POLITECNICO DI TORINO

DIPARTIMENTO DI INGEGNERIA MECCANICA E AEROSPAZIALE
(DIMEAS)

Corso di Laurea in Ingegneria Biomedica

Tesi di Laurea Magistrale

Using Wearable Technologies and Machine Learning methods to Estimate Clinical Outcomes for Acquired Brain Injury Rehabilitation

POLITECNICO
DI TORINO

HARVARD
MEDICAL SCHOOL

Relatore
Prof. Danilo Demarchi
Prof. Paolo Bonato
Correlatore:
Dr. Federico Parisi

Studente
Alberto Stella
matricola: 262981

Anno accademico 2020 – 2021

Abstract

The development of personalized rehabilitation strategies for patients with hemiparesis is fundamental to achieve the most effective outcome from the treatments. Clinicians are fully aware of the fact that the patients' responsiveness to an intervention is extremely subjective and that the need to quantitatively track their motor-gains is evident. Wearable sensing technology can meet this demand through cost-effective and flexible solutions, enabling accurate assessments of movement quality and motor impairment.
The approach proposed in this thesis relies on machine learning-based algorithms to estimate clinical scores through the analysis of wearable accelerometers data collected during the performance of Activities of Daily Living (ADLs).
The purpose of the study is to build predictive models able to mimic the evaluation criteria currently used by clinicians, in order to define the recovery trajectory of Stroke survivors and Traumatic Brain Injury patients.
Among the numerous assessment scales developed in the past years, in this project the upper limb Fugl-Meyer assessment (FMA) scale is used to quantify the severity of motor impairments, and the Functional Ability Scale (FAS) is used to evaluate the quality of movement.
3-axial accelerometers data are preprocessed with Digital Signal Processing (DSP) methods, such as segmentation and filtering, and subsequently analyzed with the aim of extracting informative features capable of defining movements properties of the study participants. Through a feature selection process, only the relevant characteristics are kept with the intention of discarding noisy and redundant data. The estimation of the clinical scores is done training and validating a regressive model using a Random Forest algorithm. Finally, the regression equation, relating the actual scores provided by the clinician and the predicted scores, is derived. In order to assess the

accuracy of the algorithms, two regression problems evaluation metrics are computed: the root-mean-square error (RMSE) and the coefficient of determination (R^2).

The results show that this approach is effective and that it is possible to estimate patients' rehabilitation outcomes in terms of both the movement quality and the motor impairment with a good level of accuracy and considering ADL tasks. The model performance are in line with the standards as R^2 of 0.83 and 0.78 are reached for FAS and FMA, respectively. This is solely achieved through the analysis of wearable accelerometers data, whereas preliminary analyses show that slightly better results can be obtained adding clinical and demographic information about patients. These findings pave the way for further studies that will presumably focus on reducing the number of sensors needed for the motor assessment and on moving the recordings from a clinical setting towards a home-based scenario.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Context

Together with the life expectancy increase, also the disability prevalence recorded an increment. It is expected that in the next years the demand for rehabilitation interventions will keep growing. Acquired brain injury (ABI), which includes stroke and traumatic brain injury (TBI), contributes a lot to the problem and it is often associated with severe disability such as upper-limb motor impairment.
The correlated loss in terms of independence and quality of life causes a significant societal burden that has to be resized with a view of facing a more critical situation soon.
Understanding if a rehabilitation program is having success with a specific patient is key, but increasing the number of neurological examinations is economically not viable. These assessments are time-consuming and it is really hard to schedule multiple clinical evaluations [1]. Thus, outcome measures are usually collected only at baseline and at the end of the treatment. It would be much better if the motor gains were estimated with a higher temporal frequency in parallel with the course of the rehabilitation program.
Although there is a plethora of treatment strategies to manage motor impairments and activity limitations caused by ABI, most are only supported by limited evidence pointing to the need for studies of improved methodological quality in this area [2].
The tailoring of medical treatment to the individual characteristics of each

patient is fundamental to address the problem of the high variability, noticed across different subjects, in response to an intervention. This is called "precision medicine" and, currently, it is a major topic in the field of rehabilitation. Wearable sensing technology has great potential in this context since it enables the collection of quantitative data in a flexible and cost-effective manner and with minimal supervision by clinical personnel, creating new opportunities for autonomous and remote monitoring outside the clinics.

1.2 Previous Studies

During the last 10 years, multiple methods to extract clinical information from wearable sensing devices, such as accelerometers, have been successfully proposed. Using the assessment scales, accurate estimates of upper-limb impairments from data collected during the performance of functional motor tasks have been reached [3]. Previous studies have shown that wearable sensing technology is suitable to monitor clinical outcomes [4] and that clinical scales, such as the Functional Ability Scale (FAS) and the Fugl-Meyer assessment (FMA) scale, fit the purpose of quantifying motor quality and impairment [5].

In most of the studies, data were collected during the performance of standardized tasks that have been considered to be able to sufficiently describe relevant movement substructures. FAS scores have been precisely predicted processing, with machine learning algorithms, the data recorded during the performance of 15 motor tasks that involved reaching and manipulating objects [3].
Later, the Wolf Motor Function Test, in particular, gained a lot of attention since, among the proposed tasks, some enable the analysis of upper extremity gross arm movements and fine motor control. The test is composed of 17 items progressing from proximal to distal and from least to most complex upper-limb movements and each item is used to assess speed and movement quality [6]. Clinicians consider those kinds of gestures really informative regarding the severity of the patients' disabilities, which is why they widely

evaluate their subjects' dexterity, strength and upper extremity function using WMFT timed functional tasks.
Using an approach based on WMFT tasks and machine learning algorithms, the derived clinical scores estimations are satisfying for both FAS and FMA [1], but this procedure requires patients' supervision and a clinical setting, which lead to costs, especially in terms of time.

1.3 Outline of the study

Tracking the outcomes of a specific rehabilitation treatment involves assessing improvements in patients' independence and in their ability to take care of themselves. These factors are key in the evaluation of a recovery program as they summarize if the quality of life of the subject is getting better.
Reduction in the time and effort spent performing real daily tasks could result in more time and energy to engage in life roles, despite a continued need for assistance [7]. Moreover, in order to reduce healthcare costs in general, not considering the independence of the patient without tasks that are enough descriptive to assess it, certainly leads to an inaccurate evaluation of the rehabilitation outcomes. Activity of Daily Living (ADL) disabilities, in fact, are often associated with higher rates of healthcare utilization, including hospitalization that, in turn, is associated with worsening ADL disability and nursing home placement [8].
The purpose of this thesis project is to extend the analysis carried out in the previous studies, concerning the estimation of clinical outcomes for rehabilitation, to Activity of Daily Living tasks. Compared to WMFTs, ADLs should lead to a more concrete estimation of the patient's independence and of his ability to perform self-care tasks. Moreover, in comparison with WMFT, ADLs tasks are less constrained as they are not standardized; this leads to a higher variability among signals recorded during the performance of the same tasks: each subject has, in fact, his manner to carry out a specific task. Leaving freedom of performing the tasks as the subjects would do on their own, certainly affects the results in terms of performance, but, on the other side, the estimations are more descriptive and realistic about the condition of the patients as they are evaluated on their way of accomplishing ADLs tasks.

Chapter 2

Motor Assessment in patients with Acquired Brain Injury

2.1 Overview

Acquired Brain Injury (ABI) is an umbrella term that includes different pathologies which share the presence of severe disabilities and the need for rehabilitation interventions. It encompasses various causes, such as vascular (e.g. stroke) and traumatic (e.g. Traumatic Brain Injury) [9].

2.2 ABI Epidemiology

2.2.1 Stroke

Stroke is a common cause of disability that is currently ranked as the fifth leading cause of death in the United States and it is estimated that it will be the fourth most common cause of disability in western countries by 2030 [10, 11].
Strokes can be distinguished between ischaemic (occlusion of a blood vessel), which constitute 80-85% of the cases, and hemorrhagic (rupture of a blood vessel). Moreover, even if the rate of recurrence reduced until the mid-2000s,

it has not changed over the last decade; the risk of recurrence 7 days post-stroke is 2% while, at 5 years, the risk of recurrence or death is 36% for small-vessel occlusion strokes and 27% for other ischemic causes [12].
In the last few years, nearly 3-4% of the healthcare costs in the Western countries can be attributable to stroke and, on average, the total lifetime expense per stroke patient is estimated around $140,000 in the United States [13]. In order to cut a part of these costs and to help stroke survivors to reduce their disabilities, rehabilitation becomes a critical aspect of the continuum of care. Thus, designing a comprehensive rehabilitation program results essential. In particular, stroke rehabilitation is a process that aims to prevent deterioration of functions, to recover, if possible, some of them, and to achieve the highest level of independence within the limits of the persistent stroke impairments [14]. Providing treatment and training to stroke survivors, many regain and relearn skills of everyday living, obtaining greater independence and improving functional capacity.

2.2.2 Traumatic Brain Injury

Traumatic brain injury (TBI) constitutes a major health and socioeconomic problem. According to the US Centers for Disease Control and Prevention (CDC), it is caused by a bump, blow, or jolt to the head or a penetrating head injury that disrupts the normal function of the brain [15]. It has been recently defined as: 'An alteration in brain function, or other evidence of brain pathology, caused by an external force' [16, 17].
The direct cost of TBI in 2000, which includes death, treatment for both hospitalized and non-hospitalized patients, was estimated to be around 9 billion US dollars [15]. In addition, it is the most common cause of disabilities in people under 35 years old [18], as it is often associated with car accidents. The long-term outcome varies according to the severity of the pathology and so the mortality of TBI: it is around 2.5% for moderate and nearly 33% for severe. Furthermore, around 1.7 million people in the USA currently suffer TBI and live with disabilities caused by it.
In this case too, the rehabilitation program is of great importance as it can help the patient to achieve the maximum degree of independence within limits imposed by their residual physical, functional and cognitive impairments.

2.3 Precision Medicine

Among a broad range of scientific areas in which precision medicine can be applied, it attained great success in the design of the rehabilitation programs in which the response of the patients is affected by high variance among subjects with the same pathology. It is the typical clinical scenario of stroke survivors and TBI patients' rehabilitation management.
In this situation the term evolves into a more suitable one, "precise rehabilitation": data-driven decisions are usually taken according to the information extracted from various sensing technologies, such as wearable sensors data. Evaluating the evolution of the patient's clinical scores over time, it is possible to understand the subject's response to a specific rehabilitation intervention. Furthermore, the more frequent the recordings are, the more precise and dynamic are the adjustments that the clinicians may apply to the patient's rehabilitation program.
Through the modern concept of precision medicine, the patient's heterogeneity is taken as advantage, using data-driven methods, to improve intervention with the purpose of providing the most suitable treatment to the right patient at the right time. In 2015, with president Barack Obama announcement about Precision Medicine Initiative, it became a priority of the United States [19].
A core concept related to the precision medicine is the "Dynamic treatment regime": decision making is defined as a sequence of decision rules, one per time interval, that characterize how the intervention will be tailored according to the response of the patient to the treatment he has been subjected up to that moment[20]. The decision timestamps depend a lot on the specific patient's features and they can be scheduled at the beginning of the treatment or time by time taking into account the patient's outcomes.
Being able to analyze big medical data is core if we want to progress in precision rehabilitation and machine learning (ML) and artificial intelligence (AI) can hugely help out.

2.3.1 Medical Big Data

During the recent decades, the fast increase in the production volume of digital data together with the development of new computational methods enabled the scientific community to use these large data sets, known as big data, for innovative purposes. Also the healthcare field has benefited from this technology, albeit with some delays compared to other disciplines. This can be explained as the consequence of several factors, such as the poor management of insights from research, the poor usage of the available evidence, the poor capture of care experience, and the difficulty in accessing medical data as they have to be treated differently from other fields [21]. Moreover, the cost in the production of clinical data may result much higher than in other contexts: clinicians are often involved, the medical instrumentation may be expensive, and the situation in which the data are collected may be not reproducible.
Among the different healthcare areas in which the use of big data may result in awesome outcomes, the potential value has been concretely expressed in:

- Precision medicine ([22], [23]) and precision rehabilitation ([24], [25])
- Analysis of medical images using computer vision methods in order to support clinicians' decision making ([26], [27], [28])
- Tailoring diagnostic, treatment decisions and telemedicine using mobile health technologies ([29], [30])
- Population health analysis ([31], [32])

2.3.2 Big Data Analysis

Data collection alone cannot result in anything useful: big data have to be processed in order to extract new insights. Machine learning (ML) and artificial intelligence (AI) algorithms emerge as the enabling technologies that make the analysis of these large data sets fruitful.
The types of learning used by computers are formally divided into 2 typologies [33]:

- Supervised Learning: it concentrates on classification, which consists of choosing among subgroups to best describe a new instance of data, on prediction, which involves estimating an unknown parameter, and, as in this thesis, on regression problems, i.e. the estimation of a continuous outcome variable. The core idea is that the computer learns how to map an input to an output based on example input-output pairs.

- Unsupervised Learning: as opposed to the previous case, the goal is to find recurring patterns or groupings within the data. Here the labels of the observations are unknown and the algorithm will use the information contained in the data to find hidden structures and organize the dataset into subgroups (also known as clusters in this field).

In this thesis, machine learning-based algorithms are used to handle a supervised regression problem. The observations in the data set, patients' data, are labeled with the clinical scores of the standardized clinical assessment scales, which are continuous variables. The purpose, in fact, is to predict the patients' clinical scores after they have received a treatment to understand if it has been successful.

2.4 Clinical Assessment Scales

The need for determining in a quantitative way the limitations in function of patients with disabilities led to the development of standardized clinical scales. Through their use, in rehabilitation, the assessment of the level of independence and the ability to perform basic daily living functions of impaired patients has been achieved [34].
In this study 2 clinical scales have been used and they will be described in the following: the Functional Ability Scale (FAS) and the Fugl-Meyer Assessment (FMA) scale.

2.4.1 Functional Ability Scale (FAS)

FAS is a 6-point standardized scale, from 0 to 5, used to visually rate the performance of upper extremity (UE) functional tasks in terms of movement quality. Typically, the scores of the single items are summed to obtain a total score. For instance, the Wolf Motor Function Test (WMFT) consists of a battery of 15 motor tasks, thus the total score is rated out of 75 points [35].
The following table 2.1 shows the rating criteria for evaluating the quality of the performed items.

Rating	Description
0	Does not attempt with more-involved upper extremity (UE)
1	More-involved UE does not participate functionally; however, attempt is made to use the UE. In unilateral tasks the less-involved UE may be used to move the more-involved UE
2	Attempts to use more-involved UE but requires assistance of the less-involved UE for minor readjustments or change of position, or requires more than 2 attempts to complete, or accomplishes very slowly. In bilateral tasks, the more-involved UE may serve only as a helper
3	Attempts to use more-involved UE, but movements are influenced to some degree by synergy or are performed slowly or with effort
4	Attempts to use more-involved UE; movement is close to normal,[a] but slightly slower; may lack precision, fine coordination, or fluidity
5	Attempts to use more-involved UE; movements appear to be normal[a]

[a]For the determination of "normal," the less-involved UE can be used as an available index for comparison, with premorbid UE dominance taken into consideration.

Table 2.1: The Functional Ability Scale (FAS) [35].

2.4.2 Fugl-Meyer Assessment (FMA) scale

The Fugl-Meyer assessment is a measure used to grade the impairment of a patient during the performance of motor tasks. Specifically, it was developed as an evaluative measure of recovery from hemiplegic stroke [36]. Hemiparesis is the most diffused disabling deficit among stroke survivors and, in fact, it affects 70% to 80% of the patients; moreover, it is often the one that needs a rehabilitation intervention the most. The purpose of its development in 1975 was to compensate for the lack of a clinical scale that considered "the neuromuscular capacity per se" [37], that standardized the patients' posture, and that took into account patient's compensatory mechanisms.
Conventionally, the impairment is defined as any loss or abnormality in psychological, physiological, or anatomical structure or function [38].
The FMA is a 3-point ordinal scale and the maximum attainable motor performance score for the upper extremity section is 66 points.
As in the previous case, the following table 2.2 shows the evaluation criteria of the concerned scale.

Rating	Description
0	Detail is not performed
1	Detail is performed partially
2	Detail is performed completely

Table 2.2: Fugl-Meyer Assessment (FMA) scale.

Upper Extremity (66 points)
Shoulder retraction
Shoulder elevation
Shoulder abduction
Shoulder abduction to 90 degrees
Shoulder adduction/internal rotation
Shoulder external rotation
Shoulder flexion 0–90 degrees
Shoulder flexion 90–180 degrees
Elbow flexion
Elbow extension
Forearm supination
Forearm pronation
Forearm supination/pronation (elbow at 0 degrees)
Forearm supination/pronation (elbow at 90 degrees, shoulder at 0 degrees)
Hand to lumbar spine
Wrist flexion/extension (elbow at 0 degrees)
Wrist flexion/extension (elbow at 90 degrees)
Wrist extension against resistance (elbow at 0 degrees)
Wrist extension against resistance (elbow at 90 degrees)
Wrist circumduction
Finger flexion
Finger extension
Extension of MCP joints, flexion of PIPs/DIPs
Thumb adduction
Thumb opposition
Grasp cylinder
Grasp tennis ball
Finger-nose speed
Finger-nose tremor
Finger-nose dysmetria
Finger flexion reflex
Biceps reflex
Triceps reflex

Table 2.3: FMA scale items [36].

2.5 Wearable Technology in Rehabilitation

Wearable technology includes a plethora of devices that share the main feature of being worn or attached to a body segment. This technology is trying to satisfy different needs that presented at the beginning of the current millennium, such as the needs for taking care of an increasing number of patients with chronic diseases, for giving care to people in areas in which the access to providers is limited, and for maximizing the independence of an increasing number of subjects with severe disabilities. Wearable technology can help to satisfy these needs with cost-effective applications in diagnostic and monitoring. Home-based patients, for instance, can be remotely and continuously monitored in order to assist them at their place, extending the reach of the experts to rural areas. Wearable technology, as presented in this thesis, can be one of the enabling technologies for precision medicine and, in particular, for assessing the effectiveness of rehabilitation in ABI patients.
Among the multitude of wearable devices, in this thesis only inertial sensors will be considered, as they are the ones used in the project.
Accelerometers have been widely implemented due to their compact size, their low-power requirement, low cost, non-intrusiveness and capacity to provide reliable data concerning the motion of people [39]. In particular, Micro Electro Mechanical System (MEMS) sensors, when implemented in microelectronic circuits, can be used to measure the acceleration. In capacitive-based MEMS accelerometers, the acceleration is calculated by measuring the change in capacitance due to a moving plate or sensing element [40]. Thanks to their high sensitivity and resolution, in respect of piezoresistive accelerometers, these devices are widely used in many commercial applications.
Inertial Measurement Units (IMUs), that are the electronic devices mainly used to measure velocity, orientation, and gravitational force, usually include both accelerometers, gyroscopes and sometimes magnetometers [41]; it is worth pointing out that, in this thesis, only the IMU accelerometers are used, since also the device energy consumption has to be considered when there is the will to perform long recordings without worrying about charging; in home-based scenarios, having a more energetically efficient device, at the expenses of extremely precise measurements, is a common choice since the high performance of combining different sensors would not be exploited

in this application, while reducing the risk of interrupting the recordings is preferred [42].
Typically, the latest devices of this kind adopt wireless communication, making their wearing and placement extremely easy even for subjects with low motor deficits.
Accelerometers, such as those used in this thesis, provide one separated data time series for each axis A_x, A_y, A_z; the magnitude of a 3-axial accelerometer can be calculated as follows: $A_m = \sqrt{A_x + A_y + A_z}$.
During the last decade, several results have been reached using accelerometers to estimate outcomes of upper limb rehabilitation programs, usually through clinical assessment scales. In 2010 [43], body-worn accelerometers data, recorded during the performance of a set of functional motor tasks, were used to estimate movement quality, in terms of FAS, through a machine learning-based algorithm; remarkable performance were achieved, in fact the predictions were extremely accurate and the model was nearly not affected by bias at all, but only stroke survivors were enrolled. In 2011 [5], wearable sensor data collected during the performance of items belonging to WMFT were used to assess motor impairments, estimating FMA clinical scores. Those predictions had a mean error of 4.74 points out of 66 of the total FMA score, which showed that the track was right, but results had to be improved. Also here, the subjects that were enrolled in the study were stroke survivors, but subsequent studies manifested the will to include other pathologies that share the fact of causing upper limb motor deficits, not focusing only on stroke, but extending also to subjects affected, for example, by traumatic brain injury. In 2020 [44], it was possible to assess movement quality through FAS even adding patients with TBI to the dataset; data were recorded during the performance of 15 standardized tasks from the WMFT. During the same year [1], a dataset, composed of stroke survivors and TBI patients, was used to accurately estimate both movement quality and motor impairment; wearable accelerometers were placed on the patients with the purpose of recording their movements during the performance of 8 standardized tasks from WMFT. FMA predictions were marked by a high coefficient of determination $R^2 = 0.86$ and a $RMSE = 3.99$ points, while for FAS predictions $RMSE = 0.38$ points, coefficient of determination $R^2 = 0.79$ were

reached. These numbers that describe the performance of the method proposed by [1], are used as a standard metric, in order to evaluate the goodness of the model proposed in this thesis.

Chapter 3

Materials and Methods

3.1 Data Collection

3.1.1 Study Partecipants

Subjects have been recruited in order to obtain a heterogeneous sample for a prospective longitudinal study: the group has been followed over time and new data have been collected.
The inclusion criteria were:

- Unilateral stroke, both ischaemic and hemorrhagic
- Focal Traumatic Brain Injury (includes scalp injury, skull fracture, and surface contusions, generally caused by contact)
- 18-80 years old at the recruitment
- Involved in an upper-limb rehabilitation program, both inpatients and outpatients
- Moderate-severe upper-limb impairment evaluated through the upper extremity total FMA score

Spaulding Rehabilitation Hospital Institutional Review Board (IRB) reviewed and approved the study procedures, which have been executed according to relevant guidelines and regulations. Furthermore, signed informed consent has been asked to each study participant or to his legally authorized delegate.

The study participants have been called for 2 recording visits: one at the beginning of the treatment (baseline) and one at the end of the treatment (discharge).

3.1.2 Wearable Accelerometers placement

During the recordings, the considered subjects have been equipped with body-worn 3-axial accelerometers (Shimmer3 by Shimmer Sensing, Dublin). The units were placed, as shown in the figure 3.1:

- Chest, sternum height
- Upper arm, mid-biceps, frontal
- Wrist, above radius and cubitus styloid, dorsal
- Index and thumb, dorsal part of the distal phalange

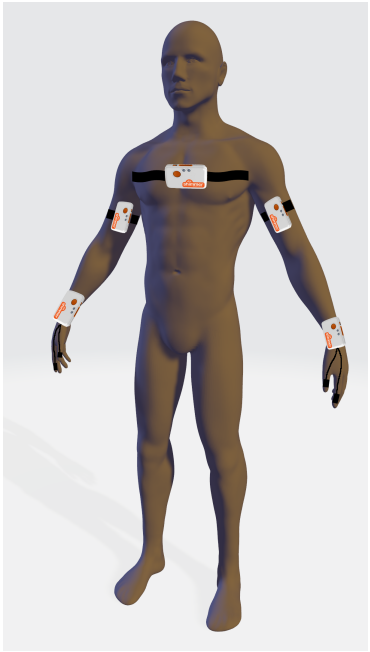

Figure 3.1: Wearable Shimmer3 sensors placement.

In the table below 3.1 the specifications of the used IMUs are reported together with a picture of the device 3.2; the integred accelerometers are well known to be ultra-compact, ultra-low power and able to provide highly accurate and scientifically reliable raw data.

Shimmer3 Accelerometers	
Range	$\pm 2g$, $\pm 4g$, $\pm 8g$, $\pm 16g$
Sensitivity	1671 LSB/g $at \pm 2g$
Numeric Resolution	$14 - bit$
Typical Operating Current	$\leq 162\mu A$
RMS Noise	$0.6mg$ $at \pm 2g$
Sampling frequency	$51.2Hz$

Table 3.1: Shimmer3 accelerometers specifications: STMicro LSM303AGR are the wide range MEMS accelerometers used in this thesis.

Figure 3.2: Shimmer3 IMU used in this thesis.

3.1.3 Activity of Daily Living (ADL) Tasks

During the data collection process, a research therapist guided the subject through a battery of tasks that belong to the macro group of Activity of Daily Living tasks. Each task has been repeated multiple times, tracking the beginning and the end of each trial with capacitive sensors.
ADL tasks aim to describe as much as possible the level of the subject's independence, thus they are not isolated movements that can be implemented during different actions, but complete activities that require the coordination of different movement patterns.

ADL TASKS ORGANIZATION		
TASK N°	NAME	TRIALS
1	open bottle	3
2	brush hair	6
3	put on/take off pen cap	3
4	unfold towel	3
5	ironing	6
6	lift box	3
7	erase board	2
8	open door	6

Table 3.2: Activity of Daily Living tasks.

3.2 Estimation Process Pipeline

As described in the figure below 3.3, patients' accelerometers data undergo different processing steps before being translated into the clinical scores:

- Signal Processing: accelerometers signals are processed with different operations in order to improve the performance of further steps.

- Feature Extraction: with the aim of reducing the amount of data that will be processed, but maintaining the information of the original dataset, some variables are extracted from the processed accelerometers signals.

- Estimation Algorithm: after reducing the dimensionality of the variables through feature selection, giving as input some observations with their corresponding clinical score, the computer learns how to predict the unknowns clinical scores of new data.

FAS are firstly estimated from the accelerometers data and, then, considering the correlation between the two clinical scores, they are given as a further input to the FMA estimation algorithm, in order to improve the performance.

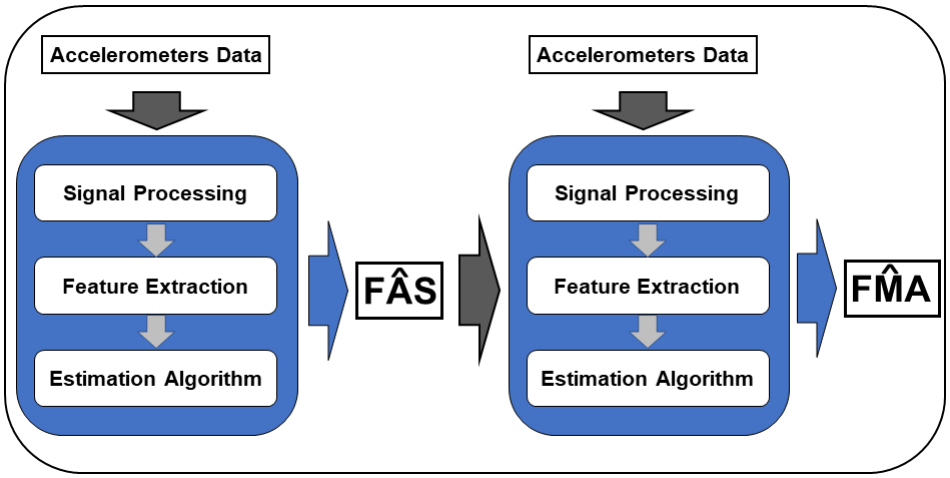

Figure 3.3: Algorithm pipeline.

3.3 Data Processing

The signal of each axis of each accelerometer is individually processed in order to keep only its relevant sections and to improve its quality. Multiple actions compose this step, starting from the filtering of the accelerometers time series, passing through the signal segmentation and finishing with the x-axis inversion.

3.3.1 Time series Filtering

The acceleration time series of each sensor and of each axis has been high pass filtered using a 6^{th} order Butterworth filter with the cutoff frequency at 0.5Hz. This was done in order to generate a gravity-free condition, to lose the orientation of the sensors that may vary among the subjects, to limit the impact of postural adjustments, and to attenuate the low-frequency integration drift: inertial systems, in fact, suffer from small errors in the measurement of acceleration, that, integrating, will result into progressively larger errors in velocity.

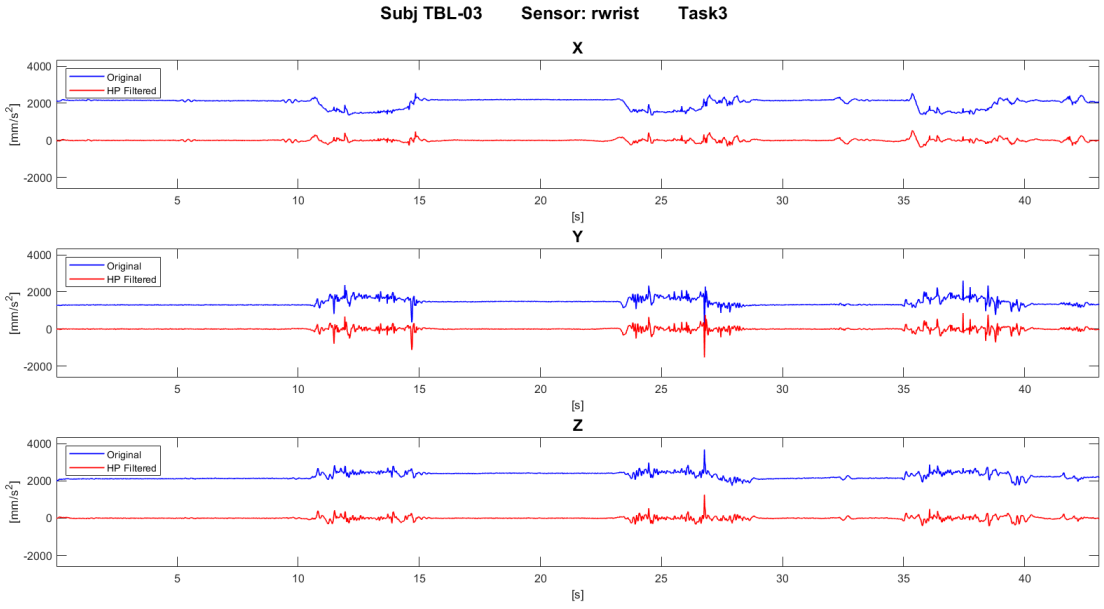

Figure 3.4: Example of right wrist high pass filtering with cutoff frequency of 0.5Hz; same subject of 3.5 performing the 3rd task.

3.3.2 Segmentation in Trials

During the recordings, using capacitive sensors, the beginning and the end of the tasks trials have been tracked. Thanks to this expedient, accelerometers time series have been automatically segmented in order to keep only the signal delimited by the couple of the reference markers. Moreover, considering that the subjects are hemiparetic, only the data related to the sensors of the affected side are taken into account, while the others are discarded.
As shown in the figure 3.5, the beginning of a certain task has been marked with a long press of the capacitive sensor, while the trials are graphically identified as the signal portions between its spikes.

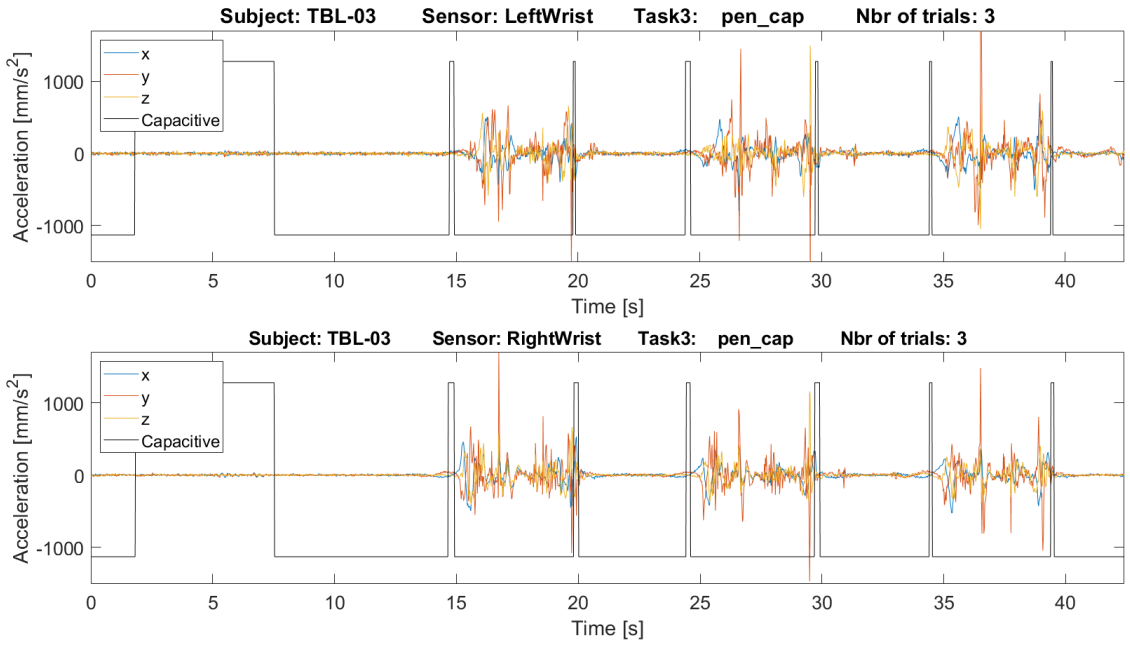

Figure 3.5: Example of 3-axial accelerometer signal segmentation: 21 aged subject with Traumatic Brain Injury and RIGHT affected side (also the dominant one).

3.3.3 X-axis inversion

Furthermore, to make the reference frame consistent, the x-axis of the sensors, positioned on the right upper arm and right wrist, has been inverted. In the figure 3.6 an example of this processing step is shown.

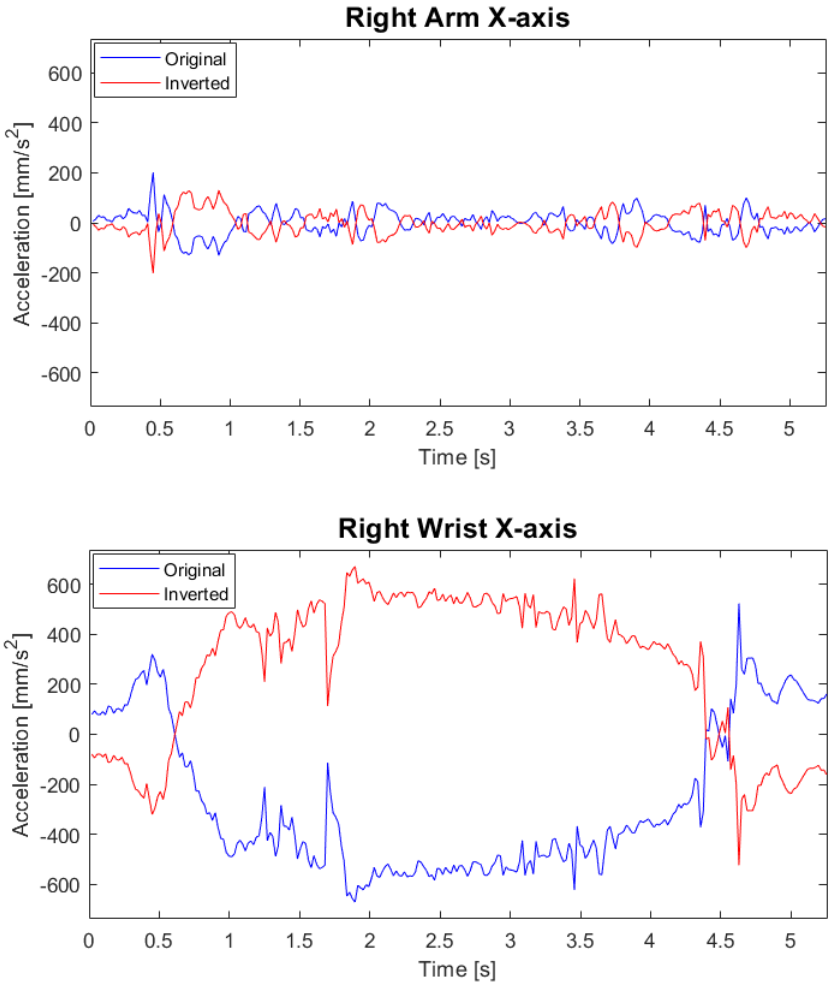

Figure 3.6: Example of right upper arm and right wrist X-axis inversion: same subject of 3.5 performing 1st trial of the 3rd task.

3.3.4 Removal of healthy side trials

Considering there is the will to keep only the data related to the affected side of the patient, it is needed to remove the trials performed with the healthy side. Tasks 2, 5, 7 and 8, which are brushing hair, ironing, erasing a board and opening a door, in fact, can be performed using only one limb. However, during the recordings, these tasks were firstly performed using the affected side and then with the healthy one.
As it is noticeable looking at the plots below, focusing only on the signal related to the affected side, it would be incorrect to include in the dataset the trials performed with the other limb since the contained information would be harmful.

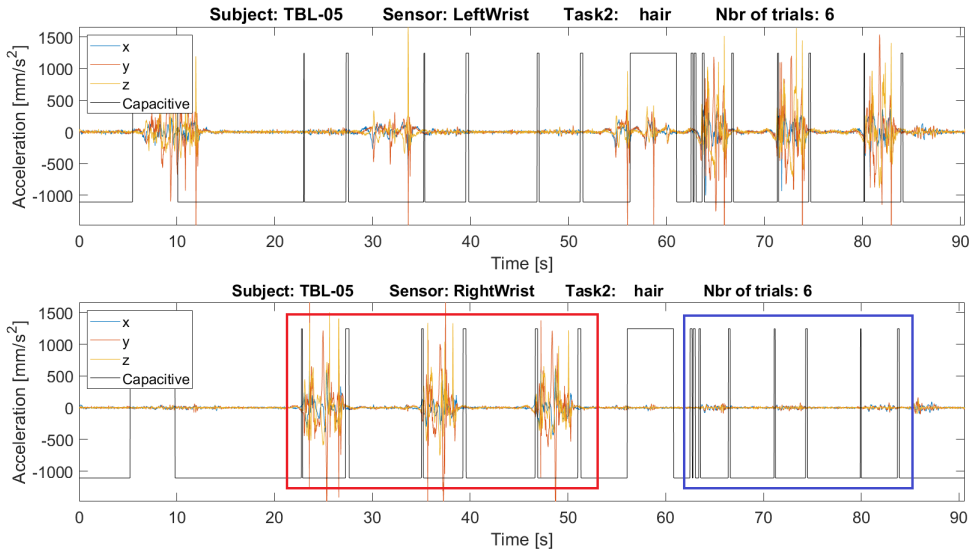

Figure 3.7: Example of removal of healthy side trials, in the blue rectangle, in a TBL subject with right affected side for task 2 (brushing hair); in the red rectangle the correct trials are shown as a comparison.

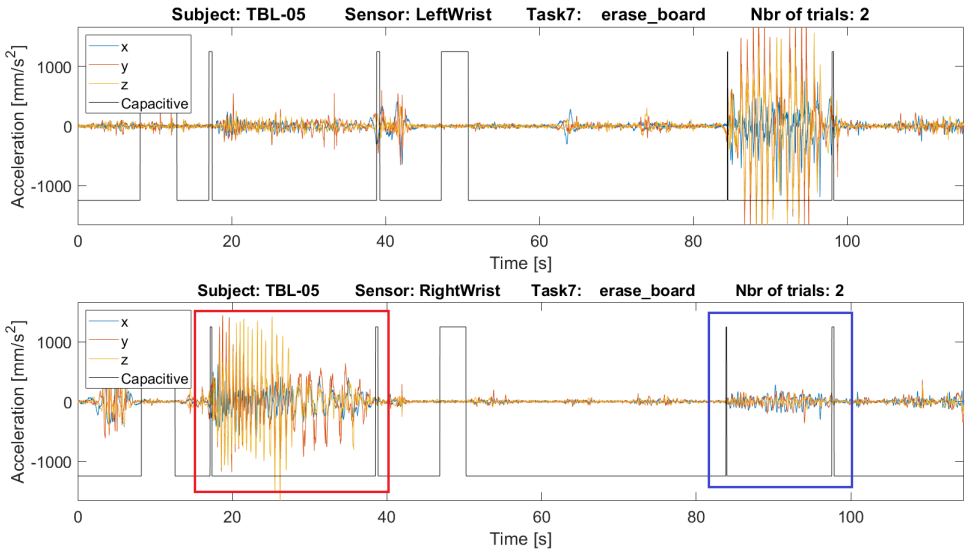

Figure 3.8: Example of removal of healthy side trials, in the blue rectangle, in a TBL subject with right affected side for task 7 (erasing a board); in the red rectangle the correct trials are shown as a comparison.

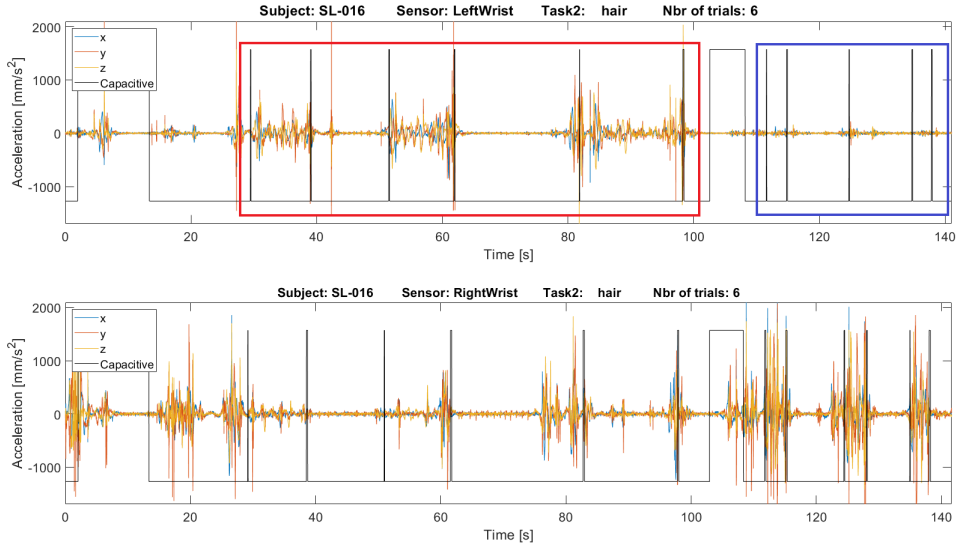

Figure 3.9: Example of removal of healthy side trials, in the blue rectangle, in a Stroke survivor with left affected side for task 2 (brushing hair); in the red rectangle the correct trials are shown as a comparison.

3.4 Feature Extraction

The step that follows the signal processing aims at reducing the data volume and capturing the main characteristics of raw signals from the accelerometers time series. For this scope, the features to extract from raw accelerometers data were chosen based on multiple previous work that showed correlation with clinical measures of functional capability ([1], [43], [45], [46], [47]).
To name a few, the use of mean and energy measures of acceleration has been shown to result in accurate recognition of certain postures and activities [48]. Then, the entropy of the signal is calculated since it may support the discrimination of activities with similar energy values [48].
In respect to the typical features belonging to the time and frequency domain, which are usually implemented in this kind of problem, further ones are added. In particular, Dynamic Time Warp (DTW) [49] is a well-known technique used to find the alignment between two time-series, warping them in a nonlinear fashion to match each other; its robustness against variation in speed or style in performing action [50] justifies the high computational cost that its calculation requires: remembering that ADLs are not standardized tasks and, so, that the way of accomplishing them is variable among different patients, DTW turns out to perfectly fit this problem.
Moreover, features measuring the correlation of acceleration between axes or sensors can improve recognition of activities involving movements of multiple body parts [48], [51]. Thus, correlation is calculated between each axis of each accelerometer and between all pairwise combinations of axes and sensors.
The total number of extracted features is 366 and all of them have been normalized to have zero mean and unity variance. In the following table 3.3 all the engineered features are listed, grouping them by field.

Feature Macrogroup	**Description**
Mean	Arithmetic Geometric Harmonic Interquantile 1st-3rd quartiles
Spread (dynamic energy)	RMS Interquantile range Absolute deviation
Power spectrum	Dominant frequency Ratio tot energy and secondary peak Ratio energy at dominant frequency and secondary peak energy Energy of secondary peak Energy at dominant frequency
Smoothness	Distance to filtered signal
Entropy	Signal entropy

Feature Macrogroup	**Description**
Kurtosis	Kurtosis
Speed	max RMS Mean
Skewness	Skewness
Dinamic Time Warp (DTW)	Distance between filtered and actual signal
Autocovariance	Range of autocovariance
Magnitude (LPF@ 8Hz)	Range magnitude acceleration Range magnitude speed Range magnitude displacement Max (speed, acceleration and displacement) RMS mean and standard deviation Entropy of acceleration Max frequency of magnitude
Magnitude derivative	Max, standard deviation Interquantile range, range

Feature Macrogroup	**Description**
Jerk	Jerk normalized by velocity RMS Max frequency jerk magnitude
Correlation	Between all sensors and all channels Between all sensors magnitudes

Table 3.3: Data features extracted from wearable sensors time series.

3.5 Feature Selection

"A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other."[52]

At first glance, having more features, that have been previously extracted to better describe the original signal, should intuitively result in more discriminating power. However, experience showed that, most of the times, the antecedent statement is not true: some features may poorly contribute to the predicting model or even harmfully in some cases, worsening the performance. The purpose of the feature selection process is to find the most descriptive subset of features among all the extracted ones. With this, the dimensionality of the dataset is reduced, the computational cost also benefits a lot. Different approaches for feature selection have been evaluated in terms of computational time saving and estimation performance:

- Correlation-based Feature Selection (CFS)
- ReliefF with Davies–Bouldin Index
- Minimum Redundancy Maximum Relevance (MRMR)
- Feature selection using Neighborhood Component Analysis (NCA)

CFS is preferred to the other methods since its execution is computationally light and since it is the most selective: the subsets include 10-15 features only, giving the possibility to make clinical intuitions interpreting which features are predictive of which task. Moreover, it enables to reach the best performance in terms of high accuracy and low estimation error.

3.5.1 Correlation-based Feature Selection (CFS)

Rational

Finding a subset of features that will be used to predict an outside variable, it is desirable that the members, which have been selected to constitute this group, will have low inter-correlations. Thus, choosing a subset of variables, with the aim of predicting another one, it is likely to select predictors that capture different aspects of the outside variable, while avoiding keeping redundant information between them.
Evaluating several composites, the correlation between the subset in analysis and the outside variable can be expressed as follows:

$$M_S = \frac{k\overline{r_{lf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{3.1}$$

- M_S : merit, correlation between the subset and the outside variable
- k: number of the features in the subset
- $\overline{r_{lf}}$: mean feature-label correlation
- $\overline{r_{ff}}$: average feature-feature intercorrelation

This equation, that results to describe the Pearson's correlation coefficient, shows that the correlation between the subset of features and the variable to predict depends on the number of the chosen variables and the magnitude of the inter-correlations among them, together with the magnitude of the correlations between the subset members and the outside variable [52].
3.1 is implemented in the feature selection process as a heuristic measure of the "merit" of feature subsets.

A Correlation-based Feature Selector

CFS belongs to the group of filter type feature selection algorithms and, using a correlation-based heuristic evaluation function, it ranks predictors subsets. Going into detail, a feature is accepted in the subset according to the extent to which it predicts the label in areas of the instance space not already covered by other features.
Analyzing 3.1, the numerator can be interpreted as an index of how predictive of the label is the subset of features in analysis, while the denominator represents the grade of redundancy among those features.
In this way, 3.1 is used to rank a subset of features that is under analysis, but it is interesting to understand how its members have been chosen: considering a high dimensional dataset, in fact, it would be particularly heavy to compute all the possible combinations of the features and so it is fundamental to understand how the subsets to be evaluated are searched and which are the starting point and the stopping criterion.
The CFS implemented in this thesis performs a greedy forward heuristic search (Greedy Stepwise) through the space of attributes subsets: it starts with no attributes in the subset, then it greedily adds one feature at a time, and it stops when the addition of any remaining attributes does not result in a higher evaluation of the merit function (3.1).
In practice, the feature selection algorithm calculates the correlation matrix of the features (K features dimensionality leads to $K \times K$ matrix) and the vector of the correlation between each feature and the label. The search starts with an empty set of features, intuitively evaluated with zero merit, then the feature with the highest correlation with the label is added and the merit of the new subset is evaluated. The following step consists in trying which one of the other features results to be the best together with the ones in the expanding subset (in this case composed of only one attribute, as it was the first iteration); this is achieved by evaluating the merits of all the possible subsets defined by the addition of a new feature: attributes which are low correlated with the ones already present in the subset but highly correlated with the label will be preferred. Until the addition of a new feature to the subset leads to an increase in the merit of the subset, the process continues, but the first time it does not, it ends and the best subset found is given as the output. The stopping criterion can be modified specifying a

threshold in the increment of the merit below which the improvement is not considered to be enough to add a new feature.
Correlation-based Feature Selector assumes that, given the label, the attributes are conditionally independent but it can perform well even if this assumption is moderately violated, while when strong feature interactions occur, it may fail to select all the relevant features [52].
In the following figure 3.10 it is schematically shown how the feature selection is usually integrated into a machine learning problem pipeline.

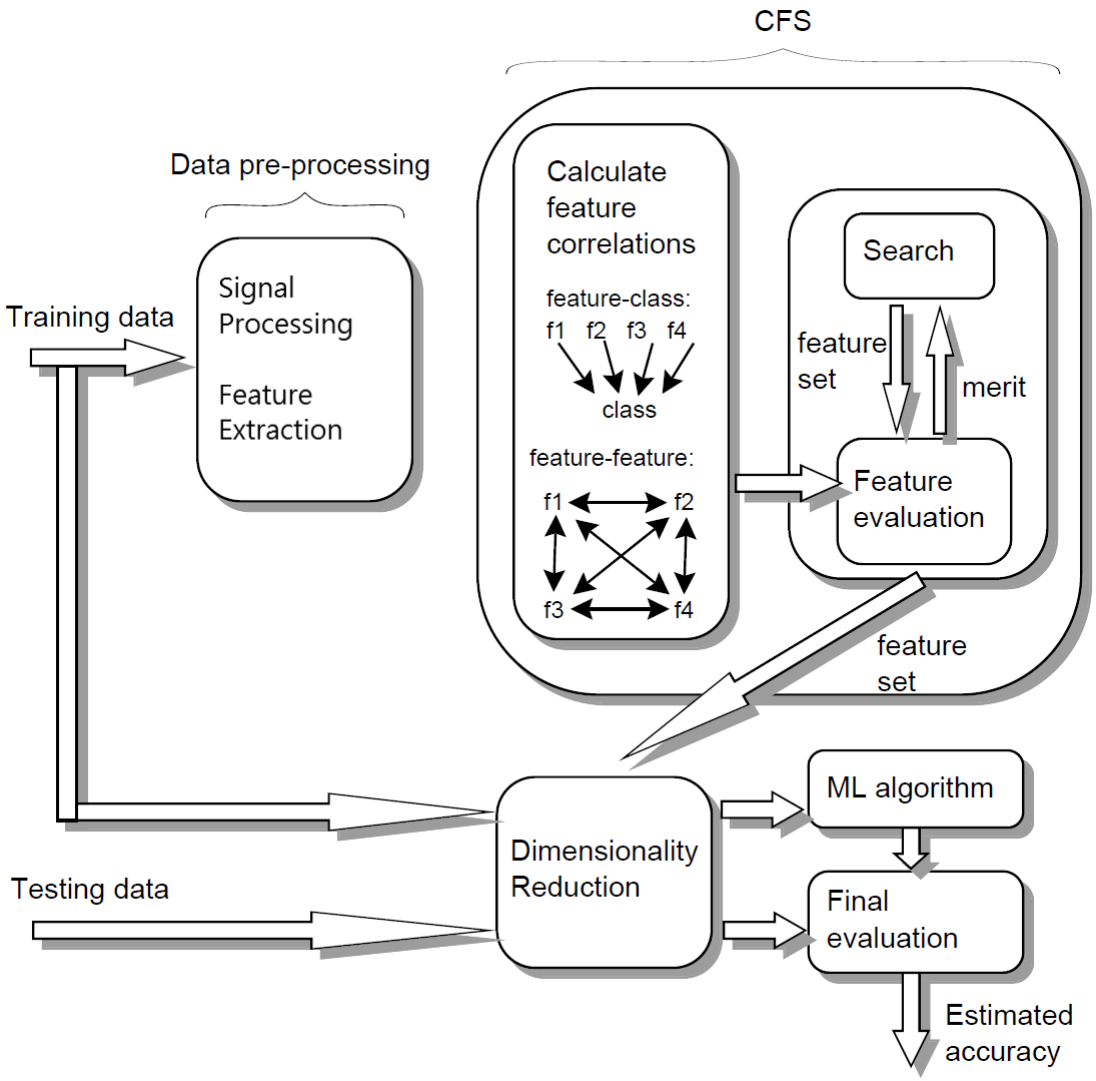

Figure 3.10: The role of the CFS in a typical Machine Learning approach. Figure adapted from [52].

3.6 Clinical Scores Estimation

Once only some of the features have been selected and so the dimensionality of the data has been reduced, it is possible to dive into the forecasting section of the pipeline. Here, using a machine learning algorithm, data can be translated into the clinical scores of our interest.
The estimation algorithm can be broken down into 2 steps:

- Single task predictions: data related to each ADL task is analyzed separately trying to estimate the clinical score of each task.

- Aggregating predictions: the single task predictions are combined in order to make a final estimation of the total clinical score that takes into account the results of all, or some, of the previous predictions.

Random forest, an ensemble of decision trees, is chosen to be used as the predicting algorithm [53].

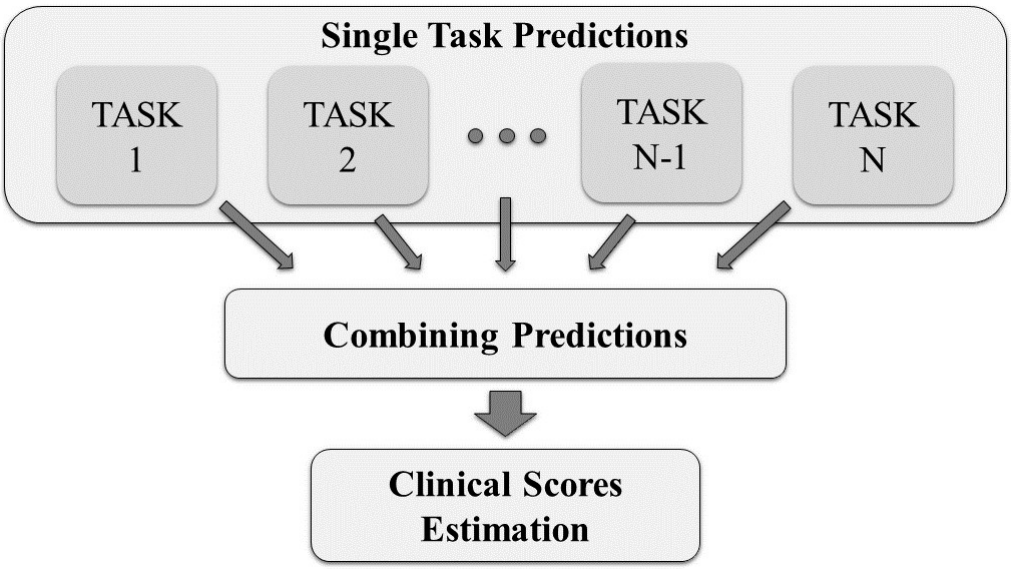

Figure 3.11: Clinical scores estimations pipeline.

3.6.1 Random Forest

Decision Trees

"A decision tree is a classifier expressed as a recursive partition of the instance space" [54]. It is composed of nodes: the first one is called "root" and has no incoming edges, while all the other nodes have only one incoming edge. If a node has outgoing edges, it is called internal, while all the other nodes without outgoing edges are called leaves, or terminal/decision nodes. In a decision tree, the role of each internal node is to split the feature space into two or more sub-spaces: in case of numeric features, decision trees can be geometrically interpreted as a collection of hyperplanes.
The tree complexity, which is inversely proportional to its comprehensibility, is commonly measured considering the total number of nodes or leaves, the tree depth and the number of features used. It can be controlled by the stopping criteria or pruning: while the first approaches resulted to be crude methods of terminating the growth, tending to degrade the tree's performance, an alternative one is to allow the tree to grow and then prune it back to an optimum size.
When evaluating which algorithm would be the most suitable for a machine learning problem, it is a good practice to think to its advantages and disadvantages; concerning to Decision Trees,the most important ones are reported in the following comparison chart 3.4:

Advantages	**Disadvantages**
Self-explanatory Convertible to a set of rules Handling datasets with errors Handling missing values	Perform poorly if there are many relevant attributes Over-sensitivity to the training set, to irrelevant attributes and to noise

Table 3.4: Decision trees characteristics.

Random Forest

A random forest is an ensemble machine learning method suitable for both classification and regression that constructs a multitude of decision trees. Typically, in a random forest with k_{trees} decision trees, these are the steps of the algorithm [53]:

1. k_{trees} bootstrap samples are taken from the dataset

2. for each of them, an unpruned tree is grown using only m_{feats} randomly sampled predictors

3. the best split is chosen among those sampled features

4. new data are predicted aggregating the predictions made from the k_{trees} trees in the forest:

    - classification problems: majority vote
    - regression problems: average of the predictions

In order to evaluate the performance of a random forest it is possible, firstly, at each bootstrap iteration, to predict data, that haven't been sampled, using a tree grown on that sample, and then, to aggregate these predictions. These predictions made on data not used to grow the tree the estimation is made from, are called "out-of-bag" (OOB) predictions. If enough trees have been grown, the OOB estimate of error rate results to be accurate [53].

Random forest is an evolution of decision trees and so it brings several benefits, but at the expense of complexity, thus a loss in terms of interpretability. In particular, two of the most remarkable advantages are its robustness to overfitting and its ability to handle small, and even incomplete, datasets. These qualities have been at the heart of the algorithm choice for this thesis, and random forest turned out to be a suitable option.

Feature Importance

Having an interpretable model is just as important as having an accurate one. Understanding the importance and the impact of each feature in the training process of the predictive model allows us to intuitively link the most relevant characteristics of the raw data to the estimated outcomes.
Basically, feature importance is used to assign a score to the input features based on how useful they are in predicting the target and so, how important their role is. Typically, the features are ranked by importance and analysis based on the impact that they have on the label predictions can be carried out.
Among different ways to evaluate the feature importance, in this thesis a permutation-based approach is used. The feature importance is calculated by analyzing how much changing the value of the variable may impact on the prediction, in particular how random re-shuffling of each predictor influences model performance. For the sake of completeness, it is necessary to specify that by randomly varying the predictor, its distribution is preserved. The consequences of the variables shuffling are evaluated in terms of the decrease in the model performance: varying the predictors breaks the relationship between them and the label, thus a drop in the model score is recorded. Since this process is not computationally light, built-in algorithms usually give the possibility to the user to choose how many permutations or repetitions have to be performed to assess the feature importance.

3.6.2 Cross-validation

Estimating the algorithm performance is key in order to assess its reliability, its generalizability and its confidence in using it. Moreover, avoiding overfitting is fundamental too.
Cross-validation is a method that allows achieving both of these goals at the same time.
It is well known that ten-folds, a special case of k-folds, one of the most used cross-validation technique that could be suitable for this project, is prone to overfitting. Thus, in this thesis a different approach has been implemented: leave-one-subject-out (LOO). It is particularly efficient when the number of observations either in the dataset or for a class value (or in a range, for regression) is small [55], and this is our case.
LOO has been implemented for the clinical scores estimation as follows: analyzing each task separately and, for each of these tasks, iterating through the subjects, the random forest model is trained with all the subjects but the one whom the prediction is made on. The excluded data compose, in fact, the test set, that is used to test the performance of the random forest model without data it has been trained with.
As shown in the figure 3.12, LOO cross-validation, together with the random sampling, of both features and observations, used in growing the trees, should ensure the robustness of the algorithm to the overfitting.

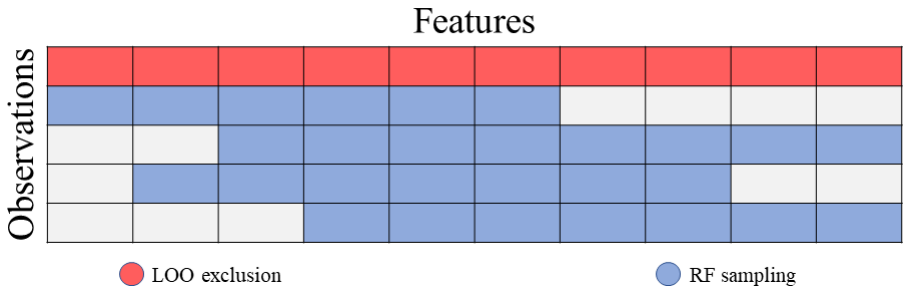

Figure 3.12: To avoid overfitting, at a given iteration, the data of one subject are excluded, while during the training of the random forest model, only some of the available observations and features are used to grow a specific tree.

Chapter 4

Results

4.1 Patients' Clinical Data Analysis

Number of patients	37
Gender	26M 11F
Diseases	16 Stroke 21 TBI
Age	42.61 ± 18.98
Chronicity	253.95 ± 451.87
Days between assessments	36.16 ± 23.65

Table 4.1: Patients' clinical and demographic data.

Analyzing the subjects' clinical data, which are resumed in the table 4.1, it is possible to say that the sample is heterogeneous according to all the variables. The main difference, in terms of clinical variables, between the stroke survivors and the patients with traumatic brain injury, is the age: TBI, in fact, can be one of the many dramatic consequences of brutal car accidents, which unfortunately are common among youngsters with little driving experience. This age difference is appreciable in the following plot 4.1 on the left.

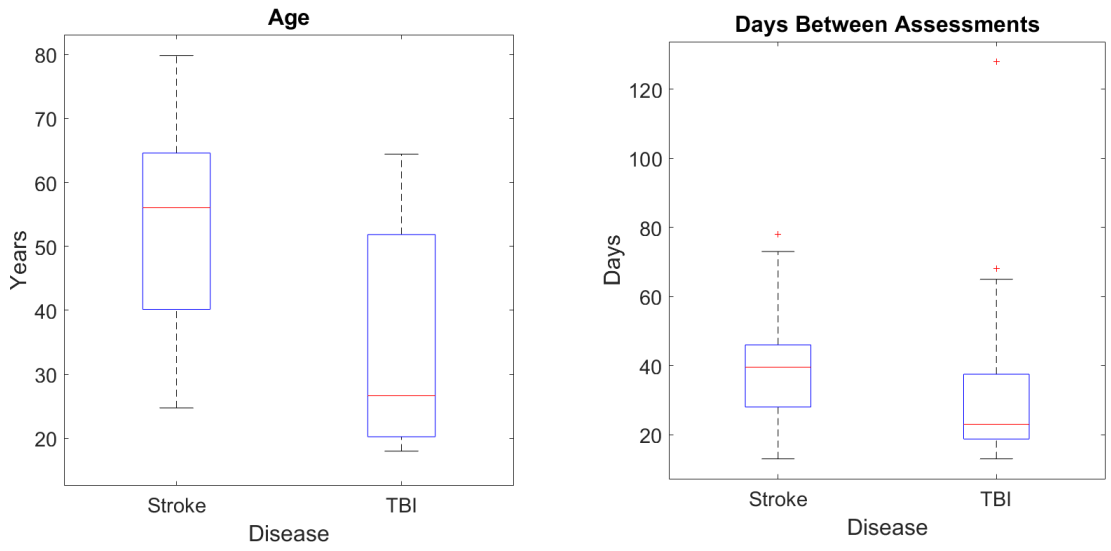

Figure 4.1: On the left it is shown the distribution of the patients' age; on the right the distribution of the days that passed between the two recordings among the patients .

From the right plot of the figure above 4.1, it is possible to see that the time intervals between the recordings are shorter compared to the ones of the Stroke subjects. Moreover, looking at the plots 4.2 and 4.3, it is reasonable to affirm that, on average, the treatments were successful. The scores related to the movement quality and to the motor impairment, in fact, are higher at the end of the treatments than at the beginning. This applies to both the diseases, but for TBI patients the outcomes turn out to be slightly better. Furthermore, remembering from the right plot of the figure 4.1 that the time elapsed between the recordings for TBI patients is less than the one elapsed in the Stroke case, it can be said that, on average, better improvements in less time are reached for patients with TBI; this may be due either to the significant age difference between the subjects of the two pathologies, appreciable in the left plot of the figure 4.1, or to different recovery time among the two diseases.

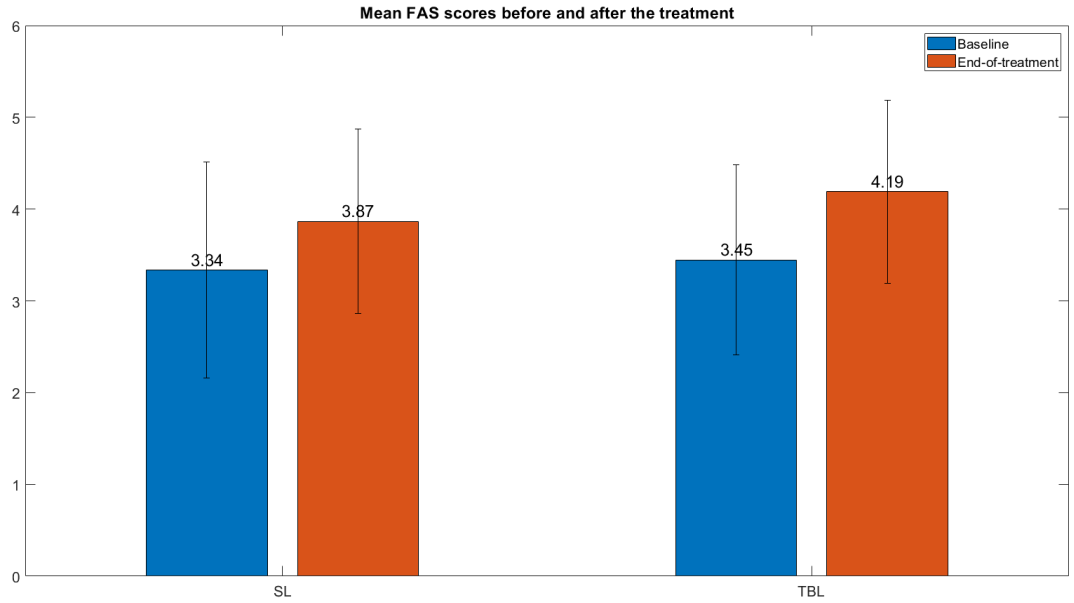

Figure 4.2: On the left are shown the mean FAS scores of single tasks for stroke survivors before and after the treatment, while on the right for patients with Traumatic Brain Lesion.

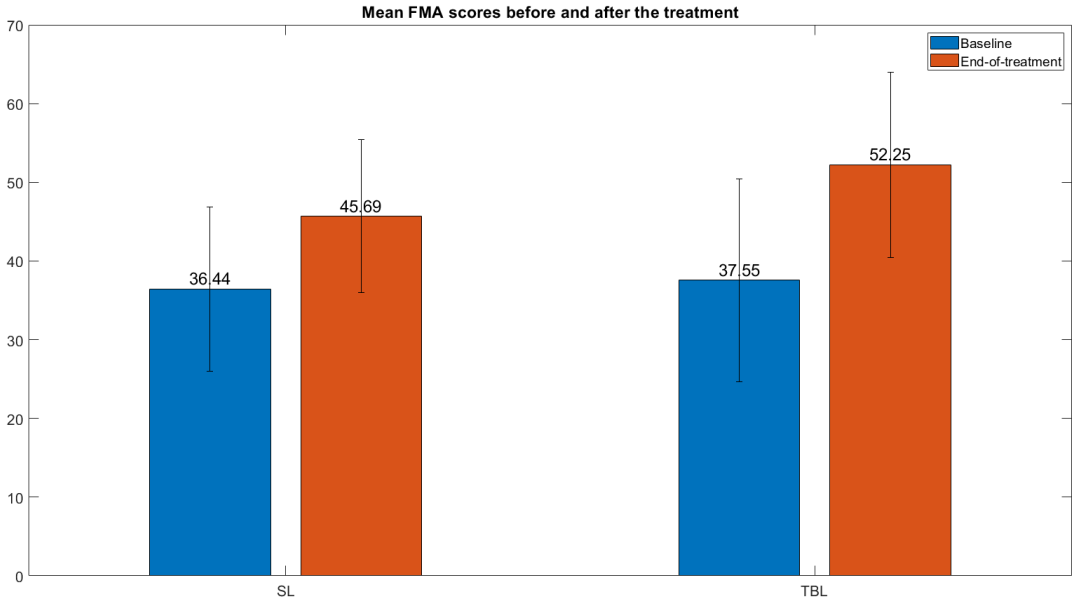

Figure 4.3: On the left are shown the mean FMA scores for stroke survivors before and after the treatment, while on the right for patients with Traumatic Brain Lesion.

4.2 FAS Scores Estimation

In order to estimate the FAS scores, a multiclass classification approach is adopted. It consists of assigning instances to one class, choosing among many. Even though the FAS is a 6 points scale, in the analyzed dataset there are not any examples of 0 scores, thus the problem is reduced to 5 classes.
Looking at the figure 4.4, where is shown the distribution of the FAS scores for each task, it can be noticed that the dataset is highly unbalanced. The target variable, in fact, has not approximately the same number of observations among all the classes, indeed for some of them there are only a few instances. Typically, in machine learning problems, this is an issue that has to be faced in order to avoid affecting the performance of the algorithms. Moreover, a model trained on an unbalanced dataset often has poor results in terms of generalization, prompting it to be biased towards the classes that have the most observations.

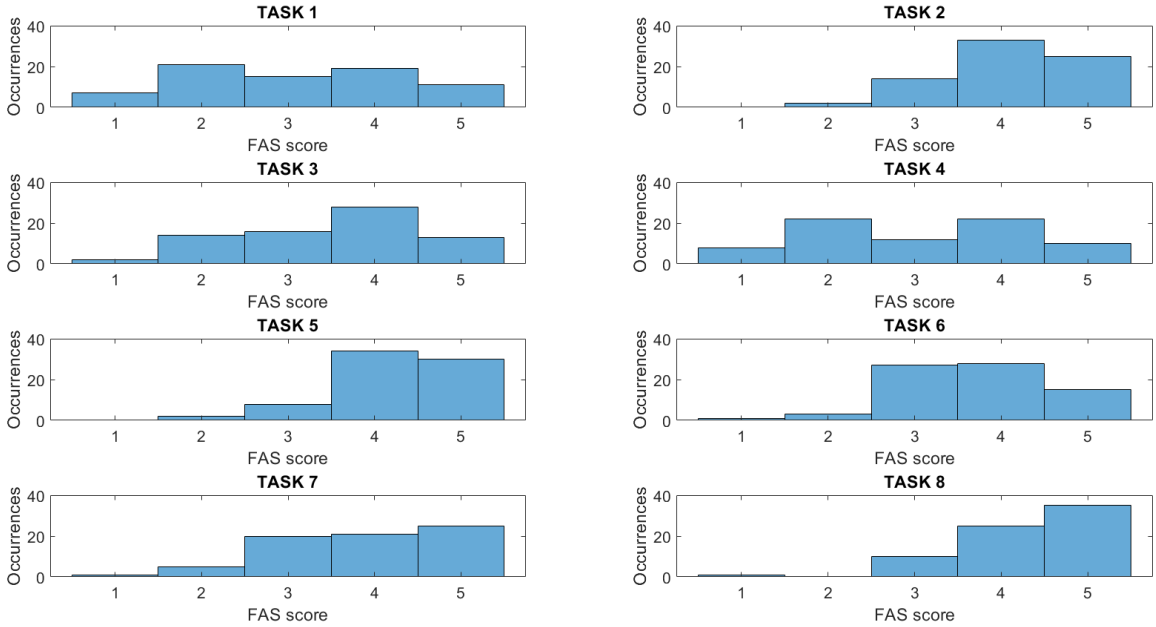

Figure 4.4: FAS scores distribution for each task. In the dataset are included both the pathologies and the recordings.

4.2.1 Dealing with an Imbalanced Dataset

Two different strategies have been compared and implemented to mitigate this problem:

- **Dataset rebalancing using Adaptive Synthetic Sampling (ADASYN)** [56]:
  as many other synthetic oversampling techniques, after having identified the majority class, which is the one with more observations, new data are synthesized taking as example the instances of the other classes and adding some variance to them. These synthetic data are labeled according to the examples from which they were generated. In order to obtain a balanced dataset, in which all the classes have approximately the same number of examples, the number of the synthesized instances depends on the degree of imbalance, defined as $d = M_{min}/M_{maj}$, where M_{min} and M_{maj} are respectively the number of observations for the minority and for the majority class. ADASYN can be seen as an extension of the Synthetic Minority Oversampling Technique (SMOTE [57]): as its distinctive feature, it allows, in fact, to generate more observations in the vicinity of the boundary between the classes, where the prediction is usually inaccurate, than in the interior of the minority class.
  Furthermore, keeping in mind the will to avoid overfitting using Leave One Out (LOO) cross-validation, synthetic data are generated at each LOO iteration so that new data do not contain any information about the excluded observations.

- **Cost-sensitive learning** [58]:
  the concept behind this approach is that, for imbalanced classification problems, misclassifying an example that belongs to a minority class is worse than incorrectly predicting one from the majority class. According to that, the first has to be more penalized than the second. Traditionally, machine learning algorithms are trained on a dataset and solve an optimization problem where they explicitly seek to minimize the error of the model. Defined the "cost" as a penalty associated with an incorrect prediction, the goal of cost-sensitive learning, instead, is to minimize the cost of a model on the training dataset, where it is assumed that different types of prediction errors have a different and known associated cost.

The costs related to each kind of misclassification error are declared in the cost matrix. It is a matrix with the same elements disposition of a confusion matrix and in this case it is a 5×5. Specifically, considering the difference in terms of scores distributions, each task has a different dataset and so its own cost matrix, which is designed with the purpose of stressing the importance of minority classes examples.

In the figures 4.5 and 4.6 it is visible how the problem of dealing with an imbalanced dataset is managed using the first approach, thus rebalancing it with ADASYN method. It is necessary to clarify that a dataset with new synthetic observations has not the same informative power of a dataset with the same number of observations but in which all of them are real: beyond a little variance, the information contained in the synthetic data points is redundant.

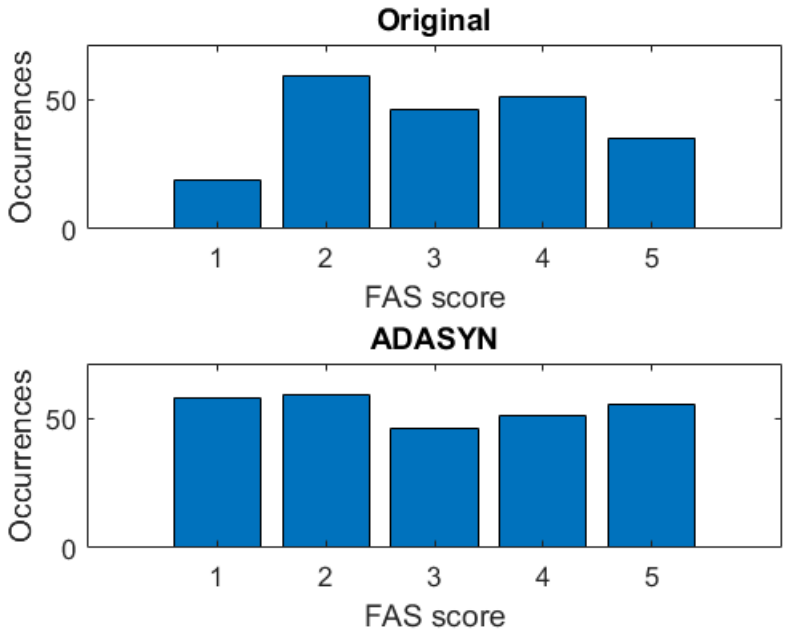

Figure 4.5: Example, for task 1, of the effects on the dataset of rebalancing it using ADASYN.

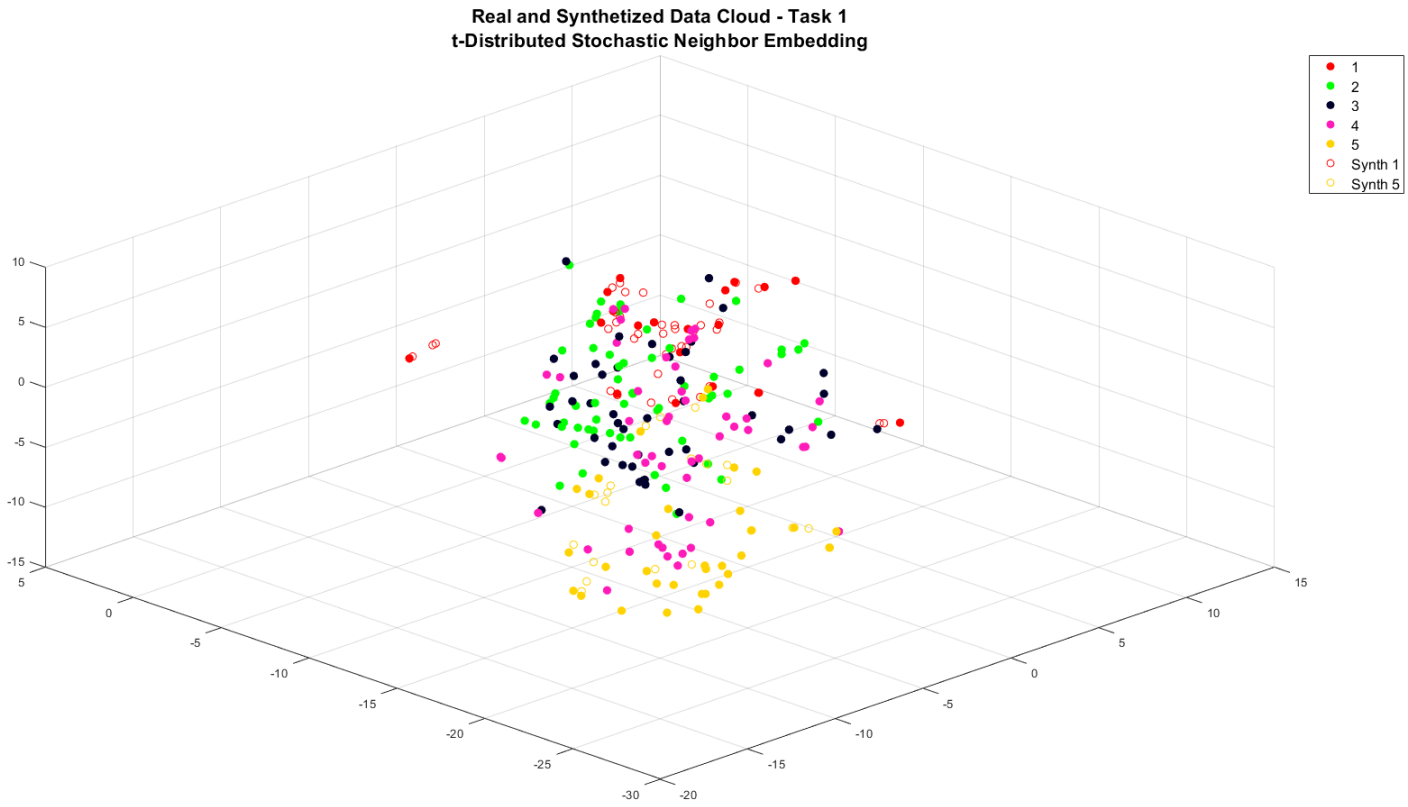

Figure 4.6: Example, for task 1, of FAS data cloud after rebalancing with ADASYN; the visualization is obtained using a dimensionality reduction algorithm, t-Distributed Stochatic Neighbor Embedding [59], and each point is colored according to the respective FAS score: for original data points the circles are filled while for synthetic data points only the edge is colored.

Comparing the confusion matrices obtained with the two approaches previously described, looking at the figure 4.7, the high imbalance problem catches the eye since it is easy to see that there are many more observations belonging to higher FAS scores. These confusion matrices derive from the sum of the ones obtained from the 8 single tasks predictions and hence they can be used to make a global comparison between the two methods. Evaluating the performance metrics for the two approaches, the cost-sensitive classification turns out to have slightly better results, but the imbalanced dataset evidently affects the sensitivity which is low in both the models. Focusing on the accuracy, it may be affirmed that the performance is poor, but, looking carefully, most of the misclassification errors are in the adjacent classes.

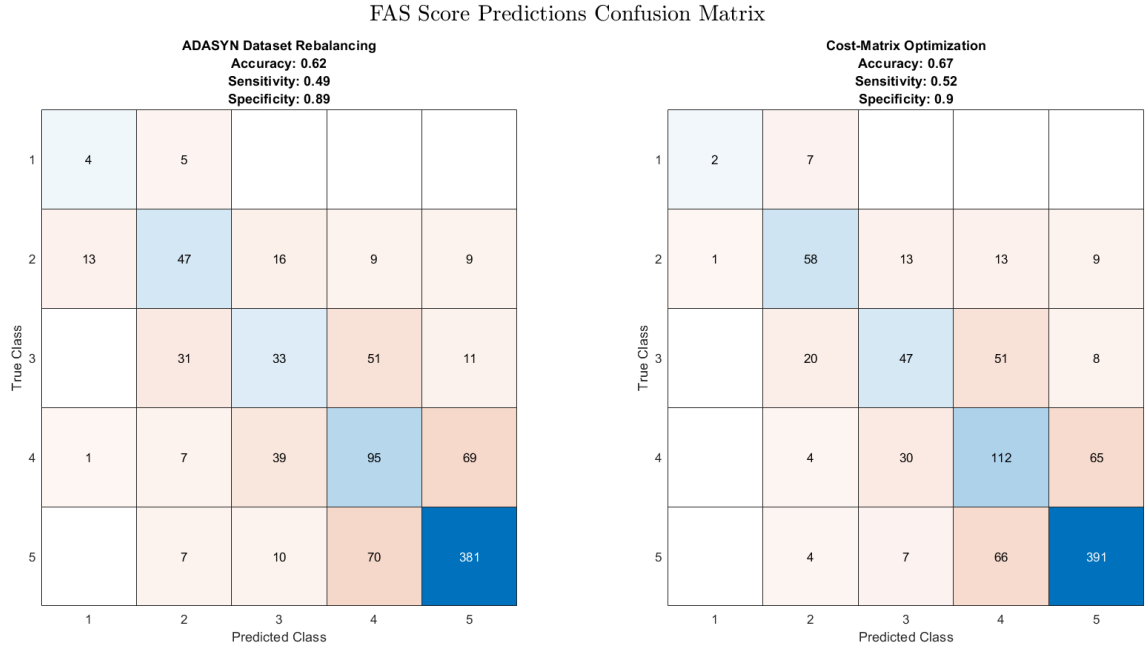

Figure 4.7: Comparison between rebalancing the dataset through ADASYN and using a Cost-sensitive classification; here are shown the confusion matrices obtained with the two techniques and the associated performance metrics.

A better way to notice that most of the missclassification errors are in the adjacent classes, which in this context is an appreciable behavior, is provided with the figure 4.8. In the related plot, it is possible to notice that around 95% of the observations are classified with an error of 1 point of the FAS. As it was anticipated previously, this is a good behavior when the classification problem can be also interpreted as an ordinal classification problem, which is this case. Since the Function Ability Scale grades the movement quality in a crescent order, the labels are sorted from 1 to 5. Hence, for instance, misclassifying an observation whose label is 4, predicting a score of 3, in this context, cannot be considered as wrong as predicting 1. The cumulative distribution function, for this kind of problems, is really explanatory about the concept just explained. Even here the cost-sensitive approach seems to work slightly better.

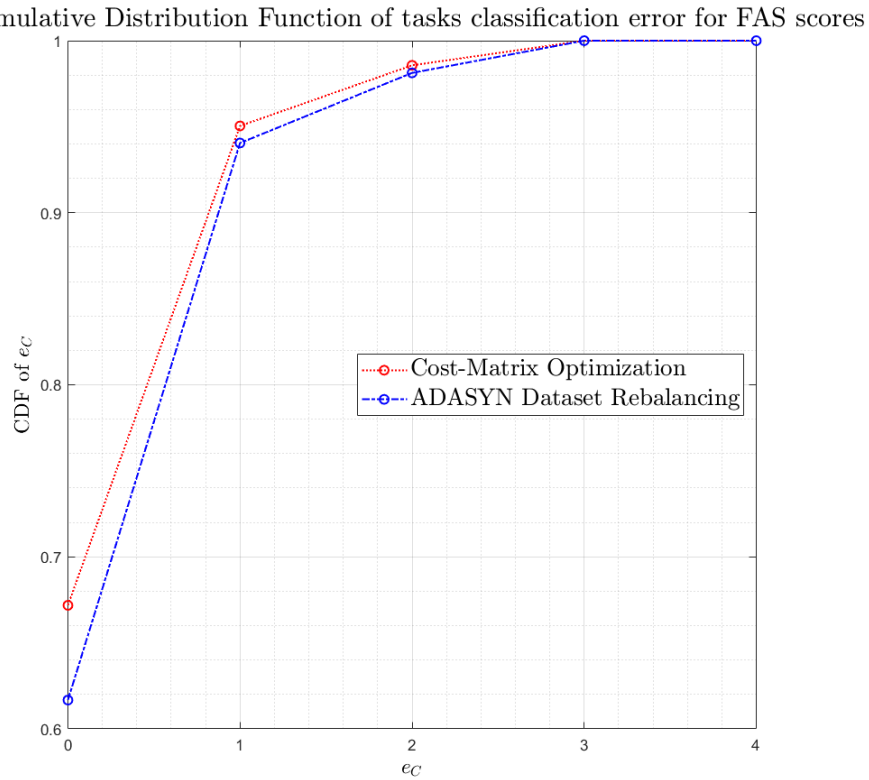

Figure 4.8: Comparison between rebalancing the dataset through ADASYN and using a Cost-sensitive classification; here are plotted the Cumulative Distribution Functions of the misclassification errors for both the approaches.

Before jumping to the conclusions regarding this comparison between the two approaches to deal with an imbalanced dataset, it is useful to analyze the Receiver Operating Characteristic plot, better known as ROC. In the ROC space it is plotted the true positive rate (TPR), also known as *Sensitivity*, against the false positive rate (FPR), also known as $1 - Specificity$. These two evaluation metrics derive from the confusion matrices obtained with the two different approaches. In this way, since each model is defined by its confusion matrix and the correspondent performance metrics, the 2 models are represented in the ROC space as 2 points. Looking at the figure 4.9, also in this case the cost-sensitive approach performs better with higher sensitivity and lower false-positive rate. However, it should be pointed out, as in the previous comparisons of these 2 approaches, that the differences in terms of performance are really thin.

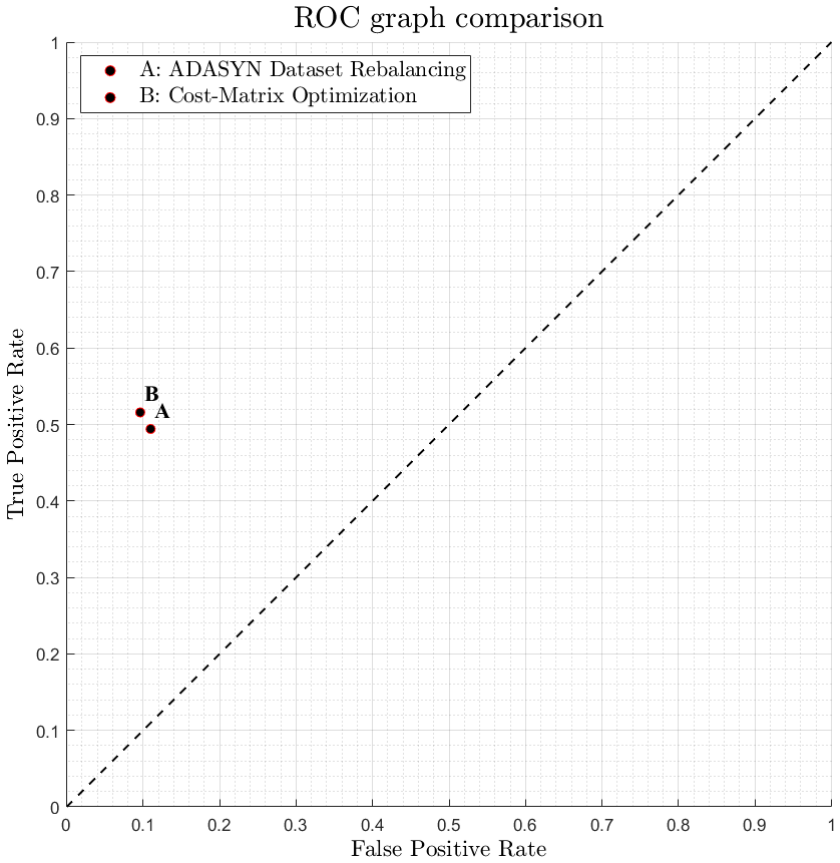

Figure 4.9: Comparison between rebalancing the dataset through ADASYN and using a Cost-sentitive classification; here are plotted in the ROC space the points related to the two approaches.

Conclusions about comparison

Finally, arriving to the conclusions about this comparison between these two methods to handle an imbalanced dataset, it is possible to say that the cost-sensitive option performs slightly better in every aspect. The problem of implementing this method is that it is dataset dependent: with a view to expand the dataset with more observations from new subjects, this method may fail since the cost-matrices are designed to make the most of low FAS scores instances, since there are few of them; thinking of enriching the dataset with low FAS observations, of which it is most lacking, in order to build a less imbalanced one, using this cost-sensitive classification would turn out to be useless or even harmful. It may be a good implementation if there were to be no further studies to improve the model and if no more data could be collected. Hence, if no more data could be added to the dataset, the cost-sensitive approach would be the best one, but considering there is the will to make this model as scalable as possible, this loss in terms of generalization cannot be counterbalanced by the small improvements in respect to the other option.
According to all of this, the chosen approach consists of rebalancing the dataset using ADASYN since it provides a good generalization and it would help the model to perform better even if the dataset had a different imbalance. Hereinafter, all the implementations and results related to the FAS scores estimation algorithm are obtained after having rebalanced the dataset with ADASYN.

4.2.2 FAS scores estimation Results

The proposed model, analyzing individually each task, gets as input the wearable accelerometers data, extracts the features, selects a subset among them, and then, with a dimensionally reduced dataset, it performs the FAS scores predictions through a random forest. As mentioned before, all the predictions are made according to Leave One Out cross-validation: the random forest for multiclassification is trained excluding one subject at each iteration. Finally, these single task predictions are added together into a final one representing the total FAS score and the regression equation, relating the actual total score provided by the clinician and the total predicted score, is derived.
The number of implemented weak learners, which in random forests are decision trees, is chosen according to the strategy proposed by Oshiro [60]. Hence, considering that:

- sometimes, increasing the number of decision trees in a random forest leads only to an increment in the computational cost without significant improvements in the performance
- typically, the evaluation metrics converge asymptotically with the increasing of the number of decision trees

after having followed the steps proposed in his paper, 100 decision trees are evaluated as the right choice since, as the number of trees increased, no further improvements could be reached. Adding up the single task predictions and considering that the highest FAS score is 5, the total FAS score is defined as a percentage of the maximum score achievable with the tasks as:

$$\hat{FAS}_{TOT} = \frac{\sum_{i=1}^{n} \hat{FAS}_{ST,i}}{5n} \times 100 \tag{4.1}$$

where $\hat{FAS}_{ST,i}$ is the i-th FAS score single task prediction and n is the number of tasks.

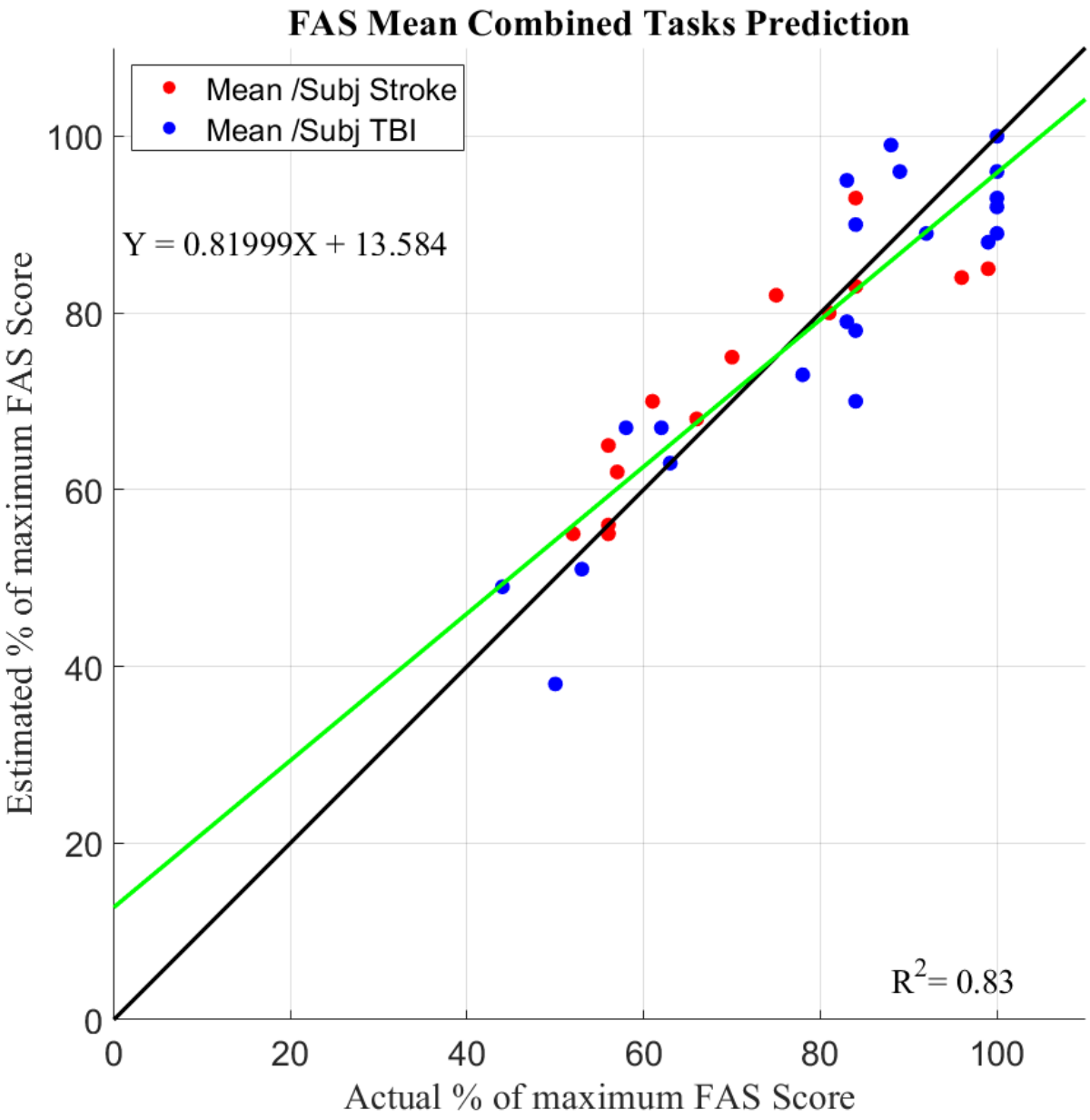

Figure 4.10: FAS total score predictions: the regression equation between the actual and the predicted total FAS scores is plotted together with the patients data points; each point represents one single patient, in red if he is a stroke survivor or in blue if he is affected by TBI; each subject's point is obtained as the average of his trials.

In the previous figure 4.10 the regression line between the actual and the predicted scores is plotted together with the patients' data points. For a better understanding of the model and of the obtained results, one should bear in mind that the purpose is to estimate the clinical outcome of a patient specific rehabilitation program; with this, the predictions are made on the recordings at the end of the treatment. Thus, remembering from the figure 4.2 that on average the treatments were successful for both the diseases, it was likely that the the data points in the figure 4.10 are shifted towards higher scores. R^2, the coefficient of determination, is used as the main evaluation metric: generally, it provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, or, in an easier way, it gives information about the goodness of fit of the model; in this context, it concretely represents how far the predicted scores are from the actual ones.
Focusing on the patients' data points in the figure 4.10, the model seems to be not affected by bias. The presence of bias in the model can be better analyzed by plotting the error for each trial of each patient as is done in the figure 4.11: evaluating the plotted points, it can be said that all of them are included in an error range of $\pm 20\%$ of the maximum total FAS score. To better comprehend the clinical aspect of the achieved performance, this error range corresponds to an error range of ± 8 grades of the FAS scale over the sum of 8 tasks scores. This can be also interpreted as if, in the worst case, at most, an error of 1 grade for each task is committed.
Comparing these results with the ones reached in the previous studies, using a mixed dataset, that included both stroke survivors and patients affected by TBI, built with data recorded during the performance of standardized WMFT tasks, it was possible to achieve a coefficient of determination $R^2 = 0.79$, while in this thesis, with the same patients but using ADL tasks, is presented a coefficient of determination $R^2 = 0.83$. The single task predictions of FAS scores and the total FAS score predictions, as has been anticipated, are used as a further input, besides accelerometers data, for the FMA estimation algorithm in order to improve those estimates.

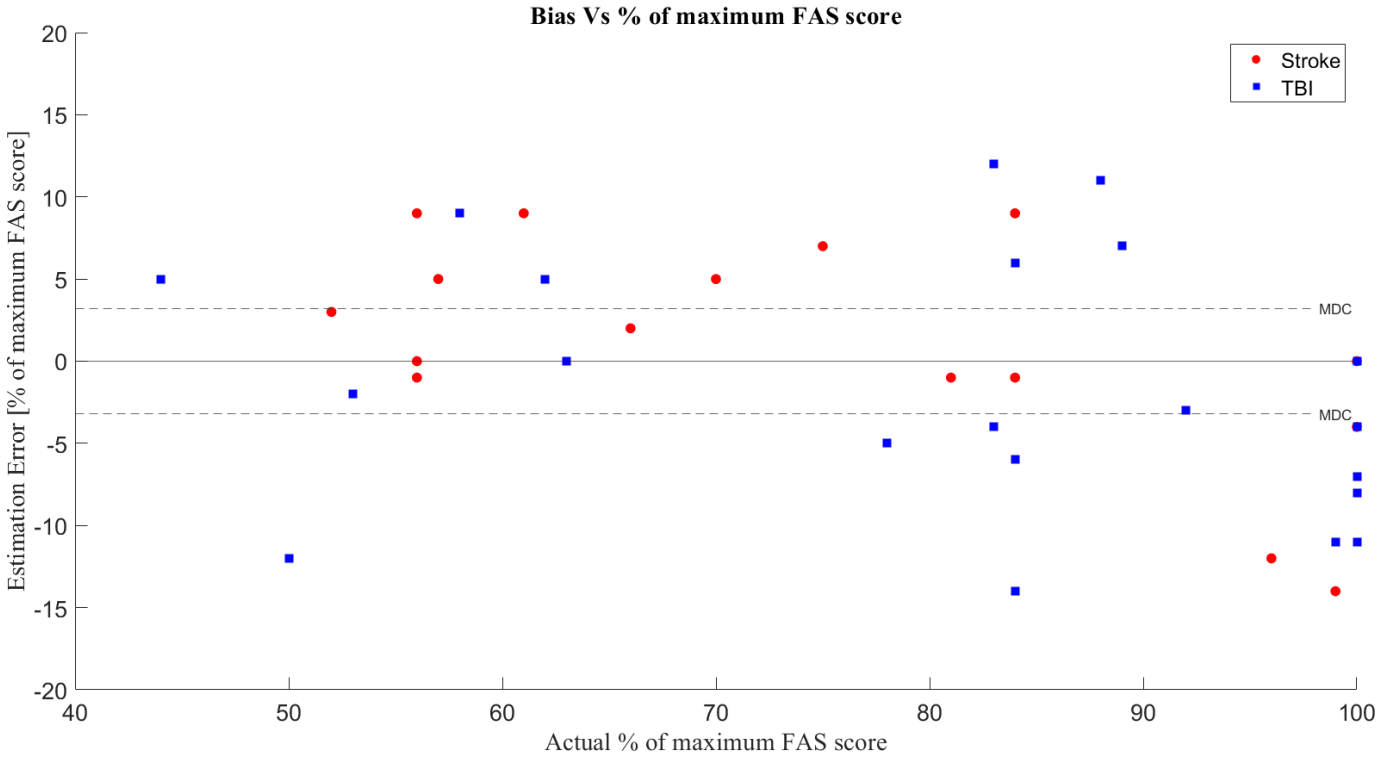

Figure 4.11: Bias analysis of % of maximum FAS score predictions: on the x-axis the percentage of the maximum FAS, while on the y-axis the subjects' estimation error, defined as $e = predicted - actual$; each point represents one single subject, in red if the patient is a Stroke survivor or in blue if he is affected by TBI.

4.3 FMA Scores Estimation

In order to provide a global view of the FMA scores estimation algorithm before going into detail, here is given a quick introductory description of its key steps and features that make it up.
The estimation of FMA scores is treated as a regression problem: the label, in fact, is continuous in a range from 21 to 66, when both the baseline and the end-of-treatment recordings are included. After extracting the features from the wearable accelerometers data, in order to perform dimensionality reduction, a subset that includes the most relevant ones is selected using a Correlation-based Feature Selector (CFS). The predictions are performed training and testing a modified random forest, which was called "Balanced Random Forest" by Adans-Dester [1].
Firstly, the estimates of FMA scores from each task are obtained individually, and then, they are combined to make a final prediction. The previous single tasks FAS score estimates are added as additional features to improve single tasks FMA predictions, while the total FAS prediction is used for the same purpose during their aggregation procedure. Moreover, to avoid overfitting, also in this case, LOO cross-validation is used and, as in the FAS estimation algorithm, it is necessary to handle an imbalanced dataset.
It was also sought the best subset of tasks, among all the possible combinations, since it was not known if all of them were descriptive and useful in predicting the FMA scores. Evaluating the performance of both single tasks and subsets of tasks, it turned out that only some of the tasks are capable of assessing the motor impairment and only a few subsets can be used to reach good performance in estimating FMA scores. In particular, the optimal subset proposed here is made up of 4 tasks, [1 3 4 6], which specifically are opening a bottle, putting on and taking of a pen cap, unfolding a towel and lifting a box.
Investigating the feature importance, calculated during the random forest training, it is possible to make some intuitions regarding the singular variables. The analysis of the algorithm performance is done evaluating the coefficient of determination R^2, the bias and, the root mean square errors $RMSE$ obtained for each subject and for different intervals of the whole FMA range. To better understand how this algorithm actually works, a schematic pipeline is provided in the following figure 4.12.

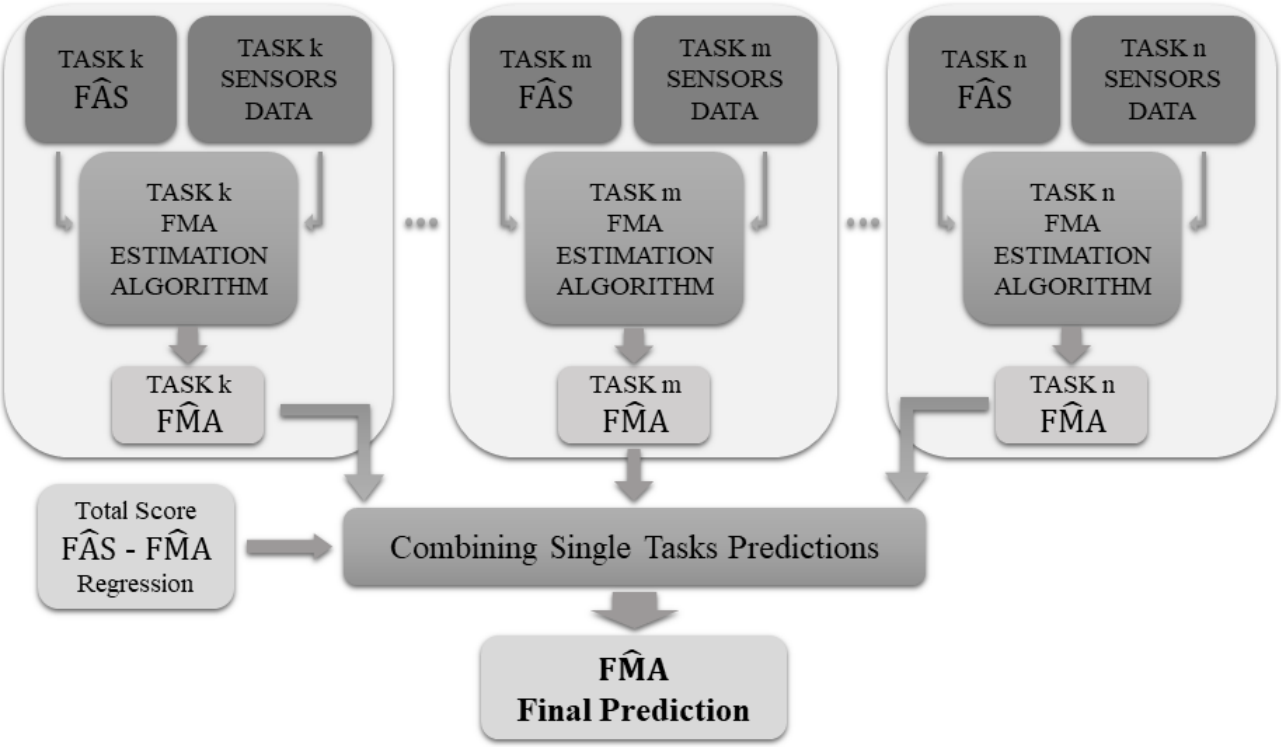

Figure 4.12: FMA score prediction algorithm pipeline. $k, m, n \in$ [1,8]; each task is analyzed separately taking as inputs the accelerometers data and the FAS predictions of the related task, and giving as output the FMA single task predictions; after the latter are calculated for all the considered tasks, they are combined together with the FAS total score predictions to make the final FMA prediction.

4.3.1 Handling Imbalanced Dataset

As in the case of FAS score predictions, also the dataset for FMA estimation is imbalanced and the consequences of using it are the same as discussed in the previous section. In order to handle this issue as best as possible, different oversampling and undersampling techniques were tried, still, the one that led to the best performance confirmed to be, also in this case, ADASYN. The big difference between how ADASYN is implemented in this situation compared to the previous one, lies in the fact that this is a regression problem, while the other is a multiclass classification and this leads to several variations in the procedure. The fundamental concept of this ADASYN implementation is based on reinterpreting a regression problem as a classification one; the key steps are listed here:

- The whole FMA range [21,66] is divided into smaller intervals of around 4-5 FMA points; according to the corresponding intervals, the observations are assigned to those new classes.
- Considering that, in order to find the examples needed to generate new observations, the ADASYN algorithm performs a neighbor search, in each FMA interval a minimum number of observations must be guaranteed. If one of these clusters has less than 7 instances, this is the chosen number, the classes are rebuilt and the observation reassigned according to larger intervals until all the clusters have at least 7 observations.
- After identifying the majority class, each other class is individually processed in order to generate as many examples as needed for rebalancing the dataset.
- In order to assign to each synthetic observation an FMA score within the boundaries of its interval, the affinities between each new point and all the original data points are calculated; for this purpose, the euclidean distance is used as the similarity measure. The FMA score of a synthetic point is thus assigned as the one from the closest example among the original observations of the same cluster.

In the following figure 4.13 are shown the results of this ADASYN implementation for a regression problem. At the end of the procedure the dataset is rebalanced and ready for further uses.

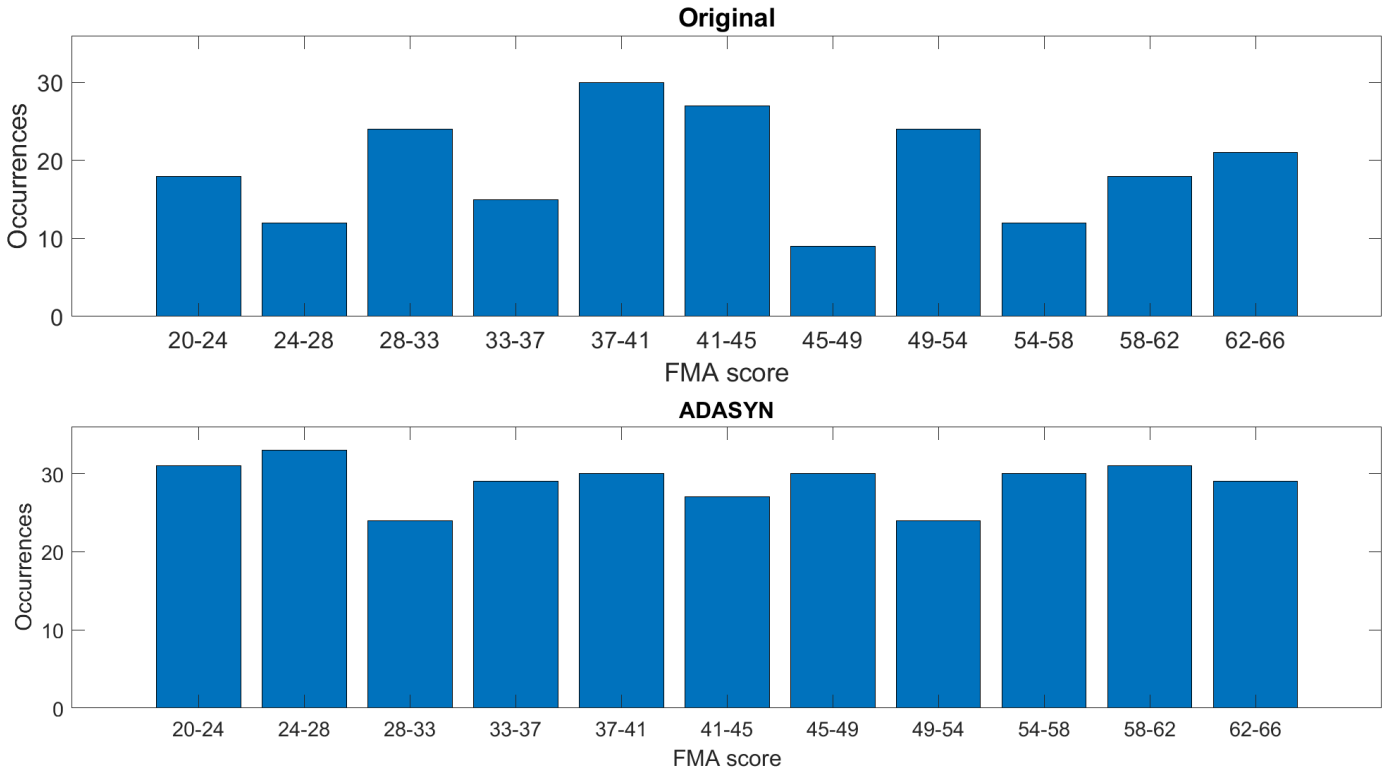

Figure 4.13: Results of rebalancing the FMA dataset with an adaptation of ADASYN algorithm that made it suitable for regression problem; the represented dataset is made up of data belonging to the optimal subset of 4 tasks with both baseline and end-of-treatment recordings.

Reducing up to 3 the dimensionality of the new dataset, that is now composed by the original and the synthetic data, it is possible to visualize in a 3D space the data cloud of both real and generated points. t-Distributed Stochastic Neighbor Embedding (t-SNE) is used for reducing the dimensionality in order to generate the following figures (4.14,4.15,4.16,4.17) related to the 4 tasks of the optimal subset.

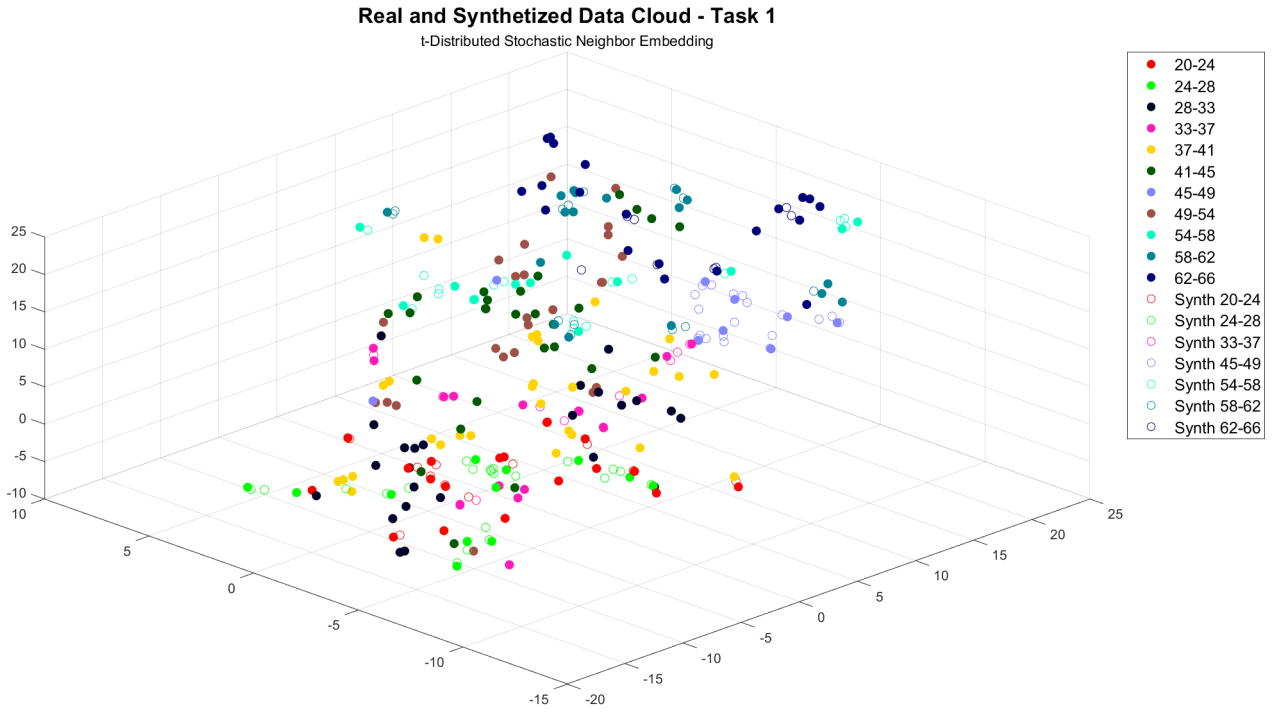

Figure 4.14: FMA data cloud for task 1.

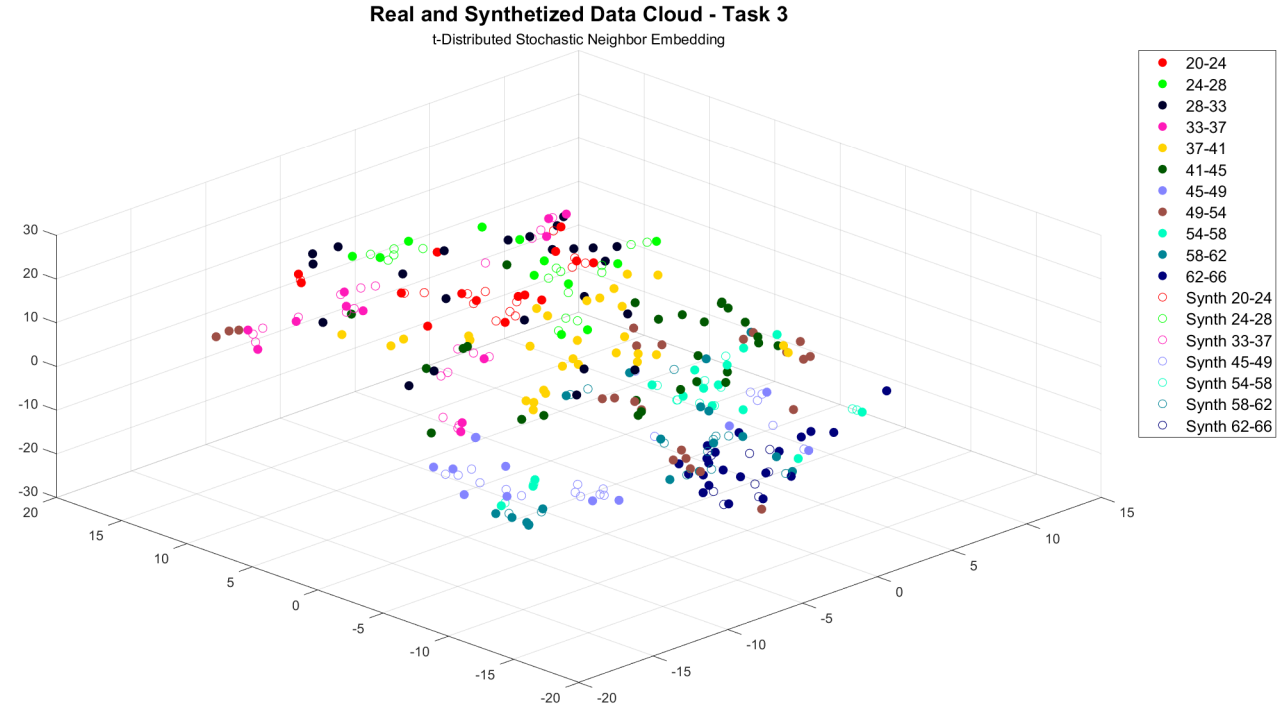

Figure 4.15: FMA data cloud for task 3.

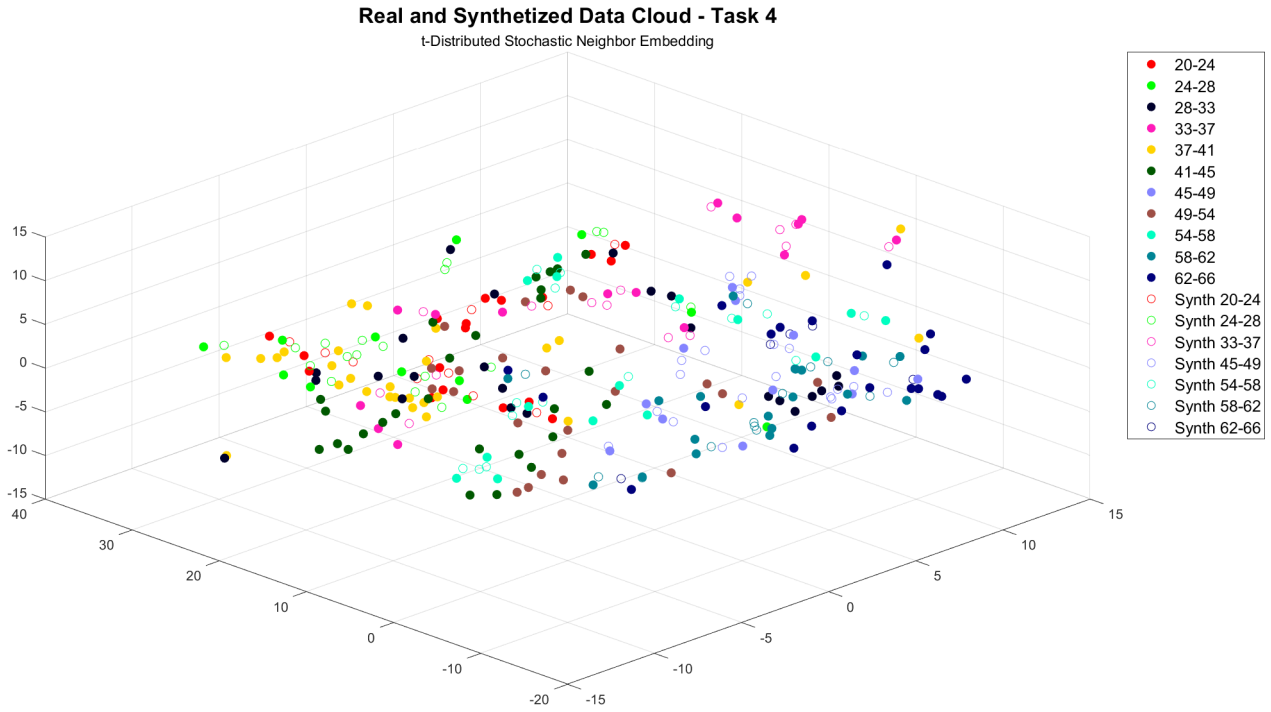

Figure 4.16: FMA data cloud for task 4.

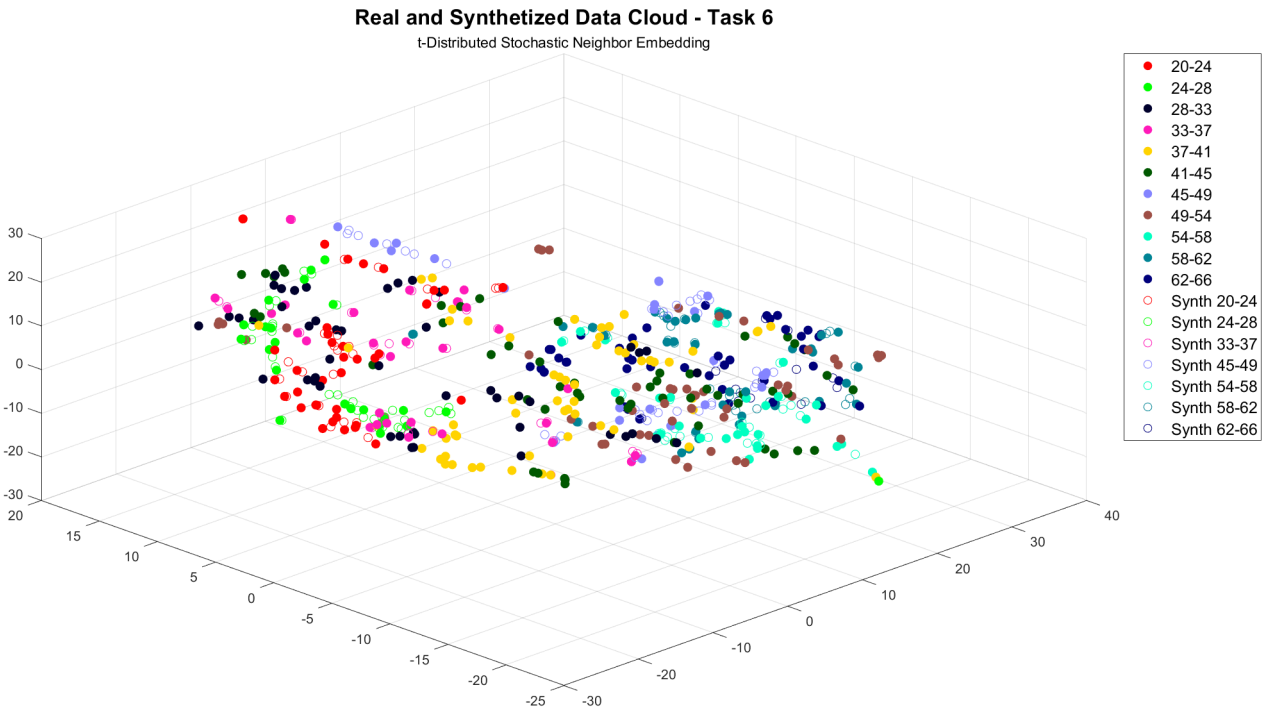

Figure 4.17: FMA data cloud for task 6.

4.3.2 Balanced Random Forest

The FMA score predictions are made training and testing Balanced Random Forests of the same type of the one described in [1]. The same approach used in the FAS scores estimation algorithm is applied here when deciding the number of decision trees for building the RF; also this case, the chosen number of weak learners is 100. The RFs utilized to generate the FMA score predictions for each of the 4 ADL tasks are trained using even more balanced datasets: observations that are used to generate the decision trees are randomly sampled picking up the same number of examples among intervals in the whole FMA range; moreover, preserving the nature of the RF, also the data features are randomly selected during this process. This further balancing of the training set in the course of the generation of the RF is obtained with the addition of a feature that labels the observation based on the correspondent FMA score at the baseline. Remembering that the purpose of making the estimates is to evaluate the clinical outcome of a patient's specific rehabilitation program, it is reasonable to keep track of the starting point, but, in order to avoid overfitting and to retain the model generalization, the information regarding the FMA score at the beginning of the rehabilitation intervention is provided by way of range. More to the point, the FMA range is divided into 5 classes as it is described in the table below 4.2:

Class	**FMA range**
1	$FMA \leq 30$
2	$30 < FMA \leq 38$
3	$38 < FMA \leq 47$
4	$47 < FMA \leq 56$
5	$FMA > 56$

Table 4.2: FMA classes for Balanced Random Forest.

4.3.3 Adding FAS estimates

In order to improve the FMA predictions, the FAS estimates are added to the FMA dataset as a further feature.
Specifically:

1. single task FAS estimates, after being unity-based normalized, are added to the correspondent single task FMA dataset;

2. FAS total score predictions are added to the 4 single task FMA estimates when aggregating them to make a final prediction:

$$\hat{FMA}_{FAS} = \frac{\hat{FAS}_{TOT}}{100} \times FMA_{maxUE}$$

   - $\hat{FMA}_{FAS}$: FMA scores derived rescaling the FAS total score estimates
   - $\hat{FAS}_{TOT}$: FAS total score estimates expressed as a percentage of the maximum achievable FAS score for 8 tasks
   - FMA_{maxUE}: FMA maximum achievable score for upper-extremity section, corresponding to 66.

The rescaling of the FAS total score estimates towards the FMA range can be justified by the partial correlation that exists among the two clinical scales; this can be observed in the figure 4.18 which shows how FMA scores can be grossly estimated using a linear regression between the obtained $\hat{FMA}_{FAS}$ and the actual FMA scores.
In the figure 4.19 are shown the results of the FMA estimation algorithm obtained without enhancing the predictions with the addition of the FAS score estimates, neither single task nor total score ones; this plot is provided in order to keep track of the performance improvement resulting from the use of $\hat{FAS}$. During the balanced RFs training, it is possible to calculate the permutation-based features importance among the ones that are selected by the Correlation-based Feature Selector (CFS) in order to assess whether adding the single task FAS estimates is impacting on the FMA predictions.
Going more into detail, looking at the figures 4.20, 4.21, 4.22 and 4.23, it can be viewed that, for the tasks 1, 3 and 6, the single task FAS estimates are present in the subset of selected features and that they turn out to have a decent importance among them.

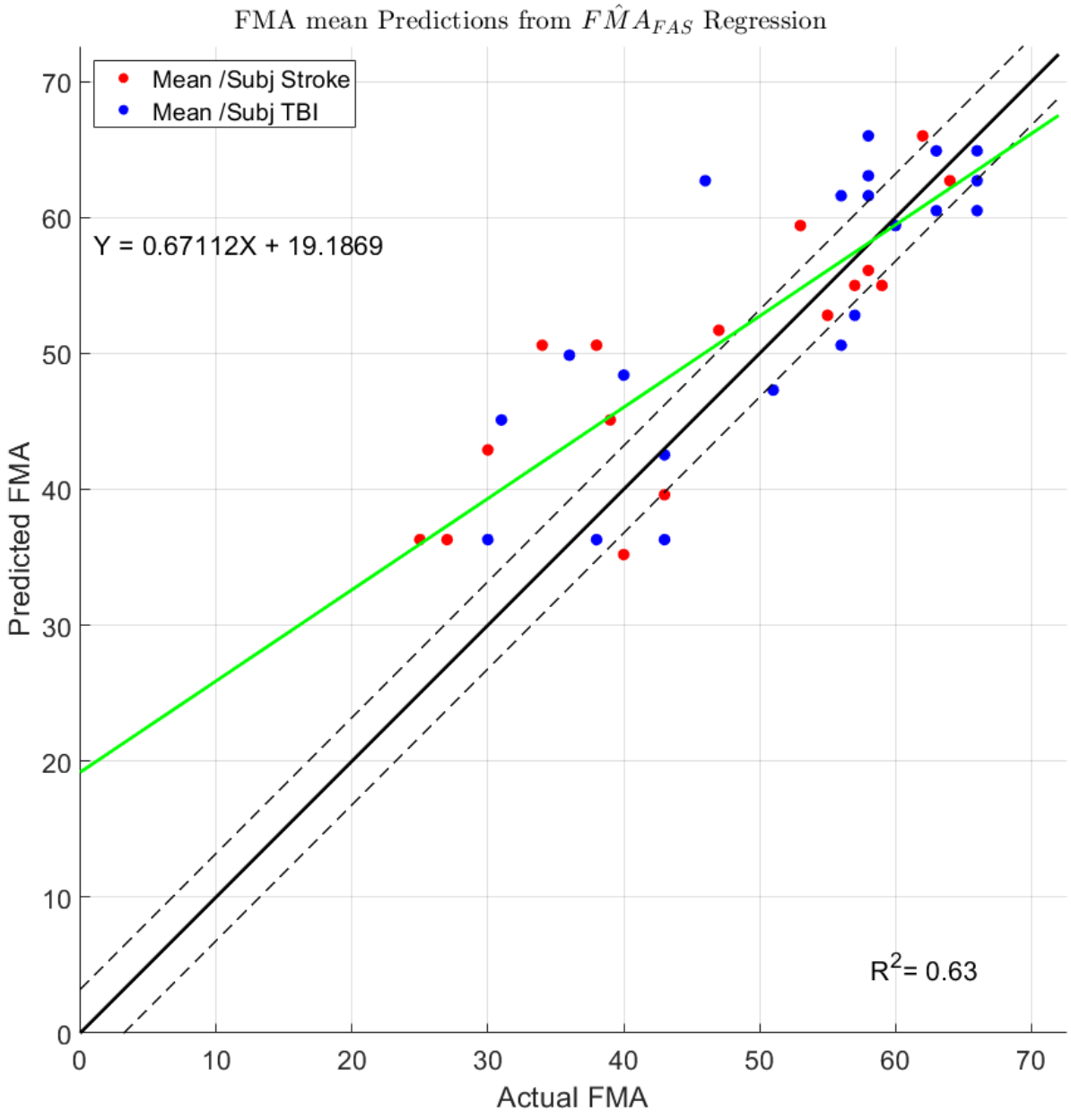

Figure 4.18: FMA scores are inaccurately predicted using the $F\hat{M}A_{FAS}$ as evidence of the correlation between the FAS and FMA scale.

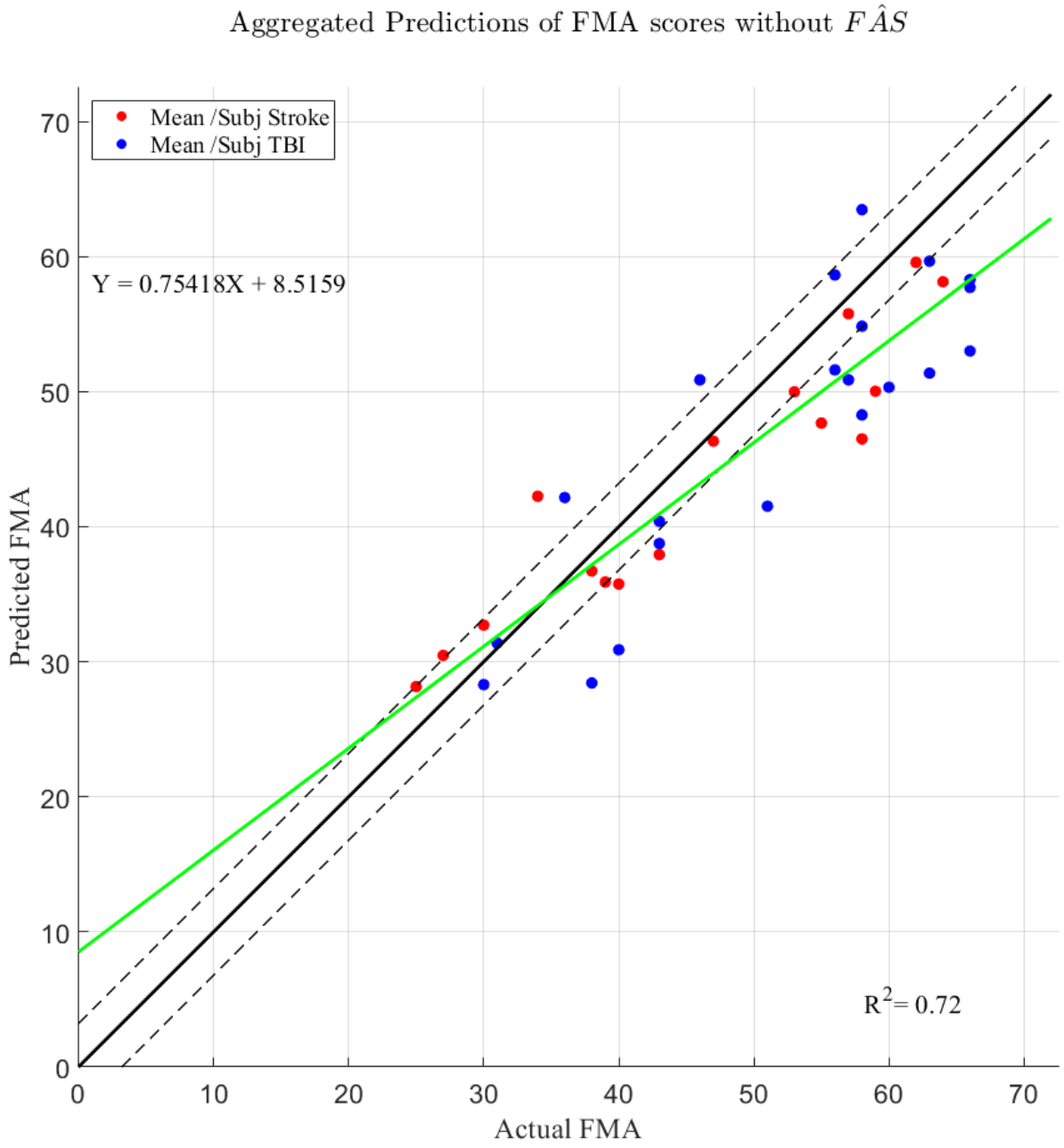

Figure 4.19: FMA prediction obtained using a model trained without $F\hat{A}S$.

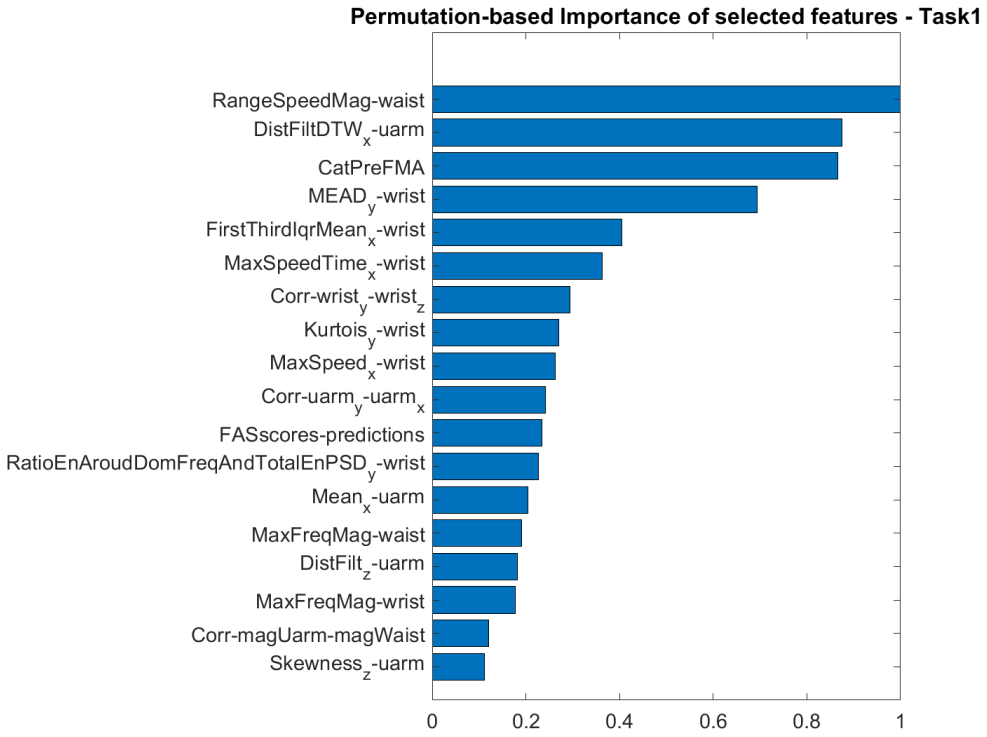

Figure 4.20: Feature Importance for task 1 FMA scores estimation.

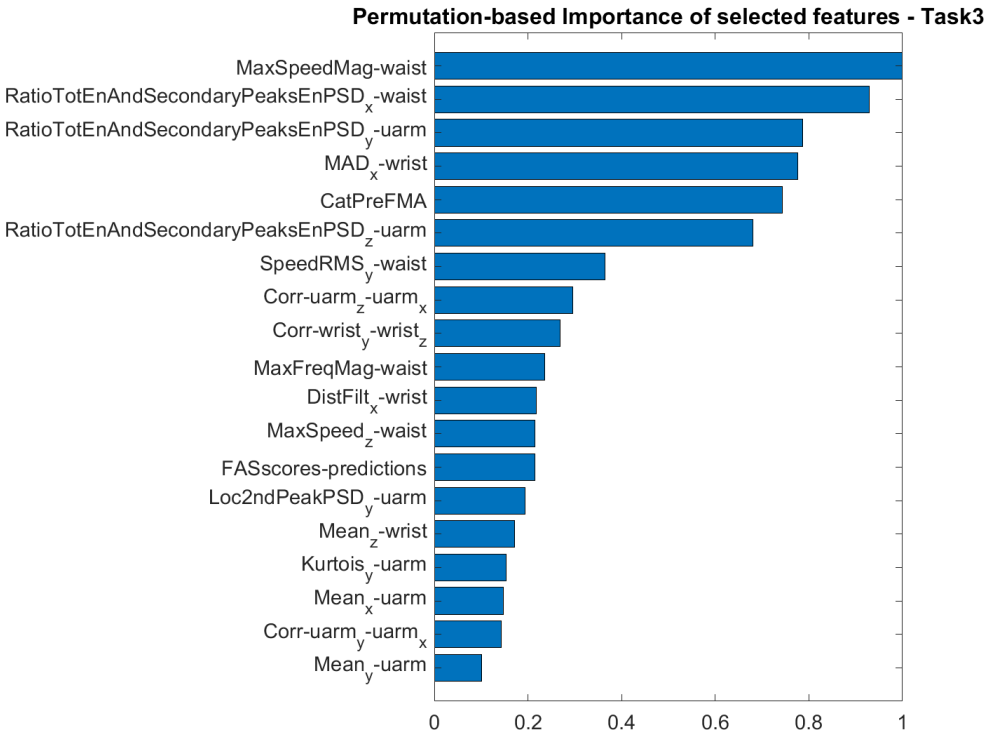

Figure 4.21: Feature Importance for task 3 FMA scores estimation.

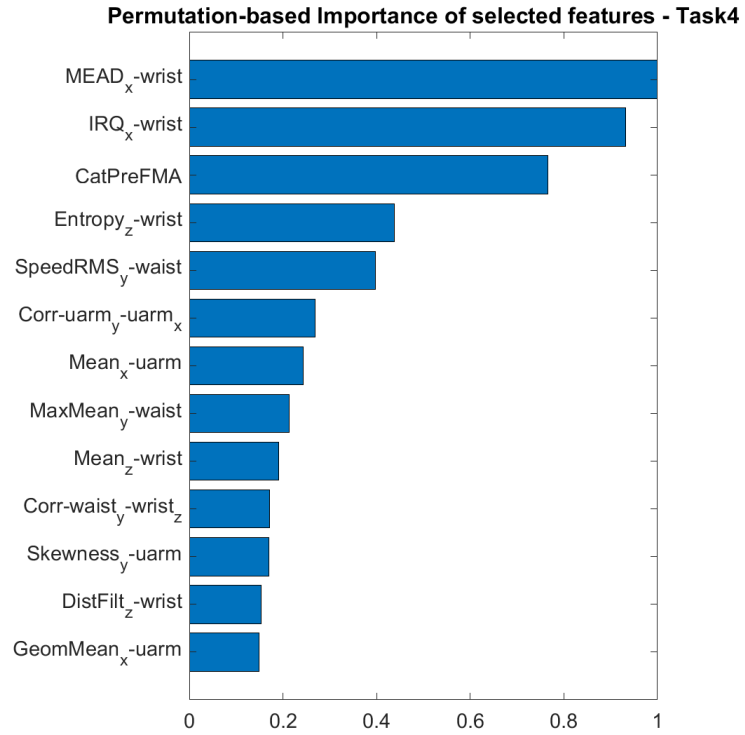

Figure 4.22: Feature Importance for task 4 FMA scores estimation.

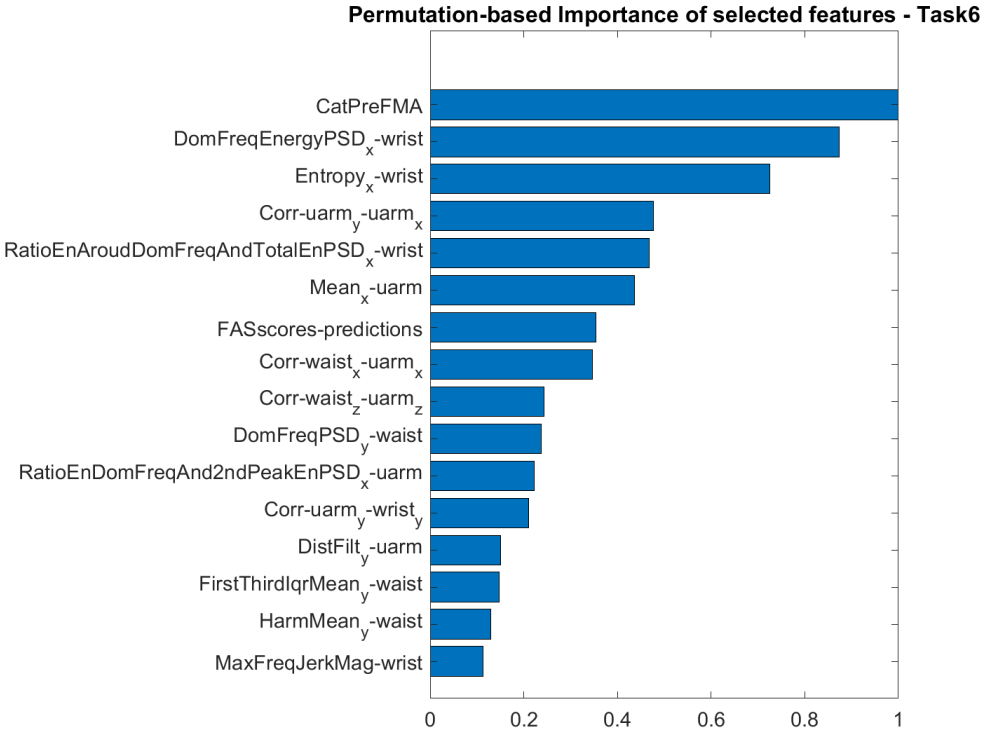

Figure 4.23: Feature Importance for task 6 FMA scores estimation.

4.3.4 FMA scores estimation Results

The results here presented are obtained processing a dataset made up of features extracted from wearable accelerometers data and of FAS estimates belonging to an optimal subset of tasks [1 3 4 6]. Both stroke survivors' and TBI patients' data are included and the recordings were made during the performance of Activities of Daily Living tasks. The FMA score predictions are generated using this dataset to train and validate a balanced random forest with 100 trees. Finally, the regression line between the predicted and the actual FMA scores is derived.
When aggregating the single task FMA predictions and the $\hat{FMA}_{FAS}$, they are averaged and the resulting estimates are given as the final outputs.
In the figure 4.24 are shown the aggregated predictions of subjects' FMA scores; the impact of the $\hat{FAS}$ addition is visible comparing this figure with the previous 4.19: the improvement in the coefficient of determination R^2 is significant since it grew by 8.3%.
Two kinds of error are calculated to evaluate the performance of the model:

- Estimation error defined as $e = predicted - actual$: it keeps track of the sign which is needed in order to evaluate the under/overestimation;
- Root Mean Square Error (RMSE) defined as $\sqrt{\sum_{i=1}^{n_{trials}} \frac{\hat{FMA}_i - FMA_i}{n_{trials}}}$: it is used to quantify the error that affects each subject.

Already from 4.24 the underestimation of the highest FMA scores can be noticed together with the little overestimation of the lowest FMA scores; this behavior suggests the presence of a little bias that is more appreciable in the figure 4.25. The estimation errors for all the FMA score predictions of the subjects graphically range between ±10 FMA grades, while analyzing each subject RMSE in the figure 4.26 a more quantitative understanding of the estimations can be obtained: despite a couple of subjects in which the RMSE reaches values around 10 FMA grades, for many other patients the predictions are really accurate. Overall, the mean RMSE is quite low, permitting to make precise FMA score estimation with, on average, a displacement of 5 FMA grades from the actual FMA score provided by clinicians. Moreover, looking at the figure 4.27, dividing the whole FMA range into intervals of 5-6 grades, it is possible to see that there are no consistent differences in terms of error among these intervals.

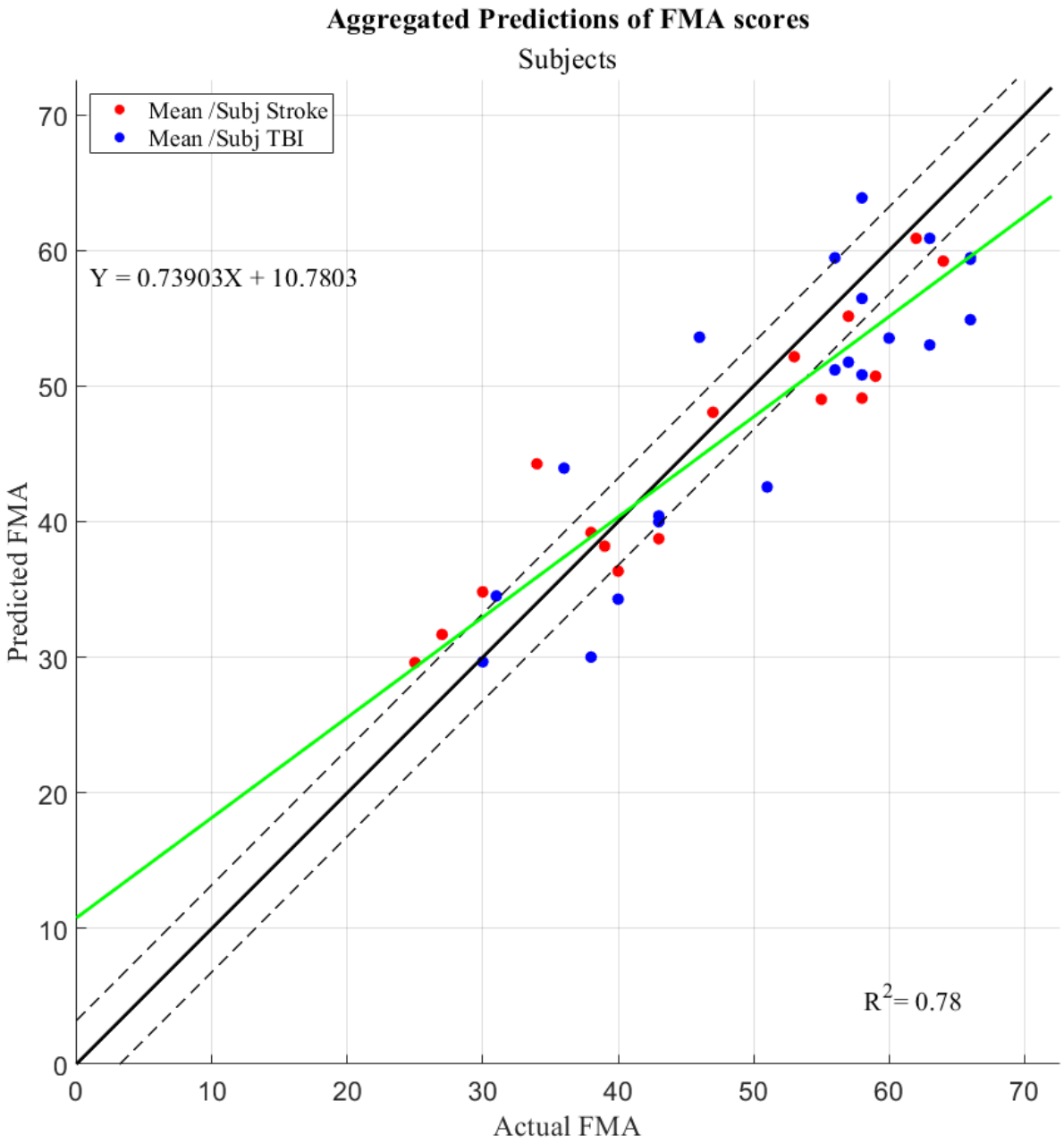

Figure 4.24: FMA predictions of subjects' scores; the regression line, in green, is derived relating the predicted FMA scores to the actual ones; in red are plotted the FMA estimates for stroke survivors, while in blue for TBI patients; the confidence interval is graphically shown with dashed lines.

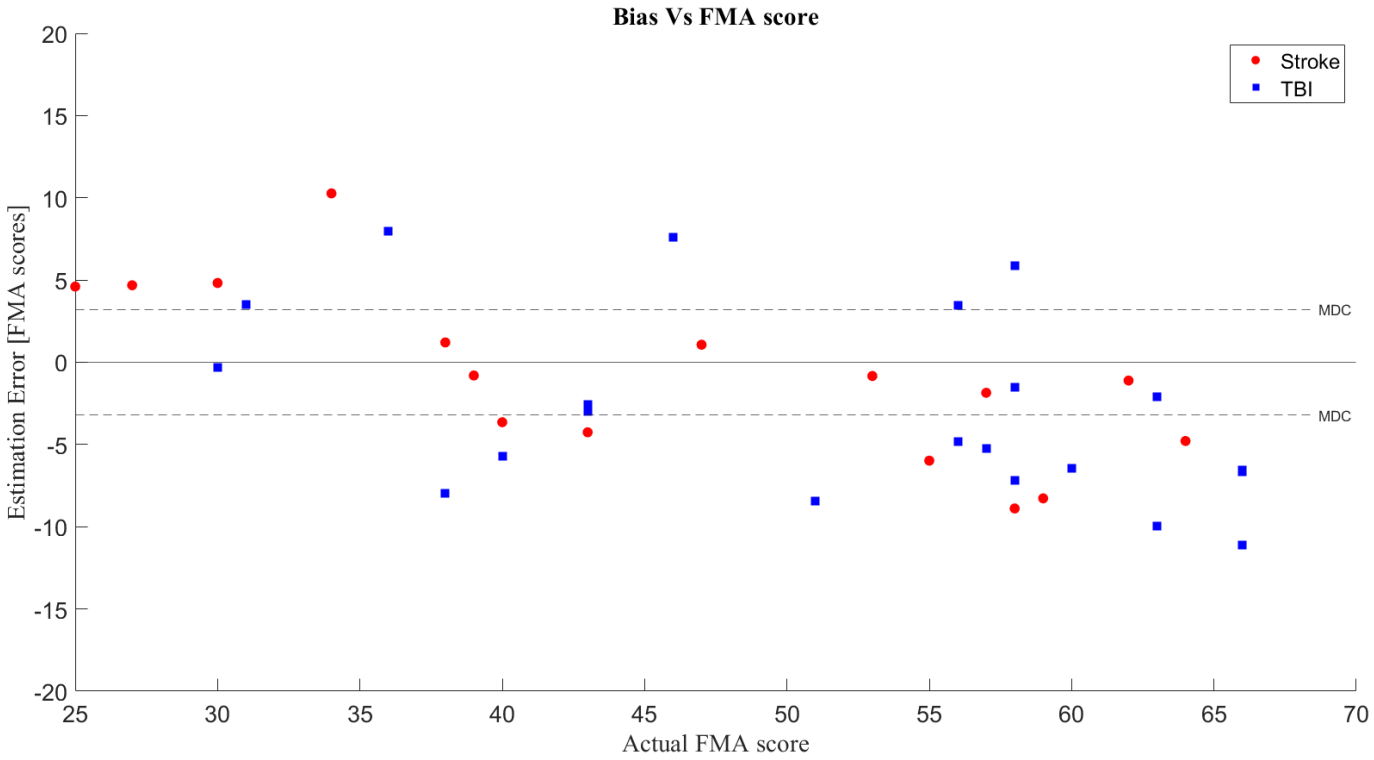

Figure 4.25: Bias analysis for FMA predictions: the estimation errors of FMA scores predictions of trials are plotted against the actual FMA scores; the range of Minimum Detectable Change (MDC) is plotted with dashed lines. The model behaves without any consistently difference towards the stroke survivors and the TBI patients.

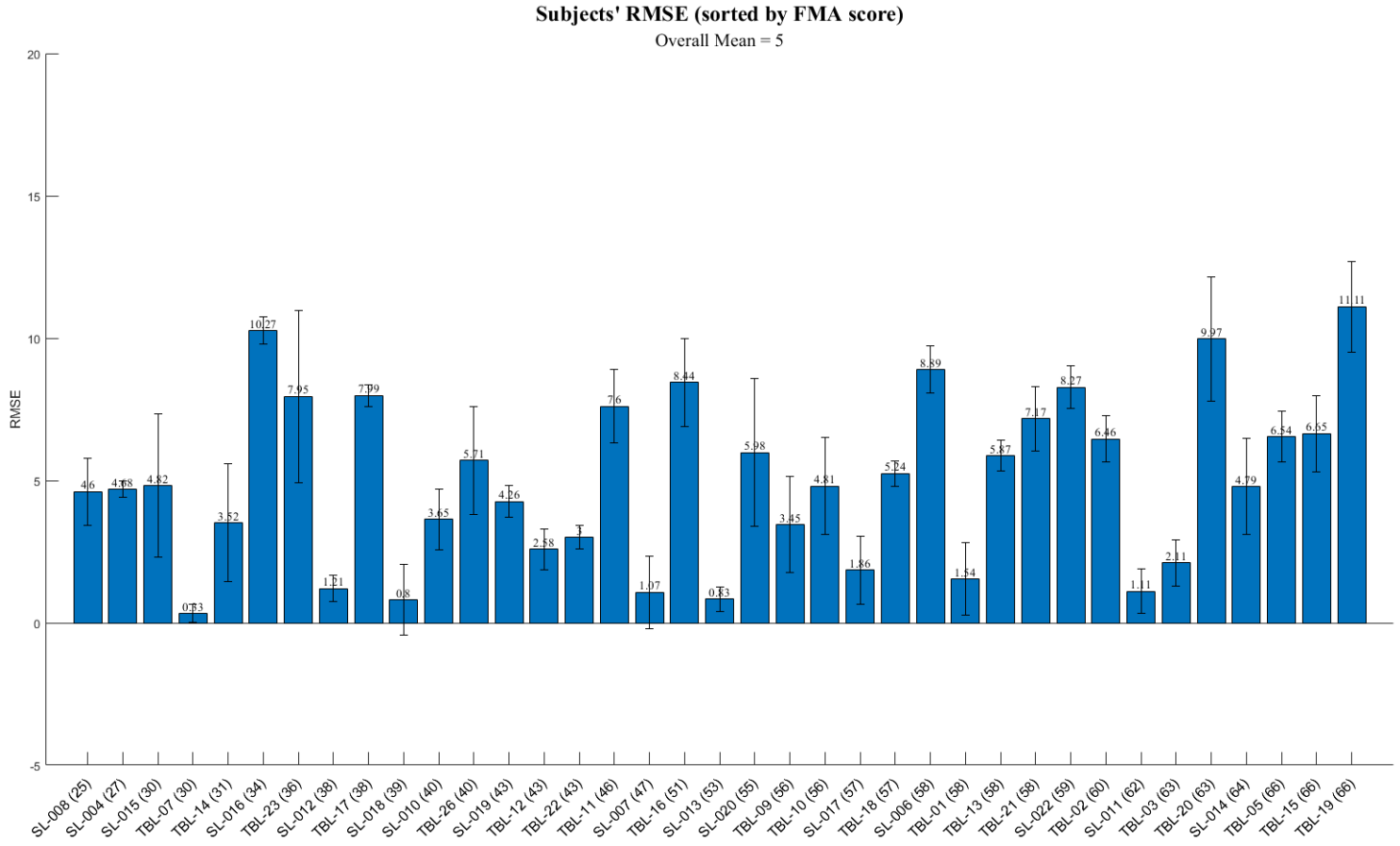

Figure 4.26: Subjects' RMSE analysis for FMA predictions: the RMSEs of each patient, calculated among their respective trials, are here shown; the patients are sorted in order of increasing actual FMA score. An overall RMSE of 5 FMA grades is here reported.

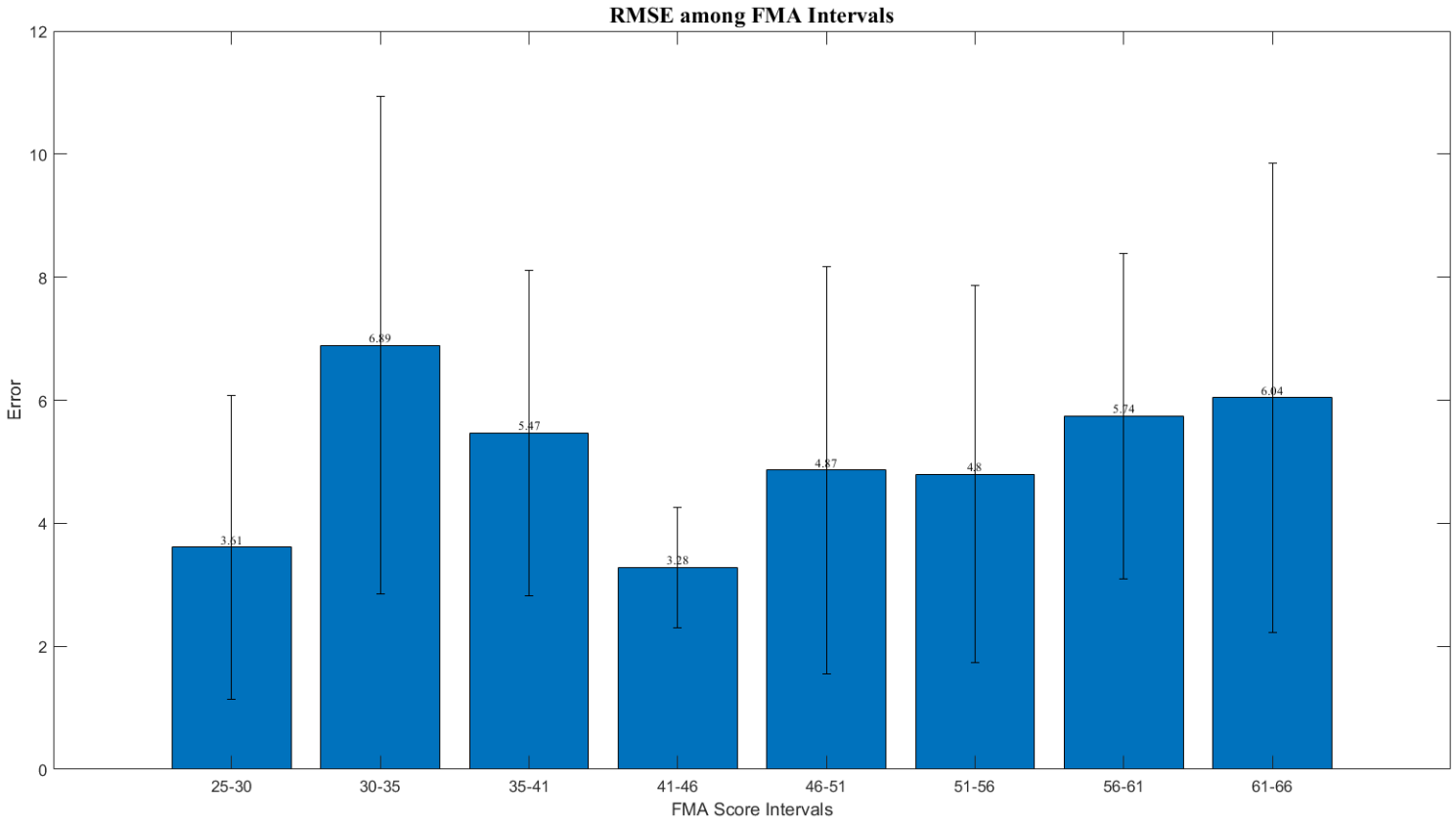

Figure 4.27: Analysis of FMA intervals RMSE for FMA predictions: the whole FMA interval of end-of-treatment recordings [25,66] is divided into intervals of 5-6 FMA points; this plot shows that no substantial difference exists in the algorithm ability to predict FMA scores that belong to different FMA scale portions.

4.4 Adding Clinical Features

This section briefly discusses, taking it one step further, how adding features based on clinical data of the patients may positively impact the previous model performance. For the sake of clarity, beyond the addition of these patients' clinical information no other changes are made to the model proposed in the last section.
In particular, the clinical features that were harvested are here described with a quick explanation of the grounds on which the decision to include them is taken:

- Age: it is reasonable to think that the response of a patient to a rehabilitation program may depend on his age and that, at the same stage of the disease, a younger patient should recover faster compared to an elder subject;
- Chronicity: it indicates how long the patient has been suffering from the disease; it gives gross information regarding the evolution of the disease and what stage the patient has reached;
- Days between assessments: it is a way of tracking how much time elapsed between the baseline and the discharge; moreover it is a way of quantifying how much time the rehabilitation program lasted;

To be more accurate, the features just described are added to the FMA scores estimation algorithm for single tasks; thus, in summary, the inputs for this algorithm are the wearable accelerometers data, the FAS estimates for single tasks and the clinical features in question.

4.4.1 Clinical Features Importance

Evaluating the permutation-based feature importance during the training of the balanced RFs, analyzing one task at a time, it is possible to understand, firstly, which clinical features are selected by the CFS and, then, which are predictive in a quantitative way. The information regarding the importance of the clinical features, that can be extracted from the figures 4.28, 4.29, 4.30 and 4.31, are summarized in the table below 4.3. Analyzing it, it is possible to say that:

- Age: even if it was considered to be relevant, the statistics show that it is selected in 1/4 tasks and with low importance;
- Chronicity: it is selected in 2/4 tasks and it has moderate importance;
- Days between assessments (DBA): it is selected in all the tasks and it has high importance.

While DBA turns out to be one of the most relevant features for all the tasks, the chronicity has a decent contribution in only 2 tasks and the age is overwhelmed by the other features. These statistics should not lead to the conclusion that the age clinical feature is not correlated with the label but rather that the extracted features are more predictive of it.

TASK	AGE	CHRONICITY	DBA
1	×	×	0.6
3	×	0.3	0.7
4	0.2	×	0.7
6	×	0.4	0.9

Table 4.3: Stats of permutation-based importance of clinical features; "days between assessments" is here abbreviated as "DBA"; a crossed box means that the clinical feature was not selected by the CFS for that task.

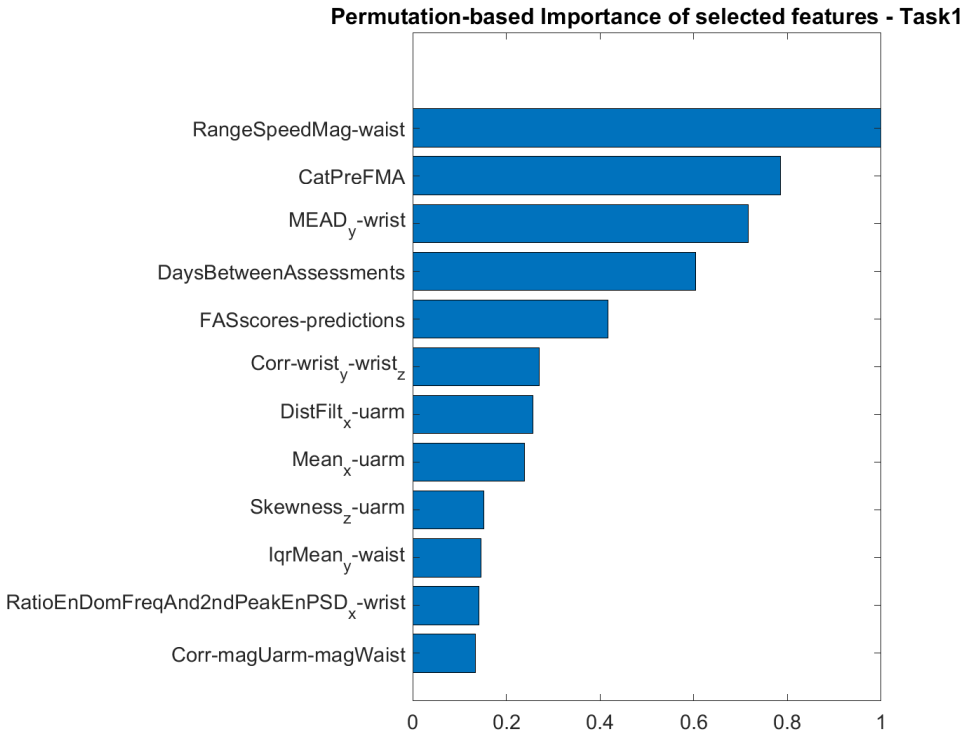

Figure 4.28: Feature Importance for task 1 (clinical features added).

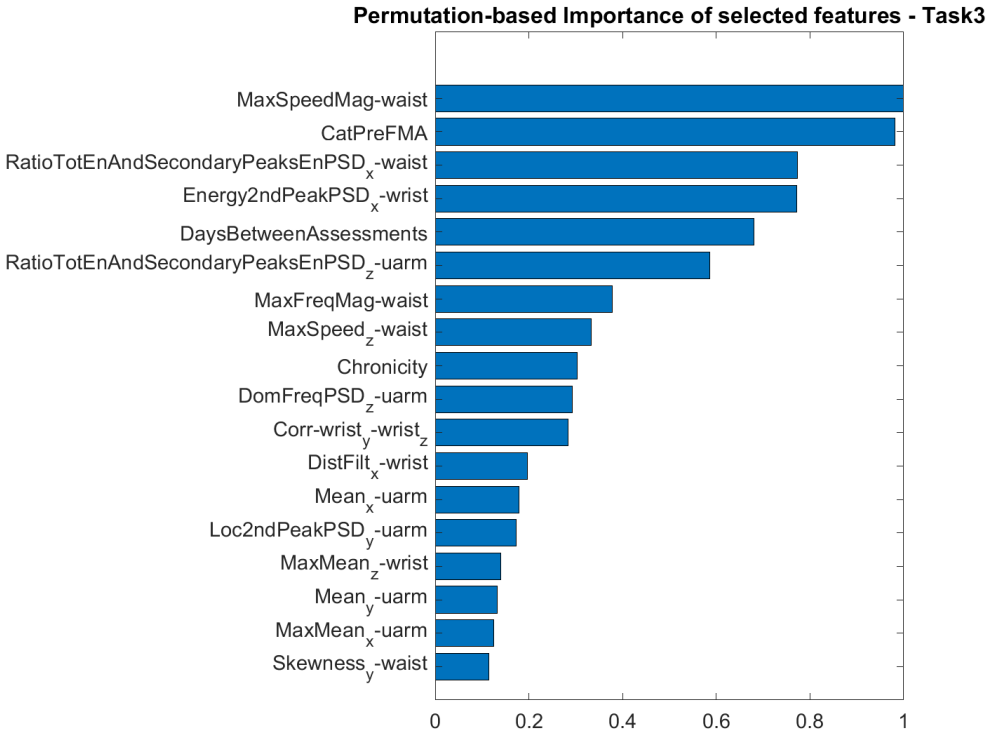

Figure 4.29: Feature Importance for task 3 (clinical features added).

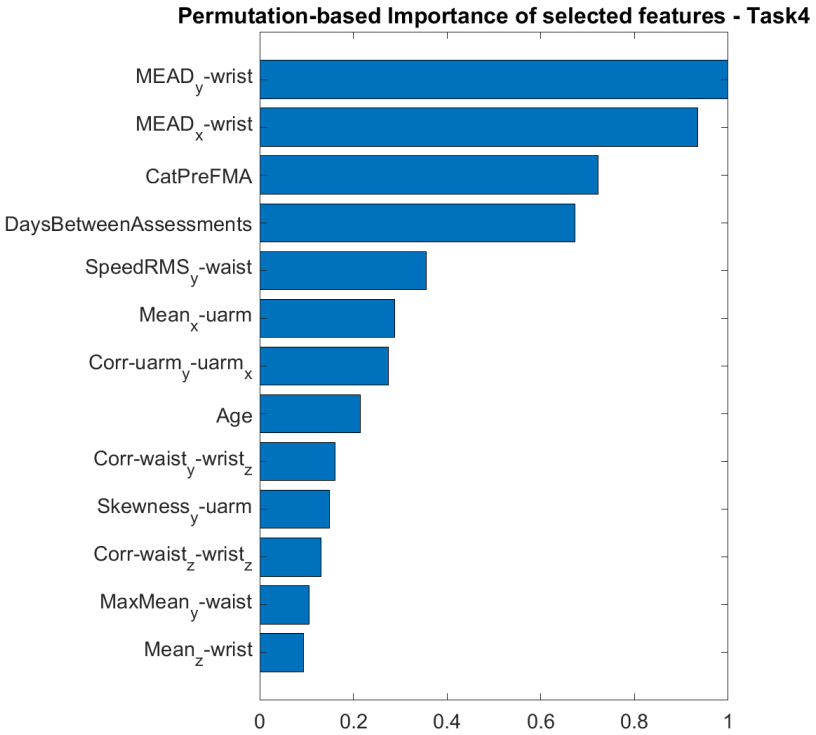

Figure 4.30: Feature Importance for task 4 (clinical features added).

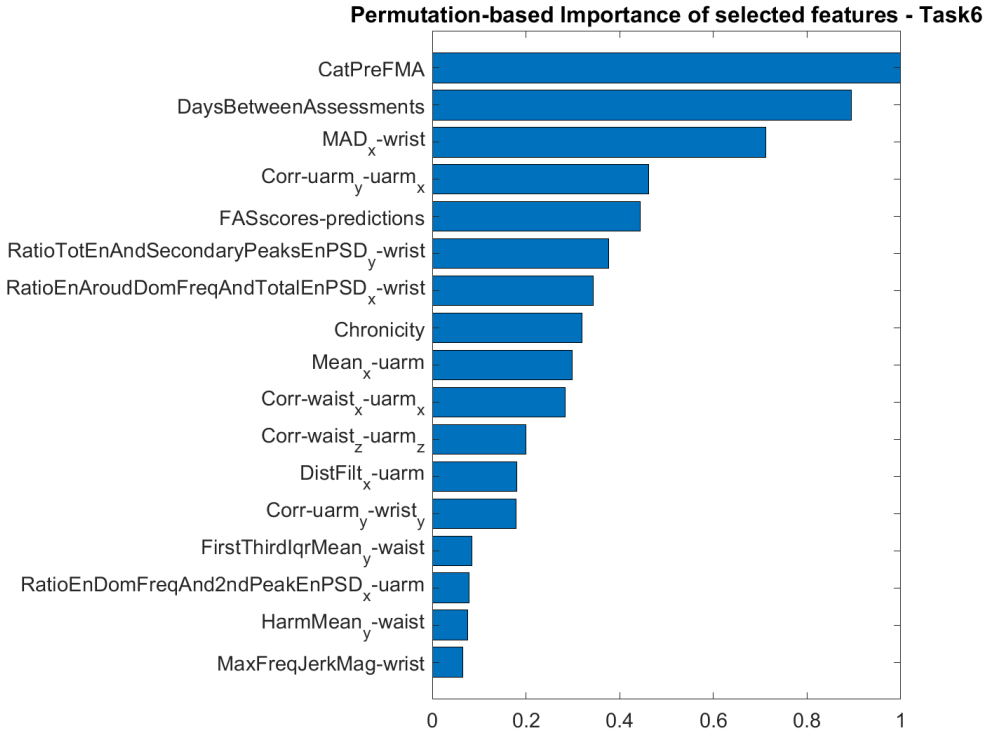

Figure 4.31: Feature Importance for task 6 (clinical features added).

4.4.2 Improvement in performance

From the figure 4.32 it is possible to assess that, thanks to the addition of the clinical features, and in particular of DBA and chronicity, the coefficient of determination R^2 slightly increased from 0.78 to 0.81. Moreover, while the estimation error increased for patients whose FMA scores are near the minimum, for the other subjects it decreased.
In the figure 4.33, comparing it with the previous 4.25, it is possible to see that there is still some bias, but from a FMA score of 40 on up, it reduces. This behavior is confirmed looking at the figure 4.34 where the subjects' RMSE are plotted: the overall error reduces from 5 to 4.34, but for lower FMA scores it increases. Plotting the RMSE among FMA intervals, as in the figure 4.35, it becomes clear that adding clinical features improves globally the FMA scores predictions at the expense of the low FMA scores estimates. Furthermore, when adding clinical features, no different behavior is noticed between the predictions of stroke survivors and patients affected by TBI. This justifies one more time the decision to include in the same dataset patients that suffer from different diseases when their data don't show evidence of distinctions.

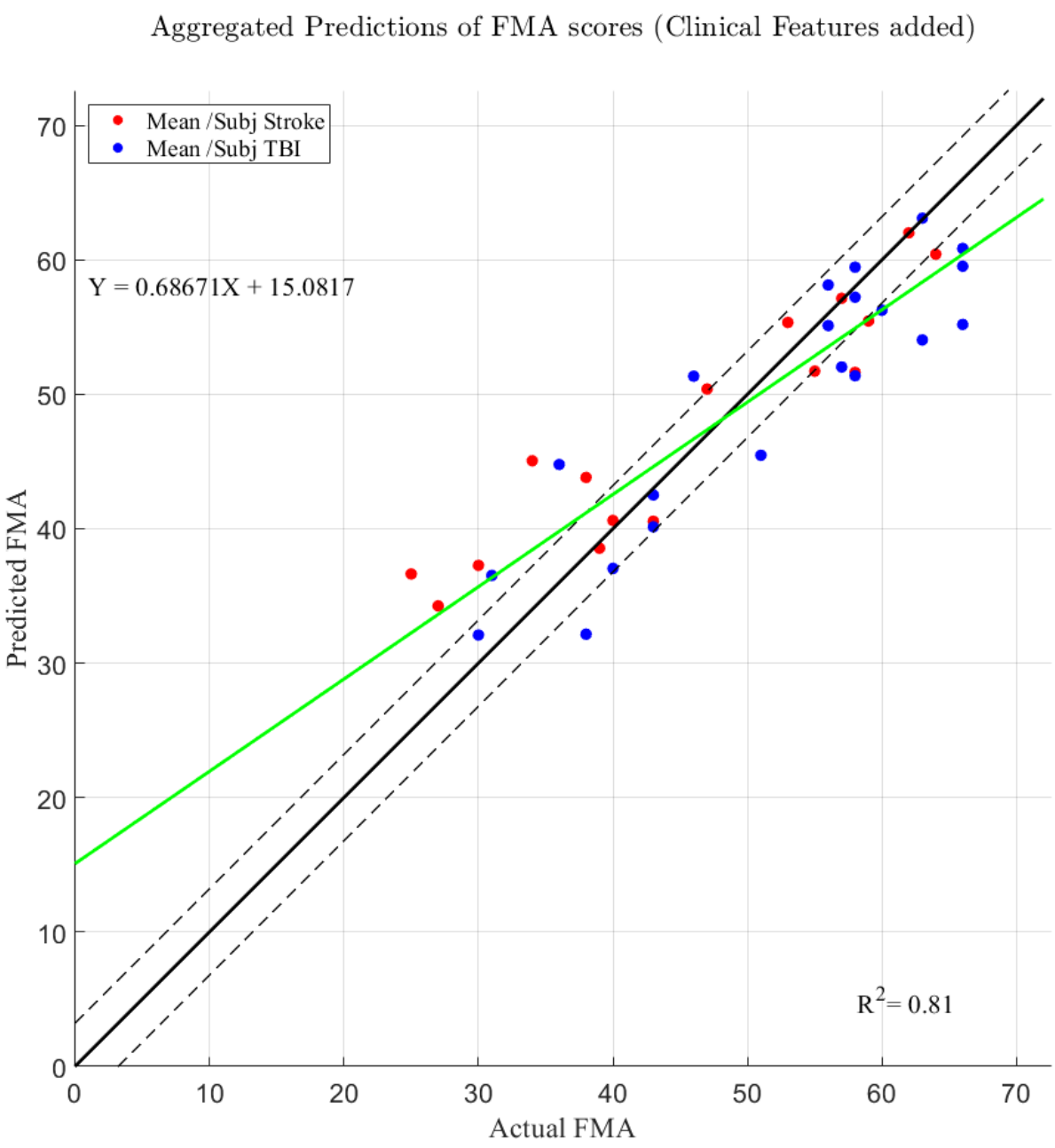

Figure 4.32: FMA predictions of subjects' scores; the regression line, in green, is derived relating the predicted FMA scores to the actual ones; in red are plotted the FMA estimates for stroke survivors, while in blue for TBI patients; the confidence interval is graphically shown with dashed lines.

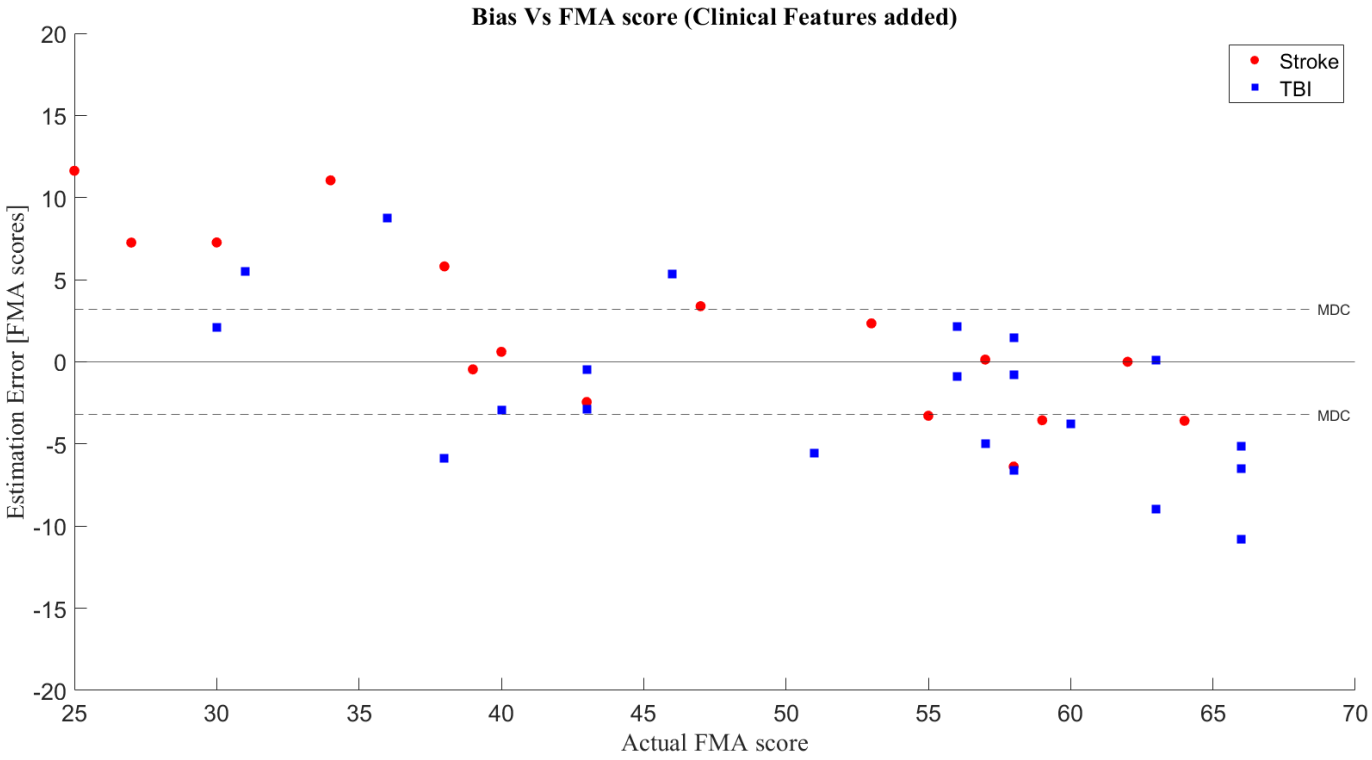

Figure 4.33: Bias analysis for FMA predictions: the estimation errors of FMA scores predictions of subjects are plotted against the actual FMA scores; the range of Minimum Detectable Change (MDC) is plotted with dashed lines. The model behaves without any consistently difference towards the stroke survivors and the TBI patients.

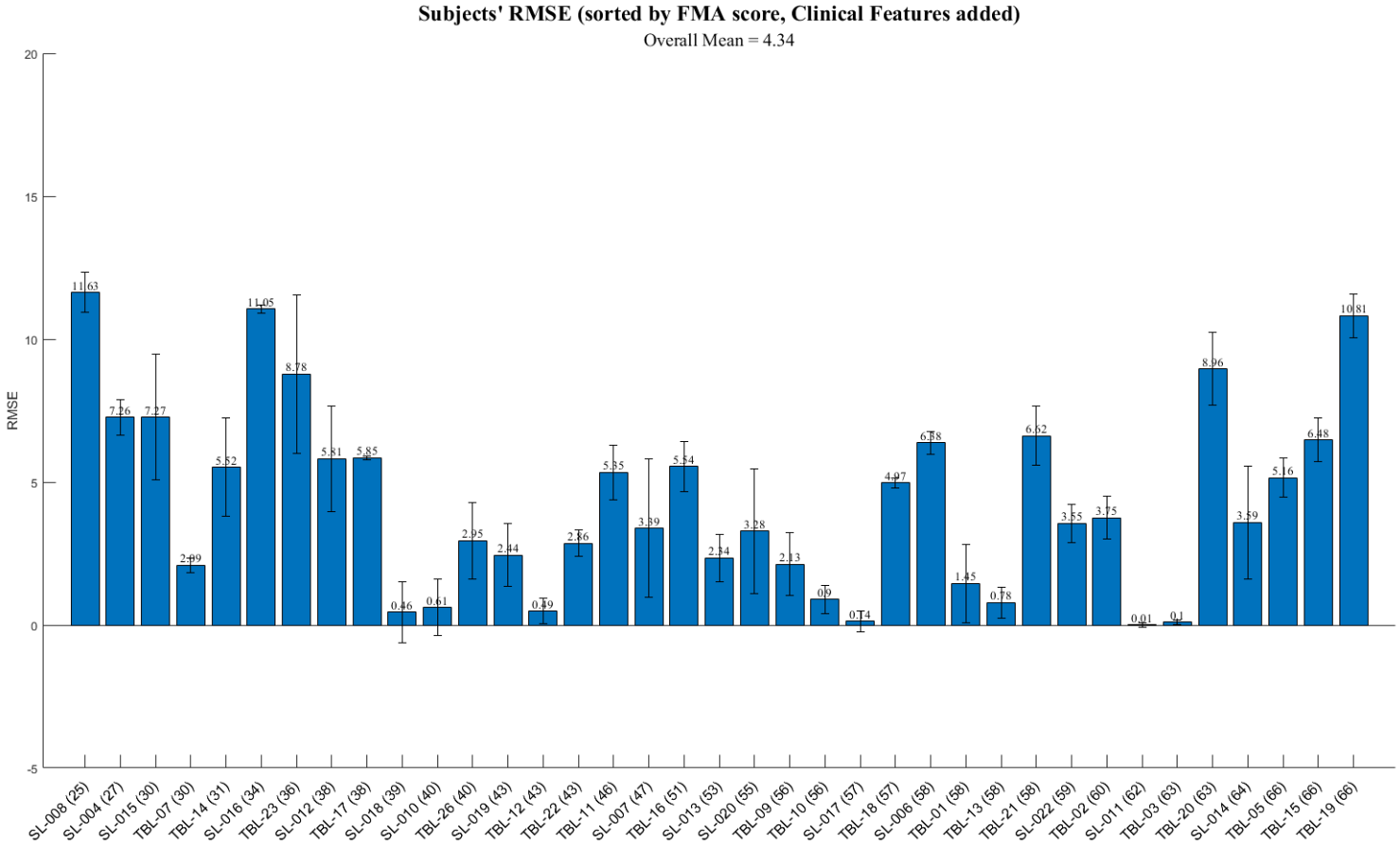

Figure 4.34: Subjects' RMSE analysis for FMA predictions: the RMSEs of each patient, calculated among their respective trials, are here shown; the patients are sorted in order of increasing actual FMA score. An overall RMSE of 4.34 FMA grades is here reported.

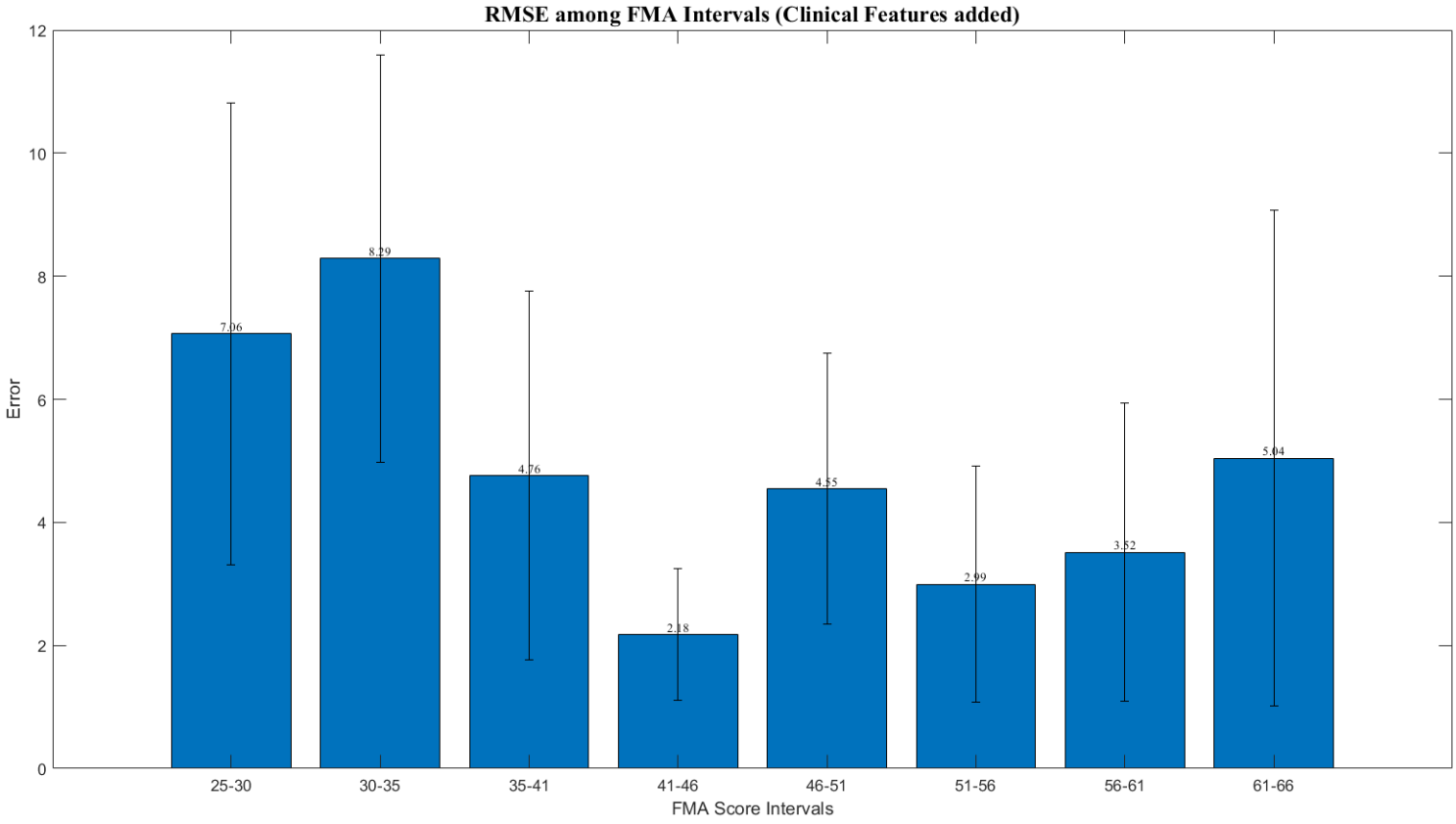

Figure 4.35: Analysis of FMA intervals RMSE for FMA predictions: the whole FMA interval of end-of-treatment recordings [25,66] is divided into intervals of 5-6 FMA points; this plot shows that no substantial difference exists in the algorithm ability to predict FMA scores that belong to different FMA scale portions.

Chapter 5

Conclusion

5.1 Conclusions

The work presented in this thesis demonstrates that data recorded with wearable accelerometers, during the performance of Activities of Daily Living tasks, can be used to assess the movement quality and motor impairment, through FAS and FMA respectively, in stroke survivors and patients suffering from traumatic brain injury.
Specifically, movement quality can be accurately estimated with a coefficient of determination R^2 of 0.83 and adding these predictions to the algorithm used for estimating the motor impairment a performance improvement of 8.3%, in terms of accuracy, is recorded in the latter. In this way, motor impairment can be assessed with R^2 of 0.78 and with a global RMSE of 5 over the sample.
In order to provide reliable evaluations of the model, leave-one-out cross-validation technique is used, since it is known to be suitable to assess the generalizability.
It has to be mentioned that the proposed model reached these performance despite the small sample size and the nonuniform distribution of the available clinical scores; the implemented Random Forest handles well this small dataset and is robust to overfitting. With a larger dataset, that includes observations all along the clinical scores scale, it would be possible to reach a quasi-uniform distribution of the estimation error; this would lead to a consistent decrease in the bias and to more accurate predictions.

Considering that these results are obtained using only wearable accelerometers data, going one step further, it is shown that adding features based on patients' clinical data has a positive impact on the proposed model, enhancing its performance. In particular, the time that elapsed between the recordings at baseline and discharge, and the disease chronicity, turned out to be predictive of the rehabilitation outcome.
Enriching the dataset with these clinical features, the model performance increased, reaching R^2 of 0.81 and reducing the global RMSE down to 4.34. Moreover, this work confirms, once again, that clinical rehabilitation outcomes of stroke survivors and patients affected by traumatic brain injury can be evaluated applying the same approach for both of them.
In order to better understand the importance of this thesis achievements and to contextualize them within the rehabilitation research, it is worth pointing out that, in the previous studies, standardized tasks belonging to the WMFT were used. On the contrary, in this work, patients performed the less constrained ADL tasks; since these tasks are not standardized, the subjects had more freedom in their movements, focusing mainly on reaching the goal required by the tasks. From the point of view of achieving high performance in terms of accuracy and error, the approach based on the analysis of data derived from ADL tasks is disadvantaged, since the variability of the signal across the subjects is much higher in respect of the standardized tasks one. On the other hand, the clinical scores predictions obtained in this way are more descriptive and realistic about the condition of the patients since they are evaluated on their way of accomplishing self-care tasks.

5.2 Future Work

Wearable technologies, as shown in this work, provide a cost-effective solution for tracking the clinical outcomes of rehabilitation programs; this suggests that, in the near future, their usage for this purpose will be widely diffused. Moreover, not far from today, it should be possible to move the recordings from a clinical setting, towards a home-based context. This would drastically reduce patients' and clinicians' burden, since, performing ADLs, the recordings do not require the presence of qualified staff, thus they could be done autonomously by the subjects at their home. More on this, an interactive platform could help the patients following them during the data collection and it could automatically upload the data to the laboratory where the analysis will be performed.
This approach would enable frequent recordings during extended rehabilitation interventions, leading to the possibility of fitting a precise motor recovery trajectory, useful for clinical decision making. The mentioned curve may be helpful in evaluating the response of a patient to an intervention, facilitating the clinicians, if necessary, to adjust and fit the treatment to him, as established by the precision rehabilitation principles.
In order to improve the model performance in predicting the clinical scores, it would be useful to include more subjects, especially with lower clinical ratings, trying to cover as much as possible all the scales range. One of the key focuses when feeding this dataset with new observations should be to strive to build it as little imbalanced as possible. Moreover, collecting more clinical information concerning, for instance, the total duration of the rehabilitation sessions between two recordings, or regarding the total recovery time among these sessions, may positively impact the model.

Bibliography

[1] C. Adans-Dester, N. Hankov, A. O'Brien, G. Vergara-Diaz, R. Black-Schaffer, R. Zafonte, J. Dy, S. I. Lee, and P. Bonato, "Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[2] S. Marshall, R. Teasell, N. Bayona, C. Lippert, J. Chundamala, J. Villamere, D. Mackie, N. Cullen, and M. Bayley, "Motor impairment rehabilitation post acquired brain injury," *Brain Injury*, vol. 21, no. 2, pp. 133–160, 2007.

[3] S. Patel, R. Hughes, T. Hester, J. Stein, M. Akay, J. G. Dy, and P. Bonato, "A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology," *Proceedings of the IEEE*, vol. 98, pp. 450–461, March 2010.

[4] P. Bonato, "Advances in wearable technology and applications in physical medicine and rehabilitation," 2005.

[5] S. Del Din, S. Patel, C. Cobelli, and P. Bonato, "Estimating fugl-meyer clinical scores in stroke survivors using wearable sensors," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5839–5842, IEEE, 2011.

[6] E. Taub, D. M. Morris, J. Crago, D. K. King, M. Bowman, C. Bryson, S. Bishop, S. Pearson, and S. E. Shaw, "Wolf motor function test (wmft) manual," *Birmingham: University of Alabama, CI Therapy Research Group*, 2011.

[7] E. E. Waehrens and A. Fisher, "Improving quality of adl performance after rehabilitation among people with acquired brain injury," *Scandinavian journal of occupational therapy*, vol. 14, no. 4, pp. 250–257, 2007.

[8] H. Xu, K. E. Covinsky, E. Stallard, J. Thomas III, and L. P. Sands, "Insufficient help for activity of daily living disabilities and risk of all-cause hospitalization," *Journal of the American Geriatrics Society*, vol. 60, no. 5, pp. 927–933, 2012.

[9] L. Jolliffe, N. A. Lannin, D. A. Cadilhac, and T. Hoffmann, "Systematic review of clinical practice guidelines to identify recommendations for rehabilitation after stroke and other acquired brain injuries," *BMJ open*, vol. 8, no. 2, p. e018791, 2018.

[10] V. L. Feigin, S. Barker-Collo, R. Krishnamurthi, A. Theadom, and N. Starkey, "Epidemiology of ischaemic stroke and traumatic brain injury," *Best Practice & Research Clinical Anaesthesiology*, vol. 24, no. 4, pp. 485–494, 2010.

[11] A. Guzik and C. Bushnell, "Stroke epidemiology and risk factor management," *CONTINUUM: Lifelong Learning in Neurology*, vol. 23, no. 1, pp. 15–39, 2017.

[12] C. Flach, W. Muruet, C. D. Wolfe, A. Bhalla, and A. Douiri, "Risk and secondary prevention of stroke recurrence: A population-base cohort study," *Stroke*, vol. 51, no. 8, pp. 2435–2444, 2020.

[13] M. Katan and A. Luft, "Global burden of stroke," in *Seminars in neurology*, vol. 38, pp. 208–211, Georg Thieme Verlag, 2018.

[14] S. R. Belagaje, "Stroke rehabilitation," *CONTINUUM: Lifelong Learning in Neurology*, vol. 23, no. 1, pp. 238–253, 2017.

[15] M. Faul and V. Coronado, "Epidemiology of traumatic brain injury," in *Handbook of clinical neurology*, vol. 127, pp. 3–13, Elsevier, 2015.

[16] H. Numminen, "The incidence of traumatic brain injury in an adult population–how to classify mild cases?," *European journal of neurology*, vol. 18, no. 3, pp. 460–464, 2011.

[17] W. Peeters, R. van den Brande, S. Polinder, A. Brazinova, E. W. Steyerberg, H. F. Lingsma, and A. I. Maas, "Epidemiology of traumatic brain injury in europe," *Acta neurochirurgica*, vol. 157, no. 10, pp. 1683–1696, 2015.

[18] K. S. Chua, Y.-S. Ng, S. G. Yap, and C.-W. Bok, "A brief review of traumatic brain injury rehabilitation," *Annals-Academy of Medicine Singapore*, vol. 36, no. 1, p. 31, 2007.

[19] M. R. Kosorok and E. B. Laber, "Precision medicine," *Annual review of*

statistics and its application, vol. 6, pp. 263–286, 2019.

[20] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 331–355, 2003.

[21] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," *Kidney research and clinical practice*, vol. 36, no. 1, p. 3, 2017.

[22] J. M. Llovet, R. Montal, D. Sia, and R. S. Finn, "Molecular therapies and precision medicine for hepatocellular carcinoma," *Nature reviews Clinical oncology*, vol. 15, no. 10, pp. 599–616, 2018.

[23] C. Pauli, B. D. Hopkins, D. Prandi, R. Shaw, T. Fedrizzi, A. Sboner, V. Sailer, M. Augello, L. Puca, R. Rosati, *et al.*, "Personalized in vitro and in vivo cancer models to guide precision medicine," *Cancer discovery*, vol. 7, no. 5, pp. 462–477, 2017.

[24] C.-Y. Chang, B. Lange, M. Zhang, S. Koenig, P. Requejo, N. Somboon, A. A. Sawchuk, and A. A. Rizzo, "Towards pervasive physical rehabilitation using microsoft kinect," in *2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops*, pp. 159–162, IEEE, 2012.

[25] G. L. Iverson, "Network analysis and precision rehabilitation for the post-concussion syndrome," *Frontiers in neurology*, vol. 10, p. 489, 2019.

[26] A. Criminisi and J. Shotton, *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.

[27] M. A. Morris, B. Saboury, B. Burkett, J. Gao, and E. L. Siegel, "Reinventing radiology: big data and the future of medical imaging," *Journal of thoracic imaging*, vol. 33, no. 1, pp. 4–16, 2018.

[28] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," *Korean journal of radiology*, vol. 18, no. 4, p. 570, 2017.

[29] Z. Lv, J. Chirivella, and P. Gagliardo, "Bigdata oriented multimedia mobile health applications," *Journal of medical systems*, vol. 40, no. 5, p. 120, 2016.

[30] R. S. Istepanian and T. Al-Anzi, "m-health 2.0: new perspectives on mobile health, machine learning and big data analytics," *Methods*, vol. 151, pp. 34–40, 2018.

[31] R. Gamache, H. Kharrazi, and J. P. Weiner, "Public and population

health informatics: the bridging of big data to benefit communities," *Yearbook of medical informatics*, vol. 27, no. 1, p. 199, 2018.

[32] A. A. Anoushiravani, J. Patton, Z. Sayeed, M. M. El-Othmani, and K. J. Saleh, "Big data, big research: implementing population health-based research models and integrating care to reduce cost and improve outcomes," *Orthopedic Clinics*, vol. 47, no. 4, pp. 717–724, 2016.

[33] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[34] C. V. Granger, A. C. Cotter, B. B. Hamilton, and R. C. Fiedler, "Functional assessment scales: a study of persons after stroke," *Archives of physical medicine and rehabilitation*, vol. 74, no. 2, pp. 133–138, 1993.

[35] M. Woodbury, C. A. Velozo, P. A. Thompson, K. Light, G. Uswatte, E. Taub, C. J. Winstein, D. Morris, S. Blanton, D. S. Nichols-Larsen, *et al.*, "Measurement structure of the wolf motor function test: implications for motor control theory," *Neurorehabilitation and neural repair*, vol. 24, no. 9, pp. 791–801, 2010.

[36] D. J. Gladstone, C. J. Danells, and S. E. Black, "The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties," *Neurorehabilitation and neural repair*, vol. 16, no. 3, pp. 232–240, 2002.

[37] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance.," *Scandinavian journal of rehabilitation medicine*, vol. 7, no. 1, pp. 13–31, 1975.

[38] J. Sanford, J. Moreland, L. R. Swanson, P. W. Stratford, and C. Gowland, "Reliability of the fugl-meyer assessment for testing motor performance in patients following stroke," *Physical therapy*, vol. 73, no. 7, pp. 447–454, 1993.

[39] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," in *Iberian conference on pattern recognition and image analysis*, pp. 289–296, Springer, 2011.

[40] A. Béliveau, G. T. Spencer, K. A. Thomas, and S. L. Roberson, "Evaluation of mems capacitive accelerometers," *IEEE Design & Test of Computers*, vol. 16, no. 4, pp. 48–56, 1999.

[41] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, and V. Kasi, "Reviews on

various inertial measurement unit (imu) sensor applications," *International Journal of Signal Processing Systems*, vol. 1, no. 2, pp. 256–262, 2013.

[42] S. Tewary, S. Chakraborty, J. Majumdar, R. Majumder, D. Kundu, S. Ghosh, and S. D. Gupta, "A novel approach towards designing a wearable smart health monitoring system measuring the vital parameters and emergency situations in real-time and providing the necessary medical care through telemedicine," in *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1–8, IEEE, 2016.

[43] S. Patel, R. Hughes, T. Hester, J. Stein, M. Akay, J. G. Dy, and P. Bonato, "A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology," *Proceedings of the IEEE*, vol. 98, no. 3, pp. 450–461, 2010.

[44] S. I. Lee, C. Adans-Dester, A. O'Brien, G. Vergara, R. M. Black-Schaffer, R. Zafonte, J. Dy, and P. Bonato, "Predicting and monitoring upper-limb rehabilitation outcomes using clinical and wearable sensor data in brain injury survivors," *IEEE Transactions on Biomedical Engineering*, 2020.

[45] T. Hester, R. Hughes, D. M. Sherrill, B. Knorr, M. Akay, J. Stein, and P. Bonato, "Using wearable sensors to measure motor abilities following stroke," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, pp. 4–pp, IEEE, 2006.

[46] Ç. B. Erdaş, I. Atasoy, K. Açıcı, and H. Oğul, "Integrating features for accelerometer-based activity recognition," *Procedia Computer Science*, vol. 98, pp. 522–527, 2016.

[47] W. Dargie, "Analysis of time and frequency domain features of accelerometer measurements," in *2009 Proceedings of 18th International Conference on Computer Communications and Networks*, pp. 1–6, IEEE, 2009.

[48] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, pp. 1–17, Springer, 2004.

[49] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*, vol. 10, pp. 359–370, Seattle, WA, USA:, 1994.

[50] S. Sempena, N. U. Maulidevi, and P. R. Aryan, "Human action recognition using dynamic time warping," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pp. 1–5, IEEE, 2011.

[51] R. Herren, A. Sparti, K. Aminian, and Y. Schutz, "The prediction of speed and incline in outdoor running in humans using accelerometry.," *Medicine and science in sports and exercise*, vol. 31, no. 7, pp. 1053–1059, 1999.

[52] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[53] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[54] L. Rokach and O. Maimon, "Decision trees," in *Data mining and knowledge discovery handbook*, pp. 165–192, Springer, 2005.

[55] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[56] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328, IEEE, 2008.

[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[58] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, pp. 973–978, Lawrence Erlbaum Associates Ltd, 2001.

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[60] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *International workshop on machine learning and data mining in pattern recognition*, pp. 154–168, Springer, 2012.