Machine Learning Based Assessment of Cognitive Performance Under Sleep Deprivation

Giulia Cisotto
Dept. of Mathematics, Informatics and Geosciences
University of Trieste, Italy
giulia.cisotto@units.it

Leonardo Badon, Beatrice Gomiero, and Leonardo Badia
Dept. of Information Engineering
University of Padova, Italy
{leonardo.badon.1,beatrice.gomiero@studenti.,leonardo.badia@}unipd.it

***Abstract*—This study investigates the integration of multiple biological signals to assess the impact of sleep deprivation on attention levels. Electrocardiogram (ECG), electroencephalogram (EEG), and electrooculogram (EOG) data from sleep-deprived healthy participants were analyzed with performance outcomes from the Psychomotor Vigilance Test (PVT), which measures response times. The primary objective was to develop a robust predictive model for the level of drowsiness based on these signals. By leveraging machine learning models, the study demonstrated the feasibility of signal-based assessments for predicting drowsiness levels. Random Forest achieved the highest accuracy when using reaction times as the true labels. It also showed promising agreement with the subjective evaluation of the alertness levels, highlighting conditions where the individuals may risk to underestimate their drowsiness. The results underscore the potential of biological signals to improve understanding of sleep deprivation's impact on cognitive performance and potentially contribute to develop robust drowsiness detection systems for practical contexts.**

***Index Terms*—Computational neuroscience; Computer aided diagnosis; Sleep deprivation; Machine learning.**

I. Introduction and State of the Art

Sleep deprivation is a growing concern in modern society, significantly impacting cognitive functions, attention, and overall performance [1]. A real-time drowsiness detection system would have the potential to mitigate these risks, enhancing safety and productivity and reducing accidents and injuries [2]. A prolific line of research deals with multi-modal systems [3]–[5] integrating different biosignals to detect fatigue and discomfort which may be predictors of drowsiness [6]–[8]. A first category of works aims to identify metabolic changes associated with drowsiness and detect disorders related to sleep [9]–[11]. For example, [9] investigates the use of signals from wearable devices and processing through AI to replace polysomnography to diagnose sleep apnea/hypopnea. In [10], a deep learning framework for sleep stage classification using physiological signals such as EEG and ECG is presented, to detect fragmented sleep patterns due to sleep apnea. The same problem is tackled in [11] using a long short-term memory network on both polysomnography and oximetry data, to possibly avoid and/or corroborate time consuming manual scoring by a professional. A second category of related papers aims to assess distress and drowsiness conditions for machinery operators or vehicle drivers, to improve safety in transportation [6], [7], [12]. Among others, [6] focused on urban rail operators and used heart rate, electro-dermal activity, and eye movements from wearable devices and cameras. They showed that fusion from multiple sensors has the ability to significantly reduce the required observation window, allowing for faster detection times. Instead, [12] proposes an integration of electronic vehicle identification with a number of physiological parameters of the drivers, ranging from fatigue state to breathing and heart rate, to improve road safety.

Following this line of research, we investigate a data-driven approach to assessing drowsiness levels by integrating multiple biological signals—electroencephalogram (EEG) [13], electrooculogram (EOG), and electrocardiogram (ECG) [14]. Using data from the DROZY dataset [15], which includes recordings of sleep-deprived individuals undergoing the Psychomotor Vigilance Test (PVT), we trained three very common machine learning models: support vector machine (SVM), random forest (RF), and neural network (NN). As a novelty, we trained them w.r.t. two labeling approaches: quantitative PVT-derived response times and the qualitative Karolinska Sleepiness Scale (KSS) scores. All achieved a reasonable level of accuracy, with the RF model performing the best (82.72%) when applied to PVT labels, highlighting the effectiveness of this method for reliable drowsiness detection. We also explored the classification accuracy outcomes when the models are trained on the KSS scores, shedding light to a promising agreement between the machine learning evaluations and the individual subjective ones [16].

By exploring the use of advanced signal processing and machine learning techniques, this study provides a foundation for developing reliable, scalable, and real-time drowsiness detection solutions applicable to various real-world scenarios [10]. It also supports the research best practice aimed at designing machine learning-based solutions able to provide explainable results [17].

II. System model

The signals undergo preprocessing to remove noise and artifacts, followed by feature extraction to identify relevant biomarkers. The extracted features serve as input for a ma-

This work has been supported by the Italian PRIN project 2022PNRR "DIGIT4CIRCLE," project code P2022788KK, and by the project D86-RIC-NA-CISOTTO funded by the University of Trieste for newly hired early and mid-career researchers.

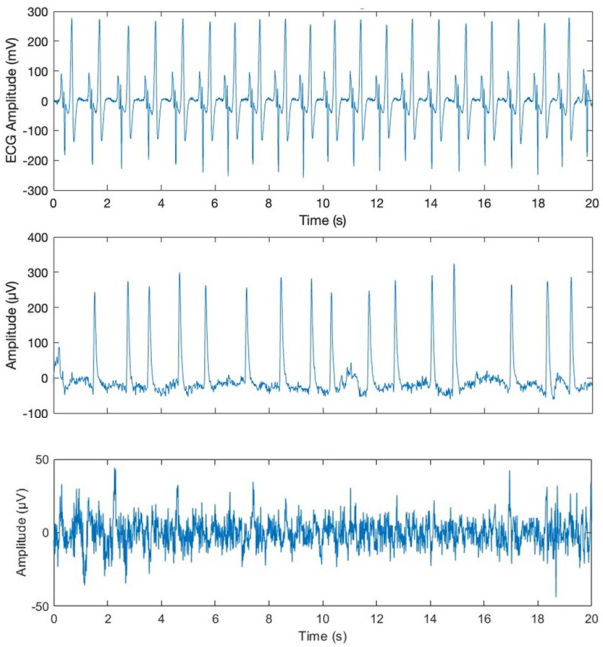

Fig. 1. A representative example of preprocessed (top) ECG signal, (middle) EOG signal, and (bottom) EEG signal.

chine learning model, which is trained to classify the level of drowsiness. We compare two different approaches in terms of chosen true labels: the first one uses the PVT outcomes (reaction times), the second one the KSS scores.

Dataset and preprocessing

We took data from the DROZY dataset, from which we extracted the polysomnography (PSG), regarded as the gold standard to study sleep and sleep stages, the PVT scores, an objective measure of vigilance and drowsiness, and the KSS scores, a self-assessment tool used to measure an individual's subjective level of sleepiness [15]. The EEG, EOG, and ECG in the PSG data were collected from 14 participants subjected to three levels of sleep deprivation. Then, each individual underwent three 10-minute sessions in which they were administered PVT: they had to press a red button in response to a yellow visual stimulus that appeared on the monitor. The time from the onset of the stimulus to the pressure of the button was recorded as *reaction time*. The visual stimuli were repeatedly shown throughout the 10 minutes at irregular intervals. At the same time, the abovementioned biological signals were measured from the individual. At the beginning of each PVT, participants were asked to estimate their level of drowsiness, according to the KSS, assigning a value from 1 to 10 (1 defined as "extremely alert" and 10 as "about to fall asleep"). All acquired signals and data are time-aligned and synchronous.

We applied a band-pass filter with a range of 0.5-60 Hz to the ECG signals to remove low-frequency noise, such as baseline wander, and high-frequency noise, such as muscle artifacts and electrical interference, thereby preserving the most informative frequency components for accurate analysis. We

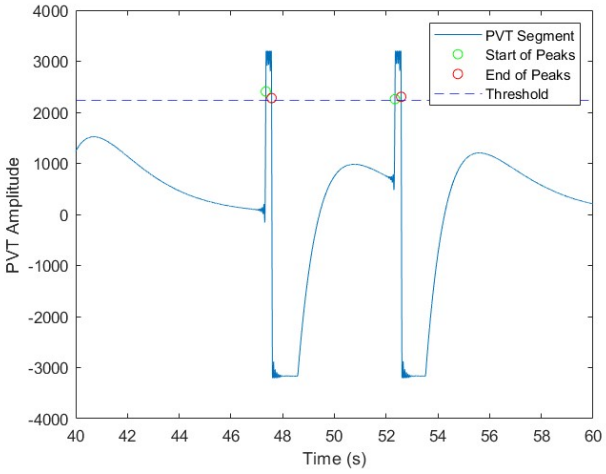

Fig. 2. A representative example of 20-second PVT signal from session 1 of participant 1. Peak detection is also shown with green and red circles.

applied a bandpass filter (0.5 to 30 Hz) to the EOG signals to remove baseline drift, electrical interference, and muscle artifacts while maintaining the integrity of the eye movement data which are fundamental for the extraction of the following features. Then, we filtered the EEG signals with a 1-45Hz Butterworth bandpass filter of fourth order to eliminate baseline drifts and high-frequency noise sources, such as EMG activity and power line interference. After preprocessing, signals appear as in Fig. 1.

Feature extraction and dataset design

Every 10 minutes session was segmented into 20 second segments, resulting in 30 segments per session. Every participant took part in 3 separate sessions. Therefore, a maximum of 1260 segments (90 segments/session $\times$ 3 sessions $\times$ 14 participants) per signal were obtained. However, due to some missing sessions in the database, the actual number of segments was only 1080. Finally, from every segment, we extracted 10 unimodal features, as described in the following.

Based on literature, we extracted from channel C3 of EEG the most relevant features associated with drowsiness [18], [19]: the power in the α, θ and β bands [3], [13]. Increases in EEG-α and EEG-θ are indicative of increases in reaction time, eye blink duration, and drowsiness, while EEG-β is indicative of alertness and concentration, with beta levels decreasing during periods of drowsiness. From EOG, we extracted the blinking duration (BD), the blinking rate (BR), the peak eyelid closing velocity (PCV), and the ratio between the amplitude and the peak closing velocity, called amplitude velocity ratio (AVR). Finally, since the ECG shows significant variations due to drowsiness, we extracted the heart rate (HR), the distances between two consecutive QRS complexes (R-R intervals, or RRI) and the signal power at both low frequencies (PLF) and high frequencies (PHF). Overall, the dataset consisted of 1080 samples and 10 features.

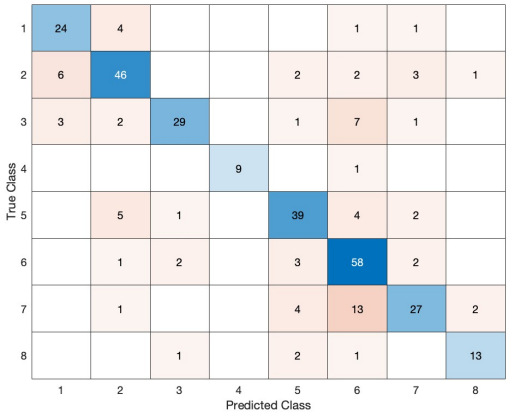

Fig. 3. Confusion matrix associated with the 8-class Random Forest classifier with KSS scores as groundtruth. Classes are numbered from 1 to 8 for convenience, but they refer to KSS scores from 2 to 9, respectively. Note that the *true class* represents the subjective evaluation of each individual of their own drowsiness level.

From the PVT, we derived a signal having the shape shown in Fig. 2 (blue line). This signal carries relevant information to assess drowsiness levels. We segmented the PVT signal into 20 second segments, consistent with the electrophysiological signals. After identifying the peaks, we extracted some relevant features, including median reaction time (med RT), standard deviation of reaction time (stdRT), minimum (minRT) and maximum reaction time (max RT), and number of lapses. Based on our analysis and the literature review, the median RT is the most suitable feature for assessing drowsiness [20].

For the groundtruth labels, we followed two different approaches, as mentioned earlier. We grouped the KSS scores into two classes, based on previous literature findings: scores of 2-6 indicate awake participants, while scores of 7-9 indicate drowsy ones. Then, we identified the maximal RT produced by the awake participants and used that value to separate the RT values into two classes: RT values corresponding to an awake reaction were 375 ms or less. Then, we classified signals w.r.t. to the above two classes of RT values. In the second approach, we simply used the KSS scores (values from 2 to 9) to build an 8-class classifier.

Machine Learning Models

We considered 3 supervised machine learning models for classification: support vector machine (SVM), random forest (RF), and neural network (NN) emerged from the literature due to their optimal trade-off between performance and data / resource consumption. They shared the same data preparation procedure: normalization (feature-by-feature) to ensure uniformity of the range, permutation to randomize the order of the samples, and finally 70%/30% training/test set split. The (hyper)parameters optimization for each classifier was obtained using cross-validation. Finally, the best models were then evaluated on the same test set to determine their relative performance.

III. Results

First, an SVM model [18] was trained using grid search to find the best values for its three main parameters, including kernel function, box constraint, and kernel scale. For this purpose, a 5-fold cross-validation was used. This model reached an accuracy of 75.13% on the test set.

Then, we considered a feedforward neural network (FFNN) model with one hidden layer. Through k-fold cross-validation, we optimized its hyperparameters, i.e., the hidden layer size and the regularization parameter (lambda). It turned out that the best values for the hidden layer size and lambda are 10 neurons and 0.001, respectively. With this model, we reached an accuracy of 80.56% in the test set.

Finally, an RF model [21] was employed. The model is characterized by two main parameters: the number of trees and the minimum leaf size. We performed a grid search optimization to find their best values, through cross-validation, finding 150 trees and a minimum leaf size of 5 as the optimal choice. This model reached an accuracy of 82.72% on the test set. As this model came out to be the best among the three candidate models (SVM, RF, NN), we also report other meaningful performance: the number of *true positive* was 44, the *true negative* were 218, while *false negative* were 47 and *false positive* were 15 (with *positive* meaning the class of drowsy individuals).

Reaction times vs subjective drowsiness perception

To assess the agreement between the machine learning models and the subjective evaluation of the drowsiness level, we performed a second classification step using the KSS scores as groundtruth labels. Every participant assigned KSS values to their drowsiness, before starting the PVT test. Although the KSS scale ranges from 1 to 10, no participant assigned the two extreme values. Thus, 8 values were chosen (from 2 to 9) and, using these new labels, we increased the granularity of the classification moving from the previous 2-class to the current 8-class problem. We decided to train another RF model, since RF was found to be the best approach in the previous classification. We applied randomization and normalization of the training data. With this second approach, we could reach an overall accuracy of 77.16% (chance level = 12.5%). The corresponding confusion matrix is reported in Fig. 3.

Although the overall accuracy decreased as compared to the values obtained in the previous classification step, here the chance level was significantly lower (the number of classes was much higher), then the performance of our RF model can be considered quite satisfactory. Moreover, it is promising that a machine learning model consistently agrees with the subjective evaluation of the drowsiness level given by a number of different individuals [16].

However, if we deepen the investigation of the confusion matrix, we can observe that the classification errors are distributed in a very similar way to the 2-class problem. Provided that KSS score 5 is defined as "Neither alert not sleepy," we can count the number of errors in the submatrix corresponding

to KSS scores 2-4 (15), the number of errors in the submatrix corresponding to KSS scores 6-9 (33), the number of errors where the participants perceived themselves as more sleepy than classified by the RF model (11), and the number of errors where the participants perceived themselves as more alert than classified by the RF model (19). This pattern reflects a similar error distribution obtained in the 2-class problem. The agreement between the machine learning model and the individuals is rather high: about 90% the predicted score was the same, or at least very close to the subjective one. On the other hand, if we believe the machine learning model captured more objective features of participants' brain activity, we can note the effect of the subjective perception of each own alertness level [16]: in fact, about 6% individuals are at risk of underestimating the degree of their drowsiness.

For this reason, and because the machine learning model's output can be obtained in real time, in a continuous modality, without distracting the individuals from their activity, the machine learning based drowsiness monitoring seems more appropriate for several applications, such as alertness monitoring during driving and attention support systems in working environments.

IV. Conclusions

Our research revealed that using simple machine learning models (support vector machines, random forests, and neural networks) to predict drowsiness levels leads to good levels of accuracy. We showed that the classification results of a random forest classifier well matched the subjective perception of the individual level of drowsiness. This encourages the development of automatic, continuous, and real-time drowsiness monitoring systems that are more objective and reactive than human awareness of their own alertness. This finding highlights the promising value of such systems in addressing sleep-related issues in various contexts, including, for example, transportation safety [6].

Despite promising results, further improvements are necessary to enhance the efficacy of drowsiness detection systems. The accuracy reached may still require improvements for a practical real-time application [22]. Our study was constrained by a limited sample size, which impacts the generality of our findings. Moreover, inter-individual differences pose a significant challenge in generalizing drowsiness detection models, suggesting the need for larger and more diverse datasets [23]. Another possible improvement can be given by adding contextual information, such as sleep patterns and environmental conditions.

References

[1] W. D. Killgore, "Effects of sleep deprivation on cognition," *Progr. Brain Res.*, vol. 185, pp. 105–129, 2010.

[2] Y. Yiğitbaşı, F. Stroppa, and L. Badia, "Optimizing real-time decision-making in sensor networks," in *Proc. IEEE Int. Conf. Devel. eSystems Eng. (DeSE)*, 2023, pp. 206–211.

[3] M. M. Hasan, C. N. Watling, and G. S. Larue, "Physiological signal-based drowsiness detection using machine learning: Singular and hybrid signal approaches," *J. Saf. Res.*, vol. 80, pp. 215–225, 2022.

[4] E. Borella, U. B. Çakmakçi, E. Gottardis, A. Buratto, T. Marchioro, and L. Badia, "Effective sensor selection for human activity recognition via Shapley value," in *Proc. IEEE Int. Wkshp Metrol. Living Env. (MetroLivEnv)*, 2024, pp. 22–27.

[5] G. Cisotto, A. V. Guglielmi, L. Badia, and A. Zanella, "Joint compression of EEG and EMG signals for wireless biometrics," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[6] H. Wu, Y. Jiao, C. Jiang, T. Wang, and J. Yu, "Fatigue state evaluation of urban railway transit drivers using psychological, biological, and physical response signals," *IEEE Access*, vol. 13, pp. 23 270–23 284, 2025.

[7] Y. Wang, "Comparative analysis of deep learning methods using multiple modal data for driver fatigue identification," in *Proc. Int. Conf. Artif. Intell. Commun. (ICAIC)*, 2024, pp. 542–552.

[8] P. S. Shedage, S. Pouriyeh, R. M. Parizi, M. Han, G. Sannino, and N. Dehbozorgi, "Stress detection using multimodal physiological signals with machine learning from wearable devices," in *Proc. IEEE Symp. Comp. Commun. (ISCC)*, 2024, pp. 1–6.

[9] D. Peng, L. Sun, Q. Zhou, and Y. Zhang, "AI-driven approaches for automatic detection of sleep apnea/hypopnea based on human physiological signals: a review," *Health Inf. Sc. Syst.*, vol. 13, no. 1, pp. 1–20, 2025.

[10] J. Ramesh, Z. Solatidehkordi, A. Sagahyroon, and F. Aloul, "Multimodal neural network analysis of single-night sleep stages for screening obstructive sleep apnea," *Appl. Sc.*, vol. 15, no. 3, p. 1035, 2025.

[11] K. Zovko, Y. Sadowski, T. Perković, P. Šolić, I. Pavlinac Dodig, R. Pecotić, and Z. Đogaš, "Advanced data framework for sleep medicine applications: Machine learning-based detection of sleep apnea events," *Appl. Sc.*, vol. 15, no. 1, p. 376, 2025.

[12] S. Zhou, N. Zhang, Q. Duan, X. Liu, J. Xiao, L. Wang, and J. Yang, "Monitoring and analyzing driver physiological states based on automotive electronic identification and multimodal biometric recognition methods," *Algorithms*, vol. 17, no. 12, p. 547, 2024.

[13] G. Cisotto and D. Chicco, "Ten quick tips for clinical electroencephalographic (EEG) data acquisition and signal processing," *PeerJ Computer Science*, vol. 10, p. e2256, 2024.

[14] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," *Fut. Gen. Comp. Syst.*, vol. 86, pp. 446–455, 2018.

[15] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *Proc. IEEE Wint. Conf. Appl. Comp. Vis. (WACV)*, 2016, pp. 1–7.

[16] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the Karolinska sleepiness scale against performance and EEG variables," *Clin. Neurophys.*, vol. 117, no. 7, pp. 1574–1581, 2006.

[17] N. Ullah, J. A. Khan, I. De Falco, and G. Sannino, "Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges," *ACM Comput. Surv.*, vol. 57, no. 4, pp. 1–36, 2024.

[18] G. Cisotto, M. Capuzzo, A. V. Guglielmi, and A. Zanella, "Feature selection for gesture recognition in internet-of-things for healthcare," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[19] D. Scapin, G. Cisotto, E. Gindullina, and L. Badia, "Shapley value as an aid to biomedical machine learning: a heart disease dataset analysis," in *Proc. IEEE Int. Symp. Cluster Cloud Internet Comput. (CCGrid)*, 2022, pp. 933–939.

[20] M. Basner and D. F. Dinges, "Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss," *Sleep*, vol. 34, no. 5, pp. 581–591, 2011.

[21] I. Wijayanto, S. Rizal, and S. Hadiyoso, "Epileptic electroencephalogram signal classification using wavelet energy and random forest," in *AIP Conference Proceedings*, vol. 2654, no. 1. AIP Publishing, 2023.

[22] A. Buratto, B. Yivli, and L. Badia, "Machine learning misclassification within status update optimization," in *Proc. IEEE Int. Conf. Commun. Netw. Satell. (COMNETSAT)*, 2023, pp. 640–645.

[23] A. Tazarv and M. Levorato, "A deep learning approach to predict blood pressure from PPG signals," in *Proc. Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2021, pp. 5658–5662.