Computer Methods and Programs in Biomedicine 243 (2024) 107925

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Validation and interpretation of a multimodal drowsiness detection system using explainable machine learning

Md Mahmudul Hasan [a,b,*], Christopher N. Watling [b,c,d], Grégoire S. Larue [b,e]

[a] School of Computer Science and Engineering, University of New South Wales (UNSW), Australia
[b] Centre for Accident Research and Road Safety - Queensland (CARRS-Q), Queensland University of Technology (QUT), Australia
[c] School of Psychology and Wellbeing, University of Southern Queensland (USQ), Australia
[d] School of Exercise and Nutrition Sciences, Queensland University of Technology (QUT), Australia
[e] Road Safety Research Collaboration, School of Law and Society, University of the Sunshine Coast (USC), Australia

ARTICLE INFO

Keywords:
Features
Physiological signals
Validation
Interpretability
SHAP analysis
Partial dependency analysis

ABSTRACT

Background and objective: Drowsiness behind the wheel is a major road safety issue with efforts focused on developing drowsy driving detection systems. However, most drowsy driving detection studies using physiological signals have focused on developing a 'black box' machine learning classifier, with much less focus on 'robustness' and 'explainability'—two crucial properties of a trustworthy machine learning model. Therefore, this study has focused on using multiple validation techniques to evaluate the overall performance of such a system using multiple supervised machine learning-based classifiers and then unbox the black box model using explainable machine learning.

Methods: Driving was simulated via a 30-minute psychomotor vigilance task while the participants reported their level of subjective sleepiness with their physiological signals: electroencephalogram (EEG), electrooculogram (EOG) and electrocardiogram (ECG) being recorded. Six different techniques, comprising subject-dependent and independent techniques were applied for model validation and robustness testing with three supervised machine learning classifiers, namely K-nearest neighbours (KNN), support vector machines (SVM) and random forest (RF), and two explainable methods, namely SHapley Additive exPlanation (SHAP) analysis and partial dependency analysis (PDA) were leveraged for model interpretation.

Results: The study identified the leave one participant out, a subject-independent validation technique to be most useful, with the best sensitivity of 70.3 %, specificity of 82.2 %, and an accuracy of 80.1 % using the random forest classifier in addressing the autocorrelation issue due to inter-individual differences in physiological signals. Moreover, the explainable results suggest most important physiological features for drowsiness detection, with a clear cut-off in the decision boundary.

Conclusions: The implication of the study will ensure a rigorous validation for robustness testing and an explainable machine learning approach to developing a trustworthy drowsiness detection system and enhancing road safety. The explainable machine learning-based results show promise in real-life deployment of the physiological-signal based in-vehicle trustworthy drowsiness detection system, with higher reliability and explainability, along with a lower system cost.

1. Introduction

Drowsiness is a critical safety issue in road transportation that demands significant attention to alleviate its impact on traffic accidents. Drowsy driving ranks among the primary factors contributing to fatalities, accounting for 20–30% of such incidents between 2011 and 2020, as reported in Australian National Road Safety Strategy [1]. In the pursuit of alleviating fatalities and enhancing safety, researchers are developing drowsiness detection using vehicular, behavioural or physiological measure. Detecting drowsiness through physiological signals shows higher reliability, however; it poses a significant challenge. This is because the development of an accurate drowsiness detection model requires the resolution of several issues, including data pre-processing, feature extraction and selection, the selection of appropriate

* Corresponding author at: School of Computer Science & Engineering (CSE), Building K17, UNSW, Sydney 2052, Australia.
E-mail address: md_mahmudul.hasan@unsw.edu.au (M.M. Hasan).

https://doi.org/10.1016/j.cmpb.2023.107925
Received 3 September 2023; Received in revised form 28 October 2023; Accepted 7 November 2023
Available online 8 November 2023

approaches and signal combinations, selecting, and designing appropriate classification models, validation, and interpretation of the models.

Previous endeavours on drowsiness detection using physiological signals are focused on feature extraction methods, used a singular validation, and mostly use unimodal approaches with traditional machine learning classifiers. For example, Taran and Bajaj [2] have utilised Hermite functions based decomposition for electroencephalogram (EEG) feature extraction, utilised six different classifiers and obtained the best accuracy of 95.45 % using extreme learning machine (ELM) with 10-fold cross-validation (CV) [2]. Sharma et al. was also focused on feature extraction using Wavelet transform from EEG signals and reported an accuracy of 95.6 % in their study (using a 10-fold CV) [3]. Khare et al. [4] have used a novel feature extraction method 'variational non-linear chirp mode decomposition' and reported an accuracy of 92.4 % using a Boosting Tree classifier. Lee at al. [5] proposed a Recurrence Plots (ReLU-RP), which shows 4–17 % better accuracy for ECG and photoplethysmogram (PPG) signals. Babaeian and Mozumdar [6] used wavelet (WT) and short Fourier transform (STFT) for extraction of features from electrocardiogram (ECG), used support vector machines (SVM) and k nearest neighbours (KNN) as classifiers and obtained the best performance of 87.5 % using SVM classifier (rando split). Similarly, Chui et al. [7] utilised EEG signal and developed electrocardiogram genetic algorithm-based support vector machine (ECG GA-SVM), where they obtained 97.01 % accuracy (using 10-fold CV).

Validation is an essential approach to unifying the model's performance and test the robustness of the machine-learning-based system. Among the validation methods utilised, the K-fold cross-validation [2,4] in the above mentioned studies overcomes the limitations of holdout validation (or train-test-split approach) by repeatedly validating with different random seeds. However, it has some limitations for imbalanced datasets, and stratified k-fold validation is used to overcome it (accuracy outcomes ranging from 91.8 to 92.13 % [e.g., 8,9]). As the validations mentioned earlier do not address autocorrelation issue for physiological signals, time series split cross-validation is employed to overcome this [10]. Nevertheless, none of the validations mentioned above address the inter-individual difference in physiological data. Leave-one-out techniques, including Leave-One-Trial-Out (LOTO) and Leave-One-Participant-Out (LOPO), address both autocorrelation [11] and inter-individual differences [12]. LOPO has been commonly used for driver drowsiness detection with multiple trials and subjects, yielding sensitivity and specificity outcomes ranging from 58.0 to 98.8 % and 98.3–98.2 %, respectively [e.g., 13,14,15]. However, Watling et al. [16] reported that the study results, especially metrics related to the overall performance of the model cannot be compared due to the varied validation methods employed.

While the above studies worked with mostly feature extraction and unimodal signals, a recent critical review by Yaacob et al. [17] suggested the necessity of using multimodal fusion and explainability for drowsiness detection. Not only in drowsiness detection, a systematic review by Khare et al. [18] in medical domain using physiological signal also suggested the need for data fusion and explainability in machine learning using such signals. In the pursuit of exploring multi-modal area, a few studies worked with hybrid combinations of physiological signals for drowsiness detection. Oliveira and colleagues [19] have performed a multi-modal machine learning based analysis and concluded that performance improved for combined method (electrooculogram (EOG) + electrocardiogram (ECG)) for drowsiness detection (10-fold CV). Amin et al. also performed a multimodal study [20] combing EEG and EOG for a real-time drowsiness detection system and reported an accuracy of 81 %, however, they have only used three participants in their study to evaluate the performance (no report on validation). Hasan et al. [21] performed an extensive comparative study of unimodal and multimodal physiological signals for drowsiness detection, which includes seven different combinations of EEG, EOG and ECG (unimodal: EEG, EOG, ECG; multimodal: EEG + EOG, EEG+ ECG, EOG + ECG, and EEG + EOG + ECG). Their results indicate that the multimodal fusion of EEG, EOG and ECG gives 7.5 % (mean) better accuracy than the unimodal signals and further helps reducing the disparity between sensitivity and specificity (8.0 %). The study found that among the multimodal combinations, fusion of EEG, EOG and ECG gives the best performance, producing 83.5 % accuracy using artificial neural networks (ANN); however, the study used a 10-fold cross validation across four different classifiers, uses a block-box machine learning approach and lacs explainability.

From the review of the previous studies, it is obvious that explanation of the identified features and the interpretation of machine learning models has received less focus than performance metrics [16–18], preventing stakeholders from understanding the reasons behind their decisions, and hindering the implementation of research solutions in the market [16]. However, explainability is an essential part of a trustworthy machine learning [22]. As such, it is important to unbox the black-box model for safety issues, to understand and explain how the system is making its predictions and help to build trust in the system and increase its reliability [23], transparency and accountability, which is especially important in safety-critical applications such as driver drowsiness detection [24]. It is also important to ensure that the predictions it makes are fair and unbiased, as it allows stakeholders to understand the factors that are being considered in the decision-making process. Second, it is unclear which validation method shows promise in evaluation of the drowsiness detection classifier, especially as very few studies have compared multiple validations methods using the same data source—raising a question on the robustness of the system. According to Raja and colleagues [22], 'explainability' and 'robustness' are two essential parts of a 'trustworthy' machine learning model, where the verification/validation is inseparable part of robustness. Therefore, these two issues remain under-addressed when dealing with drowsiness detection using physiological signals and machine learning, which raises concerns about the trustworthiness of the system.

Considering the need for explainability in machine learning model and the need to test multiple validation methods for ensuring 'trustworthiness', this study specifically focuses on explainable machine learning for drowsiness detection using multimodal physiological signals, unboxing the 'black-box' machine learning model for a real-world drowsiness detection system. To the best of our knowledge, explainable analysis using the SHapley Additive exPlanations (SHAP) and partial dependency analysis (PDA) have not yet been performed in terms of physiological signals-based drowsiness detection studies, especially for a multimodal system using electroencephalography (EEG), electrocardiography (ECG), and electrooculography (EOG), which uses PVT as drowsiness stimuli with respect to KSS as sleepiness measure. Therefore, the current study sought to assess the 'trustworthiness' of a multimodal physiological signal-based machine drowsiness detection system, with a focus on 'explainability' and 'robustness'.

2. Method

2.1. Participants

In total, 35 individuals participated in this study who were between the ages of 17–25. The study inclusion criteria also required participants to have a habitual sleep duration of more than 7 hours, a habitual bedtime no later than 12 am, no sleep-related disorders or on medication that affects sleepiness or arousal.

2.2. Study design

This was an experimental study, specifically conducted in a laboratory setting. As a stimulus to measure behavioural alertness, a 30-minute-long customised psychomotor vigilance task (PVT) based experiment was developed utilizing an experimental software for psychological study design (PEBL). Participants were presented with a

computer-screen prompt every five minutes, where they were asked to self-assess their degree of drowsiness using the Karolinska Sleepiness Scale (KSS), which was recorded by the software [21].

2.3. Instruments

2.3.1. Subjective sleepiness

The KSS is a widely used scale for the measurement of subjective sleepiness. In this measurement system, participants rate their present degree of subjective sleepiness calibrated on a 9-level Likert scale, where the higher levels in the calibration indicates higher measure of sleepiness [25]. It is used a valid and reliable measurement while detecting subjective sleepiness [26] and as such it was used as ground truth in the current study.

2.3.2. Physiological signal acquisition

Several physiological metrics can show variations in drowsiness [12], with a substantial amount of research indicating various metrics from multimodal biomedical signals (EEG, EOG, and ECG) show alterations in drowsiness, which were used in this study. The EEG electrode sites were left-central (C3-A2) and left occipital (O1-A2), being the signal recorded using gold cup electrodes. The electrodes for EOG data acquisition were positioned directly below the left eye's pupil in a vertical orientation. For ECG signals acquisition, a modified system of chest-lead was used containing a pair of electrodes, with one positioned around 3–5 cm underneath the collarbone and another on lower part of left ribcage. The ground electrode for the participant was situated beneath the left clavicle. To adhere to electrophysiological recording standards, the minimum impedance of all electrodes was no more than 5 kilo-Ohm (kΩ).

The electrophysiological data was collected using a electrophysiological wireless signal acquisition device named BioRadio 150 as well as an integrated software named BioCapture [27]. The frequency of the electrophysiological data collection was 600 Hz. It is worth to note that an online Notch filter at 50 Hz was used applied during data collection to mitigate the impact of line frequency. A band-pass filter (0.3–35 Hz for EEG as well as EOG, and 0.3–70 Hz for ECG) was applied as apart of signal preprocessing.

2.3.3. Psychomotor vigilance task (PVT)

The psychomotor vigilance task (PVT) is a widely recognised method for evaluating levels of alertness based on behavioural observations. During the task, participants are required to maintain their focus on a small fixation (cross symbol), which is displayed on the screen for a duration of 400 milliseconds. Following this, a red dot pops up at random intervals (ISI: 1–10 seconds). Participants must respond to the stimulus as quickly as possible via a keystroke. The PVT measures behavioural alertness through various parameters, including reaction time to the stimuli and the number of response lapses (i.e., response times exceeding 500 milliseconds). The task is free of learning and practice effects, making it an ideal tool for extended testing sessions. Typically, the standard PVT test lasts for a duration of 10 minutes, however, 5- and 3-minute versions have been employed as well as longer durations (e.g., 30-minute) [28]. All these methods have demonstrated their usefulness in detecting changes in alertness, with the PVT being notably reliable and valid in measuring behavioural alertness [29]. An extended version of PVT having a 30-minutes duration was utilised in the experiment of the current study, utilizing the PEBL software for the task administration.

2.4. Experimental procedure

The research experiment was performed with human subjects, with ethical clearance provided by the Queensland University of Technology (QUT) Human Research Ethics Committee. Participating in the study involved two sessions: intake-screening and a testing session. During the first session, the subjects signed a form providing their consent to collect their physiological data and were being provided with an actigraph watch, which they wore so that their normal sleep patterns can be observed for a minimum of five days and the observation can be used to ensure compliance with the study's protocols. A sleep diary was also provided to fill up sleep-awake time.

The subjects attended the second session at the laboratory at 2:00 p.m. as a part of the testing session. Before starting the testing session, they were required to have obtained their usual amount of sleep in the nights prior to the experiment. Upon arrival at the laboratory, their actigraph data were reviewed to confirm adherence to their normal sleep-awake patterns in the three days preceding the final experiment. If there were any significant deviations from their typical sleep pattern, the test session was postponed to a later date. If technical problems occurred with the actigraph, the participant's sleep diary was utilised to evaluate their sleep patterns.

Once it was determined the participant had followed the study protocol, they were provided with an explanation of the KSS ratings, which measure subjective levels of sleepiness, and were given a brief introduction to the PVT task that lasted for one minute. Following this, the participant underwent the main session, during which EEG, EOG, and ECG electrodes were attached while they performed the PVT task on a computer screen for a period of 30-minutes as part of the experiment. While completing the PVT task, participants' subjective sleepiness levels were measured every five minutes, using the KSS scale. The PEBL software controlled the PVT and KSS. A webcam was installed to observe the participants while completing the task. It was noted that some of the participants were not fully committed to the task, which was further confirmed with higher number of lapses (>25) and therefore those participants were removed from the study. As such, 26 participants data was utilised for the purpose of further analysis.

2.5. Data preprocessing, feature extraction and selection

The collected biosignals data was preprocessed and filtered prior to feature extraction that included time- and frequency-based domains. Finite Impulse Response (FIR) bandpass filter (LF: 0.3 Hz, HF: 35 Hz (EEG/EOG), 70 Hz (ECG)) with a 'Hanning window' was used for the filtering purpose. The reason behind using the FIR based Hanning window is that the Hanning window gives the best Signal to Noise ratio (SNR) [30].

Among all the EEG metrics, the fundamental EEG bands, EEG-α, EEG-θ and EEG-β power spectra, and the EEG-band ratio, i.e.,(theta + alpha)/beta, (theta + alpha)/(alpha + beta) and beta/ alpha beta/ alpha in central and Occipital channel proved to be the more sensitive indices for drowsiness detection in both non-professional and professional drivers [16,21]. It is important to note that the EEG data was collected from two electrode positions (Ch1: Central region, C3), Ch2 (Occipital Region-O1); the A2 position was used as a reference and the EEG sub-bands found from the spectral analysis are divided into five categories: delta (0.50–4.00 Hz), theta (4.01–7.00 Hz), alpha (8.00–15.00 Hz), beta (16.00–32.00 Hz), and gamma (36.00- 44.00 Hz) [21]. Among all the EOG metrics, blink duration, blink frequency, blink amplitude, peak closing velocity (PCV), and amplitude/velocity ratio of blinks (AVRs or A/PCV) proved to be the more sensitive indices than the other features for drowsiness detection as assessed by the literature [16,21]. Among all the ECG metrics, heart rate (HR) and R-R interval (RRI) have proven to be the most useful features in the time domain, and signal power at low

Table 1
The lower and upper cut off frequency for data preprocessing & filtering.

Channel	Low Freq Filter	High Freq Filter	Notch Filter
EEG	0.3 Hz	35 Hz	50 Hz
EOG	0.3 Hz	35 Hz	50 Hz
ECG	0.3 Hz	70 Hz	50 Hz

frequency (LF), high frequency (HF) and LF/HF ratio are to be the most prominent features for drowsiness detection in the frequency domain as assessed by the literature [16,21]. It is important to note that HR and RRI are in the time domain, while LF, HF and LF/HF are in the Frequency Domain Heart Rate Variability (HRV) metrics.

All the EEG features were extracted using the Acqknowledge®−4.2 software, which utilises the Welch periodogram method [31] to compute the Power Spectral Density (PSD) of the EEG-frequency bands. The Welch method calculates the Power Spectral Density of a signal by dividing it into overlapping segments, applying a window to each segment to reduce spectral leakage, and then averaging the results to obtain a more accurate representation of the signal's frequency content [31]. While Acqknowledge®−4.2 offers windowing options such as Hanning, Hamming, or Blackman for PSD computation, we specifically calculated using the 'Hanning window' to achieve a better Signal-to-Noise Ratio (SNR) [30]. Note that all the EOG features were extracted running our developed and customised MATLAB®−2018 code. For finding the EOG based eye blinks, we tuned a 'threshold' to identify the local maxima, i.e., eye blinks in EOG, which was further used for 'derivative analysis' to compute PCV and AVR [32]. This method involves taking the derivative of the EOG signal with respect to time, where the peak closing velocity corresponds to the point in the derivative where the signal changes most rapidly [32]. This was detected by finding the maximum or minimum value in the derivative. Considering in indexes of the derivative changing from negative to positive and vice-versa, the amplitude, blink duration, blinking rate and AVR was computed [32]. All the ECG features were calculated using Acqknowledge®- 4.2 software, which employs a built-in 'QRS detector' for heart rate variability analysis (HRV), based on a modified Pan-Tompkins algorithm [33]. Power Spectral Density at LF and HF were also calculated using Welch periodogram method [31] with a 'Hanning window' [30].

Based on the literature reviewed, a total of 22 features were derived from the three electrophysiological data [21] (Table 2). Considering the nature of dataset, and reviewing the previous studies [34,35], we have primarily utilised filter-based feature selection methods that rank a subset of features independently of any learning algorithm. Specifically, two univariate filter-based feature selection methods: the 'analysis of variance (ANOVA) F-Test' and 'correlation-coefficient' based ranking were applied to the dataset and later the feature subsets were aggregated using a technique stability feature selection [21].

Table 2
Extracted features from the physiological signals.

Sl	EEG	Sl	EOG	Sl	ECG
1	Alpha Central (α-EEG C3) (μV^2/Hz)	1	Blink Duration (BD) (seconds)	1	Mean Power at Low Frequency (Mean P- LF) (μV^2/Hz)
2	Theta Central (θ-EEG C3) (μV^2/Hz)	2	Amplitude Velocity Ratio (AVR)	2	Mean Power at High Frequency (Mean P-HF) (μV^2/Hz)
3	Beta Central (β-EEG C3) (μV^2/Hz)	3	Peak Closing Velocity (PCV) (degrees/s)	3	R-R Interval (RRI) milliseconds (ms)
4	Theta Occipital (θ-EEG O1) (μV^2/Hz)	4	Blinking Rate (BR)	4	Heart Rate (HR) (beats per minute- bpm)
5	Beta Occipital (β-EEG O1) (μV^2/Hz)	5	Amplitude (Amp)	5	LF/HF Ratio
6	Alpha Occipital (α-EEG O1) (μV^2/Hz)				
7	β /α Central (C3)				
8	(θ + α) / β Central (C3)				
9	(θ + α) / (α+ β) Central (C3)				
10	β /α Occipital (O1)				
11	(θ + α) / β Occipital (O1)				
12	(θ + α) / (α+ β) Occipital (O1)				

Note: EEG, electroencephalography; EOG, electrooculography; ECG, electrocardiography.

2.6. Dataset and KSS distributions

In total, 35 participants took part in the experiment. The data from a total of nine participants were excluded due to various reasons, such as a higher number of PVT lapses, frequent postural movements, or head nodding, and the data from the remaining 26 participants were used for the subsequent analysis. From this data, after removing the unexpected movement artefacts by filtering and manual screening, a total of 22 features were extracted from the EEG, EOG and ECG signals based on the existing literature (described in Section 2.5). All the proposed features were extracted from each physiological signal using a 5-second epoch length, except the blinking rate from EOG and the heart rate from ECG, which were calculated per minute. In total, total 9360 observations were taken for 26 participants, with the given epoch size. Then the mean values for all the attributes were calculated for each 5-minutes session using pivot analysis, considering a 5-minutes prediction window. The data was compressed to 156 observations for 26 participants. Among the 156 observations, 79 observations were obtained for drowsy states (positive class) and 77 observations for awake states (negative class) (Table 3 and Fig. 1). A total of seven KSS values was obtained from each subject during each 30-minutes test session, as the KSS score was considered to apply to the five minutes preceding each reported rating.

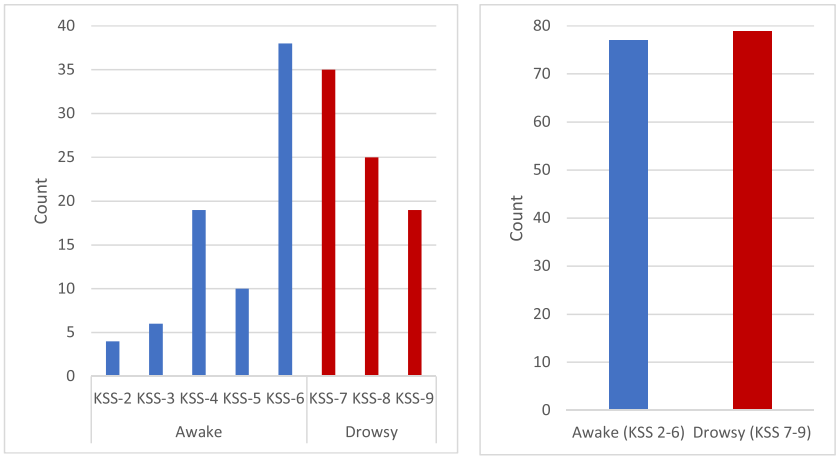
2.7. Classification and drowsiness level detection

Several studies have employed various machine learning models to detect drowsiness from biosignals, and there is no established rule on which classification model to use for specific applications or participant groups [14]. Considering the evident variations in computational expenses and intricacy as reported in previous studies, and in alignment with the existing scholarly literature, three supervised learning models were utilized for classification purposes. These models encompassed the k-nearest neighbor (KNN), support vector machines (SVM), and random forest classifier (RF) which are the most popular models used by several studies while detecting drowsiness using physiological signals [21,36]. Feature scaling was performed in terms of normalization (Min-Max scaling) and standardisation before applying KNN and SVM, respectively. All the classifiers were used to examine the utility of different validation techniques in the current study; however, random forest classifier was used for interpretability, as it does not require feature scaling and facilities explainability.

A binary classification was performed considering two classes for a set of KSS scores: drowsy and awake. Given that crash risk is highly associated with KSS 7–9 [37], this range was used as drowsy state and KSS 2–6 was the awake state. The hyperparameters for each classifier were fine-tuned through an iterative process to achieve the highest accuracy following cross-validation. In the case of KNN classifiers, the

Table 3
Score distribution across the dataset for all the eligible participants.

Participant ID	Awake state					Drowsy state		
KSS Score	KSS-2	KSS-3	KSS-4	KSS-5	KSS-6	KSS-7	KSS-8	KSS-9
KSS Count	4	6	19	10	38	35	25	19
KSS Count in each class			77				79	
Total KSS count				156				



(a) (b)

Fig. 1. Counts of KSS-scores across all the samples on the whole dataset (a) KSS categorical count over score 2–9 (b) KSS total count (drowsy and awake states).

optimal 'k' value was determined during model training, selecting the 'k' value associated with the highest training accuracy. For Gaussian SVM, the 'radial basis function (RBF) kernel' was employed, and a 'grid-search' was performed to optimize the 'C' and 'γ' parameters using the 'grid-search' algorithm, ultimately selecting the parameter values that yielded the best accuracy after cross-validation. The count of trees, referred to as 'n estimate,' and the maximum depth of each tree, known as 'max-depth,' were adjusted using the scikit-learn library's 'grid--search' approach while determining the optimum hyper parameters for the random forest model (Table 4). Python 3.6.7 was used to implement, train, and test all of the models in Google Colab platform.

2.8. Performance measures

The effectiveness of a machine learning model is evaluated based on various measures, including sensitivity, specificity and accuracy [38]. Sensitivity measures the proportion of correctly identified positive samples, while Specificity measures the same thing for the negatives [21]. Additionally, accuracy signifies the ratio of correctly detected samples overall, regardless of whether they are positive or negative [21]. These three measures (sensitivity, specificity, and accuracy) were utilised in this study to validate the drowsiness detection model (Fig. 2).

2.9. Validation

Six different cross-validation (CV) techniques were applied to the experimental data. The details of the techniques are given below.

2.9.1. Participant dependent validation

Participant dependent validations in this study encompassed various techniques. Holdout validation involved random partitioning into training (70–80 %) and test (20–30 %) datasets, utilizing 80:20 splitting. K-fold cross-validation divided the dataset into k sub-folds, iteratively using (k-1) folds for training and one for validation. Stratified k-fold validation ensured an equal distribution of target class labels (drowsy and awake) in both training and test data. Time series split cross-validation was applied for datasets with correlated time series data points, preventing the inclusion of neighbouring data points in training and test sets, addressing autocorrelation issues [10].

Table 4
Parameters considered for hyper parameter tuning for the different classifiers.

Classifiers	Parameter tuned
KNN	• Number of neighbours (k)
SVM	• 'C' value • Gamma (γ with 'rbf' kernel)
RF	• Number of trees in the forest (n_estimate) • The maximum depth of the tree (max_depth)

*Note:*KNN: K-nearest neighbours, SVM:support vector machines and RF: random forest.

2.9.2. Participant independent validation

Participant independent validation techniques in this study included Leave One Out Cross-Validation (LOO CV) and Leave One Participant Out Cross-Validation (LOPO CV). LOO CV involved N iterations, where N represents the number of instances, using (N-1) instances for training and one for validation in each iteration. For this study with $N = 156$ instances, 155 instances were used for training in each iteration. LOPO CV grouped data based on participant IDs, creating 26 groups for the 26 participants. Cross-validation was performed 26 times, with one participant's data used for validation and the remaining participants' data for training in each iteration.

2.10. Interpretation

2.10.1. Shapley additive analysis (SHAP)

SHapley Additive exPlanations, abbreviated as SHAP is a method for explaining the output of machine learning models by attributing the prediction to the features that contributed to it [39]. It does this by using the concept of Shapley values, a method from game theory, for fairly distributing the "credit" for a prediction among the features that contributed to it. SHAP values are useful to calculate the importance of each feature in a model's prediction, which help explain how the model arrived at its prediction by showing which features had the greatest influence on the final output.

Shapley values represent the importance of features in machine learning models with multicollinearity. This approach entails retraining the model on different feature subsets (S ⊆ F, F being the collection of all features), assigning importance values to each feature to gage its influence on model predictions. The calculation involves training a machine learning (ML) model ($f(x)$, x being the feature sets) both with and without a particular feature, and comparing their predictions ($f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$). As this impact is dependent on other features, these differences are computed for all possible subsets $S \subseteq F\backslash\{i\}$. The Shapley values ($\varphi_i$) are then derived from these computations and serve as feature attributions, representing a weighted average of all potential differences [39] (Eq. (1)).

$$\varphi_i = \sum_{S\subseteq F\backslash\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right] \quad (1)$$

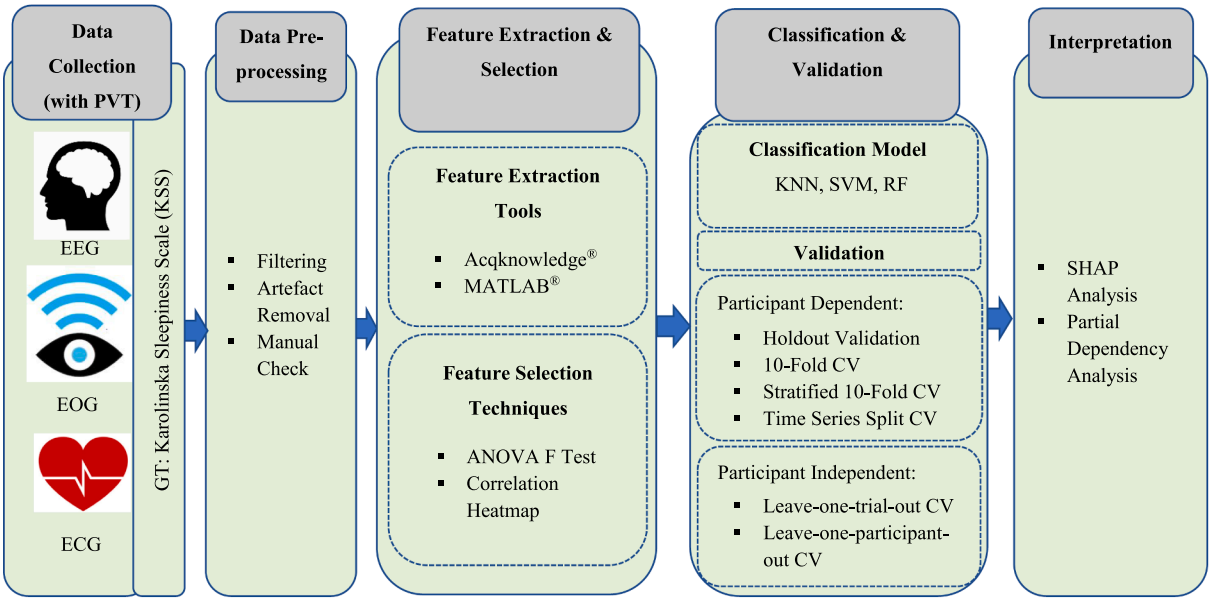


Fig. 2. Proposed methodology for validation and interpretation of multimodal physiological signal-based drowsiness detection system. Note: EEG: electroencephalography, EOG: electrooculography, ECG: electrocardiography, GT: ground truth, PVT: Psychomotor vigilance task, CV: cross-validation, KNN: K-nearest neighbours, SVM:support vector machines and RF: random forest, SHAP: SHapley Additive exPlanation.

2.10.2. Partial dependency analysis

Partial dependency analysis is a technique, which is normally used for interpreting and explaining the output of machine learning models [40]. PDA help understand the relationship between a single feature in relation to the model's output, holding all other features constant. The resulting plot from PDA presenting the correlating the attribute or feature values and the trained model's output is called partial dependency plot (PDP), which help identify the important features in the model's prediction and understand how the model is using them. Independent Component Analysis (ICA) is a statistical technique of the PDA framework, which usually helps to interpret and identify the underlying independent components of dependency.

The partial dependence (*PD*) of a machine learning algorithm's output ($g(x)$) on a specific set of variables x_S is defined as $PD(x_S)$. This is the measured by the expected value of output (g) considering the marginal distribution of all variables except those in x_S. Thus, it represents how the output (g) depends on the chosen subset of variables (x_S)—with x_C being the set of variables not included in x_S, Ex_C being the expectation (i.e., the average with respect to the distribution of the complement set x_C), $g(x_S, x_C)$ being the model's output for the given subset x_S and complement set x_C, $P x_C$ being the probability distribution of the variables in the complement set x_C, and can be expressed by the following Eq. (2) [41].

$$PD(x_S) = E_{x_C}[g(x_S, x_C)] = \int g(x_S, x_C)\, dP(x_C) \tag{2}$$

It is worth emphasizing that both Shapely and Partial Dependency Analysis techniques are applicable to any machine learning classifier. However, due to the necessity of feature scaling for K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) utilised in this study, we exclusively employed Random Forest for the explanatory analysis. Random Forest can effectively operate with unscaled or original features, thereby enhancing the interpretability of the model's explanations.

3. Results

3.1. Feature selection and classification performance

Two feature selection techniques were applied on the extracted 22 features from the electrophysiological signals (see Section 2.5). After applying the feature selection methods (ANOVA F-Test and correlation-coefficient), a total 13 features were finally selected based on stability feature selection techniques [21,36] (Table 5). Thus, two feature sets were obtained from three of the signals with associated target values, which were supplied to the machine learning classifiers as training and test data. The proposed system's performance with different validation techniques (positive and negative label distributions for holdout validation: Table 6, for 10-fold, stratified 10-fold and time series split CV: Table 7, for LOO and LOPO CV: Table 3) with two sets of features are summarised in Table 8. It is important noting that the holdout validation was applied by splitting the data once; therefore, there is no standard deviation reported for the holdout validation. For the other validation methods, the means and standard deviations were reported.

The holdout validation was performed in two ways with same random split - one with 80:20 while the another is using 70:30 representing train: test data. While using 80 % of the training instances was used to train the model holding out the rest for testing, the sensitivity, specificity, and accuracy ranged from 61.5 to 78.9 % for KNN, 63.2–76.9 % for SVM, and 76.9–78.9 % for RF using 22 features. When the selected 13 features were used, the performance was increased and was ranged from 78.9 to 84.6 % for KNN, 68.4–84.6 % for SVM 76.9–78.9 %, and 78.9–84.6 % for RF. Using a 10-Fold cross-validation technique with 22 features, the sensitivity, specificity, and accuracy ranged from 56.1 to 68.2 % for KNN, 54.1–76.4 % for SVM 76.9–78.9 %, and 63.9–66.6 % for RF. When the selected 13 features were used, performance of all the metrics were improved, ranging from 58.9 to 70.2 % for KNN, 60.5–76.4 % for SVM 76.9–78.9 %, and 67.7–70.5 % for RF. Stratified 10-fold cross-validation yields comparable performance to the standard 10-fold cross-validation. The reason could be that the data collected was having almost equal number of positive and negative samples/instances.

The time series cross validation was performed using $k = 10$ splits. While using 22 features, the mean performance metrices were ranging between 49.3–66.0 % for KNN, 43.1–87.8 % for SVM, and 38.60–76.5 % for RF. When the 13 features have been used after feature selection, specificity and accuracy improved for KNN by 3.6–23.2 %, with a considerable drop in sensitivity (25 %); sensitivity and specificity improved for SVM by 17.4–24.6 %, with a slight drop in accuracy (1.5

Table 5
Aggregating feature rankings (shaded features signifies the excluded features; exclusion criteria, ANOVA F value<2.0, Correlation coefficient <0.10).

Signals	Rank	Features	F Values	Correlation Coefficient	Selection criteria- F Value (If F-values> 2, assigned 1; else 0)	Selection criteria- - Correlation Coefficient (If the coefficient>0.1, assigned 1; else 0)
EEG	1	α C3	11.33	0.26	1	1
	2	θ C3	7.43	0.21	1	1
	3	β C3	6.15	0.20	1	1
	4	θ O1	3.93	0.16	1	1
	5	β O1	3.72	0.15	1	1
	6	α O1	3.31	0.15	1	1
	7	β /α O1	0.75	0.07	0	0
	8	β /α C3	0.5	0.06	0	0
	9	(θ + α) /(α+ β) C3	0.3	0.04	0	0
	10	(θ + α) / β O1	0.14	0.03	0	0
	11	(θ + α) / β C3	0.08	0.02	0	0
	12	(θ + α) /(α+ β) O1	0.000345	0.001497	0	0
EOG	1	BD	34.77	0.43	1	1
	2	AVR	28.15	0.39	1	1
	3	PCV	3.04	0.14	1	1
	4	BR	0.44	0.05	0	0
	5	Amp	0.33	0.05	0	0
ECG	1	Mean P- LF	10.55	0.25	1	1
	2	Mean P- HF	10.34	0.25	1	1
	3	RRI	5.64	0.19	1	1
	4	HR	4.94	0.18	1	1
	5	LF/HF Ratio	0.27	0.04	0	0

Note: α-C3, Alpha Central; θ-C3, Theta Central; β-C3, Beta Central; α-O1, Alpha Occipital; θ-O1, Theta Occipital; β-O1, Beta Occipital; BD, Blink Duration; AVR, Amplitude Velocity ratio; PCV, Peak Closing Velocity; BR, Blinking Rate; Amp, Amplitude; LF/Mean P-LF, Mean Power at Low Frequency; HF/Mean P-HF, Mean Power at High Frequency.

Table 6
Distribution of the resulting labels across holdout validation schemes.

Holdout (Test data:20 %, RS =1), Trai-test split (RS=1)					Holdout (Test data:30 %, RS =1) Trai-test split (RS=1)				
Train data		Test data		Train + Test	Train data		Test data		Train + Test
Awake Count	Drowsy Count	Awake Count	Drowsy Count	Total count	Awake Count	Drowsy Count	Awake Count	Drowsy Count	Total count
64	60	13	19	156	56	53	21	26	156

%), and the sensitivity and accuracy slightly improved by1.5–3 % for RF, though the specificity dropped marginally (0.7 %). The results from the time series split cross validation indicates high variance among the different experiments using k (=10) folds. When the LOPO validation was done in a grouped approach leaving samples of each participant as a test group and sample from the rest of the participants as training group, the performance substantially varied from 10-Fold CV and leave one instance out (LOO) validation technique. In this approach using 22 features, the sensitivity, specificity and accuracy ranged from 66.4 to 70.3 %, 50.4–66.3 %, and 59.2–60.3 %, respectively for KNN, SVM and RF. When using the selected 13 features, the sensitivity, specificity and accuracy improved drastically, ranging from 68.1 to 70.3 %, 76.2–84.8 %, 67.3–80.1 %, respectively for KNN, SVM and RF.

Table 7
Distribution of the resulting labels across 10-fold, stratified 10-fold and time series (split=10) cross-validation schemes.

Fold	10-Fold CV					Stratified 10-fold CV					Time Series Split CV (split=10)				
	Train data		Test data		Train + Test	Train data		Test data		Train + Test	Train data		Test data		Train + Test
	Awake Count	Drowsy Count	Awake Count	Drowsy Count	Total count	Awake Count	Drowsy Count	Awake Count	Drowsy Count	Total count	Awake Count	Drowsy Count	Awake Count	Drowsy Count	Total count
1	69	71	8	8	156	69	71	8	8	156	8	8	7	7	30
2	65	75	10	6	156	69	71	8	8	156	19	11	6	8	44
3	69	71	8	8	156	69	71	8	8	156	25	19	8	6	58
4	67	73	12	4	156	69	71	8	8	156	33	25	7	7	72
5	66	74	11	5	156	69	71	8	8	156	46	26	5	9	86
6	74	66	6	10	156	69	71	8	8	156	51	35	8	6	100
7	69	72	8	7	156	69	72	8	7	156	52	48	6	8	114
8	69	72	8	7	156	70	71	7	8	156	60	54	7	7	128
9	70	71	7	8	156	70	71	7	8	156	70	58	8	6	142
10	75	66	6	9	156	70	71	7	8	156	75	67	7	7	156

3.2. Participant dependent vs participant independent validation

To visualize the results obtained using multiple classifiers and validation methods, the mean performance was computed from three classifiers. After taking the mean performance from each classifier, the 10-fold cross-validation and LOPO cross-validation performance (mean and standard error) was compared in Fig. 3. The plots show that the performance for both validation techniques improved when using the selected features, except for the sensitivity in 10-fold cross validation. Interestingly, the improved performance is greater in case of LOPO approach. It is important to note that while the performance was plotted individually for each classifier, they showed the similar trend.

3.3. Interpretation

For the interpretation of the developed random forest classifier, two methods were applied, namely SHAP analysis and partial dependency analysis. The two methods are described below.

3.3.1. SHAP analysis

SHAP based feature importance. A SHAP analysis was performed on the dataset having the included features to generate a feature rank (Fig. 4(a)). Physically, the feature rank in Fig. 4 provides a clear ranking of features based on the mean of absolute Shapley values per features corresponding to total samples from each class, and sorted in a descending which quantifies their impact on model predictions. From Fig. 4(b), the positive shapley value means that the corresponding feature value is pushing the model output to be higher than the average prediction (i.e., drowsy state) and vice versa for the negatives (i.e., awake state). The color coding (blue for lower magnitudes and red for higher magnitudes) in the figure visually represents the strength of the feature's influence on different samples.

From a scientific perspective, the analysis in Fig. 4(a) and (b) suggest that the EOG features, such as Blink Duration (BD) and Amplitude Velocity Ratio (AVR) are considered as most two important features by the shapely concept. Fig. 4(b) highlights that the longer BD and the higher the AVR, the greater the chance of that sample being classified as a drowsy state. Next to the EOG features, theta and alpha EEG signal power from left-central (C3) and left-occipital (O1) region are got the highest importance among EEG features, where higher theta and alpha band power increases the chance of the samples being classified as drowsy state.

SHAP dependence plot. To understand the contribution of each feature, form each data samples to the model prediction corresponding their magnitude, a SHAP dependence plot was performed for top ranked features from EOG signal (Fig. 5(a,b)). Physically, SHAP Dependence plots interprets how specific feature values impact the predicted probability of drowsiness. For example, in Fig. 5(a), it is physically evident that when Blink Duration (BD) is less than 0.15 s, the model consistently lowers the predicted probability of being classified as drowsy, but a BD of more than 0.15 s consistently increases the predicted probability of drowsiness. Similarly, the SHAP dependence plot of AVR in Fig. 5(b) shows that below the ratio of 0.075, the model classifies the samples as awake state while it increases the probability of being classified as drowsy in the AVR is greater than or equal to 0.075.

Fig. 5(c,d) provides a nuanced and scientifically meaningful insights into the interactions between EOG and EEG features and a detailed understanding of how various features interact and influence the model's predictions. It shows a visualization for EEG Theta (channel C3) and alpha (channel O1) band powers corresponding to the strongly interacted EOG features, i.e., BD and AVR. Fig. 5(c) illustrates that in the cases where the EEG theta C3 channel power is greater than 0.00023 μV^2/Hz, the presence of longer blink durations (red color dots) increases the chances of feeling drowsy. Conversely, for shorter blink durations (blue color dots) reduces the chances of the instances being classified as drowsy. Similarly, Fig. 5(d) shows that the EEG alpha O1 band power higher than 0.0002 μV^2/Hz increases the chances of drowsiness, with higher amplitude velocity ratio (red color).

3.3.2. Partial dependency analysis

Partial dependence and individual component expectation plots. To understand the marginal effect of the feature values on the predicted output of the classification model, both individual component expectation (ICE) and partial dependence plot (PDP) plots were performed together, which are shown in Fig. 6 and Fig. 7. Physically, both figures reveal how specific feature values influence the predicted probability of drowsiness. In both cases, the y-axis represents the predicted probability of a machine learning model and x-axis presents the magnitude of feature values. The thin separate curves show the dependency of the prediction on the feature (individual components dependence from each sample) and the thick curves represents the average effect of them (mean partial dependence.

From the ICE and PDP in Fig. 6(a), it can be observed that the longer blink duration increases the probability of drowsiness. When the blink duration is between 0.125–0.150 s, a noticeable transition occurs, and beyond 0.15 s, participants are consistently classified as drowsy. Similarly, in Fig. 6(b), a physically observable effect is that higher Amplitude Velocity Ratio (AVR) is linked to an increased probability of drowsiness. Notably, the transition in this case happens when AVR is within the range of 0.07–0.08, beyond which participants are consistently classified as drowsy.

Scientifically, Fig. 6(c,d) makes it easier to estimate a threshold (BD

Table 8
Classification performance for different validation technique, reported numbers: mean ± standard deviation; the highlighted numbers represent the best performance using the specific validation technique.

Classifier			K-Nearest Neighbours (KNN)			Support Vector Machines (SVM)			Random Forest (RF)		
Validation Techniques	Distribution A (KSS 2–6), D (KSS 7–9)	Count of Features	Sensitivity Mean ± SD (%)	Specificity Mean ± SD (%)	Accuracy Mean ± SD (%)	Sensitivity Mean ± SD (%)	Specificity Mean ± SD (%)	Accuracy Mean ± SD (%)	Sensitivity Mean ± SD (%)	Specificity Mean ± SD (%)	Accuracy Mean ± SD (%)
Holdout (Test data:20 %, RS =1)	Train A: 64, D:60 Test A: 13, D: 19	22 Features	78.9	61.5	71.8	63.2	76.9	68.7	78.9	76.9	78.1
		13 Features	78.9	84.6	81.2	68.4	84.6	75.0	78.9	84.6	81.3
Holdout (Test data:30 %, RS =1)	Train A: 56, D:53 Test A: 121, D: 26	22 Features	88.4	61.9	76.6	69.2	66.6	68.1	80.8	66.7	74.5
		13 Features	65.3	80.1	72.3	84.6	61.9	74.4	84.6	66.7	76.6
10-Fold CV	A: 77	22 Features	68.2 ± 32.9	56.1 ± 33.2	56.8 ± 15.2	76.4 ± 28.1	54.1 ± 27.3	65.0 ± 22.3	66.6 ± 16.9	63.9 ± 25.8	65.5 ± 11.1
	D: 79	13 Features	58.9 ± 24.3	70.2 ± 23.8	64.7 ± 13.7	60.5 ± 17.1	76.4 ± 16.6	68.7 ± 13.1	70.5 ± 29.4	67.7 ± 26.4	69.3 ± 19.8
Stratified K fold	A: 77	22 Features	69.3 ± 15.1	54.6 ± 21.6	62.2 ± 14.3	52.8 ± 20.3	63.6 ± 18.6	58.5 ± 14.5	66.6 ± 16.9	63.9 ± 25.8	65.5 ± 11.1
	D: 79	13 Features	58.9 ± 17.3	70.2 ± 22.4	64.7 ± 12.8	60.5 ± 15.9	76.4 ± 16.1	68.7 ± 09.1	70.5 ± 29.4	67.7 ± 26.4	69.3 ± 19.8
Time Series Split (k = 10)	A: 77	22 Features	49.3 ± 38.3	66.0 ± 28.5	57.8 ± 15.5	70.4 ± 30.6	43.1 ± 33.3	64.3 ± 20.7	38.6 ± 32.1	76.5 ± 36.7	57.1 ± 17.8
	D: 79	13 Features	24.3 ± 26.5	89.2 ± 18.2	61.4 ± 14.4	87.8 ± 17.8	67.7 ± 33.6	62.8 ± 21.1	41.6 ± 35.2	75.8 ± 36.9	58.6 ± 18.9
LOO CV	A: 77	22 Features	59.1 ± 28.8	53.9 ± 27.3	68.4 ± 15.5	57.2 ± 38.3	57.5 ± 28.9	72.4 ± 44.7	70.9 ± 15.4	68.8 ± 24.3	69.9 ± 22.3
	D: 79	13 Features	60.3 ± 29.3	54.9 ± 13.9	71.2 ± 45.3	62.1 ± 28.8	72.4 ± 24.1	77.5 ± 09.1	69.6 ± 13.9	72.7 ± 10.7	71.2 ± 11.2
LOPO CV	A: 77	22 Features	68.9 ± 25.3	52.4 ± 18.2	59.2 ± 16.2	66.4 ± 31.1	66.3 ± 26.4	65.8 ± 22.4	70.3 ± 29.3	50.4 ± 26.4	60.3 ± 20.1
	D: 79	13 Features	68.4 ± 37.6	76.2 ± 30.8	67.3 ± 36.9	68.1 ± 27.5	84.8 ± 16.5	73.3 ± 28.4	70.3 ± 29.4	82.2 ± 9.8	80.1 ± 8.2

Note: A: Awake; D: Drowsy;10-Fold CV: 10-Fold Cross-Validation; LOO CV: SD: Standard Deviation; Leave One Out Cross-Validation; LOPO CV: Leave One Participant Out Cross-Validation; RS: Random Split; k: number of splits; No standard deviation was reported in holdout validation technique because the data was spitted only once in this approach.

Fig. 3. Comparison of 10-Fold Cross-validation and Leave One Participant Out (LOPO) Cross-Validation methods using mean outcome of KNN, SVM and RF) classifier, 22 F: all the 22 features included, 13F: the selected features 13 included.

threshold=0.14 s and AVR threshold=0.078) when the machine learning classifier makes a decision based on BD and AVR. The combined dependence of BD and AVR has been presented in Fig. 6(e), which visualises the effect in a 3D plot. The similar analysis was presented for EEG-based features in Fig. 7(a)–(e), which physically show that higher EEG Theta and Alpha band powers are associated with an increased probability of drowsiness. For the theta band power (central), the transition happens after the band power reaches 0.00025 μV^2/Hz, suggesting a scientific threshold for the onset of drowsiness. Similarly, for the alpha band power (occipital), it triggers sleepiness even before it reaches the 0.00025 μV^2/Hz threshold, providing a scientific insight into the predictive nature of EEG alpha band power in occipital region.

Two-way feature interaction plot. A two-way interaction plot was performed to understand the dependency of predicted outcome on multiple top ranked features and interaction between themselves, i.e., EOG and EEG features. Physically, this plot identifies a specific threshold where the two features interact to affect drowsiness probability. From the 2-way interaction plot in Fig. 8(a), when BD exceeds 0.15 s, both BD and AVR jointly influence drowsiness levels; however, below this threshold, BD alone has a significant impact, while contribution of AVR to model predictions is minimal. This means BD can only trigger AVR when it surpasses the 0.15 second threshold. Similarly, the interaction between EOG and EEG features are presented in Fig. 8(b). It shows that below 0.15 secs BD, EEG Theta C3 power is independent of BD when making predictions, but after reaching the value of 0.15 secs, BD impacts theta EEG power substantially (central C3), which combinedly influences the probability of drowsiness. That means with longer blink duration, the sleepiness increases, at the same time it triggers the theta EEG power (central). From the figures, this physical interpretation is that the specific threshold initiates the interaction between these features, leading to changes in drowsiness probability. The scientific interpretation delves into the relationship between EOG and EEG features, demonstrating that certain feature interactions become significant only under specific conditions, potentially revealing deeper insights into the processes contributing to drowsiness.

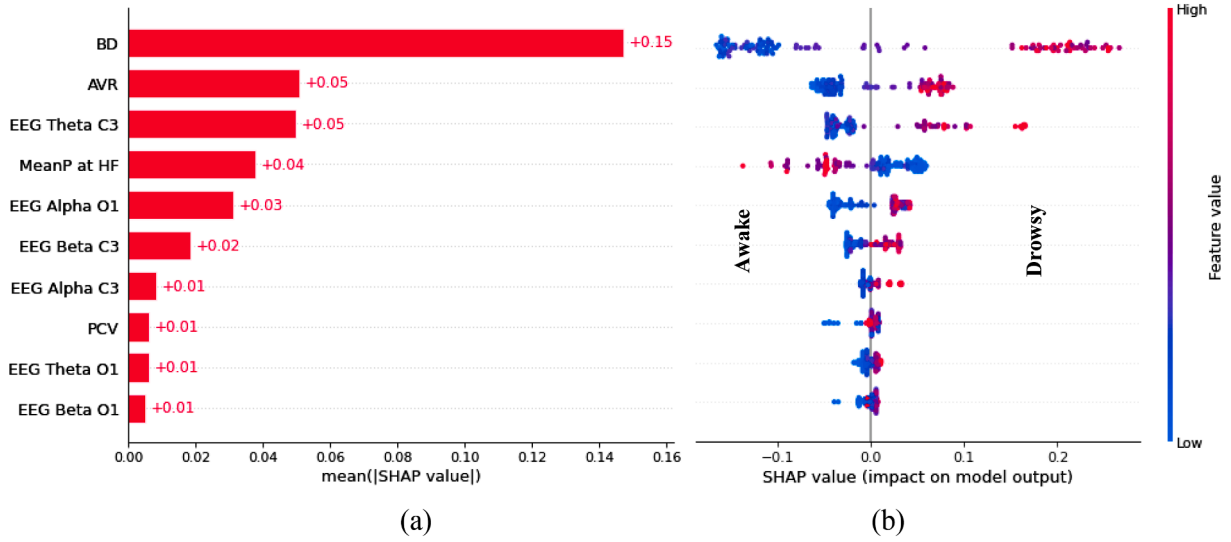


Fig. 4. SHAP based Feature ranking of top 10 features (a) SHAP feature ranking (b) SHAP summary plot. Note: BD: blink duration, EEG: electroencephalogram, C3: central channel, O1: Occipital channel. PCV: peak closing velocity, AVR: amplitude velocity ratio. The x-axis represents the shapley values while the y axis represents the included features ranking. Each blue dot corresponds to a lower magnitude of the feature for different samples, while the red dots indicate higher magnitudes of the features.

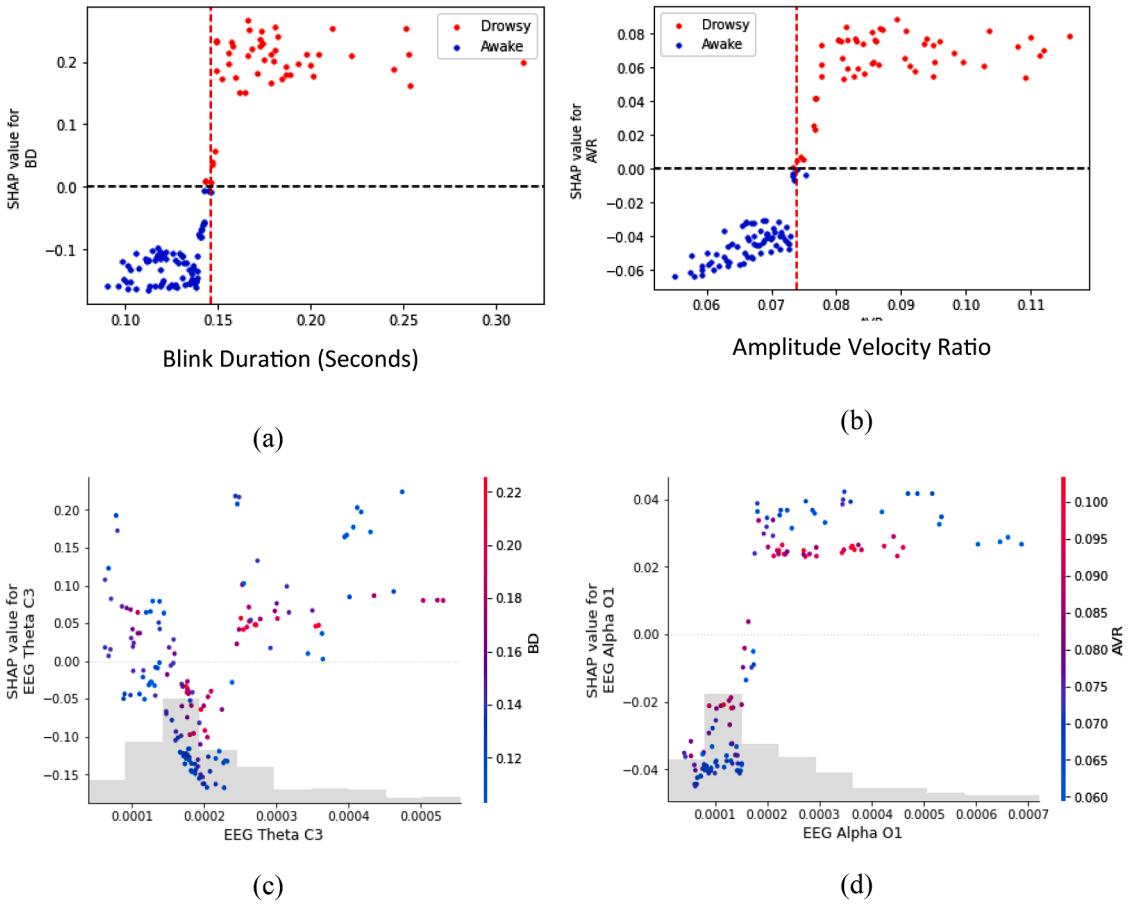


Fig. 5. SHAP feature dependence plot; cut-offs (a)BD: blink duration (0.146577 s), (b) AVR: amplitude velocity ratio (0.0739612), (c) EEG C3: electroencephalography theta central channel (0.0002386 $\mu V^2/Hz$), (d) EEG O1: electroencephalography alpha occipital channel (0.00016376 $\mu V^2/Hz$). The x-axis of the figure represents the of the feature value (magnitude) while the y-axis presents the corresponding shapley values.

Partial dependence-based feature importance. To compute average absolute partial dependence for feature ranking, we first calculated the partial dependence values for each feature. Then the average absolute partial dependence was computed by taking the mean of the absolute values of the partial dependence values. Considering the measure of the overall importance of the feature in the model, the partial dependence-

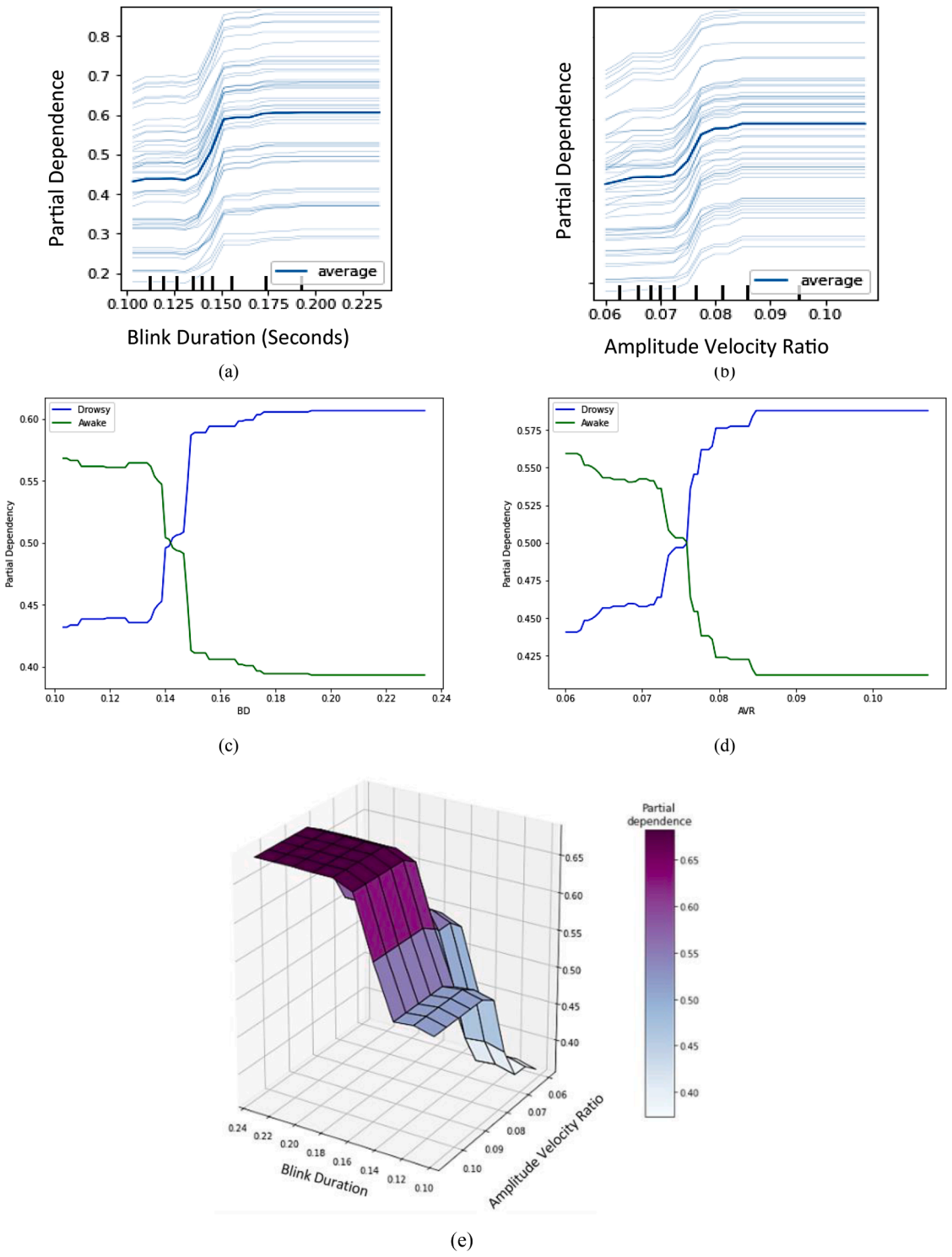


Fig. 6. (a) ICE and PDP for BD (blink duration) (b) ICE & PDP for AVR (amplitude velocity ratio) (c) decision boundary for BD (0.140086 s) (d) decision boundary for AVR (0.074356) (e) 3D plot; BD: blink duration, AVR: amplitude velocity ratio. The x-axis of the figure represents the of the feature value (magnitude) while the y-axis presents the partial dependence of the features.

based feature rank has been plotted to compare with SHAP-based feature ranking (Fig. 9). The partial dependence-based feature importance identified the blink duration, amplitude velocity ratio and EEG Theta power at C3 channel to be the top three features, which is also supports the SHAP-based feature ranking (Fig. 4).

3.3.3. Comparison of SHAP and partial dependence scores

To observe the relation between the feature ranking obtained using both explainable models, multiple statistical analysis was performed, and the results were plotted (Fig. 10). First, a correlation analysis was performed between the SHAP feature scores and mean partial dependence of the features. It resulted in a correlation coefficient of $r = 0.849$, which indicates a strong positive correlation between the SHAP scores and the partial dependence scores. This means that the two methods tend to produce similar rankings for the features. Second, the p-value was calculated between the obtained feature scores. It produces a p-

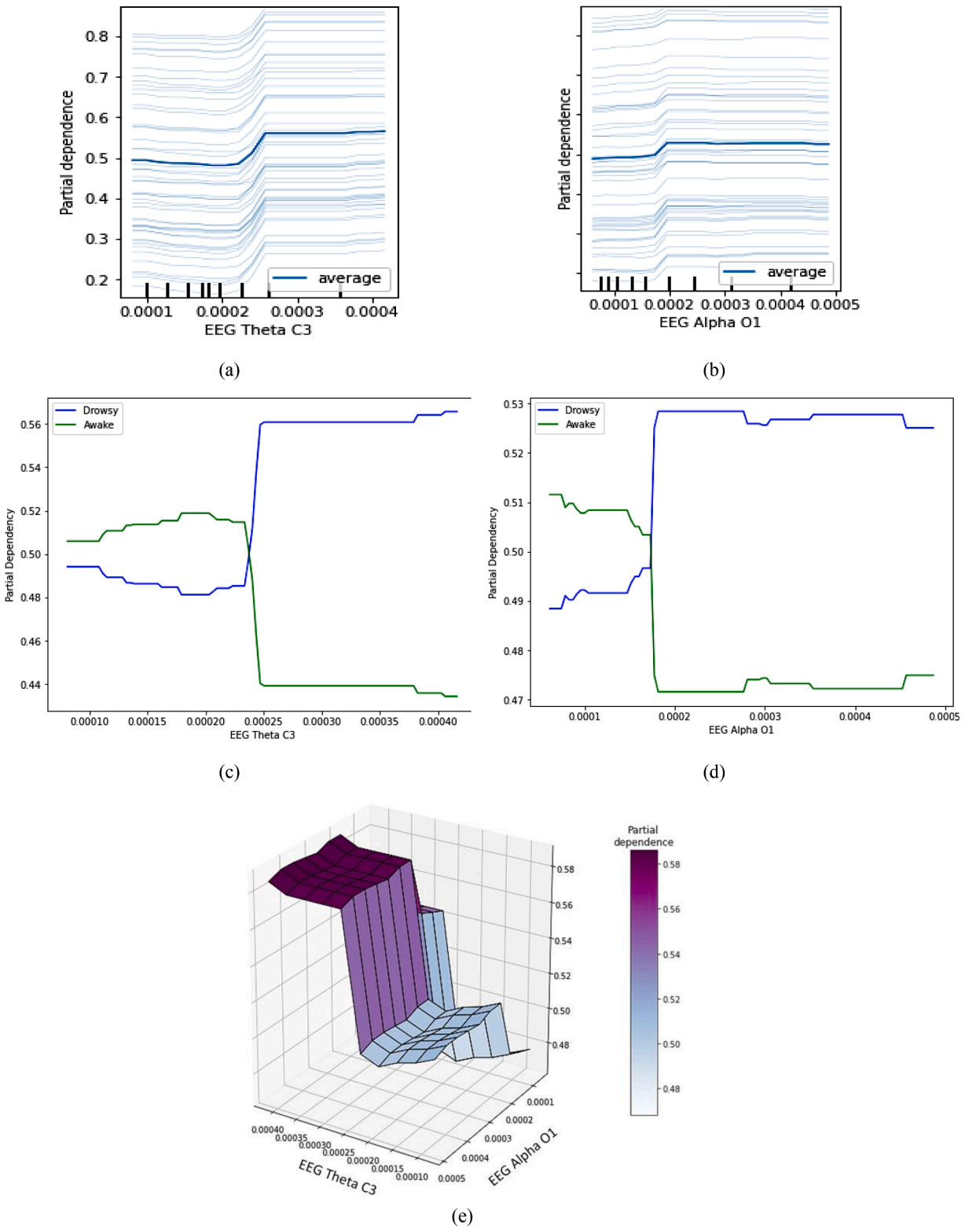


Fig. 7. (a) ICE and PDP for EEG theta C3 (b) ICE and PDP for EEG alpha O1 (c) decision boundary for EEG theta C3 (0.0002369256 μV^2/Hz) (d) decision boundary for EEG alpha O1(0.00016429 μV^2/Hz) (e) 3D plot; EEG: electroencephalography, C3: central channel, O1: occipital channel. The x-axis of the figure represents the of the feature value (magnitude) while the y-axis presents the partial dependence of the features.

value of 0.0019, which is relatively small (<0.05), indicating that the correlation observed between SHAP scores, and partial dependence scores is statistically significant. The small p-value suggests that it is unlikely to observe such a strong positive correlation by random chance. This implies that features that are important according to SHAP analysis also tend to exhibit strong partial dependence effects. It provides evidence that the relationship between SHAP and partial dependence is real and not due to random fluctuations. Third, a distribution plot was performed and both distribution plots for SHAP scores and partial dependence scores show a similar distribution, it generally means that the two

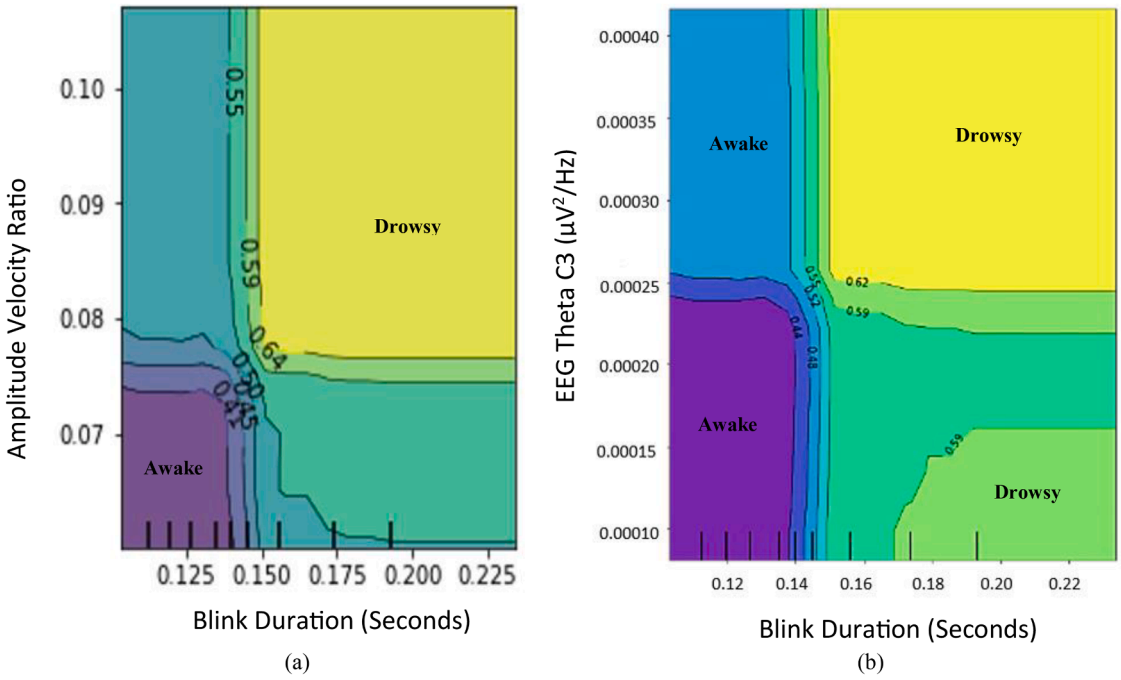


Fig. 8. Two-way Numerical PDP using random forest classifier; BD: blink duration, AVR: amplitude velocity ratio, EEG: electroencephalography, C3: central channel.

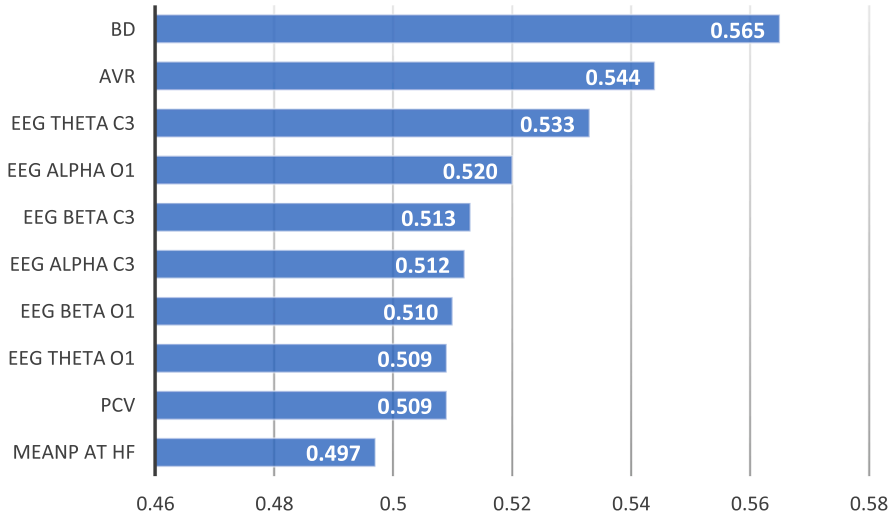


Fig. 9. Mean partial dependence-based feature ranking of top 10 features. *Note:* BD: blink duration, EEG: electroencephalogram, C3: central channel, O1: Occipital channel. PCV: peak closing velocity, AVR: amplitude velocity ratio.

methods are producing consistent and aligned results in terms of feature importance.

4. Discussion

4.1. Performance measure with different validation methods

The performance measures with the different validation techniques are described below in two sections: participant dependent and independent evaluations.

4.1.1. Participant dependent evaluation

Among the participant dependent validation methods, the results obtained from time series cross validation are surprisingly varied from the other (train-test split, 10-fold and stratified 10-fold approach). Especially, the sensitivity and accuracy values have decreased for KNN and RF classifier while using the time series cross validation; however, the mean specificity was increased and was in the range of 66.0–89.2 for KNN and 75.8–76.5 % for RF. The reason behind the lower performance in time series cross validation can be explained by the 'autocorrelation' issue. Drowsiness is a slowly progressing state, which has the consequence that neighbouring (in time) data point are highly correlated (autocorrelated). This means that when we assign data randomly to the train/test sets in train test split, k-fold and stratified k-fold techniques, neighbouring data might end up in both the training and in the test set, so we essentially have a leakage of the unseen test data in the training set. This leads to generalization issues when the developed classifier is used on a new dataset [42].

4.1.2. Participant independent evaluation

The LOPO validation used in this study was participant independent given that all the six trials of a specific participant were used to test the

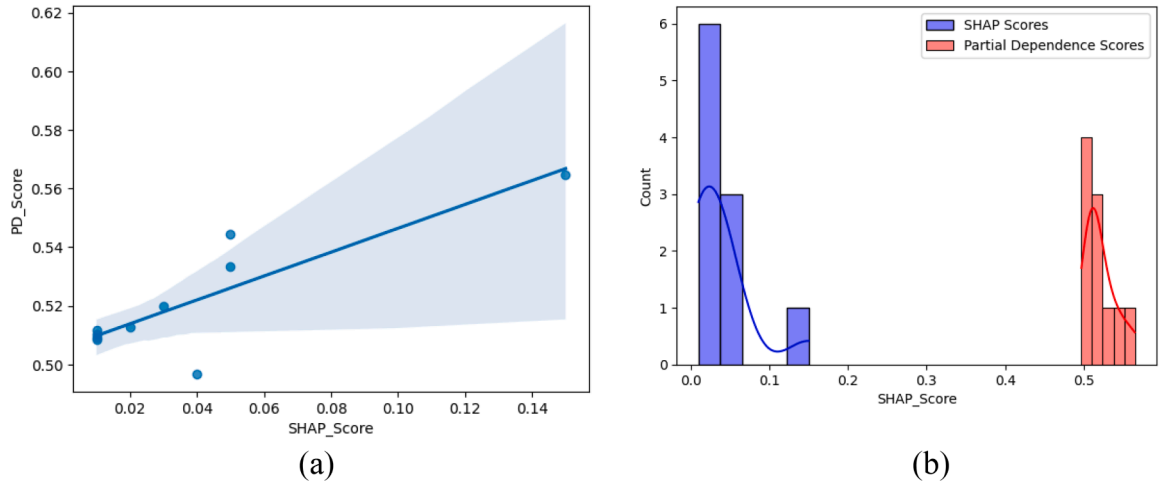


Fig. 10. (a) A scatter plot with a regression line to visualize the linear relationship between SHAP scores and partial dependence scores (PD_Score) to assess the strength and direction of the correlation. (b) A histograms or kernel density plots to visualize the distribution of SHAP scores and partial dependence scores to understand the spread and central tendency of these values.

model, other participants trials being the training data. The LOPO validation gave an accuracy of 80.1 % for RF, which was the second higher accuracy outcome, aside from the holdout validation, which was 81.3 %. As this validation is not biased towards the participants data (due to participants independency) and the test participants trials are unknown to the trained model, this validation gives the best generalised results in validation, for a participant independent evaluation. Most importantly, in the LOPO approach, the test data includes the trials of only one individual participants in each fold, thus the autocorrelation issue is also being addressed here. The obtained results are consistent with outcomes from previous drowsiness detection, using leave one participant out validation [43], although, other research has obtained much lower outcomes [13]. A factor in favor with leave one participant out is how estimates of bias are quite low with leave one out validation [44]. Thus, the utility of leave one participant out cross-validation appears to have been demonstrated with the current results.

An important consideration for the use of biosignals with drowsiness detection systems is the variations between and within individuals regarding their biosignals. For instance, EEG data is non-stationary [45] and variations between individuals also result in significant variations of the presentation with EEG biosignals [46]. The utility of LOPO has been noted as a way of beginning to address these individual differences in biosignals associated with sleepiness [47], differences which are likely to play a significant role in refraining the implementation of such systems in vehicles. Additional techniques could also improve the system's performance, for instance by adapting (re-training) the system for each individual driver [48]. Considering the combined outcomes noted above with the leave one participant out cross validation and the proposed benefits of additional training of a system, that is specified to individual drivers seemingly has the potential to improve the overall outcomes of detecting drowsiness.

4.2. Explainable results

The explainable results obtained from this study have proven the utility of some specific features, which contribute to the reliability of the system. The SHAP analysis and partial dependence analysis revealed that the features in the top of the priority list ranked by the SHAP values are strongly influence the prediction outcome but also triggers the other features for the decision-making process. For example, the blink duration feature from EOG signal is ranked highest based on the SHAP-values and theta EEG signal power at central (C3) channel is ranked the highest among EEG features. The two-way feature interaction plot showed that both features influence the prediction outcome, i.e., blink duration longer than 0.15 leads to drowsy state but also triggers the EEG theta C3 power for higher predicted probability of drowsiness. This signifies the strong correlation between the EOG and EEG signals, and the utility of using left-central (C3) channel electrode and theta band power. This implication may help reducing number of sensors, omitting other sensors corresponding the lower ranked features, such as heart rate and R-R interval from ECG.

It is important to note that most of the top features ranked by the SHAP values are consistent with the feature selection techniques mentioned, with some inconsistency in the lower ranked features [21]. This is because those univariate techniques are different from SHAP analysis, being independent of the machine learning based model, while the SHAP values are based on the predicted outcome from the machine learning model after the model is trained with the data.

4.3. Practical applications

The study findings have utility in real-world application, especially in developing a trustworthy drowsiness detection system in terms of validation and explainability. First, participant independent LOPO validation shows promise in developing such a trustworthy system, which is not biased towards the participants data due to participants independency. Although a system may produce very high performance with a participant dependent train-test split method, the underlying machine learning model may not be well-generalised and does not fulfill the criteria of being a trustworthy detection system. Second, reliability is a crucial part of a trustworthy machine learning based detection system. The explainable machine learning approach may allow the stakeholders to understand the influence of the important physiological based features (such as EOG and EEG features) in the decision-making process and add extra-reliability on the operational principle of the system. This will accelerate the proper implication of the research findings in industry applications. Furthermore, the most important features identified by the SHAP analysis help reducing number of sensors usable to the system and thus reduce system cost. The decision boundary in the partial dependency analysis shows a clearer feature cut-off, which might be utilised to develop a simple offline drowsiness detection system. The system could use the feature cut-off values for multiple important features and give decision on drowsy and awake stage. This will reduce the computational cost as well as the usable sensors. Overall, a proper implication of the study findings will offer more accurate, reliable and cost-effective trustworthy drowsiness detection system to be deployed

on-road.

4.4. Strengths, limitations and future works

Our study is centered on evaluating the 'explainability' of a multimodal physiological signal-based drowsiness detection system, induced through the Psychomotor Vigilance Task (PVT) and assessed using Karolinska Sleepiness Scale (KSS) scores. While the integration of multiple validation techniques enhances the robustness of our machine learning model, it also plays a pivotal role in establishing the model's trustworthiness [22]. Our primary emphasis on 'explainability' sets us apart from existing works in this field. Notably, we introduce 'multi--validation' to assess the system's 'trustworthiness'. To the best of our knowledge, comprehensive explainable analyses such as SHAP and Partial Dependence Analysis have not been applied to a multimodal physiological signal-based system that incorporates EEG, EOG, and ECG data to assess the marginal contribution of physiological features in response to the PVT task and well-validated KSS measures. Therefore, our study represents a significant contribution to the existing body of knowledge in this field.

The study limitations should be considered while interpreting the results. First, the Psychomotor Vigilance Task (PVT) was implemented in this study as the primary task to induce rapid drowsiness in a controlled laboratory setting, replacing the driving task. The shortfall of using the PVT is that an individual's KSS levels can decrease rapidly when performing the task for a relatively short time. On the other hand, in real-life situations, people tend to experience drowsiness more frequently after extended durations. As a result, relying solely on PVT may not provide an accurate depiction of the fluctuations and shifts in drowsiness that happen during real-world activities like driving. Second, the sample size used in this study is small, which is not enough to explore inter-individual differences or performing one trial per participants for examining utility of leave one trial out (LOO) validation. Future research could be performed using a greater number of participants, which could give better insight into the inter-individual differences and validation techniques.

Third, in this study, the most beneficial features identified in prior literature were selected for evaluation of performance measures from hybrid physiological data, using traditional supervised machine learning techniques. Typically, conventional learning algorithms are employed, wherein the process of extraction of relevant features and then those feature-specific classification are carried out as separate stages [49]. Nevertheless, in recent years, deep learning models have gained popularity in designing end-to-end systems, wherein the feature extraction process is automatic and does not require manual effort. As such, the deep learning architectures [49] are capable of extracting more resilient and abstract features, which may prove more beneficial in drowsiness detection. Future research could be performed to validate the deep architectures performance, and to interpret the results accordingly.

Fourth, the performance of the machine learning models used in the study produces a resonable metrices score with the given set of 13–22 features, with a highest sensitivity of 88.4 % (KNN), specificity of 84.6 % (KNN, SVM & RF) and accuracy of 81.3 % (RF) using participant dependent holdout validation, which further got reduced due to participant dependent cross-validation, i.e., 70.3 % sensitivity (RF), 84.8 % specificity (SVM) and 80.1 % accuracy (RF). While this performance is not comparable to other studies due to different settings, time window, stimulus to drowsiness, and sleepiness scores [16], our main focus was to assess the 'trustworthiness' rather than producing the 'best performance metrices'.

Fifth, we have included a limited number of features which reduces computational cost but also number of sensors in terms of real-world deployment. A recent systematic review and meta-analysis by Watling et al. [16] show that increasing the number of features does not necessarily enhance the performance drowsiness detection using physiological signals. In this experimental study, only two EEG channels were utilised, which produced 12 EEG features. While there are 32-channel EEG headsets are available in the market, and utilizing the extracted features could improve the system performance, it may increase the system cost and computational complexity. Also, using extensive number of features makes the model complex, which is difficult to interpret by the 'explainable' methods [50]; that is why we have stuck to a limited number of most useful features in our study. However, some more useful features with novel feature extraction methods [4] can be applied in future to improve robustness, and explainability, and trustworthiness of the system.

Last, interpretable machine learning techniques like SHAP and PDA have several limitations. They are often used with complex, black-box models and provide local, rather than global interpretability [41]. SHAP can be computationally expensive [39], and PDA assumes linear relationships, limiting their applicability to numeric features [41]. High-dimensional data with extensive features [50], subjectivity in interpretation, and data distribution sensitivity are additional challenges with these methods [39]. Despite their limitations, SHAP and PDA remain valuable for enhancing transparency in complex machine learning models with limited number of features, which helps assessing the trustworthiness of the system.

5. Conclusion

This study utilised multimodal physiological signals to detect drowsiness, assess the trustworthiness of such a system with multiple validation techniques and interpreted the results with explainable machine learning techniques. Among the validation techniques, the holdout and leave one participant out methods yield the most promising results. Especially, the leave one participant out validation method is advantageous as it provide participant-independent validation while also addressing the issue of autocorrelation. The interpretation of results shows the marginal effect of features on model prediction and influence between themselves. These findings indicate its usefulness of different validation and interpretation for a 'trustworthy' drowsiness detection system using biosignals and their applicability in different situations, based on the data structure and inter-individual differences.

CRediT authorship contribution statement

Md Mahmudul Hasan: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Software, Writing – original draft, Writing – review & editing. **Christopher N. Watling:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Grégoire S. Larue:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Declaration of Competing Interest

No conflicts of interest to report.

Acknowledgment

This research project was supported by the Australian Govt. Research Training Program (RTP) Scholarship and QUT Faculty Write Up Scholarship.

References

[1] Australian Transport Council, "National road safety strategy 2011-2020," 2011. [Online]. Available: https://www.roadsafety.gov.au/sites/default/files/2019-11/nrss_2011_2020.pdf.

[2] S. Taran, V. Bajaj, Drowsiness detection using adaptive hermite decomposition and extreme learning machine for electroencephalogram signals, IEEE Sens. J. 18 (21) (2018) 8855–8862, https://doi.org/10.1109/jsen.2018.2869775. Nov.

[3] S. Sharma, S.K. Khare, V. Bajaj, I.A. Ansari, Improving the separability of drowsiness and alert EEG signals using analytic form of wavelet transform, Appl. Acoust. 181 (2021), 108164.
[4] S.K. Khare, V. Bajaj, G. Sinha, Automatic drowsiness detection based on variational non-linear chirp mode decomposition using electroencephalogram signals, in: Modelling and Analysis of Active Biopotential Signals in Healthcare, 1, IOP Publishing Bristol, UK, 2020, pp. 5-1-5-25.
[5] H. Lee, J. Lee, M. Shin, Using wearable ECG/PPG sensors for driver drowsiness detection based on distinguishable pattern of recurrence plots, Electronics 8 (2) (2019), https://doi.org/10.3390/electronics8020192. BaselFebArt no. 192.
[6] M. Babaeian, M. Mozumdar, Driver drowsiness detection algorithms using electrocardiogram data analysis, in: Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference, 2019, pp. 1–6.
[7] K.T. Chui, K.F. Tsang, H.R. Chi, B.W.K. Ling, C.K. Wu, An accurate ECG-based transportation safety drowsiness detection scheme, IEEE Trans. Ind. Inf. 12 (4) (2016) 1438–1452, https://doi.org/10.1109/TII.2016.2573259.
[8] T. Kundinger, N. Sofra, A. Riener, Assessment of the potential of wrist-worn wearable sensors for driver drowsiness detection (in English), Sensors 20 (4) (2020), https://doi.org/10.3390/s20041029. Basel, SwitzerlandArticle.
[9] R. Tamanani, R. Muresan, A. Al-Dweik, Estimation of driver vigilance status using real-time facial expression and deep learning, IEEE Sens. Lett. 5 (5) (2021) 1–4.
[10] C. Bergmeir, J.M. Benítez, On the use of cross-validation for time series predictor evaluation, Inf. Sci. 191 (2012) 192–213. Ny.
[11] K.Le Rest, D. Pinaud, P. Monestiez, J. Chadoeuf, V. Bretagnolle, Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation, Glob. Ecol. Biogeogr. 23 (7) (2014) 811–820.
[12] S. Barua, M.U. Ahmed, C. Ahlström, S. Begum, Automatic driver sleepiness detection using EEG, EOG and contextual information, Expert Syst. Appl. 115 (2019) 121–135, https://doi.org/10.1016/j.eswa.2018.07.054.
[13] L.W. Ko, O. Komarov, W.K. Lai, W.G. Liang, T.P. Jung, Eyeblink recognition improves fatigue prediction from single-channel forehead EEG in a realistic sustained attention task, J. Neural Eng. 17 (3) (2020), 036015, https://doi.org/10.1088/1741-2552/ab909f, 2020/06/29.
[14] J. Chen, H. Wang, C. Hua, Electroencephalography based fatigue detection using a novel feature fusion and extreme learning machine, Cogn. Syst. Res. 52 (2018) 715–728, https://doi.org/10.1016/j.cogsys.2018.08.018.
[15] J. Min, P. Wang, J. Hu, Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system," (in eng), PLOS One 12 (12) (2017), e0188756, https://doi.org/10.1371/journal.pone.0188756.
[16] C.N. Watling, M.M. Hasan, G.S. Larue, Sensitivity and specificity of the driver sleepiness detection methods using physiological signals: a systematic review, Accid. Anal. Prev. 150 (2020), 105900.
[17] H. Yaacob, F. Hossain, S. Shari, S.K. Khare, C.P. Ooi, U.R. Acharya, Application of artificial intelligence techniques for brain-computer interface in mental fatigue detection: a systematic review (2011-2022), IEEE Access 11 (2023).
[18] S.K. Khare, S. March, P.D. Barua, V.M. Gadre, U.R. Acharya, Application of data fusion for automated detection of children with developmental and mental disorders: a systematic review of the last decade, Inf. Fusion 99 (2023) 101898.
[19] L. Oliveira, J.S. Cardoso, A. Lourenço, C. Ahlström, Driver drowsiness detection: a comparison between intrusive and non-intrusive signal acquisition methods, in: Proceedings of the 7th European Workshop on Visual Information Processing (EUVIP), 2018, pp. 1–6, https://doi.org/10.1109/EUVIP.2018.8611704, 26-28 Nov2018.
[20] J. Arnin et al., "Wireless-based portable EEG-EOG monitoring for real time drowsiness detection," (in eng), Conf Proc IEEE Eng Med Biol Soc, vol. 2013, pp. 4977–80, 2013, doi: 10.1109/embc.2013.6610665.
[21] M.M. Hasan, C.N. Watling, G.S. Larue, Physiological signal-based drowsiness detection using machine learning: singular and hybrid signal approaches, J. Saf. Res. 80 (2021) 215–225.
[22] R. Chatila, et al., Trustworthy AI, Reflect. Artif. Intell. Humanity 12600 (2021) 13–39.
[23] W.J. von Eschenbach, Transparency and the black box problem: why we do not trust AI, Philos. Technol. 34 (4) (2021) 1607–1622.
[24] S. Jeong, Y. Baek, S.H. Son, A hybrid V2X system for safety-critical applications in VANET, in: Proceedings of the IEEE 4th International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA), IEEE, 2016, pp. 13–18.
[25] T. Åkerstedt, M. Gillberg, Subjective and objective sleepiness in the active individual, Int. J. Neurosci. 52 (1–2) (1990) 29–37.
[26] T. Åkerstedt, A. Anund, J. Axelsson, G. Kecklund, Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function, J. Sleep. Res. 23 (3) (2014) 240–252, https://doi.org/10.1111/jsr.12158.
[27] Cleveland Medical Devices Inc, BioCapture User's Guide," pp. 1–44, 2006.
[28] C. Anderson, A.W.J. Wales, J.A. Horne, PVT lapses differ according to eyes open, closed, or looking away (in eng), Sleep 33 (2) (2010) 197–204, https://doi.org/10.1093/sleep/33.2.197.
[29] L. Arsintescu, K.H. Kato, P.F. Cravalho, N.H. Feick, L.S. Stone, E.E. Flynn-Evans, Validation of a touchscreen psychomotor vigilance task (in eng), Accid. Anal. Prev. 126 (2019) 173–176, https://doi.org/10.1016/j.aap.2017.11.041.
[30] M.A. Hassan, E.A. Mahmoud, A.H. Abdalla, A.M. Wedaa, A Comparison between windowing FIR filters for extracting the EEG components, J. Biosens. Bioelectron. 6 (4) (2015) 1–6.
[31] M. Melinda, I.K.A. Enriko, M. Furqan, M. Irhamsyah, Y. Yunidar, N. Basir, The effect of power spectral density on the electroencephalography of autistic children based on the welch periodogram method, JURNAL INFOTEL 15 (1) (2023) 111–120.
[32] M.J.S. Johns, The amplitude-velocity ratio of blinks: a new method for monitoring drowsiness, Sleep 26 (2003) no. SUPPL.
[33] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. BME-32 (3) (1985) 230–236.
[34] S. Das, Filters, wrappers and a boosting-based hybrid for feature selection, ICML 1 (2001) 74–81.
[35] J. Suto, S. Oniga, P.P. Sitar, Comparison of wrapper and filter feature selection algorithms on human activity recognition, in: Proceedings of the 6th International Conference on Computers Communications and Control (ICCCC), IEEE, 2016, pp. 124–129.
[36] M.M. Hasan, Biomedical Signal Based Drowsiness Detection Using Machine learning: Singular and Hybrid Signal Approaches, Queensland University of Technology, 2021.
[37] T. Åkerstedt, J. Connor, A. Gray, G. Kecklund, Predicting road crashes from a mathematical model of alertness regulation—The sleep/wake predictor, Accid. Anal. Prev. 40 (4) (2008) 1480–1485.
[38] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.
[39] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 1–10.
[40] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[41] Q. Zhao, T. Hastie, Causal interpretations of black-box models, J. Bus. Econ. Stat. 39 (1) (2021) 272–281.
[42] D.R. Roberts, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (8) (2017) 913–929.
[43] S. Barua, M.U. Ahmed, C. Ahlstrom, S. Begum, Automatic driver sleepiness detection using EEG, EOG and contextual information, Expert Syst. Appl. 115 (2019) 121–135, https://doi.org/10.1016/j.eswa.2018.07.054. Jan.
[44] A. Elisseeff, M. Pontil, Leave-one-out error and stability of learning algorithms with applications, NATO Sci. Ser. Sub III Comput. Syst. Sci. 190 (2003) 111–130.
[45] A.Y. Kaplan, A.A. Fingelkurts, A.A. Fingelkurts, S.V. Borisov, B.S. Darkhovsky, Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges, Signal Process. 85 (11) (2005) 2190–2212, https://doi.org/10.1016/j.sigpro.2005.07.010, 2005/11/01.
[46] H.P. Landolt, Genetic determination of sleep EEG profiles in healthy humans, Prog. Brain Res. 193 (2011) 51–61, https://doi.org/10.1016/B978-0-444-53839-0.00004-1.
[47] A. Kamrud, B. Borghetti, C. Schubert Kabban, The effects of individual differences, non-stationarity, and the importance of data partitioning decisions for training and testing of EEG cross-participant models, Sensors 21 (9) (2021) 3225 [Online]. Available, https://www.mdpi.com/1424-8220/21/9/3225.
[48] H. Martensson, O. Keelan, C. Ahlstrom, Driver sleepiness classification based on physiological data and driving performance from real road driving, IEEE Trans. Intell. Transp. Syst. 20 (2) (2019) 421–430, https://doi.org/10.1109/tits.2018.2814207. Feb.
[49] M. Hultman, I. Johansson, F. Lindqvist, C. Ahlström, Driver sleepiness detection with deep neural networks using electrophysiological data, Physiol. Meas. 42 (3) (2021), 034001.
[50] A. Zytek, I. Arnaldo, D. Liu, L. Berti-Equille, K. Veeramachaneni, The need for interpretable features: motivation and taxonomy, ACM SIGKDD Explor. Newsl. 24 (1) (2022) 1–13.