POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Application of Artificial Intelligence techniques to predict the emotional state of patients undergoing neuromotor rehabilitation with Lokomat exoskeleton

Tesi di Laurea Magistrale in
Biomedical engineering-Ingengeria biomedica

Enrico Magliulo, 976659

Abstract: Pathologies such as Cerebral Palsy (CP) and Acquired Brain Injury (ABI) may easily lead to neuromotor impairments, widely treated nowadays with robot-assisted therapy, which has shown to be highly beneficial for neuromotor recovery. Nevertheless, the psychological response to the therapy, a key aspect especially when dealing with children, to date hasn't been deeply investigated. This Master thesis aimed to predict with Artificial Intelligence tools the emotional well-being and engagement of 42 subjects undergoing neuromotor rehabilitation with Lokomat exoskeleton at IRCCS Medea. During some of the sessions, Electrodermal Activity (EDA) and Blood Volume Pulse (BVP) were recorded by means of the Empatica E4 wristband. Time and frequency domain parameters were sorted and used as input to Machine Learning (ML) and Deep Learning (DL) models.
At first labels representing the emotional state of each session were defined. Two different predictions were implemented, one relying on patient reported outcomes and another on therapist reported one.
In the ML approach, statistical tools were adopted to perform feature selection and to lower the computational cost of the models. In the DL approach, the dataset was upsampled by windowing the signals, in order to make the training more consistent. The final part was the implementation of the models and the evaluation of their metrics.
Results reveal that models are able to recognize different emotional states, suggesting the capability of AI to study the psychological response to therapy starting from recorded biosignals. ML models provided better metrics rather than DL ones, due to the low amount of available data.
In ML, *Support Vector Machine* (SVM) made better predictions than *K-Nearest Neighbors* (KNN), probably due to the high overfitting risk related to KNN use with unbalanced datasets. *Neural Networks* showed the best DL predictions.
A possible future development may be the implementation of a real time prediction system, that could customize therapy according to patient specific needing and overcome the communication limits related to subjects' young age or neurological impairments.

Advisor:
Prof. Emilia Ambrosini
Co-advisors:
Ing.Emilia Biffi
Ing.Simone Costantini
Ing.Anna Falivene
Academic year:
2022/2023

Key-words: Lokomat, Empatica E4, HRV, EDA, Machine Learning, Deep Learning, emotion recognition

Contents

1 Introduction **2**
- 1.1 Neuromotor impairments in children: cerebral palsy and acquired brain injury . . . 2
- 1.2 Neurorehabilitation in children with neuromotor impairments . . . 3
- 1.3 Mental wellbeing during rehabilitation . . . 3
- 1.4 Electrodermal Activity and Heart Rate Variability . . . 4
- 1.5 Deep Learning and Machine Learning application in emotion recognition . . . 5
- 1.6 Present work objective . . . 6

2 Materials and Methods **6**
- 2.1 Test subjects and acquisition protocol . . . 6
- 2.2 Questionnaires . . . 7
  - 2.2.1 Patient reported questionnaires . . . 7
  - 2.2.2 Therapist reported questionnaire . . . 7
- 2.3 Labelling strategy . . . 8
  - 2.3.1 Patient reported outcome . . . 8
  - 2.3.2 Therapist reported outcome . . . 8
- 2.4 HRV and EDA parameters . . . 9
- 2.5 Dataset structure and preprocessing . . . 10
- 2.6 Proposed models . . . 11
  - 2.6.1 ML models . . . 12
  - 2.6.2 DL models . . . 12
- 2.7 Metrics for evaluation . . . 15

3 Results and Discussion **15**
- 3.1 Labelling strategies . . . 15
- 3.2 Feature selection . . . 16
- 3.3 Data upsampling . . . 17
- 3.4 ML results . . . 17
  - 3.4.1 Therapist reported outcome results . . . 18
  - 3.4.2 Patient reported outcome results . . . 20
- 3.5 DL results . . . 23
  - 3.5.1 Therapist reported outcome results . . . 23
  - 3.5.2 Patient reported outcome results . . . 24

4 Conclusions **25**

1. Introduction

1.1. Neuromotor impairments in children: cerebral palsy and acquired brain injury

Cerebral palsy (CP), a neuromotor disorder that affects the development of movement, muscle tone and posture, and hence the further development of the brain during children's growth, at the present day has an incidence on new births which ranges between 1.5 and 3 per 1000 births . It is typically diagnosed by means of Magnetic Resonance Imaging (MRI) scanning [1].
CP causes may be distinguished according to 3 different birth phases. In the *prenatal* one the main risk factor is congenital brain malformation; in the *perinatal* one, CP is mainly associated to the emergencies during the labour (e.g. antepartum haemorrhage); in the *postnatal* one problems such as infections or injuries may led to develop CP [2]. In case of absence of the typical risk factors associated to this disorder, the principal way to assess the presence of CP is an abnormal neuromotor development; Patel *et al.* list a series of key motor steps that a healthy baby should face with no problems and that, on the contrary, represent a serious problem for CP affected babies [1]. The *Gross Motor Function Classification System* (GMFCS) represents with a scale the degree of impact that CP has on children (aged between 6 and 12) and on their daily life. Level I of the scale is related to patients that are able to develop all the gross motor functions, such as climbing stairs, walking or jumping, but with reduced balance and speed; level V is associated to children transported in a wheelchair in all settings [3].

Another pathology that affects neuromotor capability is the Acquired Brain Injury (ABI), that consists in the collection of all the cerebral damages happened because of external causes (such as traumas, aneurysms, hypoxia or postsurgical complications). This disturb may bring several impairments to normal motor functions, such as spasticity, a reduced Range of Motion (ROM), balance problems or muscolar weakness [4]. All over the world, pediatric ABI has an annual incidence that oscillates between between 47 and 280 per 100.000 children, according to the geographic area of interest [5].

1.2. Neurorehabilitation in children with neuromotor impairments

Both CP and ABI require long term rehabilitation, as they are chronic pathologies. The development of modern technologies has spread, in the field of rehabilitation, the adoption of the so called *robot-assisted therapy* and also of tools such as the Virtual Reality (VR).
Nowadays several tools have been implemented for performing robot-assisted therapy. An example is the Lokomat exoskeleton, which may be used for both upper limb and lower limb rehabilitation. This system has revealed to be much more effective in the gait training, improving parameters such as legs strength, gait velocity and functional mobility in a faster and more significant way with respect to conventional treatment [6].
Falzarano *et al.* conducted a literature research to establish the diffusion of robot-assisted therapy in pediatric neuromotor rehabilitation and its effectiveness, examining 14 different devices. The research concludes that, due to the fact that most of robotic systems are designed for adults and that children therapy is usually done into a playful environment, further in-dept analysis should be done in this field [7].
VR-based treatments (also called *immersive therapy*) have shown to achieve the same level of efficacy with respect to conventional techniques. Also, they embed some advantages, such as the flexibility of the environments that can be created, which is a trigger for patients' engagement (especially child-aged ones) into the task, and the capability of these systems to adapt the scenario to patient specific needs, which makes this a custom based method [8].

1.3. Mental wellbeing during rehabilitation

There is evidence that the studies concerning the psychophysical involvement into a rehabilitation treatment to date have focused on the investigation of biomechanical engagement of the patient, which can be easily quantified calculating the values of the reaction forces and the torques by means of force sensors [9]. For this reason, the involvement of these measurements is becoming common practice in modern rehabilitation treatment, especially for gait analysis.
Also mental engagement analysis has shown to play a key role in rehabilitation, since practitioners consider that a motivated patient achieves better performance and faster recover with respect to non-motivated one [10] . Nevertheless the amount of papers related is very small, since there is not a clear definition of 'motivation' and the crucial step of emotions quantification must be faced [11]. King *et al.* developed a quantitative measure, called *Pediatric Rehabilitation Intervention Measure of Engagement-Observation* (PRIME-O), to express the level of engagement in pediatric rehabilitation starting from self-assessed questionnaires answers [12].
The simplest way to measure the level of mental wellbeing of a patient during rehabilitation treatment is filling in some dedicated questionnaires aimed to define in a numerical way the emotional state of the subject. The most used tool to do so is the Likert scale, which consists in a behavioral measurement that assigns a score, in a graded scale, to a set of items (that in this application are the questions provided to the patients) [13]. There are several examples of questionnaires adopting a Likert scale approach to classify stress and mental engagement; one of the most used is the *Self-Assessment Manikin (SAM)*, whose scale ranges between 1 (negative) and 9 (positive) [14]. This tool rates 3 different aspects of a person's psychophysiological reaction to an external event, such as a physical exercise:

- The level of valence (ranging from unpleasant to pleasant)
- The level of arousal (ranging from calm to excited)
- The level of dominance (ranging from independent to dependent)

A key advantage of the SAM is the adoption of graphics to describe the scores, which facilitate its comprehension. Koenig *et al.* submitted the SAM to patients during gait rehabilitation treatment assisted with the Lokomat exoskeleton in order to estimate patients' psychological state and engagement into the therapy [15].
Once filled, questionnaires need to be interpreted in order to agglomerate the several scores into a single one and to classify the emotional state of the patient. An example of tools implemented for this scope was proposed by Russel [16] and is reported in Fig. 1. It places the several emotional states in a circular shape; the four dials of the graph are associated to as many indicators, two of which are related to positive emotions while the remaining ones to negative emotions.

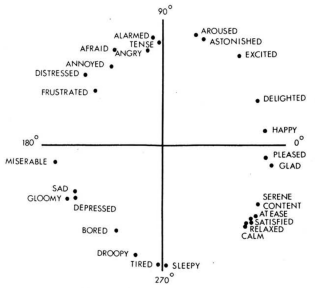


Figure 1: the Russel circumphlex model [16].

Another useful instrument, which was derived from Russell's one and shown in Fig.2, is the Valence-Arousal (VA) model, that defines, by means of plotting the valence on the x-axis and the arousal on the y-axis, 4 main emotional state areas [17]:

- **High Valence/High Arousal (HVHA)**
- **High Valence/Low Arousal (HVLA)**
- **Low Valence/High Arousal (LVHA)**
- **Low Valence/Low Arousal (LVLA)**

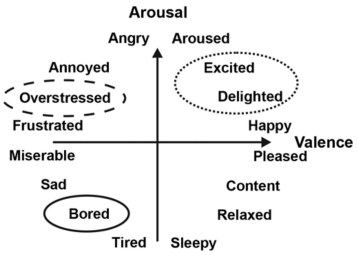


Figure 2: the valence arousal model [15].

However, despite their large use, psychometric questionnaires present some limitations: the responses may vary from subject to subject, and are also strictly linked to the capability of the patient to fully understand the questions and to his/her willingness to respond accurately [18].

1.4. Electrodermal Activity and Heart Rate Variability

Electrodermal Activity (EDA) and Heart Rate Variability (HRV) are two among the physiological signals which have been used in emotional state assessment.
The EDA can be defined as the variation of electrical conductance of the skin, and is derived by the sum of two components [19]:

- A *tonic* component, also defined *Skin Conductance Level (SCL)*, that represents the basic activity level of skin conductance. It is an assessment regarding subject's both physical and psychological state, and can be considered as a generic measurement of the activation level of the Autonomous Nervous System (ANS). In a recent study, Dawson *et al.* demonstrated that SCL tends to rise up, with respect to its value recorder in rest conditions, if the patient is experiencing a demanding physical task [20].
- A *phasic* component, also defined *Skin Conductance Response (SCR)*, that represents rapid and transient changes (in a 1-5 seconds time window) in the electrical activity of the skin. This component is often associated with emotionally significant events or stimuli that induce an arousal or stress response in the subject. It may be *event-related* (whether the stimulus is represented by a specific event or condition) or *tonic stimuli-related* (due to a variation of the enviromental conditions or of the stress level), and can be considered as the specific response of the Sympathetic Nervous System (SNN).

Fig. 3 represents the EDA signal, highlighting specifically the SCR component.

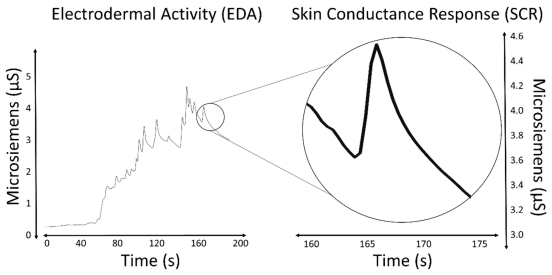


Figure 3: EDA signal with SCR component highlighted [19].

The HRV is a phenomenon that relies on the variability of the time that interleaves two consecutive heart beats [21]. The two main factors that are responsible for heart rate fluctuations are the respiration (that affects heart rate according to the respiratory rate) and low frequency (0.1 Hz) oscillations related to the so called *Mayer waves* (variations in arterial blood pressure) [22]. Generally the HRV tends to lower within the age, and its behavior is an indicator of the type of nervous system activated in a specific moment. In fact, an increasing HRV with respect to the baseline value of a subject states an activation of the Vagal System, while the opposite case assesses SNN activation [21].
Conventionally, EDA signal is recorded by means of a couple of skin electrodes, usually placed on body extremities, such as hands or feet, while HRV is derived from the ECG recording. Although it is not the gold standard, they may also be sampled thanks to the medical-grade device Empatica 4 wrist-band (Empatica®, Milan, Italy) [23] .
Both these signals may be used to predict the emotional state of a patient undergoing rehabilitation treatment session, if given as input to Artificial Intelligence (AI) models (see section 1.5).

1.5. Deep Learning and Machine Learning application in emotion recognition

Machine Learning (ML) and Deep Learning (DL) are two branches of the wide field of the AI. The first one represented the first attempt to use a machine in order to exploit a specific task [24]; the second one is a further development of the standard ML techniques [25].
The problems that may be faced by AI techniques consist in taking as input a set of data (which may be numerical values or Booleans), defined as a tensor of dimensionality larger or equal to 2, and finding a way to separate them according to some characteristics. The mathematical tool used for this purpose is called *separation hyperplane*.
The task of interest may be related to *supervised learning* or to *unsupervised learning*. The supervised learning is a paradigm of the AI whose main goal is performing the prediction of a numerical value (called *target*) related to every sample of a dataset. Two examples of this paradigm are the *classification* and the *regression*. On the other hand, in the unsupervised learning there is no target; a classical example among these kind of tasks is the *clustering*, which aims to group together samples united by similar characteristics [26].
One of the key aspects of AI is the learning process, which is performed by assigning to every sample a randomly initialized weight and updating it according to a loss function optimization. The *loss function* is a tool that provides an assessment of the goodness of the prediction, and is computed in different ways, according to the task the algorithm is facing.
The first step to deal with consists in managing dataset's variables so that the algorithm is able to recognize which of them are the most significant for the prediction; this is known as *feature extraction*. In ML, this process is performed manually by the user (*hand-crafted feature extraction*), while in DL it is performed by the algorithm itself (*data-driven feature extraction*) by the feature extractor [25]. Furthermore DL models are able to deal with very hefty dataset, rather than ML ones, and with unstrctured data, such as images, texts and signals. The first DL architecture implemented is the *Convolutional Neural Network* (CNN), which is a further development of the standard *Multi-Layer Perceptron* (MLP) [27]. In the CNN, the neurons combine the input information in a non-linear way, thanks to the so called *activation functions*. A CNN is made of two parts (see Fig. 4):

- The *feature extractor,* whose layers (which are called *hidden layers*) are made of neurons that process the information they receive and pass them through an activation function. This part ends with a *flattening layer*, that reduces the input volume to a vector (which is then fed to the second part). The typical activation functions are the sigmoid: $y(x) = 1/1 + e^{-x}$, the ReLu: $y(x) = \max(0, x)$ and the tanh: $y(x) = e^{x} - e^{-x}/e^{-x} + e^{x}$
- A standard MLP aimed to compute the task of interest

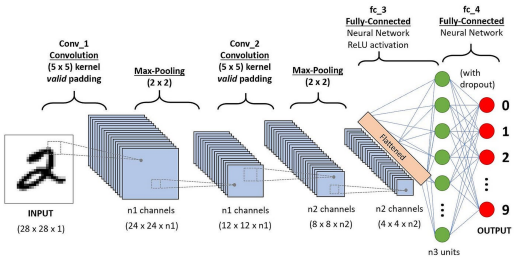


Figure 4: graphical representation of a typical CNN architecture [28]

Emotion recognition is defined by Kosti *et al.* as "understanding what a person is experiencing from her frame of reference" [29]. This task consists in predicting the emotional state of a subject while developing a specific activity (such as singing, speaking, performing a motor function). In the last years, there has been an increasingly focus on emotion recognition in the robot-assisted rehabilitation field, since the prediction of the mental state is useful to assess the level of comfortability of the treatment and hence to customize it according to the patient specific needings [30].
Recently AI models have shown to provide high performances in predicting the emotional state of a person, improving as a consequence the opportunities for the automatization of emotion recognition [31].
The majority of the related works adopted benchmark datasets and treated the emotion recognition as a supervised learning task [32–34], even though some attempts to perform it in an unsupervised way were done [35]. Thus input signals need to be provided of a target. Al Machot *et al.* [33] derived a 4 labels classification problem starting from the MAHNOB dataset, that consists in a collection of EDA signals coming from 30 healthy subjects (aged between 19 and 40) that self-evaluated their emotional state by means of the SAM. The level of valence was rated as 'negative' whether the relative score in SAM was in the range (1-5), 'positive' in the range (6-9); the arousal was labelled as 'passive' in the range (1-5), 'active' otherwise. This brought to the definition of the same 4 classes defined in [17]. The dataset was built extracting 12 time and frequency domain parameters of the EDA, related to both the components of the signal (see section 1.4), and was fed as input to a *K-Nearest Neighbors* (KNN) classifier. Dalmeida *et al.* [32] analyzed several HRV extracted parameters related to 27 recordings, taken from the PhysioNet database, in order to assess whether the patients were stressed or not. They adopted Galvanic Skin Responses (GSR) values to derive the labels: if the GSR value of a subject fell below the median one, the patient was labelled as 'not stressed', while 'stressed' data were the ones having a GSR higher than the median one.
An attempt to automatically recognize 4 different emotional states (sadness, anger, surprise, happiness) in children robot-based rehabilitation was done by Nagae *et al.*, that validated a Human-to-Robot-Interface (HRI) using a *Support Vector Machine* (SVM) classifier and EDA extracted features and achieved a 38,6% recognition rate [36].

1.6. Present work objective

The purpose of this thesis is to predict the emotional state of patients undergoing several neuromotor rehabilitation sessions with Lokomat exoskeleton starting from EDA and HRV signals recordings. The prediction was done by means of the implementation, in a methodological approach, of both ML and DL models and their training with a set of time and frequency domain parameters (explained in detail in section 2.4) extracted from the biosignals previously quoted. Two different labelling strategies were adopted, one related to self-reported questionnaires and another one to therapist-reported questionnaire (see section 2.6), and the related algorithms performances were compared. The results were evaluated by means of several metrics, described in detail in section 2.7.

2. Materials and Methods

2.1. Test subjects and acquisition protocol

This master thesis project enrolled 42 subjects (28 males and 14 females), aged between 5 and 68, who underwent several sessions of neuromotor rehabilitation with Lokomat in the Scientific Institute I.R.C.C.S. "E. Medea".
Lokomat is an active exoskeleton adopted in gait robotic rehabilitation that allows the performance of treatment in an engaging and playful enviroment, which is the main reason why it is used in pediatric rehabilitation [6].
Patients performed between 15 and 20 Lokomat sessions, whose time duration ranged between 13 and 41

minutes, according to the indications of the clinicians. Data included in this work are related to 2-3 sessions during which patients wore a wearable wrist sensor, the Empatica E4. This device was adopted in order to record the EDA (with a sampling frequency of 4 Hz), whose tonic and phasic components were successively derived, and the Blood Volume Pulse (BVP) (with a sampling frequency of 64 Hz) from which the HRV was sorted [37]. The wristband was worn by the subjects in 3 different phases of the session: before the treatment, during the treatment and after the treatment. Patients, at the beginning and at the end of each session, were asked to fill in VAS (*visual analogue scale*)-type questionnaires [38], which are described in detail in section 2.2.1, related to their emotional state. During the treatment a questionnaire regarding the level of engagement of the subjects into the task (see section 2.2.2) was compiled by the therapist.
At the end of the session, the participants were asked to fill a further questionnaire, which consisted in a comic, inspired from the one used by Phelan *et al.*, that represented graphically the patient making the physical exercise [39]. In the comic, subjects were allowed to write some sentences and also some emoticons describing their feeling about the therapy. The specific comic adopted for the purpose is reported in Fig.5.

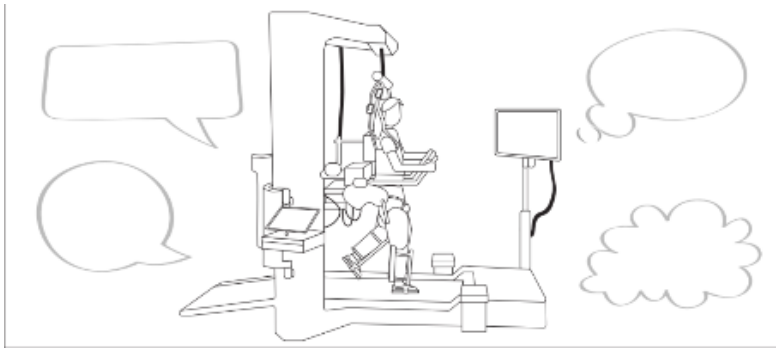


Figure 5: comic questionnaire

Each of the subjects was given an ID code in order to keep the anonymization. The IRCCS Medea ethic committee approved the protocol on the 27th of January 2022 (Prot. N. 02/22 - CE). The patients, if adults, or their guardians signed a consent. All data were pseudonymized. The present work is part of this study.

2.2. Questionnaires

2.2.1 Patient reported questionnaires

The questionnaires compiled by the patients before and after the session were made of 10 items each. The first 6 of them were the same for both the pre and the post treatment phases, and consisted in the quantification of several feelings (worry, happiness, sadness, anger, fear, boredom) by means of a 3 DoF Likert scale, with the following semantic: '1' means 'a few', '2' means 'enough', '3' means 'a lot'.
Indeed the last 4 items, that differed between the two phases, concerned the trust the patients had in Lokomat and in their lower limb functionalities recovery.

2.2.2 Therapist reported questionnaire

During the progress of the exercise, therapist filled in a questionnaire, whose items were represented by couples of boundary emotional states which are opposite each other (e.g. passive-active). The therapist had to give a score in every item, ranging from -3 to 3, to assess where the patient's state was located. The list was made of 12 items overall, which are reported below.

1. Passive-active
2. Fearful-master of the situation
3. Anxious-relaxed
4. Impulsive-thoughtful
5. Distractable - Attentive (focused)
6. Hyperactive - Quiet
7. Underestimate his/her abilities - Overestimate his/her abilities
8. He/she lacks persistence - He/she is persistent
9. Worried about failure - Does not worry about failure
10. Unable to derive satisfaction from success- Able to derive satisfaction from success
11. Manages emotions negatively - Manages emotions positively

12. Does not actively seek information to learn and update - Actively seeks information to learn and update

2.3. Labelling strategy

The key step to set up the classification task was the choice of the labelling strategy, which led to the definition of the target vector quoted in section 1.5. Due to the different structure of the patient reported and therapist reported questionnaires, two different labelling strategies were adopted. This brought to two different scenarios of prediction: one related to the mental wellbeing (patient reported outcome) and another one related to the mental engagement (therapist reported outcome).

2.3.1 Patient reported outcome

In order to derive the labels, the first attempt consisted in a literature research to check similarities between the *ad hoc* questionnaires used in this work and clinical validated ones. Several psychometric tools and emotional indices were considered:

1. the *State-Trait Anxiety Inventory* (STAI, [40]), a 20-items indicator of individual tendency to experience anxiety, and its short version (Short-STAI, [41]) that has 6 items only;
2. the *Hamilton Anxiety Rating Scale* (HAM-A) [42];
3. the *Beck Depression Inventory* (BDI) [43];
4. the *Montgomery-Asberg Depression Rating Scale* (MADRS) [44];
5. the *Profile of Mood States* (POMS) [45].

However, all of them were finally discarded, since they presented several similarities with the questionnaires, but none of them was 100% correspondent, hence the metrics used to condensate and evaluate the scores of these tools would not have been validated procedures if applied to this context. So the chosen strategy was the so called *double-blinded expert evaluation* [46]. This method required a group of 4 experts, each one providing a personal score to the answers given by the patients at every session; the evaluation was given considering all the answers to the pre/post questionnaires, since they were very similar. In addiction to the 10 items previously quoted in section 2.2.1, a new one was created. This 11^{th} item contained the polarity scores derived from the Sentiment Analysis done on the comics filled by the patients at the end of the treatment. The Sentiment Analysis is the study of the emotional polarity of a text, and belongs to the category of the *Natural Language Processing* (NLP) [47]. The final goal of this task is to express with a numerical score the polarity of the sentiments expressed in the text. This score is comprised between 1 and -1: values closer to 1 indicate a positive polarity, and vice versa for the negative ones. Recently several Sentiment Analysis tools were implemented and validated; the one adopted in this work is called *TextBlob* 0.16.0 (released in April 2020 [48]). This algorithm is able to encode the words that compose a text and understand their meaning, and its training is performed by labelling the sentences of the text under analysis (setting up a classification problem of the emotional polarity expressed by the text). TextBlob was chosen since the literature shows that it is widely used in the Sentiment Analysis field [47].

The assessments on the 11 items were expressed by means of a number, ranging from -3 to 3 (in order to maintain the coherence with the ranking done in the therapist reported questionnaire, see section 2.2.2), that was aimed to be an indicator of the emotional state of the subject. In the scale, values tending to the lower limit of the range embedded negative emotions, while values closer to the upper limit expressed positive emotions. Then the compound of the 4 evaluations was examined by an additional and external expert, in order to provide the final evaluation for every session. Then the scores were converted into the labels by dividing the overall range in 3 different areas, each one representing a specific emotional polarity:

1. scores -2 and -3 were grouped into the negative emotion area, and they were assigned the label '0';
2. scores -1, 0 and 1 were grouped into the neutral emotion area, and they were assigned the label '1';
3. scores 2 and 3 were grouped into the positive emotion area, and they were assigned the label '2';

In addiction, in this strategy the inter-rater agreement was investigated by means of the Krippendorf α coefficient calculation. The coefficient was derived by a Matlab file exchange function; although there is not an objective threshold, a value of α higher or equal than 0.8 is generally considered an assessment of high agreement [49].

2.3.2 Therapist reported outcome

A different strategy was adopted for the treatment phase in order to derive the labels. For this purpose, a model that relies on the one proposed by Russel [16] was built up with 8 of the 12 items of the checklist. Also an *Exploratory Factorial Analysis* (EFA) was applied in order to check whether the selected items were suitable for the circumphlex model. The EFA is a statistical tool, very commonly used in scale developing tasks (hence in any kind of quantification procedure) associated to a large number of features, that starting from a group of measured variables derives a set of new items (named *latent constructs*) which cannot be observed directly by

the researcher. The latent constructs are aimed to assess the presence of patterns underlying the original data; the choice of considering only some of them allows the researcher to lighten the dataset and to facilitate the study [50].
Labels were sorted taking inspiration from the procedure shown by Koenig *et al.* [15]. Once inserted into the circumphlex model, items were grouped so that each group corresponded to a quadrant and identified a specific emotional state. The 4 areas are reported in Fig. 6.

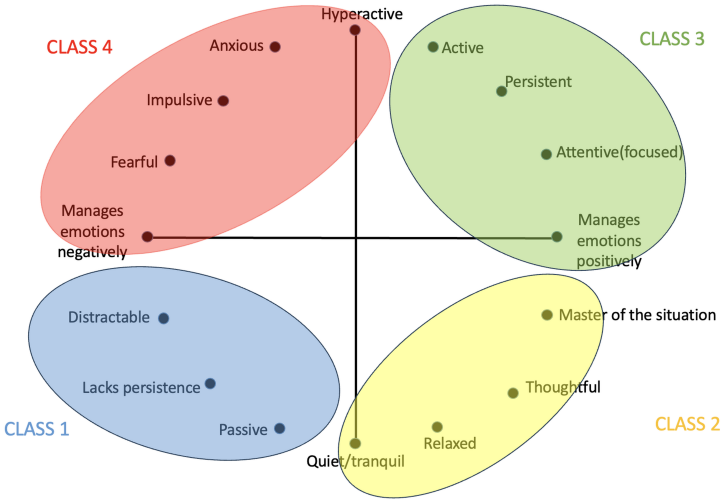


Figure 6: classes identification in the Russel model

Each class embedded a set of items of the model, and corresponded to a specific psychological meaning:

- Class 1 was related to the third quadrant of the plot, and was an expression of the low valence and low arousal samples (passive negative);
- Class 2 was related to the fourth quadrant of the plot, and was an expression of the high valence and low arousal samples (passive positive);
- Class 3 was related to the first quadrant of the plot, and was an expression of the high valence and high arousal samples (active positive);
- Class 4 was related to the second quadrant of the plot, and was an expression of the low valence and high arousal samples (active negative).

Labels were assigned to the samples by defining 4 vectors (one for each quadrant of the plot), each one having as modulus the score given to the sample in every class (calculated as the mean of the scores obtained in the items related to each class) and as direction the bisector of the quadrants. Then the vectorial sum was done, and the direction of the resultant vector stated the label of the sample.

2.4. HRV and EDA parameters

The dataset contained 35 physiological parameters, derived by means of the algorithms described by Costantini *et al.* [37]; 15 of them were related to the HRV, while the remaining 20 to the EDA. Table 1 and 2 present all the physiological parameters adopted in this work.

Mean IBI (ms)	SDNN (ms)	RMSSD (ms)	Mean HR (bpm)	SDSD (ms)
pNN50	HRV Triangular index (-)	TINN (ms)	HRV skewness	HRV kurtosis
Normalized LF Spectrum IBI (%)	Normalized HF Spectrum IBI (%)	Symphathetic modulation index	Vagal modulation index	Symphatovagal balance index

Table 1: parameters derived from the HRV signal.

Mean EDA Tonic (uS)	Standard Deviation EDA Tonic (uS)	IQR EDA Tonic (uS)	Skewness EDA Tonic (uS)	Kurtosis EDA Tonic (uS)
Max Upspeed EDA Tonic (uS/min)	Max Downspeed EDA Tonic (uS/min)	NS.EDRs (/s)	Mean EDA phasic Peak Ampl (uS)	Std EDA phasic Peak Ampl (uS)
Normalized AUC (uS*s)	Mean rise time (s)	Mean EDA phasic Peak-to-Peak distance (s)	Std EDA phasic Peak-to-Peak distance (s)	Normalized VLF spectrum EDA (%)
Normalized LF spectrum EDA (%)	Normalized HF1 spectrum EDA (%)	Normalized HF2 spectrum EDA (%)	Normalized VHF spectrum EDA (%)	II order regr a coeff

Table 2: parameters derived from the EDA signal.

2.5. Dataset structure and preprocessing

The dataset was composed of a set of rows, each one representing a session of a specific patient, and 37 columns overall, obtained adding to the 35 physiological features described in section 2.4 two variables containing age of the patient, a positive discrete variable, and sex, a categorical one represented with a binary vector whose value '0' corresponded to 'male' and '1' to 'female'.
The post-treatment recordings were used for standardization purposes. The train-test split was performed with test size= 25%; in the DL approach, also the train-validation split was done, with validation size= 15%.
Preprocessing procedure was differentiated in the ML and DL methods. Fig.7 summarizes all the steps made in the two approaches before feeding the dataset to the algorithms.

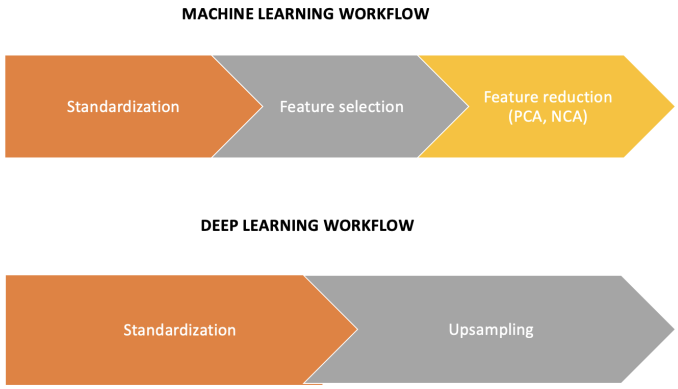


Figure 7: ML and DL approaches preprocessing workflows

One of the steps common to both the approaches was data standardization, which was made by subtracting the post-treatment physiological data (considered as the baseline) to the Lokomat ones. The resulting standardized dataset represented the variation of the parameters calculated during the treatment phase with respect to the baseline recording.
In the ML approach, feature selection was performed, in order to reduce the size of the dataset and, as a consequence, the computational cost of the models. At first the level of correlation of the variables was studied. The choice of the type of correlation coefficient to apply was taken by means of the calculation of the skewness index and the kurtosis index for all the features. The Spearman correlation analysis was chosen, because of its suitability to non-normally distributed features [51] (see section 3.1). The level of correlation was considered relevant whether the absolute value of the Spearman coefficient was higher than 0.8.
A further feature reduction was done by means of two commonly used techniques. The first one was the *Principal Components Analysis* (PCA), that consists in combining the features of a dataset in order to build up a new set of variables, called *Principal Components* (PC), such that all the PC are statistically uncorrelated each other and explain a certain fraction of the total variance (*individual explained variance*) of the data [52]. The sum of

the individual explained variances is called *cumulative variance*, and its value is essential in the choice of the number of PC to consider. In this thesis, two threshold values of cumulative variance were chosen, according to the related works available in literature:

1. Cumulative variance= 80% [53], defined as PCA_{80};
2. Cumulative variance= 90% [54], defined as PCA_{90}.

The other tool chosen for this purpose was the *Neighborhood Components Analysis* (NCA). This technique projects the original data into a lower dimensionality space, so that similar samples (similarity is measured by means of the Euclidean distance of the observations) are closer each other. The output of this analysis is a new dataset representing the optimal projection of original data into the lower dimensionality space [55]. The choice of the projections to use still relies on the cumulative variance.

In the DL approach, data augmentation was performed in order to increase the dataset size and to improve the feature extraction process of the models. The goal of this technique is to bootstrap samples belonging to the minority class until data reach the same cardinality in all the classes. At first synthetic data generators were considered, such as the *Synthetic Minority Oversampling Technique* (SMOTE) and the *Adaptive Synthetic Sampling* (ADASYN) [56]. The SMOTE is an algorithm that generates new samples interpolating the original ones belonging to the minority class of the dataset. ADASYN is a further implementation of SMOTE which focuses the synthetic data generation around the samples that are most frequently misclassified by the model. Askari *et al.* [57] adopted these tools in order to augment EDA and HRV signals in a classification task. However since in this work numerical parameters, whose distribution revealed to be mostly random, are used instead of the signals themselves, these algorithms were discarded. Current literature states that the most widespread way of balancing random datasets is the replication of samples belonging to the minority class, so that its size is increased [58]. This method was initially adopted, but since it added no further information to the dataset apart from shape augmentation, it was finally rejected.

Finally the chosen strategy consisted in performing data upsampling by segmenting the original signals, in order to obtain several windows that were considered themselves a signal (and, as a consequence, a session) and from which the parameters were calculated. The length of the windows was derived by fine tuning with a statistical analysis a range of window length values, defined combining the state of the art regarding EDA and HRV windowing with the duration of the recordings. Chaspari *et al.* sliced 5 minutes EDA recordings using 3 different window sizes: 10, 20 and 30 seconds, meaning the 3.33%, 6.66% and 10% of the original data length respectively [59]. Bigger *et al.* assessed that a time window ranging between 2 and 15 minutes is an accurate approximation of a 24 hours HRV recording and hence a good predictor of the whole signal itself [60].

In this work, a series of segmentation scenarios was investigated by means of the calculation, for every time length, of the *Multivariate Coefficient of Variation* (MCV), according to the formulation proposed by Albert&Zhang (2010) [61]. The MCV is a further development of the standard coefficient of variation that accounts for the multivariate characteristic of the data (hence for the presence of multiple variables); it is an indicator of the deviation of all the features with respect to their mean value. The Albert&Zhang formulation of the MCV is reported in equation 1:

$$\gamma(AZ) = \sqrt{\frac{\mu^t \cdot \Sigma \cdot \mu}{(\mu^t \cdot \mu)^2}} \tag{1}$$

where μ is the mean vector and Σ the covariance matrix of the dataset.

The size of the windows ranged between 5 and 12 minutes; for each of them, 0% overlapping [62] and 50% overlapping [63] [64] were considered. Before the calculation of the MCV, a standardization of the dataset was performed, since the parameters had different ranges of values. The *robust scaler* was chosen for this purpose, since it relies on the *Inter-Quartile Range (IQR)* and hence, differently from the *minmax*, is not influenced by the presence of outliers [65]. Robust scaler is defined as:

$$x\_scaled = \frac{x - median(x)}{IQR(x)} \tag{2}$$

where x_scaled is the standardized variable, x is the original one, $median(x)$ is the central value of the variable and $IQR(x)$ is its inter-quartile range, that contains the second and the third quartiles of the samples. The output of the analysis was, for every scenario, a vector of MCV values, each one related to a sample, that were condensed into a single one by calculating the mean MCV. The windowing scenario that provided the minimum mean MCV was finally chosen.

2.6. Proposed models

Starting from the dataset previously described and the labels obtained with both the strategies, the goal was to identify the ML and DL models that best predicted patients' emotional state. All the algorithms were

implemented in Python language, using the Google Colaboratory application.

2.6.1 ML models

In the ML approach, two families of models, that were used in emotion recognition based on EDA and HRV signals, were adopted for the scope: the *Support Vector Machine* (SVM) and the *K-Nearest Neighbors* (KNN). SVM is a method whose working principle consists in projecting a dataset into a different dimensionality space by means of a mathematical function (named *kernel*). This way the algorithm distinguishes data more easily, according to specific characteristics (e.g. the label in a classification problem), by means of a separation hyperplane; the best separation is provided by the hyperplane that has the maximum distance with respect to the closest sample of every class. The number of misclassifications equals the number of samples that fall beyond the bound of the hyperplane. In this task, 3 models belonging to SVM family were used; they were all implemented in the scikit-learn 1.3.1 Python library [66]:

- SVC
- NuSVC
- LinearSVC

The SVC is the standard Support Vector Classifier; it is equipped with several hyperparameters that were fine tuned by means of the *gridsearch* tool, which sets a grid of ranges of the hyperparameters values [66]. The tuning is then performed by means of the *k-fold cross-validation*, hence splitting the dataset into k subsets and using the k^{th} one as test set and the other k-1 ones as training set; in this case, k was set as its default value, which is 5. Table 3 reports the hyperparameters tuned, their meaning and the ranges of values adopted.

PARAMETER	MEANING	VALUES
C	Regularization parameter	0.001,0.01, 0.1,1.0
Kernel	Kernel type	'rbf', 'linear', 'poly','sigmoid'
Degree	Degree of 'poly' kernel	2, 3, 4, 5
γ	Kernel coefficient for 'rbf', 'poly' and 'sigmoid'	'auto', 'scale'

Table 3: hyperparameters tuning for the SVC model.

The *regularization parameter* expresses the generalization capability of the algorithm; the lower it is, the stronger the regularization of the model, hence the lower the risk of overfitting. The γ parameter is computed as $1/\#features$ if equal to 'auto' and as $1/(\#features * Var)$ if equal to 'scale', where Var is the variance of the features.
The NuSVC consists in a SVC model to which a parameter ν (that must be within the (0,1] interval) is added, whose aim is to bound the classification error made in the training phase. As a consequence, the lower ν, the higher the risk of overifitting; this parameter was fine tuned within the following range of values: $0.1, 0.2, 0.3, 0.4, 0.5$.
The LinearSVC is a SVC whose kernel is fixed to 'linear'.
The KNN is a model that performs its prediction relying on the k nearest samples to the one of interest; the degree of proximity of the samples is determined with the Euclidean distance. In classification problems, the algorithm computes the prediction applying the majority voting on the labels related to the k nearest neighbors of the sample under analysis [67]. Also in this case, a *gridsearch* was set up to fine tune the k value; the range of values was bounded between 5 and half of the training set size, with $step = 5$.

2.6.2 DL models

In the DL approach, 3 main families of architectures were considered during models implementation:

- The *Dense Neural Network* (DNN);
- The *Convolutional Neural Network* (CNN);
- The *Long-Short-Term Memory* (LSTM)

The learning process in all the architectures takes place through several iterations (also called *epochs*) and is stated by the so called *gradient descent theory* (3):

$$w(k+1) = w(k) - \eta \cdot \frac{\partial E}{\partial w} \tag{3}$$

where η is the *learning rate,* a hyperparameter that states the strength of the learning process at every iteration, and $\partial E/\partial w$ is the gradient of the loss function computed at the k^{th} step [68]. The chosen loss function was the

Categorical Cross-Entropy (CCE) (4), typically adopted in classification problems and reported in equation 4:

$$CCE = -\sum(y(i) \cdot \log(y'(i)) \tag{4}$$

where $y(i)$ represents the real label and $y'(i)$ the predicted one.
The DNN can be considered as the simplest form of DL algorithms; its feature extractor is characterized by dense layers only [69]. A *dense layer* is composed by neurons that receive as input information coming from all the neurons belonging to the previous layer. Fig. 8 shows the typical architecture of a DNN.

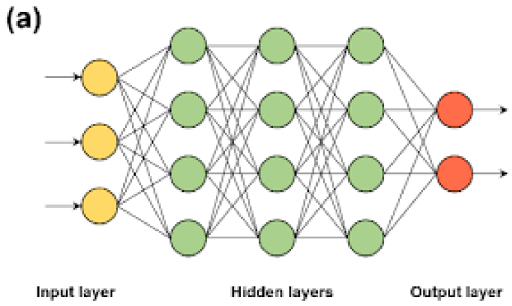


Figure 8: typical architecture of a DNN [69]

The architecture of the DNN hereby used fully matches the one adopted by Eren *et al.* [34], and is summarized in Table 10.

Layer name	Output shape	Number of parameters
input	(None, 37)	0
dense1	(None, 18)	684
dense2	(None, 14)	266
dense3	(None, 10)	150
dense4	(None, 8)	88
dense5	(None, 4)	36
output	(None, 3)	15

Table 4: summary of the proposed DNN.

In all the hidden layers of the model, the Heun kernel was set. At the end there is the output layer, a dense with 3 neurons (since the task of interest was in both cases a 3-classes classification problem, see section 3.1) where the softmax is applied.
The second proposed architecture is the CNN, summarized in Table 5.

Layer name	Kernel size	Number of filters	Padding	Activation function	Stride
conv1	(3,3)	2	'same'	ReLu	1
conv2	(3,3)	196	'same'	ReLu	1
conv3	(3,3)	92	'same'	ReLu	1
flattening	/	/	/	/	/
output	/	3	/	softmax	/

Table 5: summary of the proposed CNN.

The third architecture proposed in this thesis is a concatenation of the previously presented CNN (whose first layer activation function was modified to 'SeLu') with a DNN. Its structure and hyperparamters are shown in Table 6.

Layer name	Kernel size	Number of filters	Padding	Activation function	Stride
conv1	(3,3)	2	'same'	SeLu	1
conv2	(3,3)	196	'same'	ReLu	1
conv3	(3,3)	92	'same'	ReLu	1
flattening	/	/	/	/	/
dense1	/	1176	/	/	/
dense2	/	1024	/	/	/
output	/	3	/	softmax	/

Table 6: summary of the concatenated model.

The last family of models considered is the *Long-Short Term Memory* (LSTM) and its further implementation, the *Bidirectional LSTM* (BiLSTM). These algorithms are virtual cells of memory, where information are stored for a certain period of time [70]. They are able to encode a text by storing the words into the cells. The cells are characterized by 3 gates (activation functions):

- The *write* gate, that allows information to get into the cell;
- The *keep* gate, that states how long the information stays inside the cell;
- The *read* gate, that allows to read the cell.

The standard LSTM encodes a sequence of data in only one direction, from the beginning to the end; BiLSTM is a LSTM that provides a double direction encoding (from the beginning to the end and viceversa). The LSTM and BiLSTM models adopted in this work are shown in Table 7 and 8.

Layer name	Output shape	Number of parameters
input	(None, 37, 1)	0
lstm1	(None, 37, 128)	6650
lstm2	(None, 128)	131584
dropout	(None, 128)	0
dense	(None, 128)	16512
output	(None, 3)	387

Table 7: summary of the proposed LSTM.

Layer name	Output shape	Number of parameters
input	(None, 37, 1)	0
bilstm1	(None, 37, 256)	133120
bilstm2	(None, 256)	394240
dropout	(None, 256)	0
dense	(None, 128)	32096
output	(None, 3)	387

Table 8: summary of the proposed BiLSTM.

Every DL model was also equipped with the *class weights* function, which computes specific weights to provide to each class. Their values were calculated as the inverse of the fraction of the samples belonging to the classes with respect to the overall amount of data. The number of epochs was set to 200. During the training phase, the *Earlystopping* callback was added in order to prevent overfitting; the monitored parameter was '*val_loss*' (which is the validation set error, computed at every epoch), with *patience* = 10. When iterations start, the callback considers as default the first epoch's value as the best one; if, within the 10 following iterations, a lower *val_loss* with respect to the first epoch's one is obtained, the best value is updated, and the 10 following steps are considered. In case the best value is not updated within 10 epochs, the callback considers that model has reached the lowest validation error and stops the training.

2.7. Metrics for evaluation

Several metrics were adopted to evaluate the goodness of the predictions. As the task of interest is a non-binary classification, all the metrics, apart from the accuracy, were calculated for each class separately (see equations 5, 6, 7, 8).
In order to derive the metrics, at first for every class 4 quantities of interest were defined:

1. True Positives (TP): the samples correctly predicted to belong to the class considered;
2. True Negative (TN): the samples correctly predicted not to belong to the class considered;
3. False Positive (FP): the samples wrongly predicted to belong to the class considered;
4. False Negative (FN): the samples wrongly predicted not to belong to the class considered.

By means of these definitions, the following metrics were computed:

$$accuracy = \frac{\sum_{i=1}^{nclasses} TP(i) + \sum_{i=1}^{nclasses} TN(i)}{nsamples} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \text{ (for every class)} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \text{ (for every class)} \tag{7}$$

$$F1\ score = \frac{2 \cdot precision \cdot recall}{precision + recall} \text{ (for every class)} \tag{8}$$

Every prediction result was also visually represented by means of the confusion matrix.
In addiction, the *Receiver Operating Characteristic* (ROC) curve was computed and its *Area Under Curve* (AUC) was calculated. The ROC curve consists in a plot whose x-axis represents the False Positive Rate (FPR), derived as $FP/(FP + TN)$, and whose y-axis represents the recall; the AUC is calculated as the area underlying the ROC curve. This operation was done for every class, as the standard ROC curve was implemented for binary classification problems; in every class the labels were binarized (e.g. when dealing with class 1, the label '0' is assigned to a sample belonging to class 1 and label '1' to the samples not belonging to class 1) and the ROC curve was computed [71].
Finally for every model the p-value was obtained by means of the *permutation_test_score* Python function [72]: 1000 random permutations of the label vector were performed, and for each of them the prediction was computed and the accuracy was calculated. Naming 'C' the amount of predictions whose accuracy was larger or equal to the one obtained with the original data, the p-value was finally computed as:

$$pvalue = \frac{C + 1}{npermutations + 1} \tag{9}$$

P-value was considered significant whether lower than 0.05 [72].

3. Results and Discussion

Results were calculated according to both the patients and therapist reported outcomes, using the signals recorded during the session. The goal was to compare algorithms performances in predicting the patients emotional wellbeing and engagement with the two different labelling strategies, which despite the reciprocal differences led to a 3-labels classification problem (see section 3.1). In all the tables, the model that provided the best result is highlighted in bold.

3.1. Labelling strategies

In both cases, labels resulted to be highly unbalanced.
The labelling strategy adopted on the therapist reported outcome (see section 2.3.2) led to the following distribution: 50 samples resulted to belonged to class '2', 34 to class '3', 10 to class '1' and 4 to class '4'. Due to the extremely low number of samples belonging to class '4', it was finally decided to discard it from the analysis, setting up a 3-labels classification.
On the other hand, according to patient reported outcome (see section 2.3.1) the sessions were labelled as follows: 64 samples belonged to class '2', 25 to class '1' and 9 to class '0'. The Krippendorf α coefficient was equal to 0.8095, assessing a high inter-raters agreement.

3.2. Feature selection

The statistical analysis performed on the EDA and HRV derived parameters revealed that none of the variables showed to have a normal distribution (23 revealed a purely randomic one and 14 a lognormal one). The Spearman correlation coefficients matrix of the variables was figured as a heatmap, reported in Fig. 9.

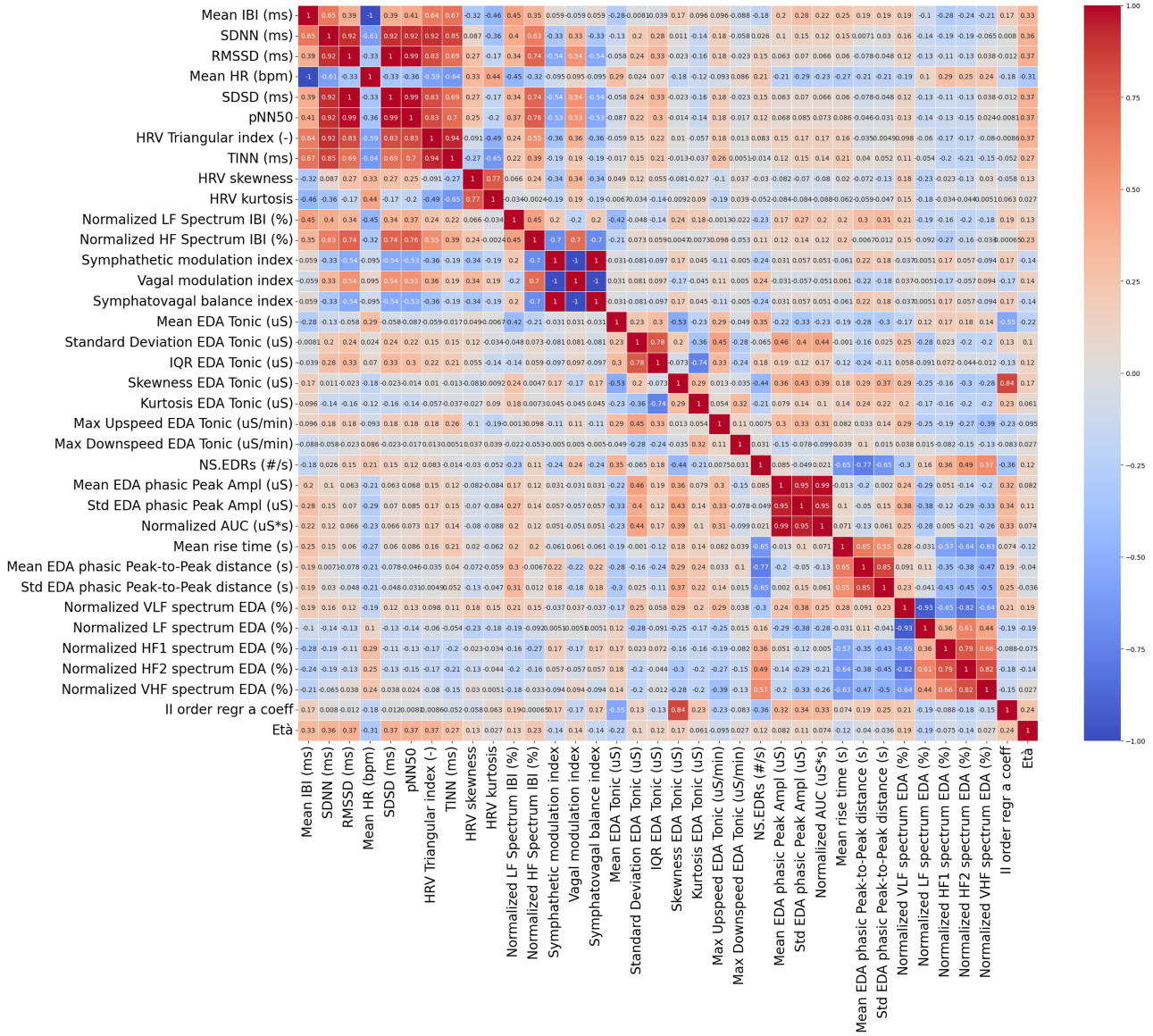


Figure 9: Spearman correlation matrix of the variables

According to these results, 8 variables were deleted: 'SDSD(ms)', 'pNN50', 'Normalized AUC (uS*s)', 'Std EDA phasic Peak Ampl (uS)', 'Symphatovagal balance index', 'Normalized VHF spectrum EDA (%)', 'Normalized HF2 spectrum EDA (%)' and 'Mean HR (bpm)'. The elimination of these features was coherent with the related literature, as 'RMSSD (ms)' is more predictive with respect to 'SDSD(ms)' and 'pNN50', as well as 'Mean EDA phasic Peak Ampl (uS)' rather than 'Normalized AUC (uS*s)' and 'Std EDA phasic Peak Ampl (uS)' and 'Normalized HF1 spectrum EDA (%)' compared to 'Normalized VHF spectrum EDA (%)' and 'Normalized HF2 spectrum EDA (%)'. Furthermore, 'Mean IBI (bpm)' and 'Mean HR (bpm)' embed very similar information, and so do 'Symphathethic balance index' and 'Symphatovagal balance index' [37].
Fig.10 shows the results of the PCA (left plot) and the NCA (right plot) performed on the dataset. For the PCA, individual explained variance and cumulative variance are displayed; in the PCA_{80} case, the first 11 PC were selected, while in the PCA_{90} case the number of PC increased up to 15.
In the NCA, the individual explained variance of every projection is plotted; projections 0, 4, 21, and 22 were chosen, as they explained the 88% of the cumulative variance of the dataset.

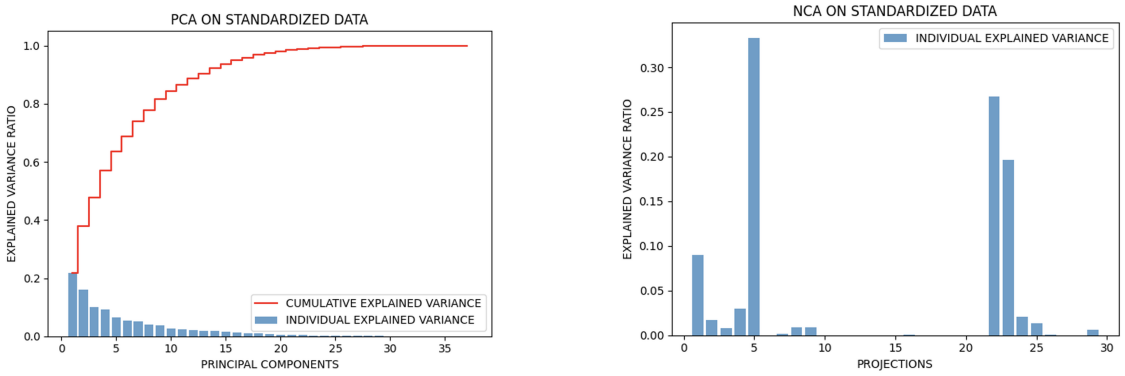


Figure 10: Principal components analysis and Neighborhood components analysis

3.3. Data upsampling

As previously seen in section 2.5, mean MCV was calculated for every window length in the 5 min-12min range, both with 0% and 50% overlapping; all the values are shown in Table 9. Basing on these results, the $12MIN, O = 0\%$ scenario was chosen to perform the upsampling.

WINDOW LENGTH	OVERLAP (%)	MEAN MCV
5 MIN	0%	1.028
5 MIN	50%	1.044
6 MIN	0%	0.995
6 MIN	50%	1.04
7 MIN	0%	0.962
7 MIN	50%	1.058
8 MIN	0%	0.993
8 MIN	50%	1.067
9 MIN	0%	0.985
9 MIN	50%	1.005
10 MIN	0%	0.994
10 MIN	50%	0.978
11 MIN	0%	0.956
11 MIN	50%	1.006
12 MIN	**0%**	**0.948**
12 MIN	50%	1.02

Table 9: MCV calculated for every windowing scenario.

Once upsampled, labels were computed with both the strategies. In the therapist reported case, the following distribution was derived: 87 samples fell in class '2', 60 in class '3', 15 in class '1' and 6 in class '4'. Also in this case, due to the low amount of samples falling in class '4', it was decided to discard it from the analysis. On the other hand, the patient reported labelling brought 110 samples belonging to class '2', 44 to class '1' and 14 to class '0'.

3.4. ML results

In the ML approach, results were obtained by applying the models to the test set after the hyperparamters tuning, hence the best parameters, obtained by the *best_param_* function of the gridsearch, were sorted for every algorithm and were set before the prediction computation [66]. Models in this approach provided better metrics with respect to the DL ones, both in absolute values and in terms of balance among the classes.
Generally feature selection techniques showed to be helpful for the algorithms, confirming that the reduction

of input size, done such that the chosen variables are a good approximation of the whole data, lowers the computational cost and allows a more accurate hand-crafted feature extraction, easing the task. This is an indicator of the key importance of dataset management in ML, as quoted in section 1.5. In detail, better results were obtained with the SVM family, probably because the KNN is more subjected to overfitting when dealing with very unbalanced labels [67].

3.4.1 Therapist reported outcome results

Table 10 summarizes the best parameters related to each model presented in section 2.6.1 for the treatment phase. From now on, the case 'Feature selection only' will be referred as 'Case 1', 'Feature selection+PCA_{80}' as 'Case 2', 'Feature selection+PCA_{90}' as 'Case 3' and 'Feature selection+NCA' as 'Case 4'.

	CASE 1	**CASE 2**	**CASE 3**	**CASE 4**
LinearSVC	$C = 0.1$	$C = 0.01$	$C = 0.01$	$C = 1.0$
SVC	$C = 1.0$, $degree = 2$, $\gamma =' scale'$, $kernel = linear$	$C = 1.0$, $degree = 2$, $\gamma =' auto'$, $kernel = rbf$	$C = 1.0$, $degree = 2$, $\gamma =' scale'$, $kernel = rbf$	$C = 0.1$, $degree = 2$, $\gamma =' scale'$, $kernel = rbf$
NuSVC	$C = 1.0$, $\nu = 0.1$	$C = 1.0$, $\nu = 0.1$	$C = 1.0$, $\nu = 0.2$	$C = 0.1$, $\nu = 0.1$
KNN	$K = 35$	$K = 30$	$K = 5$	$K = 40$

Table 10: best parameters of ML models in every analysis scenario with therapist reported labels.

The metrics described in section 2.7 were derived for every case and are presented in Table 11.

		accuracy	**AUC class 1**	**AUC class 2**	**AUC class 3**	**F1 class 1**	**F1 class 2**	**F1 class 3**	**p-value**
CASE 1	LinearSVC	0.46	0.86	0.54	0.44	0.5	0.55	0.33	0.158
	SVC	0.42	0.89	0.39	0.51	0.44	0.48	0.33	0.148
	NuSVC	0.5	0.77	0.48	0.62	0.4	0.5	0.53	0.04
	KNN	0.58	0.8	0.59	0.62	0	0.72	0.2	0.123
CASE 2	LinearSVC	0.5	0.84	0.57	0.63	0.33	0.59	0.4	0.028
	SVC	0.38	0.77	0.56	0.58	0.29	0.46	0.27	0.001
	NuSVC	0.58	0.89	0.72	0.67	0.29	0.72	0.5	0.003
	KNN	0.54	0.64	0.54	0.67	0	0.65	0.4	0.123
CASE 3	LinearSVC	0.46	0.89	0.61	0.53	0.29	0.59	0.29	0.108
	SVC	0.54	0.84	0.55	0.67	0.5	0.65	0.31	0.123
	NuSVC	0.46	0.89	0.62	0.62	0.33	0.56	0.35	0.159
	KNN	0.5	1	0.6	0.59	1	0.45	0.45	0.617
CASE 4	LinearSVC	0.42	0.34	0.55	0.6	0.33	0.22	0.58	0.379
	SVC	0.5	0.61	0.44	0.41	0.22	0.67	0.33	0.001
	NuSVC	**0.63**	**0.93**	**0.66**	**0.5**	**0.67**	**0.69**	**0.5**	**0.001**
	KNN	0.54	0.77	0.55	0.61	0	0.69	0.18	0.123

Table 11: ML results with therapist reported labels.

In all the sections related to ML results, the first row of matrices and ROC curves displayed is related to 'Case 1', the second row to 'Case 2', the third row to 'Case 3' and the fourth row to 'Case 4'. Fig. 11 represents the confusion matrices displayed for all the cases.

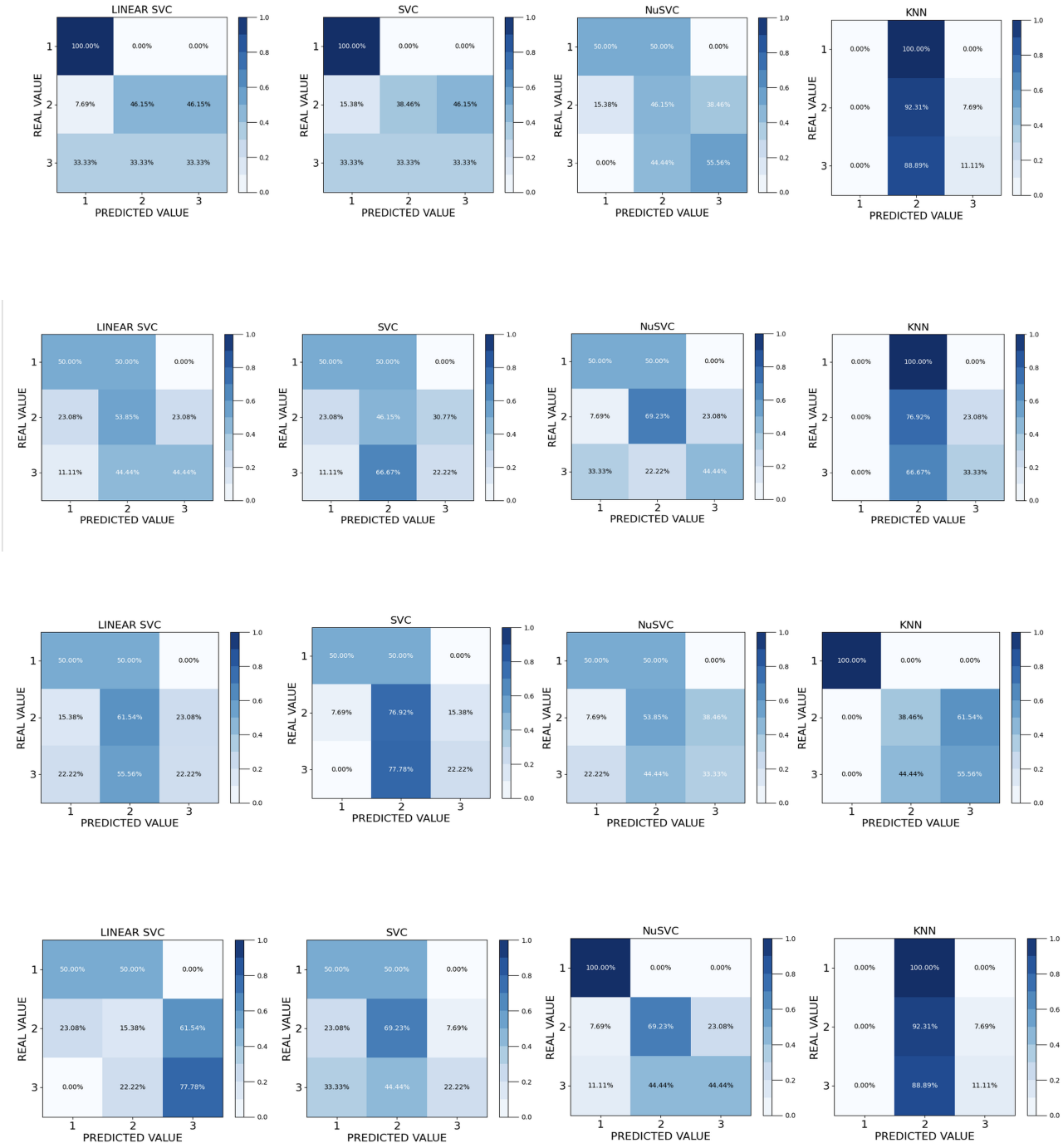
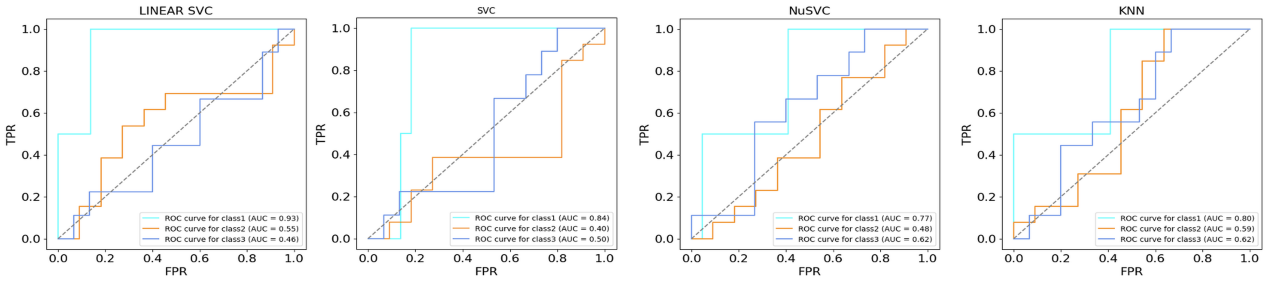


Figure 11: ML confusion matrices with therapist reported labels

Then the ROC curves were plotted, and are reported in Fig.12.



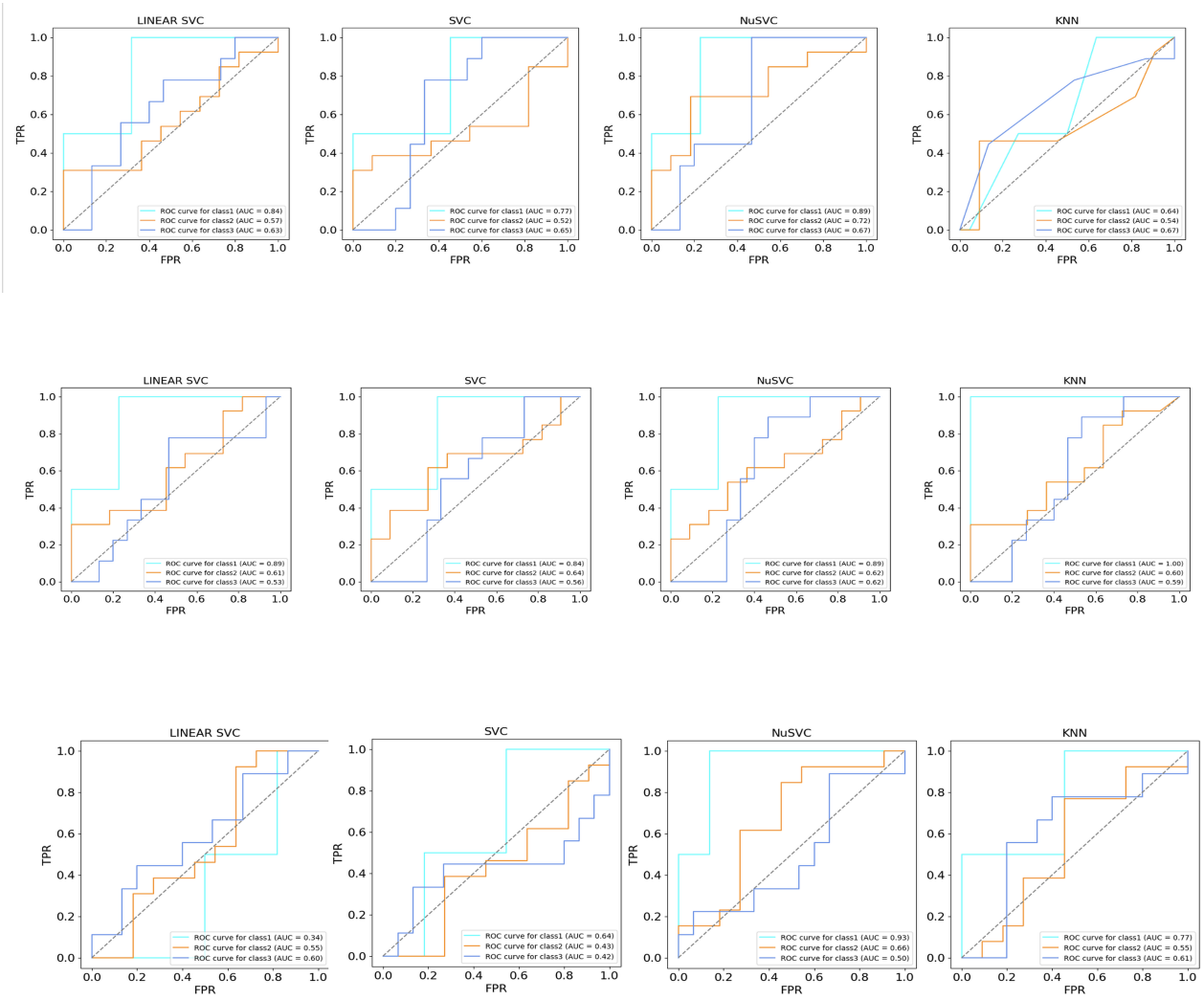


Figure 12: ML ROC curves with therapist reported labels

According to these metrics, the NuSVC-Case 4 revealed the best performance. Confusion matrices reveal the capability to distinguish all the classes, which is confirmed also by the fact that ROC curves are all above the bisector. P-value is signifcant.

3.4.2 Patient reported outcome results

Table 12 contains the best parameters obtained by fitting the models to data labelled with the patient reported outcome labelling strategy.

	CASE 1	**CASE 2**	**CASE 3**	**CASE 4**
LinearSVC	$C = 1.0$	$C = 0.01$	$C = 0.01$	$C = 0.01$
SVC	$C = 1.0$, $degree = 2$, $\gamma =' auto'$, $kernel = rbf$	$C = 0.01$, $degree = 5$, $\gamma =' scale'$, $kernel = poly$	$C = 1.0$, $degree = 2$, $\gamma =' scale'$, $kernel = rbf$	$C = 1.0$, $degree = 2$, $\gamma =' scale'$, $kernel = rbf$
NuSVC	$C = 0.1$, $\nu = 0.1$	$C = 0.1$, $\nu = 0.2$	$C = 1.0$, $\nu = 0.1$	$C = 0.01$, $\nu = 0.1$
KNN	$K = 25$	$K = 15$	$K = 5$	$K = 10$

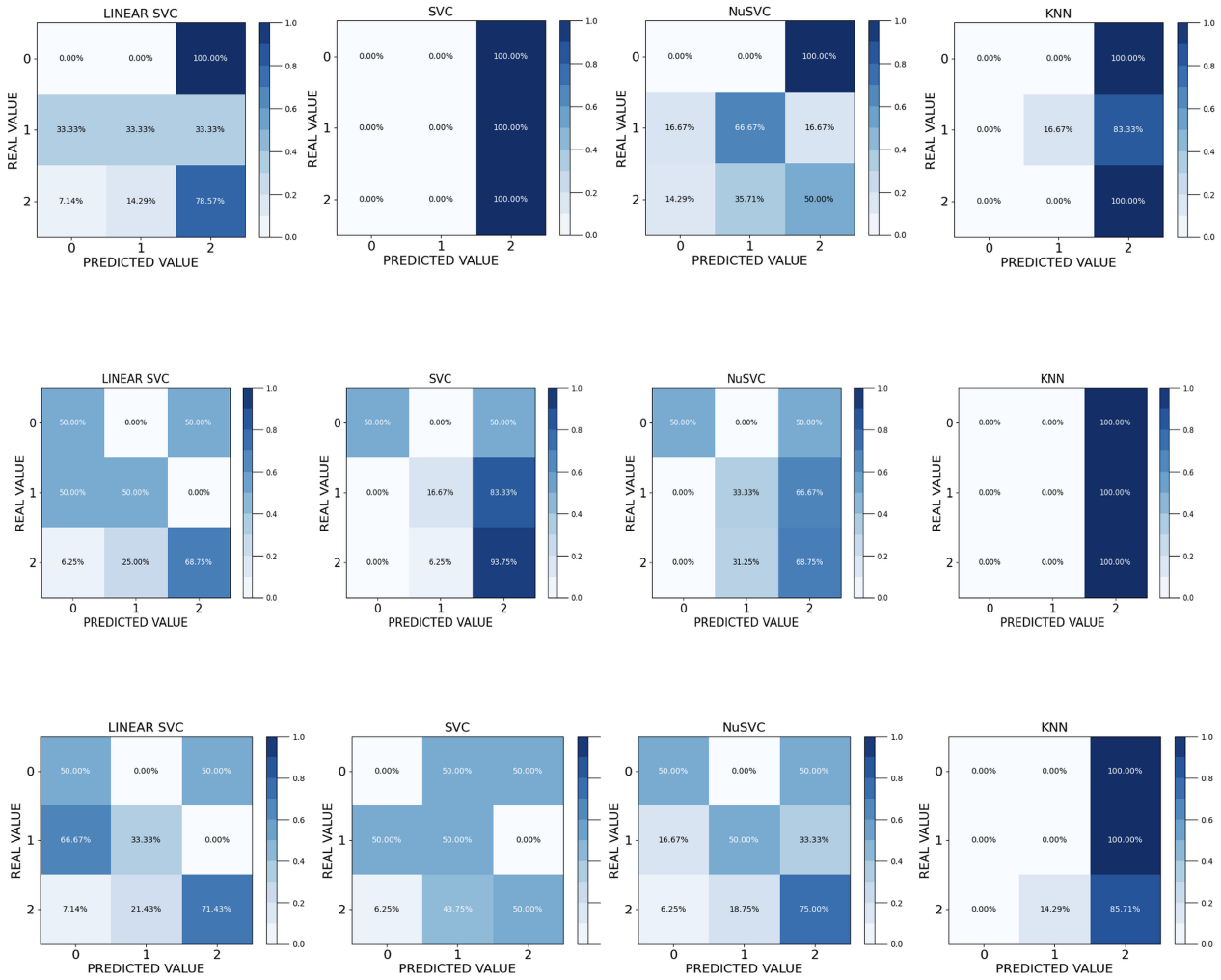
Table 12: best parameters of ML models in every analysis scenario with patient reported outcome.

The results obtained with the patient reported labels are hereby presented in Table 13.

		accuracy	AUC class 0	AUC class 1	AUC class 2	F1 class 0	F1 class 1	F1 class 2	p-value
CASE 1	LinearSVC	0.59	0.58	0.34	0.55	0	0.4	0.76	0.009
	SVC	0.67	0.45	0.6	0.55	0	0	0.8	0.001
	NuSVC	0.5	0.7	0.65	0.46	0	0.53	0.58	0.211
	KNN	0.68	0.7	0.65	0.46	0	0.29	0.8	0.002
CASE 2	LinearSVC	0.63	0.55	0.6	0.47	0.29	0.46	0.79	0.002
	SVC	**0.71**	**0.07**	**0.57**	**0.47**	**0.67**	**0.25**	**0.81**	**0.001**
	NuSVC	0.58	0.84	0.41	0.38	0.67	0.3	0.69	0.027
	KNN	0.67	0.8	0.78	0.75	0	0	0.8	0.023
CASE 3	LinearSVC	0.59	0.47	0.44	0.72	0.25	0.36	0.8	0.004
	SVC	0.46	0.45	0.61	0.72	0	0.35	0.64	0.153
	NuSVC	0.67	0.7	0.56	0.71	0.4	0.5	0.77	0.001
	KNN	0.55	0.9	0.45	0.54	0	0	0.71	0.997
CASE 4	LinearSVC	0.59	0.5	0.35	0.4	0	0.57	0.67	0.004
	SVC	0.32	0.22	0.42	0.38	0.24	0.22	0.44	0.057
	NuSVC	0.45	0.38	0.58	0.61	0.22	0.25	0.59	0.045
	KNN	0.59	0.6	0.58	0.63	0	0.2	0.75	0.870

Table 13: ML results with patient reported labels.

Also in this case the confusion matrices and the ROC curves were represented, and are reported in Fig.13 and Fig. 14.



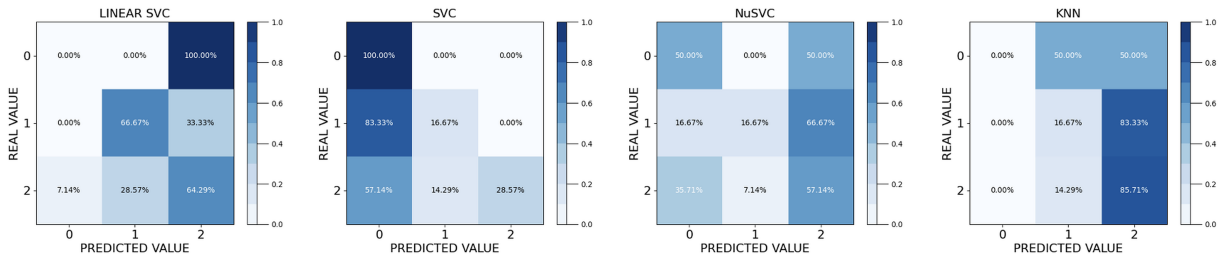


Figure 13: ML confusion matrices with patient reported labels

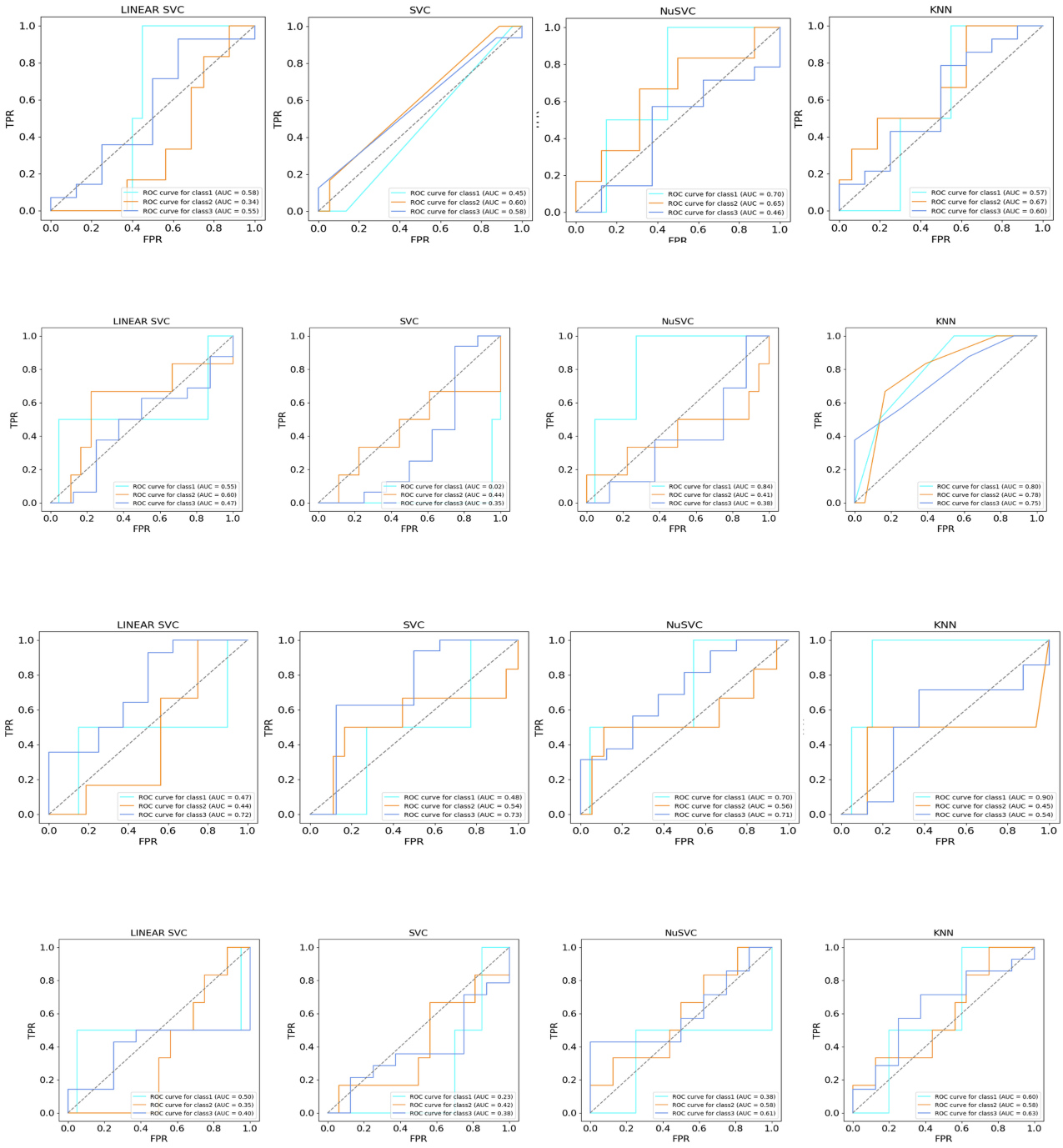


Figure 14: ML ROC curves with patient reported labels

In this case, the best results were obtained by the SVC-Case 2, that achieved high accuracy and non-null metrics in all the classes, as confirmed by the confusion matrices. P-value is signifcant.
This labelling strategy led to a larger amount of p-values lower than 0.05. This may be explained by the fact that the self-reported outcome may be more significant with respect to the therapist reported one to assess the mental state of the subjects.

3.5. DL results

In the DL approach, two scenarios are presented: the first one is related to the prediction made on the original data after the standardization (see section 2.4), which from now on will be defined as 'Case 1', while in the second one also data upsampling was applied (this will be called 'Case 2').
As previously quoted in section 3.4, this approach led to worse results. This was expected due to the poor amount of data for a DL predictioon, that did not allow an accurate data-driven feature extraction and suggests the needing of a larger amount of data in order to further improve the DL analysis.
Upsampling did not improve the metrics with respect to the ones obtained with original data. Despite the increased size of the dataset, the feature extraction worsened, and so were the performances. This may be related to the fact that, despite the low mean MCV and the equal length of the windows, the segmentation procedure could introduce a higher variability among the samples, making the feature extraction more difficult for the algorithms. Also, non significant p-values were achieved in this case, meaning that the chosen upsampling technique may introduce some noise in the data and worsen their quality.
In detail, neural networks achieved better results than LSTM, probably because the dataset was to poor to perform a consistent encoding with the virtual memory cells.

3.5.1 Therapist reported outcome results

Also in this case the distinction between the two labelling strategies was done. Table 14 contains the results obtained using the therapist reported labels.

		accuracy	**AUC class 1**	**AUC class 2**	**AUC class 3**	**F1 class 1**	**F1 class 2**	**F1 class 3**	**p-value**
CASE 1	DNN	0.42	0.41	0.56	0.33	0	0.62	0.25	0.533
	CNN	**0.58**	**0.89**	**0.52**	**0.54**	**0.5**	**0.64**	**0.53**	**0.001**
	CONCAT	0.46	0.55	0.33	0.4	0	0.54	0.42	0.146
	LSTM	0.46	0.55	0.53	0.61	0.33	0.48	0.48	0.035
	BiLSTM	0.5	0.61	0.53	0.64	0.18	0.44	0.74	0.201
CASE 2	DNN	0.34	0.39	0.44	0.44	0	0.5	0.27	0.943
	CNN	0.49	0.55	0.61	0.58	0	0.55	0.47	0.232
	CONCAT	0.51	0.7	0.38	0.38	0	0.64	0.32	0.202
	LSTM	0.51	0.77	0.62	0.56	0	0.68	0	1.000
	BiLSTM	0.42	0.91	0.49	0.52	0	0.45	0.41	0.885

Table 14: DL results with therapist reported labels.

Fig. 15 the confusion matrices and the ROC curves. In all the sections related to DL results, the first row of matrices and ROC curves displayed is related to 'Case 1' and the second one to 'Case 2'.

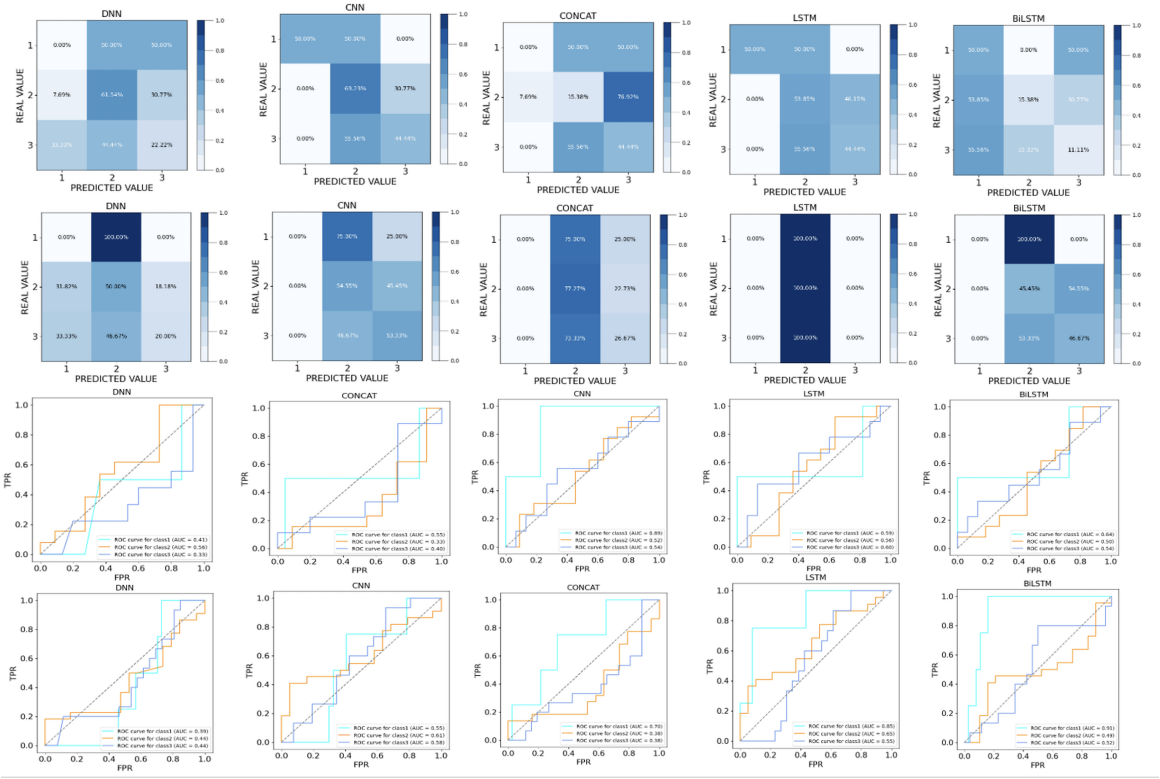


Figure 15: DL therapist reported confusion matrices and ROC curves

As can be seen in Table 14, the best model in this case is the CNN-Case 1. The confusion matrix confirms the capability to distinguish the 3 classes, and the ROC curves are all above the bisector, strengthening the validity of the results. P-value is significant.

3.5.2 Patient reported outcome results

Indeed Table 15 shows the results obtained in the DL approach with the patient reported labels.

		accuracy	AUC class 0	AUC class 1	AUC class 2	F1 class 0	F1 class 1	F1 class 2	p-value
CASE 1	**DNN**	**0.73**	**0.38**	**0.72**	**0.75**	**0**	**0.55**	**0.84**	**0.001**
	CNN	0.55	0.68	0.62	0.62	0	0.43	0.67	0.083
	CONCAT	0.59	0.8	0.58	0.55	0	0.36	0.71	0.110
	LSTM	0.64	0.75	0.33	0.34	0	0	0.78	0.001
	BiLSTM	0.5909	0.88	0.62	0.54	0	0	0.7429	0.002
CASE 2	DNN	0.5	0.53	0.63	0.56	0.2	0.21	0.66	0.940
	CNN	0.52	0.62	0.53	0.55	0	0.46	0.64	0.065
	CONCAT	0.62	0.66	0.62	0.64	0	0	0.77	0.749
	LSTM	0.64	0.72	0.48	0.55	0	0	0.78	1.000
	BiLSTM	0.43	0.74	0.52	0.54	0	0.35	0.52	1.000

Table 15: DL results with patient reported labels.

Confusion matrices and ROC curves are reported in Fig. 16.

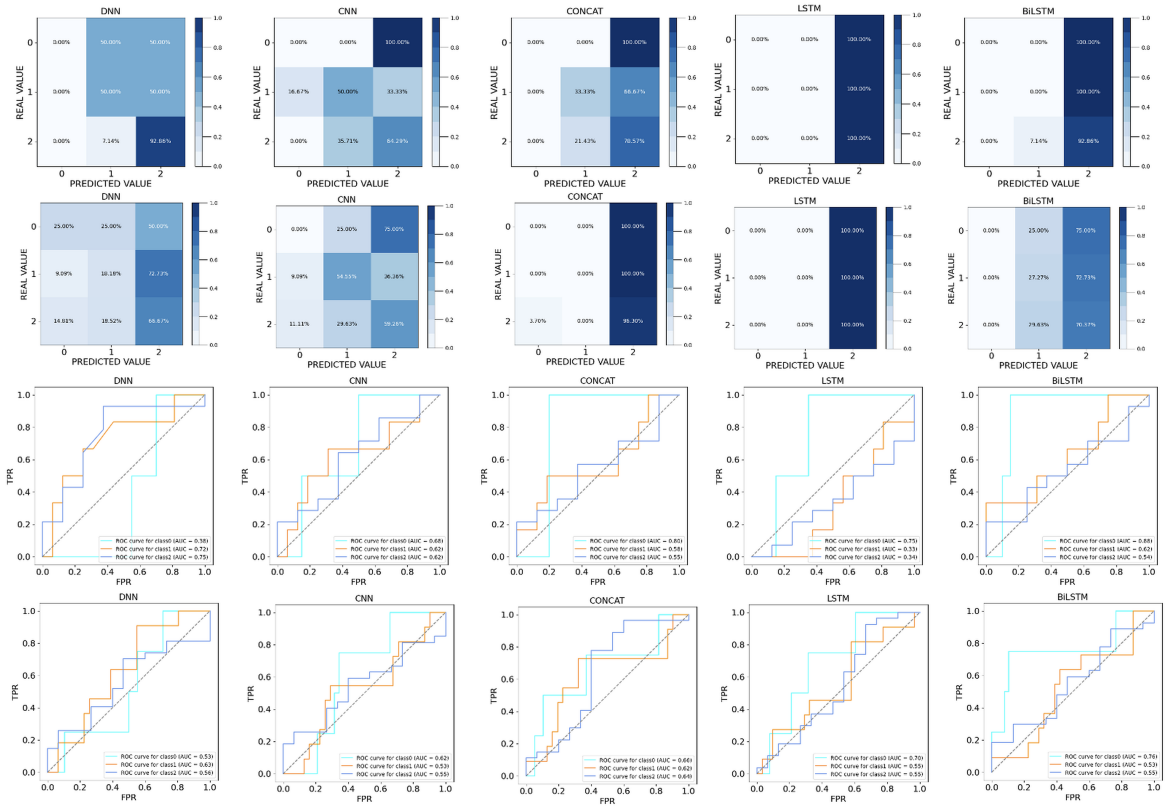


Figure 16: DL patient reported confusion matrices and ROC curves

This scenario revealed as best model the DNN-Case 1; despite the null performance on class '0', it achieved a significant p-value, high accuracy and ROC curves plots abundantly above the bisector for classes '1' and '2'.

4. Conclusions

To date there is a lack of focus, in the neuromotor rehabilitation field, on the mental wellbeing of the patients [9]. Current literature is poor of investigations of the mental engagement of the subjects undergoing this kind of therapy sessions.
In this Master thesis, time and frequency domain parameters extracted from the biosignals recorded by means of the Empatica E4 wristband [37] were used as input for AI models.
Patient reported and therapist reported questionnaires were analysed in order to derive a numerical value that was an indicator of the emotional wellbeing and engagement of the subjects during and after the treatment.
Algorithms belonging to both ML and DL were trained to classify each session in which signals were acquired.
Some models revealed the capability to distinguish different emotional wellbeing and engagement levels, which highlights the importance to further investigate the psychological response of the subjects to the treatment.
Being able to predict the emotional wellbeing and engagement of young patients during rehabilitation potentially allows to customize the therapy according to patient needs and wellbeing.
In addiction, the analysis of questionnaires related to different phases of the treatment enabled to investigate different aspects of patients psychological response to Lokomat.
Despite this, there are still some limitations to be overcome. At first the fact that both in the patient and in therapist reported case the labels were derived starting from *ad hoc* questionnaires, as previously seen in section 2.2, which do not belong to the ones adopted in literature and thus in clinical practice.
A second issue is related to the small amount of available data, which did not enable the algorithms to perform a substantial training and hence to learn accurately the relationship between the labels and the parameters. This arises the necessity of collecting more recordings to improve the models performances.
The results obtained suggest to go into detail in the application of emotion recognition to neuromotor robot-assisted rehabilitation.
One of the possible future developments of this research can be the implementation of a real-time prediction system, that may allow to assess in live the emotional state of the participant and to adjust the therapy right during its execution. This may foster increased patient's comfort and treatment reliance.

References

[1] D. R. Patel, M. Neelakantan, K. Pandher, and J. Merrick, "Cerebral palsy in children: a clinical overview," *Translational pediatrics*, vol. 9, no. Suppl 1, p. S125, 2020.

[2] D. S. Reddihough and K. J. Collins, "The epidemiology and causes of cerebral palsy," *Australian Journal of physiotherapy*, vol. 49, no. 1, pp. 7–12, 2003.

[3] A. Paulson and J. Vargus-Adams, "Overview of four functional classification systems commonly used in cerebral palsy," *Children*, vol. 4, no. 4, p. 30, 2017.

[4] P. H. McCrea, J. J. Eng, and A. J. Hodgson, "Biomechanics of reaching: clinical implications for individuals with acquired brain injury," *Disability and rehabilitation*, vol. 24, no. 10, pp. 534–541, 2002.

[5] M. C. Dewan, N. Mummareddy, J. C. Wellons III, and C. M. Bonfield, "Epidemiology of global pediatric traumatic brain injury: qualitative review," *World neurosurgery*, vol. 91, pp. 497–509, 2016.

[6] Y. Cherni, L. Ballaz, J. Lemaire, F. Dal Maso, and M. Begon, "Effect of low dose robotic-gait training on walking capacity in children and adolescents with cerebral palsy," *Neurophysiologie Clinique*, vol. 50, no. 6, pp. 507–519, 2020.

[7] V. Falzarano, F. Marini, P. Morasso, and J. Zenzeri, "Devices and protocols for upper limb robot-assisted rehabilitation of children with neuromotor disorders," *Applied Sciences*, vol. 9, no. 13, p. 2689, 2019.

[8] I. Bortone, M. Barsotti, D. Leonardis, A. Crecchi, A. Tozzini, L. Bonfiglio, and A. Frisoli, "Immersive virtual environments and wearable haptic devices in rehabilitation of children with neuromotor impairments: a single-blind randomized controlled crossover pilot study," *Journal of neuroengineering and rehabilitation*, vol. 17, no. 1, pp. 1–14, 2020.

[9] R. Banz, M. Bolliger, G. Colombo, V. Dietz, and L. Lünenburger, "Computerized visual feedback: an adjunct to robotic-assisted gait training," *Physical therapy*, vol. 88, no. 10, pp. 1135–1145, 2008.

[10] N. Maclean, P. Pound, C. Wolfe, A. Rudd, *et al.*, "A critical review of the concept of patient motivation in the literature on physical rehabilitation," *Soc Sci Med*, vol. 50, no. 4, pp. 495–506, 2000.

[11] R. J. Siegert and W. J. Taylor, "Theoretical aspects of goal-setting and motivation in rehabilitation," *Disability and rehabilitation*, vol. 26, no. 1, pp. 1–8, 2004.

[12] G. King, L. A. Chiarello, L. Thompson, M. J. McLarnon, E. Smart, J. Ziviani, and M. Pinto, "Development of an observational measure of therapy engagement for pediatric rehabilitation," *Disability and Rehabilitation*, vol. 41, no. 1, pp. 86–97, 2019.

[13] L. R., "Technique for the measure of attitudes arch. psycho," *Commun. ACM*, vol. 22, p. 140, 1932.

[14] T.-M. Bynion and M. T. Feldner, *Encyclopedia of Personality and Individual Differences*, pp. 1–3. Springer International Publishing, 2017.

[15] A. Koenig, X. Omlin, L. Zimmerli, M. Sapa, C. Krewer, M. Bolliger, F. Müller, and R. Riener, "Psychological state estimation from physiological recordings during robot-assisted gait rehabilitation," *Journal of Rehabilitation Research and Development*, vol. 48, no. 4, pp. 367–385, 2011.

[16] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[17] S. Basu, N. Jana, A. Bag, M. M, J. Mukherjee, S. Kumar, and R. Guha, "Emotion recognition based on physiological signals using valence-arousal model," in *2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 50–55, 2015.

[18] T. Razavi, "Self-report measures: An overview of concerns and limitations of questionnaire use in occupational stress research," 2001.

[19] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, p. 479, 2020.

[20] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," *Handbook of psychophysiology*, vol. 2, pp. 200–223, 2007.

[21] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[22] C. Julien, "The enigma of mayer waves: facts and models," *Cardiovascular research*, vol. 70, no. 1, pp. 12–21, 2006.

[23] N. Milstein and I. Gordon, "Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states," *Frontiers in Behavioral Neuroscience*, vol. 14, p. 148, 2020.

[24] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, pp. 255–260, 2015.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[26] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[27] A. Botalb, M. Moinuddin, U. Al-Saggaf, and S. S. Ali, "Contrasting convolutional neural network (cnn) with multi-layer perceptron (mlp) for big data analysis," in *2018 International conference on intelligent and advanced system (ICIAS)*, pp. 1–5, IEEE, 2018.

[28] S. Saha, "A comprehensive guide to convolutional neural networks — the eli5 way," *Towards Data Science*, vol. 4, 2018.

[29] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] D. E. Barkana and E. Masazade, "Classification of the emotional state of a subject using machine learning algorithms for rehabroby," in *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications*, pp. 2160–2187, IGI Global, 2017.

[31] F. A. Machot, A. Elmachot, M. Ali, E. A. Machot, and K. Kyamakya, "A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors," *Sensors*, vol. 19, no. 7, p. 1659, 2019.

[32] K. M. Dalmeida and G. L. Masala, "Hrv features as viable physiological markers for stress detection using wearable devices," *Sensors*, vol. 21, no. 8, p. 2873, 2021.

[33] F. Al Machot, M. Ali, S. Ranasinghe, A. H. Mosa, and K. Kyandoghere, "Improving subject-independent human emotion recognition using electrodermal activity sensors for active and assisted living," in *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*, pp. 222–228, 2018.

[34] E. Eren and T. S. Navruz, "Stress detection with deep learning using bvp and eda signals," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–7, IEEE, 2022.

[35] D. Huysmans, E. Smets, W. De Raedt, C. Van Hoof, K. Bogaerts, I. Van Diest, and D. Helic, "Unsupervised learning for mental stress detection," in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 4, pp. 26–35, 2018.

[36] T. Nagae and J. Lee, "Understanding emotions in children with developmental disabilities during robot therapy using eda," *Sensors*, vol. 22, no. 14, p. 5116, 2022.

[37] S. Costantini, M. Chiappini, G. Malerba, C. Dei, A. Falivene, S. Arlati, V. Colombo, E. Biffi, and F. A. Storm, "Wrist-worn sensor validation for heart rate variability and electrodermal activity detection in a stressful driving environment," *Sensors*, vol. 23, no. 20, p. 8423, 2023.

[38] C. Kahl and J. A. Cleland, "Visual analogue scale, numeric pain rating scale and the mcgill pain questionnaire: an overview of psychometric properties," *Physical therapy reviews*, vol. 10, no. 2, pp. 123–128, 2005.

[39] S. K. Phelan, B. E. Gibson, and F. V. Wright, "What is it like to walk with the help of a robot? children's perspectives on robotic gait training technology," *Disability and rehabilitation*, vol. 37, no. 24, pp. 2272–2281, 2015.

[40] K. A. Knowles and B. O. Olatunji, "Specificity of trait anxiety in anxiety and depression: Meta-analysis of the state-trait anxiety inventory," *Clinical psychology review*, vol. 82, p. 101928, 2020.

[41] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the spielberger state—trait anxiety inventory (stai)," *British journal of clinical Psychology*, vol. 31, no. 3, pp. 301–306, 1992.

[42] M. Hamilton, "The assessment of anxiety states by rating.," *British journal of medical psychology*, 1959.

[43] P. Richter, J. Werner, A. Heerlein, A. Kraus, and H. Sauer, "On the validity of the beck depression inventory: A review," *Psychopathology*, vol. 31, no. 3, pp. 160–168, 1998.

[44] M. Zimmerman, I. Chelminski, and M. Posternak, "A review of studies of the montgomery–asberg depression rating scale in controls: implications for the definition of remission in treatment studies of depression," *International clinical psychopharmacology*, vol. 19, no. 1, pp. 1–7, 2004.

[45] V. Pollock, D. W. Cho, D. Reker, and J. Volavka, "Profile of mood states: the factors and their physiological correlates," *The Journal of nervous and mental disease*, vol. 167, no. 10, pp. 612–614, 1979.

[46] R. Snodgrass, "Single-versus double-blind reviewing: An analysis of the literature," *ACM Sigmod Record*, vol. 35, no. 3, pp. 8–21, 2006.

[47] K. Denecke and D. Reichenpfader, "Sentiment analysis of clinical narratives: A scoping review," *Journal of Biomedical Informatics*, p. 104336, 2023.

[48] S. Loria *et al.*, "textblob documentation," *Release 0.15*, vol. 2, no. 8, p. 269, 2018.

[49] J. Eggink, "Krippendorff's Alpha." `https://www.mathworks.com/matlabcentral/fileexchange/36016-krippendorff-s-alpha`, November 2023. MATLAB Central File Exchange.

[50] M. Norris and L. Lecavalier, "Evaluating the use of exploratory factor analysis in developmental disability psychological research," *Journal of autism and developmental disorders*, vol. 40, pp. 8–20, 2010.

[51] C. Spearman, "The proof and measurement of association between two things.," 1961.

[52] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci*, vol. 374, no. 2065, p. 20150202, 2016.

[53] G. Pasini, "Principal component analysis for stock portfolio management," *International Journal of Pure and Applied Mathematics*, vol. 115, no. 1, pp. 153–167, 2017.

[54] D. P. Armstrong, S. P. Pretty, T. B. Weaver, S. L. Fischer, and A. C. Laing, "Application of principal component analysis to forward reactive stepping: Whole-body movement strategy differs as a function of age and sex," *Gait & Posture*, vol. 89, pp. 38–44, 2021.

[55] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood components analysis advances in neural information processing systems. 17," 2004.

[56] J. Brandt and E. Lanzén, "A comparative review of smote and adasyn in imbalanced data classification," 2021.

[57] M. R. Askari, M. Abdel-Latif, M. Rashid, M. Sevil, and A. Cinar, "Detection and classification of unannounced physical activities and acute psychological stress events for interventions in diabetes treatment," *Algorithms*, vol. 15, no. 10, p. 352, 2022.

[58] H. He and Y. Ma, "Imbalanced learning: foundations, algorithms, and applications," 2013.

[59] T. Chaspari, B. Baucom, A. C. Timmons, A. Tsiartas, L. B. Del Piero, K. J. Baucom, P. Georgiou, G. Margolin, and S. S. Narayanan, "Quantifying eda synchrony through joint sparse representation: a case-study of couples' interactions," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 817–821, IEEE, 2015.

[60] J. Bigger, J. L. Fleiss, L. M. Rolnitzky, and R. C. Steinman, "The ability of several short-term measures of rr variability to predict mortality after myocardial infarction.," *Circulation*, vol. 88, no. 3, pp. 927–934, 1993.

[61] S. Aerts, G. Haesbroeck, and C. Ruwet, "Multivariate coefficients of variation: comparison and influence functions," *Journal of Multivariate Analysis*, vol. 142, pp. 183–198, 2015.

[62] L. Zhu, P. Spachos, P. C. Ng, Y. Yu, Y. Wang, K. Plataniotis, and D. Hatzinakos, "Stress detection through wrist-based electrodermal activity monitoring and machine learning," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[63] R. Sanchez-Reolid, F. L. de la Rosa, M. T. Lopez, and A. Fernandez-Caballero, "One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity," *Biomedical Signal Processing and Control*, vol. 71, p. 103203, 2022.

[64] C.-P. Hsieh, Y.-T. Chen, W.-K. Beh, and A.-Y. A. Wu, "Feature selection framework for xgboost based on electrodermal activity in stress detection," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 330–335, 2019.

[65] R. Raghav, G. Lemaitre, and T. Unterthiner, "Compare the effect of different scalers on data with outliers," 2020.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[67] D. Shakhnarovish and Indyk, "Nearest-neighbour methods in learning and vision," *The MIT Press*, 2005.

[68] N. Ketkar, *Stochastic Gradient Descent*, pp. 113–132. Berkeley, CA: Apress, 2017.

[69] J. Baek and Y. Choi, "Deep neural network for predicting ore production by truck-haulage systems in open-pit mines," *Applied Sciences*, vol. 10, p. 1657, 03 2020.

[70] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[71] S. M. Winkler, M. Affenzeller, and S. Wagner, "Sets of receiver operating characteristic curves and their use in the evaluation of multi-class classification," (New York, NY, USA), Association for Computing Machinery, 2006.

[72] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," in *2009 Ninth IEEE International Conference on Data Mining*, 2010.

Abstract in lingua italiana

Patologie come la Paralisi Cerebrale e la Lesione Cerebrale Acquisita possono facilmente portare a deficit neuromotori, diffusamente trattati al giorno d'oggi con la terapia robotizzata, che ha dimostrato di essere molto benefica per il recupero delle funzioni neuromotorie. Ciononostante, la risposta psicologica alla terapia, aspetto fondamentale soprattutto avendo a che fare con bambini, ad oggi non è stato ancora vagliato accuratamente. Questa tesi Magistrale è finalizzata alla predizione tramite strumenti di Intelligenza Artificiale il livello di benessere e coinvolgimento emotivo di 42 soggetti sottoposti a riabilitazione neuromotoria con esoscheletro Lokomat all'IRCCS Medea. Durante alcune delle sessioni, l'Electrodermal Activity (EDA) e il Blood Volume Pulse (BVP) sono stati registrati per mezzo del braccialetto medico Empatica E4. Diversi parametri relativi al dominio del tempo e delle frequenze sono stati ricavavti e dati in input a modelli di Machine Learning (ML) e Deep Learning (DL).
In primo luogo sono state ricavate delle etichette che rappresentassero lo stato emotivo di ogni sessione. Sono state implementate due predizioni, una relativa all'outcome riportato dal paziente e un'altra relativa a quello riportato dalla terapista.
Nell'approccio ML, è stata effettuata una selezione delle variabili tramite delle analisi statistiche, al fine di diminuire il costo computazionale dei modelli. Nell'approccio DL, il dataset è stato aumentato tramite la finestratura dei segnali, per rendere il processo di allenamento dei modelli più consistente. La parte finale consisteva nell'implementazione degli algoritmi e nella valutazione delle relative metriche.
I risultati mostrano che i modelli riescono a riconoscere i diversi stati emozionali, suggerendo la capacità dell'IA di studiare la risposta psicologica alla terapia a partire da segnali fisiologici. I modelli ML hanno portato risultati migliori di quelli DL, a causa dell'esiguo numero di dati a disposizione. Il *Support Vecotr Machine* (SVM) ha fatto prediziini migliori del *K-Nearest Neighbors* (KNN), probabilemente a causa dell'alto rischio di overifitting associato all'uso del KNN con dataset sbilanciati. Le *Neural Networks* hanno ottenuto i risultati mgliori nel DL.
Un possibile sviluppo futuro può essere l'implementazione di un sistema di predizione in tempo reale, che permetterebbe di adattare la terapia ai bisogni specifici del paziente e di superare il limite di comunicazione dovuto alla giovane età o al deficit neurologico del paziente.

Parole chiave: Lokomat, Empatica E4, HRV, EDA, Machine Learning, Deep Learning, riconoscimento delle emozioni

Acknowledgements

Here you might want to acknowledge someone.