Costantini *et al.*
Journal of NeuroEngineering and Rehabilitation (2024) 21:215
https://doi.org/10.1186/s12984-024-01519-2

Journal of NeuroEngineering and Rehabilitation

RESEARCH

Open Access

Artificial intelligence tools for engagement prediction in neuromotor disorder patients during rehabilitation

Simone Costantini[1*†], Anna Falivene[2†], Mattia Chiappini[2], Giorgia Malerba[2], Carla Dei[2], Silvia Bellazzecca[2], Fabio A. Storm[2], Giuseppe Andreoni[2,3], Emilia Ambrosini[1] and Emilia Biffi[2]

Abstract

Background Robot-Assisted Gait Rehabilitation (RAGR) is an established clinical practice to encourage neuroplasticity in patients with neuromotor disorders. Nevertheless, tasks repetition imposed by robots may induce boredom, affecting clinical outcomes. Thus, quantitative assessment of engagement towards rehabilitation using physiological data and subjective evaluations is increasingly becoming vital.

This study aimed at methodologically exploring the performance of artificial intelligence (AI) algorithms applied to structured datasets made of heart rate variability (HRV) and electrodermal activity (EDA) features to predict the level of patient engagement during RAGR.

Methods The study recruited 46 subjects (38 underage, 10.3 ± 4.0 years old; 8 adults, 43.0 ± 19.0 years old) with neuromotor impairments, who underwent 15 to 20 RAGR sessions with Lokomat. During 2 or 3 of these sessions, ad hoc questionnaires were administered to both patients and therapists to investigate their perception of a patient's engagement state. Their outcomes were used to build two engagement classification targets: self-perceived and therapist-perceived, both composed of three levels: "Underchallenged", "Minimally Challenged", and "Challenged". Patient's HRV and EDA physiological signals were processed from raw data collected with the Empatica E4 wristband, and 33 features were extracted from the conditioned signals. Performance outcomes of five different AI classifiers were compared for both classification targets. Nested k-fold cross-validation was used to deal with model selection and optimization. Finally, the effects on classifiers performance of three dataset preparation techniques, such as unimodal or bimodal approach, feature reduction, and data augmentation, were also tested.

Results The study found that combining HRV and EDA features into a comprehensive dataset improved the synergistic representation of engagement compared to unimodal datasets. Additionally, feature reduction did not yield any advantages, while data augmentation consistently enhanced classifiers performance. Support Vector Machine and Extreme Gradient Boosting models were found to be the most effective architectures for predicting self-perceived engagement and therapist-perceived engagement, with a macro-averaged F1 score of 95.6% and 95.4%, respectively.

†Simone Costantini and Anna Falivene equally contributed to this work.

*Correspondence:
Simone Costantini
simone.costantini@polimi.it
Full list of author information is available at the end of the article

Conclusion The study displayed the effectiveness of psychophysiology-based AI models in predicting rehabilitation engagement, thus promoting their practical application for personalized care and improved clinical health outcomes.

Keywords Robot-Assisted Gait Rehabilitation, Engagement, Psychophysiological signals, Classification, K-Nearest Neighbors, Random forest, Extreme Gradient Boosting, Support Vector Machine, Neural Network

Introduction

Robot-assisted therapies are nowadays an established intervention, widely used for gait rehabilitation in neuromotor impaired subjects, such as patients with cerebral palsy, acquired brain injuries, and stroke [1, 2]. These therapies reduce the physical effort and time required by therapists, improve the reproducibility of physiological gait kinematics, and increase intensity, volume, and difficulty of task-oriented motor exercises compared to conventional treatments [3]. Repetitive, intensive, task-oriented, and quantifiable training is indeed an essential feature for a rehabilitation intervention to foster recovery and neuroplasticity [4, 5].

Nevertheless, the repetitive and routinized nature of robotic activity, along with potential fatigue or pain, can lead to boredom and reduced motivation of the patient towards the therapy itself [6, 7], affecting adherence and compliance to the rehabilitation programs. Therefore, interventions aimed at fostering engagement (i.e., increasing the level of attention and interest during therapy sessions) such as modulating task workload within the therapy may yield greater neuroplastic changes and functional outcomes as well as boosting motivation [8]. Indeed, motor learning theory suggests that learning rates are highest when task difficulty positively challenges and excites subjects [9, 10]: if a task is under-challenging it can be perceived as boring, while if it is too difficult it can be overly stressful. Moreover, when patients perceive challenges that match their skills, they may experience a state of flow (i.e., optimal experience), which is a state of consciousness characterized by deep concentration, positive affect, clear goals, perceived control, and autonomous motivation [11].

For this reason, fully understanding the psychophysiological state of patients allows adapting the task not only to the patients' motor performance but also to their emotional state and engagement, which is a key factor in further enhancing the success of the therapy [12].

Engagement in rehabilitation was initially defined as the patient's commitment in the rehabilitation interventions, demonstrated through active and focused participation [13]. In this context, engagement was further explained by King and collaborators as a complex multifaceted state of investment in the therapy, which comprises three separate components: affective (i.e., emotional involvement in the therapy, motivation and optimistic expectations about the final outcome), cognitive (i.e., beliefs about usefulness and efficacy of the therapy) and behavioral (i.e., active participation and collaboration during sessions) involvement [14]. Moreover, Bright and co-authors further described engagement as being present in rehabilitative activity with willingness and emotional interest. They defined the attainment of this state as a continuum ranging from merely tolerating and agreeing to the treatment to being involved in the therapy and actively collaborating [15].

Assessing patient engagement during rehabilitation sessions is crucial to optimizing treatment outcomes [16], especially in the developmental age. However, this process is not straightforward, as it can require a combination of subjective and objective methods. In the last decades, several qualitative or semi-quantitative measures of engagement have been developed, such as the Self-Assessment Manikin (SAM) [17]; the PRIME-SP [18], compiled by service providers; and the Pediatric Motivation Scale [19].

As for objective measurements of engagement, various features of heart rate variability (HRV) and electrodermal activity (EDA) have been shown to be informative for the subject's psychophysiological states [20–27]. Namely, HRV refers to oscillations between consecutive cardiac cycles [23, 26]. Heart rate and rhythm are largely under the control of the autonomic nervous system (ANS), therefore a variation in the oscillation of consecutive heart beats can reflect a higher parasympathetic or sympathetic influence on heart rate. Specifically, a reduced HRV and inhibited parasympathetic activity were reported during high cognitive workload [28], and due to changes in affective states [29]. The EDA signal is a measure of skin conductance as a reaction of sweat secretion associated with the sympathetic nervous system activity, whose arousal associated with emotion, cognition, and attention is reflected in changes in the EDA signal [30].

Several studies also analyzed the correlation between subjective and objective assessment outcomes. The SAM questionnaire was administered to groups of subjects (i.e., adults, pediatric patients, healthy participants) performing cognitive tasks, while recording physiological data, such as electrocardiogram (ECG), respiration, EDA and skin temperature to assess mental

engagement during robotic rehabilitation treatments, finding significant correlation mostly between EDA features and SAM items [22, 24, 31].

In recent years, artificial intelligence (AI) has become a widely used methodological approach for the investigation and prediction of psychophysiological states [32]. In the field of Robot-Assisted Gait Rehabilitation (RAGR), Koenig and colleagues [33] predicted mental engagement in stroke patients during gait rehabilitation. They performed a 4-class classification task using physiological data such as heart rate, breathing frequency, skin conductance and skin temperature as input vectors for a Kalman adaptive linear discriminant analysis classifier. The classification results showed an accuracy of 70%. Bhat and colleagues [34], instead, predicted cognitive load using electroencephalographic (EEG) and EDA signals collected during a virtual reality training task in healthy subjects. Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), and Artificial Neural Network (ANN) models were trained to discriminate between low and high cognitive load. The XGB classifier achieved the best performance with a F1 score of 83%. Similarly, Gogna and collaborators [35] performed cognitive workload prediction in healthy subjects during increasing levels of difficulty of cognitive tasks, employing the EEG signal as input of Discriminant Analysis, SVM, KNN, and ANN classifiers. SVM showed the best performance with an accuracy of 90%.

Although not specifically trained with data coming from a rehabilitative environment, other studies implemented interesting predictive models of affect and cognitive workload with the intention to apply them for rehabilitation purposes. For instance, Bailenson et al. trained an ANN model with videotapes of subjects' faces together with ECG, EDA and somatic activity-related signals to automatically predict emotions of sadness and amusement [36]. F1 scores of 41% and 25% were obtained for the prediction of amusement and sadness, respectively, when using only physiological data as input of the classifier, whereas the combination of facial and physiological signals yielded F1 scores of 66% and 37%. In Gümüslü et al., an XGB model was developed for emotion recognition applications in robot-assisted rehabilitation. Specifically, EEG activity, blood volume pulse (BVP), skin temperature and EDA were used to classify pleasant, unpleasant and neutral emotions. An accuracy of 93.70% was achieved when using only physiological signals [37]. Finally, Romaniszyn-Kania and colleagues performed a 12-class emotion prediction using features from physiological data, such as BVP, EDA, acceleration signals, and from a modified version of the Job-Related Affective Well-Being Scale in healthy subjects during physiotherapy exercises of varying difficulty. KNN classifier was trained and achieved an accuracy of 81.63% [38].

The aforementioned studies primarily focused on assessing or predicting affective or cognitive states for rehabilitation purposes. However, provided that these are separate but concurrent aspects of engagement [14], a comprehensive estimation of the engagement state, intended as a multidimensional construct during rehabilitative tasks was never performed. Furthermore, the literature lacks studies performing prediction of engagement in pediatric patients with neuromotor impairments during robot-assisted rehabilitation. Therefore, the main goal of the present work was to compare the performance of various artificial intelligence and machine learning architectures in predicting the engagement of patients (i.e., children and adults) with neurological disorders during RAGR. Also, a secondary goal was to evaluate the effect of various pre-processing techniques on the model performance. The objective is threefold: to analyze the impact of using unimodal data (HRV or EDA features separately) and bimodal data as inputs to different classifiers; to examine the effect of feature reduction; and to investigate the effect of data augmentation on models performance.

Materials and methods

Study design

Participants

The present study involved 46 subjects (mean age 15.5 ± 14.2; 30 males; 38 underage, mean age 10.3 ± 4.0; 8 adults, mean age 43.0 ± 19.0) who had cerebral palsy, acquired brain injury or hereditary spastic paraplegia. Demographic details regarding each participant are reported in Supplementary Table 1. This study was performed in accordance with the Declaration of Helsinki, and the Ethics Committee of Scientific Institute E. Medea approved the observational study protocol (protocol code: Prot. N. 02/22-CE; date of approval: January 27th, 2022). Patients, if adults, or their guardians signed a written informed consent. All data were pseudonymized.

Protocol

Each patient underwent 15 to 20 robot-based gait rehabilitation sessions with the Lokomat system (Hocoma AG, Volketswil, Switzerland) at IRCCS Eugenio Medea, according to the clinical plan. Data acquisition was performed only during 2 or 3 sessions distributed during the rehabilitation period not to reduce the acceptability of the therapy. Each experimental session was composed of three phases:

- Phase 1: Patient welcome and questionnaires completion.

- Phase 2: Sensors setup and rehabilitation activity with the Lokomat system.
- Phase 3: Questionnaires completion and reference signals acquisition.

During Phase 1 and Phase 3, patients were asked to fill in a 9-items three-point Likert scale questionnaire (Table 1) about their feelings and expectations on the rehabilitation activity, to which possible answers were: "not at all", "enough", "very much". Items were adapted from the pedsQL questionnaire [39], the CORE system trust [40], and the UTAUT model determinants [41]. In addition, only during Phase 3, patients were also asked to provide their thoughts about the rehabilitation activity through open-ended questions presented in four speech bubbles, coherently with Phelan et al. [42]. Hereafter, the 9-items questionnaire and open-ended questions will be referred to as self-reported outcomes.

Conversely, during Phase 2, therapists were asked to complete a questionnaire (hereafter: therapist-reported outcome), divided into 12 items, to collect a reliable evaluation about the patient's state and behaviors during the rehabilitation training (Table 2). Items selected were partially among those of the Movement Assessment Battery for Children [43], taking into consideration the section describing non-motor items that can influence the movement. Each item was scored on a 13-point semantic differential numerical scale from -6 to 6. An example of the scoring is given for the item T1, where with positive values the therapists considered the positive pole (e.g., Active) as the most suitable for describing patient's state and behaviors, while choosing a negative score, the negative pole (e.g., Passive) was preferred. A score of 0 indicated the "neutral" or "undecided" category.

Both self-reported and therapist-reported outcomes were designed to globally investigate the psychological state of patients during RAGR according to the patients and therapists perspectives, respectively.

During Phase 2 and Phase 3, patients wore the Empatica E4 wearable device (Empatica®, Milan, Italy) on the non-dominant wrist. Empatica E4 is a Class IIa medical device including a photoplethysmogram sensor that measures the BVP signal (sampling frequency: 64 Hz), two silver-coated electrodes that apply a small alternating current to the skin in order to measure the EDA signal (sampling frequency: 4 Hz) and a MEMS type 3-axis accelerometer, to capture motion-based activity (sampling frequency: 32 Hz). Data collected during Phase 3 were exploited as a reference signal, assuring that patients were relaxed and in a resting condition. For the purpose of further analyses, we treated each experimental session performed by each patient as independent.

Table 1 List of items related to the three-points Likert scale self-reported outcome

Item ID	Item definition
S1	Do you feel worried?
S2	Do you feel happy?
S3	Do you feel sad?
S4	Do you feel angry?
S5	Do you feel scared?
S6	Do you feel bored?
S7	Do you think that the therapy with the Lokomat is useful?
S8	Are you able to handle the therapy with the Lokomat?
S9	Do you think that the therapy with Lokomat is effective with respect to gait improvement?

Table 2 List of items related to the therapist-reported outcome

Item ID	Negative Pole	Positive Pole
T1	Passive	Active
T2	Fearful	Assertive
T3	Anxious	Relaxed
T4	Impulsive	Thoughtful
T5	Distracted	Focused
T6	Hyperactive	Quiet
T7	Underestimates his/her abilities	Overestimates his/her abilities
T8	Not persistent	Persistent
T9	Concerned about failure	Not concerned about failure
T10	Unable to derive satisfaction from success	Able to derive satisfaction from success
T11	Manages emotions in a negative manner	Manages emotions in a positive manner
T12	Does not actively seek information to learn	Does actively seek information to learn

Classification targets

The engagement level of patients undergoing RAGR was determined using semi-quantitative data collected through self-reported and therapist-reported outcomes, resulting in the identification of two classification targets: self-perceived and therapist-perceived engagement. To better understand the process of transitioning that characterize the engagement, as defined by Bright et al. [15], from a state of non-involvement in therapy to a full commitment, a three-class classification problem was defined for both targets: each session was assigned a label corresponding to "Underchallenged", "Minimally Challenged" or "Challenged", according to the perceived engagement level.

To label each patient's level of engagement according to both classification targets, expert reviews evaluation methods were exploited [44, 45]. The inter-rater agreement was investigated for both expert review procedure outcomes by means of the Krippendorf α coefficient [46] computed with the code developed by Eggink et al. [47]. Values of α higher than or equal to 80% were considered representative of high agreement. Finally, the difference in engagement perception between patients and therapists was explored by constructing a contingency table.

Self-perceived engagement

For the self-reported outcomes, four independent raters (i.e., biomedical engineers with 1- to 5-years experience in affective computing and psychophysiological signal processing) provided a personal judgment on the level of the self-perceived engagement (i.e., assigning one of the three engagement classes). Each rater analyzed the responses of the 9-items of Likert scale questionnaires of both Phase 1 and 3, thus also accounting for potential differences between the two phases, and the responses of speech bubbles-questionnaire of Phase 3. An external psychologist with a 4-years experience in the field of human factors examined these evaluations to provide the final decision.

Therapist-perceived engagement

The therapist-reported outcomes were processed in order to facilitate the subsequent labelling procedure: each item was split into two separate sub-items (e.g., $T1^{(-)}$ and $T1^{(+)}$ for the T1 item), for a total of 24 sub-items. If an item Ti had received a negative score, the absolute value of Ti was assigned to $Ti^{(-)}$, while a score of 0 was added to the corresponding $Ti^{(+)}$ sub-item. Conversely, for positive Ti scores, $Ti^{(+)}$ was set to the absolute value of Ti, and $Ti^{(-)}$ was set to 0. In case of a neutral score for Ti, a 0 was added to both sub-items.

Then, two raters (i.e., biomedical engineers with 1-year experience in affective computing and psychophysiological signal processing) independently linked each sub-item to only one of the three engagement classes according to each sub-item's ability to describe the meaning of the linked class. Each rater could also exclude any sub-items that in his/her opinion matched neither of the three levels. The same external psychologist who proposed the final evaluation for the self-perceived engagement classes cleared up any disagreement between the two raters and provided the final decision.

Following the expert review evaluation, the average value of the sub-items scores of each generated class was calculated for every experimental session. The session was then labelled according to the class with the highest average value, with sessions in which the average value was the same for multiple classes being excluded from further analysis.

In this context, the expert review evaluation method was employed to group the sub-items rather than to directly assign labels to each session. This approach was deemed more appropriate, assuming the difficulty of reaching agreement among raters in condensing 12 items on a 13-point scale into one single engagement assessment.

Signal processing and feature extraction

The whole processing of physiological data was carried out in MATLAB (R2022b, The MathWorks Inc., Natick, MA, USA).

Due to the high variability of the duration of acquisition phases among subjects, which can be attributed to different clinical needs, a maximum number of five 5-min windows of Phase 2 signals, with a minimum time distance of 2.5 min, were manually selected for each session according to the raw BVP and EDA signals quality, while one 2.5-min window was kept for the reference signal due to the short duration of Phase 3. The 5-min window length was consistent with literature findings and with features reliability [23, 48]. Moreover, a 2.5-min minimum distance between the selected windows was chosen to guarantee no redundancy for the final dataset during AI models training.

Each 2.5- and 5-min BVP and EDA window was processed according to the methodology explained in Costantini et al. [49]. A third-order Butterworth bandpass filter with subject-specific cut-off frequencies (i.e., defined by subtracting the acceleration spectrum to the raw BVP spectrum) was applied to BVP to deal with motion artifacts. The time-points of the BVP diastolic valleys were then detected from the conditioned BVP signal, and the HRV signal was computed as the sequence of temporal distances between consecutive valleys. The HRV signal was first cleaned of potential missing and extra beats through a dedicated pipeline for HRV

artifacts detection and correction [50], and finally was resampled at 4 Hz by means of piecewise-cubic spline interpolation. As for the raw EDA signal, it was first conditioned with a 1-s windows moving average filter, then a z-score normalization was performed, and lastly the cvxEDA algorithm [51] was used to decompose EDA into its tonic and phasic components. An artifact detection algorithm, based on the stationary Haar wavelet transform and reported in detail in [49], was finally applied to the raw EDA signal to detect and correct potential motion-related spikes that could have been mistaken for phasic responses.

From each Phase 2 and reference window, 14 HRV and 19 EDA features were extracted, in line with [23, 52–54]. Tables 3 and 4 report the HRV and EDA features, respectively.

To address inter-subject variability, normalization was performed by subtracting the reference features to the ones related to Phase 2. Features were then independently rescaled with z-score standardization. Definitively, the dataset for self-perceived and therapist-perceived engagement prediction was composed of 14 HRV and 19 EDA normalized and rescaled features related to Phase 2 of the study protocol.

Dataset preparation

To investigate the effects of various preprocessing techniques on the original dataset, three stages of dataset preparation were considered, each testing various approaches and their influence on the AI models performance separately. In line with [32], the three stages, applied to both classification targets, consisted of:

1. ANS modeling.
2. Feature reduction.
3. Data augmentation.

Thus, only the approach that, according to the median value, performed better across the AI models in each stage was transferred to the next one.

The first stage aimed at examining the AI models performance when trained with three different datasets originating from three possible ways to model the autonomic nervous system activity: two unimodal datasets, composed of either 14 HRV features or 19 EDA features, respectively, and one bimodal dataset (BD), defined by merging the HRV and EDA features. The rationale behind this preliminary assessment was to investigate whether combining HRV and EDA features could result in synergistic effects, leading to improved engagement prediction.

The second stage explored whether and how AI models performance could have benefited from feature reduction. According to [32], there are two major approaches to perform feature reduction, namely features space projection onto a lower dimensionality space, and sequential selection of individual features taking inter-features correlations into account. However, the sequential features selection approach was finally discarded since our purpose at this stage was to train each AI model for both classification targets on the same number and type of

Table 3 List of HRV features

Feature	Description
Time-domain features	
Mean HR	Mean Heart Rate frequency
SDNN	Standard Deviation of all inter-beat-intervals
RMSSD	Root Mean Square of the Successive Differences
SDSD	Standard Deviation of Successive Differences between adjacent NN
pNN50	Count of delta NN exceeding 50 ms divided by the total number of all NN
HRV Triangular index	Total number of all NN divided by the height of the histogram of all NN
TINN	Baseline width of the triangular interpolation of the highest peak of the NN histogram
HRV Skewness	Skewness of the NN
HRV Kurtosis	Kurtosis of the NN
Frequency-domain features	
HRV nLF	Normalized power in Low Frequency range [0.04–0.15] Hz
HRV nHF	Normalized power in High Frequency range [0.15–0.40] Hz
Sympathetic modulation index	LF/(Total Power–VLF)
Vagal modulation index	HF/(Total Power–VLF)
Symphatovagal balance index	LF/HF

NN Inter-Beat-Intervals, *TINN* Triangular Interpolation of NN, *LF* Low Frequency, *HF* High Frequency, *VLF* Very Low Frequency (range [0.-0.04] Hz)

Table 4 List of EDA features

Feature	Description
Time-domain tonic component features	
Mean EDA Tonic	Mean value of the tonic component
St.Dev. EDA Tonic	Standard deviation of the tonic component
IQR EDA Tonic	Interquartile range of the tonic component
Skewness EDA Tonic	Skewness of the tonic component
Kurtosis EDA Tonic	Kurtosis of the tonic component
Max Upspeed EDA Tonic	Maximum positive slope of a regression line fitted on the tonic component
Max Downspeed EDA Tonic	Maximum negative slope of a regression line fitted on the tonic component
Time-domain phasic component measures	
NS.EDRs	Frequency of Non-Specific phasic peaks
Mean EDA Phasic Peak Amplitude	Mean of the amplitude of all NS.EDRs in the interval
St.Dev. EDA Phasic Peak Amplitude	Standard deviation of the amplitude of all NS.EDRs in the interval
nAUC EDA Phasic	Mean normalized area under the curve of phasic peaks
Mean Rise Time	Mean temporal distance onset-peak
Mean EDA Phasic P-to-P distance	Mean distance phasic peak-to-peak
St.Dev. EDA Phasic P-to-P distance	Standard deviation phasic distance peak-to-peak
Frequency-domain features	
EDA Phasic nVLF	Normalized power in Very Low Frequency range [0.–0.045] Hz
EDA Phasic nLF	Normalized power in Low Frequency range [0.045–0.15] Hz
EDA Phasic nHF1	Normalized power in High Frequency range [0.15–0.25] Hz
EDA Phasic nHF2	Normalized power in High Frequency range [0.25–0.40] Hz
EDA Phasic nVHF	Normalized power in Very High Frequency range [0.4–0.5] Hz

features. Therefore, the following two approaches for feature reduction were implemented:

- Literature-based feature reduction: a literature analysis was performed to identify the most predictive and significant HRV and EDA features according to previous findings.
- Projection-based feature reduction: the Principal Components Analysis (PCA) technique was used to project the original features onto a space of uncorrelated features [55]. Principal Components (PC) were ranked in decreasing order of explained variance. Cumulative Explained Variance (CEV) was computed sequentially for each PC as the cumulative sum of the explained variances up to the k^{th} PC. Three distinct levels of CEV thresholds, namely 80%, 90%, and 95%, were used to reduce the feature space up to the first k PCs.

The third stage assessed the impact of data augmentation (DA) on AI models performance. To progressively increase dataset size, each 5-min window of HRV and EDA signals was segmented according to one of the following methods:

- Two 4-min windows with three minutes overlap.
- Three 3-min windows, with two minutes overlap.
- Four 2-min windows with one minute overlap.
- Five 1-min windows with no overlap.

The minimum window length was chosen to be consistent with the minimum reasonable lengths for HRV and EDA. For HRV, we focused on short-term HRV, neglecting ultra-low and very-low oscillations. Several works, including [23, 56], have shown that 2- to 5-min windows are sufficient for capturing low and high-frequency variations. However, considering our data primarily comes from pediatric subjects with higher heart rates, a 1-min window can still provide an adequate number of heartbeats for reliable analysis. Concerning the EDA signal, Stržinar et al. [48] stated that while a 2-min window is optimal, 1-min windows are also sufficient for gathering significant information.

Classifiers

Five different classification algorithms, namely KNN [57], RF [58], XGB [59], SVM [60], and Feed-Forward Neural Network (FFNN) [61] were selected based on previous related works reported in [32]. Each algorithm was implemented in Python (Python Software Foundation, version 3.10), using keras [62] and scikit-learn [63]

libraries. A quick overview of classification principles and structures is presented below.

The KNN algorithm classifies a data point based on the majority class of its k-nearest neighbors in the feature space according to distance metrics. RF trains multiple decision tree classifiers on subsets of the original dataset, then performs classification by leveraging on a majority voting system. XGB is an ensemble learning algorithm that uses several decision trees by combining their classification outputs in a weighted sum. SVM aims at finding a hyperplane in the feature space that best separates data points of different classes by maximizing the margin between classes, while penalizing misclassifications. FFNN for classification is a computational model that consists of interconnected nodes, organized in layers, which learns from data by adjusting weights during training according to the backpropagation technique (i.e., minimizing the sparse categorical cross-entropy loss function). It is made of an input layer, at least one hidden layer and an output layer. Except for the former, each layer has a specific activation function that allows for a non-linear integration of the amount of information that each neuron receives in input. In the present work, the output layer was composed of three neurons and equipped with the softmax activation function, while the number of hidden layers, their activation function, the number of neurons per each layer, and the L2 regularization term were set as hyperparameters to be optimized (Table 5). To face overfitting, a dropout layer, with a dropout rate of 0.05, was implemented upstream the output layer. Each FFNN was trained for a maximum of 200 epochs, with mini-batch size at 32, initial learning rate at 0.0001, Adam algorithm for sparse categorical cross-entropy minimization, and an early stopping callback with patience 10 to prevent over-fitting.

Model selection

The nested cross-validation technique [64] was used to avoid the introduction of a model selection bias error [65], to compare classifiers and approaches during the dataset preparation pipeline, and to select the best classifiers at the end of the pipeline. Specifically, this technique involved nesting a k-fold cross-validation step for hyperparameter optimization inside a stratified k-fold cross-validation procedure for model selection, thus leveraging on record-wise classification. In the present work, $k=7$ and $k=6$ were chosen for the outer and inner cross-validation loops, respectively. Regarding the hyperparameters tuning procedure, the grid search technique was used to investigate the hyperparameters space: Table 5 resumes the tuned hyperparameters and their range for each selected classifier.

The macro averaged F1 score (hereafter: F1 Macro) was used to compare the models performance and to deal with the heavily unbalanced distribution of labels in both classification targets (see Results: Classification targets).

Table 5 Tunable classifiers hyperparameters within instances of k-fold cross-validation for hyperparameters optimization

Classifier	Hyperparameter	Definition	Values
KNN	K	Number of neighbors to consider for the prediction	From 5 to 40, with step 1
	Weights	Specifies the weight function used in prediction	"uniform" or "distance"
	P	Power parameter for the Minkowski distance	1 or 2
RF	n_estimators	Number of trees in the forest	100 or 200
	max_depth	Maximum number of nodes for each tree	From 3 to 15, with step 1
	Criterion	Function to measure the quality of a split	"gini" or "entropy"
XGB	learning_rate	Controls how the step size shrinks during the tree building process	0.01, 0.02, 0.05, 0.1, 0.2, or 0.5
	max_depth	Maximum number of nodes for each tree	From 5 to 10, with step 1
	n_estimators	Number of trees in the forest	100 or 200
	reg_lambda	Controls the L2 regularization term on weights	0.001, 0.01, 0.1, 1.0
SVM	C	Regularization parameter that controls the penalization of the classification errors	0.1, 0.2, 0.5, 1.0, 2, 5, 10, 20, 50, or 100
	Kernel	Type of kernel function to transform the features space	"rbf" or "poly"
	Gamma	Controls the shape and smoothness of the decision boundary	0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, or 1
	Degree	Only for polynomial kernels, it defines the degree of the polynomial kernel function	2, 3, 4, or 5
FFNN	Activation	Type of activation function for the hidden layers	"relu" or "tanh"
	n_neurons	Number of neurons per hidden layer	32, 64 or 128
	n_layers	Number of hidden layers	3, 4 or 5
	L2	L2 regularization term on weights	10^{-2}, 10^{-3} or 10^{-4}

Each training process was run on a laptop with 16 GB of RAM, an Nvidia GeForce GTX 1650 Ti GPU, and an Intel Core i7-10750H CPU, at 2.60 GHz, dedicating 10 out of the 12 available threads exclusively to computations. The tuning process tested a different number of combinations of hyperparameters in relation to the type of classifier: 144 instances for KNN, 52 for RF, 288 for XGB, 640 for SVM, and 54 for FFNN were investigated. Thus, the computational time (hereafter: CPU Time, in seconds), normalized over the number of cycles of nested cross-validation, was reported as an additional metric to compare the classifiers performance.

Optimization and validation of the best classifiers

For both the self-perceived and therapist-perceived engagement classification targets, the better-performing classifier, paired with the appropriate dataset preparation pipeline, was subjected to further training and optimization. Thus, nine-fold cross-validation for hyperparameters tuning (see Table 5) was run over the whole dataset for the selected classifiers.

Validation of the optimized classifiers for self-perceived and therapist-perceived engagement prediction was done on the validation folds. Specifically, an average confusion matrix was obtained by aggregating the confusion matrices computed on the validation folds at each kth iteration. Similarly, Receiver Operating Characteristic (ROC) curves for each class were produced.

At last, the permutation importance algorithm was applied to the better-performing models for both classification targets to assess the weights of each feature on models predictions [58]. Within this work, the permutation feature importance was defined for each feature as the decrease in the models F1 Macro when the feature itself was randomly shuffled (hereafter: ΔF1 Macro).

Results

Classification targets

Data related to a total of 110 independent rehabilitation sessions over 46 subjects were collected. After signal processing, windowing, feature extraction and normalization, the dataset was composed of 542 records, each with 14 HRV and 19 EDA features.

The class distribution following the labelling procedure resulted to be highly unbalanced for both self-perceived and therapist-perceived engagement: for self-perceived engagement, 7.38% over 542 samples was labelled as "Underchallenged", 23.25% as "Minimally Challenged", and 69.37% as "Challenged". The Krippendorf α coefficient measured an inter-rater agreement of 81%.

On the other hand, sub-items of the therapist-reported outcome were grouped as follows:

- $T1^{(-)}$, $T5^{(-)}$, $T8^{(-)}$ and $T12^{(-)}$ fell within the "Underchallenged" class.
- $T3^{(+)}$, $T4^{(+)}$ and $T6^{(+)}$ in the "Minimally Challenged" class.
- $T1^{(+)}$,$T2^{(+)}$,$T5^{(+)}$,$T8^{(+)}$,$T9^{(+)}$, $T12^{(+)}$ in the "Challenged" class.
- Items $T2^{(-)}$, $T3^{(-)}$, $T4^{(-)}$, $T6^{(-)}$,$T7^{(+)}$, $T7^{(-)}$, $T9^{(-)}$,$T10^{(+)}$, $T11^{(+)}$, $T10^{(-)}$ and $T11^{(-)}$ were excluded since they did not match any of the three engagement levels, according to both raters.

The Krippendorf α coefficient revealed an inter-raters agreement in grouping the items of 82%. The labelling strategy adopted in this case led to the following distribution: 9 records related to two experimental sessions were not considered due to the fact that multiple engagement classes reported the same average score; 10.69% over the remaining 533 records was labelled as "Underchallenged", 53.85% as "Minimally Challenged", and 35.46% as "Challenged".

The difference in the self-perceived and therapist-perceived engagement is highlighted in the contingency table presented in Fig. 1. More in detail, 75% of the "Underchallenged" sessions according to self-perceived engagement were instead defined as "Minimally Challenged" or "Challenged" in therapist-perceived engagement. Secondly, when there was no concordance, sessions that were classified as "Minimally Challenged" in the self-perceived outcome were considered in the therapist-reported outcome as "Challenged" 64% of the time

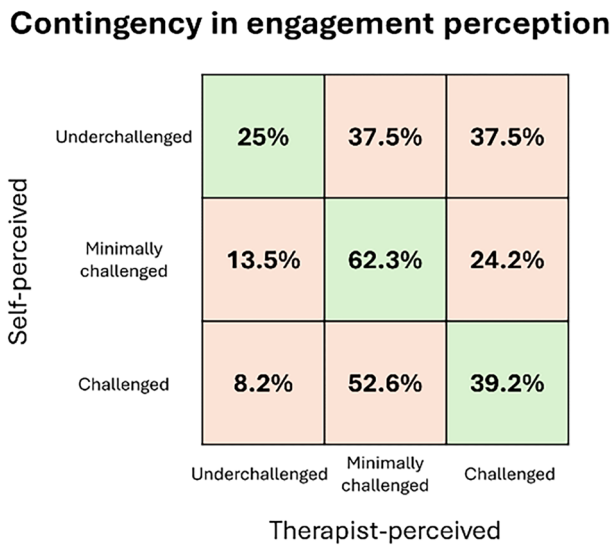

Fig. 1 Contingency table between self-perceived and therapist-perceived engagement classification. Green cells stand for concordant classification, and red cells indicate discordant classification

and as "Underchallenged" 36% of the time. Lastly, 13.5% of "Challenged" sessions according to the self-reported outcomes were identified in the therapist-reported outcome as "Underchallenged", while 86.5% of the incongruous sessions were classified as "Minimally Challenged". Supplementary Figure 1 and Supplementary Figure 2 show the time series of HRV and EDA physiological data according to the three classification targets.

Impact of dataset preparation on models performance

This section provides insights into the outcomes of classifiers across different dataset preparation approaches, divided into three sequential stages (ANS modeling, Feature reduction, Data augmentation).

ANS modeling

The impact of the ANS modeling on the classifiers performance for both classification targets is reported in Fig. 2. Accounting for the median performance of the five classifiers for both classification targets, the bimodal dataset as an input to AI models outperformed the unimodal HRV and EDA datasets. For self-perceived engagement, BD yielded a median F1 Macro of 0.78 ± 0.12, while the unimodal HRV and EDA datasets yielded 0.64 ± 0.13 and 0.67 ± 0.07, respectively. Likewise, for therapist-perceived engagement, BD obtained a median F1 Macro of 0.78 ± 0.07, while the unimodal HRV and EDA datasets reached lower values of 0.57 ± 0.06 and 0.70 ± 0.08. In addition, considering each classifier, the BD approach turned out to be the most profitable choice in terms of F1 Macro, reaching peaks of 0.83 ± 0.06 and 0.83 ± 0.05 for self-perceived and therapist-perceived engagement, both with SVM.

Supplementary Table 2 collects the training CPU Time, expressed in seconds, for each classifier, dataset and classification target. Similar patterns were observed across the three ANS modeling approaches and both classification targets. Notable variations in CPU Time across models were observed: FFNN and XGB exhibited significantly higher CPU Time, with FFNN the most computationally expensive, while KNN, RF, and SVM demonstrated comparatively lower CPU Time. In addition, RF and XGB CPU Time decreased as the number of input features decreased going from BD (i.e., 33 features) to the unimodal datasets (i.e., 14 features for unimodal HRV and 19 for unimodal EDA). Conversely, KNN, SVM, and FFNN experienced an increase in CPU Time as the number of input features decreased.

To sum up, although nothing can be claimed on the statistical superiority of the BD approach, it seems that, qualitatively, the BD approach outperformed the unimodal ones in terms of classifiers F1 Macro for both classification targets, and three out of five models reduced their CPU Time when trained on BD over the unimodal datasets. As a result, the BD approach was chosen as the most suitable for the following stages.

Feature reduction

The literature-based feature reduction approach allowed to reduce the number of features from 33 to 20 (i.e., 8 for HRV, 12 for EDA). Mean HR, SDNN, RMSSD, pNN50, Triangular Index, and HRV Skewness were selected among the time-domain HRV features according to [23, 27, 66], whereas HRV nLF and HRV nHF were chosen to represent frequency-domain HRV features, as suggested in [23]. Concerning EDA, the following features were selected:

- Mean EDA Tonic, Max Upspeed EDA Tonic, and Max Downspeed EDA Tonic to represent the time-domain tonic component features, according to [52].
- NS.EDRs, Mean EDA Phasic Peak Amplitude, nAUC EDA Phasic, and Mean Rise Time to represent the

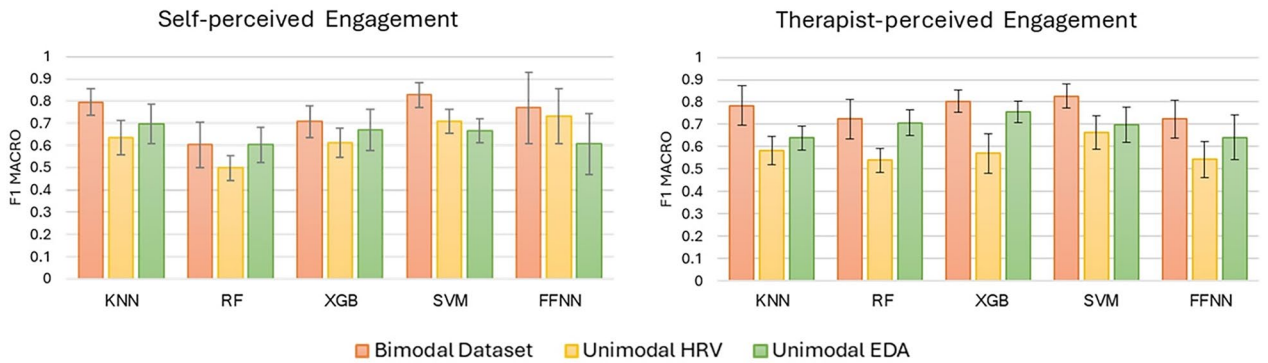

Fig. 2 Mean and standard deviation (error bars) F1 Macro as a function of classifiers and ANS modeling for both self-perceived and therapist-perceived engagement classification targets

time-domain phasic component features, in line with [53, 67].
- All five frequency-domain features, as reported in [53].

Figure 3 reports F1 Macro scores of each classifier for both classification targets in BD without feature reduction and with literature-based feature reduction. According to the median, F1 Macro values observed for BD without feature reduction (0.77 ± 0.12 for self-perceived engagement, and 0.78 ± 0.08 for therapist-perceived engagement) and with literature-based feature reduction (0.76 ± 0.12 for self-perceived engagement, and 0.77 ± 0.05 for therapist-perceived engagement) were qualitatively comparable.

As for the projection-based feature reduction, PCA served as a pivotal feature extraction technique, being applied with three distinct levels of CEV threshold retention: 80%, 90%, and 95%. PCA 80% CEV reduced the feature space to the first 9 PCs; PCA 90% CEV kept up to the first 14 PCs; and PCA 95% CEV up to the first 18. F1 Macro of all classifiers as a function of the projection-based feature reduction approach is shown in Fig. 4. Across the five classifiers, either for self-perceived or therapist-perceived engagement, BD without feature reduction reported consistently higher median F1 Macro scores compared to all PCA approaches for feature reduction. Furthermore, a decrease in median F1 Macro was observed for both classification targets when transitioning from PCA 95% CEV (0.72 ± 0.19 for self-perceived engagement, and 0.64 ± 0.16 for therapist-perceived engagement) to PCA 90% CEV (0.70 ± 0.13 for self-perceived engagement, and 0.66 ± 0.12 for therapist-perceived engagement), and further down to PCA 80% CEV scenarios (0.64 ± 0.16 for self-perceived engagement, and 0.62 ± 0.12 for therapist-perceived engagement). Moreover, it is noteworthy to highlight that the impact of PCA varied across individual classifiers. Specifically, RF and XGB exhibited a more pronounced decline in F1 Macro with the application of PCA: compared to BD without feature reduction, up to −28% loss for RF and −27% loss for XGB in self-perceived engagement,

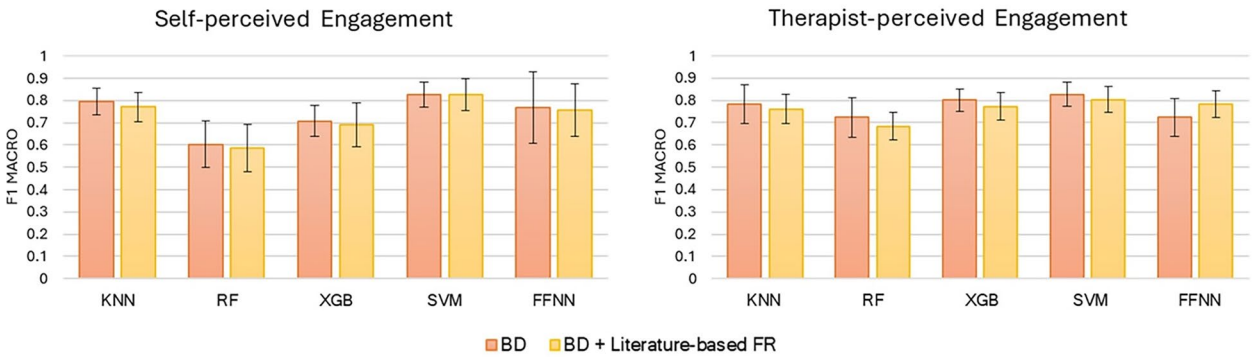

Fig. 3 Mean and standard deviation (error bars) F1 Macro for each classifier in BD and BD + literature-based feature reduction for both self-perceived and therapist-perceived engagement classification targets. *BD* Bimodal Dataset, *FR* Feature Reduction

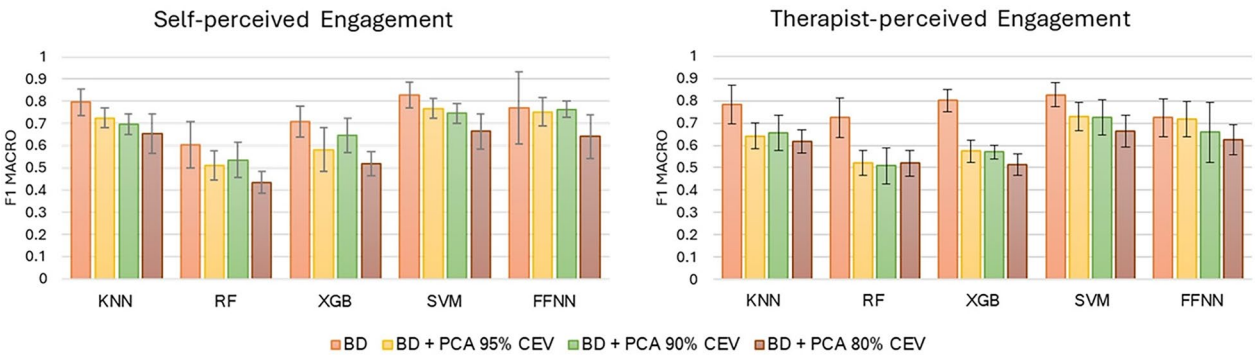

Fig. 4 Mean and standard deviation (error bars) F1 Macro for each classifier in BD without feature reduction and BD + projection-based feature reduction for both self-perceived and therapist-perceived engagement classification targets. *BD* Bimodal Dataset, *CEV* Cumulative Explained Variance

and up to − 36% loss for XGB and − 28% loss for RF in therapist-perceived engagement. Conversely, KNN, SVM and FFNN demonstrated a relatively modest impact of PCA, experiencing comparatively less degradation in performance: compared to BD without feature reduction, up to − 20% loss for SVM, − 18% for KNN, and − 17% for FFNN in self-perceived engagement, and up to − 21% loss for KNN, − 20% for SVM, and − 14% for FFNN in therapist-perceived engagement.

CPU Time results as a function of feature reduction (Supplementary Table 3) matched the previously described trends in Supplementary Table 2. On average, the most computationally expensive classifier was FFNN, followed by XGB. KNN, RF, and SVM all had quite low training CPU Time. Furthermore, both feature reduction algorithms mitigated RF and XGB CPU Time. In contrast, KNN, SVM, and FFNN did not appear to get the same computational benefits from feature reduction.

Given the foregoing results for each classification target, in terms of both F1 Macro and CPU Time, any type of feature reduction approach was excluded from the dataset preparation pipeline. As a result, the most suitable dataset preparation scenario for the next stage was set to the bimodal dataset without feature reduction.

Data augmentation

Figure 5 reports the influence of data augmentation on the F1 Macro scores for both classification targets. Each classifier exhibited similar trends in response to the incremental data augmentation applied to BD, independently of the classification target. The median F1 Macro across classifiers showed a consistent increase in models performance after switching from BD to the 4-min window DA scenario (0.90 ± 0.04 for self-perceived engagement and 0.92 ± 0.03 for therapist-perceived engagement). Moving from a 4-min to a 3-min windows DA scenario resulted in peaks in F1 Macro scores for both classification targets (0.94 ± 0.07 for self-perceived engagement and 0.93 ± 0.03 for therapist-perceived engagement). DA with 2-min windows (0.92 ± 0.06 for self-perceived engagement and 0.91 ± 0.03 for therapist-perceived engagement) and 1-min windows (0.90 ± 0.06 for self-perceived and 0.88 ± 0.06 for therapist-perceived engagement) resulted in a slight decrease in performance. Again, it is crucial to highlight that data augmentation impacted differently on classifiers. For self-perceived engagement, classifiers that benefited the most from data augmentation were RF (up to a 31% increase in F1 Macro) and XGB (+ 25%), followed by FFNN (+ 19%), and then by SVM (+ 13%) and KNN (+ 11%). Similarly, for therapist-perceived engagement, data augmentation had a smaller influence on F1 Macro scores for KNN (up to 11%) and SVM (+ 12%), a considerable improvement in XGB (+ 16%), and the biggest improvements in FFNN (+ 22%) and RF (+ 22%). Each percentage gain mentioned above refers to the transition from BD without DA to the best-performing DA scenario, which consistently turned out to be the 3-min windows DA scenario.

CPU Time results as a function of data augmentation for both classification targets are reported in Supplementary Table 4. In general, CPU Time of each classifier increased as the dataset size increased due to data augmentation. Interestingly, the growth trend of CPU Time was different across classifiers since KNN and SVM showed a more than linear increasing trend of CPU Time as a function of the dataset size, while a linear increasing trend was typical of RF, XGB, and FFNN, both in case of self-perceived engagement and therapist-perceived engagement.

Considering the median F1 Macro scores qualitatively revealing the 3-min windows DA scenario as the most favorable, coupled with the fact that the associated CPU Time increment remained within acceptable bounds, the 3-min windows data augmentation was used into the dataset preparation pipeline.

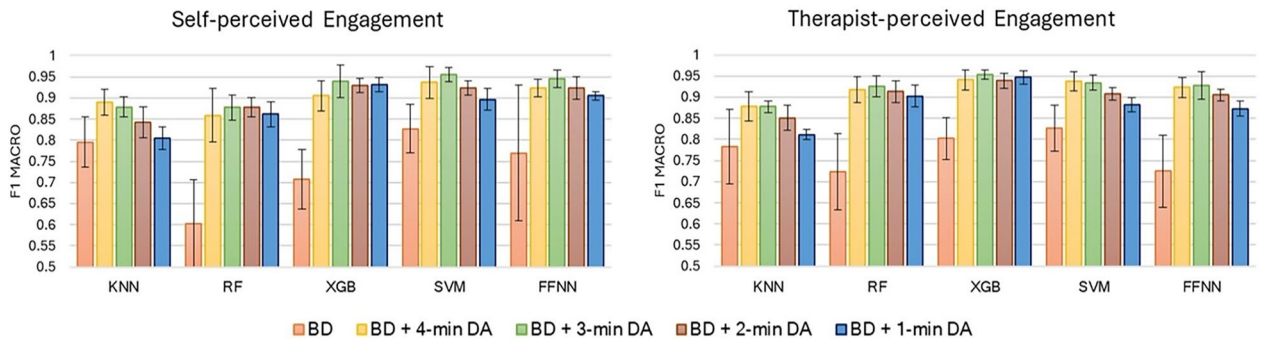

Fig. 5 Mean and standard deviation (error bars) F1 Macro as a function of classifiers and data augmentation for both self-perceived and therapist-perceived engagement classification targets. *BD* Bimodal Dataset, *t-min* t-minute window, *DA* Data Augmentation

Best classifiers

According to the performance outcomes of the nested cross-validation, the best classifier to predict the self-perceived engagement was SVM (F1 Macro: 0.96 ± 0.02), while the best model for therapist-perceived engagement prediction was XGB (F1 Macro: 0.95 ± 0.01), both trained on the bimodal dataset, with the 3-min windows data augmentation approach.

Further training on the whole dataset with a ninefold cross-validation was done to definitively tune the hyperparameters of the best models. The optimization process of the SVM model for self-perceived engagement prediction led to the following best hyperparameters: $C = 100$; $kernel = rbf$; $gamma = 0.05$. In parallel, the following set of optimized hyperparameters was obtained for XGB model for therapist-perceived engagement prediction: $learning\_rate = 0.2$; $max\_depth = 6$; $n\_estimators = 200$; $reg\_lambda = 0.1$.

Figure 6 shows the average confusion matrix on the validation folds relative to the best models for self-perceived and therapist-perceived engagement.

For self-perceived engagement, the class "Underchallenged" was wrongly predicted in 7.5% of cases, and in 100% of wrong predictions it was confused with "Challenged". Additionally, the class "Minimally Challenged" was badly predicted in 9.52% of cases, with misclassifications towards the class "Underchallenged" in 2.7% of wrong predictions, and towards the class "Challenged" in the remaining 97.3%. Lastly, the class "Challenged" was almost never wrongly predicted (0.89% of cases), in 21.7% of wrong cases it was confused with the class "Underchallenged", and in 78.3% of cases with the class "Minimally Challenged".

For therapist-perceived engagement, the class "Underchallenged" had a 12.28% misprediction rate, with 62% of errors involving misclassification as "Challenged", and 38% as "Minimally Challenged". Secondly, the class "Minimally Challenged" showed a remote 1.39% misprediction rate exclusively towards the class "Challenged" (100%). Notably, the class "Challenged" experienced a small misprediction rate (4.41%), with 96% of these inaccuracies involving confusion with "Minimally Challenged" and the remaining 4% with "Underchallenged".

Average ROC curves on the validation folds and permutation features importance outcomes of the best classifiers for self-perceived and therapist-perceived engagement are shown in Figs. 7 and 8, respectively.

Discussion

The current study assessed the performance of five AI models, namely KNN, RF, XGB, SVM, and FFNN, to predict the level of patient engagement during RAGR when applied to record-wise HRV and EDA physiological features.

The crucial role of investigating the feasibility of different artificial intelligence algorithms when dealing with classification of emotional states was widely addressed by [32], who also listed five key prerequisites for extracting representative features of the ANS activation: ANS modeling, training set preparation, feature extraction, normalization, and dimension reduction. Thus, the present study also aimed at evaluating the impact of ANS

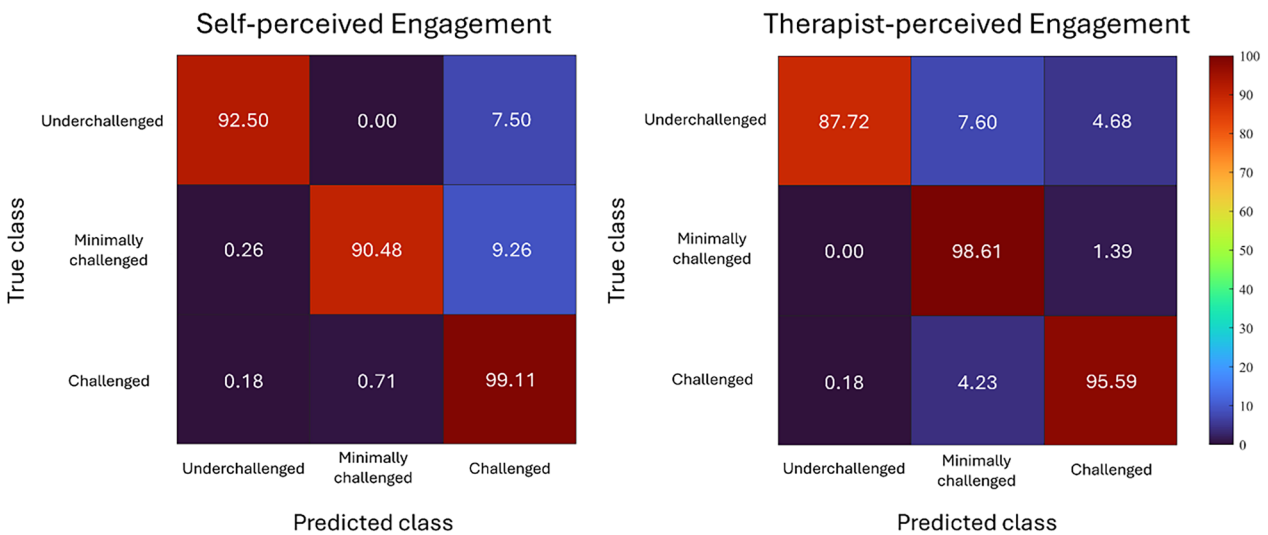

Fig. 6 Average Confusion Matrix on the validation folds of the best classifiers for self-perceived and therapist-perceived engagement classification targets. Values are expressed in percentage with respect to the size of each class

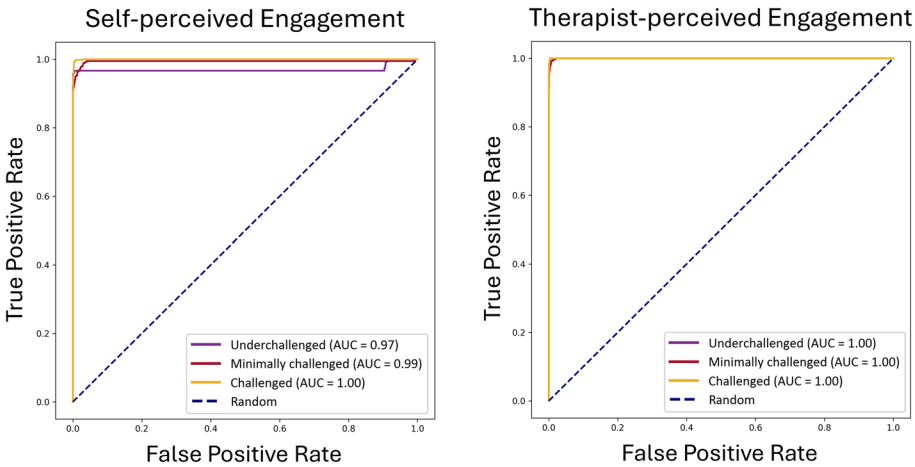

Fig. 7 Average ROC curves on the validation folds of the best classifiers for self-perceived and therapist-perceived engagement classification targets. Class-specific ROC Area Under the Curve values are reported in brackets within the legend. The label Random refers to the theorical behavior of a dumb model unable to discriminate between true and negative samples. *AUC* Area Under the Curve

modeling, feature reduction, and data augmentation on the selected AI models performance.

Classification targets

The importance of assessing and predicting both self-perceived and therapist-perceived engagement was highlighted by [18]. Specifically, they stated that although engagement can be seen as an internal state of the patient, therapists must be able to ascertain the level of engagement or disengagement in order to modify the therapy accordingly, suggesting that both patients and therapists thoughts and feelings play a crucial role in highly positive rehabilitation outcomes. Our study further explored this aspect since it preliminarily highlighted that there can be uncertainty in concordance between the patients' and therapists' perceptions of the level of engagement. This could be caused by a certain unreliability of the self-reported outcomes due to patients' early age or clinical condition. Nonetheless, the lack of agreement enhanced the decision to split the engagement prediction into two distinct classification targets, and further emphasized the significance of considering both perspectives for a comprehensive evaluation of engagement.

To date, the choice of the correct number of classes to discretize engagement is still controversial. While this study arranged three engagement levels (i.e., "Underchallenged", "Minimally Challenged" and "Challenged") according to the definition of engagement as a continuum proposed by [15], others adopted slightly different approaches. In Koenig and colleagues [22], three cognitive engagement levels were defined, namely "Underchallenged", "Challenged", and "Overchallenged". Similarly, Li and colleagues [68] proposed a three-level discretization of subject engagement in "Engaged", "Normal", and "Bored". In Gokay et al. [69], instead, the valence-arousal model representation was exploited to split both emotional and cognitive state into two binary levels, while other researchers discretized cognitive workload into as many classes as the number of task difficulty levels employed in the study protocol [70–73]. Despite the significant interest in examining the presence of elements leading patients, especially in pediatric age, to interact negatively with the rehabilitation therapy to the point of feeling overwhelmed, it is important to note that these aspects were not considered because they fall outside the definition of engagement proposed by Bright et al. [15] that was used for this study.

The expert review evaluation method employed for both classification targets proved to be a reliable choice for condensing self-reported and therapist-reported outcomes into the engagement classes, according to the Krippendorf α values reported for both self-perceived and therapist-perceived engagement.

For both classification targets, the resulting class distribution was highly unbalanced since few records of "Underchallenged" were available. Regarding the possibility of rebalancing the dataset, a deliberate decision was made to maintain it unbalanced, to reflect the matter-of-facts distribution of classes within the reference population.

Impact of dataset preparation on models performance

The impact of dataset preparation techniques on classifiers performance for self-perceived and

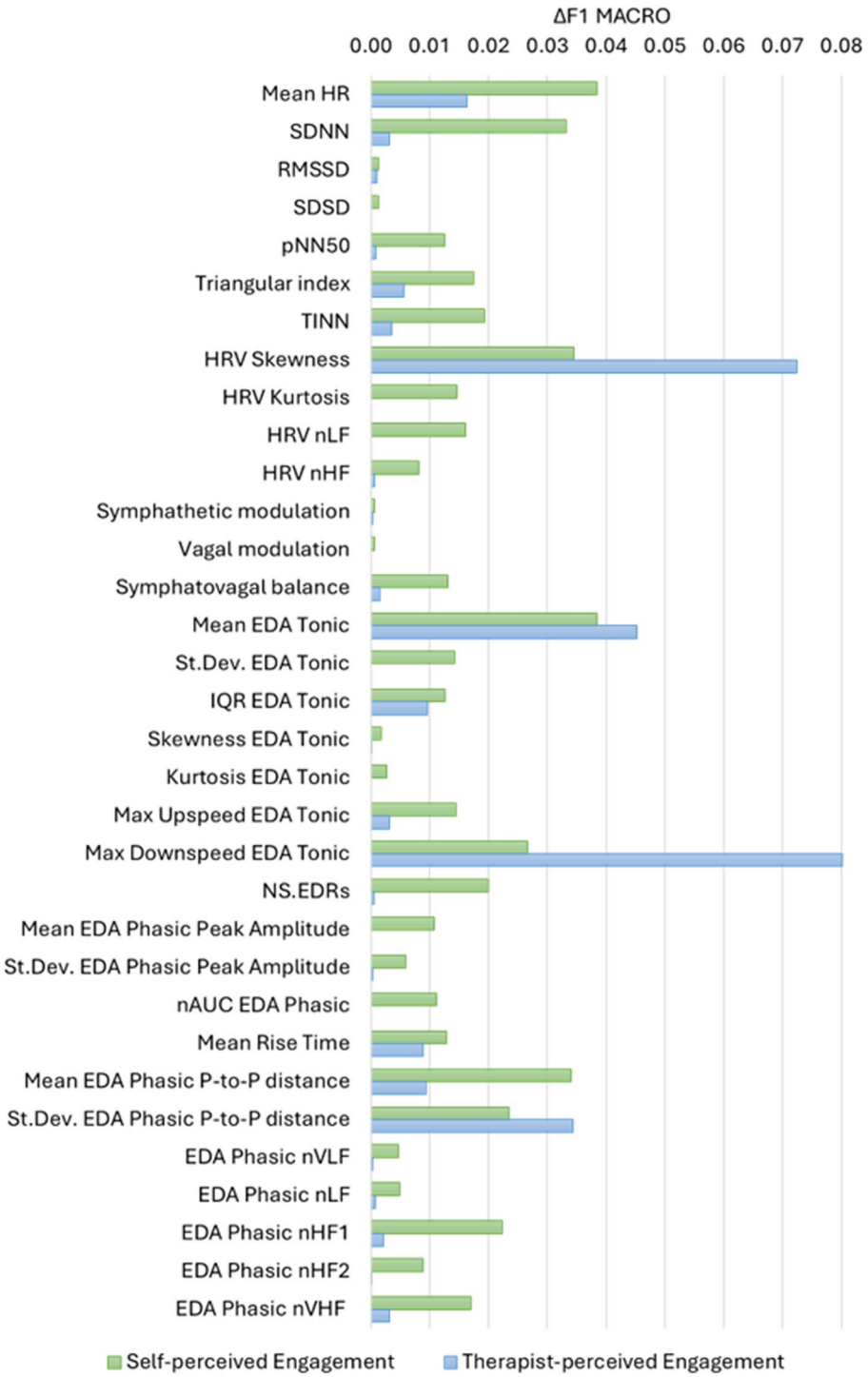

Fig. 8 ΔF1 Macro as a function of input features according to the permutation feature importance algorithm performed on the best classifiers for both self-perceived and therapist-perceived engagement classification targets

therapist-perceived engagement prediction were systematically explored, revealing valuable insights into the effectiveness of various approaches.

ANS modeling

Qualitatively speaking, the fusion of features from HRV and EDA signals within the bimodal dataset generally yielded better F1 Macro scores than the unimodal HRV and EDA datasets. This may highlight the synergistic effects of training AI models on multimodal data, as the combination of features from both physiological signals could have provided a more comprehensive representation of various aspects related to patient engagement. This result is supported by the review in [74], whose findings demonstrate a significative increase in AI models accuracies when trained on multimodal datasets rather than the unimodal ones. Indeed, there is wide evidence that HRV features better replicate emotional states [20, 25], while EDA features reflect cognitive states more effectively [22, 24]. To sum up, the bimodal dataset likely enabled a more complete representation of engagement compared to the unimodal datasets without affecting CPU Time.

Finally, the unimodal EDA consistently outperformed the unimodal HRV. This could be attributed to HRV increased susceptibility to physical activity, or to a stronger representation of the classification target by the EDA features.

Feature reduction

The use of literature-based feature reduction resulted in comparable F1 Macro with BD with no feature reduction, suggesting that the features selected according to the literature findings could express nearly all the information enclosed in the bimodal dataset.

On the other hand, projection-based feature reduction approaches led classifiers to consistently decrease F1 Macro, thus proving that PCs were less informative than the original features. Accordingly, these results suggested that, although PCs explained a huge portion of the dataset variance, they lacked, at least partially, the direct physiological interpretability in terms of engagement captured by the original features [75]. Moreover, the decreasing trend of F1 Macro scores from 95% CEV to 80% CEV highlighted the trade-off between information preservation and dimensionality reduction [75]. Notably, XGB and RF models showed greater sensitivity to projection-based feature reduction compared to other classifiers. In fact, RF and XGB are both ensemble learning methods that leverage on Decision Tree, whose aim is to capture complex relationships within the data to perform predictions [76]. The loss of features interdependence induced by PCA might have negatively influenced RF and XGB capability to capture that intricated relationship in data, resulting in a more significant decrease in performance.

As for the variation in CPU time based on the number of features, it was observed that SVM, KNN, and FFNN had higher computational costs as the number of features decreased. In contrast, RF and XGB better benefited from their more efficient parallelized training computation process, thus reporting a decrease in CPU Time as the number of features increased.

To summarize, projection-based feature reduction worsened the models performance; hence, it was excluded. Similarly, although literature-based feature reduction showed comparable performance with BD, it increased CPU Time for the majority of the AI models; therefore, its use was discouraged.

Finally, some controversy is reported in the literature regarding the impact of feature reduction on models performance. Chanel and colleagues [77] tested an ANOVA-based feature reduction approach that led to improved performance of SVM when used for emotional states prediction by means of electroencephalographic features. Amiriparian and colleagues [78], instead, tested the effects of various feature reduction approaches (i.e., PCA, correlation analysis, sequential forward selection, competitive swarm optimization) for valence and arousal prediction using HRV and EDA features compared to a baseline scenario where feature reduction was not applied at all. They found out that no feature reduction algorithm was able to outperform the baseline scenario on valence prediction, while sequential forward selection was superior to any other approach and the baseline scenario when applied to arousal prediction.

Data augmentation

Data augmentation led to a remarkable increase in F1 Macro scores for all classifiers and both classification targets, introducing diversification into the training dataset, promoting generalization and pattern recognition, and consequently mitigating the risk of overfitting, according to [79]. The positive impact of data augmentation is in line with [80, 81], where it successfully improved classification outcomes of several classifiers in emotion recognition tasks. Nonetheless, observing the median F1 Macro trends as a function of DA scenarios, from BD scenario without DA to 1-min windows DA, it seemed that the dataset size was excessively increased beyond the 3-min windows DA scenario. This may inadvertently have led to over-augmentation, wherein the introduction of excessive variance has potentially undermined the classifiers ability to generalize effectively. Thus, it is worth noting that the breakeven point in the trade-off between data augmentation and data distribution was found for both classification targets in the 3-min windows DA scenario. Again, some classifiers, such as XGB and RF, benefited from data augmentation more than others since the

increase in dataset size better enhanced their proficiency in capturing complex patterns among features.

As for the CPU Time increase as a function of dataset size (hereafter expressed as N), SVM and KNN exhibited higher computational complexity than XGB, FFNN, and RF. According to previous studies, a complexity of $O(N^3)$ was found for SVM [60]; KNN, RF, and XGB share the same time complexity of $O(NlogN)$ [57–59]; FFNN, instead, shows a complexity of $O(N)$ [82]. On this basis, KNN, SVM, and FFNN for self-perceived and therapist-perceived engagement classification had consistent time complexities with previous works, while training for RF and XGB was faster than expected, probably due to the more efficient parallelization of computation.

Models for engagement prediction

SVM for self-perceived engagement and XGB for therapist-perceived engagement, both trained on the bimodal dataset with a 3-min windows data augmentation, obtained the highest F1 Macro scores.

There is wide evidence in the literature that proved the high efficacy of various AI models for affective states or cognitive workload prediction in rehabilitation [33–35, 83]. For instance, Koenig et al. developed a Kalman Adaptive Linear Discriminant Analysis model for predicting cognitive workload in robot-assisted gait rehabilitation scenarios in adults. This model was based on performance data, heart rate, breathing rate, skin temperature, skin conductance, and reaction forces exchanged with the robot, achieving an accuracy of 75% on stroke patients [33]. Secondly, XGB was reported as the best classifier for predicting cognitive workload by means of EEG and EDA features in [34], while Gogna and colleagues [35] found out that SVM achieved the highest scores in cognitive workload prediction tasks. Furthermore, a study on stroke patients virtual rehabilitation used linear SVM models to classify affective states such as tiredness, tension, pain, and satisfaction based on 3D hand movement and finger pressure, with a ROC AUC of 71% [83].

Nevertheless, this is the first study that primarily focuses on engagement as a multidimensional construct on a group of patients including children, adolescents and few adults.

A very slight bias or preference of the self-perceived engagement model in predicting the "Challenged" class, which can be assessed from the average confusion matrix outcomes, should be further investigated. The tendency for SVM to forecast instances as “Challenged” could have been caused by some noise in the raw physiological signals, notably movement-related artifacts, which were limited, but not definitively addressed by signals conditioning, and could have been misread as cognitive involvement by SVM. As for the therapist-perceived engagement model, no relevant bias was observed, rather mispredictions were overall randomly distributed.

The optimal ROC curves indicated the remarkable proficiency of the best models in effectively distinguishing between different engagement levels within both classification targets.

The investigation of permutation feature importance revealed notable differences between SVM and XGB. SVM displayed a greater distribution of feature importance, indicating that many features consistently contributed to classification. In contrast, XGB relied on a smaller number of features. Provided that this behavior is to credit to the intrinsic characteristics of the models, the diversified handling of features by SVM may contribute to greater robustness in further generalization, whereas XGB increased emphasis on a restricted subset of features may imply less replicability. As a result, minimal fine-tuning of the current XGB model may be required before it is applied to new unseen data from various contexts of application.

Study limitations

The study has several limitations. Firstly, the high age variance among the patients due to the presence of 4 young adults and 4 elderly adults could have affected the interpretability of the HRV and EDA features. Nonetheless, we believe that this could have enhanced the variability of the dataset and improved the generalization abilities of proposed AI models for engagement prediction in RAGR.

Secondly, both self-perceived and therapist-perceived engagement levels should be intended as the most common engagement state observed throughout the entire RAGR activity, which may not fully represent dynamic engagement variations. Nevertheless, unless undergoing a total change in thinking towards unsupervised learning, supervised model training for engagement prediction inevitably requires the collection of subjective outcomes from all stakeholders (i.e., service providers and patients) during the rehabilitation activities. Provided that qualitative assessments can only be done at discrete timepoints, they necessarily must refer to specific time intervals during the rehabilitation activity.

In addition, both classification targets exhibited imbalance across different engagement classes, thus limiting exploration of additional engagement levels. Nevertheless, this dataset is representative of the real picture of RAGR.

Another limitation to take into account is the subjectivity and potential bias of labeling process. Nevertheless, a double-blind evaluation process was performed to provide the most objective data possible.

Furthermore, it is noteworthy to say that physiological data were affected by motion-related artifacts, thus introducing noise into the bimodal dataset and slightly biasing the SVM predictions, particularly towards the "Challenged" class. Nevertheless, this again is embedded in real-world data acquisition.

Moreover, the high performance of the AI models observed in this study may be partially associated with the manual selection of high-quality signal windows, suggesting that the AI models performance could decrease when applied to real-word data with higher grade of noise.

Finally, the artificial neural network utilized was not optimized to its full potential due to limited computational resources, potentially influencing its performance.

Study strengths and future prospects

This study has several strengths. First, it represents one of the pioneering efforts to comprehensively assess patient engagement towards rehabilitation therapy through the collection of physiological data related to the ANS activity. Furthermore, it marks the first attempt of its kind within the context of pediatric motor rehabilitation and introduces a novel workflow for predicting engagement using AI tools. Additionally, although this work is exclusively focused on lower-limb robot-assisted rehabilitation using the Lokomat, it can be hypothesized that our findings may offer preliminary predictive insights for upper-limb rehabilitation, according to what demonstrated by Cakmak et al. (84). Finally, the study's approach of predicting engagement towards therapy from both the patient's and therapist's perspectives represents a significant innovation.

Prospects will enable further improvements in engagement prediction during robot-assisted motor rehabilitation scenarios. First, deep learning models based on raw physiological data could be developed to provide near-real-time prediction of the patients' engagement level during rehabilitation activity, and to predict the dynamic fluctuations of engagement more accurately. Then, it could be beneficial to explore novel models capable of predicting other constructs than engagement, such as stress in patients towards the rehabilitation therapy. This will be particularly useful for therapists, who will be supported in detecting when patients are experiencing physical and psychological impediments during rehabilitation sessions. Additionally, future work may explore advanced techniques to automatically detect high-quality windows of BVP and EDA raw signals, thus limiting biases potentially introduced by the manual selection of the 5-min windows. Moreover, it would be useful to explore strategies to integrate both patient and therapist perceptions of engagement into a unified scale. Lastly, real-time engagement detection based on these AI models could serve as the foundation for more comprehensive systems aimed at providing suggestions to therapists regarding potential changes to the rehabilitation environment and equipment. This would allow for more personalization of these features to better satisfy the physical and psychological needs of patients.

Conclusions

The present study explored the performance of five artificial intelligence algorithms (i.e., KNN, RF, XGB, SVM, and FFNN) applied to structured HRV and EDA physiological features to predict the level of patient engagement during RAGR. Engagement was assessed with ad hoc reports from both the patient's and therapist's perspectives, and three levels of engagement each were defined, namely "Underchallenged", "Minimally Challenged", and "Challenged". To the best of our knowledge, this study represents the first attempt to predict engagement within an experimental group consisting mainly of pediatric subjects or young adults affected by various neuromotor disorders.

The effects of the three dataset preparation approaches, such as ANS modeling, feature reduction, and data augmentation, on the AI classifiers overall performance were also assessed. Specifically, fusing features from HRV and EDA into a comprehensive dataset enabled a more synergistic and complete representation of engagement compared to the unimodal datasets. Additionally, 3-min windows data augmentation was able to further increase models performance. SVM emerged as the most effective architecture for self-perceived engagement prediction, while XGB was found to be optimal for predicting therapist-perceived engagement, with macro-averaged F1 scores of 95.6% and 95.4%, respectively.

Overall, this study encourages the practical use of AI to improve patient care and rehabilitation outcomes, thus looking forward to more personalized treatment approaches in clinical practice.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
AUC	Area Under the Curve
BD	Bimodal Dataset
BVP	Blood Volume Pulse
CEV	Cumulative Explained Variance
CPU Time	Computational Time
DA	Data Augmentation
ECG	Electrocardiography
EEG	Electroencephalogram
EDA	Electrodermal Activity
FR	Feature Reduction
HF	High Frequency
FFNN	Feed-Forward Neural Network
HR	Heart Rate
HRV	Heart Rate Variability

IQR	Interquartile Range
KNN	K-Nearest Neighbors
LF	Low Frequency
NN	Inter-Beat intervals
NS.EDR	Non-Specific Electrodermal Response
PCA	Principal Component Analysis
PC	Principal Component
RF	Random Forest
RMSSD	Root Mean Square of the Successive Differences
ROC	Receiver Operating Characteristic
SAM	Self-Assessment Manikin
SDNN	Standard deviation of all NN intervals
SDSD	Standard Deviation of the Successive Differences
St.Dev.	Standard Deviation
SVM	Support Vector Machine
TINN	Triangular Interpolation of NN
VHF	Very High Frequency
VLF	Very Low Frequency
XGB	Extreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12984-024-01519-2.

Supplementary Material 1.

Acknowledgements
The authors acknowledge Roberta Morganti for her support in training the patients, Enrico Magliulo for his preliminary data preparation, Ettore Micheletti, for having performed the self-reported and therapist-reported data entry, and all the patients and families that accepted to participate to the study.

Author contribution
S.C. and A.F. performed the study conceptualization and formal analysis; S.C, A.F., M.C., G.M., C.D., and S.B. took part to data curation; S.C., A.F. and E.B. performed preliminary investigation; S.C., A.F., M.C., and G.M. developed and refined the methodology; S.C. and A.F. prepared all tables and figures; C.D. recruited the subjects; E.B. supervised the study and handled funding acquisition; S.C. and A.F. wrote the main manuscript text; all authors reviewed the manuscript.

Funding
This study was supported by the Italian Ministry of Health (Ricerca Corrente 2021/2024 to E. Biffi); by the Italian Ministry of University and Research (Doctoral Scholarship awarded to S. Costantini); by the Italian Ministry of University and Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (# PNC0000007)." The authors declare no conflicts of interest with the added funding.

Availability of data and materials
Data is available on Zenodo: https://doi.org/https://doi.org/10.5281/zenodo.10812450.

Declarations

Ethics approval and consent to participate
The study was performed in accordance with the Declaration of Helsinki, and the Ethics Committee of Scientific Institute E. Medea approved the observational study protocol (protocol code: Prot. N. 02/22-CE; date of approval: January 27th, 2022). Patients, if adults, or their guardians signed a written informed consent.

Consent for publication
All subjects consented for their data to be published.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Electronics Information and Bioengineering, Politecnico di Milano, Milan, Italy. [2]Scientific Institute, IRCCS "E. Medea", Bosisio Parini, Italy. [3]Department of Design, Politecnico di Milano, Milan, Italy.

Received: 10 May 2024 Accepted: 27 November 2024
Published online: 19 December 2024

References

1. Beretta E, Storm FA, Strazzer S, Frascarelli F, Petrarca M, Colazza A, et al. Effect of robot-assisted gait training in a large population of children with motor impairment due to cerebral palsy or acquired brain injury. Arch Phys Med Rehabil. 2020;101(1):106–12.
2. Mehrholz J, Thomas S, Kugler J, Pohl M, Elsner B, Mehrholz J, et al. Electromechanical-assisted training for walking after stroke (Review). Cochrane Database Syst Rev. 2020. https://doi.org/10.1002/14651858.CD006185.pub5.
3. Calabrò RS, Cacciola A, Bertè F, Manuli A, Leo A, Bramanti A, et al. Robotic gait rehabilitation and substitution devices in neurological disorders: where are we now? Neurol Sci. 2016;37:503–14.
4. Valè N, Gandolfi M, Vignoli L, Botticelli A, Posteraro F, Morone G, et al. Electromechanical and robotic devices for gait and balance rehabilitation of children with neurological disability: a systematic review. Appl Sci. 2021;11(24):12061.
5. Sharma N, Classen J, Cohen LG. Neural plasticity and its contribution to functional recovery. 1st ed. Vol. 110, Handbook of Clinical Neurology. Elsevier B.V.; 2013. 3–12 p. https://doi.org/10.1016/B978-0-444-52901-5.00001-0
6. Bracke P, Bruynooghe K, Verhaeghe M. Boredom during day activity programs in rehabilitation centers. Sociol Perspect. 2006;49(2):191–215.
7. Zbogar D, Eng JJ, Miller WC, Krassioukov AV, Verrier MC. Movement repetitions in physical and occupational therapy during spinal cord injury rehabilitation. Spinal Cord. 2017;55(2):172–9. https://doi.org/10.1038/sc.2016.129.
8. Danzl MM, Etter NM, Andreatta RD. Facilitating neurorehabilitation through principles of engagement. Article J All Health. 2012.
9. Guadagnoll MA, Lee TD. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. J Motor Behav. 2004;36:212–24.
10. Krakauer JW, Hadjiosif AM, Xu J, Wong AL, Haith AM. Motor learning. Compr Physiol. 2019;9:613–63.
11. Bassi M, Ferrario N, Ba G, Delle Fave A, Viganò C. Quality of experience during psychosocial rehabilitation: a real-time investigation with experience sampling method. Psychiatr Rehabil J. 2012;35(6):447–53.
12. Novak Vesna D. and Koenig AC and RR. Psychophysiological Integration of Humans and Machines for Rehabilitation. In: Reinkensmeyer David J. and Marchal-Crespo L and DV, editor. Neurorehabilitation Technology. Cham: Springer International Publishing; 2022. p. 207–21. https://doi.org/10.1007/978-3-031-08995-4_10
13. Lequerica AH, Kortte K. Therapeutic engagement: a proposed model of engagement in medical rehabilitation. Am J Phys Med Rehabil. 2010;89(5):415–22.
14. King G, Currie M, Petersen P. Child and parent engagement in the mental health intervention process: a motivational framework. Child Adolesc Ment Health. 2014;19(1):2–8.
15. Bright FAS, Kayes NM, Worrall L, McPherson KM. A conceptual review of engagement in healthcare and rehabilitation. Disabil Rehabil. 2015;37(8):643–54.
16. Zhong B, Niu W, Broadbent E, McDaid A, Lee TMC, Zhang M. Bringing psychological strategies to robot-assisted physiotherapy for enhanced treatment efficacy. Front Neurosci. 2019;18:13.
17. Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. I. B&w Thu Exp Psychrar. 1994;25
18. King G, Chiarello LA, Thompson L, McLarnon MJW, Smart E, Ziviani J, et al. Development of an observational measure of therapy engagement for pediatric rehabilitation. Disabil Rehabil. 2019;41(1):86–97.
19. Tatla SK, Jarus T, Virji-Babul N, Holsti L. The development of the pediatric motivation scale for rehabilitation. Can J Occup Ther. 2015;82(2):93–105.

20. Denson T, Grisham J, Moulds M. Cognitive reappraisal increases heart rate variability in response to an anger provocation. Motiv Emot. 2011;35:14–22.
21. Giannakakis G, Grigoriadis D, Giannakaki K, Simantiraki O, Roniotis A, Tsiknakis M. Review on psychological stress detection using biosignals. IEEE Trans Affect Comput. 2022;13(1):440–60.
22. Koenig A, Omlin X, Bergmann J, Zimmerli L, Bolliger M, Müller F, et al. Controlling patient participation during robot-assisted gait training. J Neuroeng Rehabil. 2011;8(1):14.
23. Malik M. Heart rate variability Standards of measurement, physiological interpretation, and clinical use. 1996. https://academic.oup.com/eurheartj/article/17/3/354/485572
24. Novak D, Ziherl J, Olenšek A, Milavec M, Podobnik J, Mihelj M, et al. Psychophysiological responses to robotic rehabilitation tasks in stroke. IEEE Trans Neural Syst Rehabil Eng. 2010;18(4):351–61.
25. Pinna T, Edwards D. A systematic review of associations between interoception, vagal tone, and emotional regulation: potential applications for mental health, wellbeing, psychological flexibility, and chronic conditions. Front Psychol. 2020. https://doi.org/10.3389/fpsyg.2020.01792.
26. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health. 2017. https://doi.org/10.3389/fpubh.2017.00258/full.
27. Shaffer F, McCraty R, Zerr CL. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. Front Psychol. 2014;30:5.
28. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 2018;15(3):235–45.
29. Mauri M, Magagnin V, Cipresso P, Mainardi L, Brown EN, Cerutti S, et al. Psychophysiological signals associated with affective states. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10. 2010. p. 3563–6.
30. Bari DS. Psychological correlates of nonspecific electrodermal responses. J Electr Bioimpedance. 2019;10(1):65–72.
31. Caldas OI, Aviles OF, Rodriguez-Guerrero C. Effects of presence and challenge variations on emotional engagement in immersive virtual environments. IEEE Trans Neural Syst Rehabil Eng. 2020;28(5):1109–16.
32. Novak D, Mihelj M, Munih M. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. Interact Comput. 2012;24(3):154–72.
33. Koenig A, Novak D, Omlin X, Pulfer M, Perreault E, Zimmerli L, et al. Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. IEEE Trans Neural Syst Rehabil Eng. 2011;19(4):453–64.
34. Bhat SS, Dobbins C, Dey A, Sharma O. Multi-modal classification of cognitive load in a VR-based training system. In: 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). 2023. p. 503–12.
35. Gogna Y, Tiwari S, Singla R. Cognitive load with different origins: an EEG model-based explanation. NeuroQuantology. 2022;20(14):743–8. https://doi.org/10.4704/nq.2022.20.14.NQ880103.
36. Bailenson JN, Pontikakis ED, Mauss IB, Gross JJ, Jabon ME, Hutcherson CAC, et al. Real-time classification of evoked emotions using facial feature tracking and physiological responses. Int J Hum Comput Stud. 2008;66(5):303–17.
37. Gümüslü E, Erol Barkana D, Köse H. Emotion recognition using EEG and physiological data for robot-assisted rehabilitation systems. In: ICMI 2020 companion—companion publication of the 2020 international conference on multimodal interaction. Association for Computing Machinery, Inc; 2020. p. 379–87.
38. Romaniszyn-Kania P, Pollak A, Bugdol MD, Bugdol MN, Kania D, Mańka A, et al. Affective state during physiotherapy and its analysis using machine learning methods. Sensors. 2021;21(14):4853.
39. Alderfer Melissa A. and Marsac ML. Pediatric quality of life inventory (PedsQL). In: Gellman Marc D. and Turner JR, editor. Encyclopedia of behavioral medicine. New York: Springer New York; 2013. p. 1448–9. https://doi.org/10.1007/978-1-4419-1005-9_974
40. Barkham M, Mellor-Clark J, Connell J, Evans C, Evans R, Margison F. Clinical Outcomes in Routine Evaluation (CORE) – The CORE measures and system: measuring, monitoring and managing quality evaluation in the psychological therapies. In: Developing and delivering practice-based evidence. John Wiley & Sons, Ltd; 2010. p. 175–219. https://doi.org/10.1002/9780470687994.ch8
41. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q. 2003;27(3):425–78.
42. Phelan SK, Gibson BE, Wright FV. What is it like to walk with the help of a robot? Childrens perspectives on robotic gait training technology. Disabil Rehabil. 2015;37(24):2272–81.
43. Biancotto M, Guicciardi M, Pelamatti GM, Santamaria T, Zoia S, others. Movement Assessment Battery for Children–Second Edition. Standardizzazione Italiana. 2017;
44. Olson K. An examination of questionnaire evaluation by expert reviewers. Field Methods. 2010;22(4):295–318.
45. Mangione TW, Fowler FJ, Louis TA. Question characteristics and interviewer effects. J Off Stat. 1992;8(3):293–307.
46. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Methods Meas. 2007;1(1):77–89.
47. Eggink J. Krippendorff's Alpha. 2023. https://www.mathworks.com/matlabcentral/fileexchange/36016-krippendorff-s-alpha,
48. Stržinar Ž, Sanchis A, Ledezma A, Sipele O, Pregelj B, Škrjanc I. Stress detection using frequency spectrum analysis of wrist-measured electrodermal activity. Sensors. 2023;23(2):963.
49. Costantini S, Chiappini M, Malerba G, Dei C, Falivene A, Arlati S, et al. Wrist-worn sensor validation for heart rate variability and electrodermal activity detection in a stressful driving environment. Sensors (Basel). 2023;23(20):8423.
50. Berntson GG, Quigley KS, Jang JF, Boysen ST. An approach to artifact identification: application to heart period data. Psychophysiology. 1990;27(5):586–98. https://doi.org/10.1111/j.1469-8986.1990.tb01982.x.
51. Greco A, Valenza G, Lanata A, Scilingo EP, Citi L. CvxEDA: a convex optimization approach to electrodermal activity processing. IEEE Trans Biomed Eng. 2016;63(4):797–804.
52. Gjoreski M, Luštrek M, Gams M, Gjoreski H. Monitoring stress with a wrist device using context. J Biomed Inform. 2017;1(73):159–70.
53. Posada-Quintero HF, Florian JP, Orjuela-Cañón AD, Aljama-Corrales T, Charleston-Villalobos S, Chon KH. Power spectral density analysis of electrodermal activity for sympathetic function assessment. Ann Biomed Eng. 2016;44(10):3124–35.
54. Braithwaite JJ, Watson DPZ, Jones RO, Rowe MA. Guide for analysing electrodermal activity & skin conductance responses for psychological experiments. CTIT technical reports series. 2013;
55. Pearson FRSK. On lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philos Magaz J Sci. 1901;2(11):559–72.
56. Li K, Rüdiger H, Ziemssen T. Spectral analysis of heart rate variability: Time window matters. Front Neurol. 2019. https://doi.org/10.3389/fneur.2019.00545.
57. Cunningham P, Delany SJ. k-nearest neighbour classifiers: 2nd Edition (with Python examples). 2020. http://arxiv.org/abs/2004.04523
58. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
59. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2016. p. 785–94.
60. Cortes C, Vapnik V, Saitta L. Support-vector networks editor. Mach Learn. 1995. https://doi.org/10.1007/BF00994018.
61. Mcculloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. BULL MATH BIOPHYS. 1943. https://doi.org/10.1007/BF02478259.
62. Chollet F, others. Keras. GitHub; 2015. https://github.com/fchollet/keras
63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
64. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Series B Stat Methodol. 1974;36(2):111–47.
65. Cawley GC, Talbot NLC. On Over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11.
66. Gruden T, Stojmenova K, Sodnik J, Jakus G. Assessing drivers' physiological responses using consumer grade devices. Appl Sci. 2019;9(24):5353.

67. Greco A, Valenza G, Citi L, Scilingo EP. Arousal and valence recognition of affective sounds based on electrodermal activity. IEEE Sens J. 2017;17(3):716–25.
68. Li C, Rusák Z, Horváth I, Ji L. Influence of complementing a robotic upper limb rehabilitation system with video games on the engagement of the participants: a study focusing on muscle activities. Int J Rehabil Res. 2014;37(4):334–42.
69. Gokay R, Masazade E, Aydin C, Erol-Barkana D. Emotional state and cognitive load analysis using features from BVP and SC sensors. 2015.
70. Badesa FJ, Morales R, Garcia-Aracil NM, Sabater JM, Zollo L, Papaleo E, et al. Dynamic adaptive system for robot-assisted motion rehabilitation. IEEE Syst J. 2016;10(3):984–91.
71. Knaepen K, Marusic U, Crea S, Rodríguez Guerrero CD, Vitiello N, Pattyn N, et al. Psychophysiological response to cognitive workload during symmetrical, asymmetrical and dual-task walking. Hum Mov Sci. 2015;1(40):248–63.
72. Ozkul F, Barkana DE, Masazade E. Dynamic difficulty level adjustment based on score and physiological signal feedback in the robot-assisted rehabilitation system rehabroby. IEEE Robot Autom Lett. 2021;6(2):447–54.
73. Rodriguez-Guerrero C, Knaepen K, Fraile-Marinero JC, Perez-Turiel J, Gonzalez-de-Garibay V, Lefeber D. Improving challenge/skill ratio in a multimodal interface by simultaneously adapting game difficulty and haptic assistance through psychophysiological and performance feedback. Front Neurosci. 2017. https://doi.org/10.3389/fnins.2017.00242.
74. D'Mello S, Kory J. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. New York: Association for Computing Machinery; 2012. p. 31–8. (ICMI '12). https://doi.org/10.1145/2388676.2388686
75. Björklund M. Be careful with your principal components. Evolution. 2019. https://doi.org/10.1111/evo.13835.
76. Fürnkranz J. Decision tree. In: Claude S, Webb GI, editors. Encyclopedia of machine learning. Boston: Springer, US; 2010. p. 263–7.
77. Chanel G, Rebetez C, Bétrancourt M, Pun T. Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Trans Syst Man Cybern Part A Syst Hum. 2011;41(6):1052–63.
78. Amiriparian S, Freitag M, Cummins N, Shuller B. Feature selection in multimodal continuous emotion prediction. In: Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). 2017. p. 30–7.
79. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Vol. 16, Array. Elsevier B.V.; 2022.
80. Kalashami MP, Pedram MM, Sadr H. EEG feature extraction and data augmentation in emotion recognition. Comput Intell Neurosci. 2022. https://doi.org/10.1155/2022/70285172.
81. Patwardhan AS, Knapp GM. augmenting supervised emotion recognition with rule based decision model. ArXiv. 2016. https://api.semanticscholar.org/CorpusID:7197280. Accessed 24 Jan 2024.
82. Yu H. Network complexity analysis of multilayer feedforward artificial neural networks. In: Johann S, Liu Y, editors. Applications of neural networks in high assurance systems. Berlin: Springer, Berlin Heidelberg; 2010. p. 41–55 (**10.1007/978-3-642-10690-3_3**).
83. Rivas JJ, Orihuela-Espina F, Sucar LE, Palafox L, Hernández-Franco J, Bianchi-Berthouze N. Detecting affective states in virtual rehabilitation. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). 2015. p. 287–92.
84. Cakmak YO, Olcek C, Ozsoy B, Khwaounjoo P, Kiziltan G, Apaydin H, et al. Hand pronation-supination movement as a proxy for remotely monitoring gait and posture stability in Parkinson's disease. Sensors. 2022;22(5):1827.

Publisher's Note