Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-020-02199-5

# Facial features and head movements obtained with a webcam correlate with performance deterioration during prolonged wakefulness

Youngsun Kong[1] · Hugo F. Posada-Quintero[1] · Matthew S. Daley[2] · Ki H. Chon[1] · Jeffrey Bolkhovsky[2]

Accepted: 1 November 2020

## Abstract

We have performed a direct comparison between facial features obtained from a webcam and vigilance-task performance during prolonged wakefulness. Prolonged wakefulness deteriorates working performance due to changes in cognition, emotion, and by delayed response. Facial features can be potentially collected everywhere using webcams located in the workplace. If this type of device can obtain relevant information to predict performance deterioration, this technology can potentially reduce serious accidents and fatality. We extracted 34 facial indices, including head movements, facial expressions, and perceived facial emotions from 20 participants undergoing the psychomotor vigilance task (PVT) over 25 hours. We studied the correlation between facial indices and the performance indices derived from PVT, and evaluated the feasibility of facial indices as detectors of diminished reaction time during the PVT. Furthermore, we tested the feasibility of classifying performance as normal or impaired using several machine learning algorithms with correlated facial indices. Twenty-one indices were found significantly correlated with PVT indices. Pitch, from the head movement indices, and four perceived facial emotions—anger, surprise, sadness, and disgust—exhibited significant correlations with indices of performance. The eye-related facial expression indices showed especially strong correlation and higher feasibility of facial indices as classifiers. Significantly correlated indices were shown to explain more variance than the other indices for most of the classifiers. The facial indices obtained from a webcam strongly correlate with working performance during 25 hours of prolonged wakefulness.

**Keywords** Prolonged wakefulness · Fatigue · Sleep deprivation · Reaction time · Artificial intelligence · Facial expressions · Facial emotions

Prolonged wakefulness induces adverse changes in cognitive performance (Alhola & Polo-Kantola, 2007; Killgore et al., 2008; Lim & Dinges, 2008). These changes are well established in the literature and include increased likelihood of errors, increased time to complete tasks or react to stimuli, and impaired attention, memory, and decision-making, among others (Åkerstedt & Philip, 2015; Dinges, 1995; Griffith & Mahadevan, 2015; Hafner, Stepanek, Taylor, Troxel, & van Stolk, 2017). Many studies have reported that prolonged wakefulness increases human-error related accidents such as traffic accidents and chemical safety accidents (Åkerstedt & Philip, 2015; Dinges, 1995; Griffith & Mahadevan, 2015; Hafner et al., 2017; Philip et al., 2014). Workers such as transportation drivers, medical providers, and military personnel may be required to work long hours or at night, leaving them susceptible to prolonged wakefulness and resulting negative impacts on their job performance. Decreased productivity and critical mistakes that result from prolonged wakefulness can be costly at both individual and societal levels (Akrout & Mahdi, 2013; Philip et al., 2014; Werner, Al-Hamadi, Limbrecht-Ecklundt, Walter, & Traue, 2018). To address these issues, researchers have been working to develop effective methods to detect and/or predict decrements in performance due to prolonged wakefulness before costly mistakes are made. These efforts include vocal features, electrodermal activity signals, and heart rate variability (McGlinchey et al., 2011; Posada-Quintero, Bolkhovsky, Qin, & Chon, 2018;

✉ Youngsun Kong
youngsun.kong@uconn.edu

[1] Department of Biomedical Engineering, University of Connecticut, 407 Engineering and Science Building, 67 N Eagleville, Storrs, CT 06269, USA

[2] Naval Submarine Medical Research Laboratory, Groton, CT, USA

Sloboda et al., 2018). However, vocal features are not easy to document in noise environment, and the other biosignal methods are limited in that they are sensitive to motion artifacts, expensive, and require electrodes to be attached to the skin/body.

One promising solution to detecting and predicting cognitive performance impairments during prolonged wakefulness is through tracking facial features and head movements. Humans express their social and emotional states through moving the facial muscles and the head, both consciously and unconsciously (Damasio, 1998; Dimberg, Thunberg, & Elmehed, 2000; Magai & McFadden, 1996; Sauter, Eisner, Ekman, & Scott, 2010). These emotional states vary from the six Ekman's basic emotions (Ekman, 1999), defined as anger, disgust, fear, happiness, sadness, and surprise. In addition to the six basic emotions, the composite emotions such as positive and negative valence (Adolph & Alpers, 2010; Cordaro et al., 2018) can also be estimated using movements of facial muscles or expressions (R. Ekman, 1997; Friesen & Ekman, 1983). Fatigue, often caused by prolonged wakefulness (National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Transportation Research Board; Committee on National Statistics; Board on Human-Systems Integration; & Panel on Research Methodologies and Statistical Approaches to Understanding Driver Fatigue Factors in Motor Carrier Safety and Driver Health, 2016), has also been suggested as an emotional state that may lead to adjustment of homeostatic balance and peripheral physiological changes (Gibson et al., 2003; LeDoux, 1998; Noakes, 2012). Noakes (2012) suggested that fatigue can be considered as an emotion regulating exercise behavior to protect the homeostasis of the body.

Therefore, facial expressions, facial emotions, and head movements can be considered in assessing prolonged wakefulness. Sundelin et al. (2013) reported that prolonged wakefulness affects facial features and may in certain cases make a face look sad. They found that prolonged wakefulness affects eyelids, exacerbates wrinkle formation and lines around eyes, and makes the corners of the mouth droop. Based on their discovery, Peng, Luo, Glenn, Zhan, and Liu (2017) used machine-learning techniques with facial features (including eyelids, wrinkle around eyes, and droopy corner mouth) to measure the degree of fatigue using self-taken photos or selfies from social media. Also, facial information has often been exploited to detect fatigue while driving (B. Lee & Chung, 2012; Y. Zhang & Hua, 2015). B. Lee and Chung (2012) used the eye-closure feature to detect fatigue while driving along with photoplethysmography and eye-blinking features . Y. Zhang and Hua (2015) trained a support vector machine (SVM) classifier with facial features extracted around the mouth and eye region to detect drivers' fatigue. Also, several studies have used head movements to detect drivers' drowsiness using a camera (Akrout & Mahdi, 2013; Friedrichs & Yang, 2010; Mittal, Kumar, Dhamija, & Kaur, 2016). The studies aimed to analyze and detect fatigue caused by prolonged wakefulness using facial features; however, it is still unclear if prolonged wakefulness detected through facial features correlates with performance deterioration. Therefore, the facial features need to be quantitatively evaluated with working or cognitive performance during prolonged wakefulness.

To summarize, prolonged wakefulness affects both cognitive performance and facial expression, facial emotions, head movements. As mentioned earlier, many methods have exploited facial features and head movements to detect prolonged wakefulness. However, no study has examined the facial features in parallel with working or cognitive performance during prolonged wakefulness. If facial features obtained with an easily accessible and relatively low-cost device like a webcam can provide information that can be used to predict the deterioration of performance produced by prolonged wakefulness, the technology could be used to alleviate the harmful or even fatal consequences of performance impairment. Hence, we aimed to investigate and compare the changes in facial features (facial expressions, facial emotions, and head movements) and the deterioration of working and cognitive performance during prolonged wakefulness. We tested and analyzed facial features obtained using a webcam, while measuring working performance using the psychomotor vigilance task (PVT) for 25 hours. The PVT has often been used to study overall performance during prolonged wakefulness due to its reliability and limited confounding effects of aptitude and learning (Basner & Dinges, 2011; Basner, Mollicone, & Dinges, 2011; Kripke, Marler, Calle, Marler, & Calle, 2004; Lim & Dinges, 2008). The PVT enables a researcher to obtain neurobehavioral changes in vigilant attention, state stability, and impulsivity by measuring the time required to press a button in reaction to a visual stimulus (Goel, 2017). In addition, PVT indices have shown strong correlation with duration of awake time during prolonged wakefulness (Lim & Dinges, 2008). We then tested machine learning methods to test feasibility of classifying deteriorated performance based on PVT using facial features, and to examine predictive validity of a set of facial features correlated with PVT performance.

## Methods

### Participants

A total of 20 healthy participants were recruited (13 males and seven females, 19–32 years of age). Participants were paid hourly and offered extra compensation if they completed the study, to motivate them to finish the experiment. Signed consent forms were collected before the experiments. Our choice of 20 subjects allows greater than 95% confidence interval to observe a significant effect ($p < .05$; Faul, Erdfelder, Lang, & Buchner, 2007; Sundelin et al., 2013). Furthermore, this sample size is sufficient to detect a correlation of at least 0.8 between PVT and facial indices at the 0.05 level of significance (Dorrian, Rogers, & Dinges, 2005; Posada-Quintero et al., 2018; Zar, 1999).

## Stimuli and materials

### Psychomotor vigilance task

The 10-min duration PVT test was performed using PC-PVT (a MATLAB-based tool) on a desktop computer (Khitrov et al., 2014). Participants were asked to click the left button of a mouse as fast as they could when a number indicating elapsed time appeared on a black background screen. Four PVT indices were calculated: average reaction time (AvRT), the number of major lapses (MaL, RT > 1 s), the number of minor lapses (MiL, 1 s ≥ RT > 0.5 s), and the number of false starts (FS) when participants clicked the mouse button before the number appeared. Reaction time is defined as the lapse between the time the stimuli was shown on the screen, and the time the participant clicked the mouse. Many studies that conducted PVT during prolonged wakefulness have reported an increase of AvRT, MiL, and FS on the PVT (Basner & Dinges, 2011; Basner et al., 2011; Doran, Van Dongen, & Dinges, 2001; Posada-Quintero et al., 2018). MiL is significantly associated with physical fatigue (I.-S. Lee, Bardwell, Ancoli-Israel, & Dimsdale, 2010). Many studies showed deterioration of MaL during prolonged wakefulness (Anderson, Wales, & Home, 2010; Posada-Quintero et al., 2018; C. Zhang et al., 2012).

### Facial indices

Facial video recordings were obtained using a Logitech C920 HD webcam, placed in front of the participants, on top of the screen. Facial indices were estimated using iMotions with Affectiva, as shown in Table 1 (McDuff et al., 2016). iMotions with Affectiva is a commercially available software based on Affdex software development kit (SDK) that extracts perceived facial emotion, expressions, and head movements. Affectiva claims accuracy of key emotion detection in the high 90th percentile, validated using 6 million facial videos from more than 87 countries (Affectiva, 2017). A total of 34 facial features were collected within four categories: head movement, facial expression, perceived facial emotion, and composite indices. The head movement indices consist of three head movement orientations (Yaw, Pitch, and Roll) determined by estimating the head position in a 3D space in Euler angles, as shown in Fig 1, and interocular distances between the two outer eye corners, essentially indicating a combination of the yaw and the movement between the face and the screen. Twenty facial expressions and seven perceived emotion indices were collected, all ranging between 0 and 100. Perceived facial emotion was based on the emotional facial action coding system (EMFACS; Friesen & Ekman, 1983; McDuff et al., 2016). EMFACS determines the likelihood of perceived emotions (not real emotions) based on facial expression changes (called "action units") without bias by the investigators or techniques (Wolf, 2015). Moreover, valence, engagement, and attention were calculated using iMotions with Affectiva software based on the perceived emotion indices and head orientations (Yaw, Pitch, and Roll), as shown in Table 2. Engagement is a measure of facial muscle activation ranging between 0 and 100, and valence indicates the intrinsic positivity or negativity in emotions ranging between −100 and +100 (Frijda, 1986).

## Design and procedure

All participants were asked to maintain consistent sleep prior to the day of the experiment, in addition to recording sleep diaries for a week before the experiment day. They were asked to avoid any stimulating or caffeine-containing drink or food starting 48 hours before their experiment day. They were also asked to bring their food on the experiment day, and the food was checked by experimenters to ensure compliance. Participants completed a medical screening questionnaire to ensure there would be no medical issues and to prevent unexpected accidents and minimize confounding factors, such as

**Table 1** Facial indices obtained using iMotions with Affectiva. Index names are from Affectiva

| Category | Indices |
|---|---|
| Head Movement | Pitch, Yaw, Roll, Interocular Distance |
| Facial Expression | Brow Furrow, Inner Brow Raise, Brow Raise, Lid Tighten, Smirk,<br>Eye Closure, Eye Widen, Smile, Nose Wrinkle, Cheek Raise,<br>Upper Lip Raise, Lip Pucker, Lip Press, Lip Stretch, Lip Suck, Mouth Open, Chin Raise, Lip Corner Depressor, Dimpler (Dimpling), Jaw Drop |
| Perceived Facial Emotion | Joy, Anger, Surprise, Fear, Sadness, Disgust, Contempt |
| Composite | Valence, Engagement, Attention |

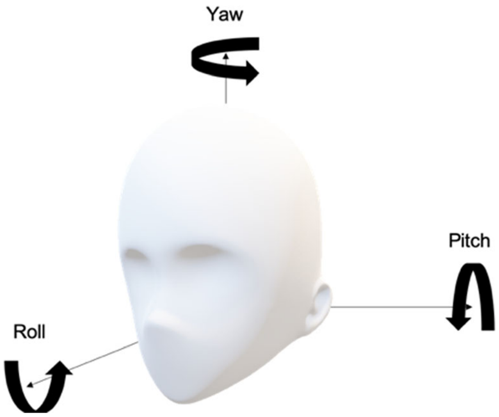

**Fig. 1** Head orientations

certain medications that can affect prolonged wakefulness. Participants received a 30-minute training at some point within two days before the start of the experiment.

Participants were asked to wake up at 6 AM and arrive at the building within 2 hours of waking up. The experiments were conducted in a 3 × 3-meter lab in the Engineering and Science building on the Storrs campus of the University of Connecticut. The room temperature was adjusted to the preference of each participant. The participants stayed with experimenters inside the building during the experiment. A total of 13 sessions per participant was performed every 2 hours for 25 hours. In every session, PVT test was performed, after a 4-minute baseline recording without any test (Khitrov et al., 2014; Loh, Lamond, Dorrian, Roach, & Dawson, 2004). Facial indices were obtained in real time during each session. Experimenters monitored the participants to ensure they were awake throughout the study. This research complied with tenets of the Declaration of Helsinki and was approved by the Institutional Review Board at the University of Connecticut.

## Statistics

From the experiments, the facial indices and working performance during the PVT were obtained for the 13 runs during the 25 hours of prolonged wakefulness, for the 20 subjects. Each participant's facial index was divided by the Euclidean norm of each session vector (1–13) of each participant, in order to accommodate the differences among subjects (Horn & Johnson, 1990). The Kolmogorov–Smirnov test was used to check normality of each PVT index and each feature index of each participant. The significant differences were calculated using the one-way analysis of variance (ANOVA) for normally distributed variables, while for nonnormally distributed variables we used Dunn's test due to the existence of missing data. For these analyses, the Bonferroni method was used for the purpose of multiple comparison correction.

For correlation analysis, the intersubject average value for each session was obtained for each facial index and measure of performance. The correlation coefficients were then calculated

**Table 2** Relation between indices (Friesen & Ekman, 1983; Affectiva, 2017)

| **Perceived emotion indices** | **Facial expression indices** |
|---|---|
| Joy | Smile |
| | Brow Raise, Brow Furrow |
| Anger | Brow Furrow, Lid Tighten, Eye Widen, Chin Raise, Mouth Open, Lip Suck |
| | Inner Brow Raise, Brow Raise, Smile |
| Surprise | Nose Wrinkle, Upper Lip Raise |
| | Lip Suck, Smile |
| Fear | Inner Brow Raise, Brow Raise, Eye Widen, Jaw Drop |
| | Brow Furrow |
| Sadness | Inner Brow Raise, Brow Furrow, Eye Widen, Lip Stretch |
| | Brow Raise, Lip Corner Depressor, Jaw Drop, Smile |
| Disgust | Inner Brow Raise, Brow Furrow, Lip Corner Depressor |
| | Brow Raise, Eye Widen, Lip Press, Mouth Open, Lip Suck, Smile |
| Contempt | Brow Furrow, Smirk |
| | Smile |
| **Composite indices** | **Facial expression or head movement indices** |
| Valence | Smile, Cheek Raise |
| | Inner Brow Raise, Brow Furrow, Nose Wrinkle, Upper Lip Raise, Lip Corner Depressor, Chin Raise, Lip Press, Lip Suck |
| Engagement | Brow Raise, Brow Furrow, Nose Wrinkle, Lip Corner Depressor, Chin Raise, Lip Pucker, Lip Press, Mouth Open, Lip Suck, Smile |
| Attention | Pitch, Yaw, Roll |

between each PVT index (AvRT, MaL, MiL, and FS) and each facial index. All facial indices and three PVT indices, AvRT, MiL, and FS were normally distributed, while MaL was nonnormally distributed. Thus, we calculated the Pearson and Spearman correlation coefficients for three PVT indices (AvRT, MiL, and FS) and MaL, respectively.

Each PVT value larger than mean + standard deviation was set as a deterioration (Class 1), which represents approximately 84.1% of the data (given the normality of the data), while the rest were set as normal (Class 0), as follows:

$$\begin{cases} \mu_f + \sigma_f > PVT_f & class1 \\ \quad otherwise & class0 \end{cases}, f \ni \{AvRT, MaL, MiL, FS\}. \tag{1}$$

The receiver operating characteristic (ROC) curves were obtained for each facial index and the PVT indices of each participant in order to evaluate the feasibility of facial indices as detectors of performance deterioration during prolonged wakefulness (Fan, Upadhye, & Worster, 2006). ROC curves calculate true positive and false positive rates of a classification model at all classification thresholds that determine two classes (Class 1 and Class 0). The thresholds can be either probability of prediction or feature values when a single feature is evaluated. For evaluating each feature, we used all possible values of each normalized facial index as thresholds. Table 3 shows an example of ROC curve using a feature threshold of 0.2. To evaluate the ROC curves, area under the curves (AUC, ranging between 0 and 1) were computed, which indicate performance across all possible classification thresholds. The higher AUC indicates more sensitivity of each facial feature to detecting performance deterioration caused by prolonged wakefulness. Figure 2 shows an example of ROC curve and AUC for eye closure.

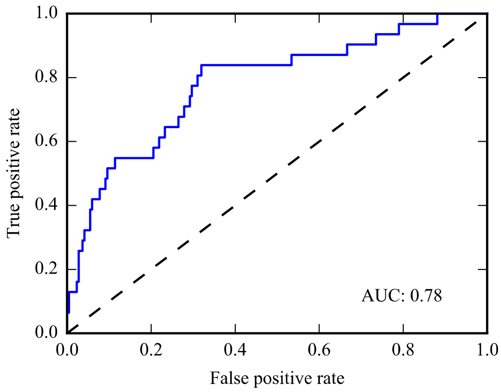

**Fig. 2** An example of ROC curve (eye closure)

Finally, we performed machine learning to detect performance deterioration with facial features. We tested two data sets that consisted of (1) all facial features and (2) facial features that highly correlated with PVT indices only. We used six different machine-learning models: support vector machine (SVM) with the linear, with the radial basis function (RBF) kernel, and with a third-order polynomial kernel, logistic regression, random forest, and k-nearest neighbors (KNN). Each classification method has its strengths and weaknesses. SVM is one of the most popular machine-learning methods that classifies or regresses linear data sets by maximizing boundaries between classes, and nonlinear problems can be solved using different types of kernels (e.g., RBF, polynomial kernels; Cortes & Vapnik, 1995). SVM has been used in many applications to date in part due to its robustness to high dimensional data, but it may not perform well for large data sets or noisy data sets. Logistic regression is a generalized linear classifier that estimates the probability of a class (such as deterioration) using a sigmoid function (McCullagh, 2019). Logistic regression is simpler and requires fewer parameter to tune, but it cannot solve nonlinear problems. Random forest

**Table 3** An example of ROC curve calculation when a threshold (facial index) is 0.2

| Subject | Session | Time | Normalized facial index (eye closure) | Predicted class (threshold: 0.2) | Target AvRT subthreshold: 507.77 ms | |
|---|---|---|---|---|---|---|
| | | | | | *PVT AvRT (ms)* | *Class* |
| Sub 1 12272018 | 1 (0 h) | 08 AM | 0.0683 | 0 | 287.50 | 0 |
| | 2 (2 h) | 10 AM | 0.2255 | 1 | 307.64 | 0 |
| | 3 (4 h) | 12 PM | 0.1559 | 0 | 315.45 | 0 |
| | 4 (6 h) | 02 PM | 0.2225 | 1 | 330.43 | 0 |
| | 5 (8 h) | 04 PM | 0.3873 | 1 | 323.17 | 0 |
| | 6 (10 h) | 06 PM | 0.3218 | 1 | 325.11 | 0 |
| | 7 (12 h) | 08 PM | 0.1706 | 0 | 319.93 | 0 |
| | 8 (14 h) | 10 PM | 0.2589 | 1 | 367.77 | 0 |
| | 9 (16 h) | 12 AM | 0.0763 | 0 | 315.96 | 0 |
| | 10 (18 h) | 02 AM | 0.0708 | 0 | 408.07 | 0 |
| | 11 (20 h) | 04 AM | 0.3090 | 1 | 614.36 | 1 |
| | 12 (22 h) | 06 AM | 0.5946 | 1 | 360.74 | 0 |
| | 13 (24 h) | 08 AM | 0.2471 | 1 | 693.67 | 1 |
| Sub 2 09082018 | 1 (0 h) | 08 AM | 0.4677 | 1 | 230.28 | 0 |
| … | | | | | | |

classifier is an ensemble learning method that is based on voting among multiple decision trees generated with different criteria (e.g., number of samples, features; Ho, 1995). It is suitable for high dimensional data and nonlinear data; it also provides low bias and moderate variance with the voting strategy. However, several parameters have to be tuned accordingly to avoid overfitting problems. In KNN classification, each sample's class is determined by voting among the training data set's closest K number (Altman, 1992). It is robust to noisy data sets, but sensitive to irrelevant features (i.e., features have to be properly selected).

Data were standardized with zero mean and unit variance. Class weights are applied for the methods due to the imbalance of the data set (210 samples for normal class and 40 samples for deterioration class). SVM parameters were set as 1 and 0.5 of C and gamma, respectively, for all kernels. Logistic regression was performed with Broyden–Fletcher–Goldfarb–Shannon's optimizer (Fletcher, 2013). Random forest was run with 10 estimators. Finally, KNN was performed with K = 5. All models were trained both with all indices, and with a set of indices that was highly correlated with the PVT outcomes. We evaluated using the leave-one-subject-out (LOSO) cross-validation approach (Koul, Becchio, & Cavallo, 2018). The LOSO cross-validation approach leaves all samples from one subject to be a test data set and uses the samples from all other subjects to be a training data set, and the procedure is repeated until all subjects are tested. This helps prevent overfitting and avoid subject bias and maximize the number of training sets (Dietterich, 1995; Ng, 1997). We then calculated the geometric mean scores (the squared root of the product of the sensitivity and specificity) of each method due to the imbalance of the data set. Geometric mean score measures the balance between classification performance of major and minor classes effectively, by maximizing both classes' accuracy equally by using both sensitivity and specificity (Akosa, 2017). Also, we used the SHapley Additive exPlanations (SHAP) to evaluate the importance of each feature in terms of the degree of its contribution (Lundberg et al., 2019; Lundberg & Lee, 2017), which is a good tool to evaluate feature importance of machine learning models using a game theory and related statistic methods.

## Results

Overall, all PVT indices showed similar trends, with significantly higher values for the last two to four sessions when compared to the rest of the sessions. Figures 3, 4, 5, 6, 7 and 8 display the mean ± standard error of the mean (*SEM*) of each PVT index and each facial index highly correlated with PVT indices. Significant differences between sessions are shown with the numbers in the figures, obtained using multicomparison tests. In Fig. 3, all PVT indices showed increasing trends through all sessions, with

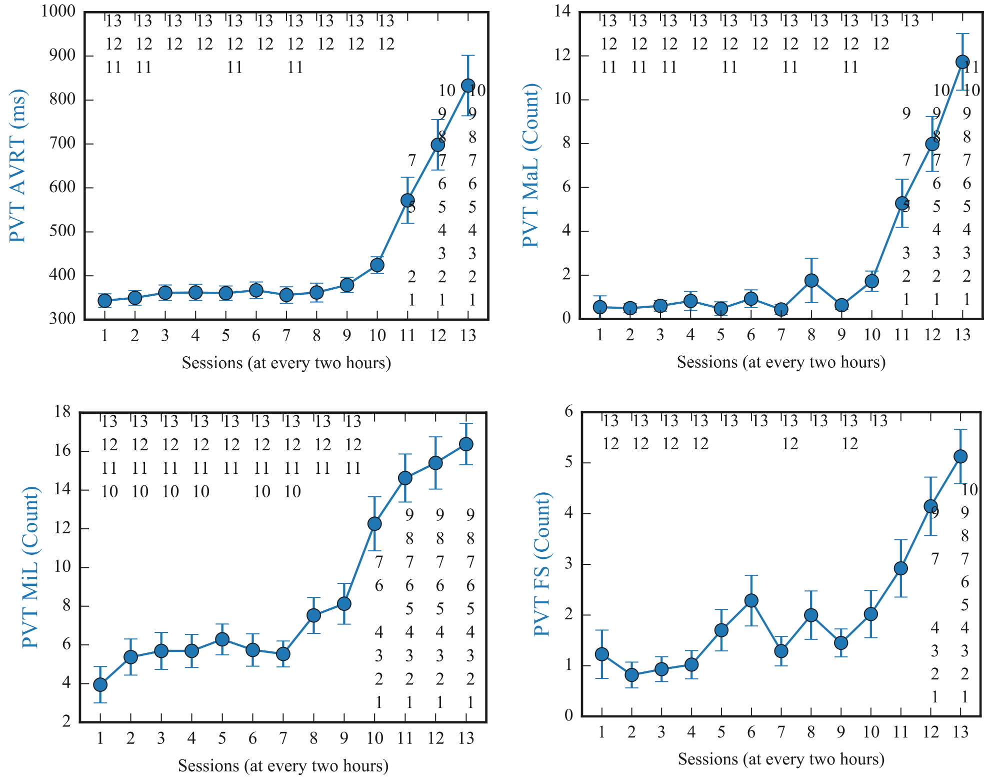

**Fig. 3** Indices of performance on PVT. Mean ± *SEM*. The column numbers indicate their significant differences to each corresponding vertical *x*-axis session ($p < .05$)

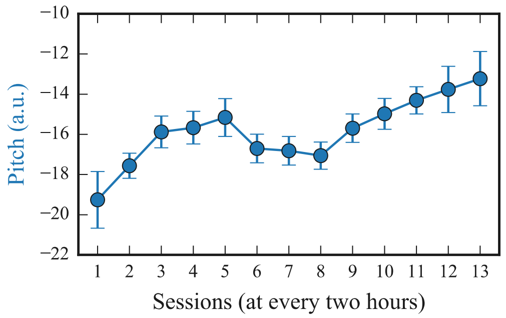

**Fig. 4** Pitch of the head movement indices. Mean ± *SEM*

the highest values in the last session. AvRT and MaL exhibited that the last three sessions were significantly higher than the rest of the sessions. Likewise, they both showed noticeable drops in the last three sessions. On the other hand, MiL and FS showed that the last four and last two sessions were significantly higher than the other sessions, respectively. Also, MiL monotonically increased until the ninth session, and FS showed an increasing trend during the first nine sessions with noticeable performance drops.

Tables 4 and 5 show correlation coefficients and the AUCs derived from the ROC curves, respectively, between each PVT index and each facial index. Twenty-one facial indices from all four categories (AvRT, MaL, MiL, and FS) showed high correlation coefficients with PVT indices. Pitch from head movement indices showed a significant correlation with PVT indices. Among facial expressions, Brow Furrow, Brow Raise, Inner Brow Raise, Eye Closure, Lid Tighten, Lip Corner Depress, Upper Lip Raise, Mouth Open, Lip Pucker, Dimpler, Jaw Drop, Nose Wrinkle, and Chin Raise were highly correlated with PVT indices. Finally, anger, surprise, sadness, disgust of perceived facial emotion indices and all composite indices were highly correlated with PVT indices.

Among head movement indices, Pitch showed significant correlation with the PVT indices AvRT, MaL, and FS (0.77, 0.64, and 0.73, $p < .05$) and MiL (0.84, $p < .001$). Pitch showed an increasing trend, except for slight decreases between the fifth and eighth sessions, as shown in Fig. 4. As shown in Table 5, it also had the highest AUC with AvRT, MaL, MiL, and FS (0.62, 0.62, 0.61, and 0.62) among the head movement indices. Yaw and roll negatively correlated with all PVT indices.

Figure 5 shows the facial expression indices for the upper part of the face, which strongly correlated with the PVT

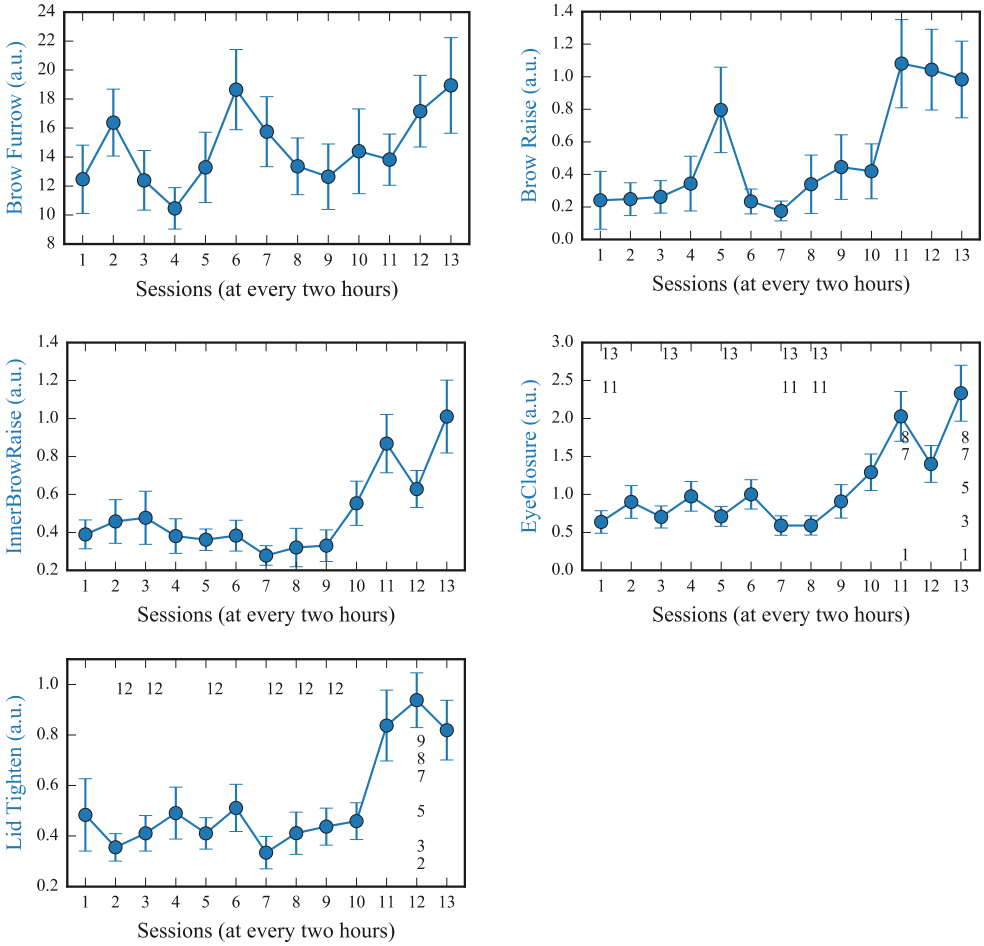

**Fig. 5** The upper facial expression. The column numbers indicate their significant differences to each corresponding vertical *x*-axis session ($p < .05$)

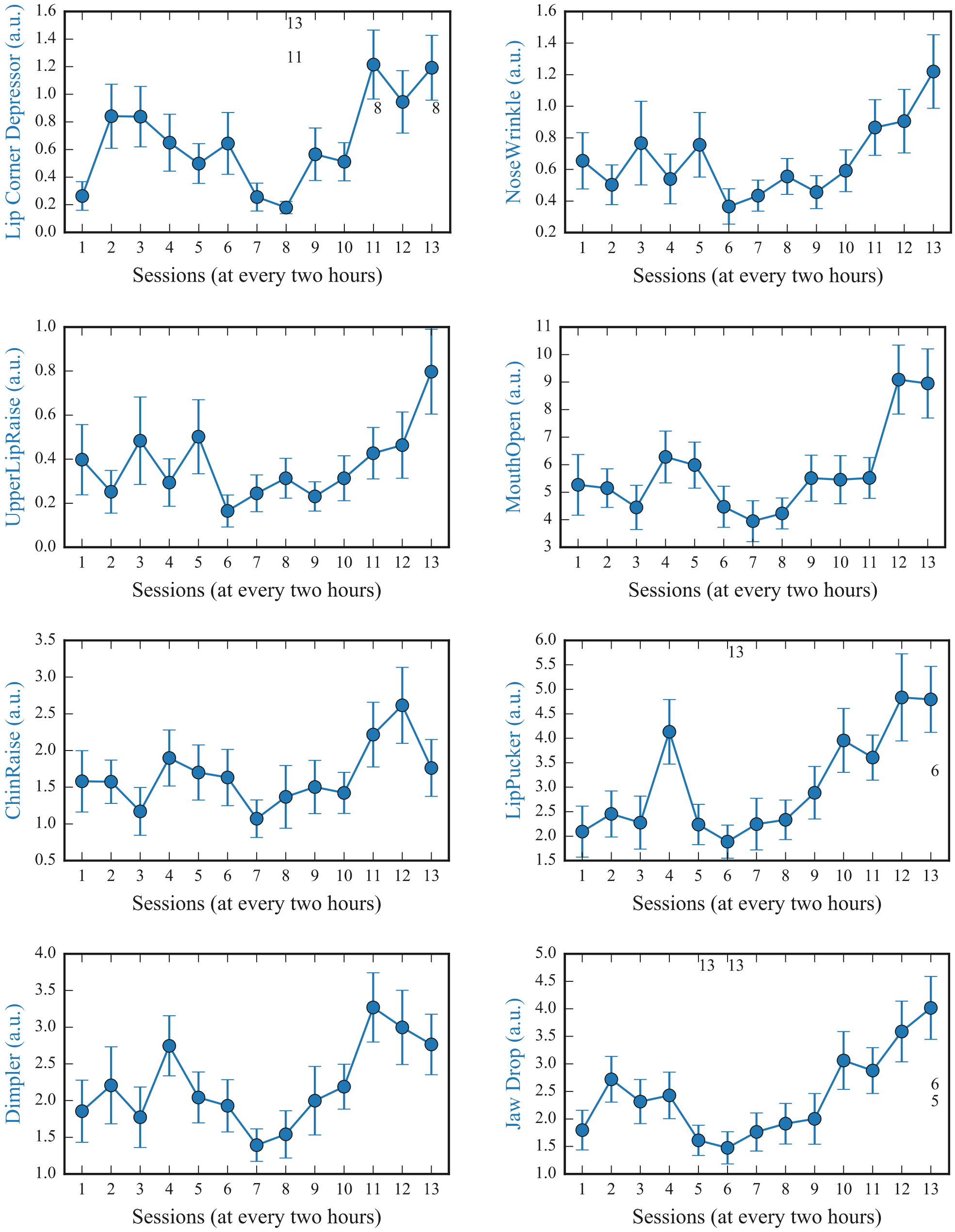

**Fig. 6** The lower facial expression indices strongly correlated with the PVT indices. Mean ± *SEM*. The column numbers indicate their significant differences to each corresponding vertical *x*-axis session ($p < .05$)

indices. Although showing strong correlations with AvRT and FS (0.56 and 0.65, $p < .05$), brow furrow exhibited the lowest AUC values with all PVT indices among the indices, as the last two values are similar to the second and sixth sessions in Fig. 5. Brow raise showed noticeably higher values in the last three sessions except for the particularly high value in the fifth session, but not significantly different (possibly due to the high *SEM*). Inner brow raise, eye closure, and lid tighten showed higher values in the last three sessions than those in the other sessions. Inner brow raise and eye closure showed a noticeable drop in the 12th session, while lid tighten showed the slightly higher value in the 12th session than in the two adjacent sessions. The 11th and 13th sessions of eye closure were significantly different from first, third, fifth, seventh, and eighth sessions, and the 12th session of lid tighten was significantly different from second, third, fifth, and seventh–ninth sessions. Lid tighten was revealed to have the highest correlation coefficients with AvRT, MaL, and FS (0.89, 0.83, and 0.87, $p < .001$) among facial expression indices (even MiL showed a significant correlation of 0.86, $p < .001$) and the highest AUC with AvRT (0.79).

Figure 6 shows that the lower facial expression indices (near the mouth) strongly correlated with the PVT indices. These indices exhibited a similar pattern—the changes of the

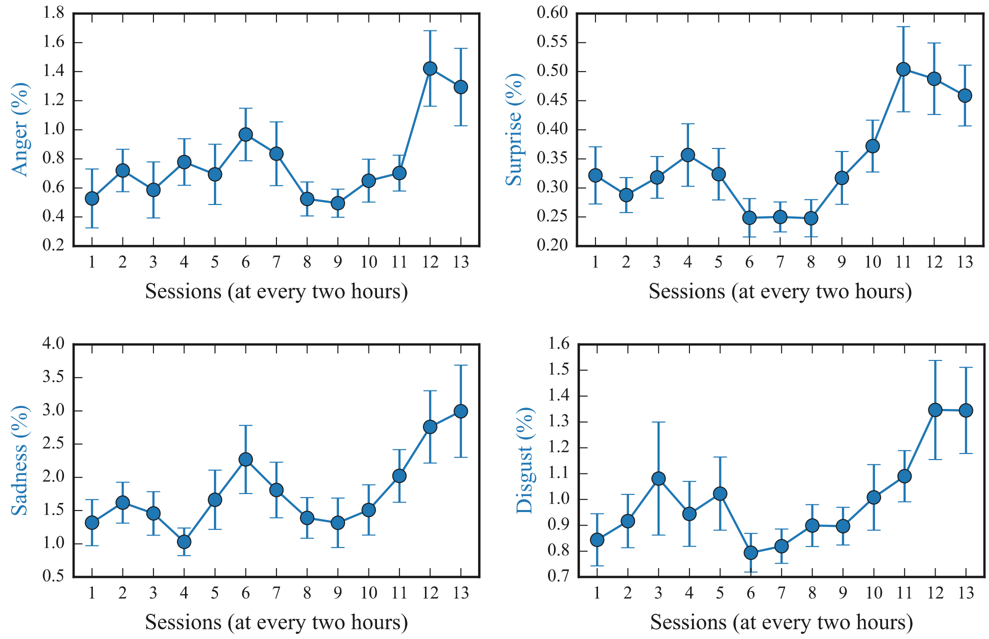

**Fig. 7** Perceived emotions. Mean ± *SEM*

first six or seven sessions are irregularly lower than those of the last one to three sessions, followed by stable movements during the next two to four sessions showing noticeable increases in the rest of the sessions. The values in the last three sessions of Lip Corner Depress were higher than those of the other sessions with a dip in the 12th session, and the values in the 11th and 13th sessions of that were significantly different from those in the eighth session. Nose Wrinkle and Upper Lip Raise exhibited similar changes throughout all sessions, with noticeably higher value in the last session. Mouth Open showed higher values in the last two sessions, possibly due to yawning. Although shown high *SEM*, chin raise showed higher values in the 11th and 12th sessions with recovering in the 13th session, possibly due to nodding off. Both Lip Pucker and Dimpler (Dimpling) showed a particularly high value in the fourth session. Lip pucker showed the higher values in the last two sessions, and the value in the last session was significantly higher than in the sixth session. Dimpler (Dimpling) showed the higher values in the last three sessions. Jaw Drop showed noticeable higher values in the last two sessions, and the value in the last session was significantly higher than in the fifth and sixth sessions.

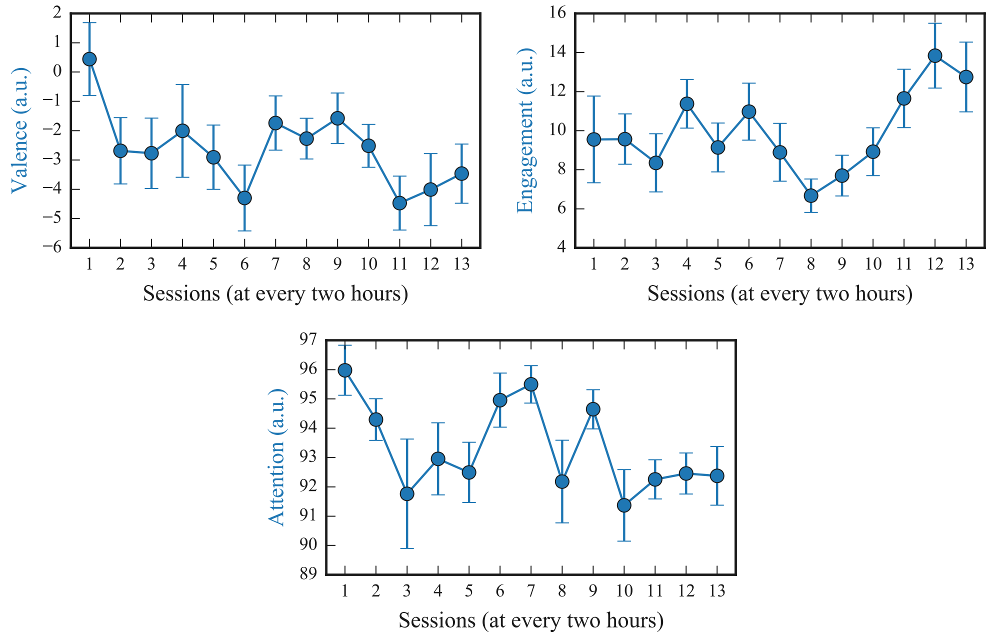

**Fig. 8** Composite indices. Mean ± *SEM*

**Table 4** Correlation coefficients between PVT and facial indices

| | Average RT | | Major lapses | | Minor lapses | | False starts | |
|---|---|---|---|---|---|---|---|---|
| | *R* | 95% CI | *R* | 95% CI | *R* | 95% CI | *R* | 95% CI |
| | **Head movement** | | | | | | | |
| Pitch | 0.77* | 0.38, 0.93 | 0.64* | 0.14, 0.88 | 0.84** | 0.54, 0.95 | 0.73* | 0.30, 0.91 |
| Yaw | −0.21 | −0.68, 0.39 | −0.18 | −0.67, 0.41 | −0.23 | −0.69, 0.37 | −0.04 | −0.58, 0.52 |
| Roll | −0.44 | −0.80, 0.15 | −0.38 | −0.77, 0.22 | −0.46 | −0.81, 0.12 | −0.36 | −0.76, 0.24 |
| Interocular Distance | 0.16 | −0.43, 0.65 | 0.01 | −0.54, 0.56 | 0.00 | −0.55, 0.55 | −0.02 | −0.56, 0.54 |
| | **Facial expression** | | | | | | | |
| Brow Furrow | 0.56* | 0.01, 0.85 | 0.39 | −0.21, 0.77 | 0.43 | −0.16, 0.79 | 0.65* | 0.15, 0.88 |
| Brow Raise | 0.82** | 0.49, 0.94 | 0.63* | 0.12, 0.88 | 0.85** | 0.56, 0.95 | 0.80* | 0.45, 0.94 |
| Inner Brow Raise | 0.89** | 0.67, 0.97 | 0.62* | 0.10, 0.87 | 0.86** | 0.59, 0.96 | 0.80* | 0.45, 0.94 |
| Eye Closure | 0.88** | 0.64, 0.96 | 0.75* | 0.34, 0.92 | 0.89** | 0.67, 0.97 | 0.82** | 0.49, 0.94 |
| Lid Tighten | 0.89** | 0.67, 0.97 | 0.83** | 0.51, 0.95 | 0.86** | 0.59, 0.96 | 0.87** | 0.61, 0.96 |
| Eye Widen | 0.20 | −0.39, 0.68 | 0.24 | −0.36, 0.70 | 0.14 | −0.45, 0.64 | 0.16 | −0.43, 0.65 |
| Cheek Raise | −0.07 | −0.60, 0.50 | −0.19 | −0.67, 0.40 | −0.26 | −0.71, 0.34 | −0.15 | −0.65, 0.44 |
| Lip Corner Depress | 0.72* | 0.28, 0.91 | 0.52 | −0.04, 0.83 | 0.67* | 0.19, 0.89 | 0.59* | 0.06, 0.86 |
| Nose Wrinkle | 0.85** | 0.56, 0.95 | 0.52 | −0.04, 0.83 | 0.73* | 0.30, 0.91 | 0.76* | 0.36, 0.92 |
| Upper Lip Raise | 0.75* | 0.34, 0.92 | 0.30 | −0.30, 0.73 | 0.58* | 0.04, 0.86 | 0.67* | 0.19, 0.89 |
| Mouth Open | 0.86** | 0.59, 0.96 | 0.49 | −0.08, 0.82 | 0.73* | 0.30, 0.91 | 0.79* | 0.42, 0.93 |
| Chin Raise | 0.61* | 0.09, 0.87 | 0.51 | −0.06, 0.83 | 0.61* | 0.09, 0.87 | 0.59* | 0.06, 0.86 |
| Lip Pucker | 0.80** | 0.45, 0.94 | 0.66* | 0.17, 0.89 | 0.83** | 0.51, 0.95 | 0.70* | 0.24, 0.90 |
| Dimpler (Dimpling) | 0.70* | 0.24, 0.90 | 0.57* | 0.03, 0.85 | 0.73* | 0.30, 0.91 | 0.59* | 0.06, 0.86 |
| Smirk | 0.44 | −0.15, 0.80 | 0.37 | −0.23, 0.77 | 0.38 | −0.22, 0.77 | 0.27 | −0.33, 0.71 |
| Lip Suck | 0.12 | −0.46, 0.63 | 0.23 | −0.37, 0.69 | 0.21 | −0.39, 0.68 | 0.05 | −0.52, 0.58 |
| Lip Press | 0.51 | −0.06, 0.83 | 0.47 | −0.11, 0.81 | 0.51 | −0.06, 0.83 | 0.39 | −0.21, 0.77 |
| Lip Stretch | 0.23 | −0.37, 0.69 | 0.20 | −0.39, 0.68 | 0.34 | −0.26, 0.75 | 0.07 | −0.50, 0.60 |
| Jaw Drop | 0.86** | 0.59, 0.96 | 0.66* | 0.17, 0.89 | 0.85** | 0.56, 0.95 | 0.72* | 0.28, 0.91 |
| Smile | 0.11 | −0.47, 0.62 | −0.07 | −0.60, 0.50 | −0.08 | −0.60, 0.49 | 0.01 | −0.54, 0.56 |
| | **Perceived facial emotion** | | | | | | | |
| Joy | −0.01 | −0.56, 0.54 | −0.22 | −0.69, 0.38 | −0.21 | −0.68, 0.39 | −0.14 | −0.64, 0.45 |
| Anger | 0.79* | 0.42, 0.93 | 0.29 | −0.31, 0.73 | 0.61* | 0.09, 0.87 | 0.80* | 0.45, 0.94 |
| Surprise | 0.82** | 0.49, 0.94 | 0.50 | −0.07, 0.82 | 0.86** | 0.59, 0.96 | 0.71* | 0.26, 0.91 |
| Fear | 0.45 | −0.13, 0.80 | 0.40 | −0.19, 0.78 | 0.32 | −0.28, 0.74 | 0.53 | −0.03, 0.84 |
| Contempt | 0.45 | −0.13, 0.80 | 0.35 | −0.25, 0.76 | 0.38 | −0.22, 0.77 | 0.34 | −0.26, 0.75 |
| Sadness | 0.85** | 0.56, 0.95 | 0.38 | −0.22, 0.77 | 0.71* | 0.26, 0.91 | 0.90** | 0.69, 0.97 |
| Disgust | 0.89** | 0.67, 0.97 | 0.54 | −0.02, 0.84 | 0.83** | 0.51, 0.95 | 0.79* | 0.42, 0.93 |
| | **Composite** | | | | | | | |
| Valence | −0.52 | −0.83, 0.04 | −0.52 | −0.83, 0.04 | −0.58* | −0.86, −0.04 | −0.57* | −0.85, −0.03 |
| Engagement | 0.75* | 0.34, 0.92 | 0.48 | −0.10, 0.82 | 0.60* | 0.07, 0.87 | 0.70* | 0.24, 0.90 |
| Attention | −0.39 | −0.77, 0.21 | −0.54 | −0.84, 0.02 | −0.56* | −0.85, −0.01 | −0.36 | −0.76, 0.24 |

*$p < .05$. **$p < .001$. Bonferroni correction was used

Figure 7 shows the four perceived emotions—Anger, Surprise, Sadness, and Disgust—highly correlated with the PVT indices except for MaL. Perceived emotion indices showed positive correlation coefficients with PVT indices, except for Joy. Especially, Surprise, Sadness, and Disgust correlated highly with all PVT indices, except for the MaL (0.82, 0.85, and 0.89, respectively, $p < .001$). Anger, Sadness, and Disgust exhibited noticeable drops in the last two sessions, while values in the last three sessions of Surprise were higher than in the other sessions.

Figure 8 shows the composite indices, valence, engagement, and attention are highly correlated with all PVT indices, except for MaL. All composite indices had no significant difference between sessions. Valence and attention showed

**Table 5** AUCs of facial indices with PVT indices

| | Average RT | | Major lapses | | Minor lapses | | False starts | |
|---|---|---|---|---|---|---|---|---|
| | AUC | 95% CI | AUC | 95% CI | AUC | 95% CI | AUC | 95% CI |
| | **Head movement** | | | | | | | |
| Pitch | 0.62 | 0.51, 0.73 | 0.62 | 0.51, 0.73 | 0.61 | 0.52, 0.70 | 0.62 | 0.52, 0.72 |
| Yaw | 0.51 | 0.40, 0.62 | 0.52 | 0.41, 0.63 | 0.42 | 0.33, 0.51 | 0.49 | 0.39, 0.59 |
| Roll | 0.46 | 0.35, 0.57 | 0.45 | 0.35, 0.55 | 0.46 | 0.37, 0.55 | 0.47 | 0.37, 0.57 |
| Interocular Distance | 0.58 | 0.47, 0.69 | 0.62 | 0.51, 0.73 | 0.59 | 0.50, 0.68 | 0.58 | 0.48, 0.68 |
| | **Facial expression** | | | | | | | |
| Brow Furrow | 0.50 | 0.39, 0.61 | 0.57 | 0.46, 0.68 | 0.49 | 0.40, 0.58 | 0.54 | 0.44, 0.64 |
| Brow Raise | 0.76 | 0.66, 0.86 | 0.79 | 0.69, 0.89 | 0.69 | 0.60, 0.78 | 0.70 | 0.60, 0.80 |
| Inner Brow Raise | 0.73 | 0.62, 0.84 | 0.73 | 0.63, 0.83 | 0.69 | 0.60, 0.78 | 0.64 | 0.54, 0.74 |
| Eye Closure | 0.78 | 0.68, 0.88 | 0.77 | 0.67, 0.87 | 0.72 | 0.63, 0.81 | 0.71 | 0.61, 0.81 |
| Lid Tighten | 0.79 | 0.69, 0.89 | 0.77 | 0.67, 0.87 | 0.71 | 0.62, 0.80 | 0.66 | 0.56, 0.76 |
| Eye Widen | 0.60 | 0.49, 0.71 | 0.60 | 0.49, 0.71 | 0.57 | 0.48, 0.66 | 0.59 | 0.49, 0.69 |
| Cheek Raise | 0.54 | 0.43, 0.65 | 0.54 | 0.43, 0.65 | 0.55 | 0.46, 0.64 | 0.53 | 0.43, 0.63 |
| Lip Corner Depress | 0.75 | 0.65, 0.85 | 0.76 | 0.66, 0.86 | 0.65 | 0.56, 0.74 | 0.64 | 0.54, 0.74 |
| Nose Wrinkle | 0.70 | 0.59, 0.81 | 0.70 | 0.60, 0.80 | 0.65 | 0.56, 0.74 | 0.66 | 0.56, 0.76 |
| Upper Lip Raise | 0.68 | 0.57, 0.79 | 0.68 | 0.57, 0.79 | 0.62 | 0.53, 0.71 | 0.64 | 0.54, 0.74 |
| Mouth Open | 0.71 | 0.60, 0.82 | 0.70 | 0.60, 0.80 | 0.64 | 0.55, 0.73 | 0.70 | 0.60, 0.80 |
| Chin Raise | 0.56 | 0.45, 0.67 | 0.56 | 0.45, 0.67 | 0.60 | 0.51, 0.69 | 0.59 | 0.49, 0.69 |
| Lip Pucker | 0.75 | 0.65, 0.85 | 0.76 | 0.66, 0.86 | 0.66 | 0.57, 0.75 | 0.71 | 0.61, 0.81 |
| Dimpler (Dimpling) | 0.63 | 0.52, 0.74 | 0.65 | 0.54, 0.76 | 0.62 | 0.53, 0.71 | 0.66 | 0.56, 0.76 |
| Smirk | 0.64 | 0.53, 0.75 | 0.65 | 0.54, 0.76 | 0.62 | 0.53, 0.71 | 0.59 | 0.49, 0.69 |
| Lip Suck | 0.49 | 0.38, 0.60 | 0.53 | 0.42, 0.64 | 0.58 | 0.49, 0.67 | 0.59 | 0.49, 0.69 |
| Lip Press | 0.65 | 0.54, 0.76 | 0.65 | 0.54, 0.76 | 0.62 | 0.53, 0.71 | 0.63 | 0.53, 0.73 |
| Lip Stretch | 0.54 | 0.43, 0.65 | 0.56 | 0.45, 0.67 | 0.62 | 0.53, 0.71 | 0.56 | 0.46, 0.66 |
| Jaw Drop | 0.74 | 0.64, 0.84 | 0.75 | 0.65, 0.85 | 0.69 | 0.60, 0.78 | 0.66 | 0.56, 0.76 |
| Smile | 0.61 | 0.50, 0.72 | 0.60 | 0.49, 0.71 | 0.61 | 0.52, 0.70 | 0.57 | 0.47, 0.67 |
| | **Perceived facial emotion** | | | | | | | |
| Joy | 0.59 | 0.48, 0.70 | 0.58 | 0.47, 0.69 | 0.58 | 0.49, 0.67 | 0.58 | 0.48, 0.68 |
| Anger | 0.55 | 0.44, 0.66 | 0.62 | 0.51, 0.73 | 0.54 | 0.45, 0.63 | 0.62 | 0.52, 0.72 |
| Surprise | 0.73 | 0.62, 0.84 | 0.72 | 0.62, 0.82 | 0.68 | 0.59, 0.77 | 0.69 | 0.59, 0.79 |
| Fear | 0.54 | 0.43, 0.65 | 0.60 | 0.49, 0.71 | 0.47 | 0.38, 0.56 | 0.59 | 0.49, 0.69 |
| Contempt | 0.61 | 0.50, 0.72 | 0.62 | 0.51, 0.73 | 0.62 | 0.53, 0.71 | 0.57 | 0.47, 0.67 |
| Sadness | 0.54 | 0.43, 0.65 | 0.62 | 0.51, 0.73 | 0.54 | 0.45, 0.63 | 0.55 | 0.45, 0.65 |
| Disgust | 0.67 | 0.56, 0.78 | 0.66 | 0.55, 0.77 | 0.66 | 0.57, 0.75 | 0.63 | 0.53, 0.73 |
| | **Composite** | | | | | | | |
| Valence | 0.45 | 0.34, 0.56 | 0.44 | 0.34, 0.54 | 0.46 | 0.37, 0.55 | 0.45 | 0.35, 0.55 |
| Engagement | 0.58 | 0.47, 0.69 | 0.64 | 0.53, 0.75 | 0.57 | 0.48, 0.66 | 0.63 | 0.53, 0.73 |
| Attention | 0.37 | 0.27, 0.47 | 0.38 | 0.28, 0.48 | 0.42 | 0.33, 0.51 | 0.35 | 0.26, 0.44 |

negative correlation coefficients with PVT indices, while engagement showed positive correlation coefficients with PVT indices. Valence showed a noticeable decrease in not only the last three sessions but also the sixth session. Engagement exhibited higher values in the last three sessions than in the other sessions, with a slight drop at the last session. Attention did not show significant correlation coefficients with AvRT and FS (−0.39 and −0.36).

Table 6 shows results of machine learning models evaluated with LOSO cross validation. Note that geometric mean scores are in the range of 0–1, and the higher values indicate more accurate predictive power for both classes (performance deterioration vs. normal). Linear classifiers (SVM with a linear kernel and logistic regression) performed better than did the other machine learning models throughout all PVT indices. We discovered that the best geometric mean scores for

**Table 6** Geometric mean score of machine learning models. Boldface fonts represent the highest score of each PVT index

| Feature sets | | SVM linear | SVM RBF | SVM poly | Random forest | KNN (K = 5) | Logistic regression | Number of features |
|---|---|---|---|---|---|---|---|---|
| All indices | AvRT | 0.6850 | 0.5428 | 0.5601 | 0.1792 | 0.3567 | **0.7302** | 34 |
| | MaL | 0.6696 | 0.6237 | 0.5137 | 0.4244 | 0.4509 | 0.6556 | 34 |
| | MiL | 0.6294 | 0.6004 | 0.3981 | 0.4354 | 0.3529 | 0.6317 | 34 |
| | FS | **0.6330** | 0.4459 | 0.3468 | 0.3845 | 0.1558 | 0.6036 | 34 |
| Correlation-based indices ($p < .05$) | AvRT | 0.7058 | 0.6157 | 0.6326 | 0.4719 | 0.4369 | 0.7241 | 19 |
| | MaL | **0.8062** | 0.7156 | 0.6011 | 0.5389 | 0.5389 | 0.7702 | 7 |
| | MiL | 0.7001 | 0.6279 | 0.4837 | 0.3792 | 0.4705 | **0.7159** | 20 |
| | FS | 0.6124 | 0.4732 | 0.4332 | 0.2215 | 0.1551 | 0.5747 | 20 |

PVT indices AvRT, MaL, MiL, and FS were logistic regression with all facial indices (0.7302), linear SVM with correlation-based facial indices (0.8062), logistic regression with correlation-based facial indices (0.7159), and linear SVM with all facial indices (0.6330). SVM with nonlinear kernels (RBF and polynomial kernels) showed lower geometric scores with all PVT indices than linear classifiers, for both features sets consisting of all indices and correlation-based indices. Random forest and KNN showed poorer performance for all PVT indices when compared with any other classifiers, achieving less than 0.5 geometric scores for AvRT, MiL, and FS with both feature sets, except for the value of 0.5389 for MaL with the correlation-based facial indices.

Moreover, most feature sets trained using the correlation-based indices showed higher performance than using all indices. SVM with a linear kernel showed higher geometric scores for PVT indices AvRT, MaL, and MiL with correlation-based indices feature set (0.7058, 0.8062, and 0.7001, respectively) than with all indices' feature set (0.6850, 0.6696, and 0.6294, respectively); the geometric scores for FS were 0.6330 and 0.6124 with all indices and correlation-based indices, respectively. Logistic regression exhibited higher geometric scores with correlation-based indices for PVT indices MaL and MiL (0.7702 and 0.7159, respectively) than with all indices (0.6556 and 0.6317). Geometric scores of logistic regression for AvRT were 0.7302 and 0.8241 with all indices and correlation-based indices, respectively, and 0.6036 and 0.5747 were obtained for FS. SVM with RBF and polynomial kernels showed higher geometric scores with correlation-based indices than with all indices. For random forest and KNN, classifiers trained with the correlation-based indices showed higher geometric scores for AvRT and MaL (also KNN for MiL).

To evaluate the best models for each index, we calculated Shapley additive explanations (SHAP) values of the models, as they show the best performance for each PVT index, as shown in Fig. 9. The features are sorted by the mean of absolute SHAP values in descending order, indicating the importance of each feature. Eye closure was the most important feature for MaL and was the top four most important feature for the others. Lip Suck, Engagement, and Lip Pucker were shown to be the most important features for AvRT, MiL, and FS, respectively. Although Lid Tighten was the five most important feature for AvRT, MaL, and MiL, it was less essential than 10 features for FS. Lip Pucker showed as an important feature in the top five for all PVT indices.

## Discussions and conclusions

In this paper, we analyzed the correlation between facial features obtained using a webcam and PVT indices of performance. We observed 34 facial indices obtained using a webcam, including head movements, facial expressions, perceived facial emotion, and composite indices. A total of 21 out of 34 indices were highly correlated with at least one PVT index, including eye-related, mouth-related expressions, Pitch from head movements, and Anger, Surprise, Sadness, and Disgust from perceived emotion features. Similar to other studies, our work also showed deterioration of PVT indices during prolonged wakefulness (Basner & Dinges, 2011; Basner et al., 2011; Posada-Quintero et al., 2018). Our PVT results exhibit stable PVT performance during the first nine to 11 sessions (0 to 17–21 hours awake), followed by significant performance deterioration in the last two to four sessions (following 25 hours awake). Our results show that facial indices are effective measurements to assess individuals' deterioration of performance and cognition during prolonged wakefulness.

To date, there have been only a few publications that compared facial features and fatigue (Knoll, Attkiss, & Persing, 2008; Sundelin et al., 2013). Sundelin et al. (2013) showed that hanging eyelids and more droopy corners of the mouth correlated with prolonged wakefulness. Similarly, our results also show that Lid Tighten and Lip Corner Depress (corresponding to the features hanging eyelids and droopy corners

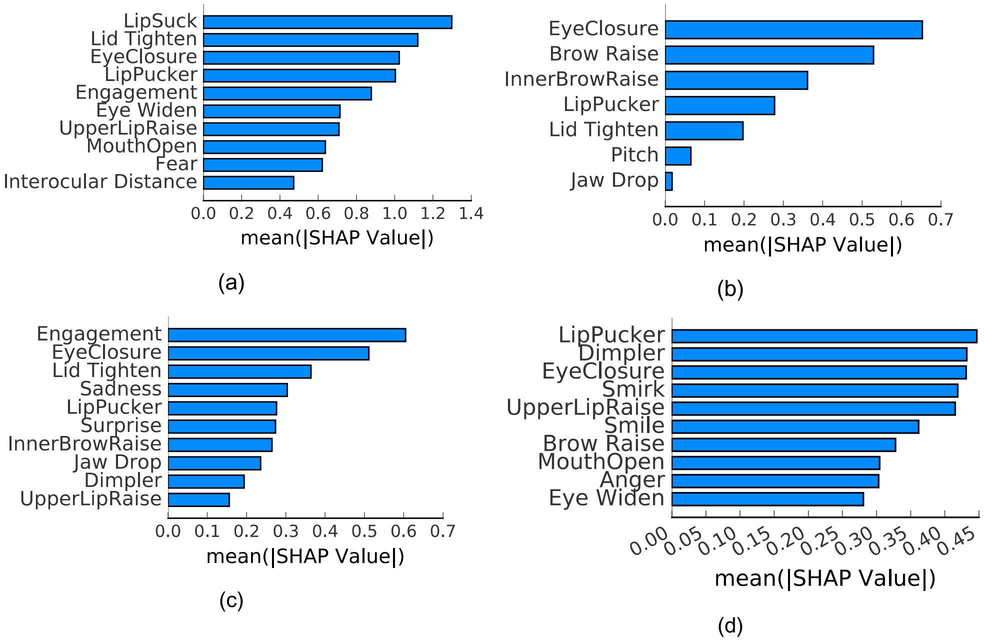

**Fig. 9** SHAP summary plots showed the best model for each PVT index, sorted in descending order by the average of the absolute SHAP values. **a** Logistic regression with all facial indices for AvRT. **b** SVM-linear kernel with correlation-based facial indices for MaL. **c** Logistic regression with correlation-based facial indices for MiL. **d** SVM-linear kernel with all facial indices for FS

of the mouth) strongly correlated with PVT indices AvRT and FS. Sundelin et al. (2013) also found that sadness was significantly associated with fatigue rating which is in agreement with our finding since there were high correlations between Sadness and PVT indices. However, their study examined only one perceived emotion, Sadness, and did not measure cognitive performance. Knoll et al. (2008) modified photographs of an upper face using digital imaging software to exam the influence of eyebrow position and shape, eyelid position, and facial rhytids with the perception of tiredness. They observed significant differences between tiredness scores and the two modifications on the face: lowering the upper eyelid and depressing the lateral brow. These two features correspond to our features of Lid Tighten and Brow Raise, and they were highly correlated with PVT indices. However, their study investigated only eye-related indices and did not measure cognitive performance at all.

In our study, we found that many eye-related features (Brow Furrow, Brow Raise, Inner Brow Raise, Eye Closure, Lid Tighten) and mouth-related features (Lip Corner Depress, Upper Lip Raise, Mouth Open, Lip Pucker, Dimpler, Jaw Drop) were significantly correlated with PVT indices. Chin Raise and Nose Wrinkle were also significantly correlated with PVT indices AvRT, MiL, and FS. Five indices from facial expressions—Eye Closure, Lid Tighten, Lip Corner Depressor, Lip Pucker, Jaw Drop—showed that a few values in the last three sessions were significantly different from a few of the first nine sessions. Especially, Lid Tighten and Eye Closure showed more than five stable sessions significantly different from the last three sessions. With regard to using head movement as a detector of prolonged wakefulness-induced performance degradation, Pitch is the only index that we tested that shows practical promise. Not surprisingly, the correlation between pitch and PVT indices is significant (0.84 for MiL, $p < .001$ and .77, .64 and .72 for AvRT, MaL, and FS, $p < .05$) since it is affected by nodding off. No significant difference was observed between sessions of Pitch.

Four perceived facial emotion indices (Anger, Surprise, Sadness, and Disgust) showed high correlation coefficients with the PVT indices. Note that this does not mean that participants genuinely felt emotions, as the emotion indices indicate the likelihood of perceived emotions based on the emotional facial action coding system (Friesen & Ekman, 1983). The explanation of these four perceived facial emotions—Anger, Surprise, Sadness, and Disgust—which were highly correlated with PVT indices, can be expanded to intense levels of annoyance, distraction, pensiveness, and boredom, respectively, according to Robert Plutchik's Wheel of Emotions (Plutchik, 2001). Interestingly, these expanded emotions are also known to be affected by sleepiness (Anderson & Horne, 2006; Bodin, Björk, Ardö, & Albin, 2015; Li et al., 2017; Weinger, 1999). The composite indices Valence and Attention showed strong negative correlation with the PVT indices. The composite index Engagement correlated with the PVT indices. No significant difference was observed between sessions of each feature from perceived emotion indices and composite indices.

Our machine learning results exhibited feasibility of classifying performance deterioration during prolonged wakefulness. We tested six machine-learning classifiers and found that linear classifiers (SVM with a linear kernel and logistic regression) outperformed others for all PVT indices, with 73.02, 80.62, 71.59, 63.30 % of geometric mean scores of AvRT, MaL, MiL, and FS, respectively. The linear classifiers with correlated features resulted in higher geometric mean scores for MaL, MiL than with all features. Although we found lower geometric mean scores with correlated features for AvRT and FS, the geometric mean scores with all features and correlated features for AvRT and FS were comparable with less than 3% of difference. We then calculated feature importance of the linear classifiers using SHAP. The rankings of the importance were different from those of the correlation coefficients in machine learning models. For example, correlation coefficients of MaL index with Eye Closure and Lid Tighten were 0.75 and 0.83, respectively; however, it was shown that for MaL the importance of Eye Closure is higher than that of Lid Tighten in Fig. 9b. Likewise, Interocular Distance, Eye Widen, Smirk, Lip Suck from facial expression indices showed no significant correlation with PVT indices, but were the top 10 important features in the classifiers. This is possibly because redundant features rather than correlation are more important for machine learning approaches.

We found that some facial indices correlate with working and cognitive performance deterioration during prolonged wakefulness, which was also found with other indices such as electrodermal activity (EDA), electrocardiogram (ECG; Posada-Quintero et al., 2018), and voice (McGlinchey et al., 2011). Facial features are practical in applications since they can be obtained using a webcam that is noninvasive and easy to collect data. Many facial indices highly correlated with working and cognitive performance on PVT. However, we need more careful approaches to select features rather than relying on only correlation in practice. For instance, using Pitch from head movement indices to detect and predict performance deterioration in the driving situation may not be a good option, as observing frequent Pitch during prolonged wakefulness can cause accidents before detection. Also, external validity of our classifiers may be limited as some real-life tasks may engage specific facial features that PVT may not invoke. For example, some tasks by medical providers (e.g., surgeon) may include frequent head movements. Moreover, some facial features cannot be observed in some cases (e.g., mouth-related features from surgeons wearing masks). This must be properly considered in the feature selection criteria as well. By comparing our indices directly to the working performance on PVT, future works can use highly correlated facial indices found in this work to detect and predict the deterioration of working and cognitive performance in practical operations (e.g., driving) so that irrevocable consequences are prevented.

**Acknowledgements** The views expressed in this article reflect the results of research conducted by the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. The study protocol was approved by the University of Connecticut Institutional Review Board and Naval Submarine Medical Research Laboratory Human Research Protections Official in compliance with all applicable Federal regulations governing the protection of human subjects. Jeffrey Bolkhovsky is an employee of the U.S. Government. This work was prepared as part of his official duties. Title 17 U.S.C. §105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

**Funding** This work is supported by funding work unit number F 1103. This work was sponsored by the Defense Health Program.

## References

Adolph, D., & Alpers, G. W. (2010). Valence and arousal: A comparison of two sets of emotional facial expressions. *The American Journal of Psychology*, *123*(2), 209–219.

Affectiva. (2017). *Emotion AI 101: All about emotion detection and Affectiva's emotion metrics* [Blog post]. Retrieved from https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics

Åkerstedt, T., & Philip, P. (2015). Sleep disorders, cognition, accidents, and performance. In S. Chokroverty & M. Billiard (Eds.), *Sleep medicine: A comprehensive guide to its development, clinical milestones, and advances in treatment* (pp. 487–494). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-2089-1_55

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. *Proceedings of the SAS Global Forum*. Retrieved from https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

Akrout, B., & Mahdi, W. (2013). Vision Based approach for driver drowsiness detection based on 3D head orientation. In J. H. Park, J. K.-Y. Ng, H.-Y. Jeong, & B. Waluyo (Eds.), *Multimedia and ubiquitous engineering* (pp. 43–50). Heidelberg, Germany: Springer.

Alhola, P., & Polo-Kantola, P. (2007). Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric Disease and Treatment*, *3*(5), 553–567.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Anderson, C., & Horne, J. A. (2006). Sleepiness enhances distraction during a monotonous task. *Sleep*, *29*(4), 573–576.

Anderson, C., Wales, A. W., & Home, J. A. (2010). PVT lapses differ according to eyes open, closed, or looking away. *Sleep*, *33*(2), 197–204.

Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, *34*(5), 581–591. https://doi.org/10.1093/sleep/34.5.581

Basner, M., Mollicone, D., & Dinges, D. F. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronautica*, *69*(11/12), 949–959. https://doi.org/10.1016/j.actaastro.2011.07.015

Bodin T., Björk, J., Ardö, J., & Albin, M. (2015). Annoyance, sleep and concentration problems due to combined traffic noise and the benefit of quiet side. *International Journal of Environmental Research and Public Health*, *12*(2), 1612–1628.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. *Brain Research Reviews*, *26*(2/3), 83–86.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)*, *27*(3), 326–327.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science*, *11*(1), 86–89. https://doi.org/10.1111/1467-9280.00221

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(s2), 4–14. https://doi.org/10.1111/j.1365-2869.1995.tb00220.x

Doran, S. M., Van Dongen, H. P. A., & Dinges, D. F. (2001). Sustained attention performance during sleep deprivation: Evidence of state instability. *Archives Italiennes de Biologie*, *139*(3), 253–267.

Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). *Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss* (Doctoral dissertation). New York, NY: Marcel Dekker.

Ekman, P. (1999). Basic emotions. *Handbook of Cognition and Emotion*, *98*(45/60), 16.

Ekman, R. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19–20.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fletcher, R. (2013). *Practical methods of optimization*. New York, NY: John Wiley & Sons.

Friedrichs, F., & Yang, B. (2010). Camera-based drowsiness reference for driver state classification under real driving conditions. *2010 IEEE Intelligent Vehicles Symposium*, 101–106. https://doi.org/10.1109/IVS.2010.5548039

Friesen, W. V., & Ekman, P. (1983). EMFACS-7: Emotional facial action coding system. *Unpublished Manuscript, University of California at San Francisco*, *2*(36), 1.

Frijda, N. H. (1986). *The emotions*. Cambridge, UK: Cambridge University Press.

Gibson, A. S. C., Baden, D. A., Lambert, M. I., Lambert, E. V., Harley, Y. X. R., Hampson, D., … Noakes, T. D. (2003). The conscious perception of the sensation of fatigue. *Sports Medicine*, *33*(3), 167–176. https://doi.org/10.2165/00007256-200333030-00001

Goel, N. (2017). Neurobehavioral effects and biomarkers of sleep loss in healthy adults. *Current Neurology and Neuroscience Reports*, *17*(11), 89. https://doi.org/10.1007/s11910-017-0799-x

Griffith, C. D., & Mahadevan, S. (2015). Human reliability under sleep deprivation: Derivation of performance shaping factor multipliers from empirical data. *Reliability Engineering & System Safety*, *144*, 23–34. https://doi.org/10.1016/j.ress.2015.05.004

Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., & Van Stolk, C. (2017). Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis. *Rand health quarterly*, *6*(4).

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282.

Horn, R. A., & Johnson, C. R. (1990). *Matrix analysis*. Cambridge, UK: Cambridge University Press.

Khitrov, M. Y., Laxminarayan, S., Thorsley, D., Ramakrishnan, S., Rajaraman, S., Wesensten, N. J., & Reifman, J. (2014). PC-PVT: A platform for psychomotor vigilance task testing, analysis, and prediction. *Behavior Research Methods*, *46*(1), 140–147. https://doi.org/10.3758/s13428-013-0339-9

Killgore, W. D. S., Kahn-Greene, E. T., Lipizzi, E. L., Newman, R. A., Kamimori, G. H., & Balkin, T. J. (2008). Sleep deprivation reduces perceived emotional intelligence and constructive thinking skills. *Sleep Medicine*, *9*(5), 517–526. https://doi.org/10.1016/j.sleep.2007.07.003

Knoll, B. I., Attkiss, K. J., & Persing, J. A. (2008). The influence of forehead, brow, and periorbital aesthetics on perceived expression in the youthful face. *Plastic and Reconstructive Surgery*, *121*(5), 1793–1802.

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, *9*, 1117.

Kripke, D. F., Marler, M. R., & Calle, E. E. (2004). Epidemiological health impact. In C. A. Kushida (Ed.). Sleep Deprivation: *Clinical Issues, Pharmacology, and Sleep Loss Effects* (pp. 223-238). CRC Press. https://doi.org/10.3109/9780203998007-10

LeDoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life.* New York, NY: Touchstone.

Lee, B., & Chung, W. (2012). Driver alertness monitoring using fusion of facial features and bio-signals. *IEEE Sensors Journal*, *12*(7), 2416–2422. https://doi.org/10.1109/JSEN.2012.2190505

Lee, I.-S., Bardwell, W. A., Ancoli-Israel, S., & Dimsdale, J. E. (2010). Number of lapses during the psychomotor vigilance task as an objective measure of fatigue. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *6*(2), 163–168. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2854704/

Li, Y., Vgontzas, A., Kritikou, I., Fernandez-Mendoza, J., Basta, M., Pejovic, S., . . . Bixler, E. O. (2017). Psychomotor vigilance test and its association with daytime sleepiness and inflammation in sleep apnea: Clinical implications. *Journal of Clinical Sleep Medicine*, *13*(9), 1049–1056.

Lim, J., & Dinges, D. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, *1129*(1), 305.

Loh, S., Lamond, N., Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 339–346. https://doi.org/10.3758/BF03195580

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2019). *Explainable AI for trees: From local explanations to global understanding*. arXiv preprint arXiv:1905.04610.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc. http://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Magai, C., & McFadden, S. H. (Eds.). (1996). *Handbook of emotion, adult development, and aging.* Academic Press.

McCullagh, P. (2019). *Generalized linear models*. New York, NY: Routledge.

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (pp. 3723-3726).

McGlinchey, E. L., Talbot, L. S., Chang, K., Kaplan, K. A., Dahl, R. E., & Harvey, A. G. (2011). The effect of sleep deprivation on vocal expression of emotion in adolescents and adults. *Sleep*, *34*(9), 1233–1241. https://doi.org/10.5665/SLEEP.1246

Mittal, A., Kumar, K., Dhamija, S., & Kaur, M. (2016, March). Head movement-based driver drowsiness detection: A review of state-of-art techniques. In *2016 IEEE International Conference on*

*Engineering and Technology (ICETECH)* (pp. 903-908). IEEE. https://doi.org/10.1109/ICETECH.2016.7569378.

National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Transportation Research Board; Committee on National Statistics; Board on Human-Systems Integration; & Panel on Research Methodologies and Statistical Approaches to Understanding Driver Fatigue Factors in Motor Carrier Safety and Driver Health (2016). Consequences of fatigue from insufficient sleep. In *Commercial motor vehicle driver fatigue, long-term health, and highway safety: Research needs* (Consensus study report). Washington, DC: National Academies Press. https://doi.org/10.17226/21921

Ng, A. Y. (1997). Preventing "overfitting" of cross-validation data. *ICML*, *97*, 245–253.

Noakes, T. D. O. (2012). Fatigue is a brain-derived emotion that regulates the exercise behavior to ensure the protection of whole body homeostasis. *Frontiers in Physiology*, *3*, 82.

Peng, X., Luo, J., Glenn, C., Zhan, J., & Liu, Y. (2017, July). Large-scale sleep condition analysis using selfies from social media. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 151-161). Springer, Cham.

Philip, P., Chaufton, C., Orriols, L., Lagarde, E., Amoros, E., Laumon, B., . . . Sagaspe, P. (2014). Complaints of poor sleep and risk of traffic accidents: A population-based case-control study. *PLOS ONE*, *9*(12), e114102. https://doi.org/10.1371/journal.pone.0114102

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, *89*(4), 344–350.

Posada-Quintero, H. F., Bolkhovsky, J. B., Qin, M., & Chon, K. H. (2018). Human performance deterioration due to prolonged wakefulness can be accurately detected using time-varying spectral analysis of electrodermal activity. *Human Factors*, *60*(7), 1035–1047. Retrieved from https://journals.sagepub.com/doi/full/10.1177/0018720818781196

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2408–2412.

Sloboda, J., Lammert, A., Williamson, J., Smalt, C., Mehta, D. D., Curry, C. I., . . . Quatieri, T. (2018). Vocal biomarkers for cognitive performance estimation in a working memory task. *Environments*, *5*, 10.

Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E. J. W., Olsson, A., & Axelsson, J. (2013). Cues of fatigue: Effects of sleep deprivation on facial appearance. *Sleep*, *36*(9), 1355–1360. https://doi.org/10.5665/sleep.2964

Weinger, M. B. (1999). Vigilance, boredom, and sleepiness. *Journal of Clinical Monitoring and Computing*, *15*(7/8), 549.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., & Traue, H. C. (2018). Head movements and postures as pain behavior. *PLOS ONE*, *13*(2), e0192767. https://doi.org/10.1371/journal.pone.0192767

Wolf, K. (2015). Measuring facial expression of emotion. *Dialogues in Clinical Neuroscience*, *17*(4), 457–462. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4734883/

Zar, J. H. (1999). *Biostatistical analysis*. New Delhi, India: Pearson Education India.

Zhang, C., Varvarigou, V., Parks, P. D., Gautam, S., Bueno, A. V., Malhotra, A., & Kales, S. N. (2012). Psychomotor vigilance testing of professional drivers in the occupational health clinic: A potential objective screen for daytime sleepiness. *Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine*, *54*(3), 296.

Zhang, Y., & Hua, C. (2015). Driver fatigue recognition based on facial expression analysis using local binary patterns. *Optik*, *126*(23), 4501–4505.

**Open practice statement** The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.