



# Studying Drowsiness Detection Performance While Driving Through Scalable Machine Learning Models Using Electroencephalography

José Manuel Hidalgo Rogel<sup>1</sup> · Enrique Tomás Martínez Beltrán<sup>1</sup> · Mario Quiles Pérez<sup>1</sup> · Sergio López Bernal<sup>1</sup> · Gregorio Martínez Pérez<sup>1</sup> · Alberto Huertas Celdrán<sup>2</sup>

Received: 1 March 2023 / Accepted: 2 December 2023 / Published online: 11 January 2024  
© The Author(s) 2024

## Abstract

Driver drowsiness is a significant concern and one of the leading causes of traffic accidents. Advances in cognitive neuroscience and computer science have enabled the detection of drivers' drowsiness using Brain-Computer Interfaces (BCIs) and Machine Learning (ML). However, the literature lacks a comprehensive evaluation of drowsiness detection performance using a heterogeneous set of ML algorithms, being also necessary to study the performance of scalable ML models suitable for groups of subjects. To address these limitations, this work presents an intelligent framework employing BCIs and features based on electroencephalography for detecting drowsiness in driving scenarios. The SEED-VIG dataset is used to evaluate the best-performing models for individual subjects and groups. Results show that Random Forest (RF) outperformed other models used in the literature, such as Support Vector Machine (SVM), with a 78% f1-score for individual models. Regarding scalable models, RF reached a 79% f1-score, demonstrating the effectiveness of these approaches. This publication highlights the relevance of exploring a diverse set of ML algorithms and scalable approaches suitable for groups of subjects to improve drowsiness detection systems and ultimately reduce the number of accidents caused by driver fatigue. The lessons learned from this study show that not only SVM but also other models not sufficiently explored in the literature are relevant for drowsiness detection. Additionally, scalable approaches are effective in detecting drowsiness, even when new subjects are evaluated. Thus, the proposed framework presents a novel approach for detecting drowsiness in driving scenarios using BCIs and ML.

**Keywords** Brain-computer interface · Electroencephalography · Framework · Machine learning

## Introduction

Drowsiness is defined as a person's tendency to fall asleep. This situation is especially critical in driving scenarios, where the dangerous combination of driving and sleepiness commonly happens [1]. Particularly, the National Highway Traffic Safety Administration (NHTSA) reported between 2013 and 2019 a total of 5593 fatalities in motor vehicle crashes involving drowsy drivers. In 2017, exclusively in the USA, 91,000 police-reported crashes involved drowsy drivers, which led to about 50,000 people being injured [2].

In the past years, drowsiness assessment has become a topic of interest for researchers. In this sense, cognitive neuroscience, the area of knowledge responsible for studying the nervous system that supports mental functions [3], including drowsiness, has proposed different techniques for its quantification [4]. The first ones are based on monitoring subjects' behavior such as facial expressions, heart rate, and yawning in order to assess drowsiness. Although these techniques

---

✉ Sergio López Bernal  
slopez@um.es

José Manuel Hidalgo Rogel  
josemanuel.hidalgor@um.es

Enrique Tomás Martínez Beltrán  
enriquetomas@um.es

Mario Quiles Pérez  
mqp@um.es

Gregorio Martínez Pérez  
gregorio@um.es

Alberto Huertas Celdrán  
huertas@ifi.uzh.ch

<sup>1</sup> Department of Information and Communications Engineering, University of Murcia, Campus de Espinardo, Murcia 30100, Murcia, Spain

<sup>2</sup> Communication Systems Group CSG, Department of Informatics IFI, University of Zurich UZH, Zürich 8050, Zürich, Switzerland

represent an advance in safety, they have significant limitations since they produce false positives and false negatives, not always being able to measure attributes related to fatigue or drowsiness.

Next, solutions based on self-assessment with scales emerged. This approach consists in asking subjects to describe how drowsy they felt in the previous minutes. Examples of these tests are the Karolinska Sleepiness Scale (KSS) [5] and the NASA Task Load Index (NASA-TLX) [6]. However, this self-evaluation process introduces a subjectivity factor that represents the main drawback of these methods. Thus, the need to objectively quantify the sleepiness of an individual arises. For this reason, neurophysiological tests have been developed, based on monitoring patients' brain signals to precisely identify drowsiness.

Brain signals are commonly obtained by electroencephalography (EEG), which measures the electrical activity produced in the brain through electrodes acting as sensors [7]. The different levels of brain activity are related to the different cognitive states of the subject. Due to this, it is necessary to study the EEG signals in different frequency bands, being the lower frequency rhythms (delta, theta, and alpha) directly related to the states of relaxation and drowsiness, and the higher rhythms (beta and gamma) related to concentration and moderate mental load, and even stressful situations in the case of the gamma band [8, 9].

Brain-Computer Interfaces (BCIs) are normally used when studying EEG, where two categories are distinguished depending on the degree of invasiveness of the electrodes. On the one hand, invasive BCIs locate the electrodes within the skull, requiring a surgical process. On the other hand, non-invasive BCIs place their electrodes directly on the subject's scalp, avoiding a surgical procedure. Nevertheless, non-invasive BCIs data must be processed afterwards to remove artifacts caused by the subjects' activity, such as eye blinking or body movements [10, 11]. Due to their advantages and feasibility of experimenting with subjects, non-invasive BCIs are the ones commonly used in the drowsiness detection scenario. In addition to non-invasive BCIs, Machine Learning (ML) models are also used to assess drowsiness using the data collected by the BCI. For this purpose, the BCIs acquire the brain signals when the subject is driving. Then, they are processed to eliminate the noise from the signals added during the acquisition using certain techniques such as Notch and band-pass filters, sample reduction, and Independent Component Analysis (ICA).

After that, features are extracted from the signals, allowing ML algorithms to classify these characteristics according to patterns identified in the data and, therefore, to predict drowsiness. It is relevant to highlight that Deep Learning (DL) is gaining popularity in identifying drowsiness while driving. However, DL approaches present several disadvantages, such as the amount of data required to train the models,

the limited speed in training and evaluating models compared to traditional ML approaches, or the difficulty in explaining the decisions of the model [12, 13].

Despite the advances and contributions of existing studies combining BCIs and ML to detect drowsiness while driving, there is a lack of literature analyzing the performance of customized and heterogeneous ML algorithms. The current literature presents a substantial amount of studies using ML, but in most of them, Support Vector Machine (SVM) is used without analyzing and comparing other well-known and relevant algorithms. In addition, the state of the art only explores the performance of customized and individual models trained with data from single subjects, presenting significant scalability issues for new subjects since a new training process per user is needed. In this sense, scalable models combining the brain activity of several subjects should be explored and analyzed to determine if they effectively detect sleepiness in various subjects, even if the models were not trained with their data.

To improve the previous challenges, this work presents the following main contributions:

- The design of a BCI and ML-based framework for drowsiness detection in driving scenarios employing EEG and Electrooculography (EOG) as features. The proposed framework considers ML classifiers and regressors for detecting different drowsiness levels in both individual users and groups of them.
- The creation of a personalized algorithm for Percentage of Eye Closure (PERCLOS) discretization to improve drowsiness labeling, which takes into account the subject behavior to establish the thresholds between three drowsiness levels.
- The deployment and evaluation of the framework using a publicly available dataset, SEED-VIG [14], modeling the EEG of 21 subjects while driving. The following ML algorithms have been trained and evaluated with different amounts of subjects and features for regression and three-class classification tasks: SVM, k-Nearest Neighbors (kNN), Decision Trees (DT), Random Forest (RF), and Gaussian Processes (GP).
- The obtained results indicate that algorithms such as RF or kNN, which are not widely explored in the literature, can improve the performance of the most commonly used techniques, such as SVM. In particular, within individual models, RF performed the best with a mean f1-score of 78% compared to SVM with 58%. Similarly, RF is also the most promising alternative for scalable models, reaching an f1-score of 79% while SVM obtained 52%.

Despite the contributions of this work, it presents limitations in terms of the amount of data used to train ML models. Using richer datasets would be useful to generalize the results

obtained to a greater portion of the population, allowing to explore more complex intelligent approaches, such as DL approaches.

The rest of this paper is organized as follows. “[Related Work](#)” section presents the state of the art from drowsiness detection in driving scenarios using BCIs. Subsequently, “[Proposed Solution](#)” section presents the design of the proposed framework, followed by “[Experiments and Results](#)” section which states the results of detecting drowsiness using the framework. Finally, “[Conclusions](#)” section presents some conclusions and potential future work.

## Related Work

This section analyzes how drowsiness assessment techniques using BCIs are implemented in the literature and what methodology is followed by each study. In particular, it documents what biosignals and data processing data are utilized, what features are extracted from the signals, the algorithms and models used to classify the signals, and their performance. In the literature, both drowsiness and fatigue are related to the same concept of a person’s tendency to fall asleep. Every study analyzed shares the same starting point, an existing dataset. Some of them decide to generate their own data, while others opt for a public dataset [15–17]. After that, it is necessary to apply data processing techniques to improve the quality of the signals, such as removing artifacts [18, 19].

Features are then extracted from different sources. In the case of EEG, each source corresponds to a transformed domain where EEG signals can be studied. Each study analyzed chooses certain features that may differ from the rest. Firstly, time-domain features are based on mathematical models and other algebraic operations, where the most popular and widespread is the Autoregressive Model (AR) [20, 21]. It is also common to extract statistical values from the signals such as variance, standard deviation and quantiles, or Hjorth parameters [15, 19, 21].

The second source commonly used in EEG research comes from frequency domain features, where Fast Fourier Transform (FFT) enables the analysis of the predominant frequencies in the original EEG signals and their amplitude. Using FFT, the Power Spectral Density (PSD) is widely employed to measure the energy in each frequency band of the brain signals, providing good results when estimating drowsiness [15, 18, 22–24].

Thirdly, time-frequency domain features, due to the non-stationary, non-linear and non-Gaussian behavior exhibited by the EEG signals, are useful to have a representation and decomposition of the frequency information of the signals linked to the temporal domain. This is why methods such as Discrete Wavelet Transform (DWT) are used [20, 25]. In addition to EEG features, it is common to combine them

with other features which are extracted from the subject’s behavior. These include heart rate (HR), blink rate, or the number of blinks [26, 27]. The blink rate determines the frequency or speed of blinking, while the number of blinks refers to the total number of blinks performed within a particular time interval.

Finally, after feature extraction, the signals are classified. There are two common aspects in the analyzed studies while classifying. First, most works use a supervised learning approach and, second, they use a limited range of algorithms which are known to provide good results, being SVM the most popular and widespread technique [15, 17, 25, 27, 28]. This algorithm is followed in popularity by linear models, such as Ridge Regression, Logistic Regression, Lasso Regression, Naive Bayes and kNN [29]. To a lesser extent, and with more popularity in other areas of EEG analysis, Linear Discriminant Analysis (LDA), DT and RF are also chosen [16, 22].

Regarding Deep Learning (DL), the most widely used neural networks are Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Extreme Learning Machines (ELMs) and Recurrent Self-Evolving Fuzzy Neural Networks (RSEFNNs). They are gaining relevance as they produce better results, in many cases, compared to traditional ML methods in drowsiness assessment [20, 25, 26].

When estimating sleepiness with supervised learning, the labels used for regression models are the values measured by self-assessment test, such as KSS, NASA-TLX, and Auditory Vigilance Task (AVT), or PERCLOS values. If the problem is approached with a classification model, the values of the labels used in the regressive methods are discretized to different levels of drowsiness [21, 30, 31].

## Performance of Literature Works

This section presents an in-depth examination of the literature to identify how the algorithms perform while also considering the processing techniques and features researchers adopt when estimating drowsiness. Focusing first on works employing ML approaches, Chen et al. [25] acquired EEG and EOG signals from 16 subjects using a nine-electrode BCI with a sampling rate of 256 Hz. Then, neurologists removed data artifacts and labeled the signals by visual inspection. Moreover, the authors extracted features from EEG using Discrete Wavelet Transformations (DWT) and combined them with EOG features. After that, the authors used SVM for classification, reaching an accuracy of 94.7%.

Gwak et al. [16] used ML to detect drowsiness at the wheel, analyzing different physiological signals and driving behaviors in a driving simulation for 16 subjects. This work used a 16-channel BCI with a sampling rate of 500 Hz, applying a band-pass filter between 1 and 40 Hz and ICA to remove artifacts. The authors considered 32 features obtained

from PSD in EEG signals, ECG characteristics, eye movement, seat pressure, and driving simulation parameters. This study trained LR, SVM, kNN, and RF classifiers, where RF obtained 81.4% accuracy in binary classification, in contrast to SVM, which obtained 78.6% accuracy.

The work performed by Li et al. [32] is relevant to the present study since the dataset employed is also SEED-VIG. The authors applied ICA and downsampling of 125 Hz to the EEG signals, obtaining 100 features related to differential entropy, while this work obtained 36 EOG features from horizontal and vertical channels. After that, the paper employed a Support Vector Regressor (SVR) as a baseline, resulting in a model with an RMSE of 0.17 and CC of 0.76.

Wei et al. [22] used a 32-channel BCI with a sampling rate of 500 Hz to acquire EEG signals from ten participants, utilized to predict drowsiness in a virtual driving environment. The authors processed the EEG using a band-pass filter between 1 and 50 Hz, a notch filter at 60 Hz, downsampling to 250 Hz, and Artifact Subspace Reconstruction (ASR). This work employed three-second epochs to obtain PSD features from theta, alpha, and beta waves from EEG. This research used LDA, kNN, and SVM algorithms for classification, where SVM obtained the best results, with an accuracy of 80%.

Akbar and Igasaki [15] used an EEG BCI with 19 electrodes and a sampling rate of 500 Hz. The authors applied a band-pass filter between 0.5 and 50 Hz, then extracted Hjorth parameters and PSD from the frequency domain and KSS to self-assess drowsiness. The algorithm for classifying was SVM, achieving an RMSE of 0.15 and a  $R^2$  of 0.83.

Qian et al. [33] studied the detection of drowsiness during daytime short naps using EEG data obtained from 25 subjects with a sampling rate of 100 Hz. The authors selected frequencies under 30 Hz and then extracted features from EEG frequency bands using FFT. Finally, the authors studied several models, where the most promising alternative was a Bayesian-Copula Discriminant Classifier (BCDC) with 94.3% accuracy, followed by Gaussian SVM (GSVM) with 93.7% accuracy.

Arefnezhad et al. [34] proposed an encoder-decoder method for drowsiness detection in driving scenarios, using EEG signals obtained from 13 subjects using a BCI with eight channels and a sampling rate of 500 Hz. This work used ICA for data processing, then extracting EEG features consistent between subjects: skewness of alpha, delta power, theta power, and Hjorth mobility of delta. For classification, the encoder uses a series of equations to relate the EEG features obtained with PERCLOS values, resulting in relevant biomarkers in the EEG. In contrast, the decoder uses Bayes filtering and biomarkers to predict PERCLOS values.

Arif et al. [35] utilized various ML algorithms to detect drowsiness. In particular, they used a BCI device with 16 channels and a sampling rate of 125 Hz on 12 subjects.

Then, this work used a band-pass filter between 0.5 and 40 Hz and a notch filter on 50 and 60 Hz frequencies, obtaining eight features from PSD and four from the band power ratio indices. Finally, they used decision trees, discriminant analysis, logistic regression, Naïve Bayes, SVM, kNN, and an ensemble classifier (bagged trees) for classification. This work concluded that the best approach was using an ensemble classifier, obtaining 85.6% accuracy.

Besides traditional ML approaches, the literature has explored the use of DL. In particular, Chakladar et al. [21] performed a workload analysis, exploring both ML and DL approaches. This work used an EEG dataset with 14 channels, with a sampling rate of 128 Hz, obtained from 48 participants. The subjects were recorded when no task was performed and during a simultaneous capacity multitasking activity, identifying three workload levels: low, moderate, and high. This research applied a band-pass of 4–32 Hz over the EEG, then extracted different features: PSD, hurst exponent, signal statistics (mean, standard deviation, skewness, kurtosis), approximate entropy, and autoregressive coefficient. Finally, three classification algorithms were tested: SVM, RF, and a hybrid approach of a Long Short-Term Memory (LSTM) with a Bidirectional LSTM, known as BLSTM-LSTM. This latter algorithm offered the best results, with 86.33% accuracy. Moreover, Cheng et al. [24] compared the performance of SVM with a CNN using EEG signals. The former obtained an accuracy of 64.05%, while the latter achieved an accuracy of 69.19%. In both cases, PSD features were utilized.

Cui et al. [19] used an explainable CNN with data from 27 subjects to detect drowsiness. The BCI has 32 electrodes and a sampling rate of 500 Hz. After the acquisition, the data is band-filtered between 1 and 50 Hz, removing artifacts using AAR, following a downsampling process to 128 Hz. After that, three-second epochs are used as raw inputs to a CNN network. The results, calculated for each subject, present an overall inter-subject accuracy of 73.22%.

Paulo et al. [36] used EEG signals from 27 subjects obtained from a 32-channel BCI with a sampling rate of 500 Hz. The authors applied a band-pass filter between 1 and 50 Hz and blink and muscular artifacts using AAR. This work trained a CNN with one convolutional layer and three dense layers with three-second temporal windows. The drowsiness problem was approached as an image classification problem, where the images represent spatiotemporal image encoding representations in the form of recurrence plots or gramian angular fields. The overall performance between individual models was 75.87% accuracy.

Shen et al. [23] evaluated multiple DL approaches to detect drowsiness in driving scenarios. This study used data obtained from a 32-channel EEG BCI with a sampling rate of 500 Hz, obtained from 11 subjects. After performing a band-pass filter between 1 and 50 Hz, Automatic Artifact Removal (AAR), and downsampling to 120 Hz, the authors calcu-



lated the PSD over each EEG channel, labeling the different experimental sessions as drowsy or alert. This work explored different classification approaches, where the most promising was their proposed method, consisting in the use of multi-source signal alignment with a tensor network, reaching a 71.97% accuracy in leave-one-subject-out cross-validation.

Zhu et al. [18] used an eight-channel EEG-based BCI with a sampling rate of 256 Hz on ten subjects. This article applied a band-pass filter between 1 and 60 Hz, a notch filter on 50 Hz, ICA, and the wavelet threshold method. Then, the authors trained a CNN to predict drowsiness while driving, evaluating either the application of an Inception or an AlexNet module. The use of the Inception module offered an accuracy of 95.59%, while the use of the AlexNet approach reached 94.68% accuracy.

In particular, Table 1 presents a summary of the most related studies, specifying the processing techniques, set of features, data labeling and the algorithms used. If a work uses regression, the results are expressed by the Root Mean Square Error (RMSE). In addition to RMSE, the Pearson correlation coefficient (CC) or the coefficient of determination ( $R^2$ ) is used depending on the paper. On the other hand, classification models are characterized by accuracy as performance metric. After analyzing the literature it can be seen that SVM is generally present in the literature but, at the same time, there is evidence of other approaches, such as neural networks or other ML algorithms, that offer similar or even better results. In addition, there is also a lack of studies that consider scalable models since all of the identified studies focus on individual models which can only detect drowsiness in a specific subject.

## Proposed Solution

This section describes the design and implementation of the proposed framework to detect drowsiness while driving, related to the first contribution indicated in the “Introduction” section. An overview of the framework is shown in Fig. 1, presenting its different components. Starting from the upper side, the first two components refer to the acquisition of data and its processing. Next, a feature extraction stage selects the most relevant aspects of the acquired data. Finally, the framework includes a data classification block, where individual models for each subject and scalable models with data from several users are implemented based on different ML algorithms.

This framework differentiates from existing platforms, focusing on the particularities of EEG and EOG signals offering specific processing capabilities for drowsiness detection. Moreover, the framework implements a novel PERCLOS discretization approach able to adapt to the particularities of each subject. Finally, the proposed framework tests a substan-

tial variety of ML algorithms to offer a detailed comparison between them in terms of well-known performance metrics.

It is worth mentioning that the structure of the proposed framework is aligned with existing frameworks using EEG signals to predict particular dynamics of the human brain [37, 38]. Moreover, the modules of these frameworks have a direct association with the phases of the BCI life cycle and traditional ML methodologies, which represent the stages required to acquire biosignals from the brain, their transformation to be understood by computers, and, finally, the use of learning techniques to predict specific events within the signals [39].

## Data Acquisition

The design and implementation of the proposed framework is generic enough to be compatible with different datasets, as well as data coming in real-time from a BCI. Nevertheless, this work used the SEED-VIG dataset [14] due to its realistic conditions, the suitability with the study purpose, and the amount and quality of the data provided.

More in detail, the SEED-VIG dataset consists of 23 experiments over 21 different subjects (two subjects repeated the experiment). Each experiment has about two hours of EEG signals recorded while the subjects were using a driving simulator. The experiments acquired data from 17 electrode channels according to the 10-20 system (see Fig. 2), using a sample rate of 200 Hz. Particularly, the *Neuroscan* BCI device was in charge of acquiring EEG and EOG biosignals [40]. This dataset provides the raw data from the different experiments, together with a variety of already processed data. Particularly, the present study uses the following data subsets: (1) raw EEG data from the 17 EEG channels, (2) average PSD relative to the five frequency bands of the brain signals and, (3) raw data from the EOG vertical channel.

The dataset was labeled every eight seconds with subjects' PERCLOS values obtained by an eye-tracking device from *SensoMotoric Instruments* [41]. PERCLOS is a psychophysiological measure of the subject that quantifies the percentage of time that a subject has been with the eyes at least 80% closed during the time interval of measurement [42].

## Data Processing

As a result of using a non-invasive BCI, the EEG signals obtained contain artifacts, so they must be filtered following the process presented in Fig. 3. Initially, the signals are processed with two filtering techniques. First, a Notch filter applied at 60 Hz eliminates the noise introduced by the power grid. Secondly, a band-pass filter between one and 30 Hz was applied since this is the frequency range of interest for the study of drowsiness [8]. The signals are then downsampled to 60 Hz following the Nyquist-Shannon sampling theorem

**Table 1** Summary of the literature works reviewed and their results. The results are expressed in multiple metrics, where *Acc* represents Accuracy, *CC* means the Pearson correlation coefficient, *RMSE* is the Root Mean Square Error, and  $R^2$  indicates the coefficient of determination

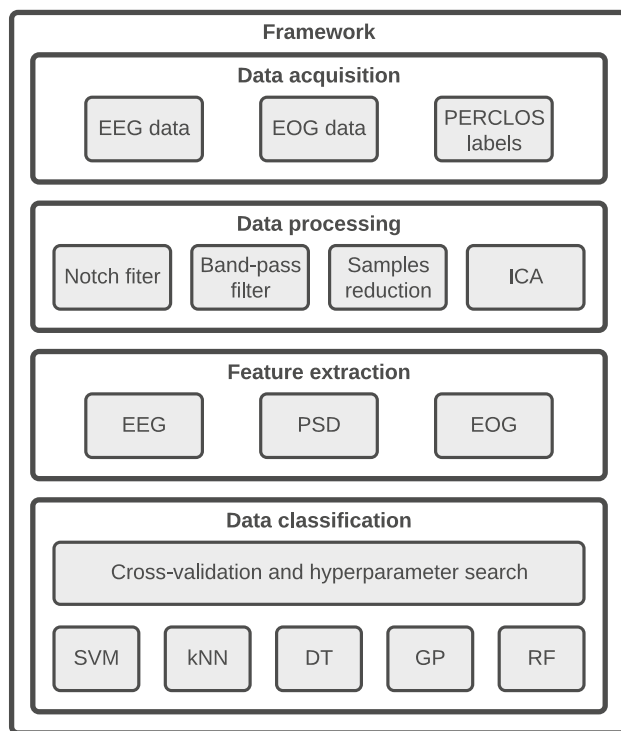
Reference	Acquisition	Processing	Features	Classification	Results
Chen et al. (2015) [25]	EEG, EOG	N/A	Four from DWT, blinking	SVM	Acc: 96.90%
				ELM	Acc: 97.30%
Cheng et al. (2018) [24]	EEG	N/A	PSD	SVM	Acc: 64.05%
				CNN	Acc: 69.19%
Gwak et al. (2018) [16]	EEG, ECG, EOG	Band-pass 1–40 Hz, ICA	PSD, ECG, EOG simulation data	CNN	Acc: 75.30%
				SVM	Acc: 78.60%
				RF	Acc: 81.40%
Li et al. (2018) [32]	EEG, EOG	ICA, downsampling 125 Hz	100 from EEG, 36 from EOG	SVR	CC: 0.76 RMSE: 0.17
Wei et al. (2018) [22]	EEG	Band-pass 1–50 Hz, notch 60 Hz, downsampling 250 Hz, ASR	PSD	kNN	Acc: 77.3%
				LDA	Acc: 79.4%
				SVM	Acc: 80.0%
Akbar and Igasaki (2019) [15]	EEG	Band-pass 0.5–50 Hz	Hjorth parameters, PSD	SVM	$R^2$ : 0.64, RMSE: 0.56
				Recurrent SVM	$R^2$ : 0.83, RMSE: 0.15
Chakladar et al. (2020) [21]	EEG	Band-pass 5–32 Hz	PSD, mean, SD, skewness, kurtosis, AR, entropy	RF	Acc: 83.00%
				SVM	Acc: 83.33%
Cui et al. (2021) [19]	EEG	Band-pass 1–50 Hz, AAR, downsampling 128 Hz	Oz channel	DT	Acc: 60.70%
				RF	Acc: 62.60%
				kNN	Acc: 63.42%
				Gaussian Naïve Bayes	Acc: 67.44%
				SVM	Acc: 69.72%
Paulo et al. (2021) [36]	EEG	Band-pass 1–50 Hz, AAR	Raw EEG	CNN	Acc: 75.87%
Shen et al. (2021) [23]	EEG	Band-pass 1–50 Hz, AAR, downsampling 120 Hz	PSD	SVM	Acc: 62.51%
Qian et al. (2021) [33]	EEG	Band-pass <30 Hz	PSD	BCDC	Acc: 94.3%
				GSVM	Acc: 93.7%
Zhu et al. (2021) [18]	EEG	Band-pass 1–60 Hz, notch 50 Hz, ICA, wavelet threshold	Raw EEG	CNN-Inception	Acc: 93.6%
				CNN-AlexNet	Acc: 94.68%
Arefnezhad et al. (2022) [34]	EEG	ICA	Skewness alpha, delta power, theta power, Hjorth delta mobility	Encoder-decoder	RMSE: 0.117
Arif et al. (2023) [35]	EEG	Band-pass 0.5–40 Hz, Notch	PSD, band power ratio indices	Discriminant analysis	Acc: 63.5%
				LR	Acc: 63.6%
				Naïve Bayes	Acc: 67.4%
				SVM	Acc: 75.7%
				DT	Acc: 77.4%
				kNN	Acc: 78.5%
				Ensemble classier	Acc: 85.6%

to reduce the size of the data and speed up its subsequent classification without losing information. Finally, ICA permits to remove the remaining artifacts, such as subjects' eye blinks, while the essential information for detecting drowsiness is preserved. Once the artifacts are removed from the initial raw data, it is also necessary to split the signals in portions (Epochs) of eight seconds. This allows to perform

a correct feature extraction since there is a PERCLOS value every eight seconds.

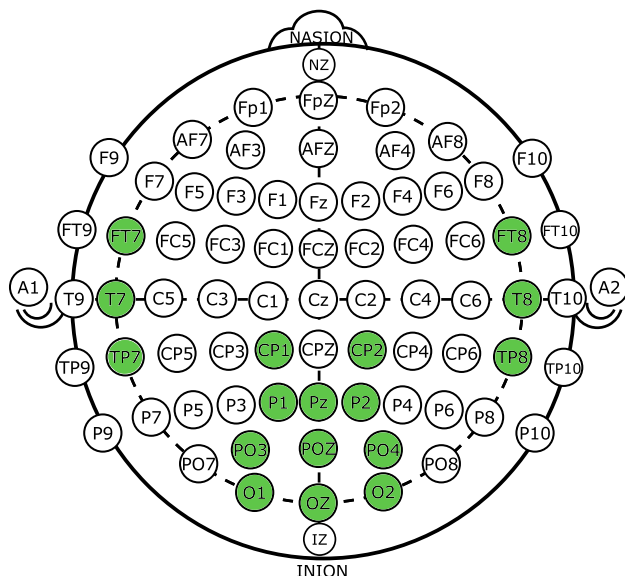
### Feature Extraction

Table 2 shows the sources from which features are extracted, their description, and the total number of features calculated.

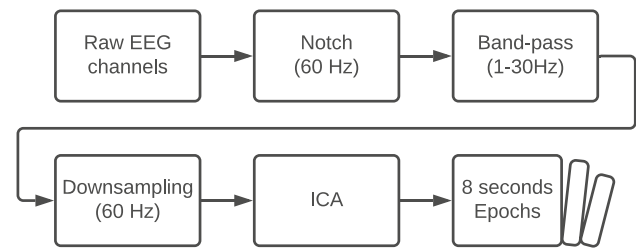


**Fig. 1** Framework overview

Focusing on EEG features, the eight extracted features for each channel are: mean, standard deviation, variance, 5th percentile, first quartile, median, third quartile, and 95th percentile. Thus, a total of  $8 \times 17 = 136$  features are obtained. Moreover, five features using PSD are calculated, one per frequency band among the 17 EEG channels. Finally, this phase



**Fig. 2** Placement of the EEG electrodes used in the SEED-VIG dataset, highlighted in color



**Fig. 3** EEG signals processing phase

calculates a final EOG feature. Then, the classification stage of the framework receives three feature vectors corresponding to the different combinations of features tested: (1) the use of the 136 EEG features; (2) the use of the five PSD features alone; and (3) a combination of PSD and EOG features.

### PERCLOS Discretization Algorithm and Drowsiness Classification

There are two main categories of supervised learning techniques: regression, which predicts numerical values (PERCLOS values in this study); and classification, which produces class assignments. Both categories are used in the framework since either approaches are used in the literature, thus facilitating subsequent comparison of the results.

Since PERCLOS values range from zero to one, it is necessary to map them into three levels of sleepiness, as recommended by Trejo et al. [43] and Chang et al. [44]. Regarding the literature, fixed thresholds are commonly chosen to divide the PERCLOS range of values into the levels of sleepiness. Nevertheless, Gu et al. [45] stated that it is not possible to directly use the thresholds of other studies since they are related to the different detection methods used by different researchers, concluding that the PERCLOS thresholds should be obtained from experiments themselves.

Based on the above, the proposed framework applies a dynamic PERCLOS discretization algorithm to calculate the thresholds between classes for each subject. With this algorithm, the physiological particularities of each subject

**Table 2** Feature sources and description for each Epoch

Feature sources	Description
EEG	Eight features representing the behavior of the signal for each particular electrode in a summarized version via statistical measures. Therefore, 136 features are obtained.
PSD	Five features, each one corresponding to the averaged power of a particular frequency band in the 17 EEG channels.
EOG	One feature corresponding to the number of blinks from the subject.

are taken into account, thus obtaining a personalized division of drowsiness levels that improves data labeling. The threshold between the *minor* and *moderate* drowsiness levels ( $th\_minor$ ) is calculated with Eq. (1) while the threshold between moderate and severe drowsiness levels ( $th\_moder$ ) is obtained by Eq. (2).

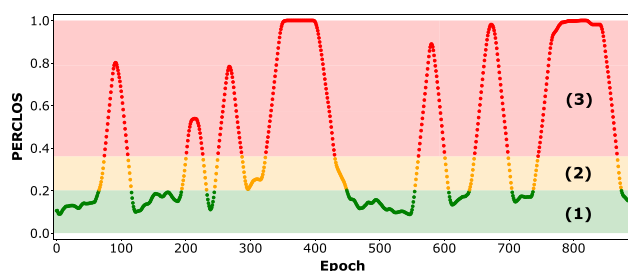
$$\begin{aligned} th\_minor = & \min(PERCLOS) \\ & + (\max(PERCLOS) \\ & - \min(PERCLOS)) * 0.125 \end{aligned} \quad (1)$$

$$\begin{aligned} th\_moder = & \min(PERCLOS) \\ & + (\max(PERCLOS) \\ & - \min(PERCLOS)) * 0.30 \end{aligned} \quad (2)$$

Concerning the static threshold values in the equations, the literature establishes values between 7.5% and 15% for the minor threshold [46, 47]. In particular, this manuscript considers the work performed by Bowman et al. [48], which defined a 12.5% threshold as being an intermediate value within the range. For the moderate range, the literature documents values between 15% and 30% [46, 47, 49]. Based on that, this publication opted for a conservative approach, selecting 30% for Eq. (2). These aspects are aligned with the second contribution of the paper, focused on the creation of a personalized PERCLOS algorithm.

A visual example of the PERCLOS discretization with the proposed algorithm in this study is shown in Fig. 4. The green zone, marked as (1), contains the values where the subject's drowsiness is considered *minor* or fully awake. Subsequently, the yellow zone, marked as (2), indicates *moderate* drowsiness while the red zone, highlighted by (3), represents severe drowsiness.

During the classification stage, the framework uses two different ML model approaches. The first one focuses on training individual and customized models for each user. The second category is based on training scalable models suitable for groups of subjects. Particularly, the two best-performing combinations in the individual models, together with the best



**Fig. 4** Output of the PERCLOS discretization algorithm with tree levels of drowsiness

one from SVM, are used for the scalable analysis, aiming to reduce the complexity of the experimentation. A combination is defined as a ML algorithm together with a vector of features. It is also essential to highlight that this framework does not implement DL algorithms due to the limitations indicated in the “Introduction” section, as the dataset used has a limited size.

To train each model, the PSD and EOG features are normalized using a *MinMax scaler*. Then, the framework shuffles the data before performing the splitting process, which varies according to the categories of models used. In individual models, the 75% of the data defines the training set, while the remaining 25% is used for testing. In contrast, the different combinations of scalable models have their own evaluation sample proportion. Moreover, tenfold cross-validation together with hyperparameter search allows finding the best configuration parameters of a model and achieving the best performance while avoiding overfitting. The algorithms of choice are SVM, kNN, DT, GP, and RF. From them, DT, kNN, and RF were selected based on the literature review previously presented, as these algorithms offer promising results. Finally, GP is selected because, although its behavior is non-Gaussian in contrast to EEG signals, it is interesting to evaluate its performance. It is worth mentioning that these methodological considerations are related to the third contribution presented in the Introduction.

## Experiments and Results

This section presents a set of experiments aiming to evaluate the drowsiness detection performance of individual and scalable ML models using regression and three-class classification techniques, covering the last contribution of the article. Concerning trained models, there is one type of individual models while three types are explored for scalable models:

- Individual models: Personalized classifier and regressor trained and evaluated for each subject.
- 100 models: General classifier and regressor trained and evaluated with the 100% of subjects, where the 75% of the data across subjects is used for training and a 25% for testing.
- 90-10 models: General classifier and regressor trained with the 90% of subjects and evaluated with the remaining 10%.
- 70-30 models: General classifier and regressor trained with the 70% of subjects and evaluated with the remaining 30%.

Regarding regressive models, two metrics are used to measure the quality of the results: RMSE and  $R^2$ . Moreover, four metrics allow to measure the performance of classification



**Table 3** Regression performance for the individual models

Algorithm	Features	RMSE	$R^2$
SVM	EEG	$0.16 \pm 0.05$	$0.42 \pm 0.22$
	PSD	$0.12 \pm 0.04$	$0.67 \pm 0.20$
	PSD+EOG	$0.12 \pm 0.04$	$0.65 \pm 0.21$
kNN	EEG	$0.15 \pm 0.05$	$0.49 \pm 0.20$
	PSD	$0.09 \pm 0.03$	$0.82 \pm 0.09$
	PSD+EOG	$0.10 \pm 0.04$	$0.75 \pm 0.18$
DT	EEG	$0.21 \pm 0.06$	$0.06 \pm 0.39$
	PSD	$0.12 \pm 0.04$	$0.68 \pm 0.19$
	PSD+EOG	$0.12 \pm 0.03$	$0.68 \pm 0.17$
GP	EEG	$0.21 \pm 0.06$	$-0.07 \pm 0.69$
	PSD	$0.13 \pm 0.04$	$0.55 \pm 0.26$
	PSD+EOG	$0.17 \pm 0.05$	$0.26 \pm 0.54$
RF	EEG	$0.14 \pm 0.05$	$0.56 \pm 0.17$
	PSD	$0.09 \pm 0.03$	$0.83 \pm 0.09$
	<b>PSD+EOG</b>	<b><math>0.08 \pm 0.02</math></b>	<b><math>0.83 \pm 0.09</math></b>

Text highlighted in bold indicates the model presenting the best performance

models: accuracy, precision, recall, and f1-score. Particularly, f1-score is prioritized because it involves both precision and recall, making it the most robust and meaningful metric for the analysis.

Since the results of individual, 90-10, and 70-30 models present multiple combinations of different algorithms and subjects tested, the results are presented averaged, subsequently indicated with the following format: Mean  $\pm$  STD.

In contrast, there is no need to do this for 100 models, since there is only one test set with the data reserved from every experiment.

### Individual Models

The performance of the trained regressive individual models is shown in Table 3, where the three assembled feature vectors (EEG, PSD, and EOG+PSD) are used to train each model to observe the performance of each one together with the different ML algorithms evaluated. Generally, it is observed that the lowest RMSE occurs in most cases when only the PSD features are used, followed closely by those using PSD together with EOG and, finally, those utilizing only EEG data. It should also be noted that, although EEG provides the worst results in all cases, these results are acceptable to obtain a good prediction of sleepiness.

As expected, GP performs the worst for all three feature sets since this algorithm is based on the probabilistic theory of the Gaussian distribution, as discussed above. In contrast, SVM and DT offer similar results in terms of their error, improving the results of GP. Finally, kNN and RF are the algorithms with the lowest RMSE. The combination offering the best performance is RF with PSD and EOG features both in RMSE ( $0.08 \pm 0.02$ ) and  $R^2$  ( $0.83 \pm 0.09$ ).

In the same way, the best combinations for classification are quite similar to those previously shown for regression since the algorithms are the same but focused on classification. Particularly, they are presented in Table 4. Nevertheless, the metrics used are different and introduce a series of con-

**Table 4** Classification performance for the individual models

Algorithm	Features	Accuracy	Precision	Recall	f1-score
SVM	EEG	$0.72 \pm 0.10$	$0.71 \pm 0.16$	$0.51 \pm 0.12$	$0.50 \pm 0.13$
	PSD	$0.76 \pm 0.10$	$0.78 \pm 0.12$	$0.58 \pm 0.13$	$0.58 \pm 0.14$
	PSD+EOG	$0.76 \pm 0.09$	$0.074 \pm 0.12$	$0.58 \pm 0.12$	$0.58 \pm 0.13$
kNN	EEG	$0.71 \pm 0.09$	$0.61 \pm 0.13$	$0.52 \pm 0.10$	$0.51 \pm 0.11$
	PSD	$0.85 \pm 0.05$	$0.81 \pm 0.50$	$0.77 \pm 0.08$	$0.78 \pm 0.07$
	PSD+EOG	$0.79 \pm 0.08$	$0.73 \pm 0.12$	$0.63 \pm 0.12$	$0.64 \pm 0.12$
DT	EEG	$0.67 \pm 0.11$	$0.63 \pm 0.14$	$0.47 \pm 0.09$	$0.46 \pm 0.10$
	PSD	$0.80 \pm 0.08$	$0.73 \pm 0.09$	$0.71 \pm 0.08$	$0.71 \pm 0.09$
	PSD+EOG	$0.80 \pm 0.09$	$0.74 \pm 0.09$	$0.72 \pm 0.07$	$0.72 \pm 0.08$
GP	EEG	$0.71 \pm 0.10$	$0.70 \pm 0.13$	$0.48 \pm 0.09$	$0.46 \pm 0.09$
	PSD	$0.70 \pm 0.11$	$0.79 \pm 0.17$	$0.46 \pm 0.12$	$0.42 \pm 0.13$
	PSD+EOG	$0.71 \pm 0.10$	$0.77 \pm 0.12$	$0.48 \pm 0.11$	$0.45 \pm 0.12$
RF	EEG	$0.74 \pm 0.09$	$0.74 \pm 0.12$	$0.54 \pm 0.10$	$0.54 \pm 0.10$
	<b>PSD</b>	<b><math>0.86 \pm 0.06</math></b>	<b><math>0.83 \pm 0.06</math></b>	<b><math>0.76 \pm 0.08</math></b>	<b><math>0.78 \pm 0.07</math></b>
	PSD+EOG	$0.85 \pm 0.06$	$0.83 \pm 0.05$	$0.74 \pm 0.08$	$0.76 \pm 0.07$

Text highlighted in bold indicates the model presenting the best performance

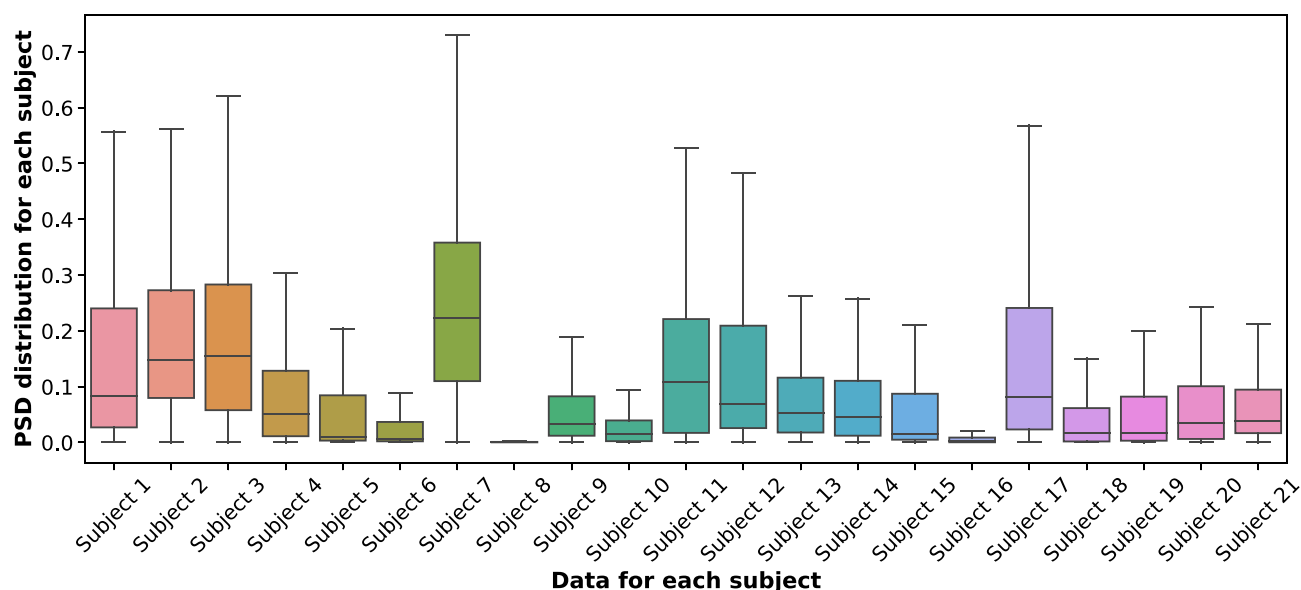


Fig. 5 PSD distribution for each of the subjects included in the SEED-VIG dataset

**Table 5** Example of the variability between subjects, presenting three configurations with different numbers of subjects used in the training set

Configuration	Accuracy	Precision	Recall	f1-score
Train: Subject 1 Test: Subject 21	0.35	0.49	0.46	0.33
Train: Subjects 1–11 Test: Subject 21	0.51	0.50	0.50	0.46
Train: Subjects 1–19 Test: Subject 21	0.60	0.50	0.54	0.52

siderations that cannot be studied from the regressive point of view. In this case, RF with PSD obtains the best performance, with an f1-score of  $0.78 \pm 0.07$ , closely followed by kNN using PSD ( $0.85 \pm 0.05$ ).

It is also relevant to study the variability of the EEG data available in the SEED-VIG dataset for each subject. In particular, Fig. 5 depicts the PSD distribution for each of the 21 subjects in the dataset, highlighting a high inter-subject variability that could affect applying personalized models trained

**Table 7** Classification performance for the 100 models

Algorithm	Features	Accuracy	Precision	Recall	f1-score
SVM	PSD	0.63	0.59	0.51	0.52
kNN	PSD	0.79	0.76	0.76	0.76
<b>RF</b>	<b>PSD</b>	<b>0.83</b>	<b>0.80</b>	<b>0.78</b>	<b>0.79</b>

Text highlighted in bold indicates the model presenting the best performance

with data from one person to another user. To better study these variations, Table 5 presents three examples of models trained with different numbers of subjects and validated with data from Subject 21. Thus, training an individual model for the first subject and testing it on Subject 21 offered a 35% accuracy. In contrast, a model trained with the first 11 subjects offered an accuracy of 51% when evaluated with Subject 21. Finally, a model including the first 19 subjects resulted in a performance of 60% accuracy when evaluated over the last subject. These results indicate that increasing the training set would improve the quality of the predictions for new subjects, thus justifying the need for scalable models.

**Table 6** Regression performance for the 100 models

Algorithm	Features	RMSE	$R^2$
SVM	PSD	0.21	0.43
kNN	PSD	0.14	0.75
<b>RF</b>	<b>PSD+EOG</b>	<b>0.12</b>	<b>0.80</b>

Text highlighted in bold indicates the model presenting the best performance

**Table 8** Regression performance for the 90-10 models

Algorithm	Features	RMSE	$R^2$
SVM	PSD	$0.26 \pm 0.06$	$-1.91 \pm 0.10$
kNN	PSD	$0.20 \pm 0.08$	$-0.66 \pm 0.04$
<b>RF</b>	<b>PSD+EOG</b>	<b><math>0.16 \pm 0.07</math></b>	<b><math>0.03 \pm 0.18</math></b>

Text highlighted in bold indicates the model presenting the best performance

**Table 9** Classification performance for the 90-10 models

Algorithm	Features	Accuracy	Precision	Recall	f1-score
SVM	PSD	0.57 ± 0.23	0.37 ± 0.14	0.41 ± 0.18	0.38 ± 0.15
kNN	<b>PSD</b>	<b>0.60 ± 0.17</b>	<b>0.46 ± 0.12</b>	<b>0.48 ± 0.12</b>	<b>0.46 ± 0.12</b>
RF	PSD	0.61 ± 0.19	0.44 ± 0.18	0.45 ± 0.19	0.43 ± 0.17

Text highlighted in bold indicates the model presenting the best performance

**Table 10** Regression performance for the 70-30 models

Algorithm	Features	RMSE	$R^2$
SVM	PSD	0.26 ± 0.07	−1.90 ± 3.02
kNN	PSD	0.22 ± 0.05	−0.70 ± 0.65
<b>RF</b>	<b>PSD+EOG</b>	<b>0.18 ± 0.05</b>	<b>−0.17 ± 0.45</b>

Text highlighted in bold indicates the model presenting the best performance

## Scalable Models

Once the results of the individual models are available, the two best-performing algorithms in the individual approach (kNN and RF) and the most promising features for each one are selected for further study. In addition, the best combination for SVM is also included due to its large presence in the literature. These three combinations are used to evaluate further models. Subsequently, each scalable model created is presented along with its performance.

### 100 Models

Regression performance is shown in Table 6 where it can be seen that both kNN with PSD and RF with PSD+EOG have a fairly good RMSE and  $R^2$ . SVM with PSD, however, provides inferior performance compared to the other options. These results follow the same trend as the individual models presented in Table 3.

As for the scalable classification models (Table 7), and in the same way as the regression models, the best options are again kNN and RF, in this case, both using PSD as features. Similarly to the results evaluating individual models, SVM has been the worst in performance.

As presented in Table 8, RF using PSD and EOG data is the most promising combination, followed by kNN using

PSD. Finally, SVM offers the worst performance of the three combinations studied.

### 90-10 Models

Since there are a total of 23 experiments over 21 different subjects in the dataset, two subjects (avoiding those who had more than one experiment) corresponding to ~10% of the total are reserved for the evaluation of the model. Subsequently, 21 experiments from a total of 19 subjects are used for training the models.

Relative to the 90-10 classification models (see Table 9), it is important to note that, in this case, kNN with PSD as features performs slightly better than RF with PSD. In contrast, SVM together with PSD offers results almost similar to the last two combinations mentioned, but always slightly worse.

### 70-30 Models

Analogous to the reasoning in the previous models, in this case, 16 experiments (from 14 different subjects) are assigned to model training while the remaining seven experiments, from seven different subjects, are reserved for evaluation.

Table 10 presents the regression results while those corresponding to classification are shown in Table 11. In both approaches, SVM is always the worst of the three combinations. In regression, RF with PSD and EOG remains the best alternative, followed by kNN with PSD. Moving to classification, both kNN and RF with PSD are alternatives to consider, as RF offers a better accuracy compared to kNN but the second one slightly outperforms it in the rest of the metrics. It is interesting to mention that the average f1-score has fallen in all three cases below 40%, which makes this set of models not as interesting as others presented above in terms of performance.

**Table 11** Classification performance for the 70-30 models

Algorithm	Features	Accuracy	Precision	Recall	f1-score
SVM	PSD	0.40 ± 0.20	0.38 ± 0.07	0.41 ± 0.09	0.29 ± 0.14
kNN	<b>PSD</b>	<b>0.45 ± 0.16</b>	<b>0.44 ± 0.07</b>	<b>0.44 ± 0.11</b>	<b>0.37 ± 0.11</b>
<b>RF</b>	<b>PSD</b>	<b>0.46 ± 0.15</b>	<b>0.41 ± 0.06</b>	<b>0.41 ± 0.07</b>	<b>0.35 ± 0.08</b>

Text highlighted in bold indicates the model presenting the best performance

**Table 12** Comparison between the literature works reviewed and the results from the present work. The results are expressed in multiple metrics, where *Acc* represents Accuracy, *CC* means the Pearson correlation coefficient, *RMSE* is the Root Mean Square Error, and  $R^2$  indicates the coefficient of determination

Reference	Acquisition	Processing	Features	Classification	Results
Chen et al. (2015) [25]	EEG, EOG	N/A	Four from DWT, blinking	SVM	Acc: 96.90%
				ELM	Acc: 97.30%
Cheng et al. (2018) [24]	EEG	N/A	PSD	SVM	Acc: 64.05%
				CNN	Acc: 69.19%
Gwak et al. (2018) [16]	EEG, ECG, EOG	Band-pass 1–40 Hz, ICA	PSD, ECG, EOG simulation data	CNN	Acc: 75.30%
				SVM	Acc: 78.60%
				RF	Acc: 81.40%
Li et al. (2018) [32]	EEG, EOG	ICA, downsampling 125 Hz	100 from EEG, 36 from EOG	SVR	CC: 0.76
				SVR	RMSE: 0.17
Wei et al. (2018) [22]	EEG	Band-pass 1–50 Hz, notch 60 Hz, downsampling 250 Hz, ASR	PSD	kNN	Acc: 77.3%
				LDA	Acc: 79.4%
				SVM	Acc: 80.0%
Akbar and Igasaki (2019) [15]	EEG	Band-pass 0.5–50 Hz	Hjorth parameters, PSD	SVM	$R^2$ : 0.64, RMSE: 0.56
				Recurrent SVM	$R^2$ : 0.83, RMSE: 0.15
Chakladar et al. (2020) [21]	EEG	Band-pass 5–32 Hz	PSD, mean, SD, skewness, kurtosis, AR, entropy	RF	Acc: 83.00%
				SVM	Acc: 83.33%
Cui et al. (2021) [19]	EEG	Band-pass 1–50 Hz, AAR, downsampling 128 Hz	Oz channel	DT	Acc: 60.70%
				RF	Acc: 62.60%
				kNN	Acc: 63.42%
				Gaussian Naïve Bayes	Acc: 67.44%
				SVM	Acc: 69.72%
Paulo et al. (2021) [36]	EEG	Band-pass 1–50 Hz, AAR	Raw EEG	CNN	Acc: 75.87%
Shen et al. (2021) [23]	EEG	Band-pass 1–50 Hz, AAR, downsampling 120 Hz	PSD	SVM	Acc: 62.51%
Qian et al. (2021) [33]	EEG	Band-pass <30 Hz	PSD	BCDC	Acc: 94.3%
				GSVM	Acc: 93.7%
Zhu et al. (2021) [18]	EEG	Band-pass 1–60 Hz, notch 50 Hz, ICA, wavelet threshold	Raw EEG	CNN-Inception	Acc: 93.6%
				CNN-AlexNet	Acc: 94.68%
Arefnezhad et al. (2022) [34]	EEG	ICA	Skewness alpha, delta power, theta power, Hjorth delta mobility	Encoder-decoder	RMSE: 0.117
Arif et al. (2023) [35]	EEG	Band-pass 0.5–40 Hz, Notch	PSD, band power ratio indices	Discriminant analysis	Acc: 63.5%
				LR	Acc: 63.6%
				Naïve Bayes	Acc: 67.4%
				SVM	Acc: 75.7%
				DT	Acc: 77.4%
				kNN	Acc: 78.5%
				Ensemble classier	Acc: 85.6%
This work	EEG, EOG	Notch 60 Hz, band-pass 1–30 Hz, downsampling 60 Hz, ICA	EEG, PSD, EOG	Gaussian Process	Acc: 71%
				SVM	Acc: 76%
				Decision Trees	Acc: 80%
				kNN	Acc: 85%
				Random Forest	Acc: 86%



## Discussion

The results for both individual and scalable models suggest that there are ML alternatives to SVM when estimating subjects' drowsiness, although the literature mainly uses this algorithm, sometimes without exploring other ML options. In the case of classification algorithms, to make a fair comparison with the literature, this section relies on accuracy and not on f1-score.

Regarding individual models, and comparing the metrics with the existing literature, Li et al. [32], who used the same dataset but different features, the RMSE obtained in almost every combination of algorithm and features in the framework improves the RMSE of 0.17 provided in their research with SVR. Moreover, the accuracy of 93.6% obtained by Zhu et al. [18] is close to the 86% obtained by the best combination in the framework. Additionally, this work improves the results of many of the works studied from the literature using ML approaches).

As can be seen in Table 12, the best results for the trained individual models are in line with the claims of Gwak et al. [16] where RF performed better than SVM. However, the results contradict Cui et al. [19] and Chakladar et al. [21] since in both studies, SVM performed better or similarly than the other tested ML-based algorithms. This, which may look controversial, can be explained by the features employed by Gwak et al. [16] and the present study, where PSD and EOG features are used. Cui et al. [19] used an entire EEG channel as feature and Chakladar et al. [21] combined PSD with time-domain features.

Therefore, a common pattern is observed: if PSD is used, the model performance obtained is increased compared to not using it and, thus, algorithms such as RF tend to perform better or, at least, similar to SVM. This pattern is observed in studies like Gwak et al. [16] with an accuracy of 81.40% using RF and Chakladar et al. [21] with 83.33% in SVM and 83.00% in RF. This may contribute to a clearer understanding of which features and algorithms should be taken into account when considering training a model for the prediction of drowsiness while driving.

The relevance of PSD as a feature can be clearly explained by how the different EEG brainwaves change between cognitive states. In particular, beta and gamma waves are predominant in demanding cognitive states, such as problem-solving, focused attention, or information processing. In contrast, during drowsy states, theta and alpha are the most common waves. In particular, theta waves are related to relaxation, drowsiness, and early stages of sleep, while alpha activity is predominant when subjects are awake but relaxed [9]. Focusing on the present work, Zheng et al. [14] published detailed information regarding the SEED-VIG dataset, cor-

roborating that the employed dataset predominates theta and alpha brain waves during drowsy states. This situation is also contrasted by the importance of the features of RF used in the present study, where theta, alpha, and gamma frequency bands were the most representative in all models studied.

Concerning scalable models, the 100 model performance is similar to the individual models, implying that having just one model for all users could be enough, compared to having one model per subject. Moreover, the 90-10 and 70-30 models reach an accuracy of  $0.60 \pm 0.17$  and  $0.46 \pm 0.15$ , respectively. In both cases, the performance is greater than 33%, which represents the accuracy of predicting the level of sleepiness randomly. Because of that, these results suggest that it could be possible to develop a scalable model which can predict drowsiness in subjects that are not involved in the experimentation and training phase of the model, although this may depend on the similarity of the subject's features distribution to those used during training.

Despite the promising results obtained, this research presents certain limitations. First, the results depend on the amount and quality of the data used. In particular, the models were trained with a specific group of 21 subjects, which could be insufficient to reach a substantial model generalization. Secondly, this research is limited to exploring the performance of ML algorithms. It is convenient to have access to a larger dataset to explore more complex models, such as those based on DL, able to detect more sophisticated patterns and, thus, achieve better performance.. Based on these limitations, further research is needed to establish the generalization of the findings, using a larger number of subjects during the training and testing phases.

## Conclusions

Drowsiness while driving is a major source of accidents and fatalities. To try to improve this situation, this research presents a framework for drowsiness detection in driving scenarios employing BCIs based on EEG, where different algorithms and feature vectors are used for regression and three-class classification. This is done for both individual and scalable models, where the first ones offer predictions for just one subject, whereas the latter are capable of estimating sleepiness in various subjects despite not having been trained with data from them. In particular, three configurations of scalable models are evaluated, based on the percentage of users employed to evaluate the models that are not included in the training phase. To validate the framework, the SEED-VIG dataset is used, which contains a total of 23 experiments performed in a driving simulator involving 21 different subjects. The labels to be predicted are PERC-

LOS values whose discretization is obtained via a dynamic PERCLOS discretization algorithm, taking into account the physiological particularities of each subject.

The results obtained suggest that PSD features are highly relevant when estimating drowsiness since the best performance for almost every tested algorithm involved PSD, regardless of the learning technique or type of model used. Also, this research illustrates that algorithms such as kNN, RF, or DT may perform equal or better than SVM, the most used algorithm in the literature. Furthermore, GP algorithms are the worst in performance, due to the intrinsic properties of the EEG signals. Lastly, looking at the drowsiness detection performance of the different trained models, the individual models offer the best results, with the limitation of being restricted to a single subject, not being scalable and valid for new users. Next, 100 models, which use the 100% of the subjects for training and testing, provide remarkably similar results to the previous ones while reducing the complexity of the experimentation into a single general model. Finally, the performance of 90-10 and 70-30 models, which reserve the 10% and 30% of subjects for evaluating the models, respectively, show the possibility of predicting drowsiness in subjects not involved during the training phase of the model, although they present a degradation in performance.

As future work, this study first proposes the generation of a new dataset using a BCI, aiming to compare the current results with those obtained from using a larger dataset. Next, it is intended to apply deep learning algorithms for drowsiness estimation, as they are becoming increasingly popular in the literature and could provide better results. Lastly, it is planned to continue working with the scalable 90-10 and 70-30 models to obtain more realistic and robust models capable of predicting drowsiness on a larger set of new subjects.

**Funding** This work has been partially supported by (a) 21628/FPI/21 and 21629/FPI/21 Fundación Séneca, cofunded by Bit & Brain Technologies S.L. Región de Murcia (Spain), (b) Bit & Brain Technologies under the project CyberBrain: Cybersecurity in BCI for Advanced Driver Assistance, associated with the University of Murcia (Spain), (c) the strategic project CDL-TALENTUM from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU, (d) the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberTracer (CYD-C-2020003) project, and (e) the University of Zürich UZH.

**Data Availability** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Competing Interests** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as

long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Kamran MA, Mannan MMN, Jeong MY. Drowsiness, fatigue and poor sleep's causes and detection: A comprehensive study. *IEEE Access*. 2019;7:167172–86. <https://doi.org/10.1109/ACCESS.2019.2951028>.
2. Institute II. Facts + statistics: Drowsy driving. 2022. <https://www.iii.org/fact-statistic/facts-statistics-drowsy-driving>.
3. Shepherd GM. *Neurobiology*. Oxford University Press; 1988.
4. Ibáñez V, Silva J, Cauli O. A survey on sleep assessment methods. *PeerJ*. 2018;6. <https://doi.org/10.7717/peerj.4849>.
5. Shahid A, Wilkinson K, Marcu S, et al. Karolinska sleepiness scale (KSS). 2011. [https://doi.org/10.1007/978-1-4419-9893-4\\_47](https://doi.org/10.1007/978-1-4419-9893-4_47).
6. Hart SG, Staveland LE. Development of NASA-TLX (task load index): Results of empirical and theoretical research. 1988. [https://doi.org/10.1016/s0166-4115\(08\)62386-9](https://doi.org/10.1016/s0166-4115(08)62386-9).
7. Malmivuo J, Plonsey R. Chapter 13 electroencephalography. In: *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press; 1995. p. 247–264.
8. Ward J. *The student's guide to cognitive neuroscience*. Routledge; 2019.
9. Ramadan RA, Vasilakos AV. Brain computer interface: Control signals review. 2017. <https://doi.org/10.1016/j.neucom.2016.10.024>.
10. Nicolas-Alonso LF, Gomez-Gil J. Brain computer interfaces, a review. *Sensors*. 2012;12(2). <https://doi.org/10.3390/s120201211>.
11. López Bernal S, Huertas Celadrán A, Martínez Pérez G, et al. Security in brain-computer interfaces. *ACM Comput Surv*. 2021;54(1):2–3. <https://doi.org/10.1145/3427376>.
12. Ahmed M, Masood S, Ahmad M, et al. Intelligent driver drowsiness detection for traffic safety based on multi CNN deep model and facial subsampling. *IEEE Trans Intell Transp Syst*. 2022;23(10):19743–52. <https://doi.org/10.1109/TITS.2021.3134222>.
13. Zhang X, Yao L, Wang X, et al. A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers. *J Neural Eng*. 2021;18(3):031002. <https://doi.org/10.1088/1741-2552/abc902>.
14. Zheng WL, Lu BL. A multimodal approach to estimating vigilance using EEG and forehead EOG. 2017. <https://doi.org/10.1088/1741-2552/aa5a98>.
15. Akbar IA, Igasaki T. Drowsiness estimation using electroencephalogram and recurrent support vector regression. *Information*. 2019;10(6). <https://doi.org/10.3390/info10060217>.
16. Gwak J, Shino M, Hirao A. Early detection of driver drowsiness utilizing machine learning based on physiological signals, behavioral measures, and driving performance. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE; 2018. p. 1794–1800. <https://doi.org/10.1109/itsc.2018.8569493>.
17. Lin CT, Chuang CH, Huang CS, et al. Wireless and wearable EEG system for evaluating driver vigilance. *IEEE Trans*

- Biomed Circ Syst. 2014;8(2):165–76. <https://doi.org/10.1109/TBCAS.2014.2316224>.
18. Zhu M, Chen J, Li H, et al. Vehicle driver drowsiness detection method using wearable EEG based on convolution neural network. *Neural Comput Appl*. 2021;33(20):13965–80. <https://doi.org/10.1007/s00521-021-06038-y>.
  19. Cui J, Lan Z, Liu Y, et al. A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods*. 2021. <https://doi.org/10.1016/j.ymeth.2021.04.017>.
  20. Garcés Correa A, Orosco L, Laciár E. Automatic detection of drowsiness in EEG records based on multimodal analysis. *Med Eng Phys*. 2014;36(2):244–9. <https://doi.org/10.1016/j.medengphys.2013.07.011>.
  21. Chakladar DD, Dey S, Roy PP, et al. EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. *Biomed Signal Process Control*. 2020;60:101989. <https://doi.org/10.1016/j.bspc.2020.101989>.
  22. Wei CS, Wang YT, Lin CT, et al. Toward drowsiness detection using non-hair-bearing EEG-based brain-computer interfaces. *IEEE Trans Neural Syst Rehab Eng*. 2018;26(2):400–6. <https://doi.org/10.1109/tnsre.2018.2790359>.
  23. Shen M, Zou B, Li X, et al. Multi-source signal alignment and efficient multi-dimensional feature classification in the application of EEG-based subject-independent drowsiness detection. *Biomed Signal Process Control*. 2021;70:103023. <https://doi.org/10.1016/j.bspc.2021.103023>.
  24. Cheng EJ, Young KY, Lin CT (2018) Image-based EEG signal processing for driving fatigue prediction. In: International Automatic Control Conference (CACS). IEEE. 2018. <https://doi.org/10.1109/cacs.2018.8606734>.
  25. Chen L, Zhao Y, Zhang J, et al. Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Syst Appl*. 2015;42(21):7344–55. <https://doi.org/10.1016/j.eswa.2015.05.028>.
  26. Jacobé de Naurois C, Bourdin C, Bougard C, et al. Adapting artificial neural networks to a specific driver enhances detection and prediction of drowsiness. *Accid Anal Prev*. 2018;121:118–28. <https://doi.org/10.1016/j.aap.2018.08.017>.
  27. Hu S, Zheng G. Driver drowsiness detection with eyelid related parameters by support vector machine. *Expert Syst Appl*. 2009;36(4):7651–8. <https://doi.org/10.1016/j.eswa.2008.09.030>.
  28. Yeo MV, Li X, Shen K, et al. Can SVM be used for automatic EEG detection of drowsiness during car driving? *Saf Sci*. 2009;47(1):115–24. <https://doi.org/10.1016/j.ssci.2008.01.007>.
  29. Cui Y, Xu Y, Wu D. EEG-based driver drowsiness estimation using feature weighted episodic training. *IEEE Trans Neural Syst Rehab Eng*. 2019;27(11):2263–73. <https://doi.org/10.1109/TNSRE.2019.2945794>.
  30. Zhuang Q, Kehua Z, Wang J, et al. Driver fatigue detection method based on eye states with pupil and iris segmentation. *IEEE Access*. 2020;8:173440–9. <https://doi.org/10.1109/access.2020.3025818>.
  31. Savas BK, Becerikli Y. Real time driver fatigue detection system based on multi-task ConNN. *IEEE Access*. 2020;8:12491–8. <https://doi.org/10.1109/access.2020.2963960>.
  32. Li H, Zheng WL, Lu BL. Multimodal vigilance estimation with adversarial domain adaptation networks. 2018. <https://doi.org/10.1109/ijcnn.2018.8489212>.
  33. Qian D, Wang B, Qing X, et al. Drowsiness detection by Bayesian-copula discriminant classifier based on EEG signals during daytime short nap. *IEEE Trans Biomed Eng*. 2017;64(4):743–54. <https://doi.org/10.1109/TBME.2016.2574812>.
  34. Arefnezhad S, Hamet J, Eichberger A, et al. Driver drowsiness estimation using EEG signals with a dynamical encoder-decoder modeling framework. *Sci Rep*. 2022;12(1):2650. <https://doi.org/10.1038/s41598-022-05810-x>.
  35. Arif S, Munawar S, Ali H. Driving drowsiness detection using spectral signatures of EEG-based neurophysiology. *Front Physiol*. 2023;14. <https://doi.org/10.3389/fphys.2023.1153268>.
  36. Paulo JR, Pires G, Nunes UJ. Cross-subject zero calibration driver's drowsiness detection: Exploring spatiotemporal image encoding of EEG signals for convolutional neural network classification. *IEEE Trans Neural Syst Rehab Eng*. 2021;29:905–15. <https://doi.org/10.1109/TNSRE.2021.3079505>.
  37. Martínez Beltrán ET, Quiles Pérez M, López Bernal S, et al. Safe-car: A brain-computer interface and intelligent framework to detect drivers' distractions. *Expert Syst Appl*. 2022;203:117402. <https://doi.org/10.1016/j.eswa.2022.117402>.
  38. Xu G, Shen X, Chen S, et al. A deep transfer convolutional neural network framework for EEG signal classification. *IEEE Access*. 2019;7:112767–76. <https://doi.org/10.1109/ACCESS.2019.2930958>.
  39. López Bernal S, Huertas Celdrán A, Martínez Pérez G, et al. Security in brain-computer interfaces: State-of-the-art, opportunities, and future challenges. *ACM Comput Surv*. 2021;54(1).
  40. Neuroscan C. 64-channel quik-cap. 2022. <https://compumedicsneuroscan.com/product/64-channels-quik-cap-synamps-2-rt/>.
  41. Imotions. SMI eye tracking glasses - imotions. 2022. <https://imotions.com/hardware/smi-eye-tracking-glasses/>.
  42. Dingus TA, Hardee HL, Wierwille WW. Development of models for on-board detection of driver impairment. *Accid Anal Prev*. 1987;19(4):271–83. [https://doi.org/10.1016/0001-4575\(87\)90062-5](https://doi.org/10.1016/0001-4575(87)90062-5).
  43. Trejo LJ, Knuth K, Prado R, et al. Eeg-based estimation of mental fatigue: Convergent evidence for a three-state model. In: Schmorow DD, Reeves LM, editors., et al., Foundations of Augmented Cognition. Berlin Heidelberg, Berlin, Heidelberg: Springer; 2007. p. 201–11.
  44. Chang BC, Lim JE, Kim HJ, et al. A study of classification of the level of sleepiness for the drowsy driving prevention. In: SICE Annual Conference 2007. IEEE; 2007. <https://doi.org/10.1109/sice.2007.4421521>.
  45. Gu WH, Zhu Y, Chen XD, et al. Hierarchical CNN-based real-time fatigue detection system by visual-based technologies using MSP model. *IET Image Proc*. 2018;12(12):2319–29. <https://doi.org/10.1049/iet-ipr.2018.5245>.
  46. Nguyen TP, Chew MT, Demidenko S. Eye tracking system to detect driver drowsiness. In: 2015 6th International Conference on Automation, Robotics and Applications (ICARA). 2015. p. 472–7. <https://doi.org/10.1109/ICARA.2015.7081194>.
  47. Celecia A, Figueiredo K, Vellasco M, et al. A portable fuzzy driver drowsiness estimation system. *Sensors*. 2020;20(15). <https://doi.org/10.3390/s20154093>.
  48. Bowman D, Hanowski R, Alden A, et al. Perclos+: Development of a robust field measure of driver drowsiness. In: 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting, 2008.
  49. Selvakumar K, Jerome J, Rajamani K, et al. Real-time vision based driver drowsiness detection using partial least squares analysis. *J Signal Process Syst*. 2016;85(2):263–74. <https://doi.org/10.1007/s11265-015-1075-4>.