POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Enrico Magliulo

Advisor: Prof. Emilia Ambrosini

Co-advisors: Emilia Biffi, Simone Costantini, Anna Falivene

Academic year: 2022-2023

1. Introduction

1.1. Mental state assessment methods in children neurorehabilitation

Cerebral palsy (CP) and Acquired Brain Injury (ABI) are two neurmotor disorders widespreadly treated by means of robot-assisted therapy tools. An example is the Lokomat exoskeleton, which has proved to be at least as effective as conventional treatment with advantages in terms of motivation and therapists effort [1].
Although it is one of the key aspects in rehabilitation, mental engagement still needs to be deeply investigated, since the crucial step of emotions quantification must be faced. Questionnaires are the instruments typically adopted for this purpose; they use tools such as the Likert scale to rate a series of items related to patient's emotional state. Once filled, questionnaires scores can also be aggregated into a single value assessing patient's emotional state. Tools that can be used for this purpose are the circumphlex model implemented by Russel [2].
However, despite their large use, psychometric questionnaires are strictly dependent on the subject's willingness and capability to respond to the questions. Quantitative measurements can be used instead, as they contain relevant information for the purpose.

1.2. Electrodermal Activity (EDA) and Heart Rate Variability (HRV)

EDA and HRV are the two most relevant markers of a person psychophysical state.
EDA can be defined as the variation of electrical conductance of the skin, and is derived by the sum of the *Skin Conductance Level* (SCL), that represents a generic measurement of the activation level of the Autonomous Nervous System (ANS), and the *Skin Conductance Response* (SCR), which represents transient changes in skin electrical activity and can be considered as the specific response of the Sympathetic Nervous System (SNN).
HRV consists in the variability of the time that interleaves two consecutive heart beats. Its behavior is an indicator of the type of nervous system activated in a specific moment; a higher HRV with respect to the baseline value assesses Vagal System activation, while a HRV lower

than the baseline states SNN activation.
These two signals may be used in emotion recognition field as input to Artificial Intelligence (AI) models.

1.3. Deep Learning and Machine Learning in emotion recognition

Machine Learning (ML) and Deep Learning (DL) are two AI branches commonly used to perform classification. A key step of this task is the selection of the most significant features for the prediction (*feature extraction*), performed by the user (*hand-crafted feature extraction*) in the ML, and by the algorithm itself (*data-driven feature extraction*), by the feature extractor, in the DL. Emotion recognition consists in predicting the emotional state of a subject while developing a specific activity, and can be performed by means of AI models. The majority of the related works adopted benchmark datasets for the scope. Dalmeida *et al.* [3] 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 the median Galvanic Skin Responses (GSR) value as a threshold to label the patient as stressed or not stressed, reaching a macro averaged F1 score equal to 0.74. Nagae *et al.* validated a system to automatically recognize sadness, anger, surprise and happiness in children robot-based rehabilitation using a SVM classifier and EDA extracted features; 38.6% of accuracy was achieved [4].

1.4. Objective of the thesis

The purpose of this thesis is to predict the emotional wellbeing and engagement of patients undergoing neuromotor rehabilitation with Lokomat exoskeleton by training both ML and DL models with a set of time and frequency domain parameters extracted from EDA and Blood Volume Pulse (BVP) recordings. Two different labelling strategies were adopted, one related to self-reported outcomes and another one to therapist-reported outcome, and the related algorithms performances were compared.

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 between 15 and 20 sessions of neuromotor rehabilitation with Lokomat in the Scientific Institute I.R.C.C.S. "E. Medea". Data included in this work are related to 2-3 sessions during which Empatica E4 wristband (Empatica®, Milan, Italy) was worn by the subjects, The device recorded the EDA, whose components were successively derived, and the BVP, from which the HRV was sorted [5]. *Visual Analogue Scale* (VAS)-type questionnaires related to their emotional wellbeing were submitted to patients at the beginning and at the end of each session. During the treatment a questionnaire regarding the level of engagement of the subjects into the task was compiled by the therapist. At the end of the session, the participants were asked to fill a comic so that they could express their feeling about the session with sentences and emoticons.
The VAS-type questionnaires were made of 10 items each. The first 6 of them rated feelings of worry, happiness, sadness, anger, fear and boredom by means of a 3-level Likert scale. Instead the last 4 items concerned the trust the patients had in Lokomat therapy.
Questionnaire filled in by the therapist was made of 12 items representing opposite emotional states; therapist had to give a score in every item, ranging from -3 to 3, to assess where the patient's emotional response was located.

2.2. Labelling strategy

Due to the different structures of the questionnaires, two different labelling strategies were adopted: one related to the mental wellbeing prediction (patient reported outcome) and another related to the mental engagement prediction (therapist reported outcome).
In patient reported outcome case, in order to account for all the information provided by the patient thanks to both VAS questionnaires and comic, a *double-blinded expert evaluation* was performed. A group of 4 experts, none of whom was psychologist, provided a personal score of all the answers given by the patients within the self-reported outcome at every session. In addiction to the 10 previously quoted ones, a new item was

created performing Sentiment Analysis (SA) on the output of the comic questionnaire. SA is the study of the emotional polarity of a text; the SA tool adopted in this work is the TextBlob 0.16.0, an algorithm able to encode the words that compose a text and understand their meaning, which was selected due to its wide use in literature.
The assessments on the 11 items were done to group patients in 3 different emotional classes: negative (with label '0'), neutral (with label '1') and positive (with label '2'). The inter-rater agreement was investigated by means of the Krippendorf α coefficient calculation; typically, 0.8 is considered the threshold for high agreement.
In therapist reported outcome case, a simil-Russel model was built up with questionnaire's items [2]. Four groups of items, corresponding to each quadrant of the model, were identified. The groups are shown in Fig.1.

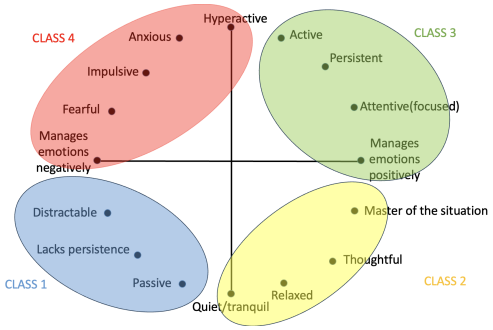

Figure 1: Classes identification in the Russel model

An *Exploratory Factorial Analysis* (EFA) was applied in order to check whether the items of the therapist's questionnaire could be placed in a circular shaped plot, as in the Russel model. Sessions were labelled by defining a vector for each quadrant, having as modulus 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.

2.3. Dataset structure and preprocessing

The dataset fed to the models contained 37 columns overall: 15 related to HRV parameters, 20 to EDA ones (both derived by means of the algorithms described by Costantini *et al.* [5]) and then patients' sex and age.
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.
Data standardization, done in both the approaches, was made by subtracting the post-treatment physiological data (considered as the baseline) to the Lokomat ones.
In the ML approach, in order to reduce the computational cost of the models, feature selection was performed by calculating the Spearman correlation coefficient of the variables, due to its suitability to non-normally distributed features; 0.8 was chosen as threshold value.
Afterwards two separated further feature reductions were 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), that explain a certain fraction of the total variance (*individual explained variance*) of the data. In this thesis, the groups of PC that explained the 80% of the overall dataset variance (defined as PCA_{80}) and the 90% (defined as PCA_{90}) were chosen.
The other algorithm chosen for this purpose is the *Neighborhood Components Analysis* (NCA), that creates projections of the original features into a lower dimensionality space such that similar samples are closer each other. Projections to use were chosen such that they explained the 88% of the overall variance of the data.
In the DL approach, data were upsampled by segmenting the original signals; every window was considered itself a signal and EDA and HRV parameters were computed over each segment. In this work, a series of segmentation scenarios was investigated by means of the calculation, for each of them, of the *Multivariate Coefficient of Variation* (MCV) with the Albert&Zhang (2010) formulation, reported in equation 1 [6]. MCV is an indicator of the deviation of all the features with respect to their mean value.

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

In equation 1, 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 and 50% overlapping were considered. Before the calculation of the MCV, a standardization of the dataset was performed, since the parameters had different ranges of values. The chosen scaler was the Robust Scaler, since it relies on the *Inter-Quartile Range* (IQR) and hence is not influenced by the presence of outliers. The output of the analysis was, for every scenario, a vector of MCV values, with length equal to the number of samples, whose mean value was computed. The windowing scenario that provided the minimum mean MCV was finally chosen.

2.4. Proposed models

In the ML approach, two families of models were adopted for the scope: the *Support Vector Machine* (SVM) and the *K-Nearest Neighbors* (KNN).
In this task, 3 models belonging to SVM family were used: SVC, NuSVC and LinearSVC.
The SVC is the standard *Support Vector Classifier*. It is equipped with several hyperparameters that were fine tuned, as: the *regularization parameter*, (whose values were 0.001, 0.01, 0.1 and 1.0), the *kernel type* (among 'poly', 'linear', 'rbf' and 'sigmoid'), the *degree* (between 2 and 5) of the polynomial kernel and the γ parameter, that is a kernel coefficient (computed as $1/\#features$ or as $1/[\#features \cdot \mathrm{Var(features)}]$).
The NuSVC consists in a SVC to which a parameter ν (that must be within the (0,1] interval) is added to bound the training phase error. 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'.
In the KNN, the k value fine tuned between 5 and half of the training set size, with $step = 5$.
In the DL approach, 3 main families of architectures were considered during models implementation: the *Dense Neural Network* (DNN), the *Convolutional Neural Network* (CNN) and the *Long-Short-Term Memory* (LSTM) with its further implementation, the *Bidirectional LSTM* (BiLSTM)
The architecture of the DNN hereby used was made of 5 dense layers, having 684, 266, 150, 88 and 36 neurons respectively.
The second proposed architecture was the CNN, having 3 convolutional layers with 2,196 and 92 neurons respectively and 'ReLu' acitvation function.
The third proposed model consisted in the same architecture of the CNN previously quoted (with the first layer's activation function turned to 'SeLu') to which two dense layers with 1176 and 1024 layers were concatenated.
LSTM and BiLSTM were both composed of two layers, each one having 128 cells, and a dropout one with coefficient=0.5 (meaning that in half of the model neurons the output is nullified).
Every DL model was also equipped with the class weights, 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*' (the validation set error, computed at every epoch), with $patience = 10$ (meaning that if, within 10 consecutive epochs, a *val_loss* lower than the first epoch's one is not reached, the training is stopped).
The metrics adopted for performance evaluation were: the *accuracy* (acc) and the *F1 score*. In addiction, the *Receiver Operating Characteristic* (ROC) curve and its *Area Under Curve* (AUC) were computed, and also the significancy of the results was evaluated by calculating the p-value (considered significant whether lower than 0.05); 1000 random permutations of the label vector were performed and, for each of them, 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: $pvalue = C + 1/1001$. Both F1 score and AUC were macro averaged.

3. Results and Discussions

Results were calculated according to both the patients and therapist reported outcomes, using the HRV and EDA parameters recorded during the session.

3.1. Labelling strategies

In both cases, labels resulted to be highly unbalanced.
The labelling strategy adopted on the therapist

reported outcome led to the following distribution: 50 samples in class '2', 34 in class '3' , 10 in class '1' and 4 in class '4'. Due to the extremely low number of samples of class '4', it was finally decided to discard it from the analysis.
On the other hand, according to patient reported outcome the sessions were labelled as follows: 64 samples in class '2', 25 in class '1' and 9 in class '0'. In this analysis, the Krippendorf's α coefficient was equal to 0.8095, confirming a high level of inter-raters agreement.

3.2. Feature selection

The statistical analysis performed on the EDA and HRV parameters revealed that 23 variables had a purely randomic distribution and 14 a lognormal one. According to the Spearman correlation coefficients calculated, 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)', coherently with the related literature [5].
Regarding PCA, in the PCA_{80} case the first 11 PC were selected, while in the PCA_{90} one the first 15 were considered.
In the NCA, projections 0, 4, 21, and 22 were chosen.

3.3. Data upsampling

The 12 minutes window with no overlap had the lowest MCV value (0.948) and thus was selected. After upsampling, in the therapist reported case, the following distribution was derived: 87 samples 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, it was decided to discard class '4'. On the other hand, the patient reported labelling brought 110 samples in class '2', 44 in class '1' and 14 in 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. Algorithms in this approach provided better metrics with respect to the DL ones. Generally feature selection techniques showed to be helpful for the algorithms, confirming that the reduction of input size allows a more accurate hand-crafted feature extraction, easing the task. In detail, better results were obtained with the SVM family, as the KNN is more subjected to overfitting when dealing with very unbalanced labels.
With therapist reported labels, the best performance was achieved by the NuSVC with NCA (acc=0.63, F1=0.62, AUC=0.7, p-value=0.001)
On the other hand, in patient reported labelling the best result was reached by SVC with PCA_{80} (acc=0.71, F1=0.58, AUC=0.37, p-value=0.001). Confusion matrices and ROC curves of both strategies are reported in Fig. 2 and 3.

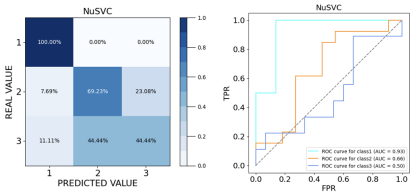

Figure 2: Metrics of ML best therapist reported prediction

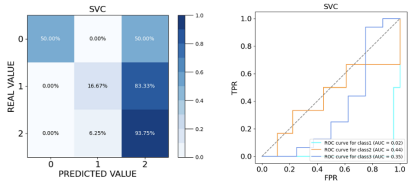

Figure 3: Metrics of ML best patient reported prediction

The patient reported labelling strategy led to a larger amount of significant predictions (hence of p-values lower than 0.05). This can be explained by the fact that the self-reported outcome may be more significant with respect to the therapist reported one to assess the emotional 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, while in the second one also data upsampling was applied.
This approach led to worse results rather than ML ones, which was expected due to the poor amount of data for a DL analysis, that does not allow an accurate data-driven feature extraction.

Upsampling did not improve the metrics with respect to the ones obtained with original data, which may be related to the fact that the segmentation could introduce a higher variability among the samples, and also led to non significant predictions, meaning that the chosen upsampling technique may introduce some noise in the data. In detail, neural networks achieved better results than LSTM, probably because the dataset was not hefty enough to perform a consistent encoding with the virtual memory cells.
Also in this case the distinction between the two labelling strategies was done. The best therapist reported labels prediction was provided by the CNN with no upsampling (acc=0.58, F1=0.56, AUC=0.65, p-value=0.001), while patient reported labels experienced the best prediction with the DNN with upsampling (acc=0.73, F1=0.46, AUC=0.62, p-value=0.001), although it obtained a null performance on label '0'. Fig. 4 and 5 show the confusion matrices and the ROC curves for both the strategies.

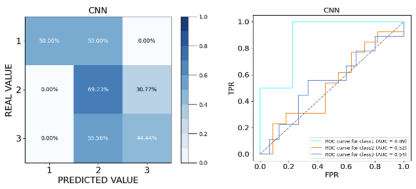

Figure 4: Metrics of DL best therapist reported prediction

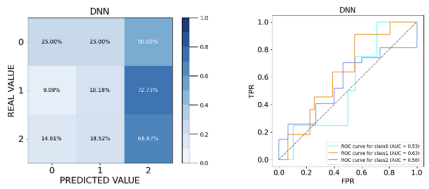

Figure 5: Metrics of DL best patient reported prediction

4. Conclusions

This master thesis work was developed due to the lack of deep investigation, in the neuromotor rehabilitation field, on the mental wellbeing of the patients.
Patient and therapist reported questionnaires were analyzed in order to derive a numerical label of patient emotional wellbeing and engagement. Algorithms belonging to both ML and DL were trained with EDA and HRV parameters to classify each session in which signals were acquired. Some models revealed the capability to distinguish different emotional states, achieving significantly higher accuracies than Nagae *et al.*; on the other hand, a lower macro avergaed F1 score with respect to Dalmeida *et al.* was obtained [3, 4]. This highlights the importance to further investigate the psychological response of the subjects to the treatment.
Despite this, there are still some limitations. Both in patient and in therapist reported case the labels were derived starting from *ad hoc* questionnaires, which do not belong to clinical practice. Also, the amount of available data was small, which arises the necessity of collecting more recordings to improve the performances, especially in DL.
A possible future development of this research may consist in the implementation of a real-time prediction system, so that therapy is customized to patient's needs, increasing his/her comfort.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.