment following sleep deprivation differs according to age. *Sci Rep.* 2021;**11**(1):21561. doi:10.1038/s41598-021-99133-y
40. Anderson C, Chang AM, Sullivan JP, Ronda JM, Czeisler CA. Assessment of drowsiness based on ocular parameters

detected by infrared reflectance oculography. *J Clin Sleep Med*. 2013;**9**(9):907–920. doi:10.5664/jcsm.2992
41. Ingre M, Akerstedt T, Peters B, Anund A, Kecklund G. Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *J Sleep Res*. 2006;**15**(1):47–53. doi:10.1111/j.1365-2869.2006.00504.x
42. Johns MW, Tucker A, Chapman R, Crowley K, Michael N. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie*. 2007;**11**(4):234–242. doi:10.1007/s11818-007-0311-y
43. Soleimanloo SS, Wilkinson VE, Cori JM, *et al*. Eye-blink parameters detect on-road track-driving impairment following severe sleep deprivation. *J Clin Sleep Med*. 2019;**15**(9):1271–1284.
44. Ahlstrom C, Zemblys R, Jansson H, Forsberg C, Karlsson J, Anund A. Effects of partially automated driving on the development of driver sleepiness. *Accid Anal Prev*. 2021;**153**:106058. doi:10.1016/j.aap.2021.106058
45. Ftouni S, Sletten TL, Howard M, *et al*. Objective and subjective measures of sleepiness, and their associations with on-road driving events in shift workers. *J Sleep Res*. 2013;**22**(1):58–69. doi:10.1111/j.1365-2869.2012.01038.x
46. Wilkinson VE, Jackson ML, Westlake J, *et al*. Assessing the validity of eyelid parameters to detect impairment due to benzodiazepines. *Hum Psychopharmacol*. 2020;**35**(2):e2723. doi:10.1002/hup.2723
47. van den Berg J. Sleepiness and head movements. *Ind Health*. 2006;**44**(4):564–576. doi:10.2486/indhealth.44.564
48. King DE. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*. 2009;**10**:1755–1758.
49. Soukupova T, Cech J. Eye blink detection using facial landmarks. 21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia, 2016.
50. Nakagawa S, Schielzeth H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol Evol*. 2013;**4**(2):133–142.
51. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;**45**(1):255–268.
52. Carstensen B, Plummer M, Laara E, Hills M. Epi: a package for statistical analysis in epidemiology. R package version 2.59. 2024. https://CRAN.R-project.org/package=Epi, Accessed on May 13, 2025.
53. Dinges DF, Mallis MM, Maislin G, Powell JW. *Evaluation of Techniques for Ocular Measurement as an Index of Fatigue and as the Basis for Alertness Management*. Washington, DC: United States Department of Transportation National Highway Traffic Safety Administration; 1998.
54. Lim D, Jeong J, Song YM, *et al*. Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features. *NPJ Digit Med*. 2024;**7**(1):324. doi:10.1038/s41746-024-01333-z
55. Zhang J, Jun T, Frank J, Nirenberg S, Kovatch P, Huang K-L. Prediction of individual COVID-19 diagnosis using baseline demographics and lab data. *Sci Rep*. 2021;**11**(1):13913. doi:10.1038/s41598-021-93126-7
56. Schmidtke K, Büttner-Ennever JA. Nervous control of eyelid function. A review of clinical, experimental and pathological data. *Brain*. 1992;**115 Pt 1**:227–247. doi:10.1093/brain/115.1.227
57. Qin L, Wang M, Deng C, *et al*. SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Trans Circuits Syst Video Technol*. 2024;**34**(4):2223–2234. doi:10.1109/tcsvt.2023.3304724
58. Reifman J, Priezjev NV, Vital-Lopez FG. Can we rely on wearable sleep-tracker devices for fatigue management? *Sleep*. 2024;**47**(3):zsad288. doi:10.1093/sleep/zsad288