POLITECNICO DI TORINO

Dipartimento di Elettronica e delle Telecomunicazioni
Master Degree "ICT for Smart Societies"

Master Degree Thesis

Machine Learning Predictive System Based on Transaxial Mid-femur Computed Tomography Images

Supervisors
Prof.ssa Monica VISINTIN
Prof. Paolo GARGIULO (Reykjavik University)

Candidate
Marco RECENTI

April 2019

HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

POLITECNICO
DI TORINO

„Allir sem búa hér er ókeypis. Það er ennþá eitthvað.
Hvað er lok heimsins?
Hvað er lok heimsins fyrir þig, ég er heima.“

“Chiunque abita qui è libero. È pur sempre qualcosa.
La fine del mondo, che cos'è?
Quella che per te è la fine del mondo, per me è casa.”

Jón Kalman Stefànsson

Contents

List of Figures

List of Tables

NOMENCLATURE

ADA - ADA Boosting Algorithms

AGES - Age Gene/Environment Susceptibility Study

AI - Artificial Intellingence

BMI - Body Mass Index

CHD - Coronary Heart Disease - CHDEVENTB

CHF - Coronary Heart Failure - CHFBAGES

CHOL - Cholesterol

CVD - Coronary Vascular Disease - CVDEVENTB

CT - Computed Tomography

DM - Diabete Medication of type 1

DM2 - Diabete Medication of type 2

EX-T - Extremely Randomized Tree Algorithm

FGait - Gait Time Fast speed - TIMEFAST

GRAD-B - Gradient Tree Boosting Algorithm

HU - Hounsfield unit

ISO - Max Strenght in Leg - ISSOMASTLEG

JI - Jaccard Accuracy Index

ML - Machine Learning

N - Amplitude

NaN - Not A Number

NGait - Gait Time Normal speed - TIMENORMAL

NTRA - Nonlinear trimodal regression analysis

NTRAS1 - NTRA Selection 1 for BMI

NTRAS2 - NTRA Selection 2 for ISO

PDF - Probability Density Function

R^2 - Coefficient of Determination

RF - Random Forest Algorithm

SL - Scikit-Learn

TUG - Time Up and Go - TUGOSEC

μ - Location

σ - Width

α - Skewness

Acknowledgment

A thanks to all the staff and the participants of the AGES Reykjavik study for their important contribution: The Age, Gene/Environment Susceptibility Reykjavik Study has been funded by NIH contract N01-AG12100, the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association), and the Althingi (the Icelandic Parliament).

Chapter 1

Introduction

1.1 Overview

It is common knowledge that, nowadays, Artificial Intelligence (AI) technologies, in particular Machine Learning (ML) algorithms, are often used in healthcare applications in order to help physicians in diagnosis or to find possible relations between measured biomedical parameters. The increasing availability of healthcare data and the development of big data analytic methods has made possible the success of ML in different healthcare applications. In this project Artificial Intelligence Machine Learning Technologies will be used to link parameters that are apparently distant each other. Some relevant clinical information can be hidden inside the large quantity of data of many patients and cannot be easily visible even for an expert physician. In these cases a machine can find connections not predictable without the help of the computational power of the actual algorithmic technologies [1] [2]. There are several different specialties in medicine that have shown an increase in research regarding AI and ML in particular. Some of them are radiology and the related image processing, telemedicine, voice and speech recognition, drug development, personalized medicine, genetics, robot-assisted surgery and many other (see fig.1.1) [3] [4]. There are many examples in literature encouraging the implementation of data mining algorithms in different fields of medicine in order to discover hidden patterns or new information in large dataset [5] [6].
In this thesis, starting from the AGES-Reykjavik database, predictive analysis is done using supervised regression and classification ML algorithms. Age Gene/Environment

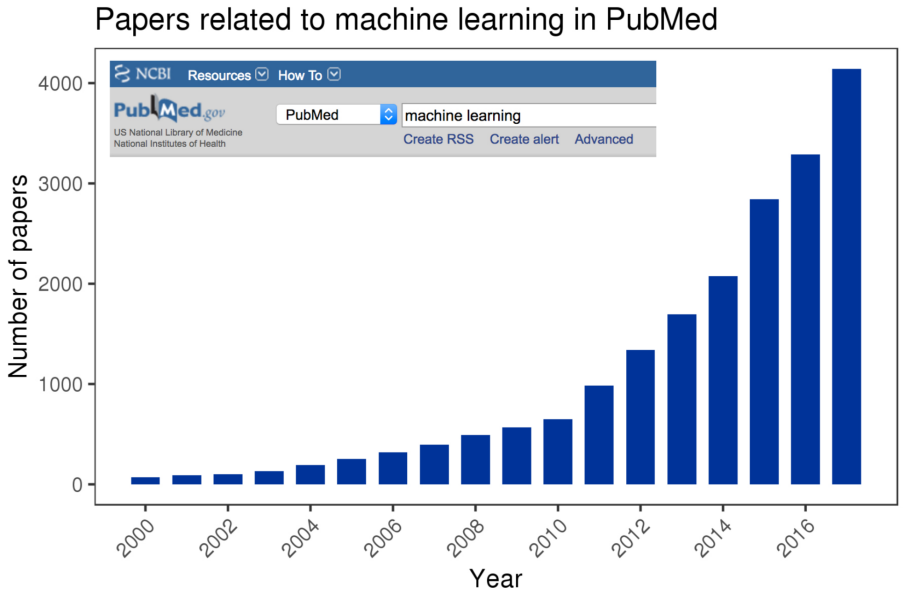

Figure 1.1: Growth of PubMed articles on machine learning [8]

Susceptibility Study (AGES) is a large dataset, designed to examine risk factors and gene/environment interaction, in relation to disease and disability in old age [7]. It was never investigated through machine learning methodologies. The database is composed of 11 parameters extracted from Computed Tomography (CT) scans of mid-femur section of a 65-95 years old population, (4 related to muscle tissues, 4 to the fat, and 3 to the connective tissues) and by 20 measurements of which the most relevant are Body Mass Index (BMI), Cholesterol (CHOL) and LEF biometric parameters (normal/fast gait speed, time up-to-go, and isometric leg strength).
Regression and classification are applied in order to predict and classify at first BMI (parameter used to test the methodology) using tree-based algorithms and then the algorithms are extended also to other measurements.

1.2 Goal of the thesis

The main goal of the work is to find links between the 11 parameters extracted from the CT scans and the other measurements to prevent or predict possible cardio-circulatory

or motor disease in elderly people. Machine learning solutions are experimented not directly on the pixels of the CT images but using Amplitude, Location, Width and Skewness of the fat, muscle and connective tissue and link these data to biomechanical measurements, BMI and Cholesterol which are apparently distant from a mid-femur CT scan.
Different methodology will be tested to predict these physiological measurements. Considering regression the feature analysis can be useful to understand which are the most relevant regressors between fat, muscle and connective tissue parameters. Hypothetically, if the regression results are good, also the classification ones should be good as well, and the latter can be considered as a confirmation of the strong or weak link between the measurement itself and the initial 11 features.

1.3 Thesis organization

A brief description of the contents of each chapter follows.

Chapter 2: Database Description and Elaboration

The database used in this thesis work is described in all its details.
The main features (columns) are analyzed and all the patients (rows) changes and cancellations are explained. In addition, outliers are analyzed and processed and the different solutions adopted in case of Not-A-Number (NaN) values in the database are listed.

Chapter 3: Supervised Machine Learning

The theory of the considered algorithms used is described with a particular focus on the methodologies that can be applied to divide the database in training and testing sets.

Chapter 4: Predictive System

The methodology used to obtain the best possible results is described. Starting from BMI prediction and classification, all the other parameters are later analyzed following appropriate processes.

Chapter 5: Regression and Classification Results

All the results are presented, starting from those obtained with the BMI, both for regression and classification, and continuing with the other parameters. Particular relevance is given to the related deductible biomedical considerations. A sub-section is completely dedicated to the leg strength results, which are better than all the others obtained from the LEF parameters and from the cholesterol.

Chapter 6: Conclusions

In this chapter the main conclusive considerations are shown following the obtained results.

Chapter 2

Database Description and Elaboration

The database used in this thesis is denominated AGES (Age Gene/Environment Susceptibility Study) and it is provided by Icelandic Hearth Association [1] [7] [9]. It is composed of 11 Nonlinear trimodal regression analysis (NTRA) parameters extracted from Computed Tomography (CT) scans of mid-femur section of a 65-95 years old population, (4 related to muscle tissues, 4 to the fat, and 3 to the connective tissues) and by 25 measurements of which the most relevant are Body Mass Index (BMI), Cholesterol (SCHOL) and LEF biometric parameters (normal/fast gait speed, time up-to-go, and isometric leg strength). There are 3157 patients in AGES I and the same number in AGES II (same measurements on the same patients taken 5-6 after AGES I), so in total 6314.
The related insights about the features and the number of patients are described in the following sections.

2.1 AGES Database - Features (columns)

The features of the database are mainly divided in two parts: 11 NTRA parameters extracted from the CT scans and 20 physiological measurements.

[1] The AGES dataset cannot be made publicly available, since the informed consent signed by the participants prohibits data sharing on an individual level, as outlined by the study approval by the Icelandic National Bioethics Committee

2.1.1 NTRA parameters

The 11 NTRA parameters are derived from a mid femur CT Scans (example in fig. 2.1) as described in details in [10] [11].

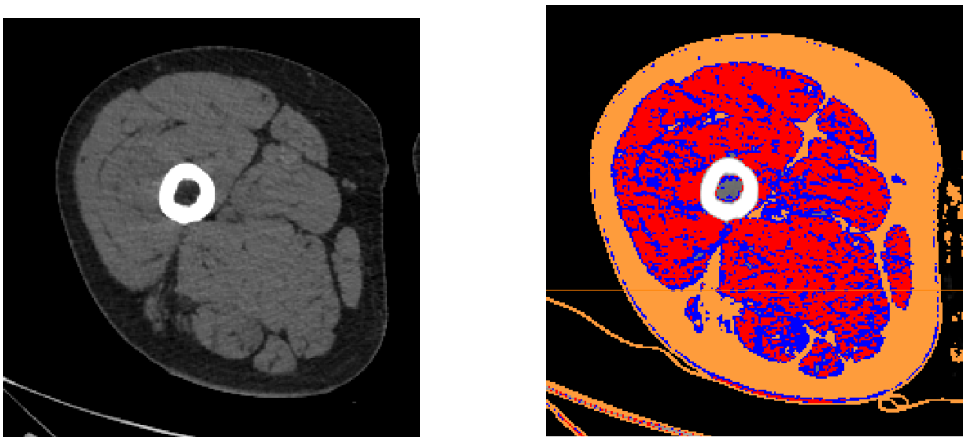

Figure 2.1: Mid-Femur CT scan: on the right FAT is orange, MUSCLE is red and CONNECTIVE TISSUE is blu

The localized scanning region extended from the iliac crest to the knee. For each patient, a single 10-mm thick transaxial mid-femur section was used in order to generate HU distributions and calculate fat and muscle cross-sectional area extensions [11]. HU is the *Hounsfield unit* scale [12]: it is a linear transformation of the original linear attenuation coefficient measurement into one in which the radiodensity of distilled water at standard pressure and temperature (STP) is defined as zero Hounsfield units (HU), while the radiodensity of air at STP is defined as -1000 HU. In a voxel with average linear attenuation coefficient μ , the corresponding HU value is:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{2.1}$$

As described in [10], for each patient, HU distribution were derived from each pixel's CT number value following the expression:

$$HU = CT \times 2,26625 - 190 \tag{2.2}$$

After this operation, HU values were binned into 128 bins and probability density functions (PDF) were derived consequently [13]. Each PDF was then exported for

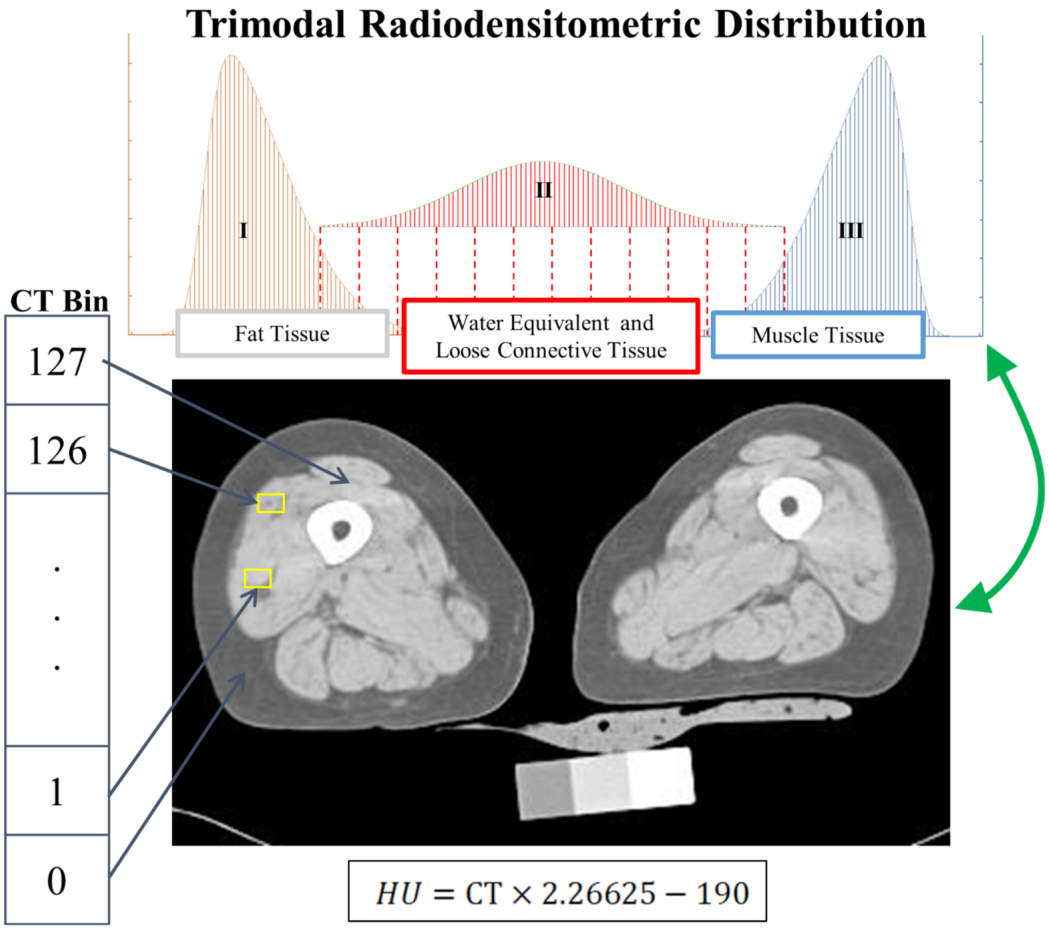

Figure 2.2: From CT scan to PDF [10]

NTRA regression analysis.

NTRA method was developed and described in details in [10]. After a series of operations, this method can give as results 11 parameters: 4 related to the *FAT*, 4 to the *MUSCLE* and 3 to the *CONNECTIVE TISSUE* (fig. 2.1) The 4 parameters are shown in the PDFs of fig. 2.3 and they are:

- N: Amplitude
- μ: Location
- σ: Width

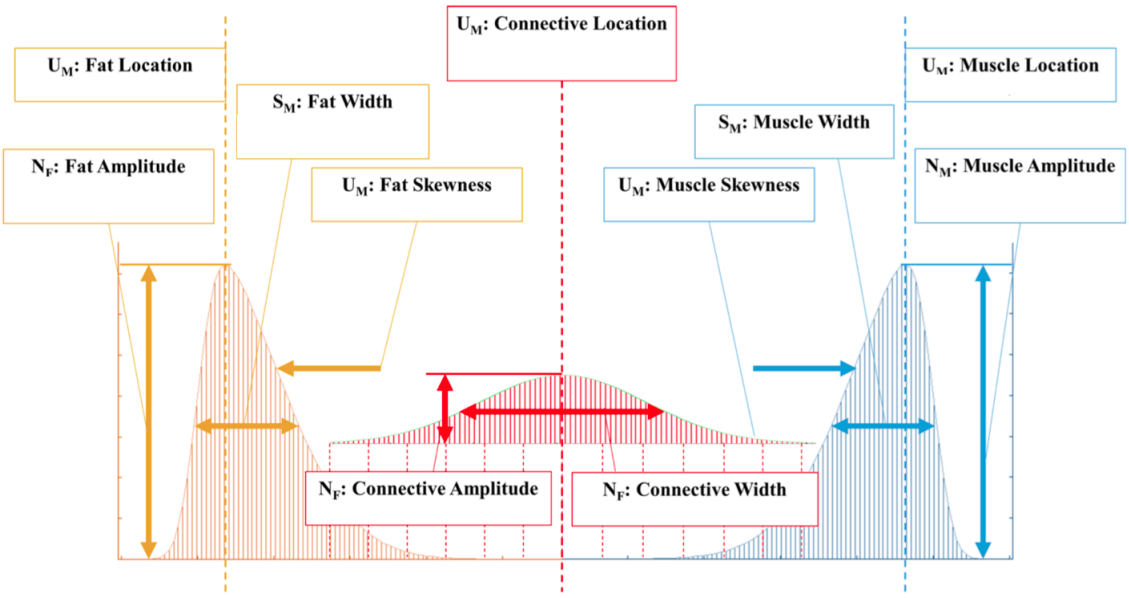

Figure 2.3: 11 NTRA parameters represented on relative PDFs [10]

- α: Skewness (not for the Connective Tissue)

Each of these 11 parameters is represented by a real number value. table 2.1 is presented in order to better understand the distributions of this values and the operations on the outliers that are going to be described in section 2.3. The values are related to AGES I, but there are not significant differences on AGES II. The reason why these values are very distant from each other is due to the CT scan itself whose pixels are different depending on whether you refer to fat, muscles or connective tissue.

2.1.2 Physiological Measurements

In addition the the 11 NTRA parameters the database is composed by other 20 measurements: all of them are numerical, both discrete and continuous values. Here all 20 measurements will be listed, the ones highlighted in **bold** will be those most used in subsequent chapters of this thesis :

- Age (on table 2.2 the age difference between the first visit in AGES I and the second one in AGES II for each patient) - Integer, min=65, max=95

AGES I	
NTRA parameter	mean value
N fat	61,9788
μ fat	-117,8241
σ fat	8,2447
α fat	-2,4914
N muscle	78,0249
μ muscle	61,4488
σ muscle	8,6205
α muscle	2,8307
N conn	41,6360
μ conn	-24,0571
σ conn	25,1132

Table 2.1: NTRA parameters mean in AGES I

AGE Difference	
AGE difference	Number of Patients
2	1
3	0
4	22
5	2537
6	568
7	25
8	3
9	1

Table 2.2: Age difference between AGES I and AGES II

- Sex (Men=1, Woman=2)
- Smoking Status (never smoked=0, smoke regularly=1, current smoker=2)
- **BMI** (Body Mass Index= $Weight/Height^2(kg/m^2)$) - Float, min=15,57, max=47,12
- DM (Diabete Medication of type 1) (No Diabete=0, otherwise 1)
- DM2 (Diabete Medication of type 2) (No Diabete=0, otherwise 1)
- **CHOL** (Cholesterol= mmol/L) - Float, min=2,3, max=9,74
- Hypertension (No Hypertension=0, Pre Hypertension=1, Hypertension=2)
- PHYSACTPAST (Category of total past moderate or vigorous physical activity score - never=1, rarely=2, occasionally=3, moderate=4, high=5)
- PHYSACTPRES (Category of frequency of moderate or vigorous physical activities in the past 12 months - never=1, rarely=2, occasionally=3, moderate=4, high=5)
- **TOGOSEC - (TUG)** (Time up and go: sec - Time to stand up from a chair, walk 3 m, go back to the chair and sit down, float value) - Float, min=6,73, max=25,8 [14]
- **TIMEFAST - (FGait)** (Walk 6M Time at fast speed: sec - 2 trial average, float value) GAIT FAST - Float, min=3,6, max=8,4 [15]
- **TIMENORMAL - (NGait)** (Average time to walk 6m at standard speed: sec, float value) GAIT NORMAL - Float, min=4,725, max=12,15 [15]
- **ISSOMASTLEG - (ISO)** (STRENGTH max strength in leg: Newtons, float value) - Float, min=212,8, max=245,7 [16]
- CHDEVENTB - CHD (coronary heart disease event before entering AGES - yes=1, no=0)
- CVDEVENTB - CVD (coronary vascular event before entering AGES (this includes stroke) - yes=1, no=0)

Database	Num Patients	Deleted ID	Total
NTRA I	3160	2923 3268 3672	**3157**
Meas I	3162	2923 3268 3452 3672 5065	**3157**

Table 2.3: **AGES I** - Patient ID cross-check

- CHFBAGES - CHF (coronary heart failure before entering AGES - yes=1, no=0)
- HEALLUNG (Has a doctor or other health provider ever told you that you had chronic lung disease, chronic bronchitis or emphysema? - yes=1, no=0)
- HEALPRKN (Has a doctor or health professional ever told you that you had Parkinson's disease? - yes=1, no=0)
- Mortality (Living=0 or Death=1)

The **bold** measurements are used both to predict (together with NTRA) and to be predicted during the machine learning predictive section. BMI is going be used as "test parameter" for the search for the optimal strategy for methodology of ML algorithms and ML train-test division.
All the **bold** measurements are continuous values and can be obviously different from the first visit of AGES I to the second one in AGES II.
Normal/fast gait speed, time up-to-go, and isometric leg strength are together called LEF measurements.

2.2 AGES Database - Patients (rows)

The provided databases are divided in AGES I and AGES II. Each of these is furthermore divided in NTRA (I and II) and Measurements (I and II). Each of these four

Database	Num Patients	Deleted ID	Total
NTRA II	3158	5065	**3157**
Meas II	3160	3268 3452 5065	**3157**

Table 2.4: **AGES II** - Patient ID cross-check

groups have a different number of patients. Patients are uniquely identified by ID number, not in a sequential order from 2 to 5859.

To avoid mistakes, the same number of patients, with the same identical ID numbers, should be present in all the four groups. A cross-check between the patient indices was made, as shown in table 2.3 and 2.4. At the end each database has 3157 patients for a total of 6314.

2.3 Outliers Management

Following an analysis of the distribution of each of the values in the database it was possible to ascertain the presence of outliers or values far from the average value of the parameter itself (Table 2.1). The outliers can affect the quality of the prediction but are absolutely relevant because they have, in this case, a specific biomedical meaning, referred to the muscle or fat or connective tissue of the patient taken from the CT scan. Anyway it is possible to spot some values that are extremely distant from the mean value, especially in the NTRA parameters, so a distinction has to be done. The outliers can be considered either as wrong values or as simple outliers. The wrong values must be manually changed. For example in the μconn of the AGES I a single value was more that 1000 with a mean of -24,057, as shown in fig. 2.4. After a manual modification of the single value the distribution is more regular even if some outliers can be individuated for example in μmusc as shown in fig. 2.5. The manual modification consists in a real manual change of the single wrong value: as shown in the example, the value above one thousand was 1027,34: the first two digits were eliminated in order to obtain 27,34, which is a more reasonable value given the distribution and the average of the

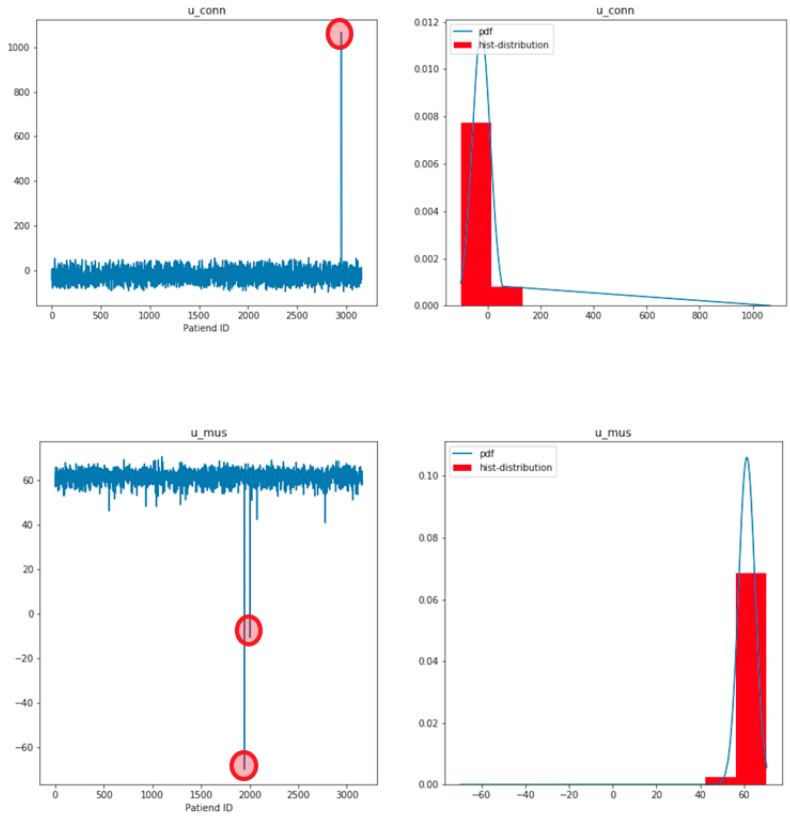

Figure 2.4: Red circles indicate **WRONG VALUES** in μconn and μmusc

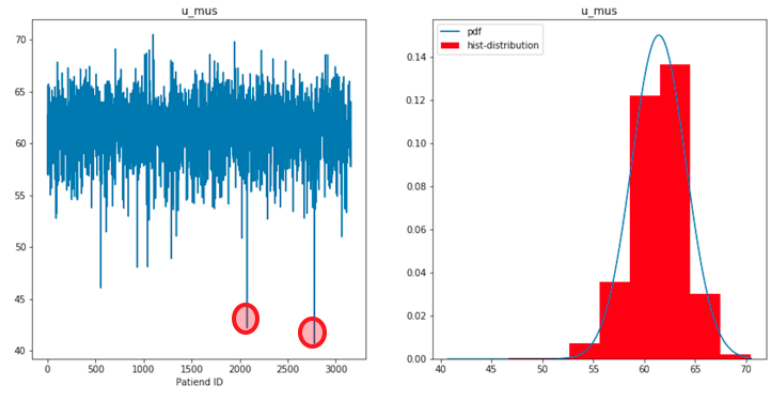

Figure 2.5: Red circles indicate **OUTLIERS** in μmusc after the manual modification of the wrong values

Parameter	AGES-I	AGES-II	AGES I+II
BMI	1	7	8
CHOL	0	0	0
TUG	39	106	145
FGait	122	345	467
NGait	44	109	153
ISO	193	229	422

Table 2.5: $N°$ NaN for all the measurements in the database

data considered.
The wrong values, which could be caused by a bad transcription, have been identified mainly in the NTRA values. As far as the other measurements are concerned, the values distant from the average value are not considered as wrong values but as simple outliers.

2.4 NaN Values

One of the main problem of the databases used to approach ML algorithms are the "Not A Number Values" (NaN). In many cases, especially in the medical databases, where lots of data are collected for thousand of patients, some of them are missing due to many possible reasons. For the AGES database this happens too: some patients do not have all the measurements, so the missing ones are considered as NaN values.

The NTRA values are always present for each patient, while for the other measurements is possible to see in Table 2.5 that some measurements as FGait or ISO have more than 400 NaN values in total. Fig. 2.6 shows the number of patients for AGES I and AGES II with at least one NaN value.

Two approaches are used to solve the NaN value problem [17]:

- **MEAN**: if there are not so many NaN values and the data are continuous: in this case the NaN value is substituted with the mean. This solution was adopted for BMI (8 NaN values in total).

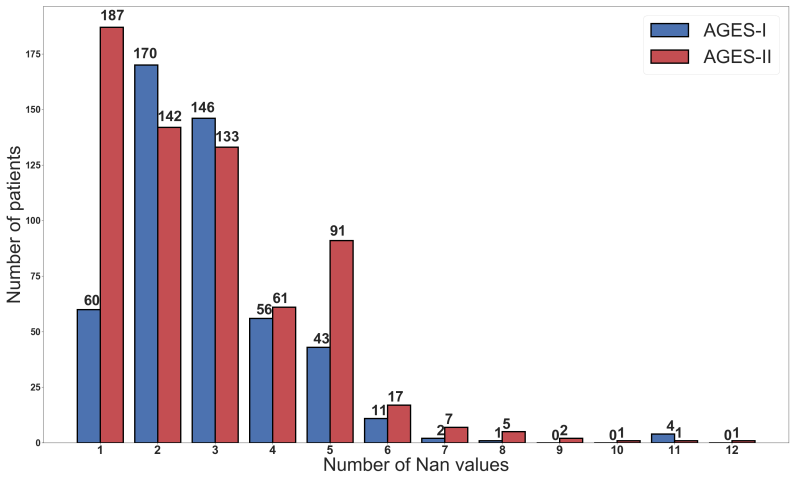

Figure 2.6: *N°* of Patients with at least one Nan value

- **DELETE**: if there are too many NaN values the best solution is to delete the patient's row otherwise the final prediction results can be compromised if Mean solution is adopted. This solution was adopted for TUG, FGait, NGait, ISO.

Chapter 3

Supervised Machine Learning

Machine learning is the discipline that studies how computers and machines learn from data. It is as a section of the Artificial Intelligence field of the Computer Science Area. [18] The two different types of learning used by machines are called supervised and unsupervised learning.
In **unsupervised learning** there are no outputs to predict: the machine divides the database in n groups or clusters trying to find commonalities in the data, not knowing if it is right or wrong. These types of algorithms are not going to be used in this thesis.
Supervised learning starts with the goal of predicting an output or a target that is already known: the goal of supervised learning is to learn a function that, given a sample of data and known outputs, best approximates the relationships between input and output which are observable in the data. [19]
Supervised learning is mainly divided in two categories:

- **REGRESSION**: these algorithms predict a continuous value. It is basically a statistical approach to find the possible relationships between variables and predict an outcome of an event based on those relationships.
- **CLASSIFICATION**: these algorithms classify the data into subsets (classes) of the database itself.

3.1 Regression

Regression models are used to predict target variables on a continuous scale, which makes them attractive for addressing many questions in science as well as applications as relationships between variables, evaluating trends, making forecasts or, like in this thesis, healthcare application [17].
From the known values $y(n)/n = 1, ...N$ we go back to $x(n)$ that can be used later to predict $y(n), n > N$ using $x(n), n > N$.
$x(n)$ is the independent variable while $y(n)$ is the dependent one. The relationship between those two is unknown and we can define them as:

- $y(n)$ **regressand**: the continuous value that has to be predicted
- $x(n)$ **regressors**: all the features that are used to predict the regressand

"When we regress Y on X, we use the values of variable X to predict those of Y" [20].

In order to evaluate the results of the prediction and to evaluate the model performances the **Coefficient of Determination** is computed: it can be indicated as $\boldsymbol{R^2}$.
It provides a measure of how well future samples are likely to be predicted by the regression model [1]. Best possible score is 1 and the final result can also be negative (because the regression model can be arbitrarily worse). So if R^2 is the unity, all variation has been explained and there is a perfect fit. If the coefficient is zero, the regression does not explain anything and the prediction is bad [21]. The definition of the Coefficient of Determination is:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1}(y_i - \bar{y})^2} \tag{3.1}$$

where:

- $\hat{y}_i$ is the predicted value of the i-th sample

[1] It is often expressed as a percentage by multiplying by 100

- y_i is the corresponding true value

- $\bar{y} = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i)$

3.2 Classification

Classification is the problem of identifying to which set of the database an observation belongs, on the basis of a training set of the database itself composed by features and containing observations whose category membership is known.

The classification can be binary or with multiple classes (this is going to be the case addressed in this thesis). Multi-class classification is more difficult from a computational point of view and usually does not give better results compared to the binary one and in literature is possible to find solutions to reduce multi-class classification to binary [22] [23]. In this thesis the classes must be considered as more than two because the binary classification is completely useless for the planned goals: in the healthcare field the binary classification is usually done in order to identify healthy and sick patients, which is not the case treated here as explained in the following subsection.
Each measurements, starting from the the BMI, is divided in classes based on the distribution of the data. Each parameter is divided in 3 (Low, Medium, High) and 5 (Low, Medium-Low, Medium, Medium-High, High) classes (details is section: 3.2.1).

In order to evaluate the results of the classifications the Accuracy Index is used: it is also called **Jaccard Index (JI)** [24]. JI measures similarity between finite sets: it is defined as the size of the intersection divided by the size of the union of the sets. It's a measure of similarity for the two sets of data, with a range from 0% to 100% (if multiplied by 100). The higher the percentage, the more similar the two sets.

$$JI(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.2}$$
$$\text{where } 0 \leq JI(A, B) \leq 1$$

3.2.1 Measurements Classification

As already mentioned previously, to proceed with multi-class classification all the measurements has to be manually divided in 3 (Low, Medium, High) and 5 (Low, Medium-Low, Medium, Medium-High, High) classes following the distribution of the data themself. The three-class classification is provided by the AGES database itself, while the 5-class classification was performed based on the informations obtainable from the classification already provided.
All the classes for BMI, CHOL, TUG, NGait and ISO are shown in fig. 3.1, 3.2, 3.3, 3.4, 3.5 and tables 3.1, 3.2, 3.3, 3.4, 3.5.
For FGait the classification is not used, because it is useless from a purely biomedical point of view: we only try to classify NGait, which is more indicative for elderly patients.

3 Classes	5 Classes
	15,65 → 25
< 25	20 → 25
25 → 30	25 → 30
> 30	30 → 35
	35 → 47

Table 3.1: **BMI** - Class division

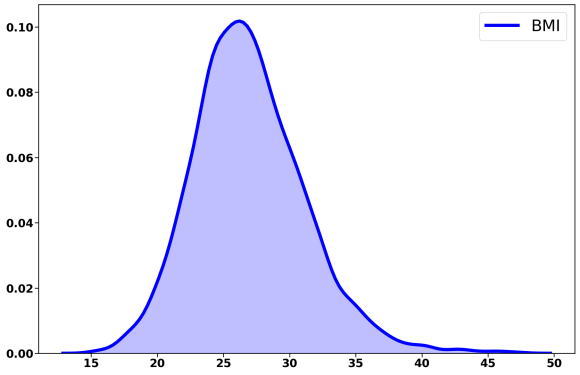

Figure 3.1: **BMI** distribution

3 Classes	5 Classes
	2,30 → 4
< 5	4 → 5
5 → 6	5 → 6
> 6	6 → 8
	8 → 9,69

Table 3.2: **CHOL** - Class division

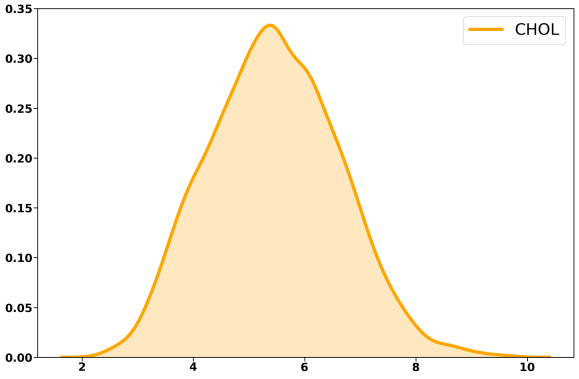

Figure 3.2: **CHOL** distribution

3 Classes	5 Classes
	5,15 → 7,5
< 10	7,5 → 10
10 → 12	10 → 12
> 12	12 → 17
	17 → 37,33

Table 3.3: **TUG** - Class division

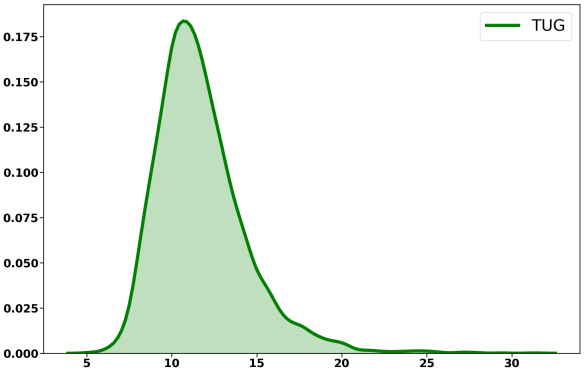

Figure 3.3: **TUG** distribution

3 Classes	5 Classes
	3,46 → 4,5
< 5, 5	4,5 → 5,5
5,5 → 6,5	5,5 → 6,5
> 6, 5	6,5 → 10
	10 → 31,19

Table 3.4: **NGait** - Class division

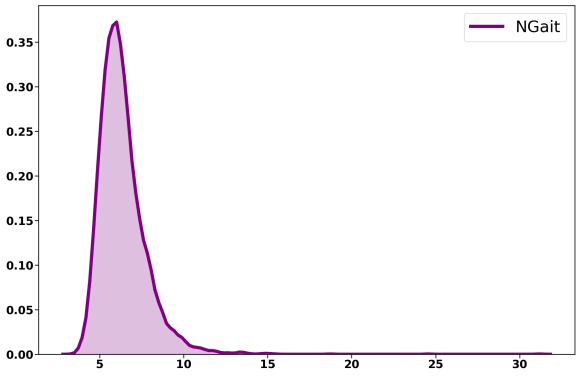

Figure 3.4: **NGait** distribution

3 Classes	5 Classes
	34,5 → 200
< 275	200 → 275
275 → 375	275 → 375
> 375	375 → 500
	500 → 781,2

Table 3.5: **ISO** - Class division

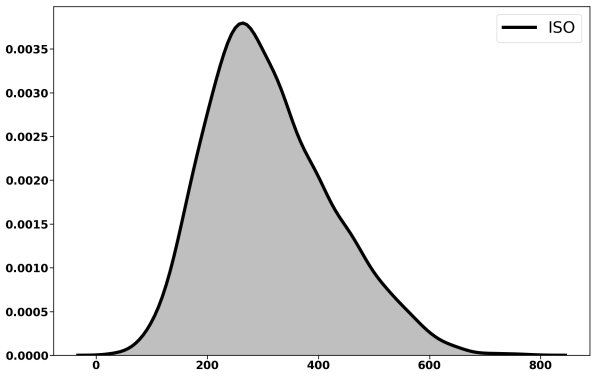

Figure 3.5: **ISO** distribution

3.3 Scikit-Learn library of Python

Scikit-Learn (SL) is the free ML library for the Python programming language used in this master thesis. It was created and ideated by David Cournapeau in 2007 and it features various classification, regression and clustering algorithms. It is designed to interoperate with NumPy and SciPy, the python numerical and scientific libraries [25]. Scikit-Learn 0.20 is the latest version to support Python 2.7.

In the following sections the algorithms used are going to be described in details. All of them can be found on the SL library with all the default values consulted on [25].

3.4 Tree Based ML Algorithms

In literature lots of supervised ML algorithms are present, more or less effective depending on the use and applications in healthcare [1]. In this thesis 4 **Tree Based Algorithms** are considered both for regression and classification analysis. They are denominated Tree Based because the basic units on which they are built are regression and classification trees **(Decision Trees)** [26]. Decision Trees are non-parametric supervised learning methods used for classification and regression. The main goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision tree is a greedy algorithm that performs a recursive binary partitioning of the feature space. The tree predicts the same label for each leaf partition. Each partition is chosen greedily by selecting the best split from a set of possible splits, in order to maximize the information gain at a tree node. The main advantages of the Decision Trees are:

- Requires little data preparation, but does not support missing values (do not require normalization or standardization of the initial data)
- The cost of using the tree is logarithmic in the number of data used to train the tree
- The feature importance is easy to extract
- Simple to understand and to interpret. Trees can be visualised (not in this case because they are so much big and deep)

In the following subsection the 4 Tree Based algorithms (**Random Forest (RF), Extremely Randomized Tree (EX-T), ADA Boosting (ADA-B), Gradient Tree Boosting (GRAD-B)**) used in this thesis are described and all their default parameters are listed in [25]. The most important parameters are the following:

- **n_estimators**: The number of trees in the forest. The larger the better, but also the longer it will take to compute. It is expected that the results will stop getting significantly better beyond a critical number of trees.

- **max_features**: The size of the random subsets of features to consider when splitting a node. The lower it is, the greater the reduction of variance, but also the greater the increase in bias.
- **max_depth**: The maximum depth of the decision trees. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- **min_samples_split**: The minimum number of samples required in a node to be considered for splitting.
- **min_samples_leaf**: The minimum number of samples required at each leaf node.

3.4.1 Random Forest

Random Forests (RF) are ensambles of Decision Trees [27] [28]. They share their same basic properties and capabilities, and, moreover, the trees combination is helpful to reduce over-fitting. The training of the set of used decision trees is done separately so that it can be executed in parallel with the others, but some randomness is injected in the training process to reduce the variance of the predictions. Randomness is injected by subsampling the original dataset on each iteration to get a different training set or considering different random subsets of features to split on at each tree node. To make a prediction on a new instance, a random forest must aggregate the predictions from its set of decision trees. In the case of classification, the aggregation is done by majority vote. Each prediction is counted as a vote for one class. The label is predicted to be the class which receives the most votes. In the case of regression the mean of the obtained predictions of the individual trees is evaluated as output [25].

3.4.2 Extremely Randomized Tree

In extremely randomized trees (EX-T), randomness goes one step further in the way splits are computed [29]. A random subset of candidate features is used exactly like in Random Forest, but, instead of looking for the most discriminative thresholds, thresholds are drawn randomly for each possible feature and the best of these random thresholds is picked as the splitting rule. The random selection of the threshold allows to further reduce the variance of the model [25].

3.4.3 ADA Boosting

AdaBoost (ADA-B) is an ensemble method that belongs to the boosting family [30] [31]. The principle of the boosting algorithms is to convert weak learners (classifiers or regressors with accuracy just above the random guessing) into strong learners that predict with high accuracy. The AdaBoost training selects only the features that improve the predictive power of the model, reducing complexity in terms of dimension and improving execution time as the not relevant features do not need to be processed. As weak learner the Decision Tree is often used, as done in this thesis. The data modifications at each boosting iteration consist of applying weights $w_1, w_2 ... w_N$ to every training samples. Initially, those weights are all set to $w_i = 1/N$, so that the first step of the algorithms trains a weak learner on the original training data. For all the other successive iterations, the sample weights are modified and the learning algorithm is applied again to the data with the new weight. At a given step, those training examples that were wrongly predicted by the boosted model at the previous step have their weights increased, whereas the weights are decreased for those examples which were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is then forced to concentrate on the examples that are missed by the previous ones [25].

3.4.4 Gradient Tree Boosting

Gradient Tree Boosting (GRAD-B) is a generalization of boosting to arbitrary differentiable loss functions [32] [33]. The major difference between AdaBoost and Gradient Boosting Algorithm is how each of them identifies the lacks of weak learners. While the ADA-B model identifies the lacks by using high weight points, GRAD-B performs the same by using gradients in the loss function which is a measure indicating how good the coefficients of the model are at fitting the data. It supports a number of different loss functions: the default one for regression is ”least squares”, others are for example ”least absolute deviation” and ”huber” which is a combination of the previous ones [25]. The default least squares is used in this thesis. The advantages of GRAD-B are that it can handle heterogeneous features and is robust to outliers in output space, while the disadvantage is that it is hard to scale due to its sequential nature [34].

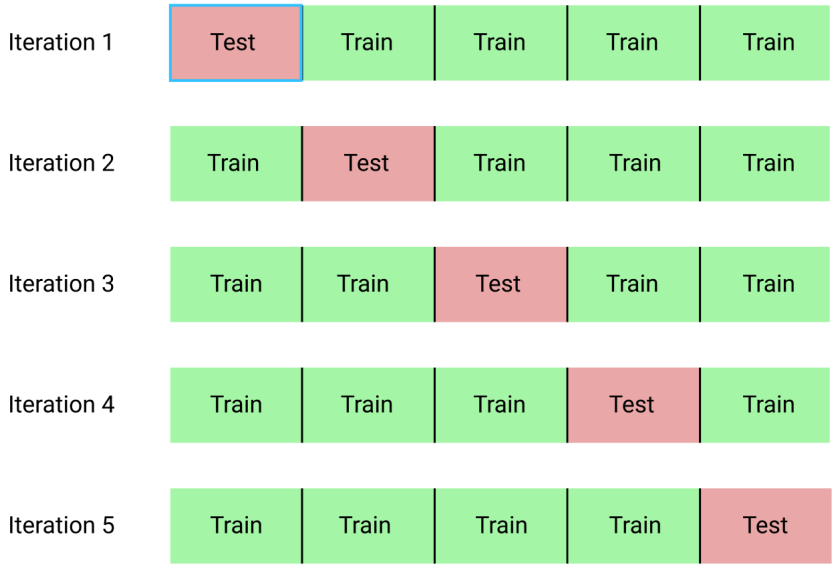

Figure 3.6: K-Fold division [35]

3.5 Train - Test Division

One of the main steps of the ML algorithms is the division of the dataset between Train and Test. The training set is made of data used for learning while the testing set is used to evaluate the quality of the regressor or of the classifier. Usually the training set is between 70% and 80% of a random selection of the data while, at the opposite, the testing part is usually between 30% and 20%. A training set too small does not allow the algorithm to learn enough from the data while a testing set too small does not allow to verify with confidence the success of the prediction, so these percentages are a good compromise. However, by partitioning the available data into training and testing, the number of samples which can be used for learning the model is drastically reduced, and the results can depend on a single particular random choice for the pair of sets. One possible solution to this problem is the Cross Validation. There are different types of Cross Validation Techniques but the overall concept remains the same: they firstly partition the data in a number of subsets, then hold out a set, used for test, and the model is trained on the remaining sets [35].

3.5.1 Cross Validation - K-Fold Division

The K-Fold Division is the most common way to apply Cross Validation. This procedure uses a single parameter *k* that is the number of splits applied to the entire dataset. As such, the procedure is called k-fold cross-validation. If, for example, like in fig. 3.6, k=5, the dataset is divided into 5 equal splits and the process will run 5 times, each time with a different holdout set. The selected group is used for the test and the remaining ones to fit the model, so iteratively the complete dataset is trained and every single data is used also for the testing. The value for k is chosen such that each train/test group of data samples is large enough to be statistically representative. In this thesis k was chosen equal to 8,12,16 and 18. The results obtained with k=8,12 are statistically more significant. With k=16,18 the results can, usually, be better because the training set is bigger but they can not be considered too much significant from a statistical point of view because the test set would be too little. Cases 16 and 18 are in any case considered for a greater overview of all the possible results [36].

Chapter 4

Predictive System

4.1 Organization of the starting system - BMI

The first parameter considered for the regression and classification is the BMI, as, hypothetically, it could give good results, being intuitively very probable that using features related to muscle and fat, we can predict the weight to square height ratio, especially in elderly people.

Considering regression, at first, the 11 NTRA features are taken in consideration as regressors. In order to reach the best R^2, a lot of combinations have been tried using the 3 databases, the 4 different k-fold divisions and the 4 Tree Based ML algorithms (fig. 4.1).
For example, using the NTRA features on the AGES I+II database with a k-fold division of 8 sets using the ADA-B algorithm, 8 different R^2 are obtained, one for each test set. The same methodology is applied also for the classification obtaining a JI as result from all the possible combinations.

All the following results are obtained using as Database AGES I+II: employing only 1 database, the number of patients is about the half, and the results of R^2 and JI are worse than the ones obtained using AGES I+II. The only exception is with the prediction of ISO for which the best database results AGES I.

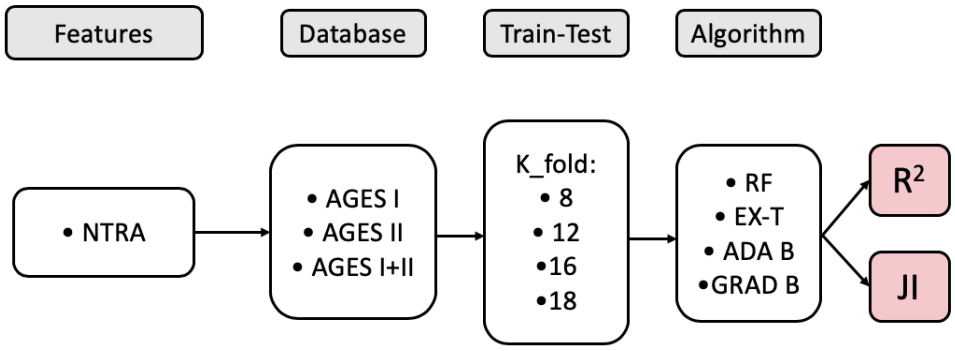

Figure 4.1: Basic Prediction System for BMI

4.1.1 BMI - Regression - features=NTRA

This section presents the possible representation of the results using as example R^2 results for the BMI using only the 11 NTRA as regressors.
All comments and considerations regarding the results themselves will be further discussed in the next chapter. The graphs used will be the same for all the following results.
The same representations can be done for the classification with 3 and 5 classes considering JI instead of R^2.

For all the types of representations each algorithm has a color: blue for RF, orange for EX-T, green for ADA-B and red for GRAD-B. Fg. 4.2 shows the value of R^2 for each test set with K=1,2,...,k (in this specific case k=16) with the respective mean value which is visible also in the histogram in fig. 4.3 together with the minimum and the maximum value obtained for each algorithm. In fig. 4.4 the distribution of the data printed in fig. 4.2 is represented through a Violin Plot in which the dots are the R^2 values obtained for each fold.

4.1.2 Feature importance

For each prediction it is possible to extract the feature importance (see fig. 4.5): it gives a score for each feature of your data, the higher the score more important or relevant is

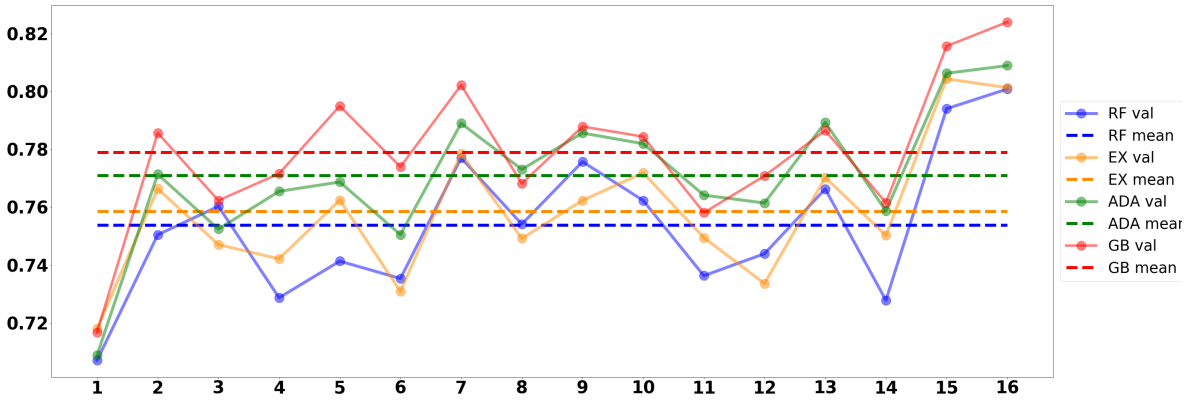

Figure 4.2: R^2 BMI - Kfold=16 - features=NTRA

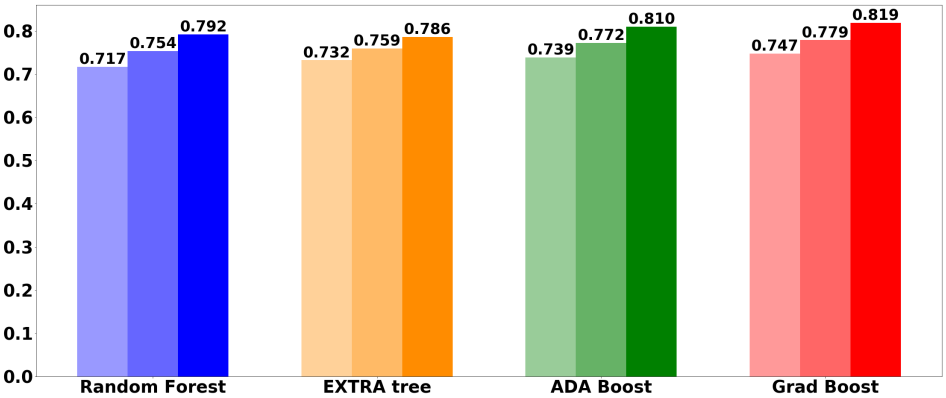

Figure 4.3: R^2 Min-Mean-Max BMI - Kfold=8 - features=NTRA

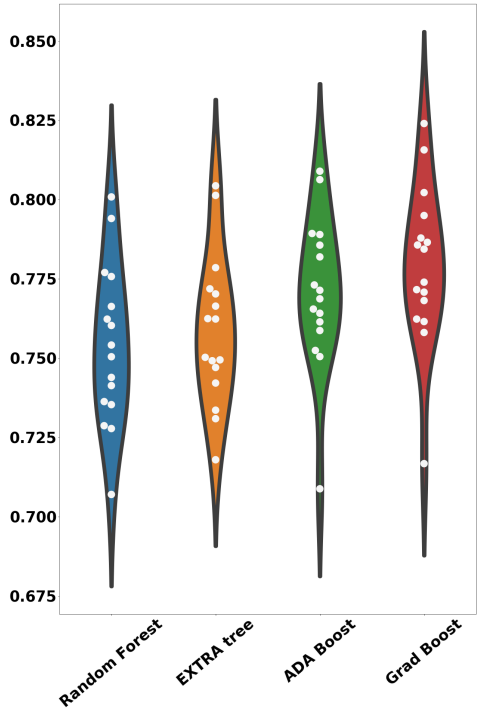

Figure 4.4: R^2 distribution BMI - Kfold=16 - features=NTRA

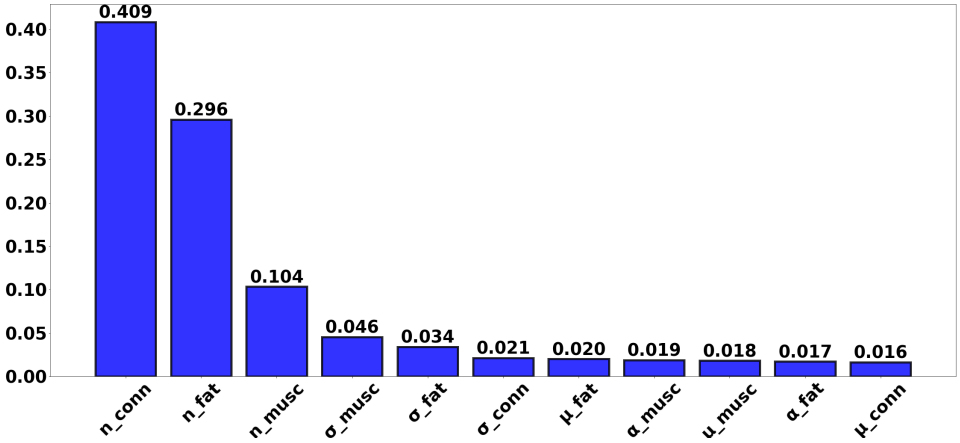

Figure 4.5: RF - Feature importance BMI - K_fold=8 - features=NTRA

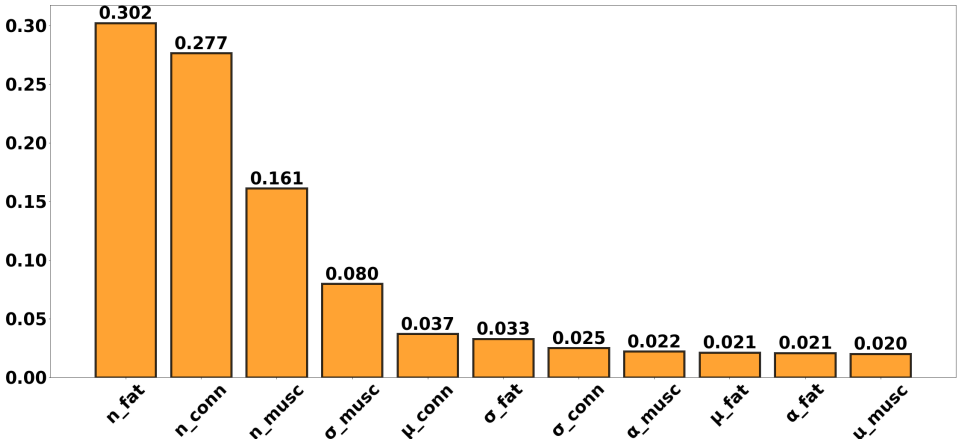

Figure 4.6: EX-T - Feature importance BMI - K_fold=8-features=NTRA

the feature towards your output variable.
The relative rank of a single feature used as a node in a decision tree can be used to assess the relative importance of that particular feature with respect to the predictability of the target parameter. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to, can be used as an estimate of the relative importance of the features [37]. In practice those estimates are stored as an attribute in scikit named $feature\_importances$ on the model. This is an array with shape N=number of features, whose values are positive and sum to 1.0 [25]. Fig. 4.5, 4.6, 4.7, 4.8 show the feature importances of the BMI Regression using only the 11 NTRA parameters as regressors with k-fold=8. There is one graph for each of the 4 algorithms. Also in this case all comments and considerations regarding the results themselves will be discussed in the next chapter and extended also to other parameters in addition to BMI.
The analysis of the feature importance is useful only when the values of R^2 are sufficiently high: if R^2 is less than 0.40 or even negative, this extraction is useless as it does not add any useful information.

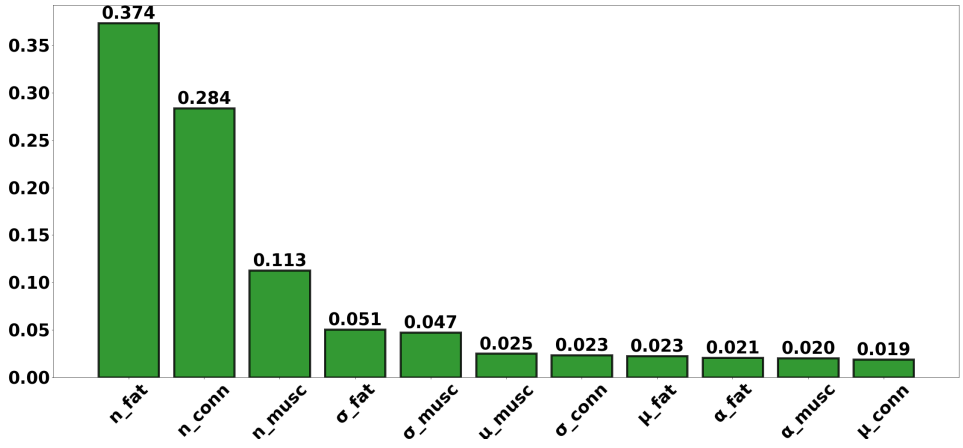

Figure 4.7: ADA-B - Feature importance BMI - K_fold=8-features=NTRA

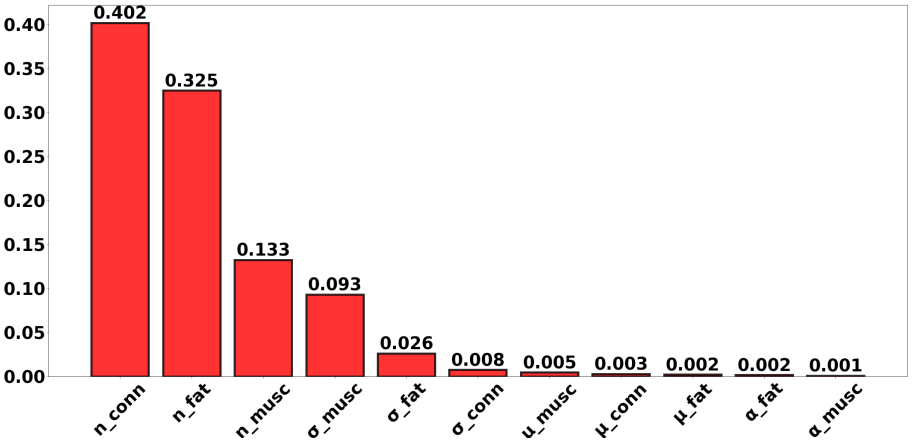

Figure 4.8: GRAD-B - Feature importance BMI - K_fold=8-features=NTRA

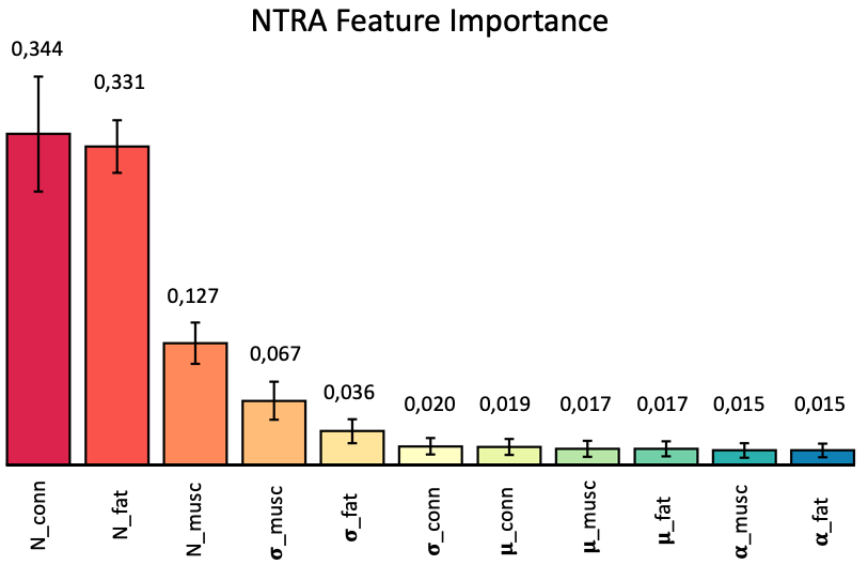

Figure 4.9: BMI prediction - Feature Importance mean (with relative std) from 0 to 1 using only NTRA as regressors.

4.2 Evolution of the system

The approach used to predict BMI just described is extended also to other parameters both for regression and classification. The measurements considered for the prediction are CHOL, TUG, NGait and ISO. The scheme of fig. 4.1 described for BMI is the same also for these other parameters.

In order to improve the system and the possible results both in terms of classification and regression, new sets of initial features can be created starting from the 11 NTRA parameters.
Obviously this further selection of features has been carried out in the case in which the results obtained using only the 11 NTRA are acceptable and not negative. Anticipating the results that will be described in the next chapter, this section describes the feature selections to predict BMI and ISO. For CHOL, TUG and NGait, since the results were negative with only 11 features, no further investigation was carried out.

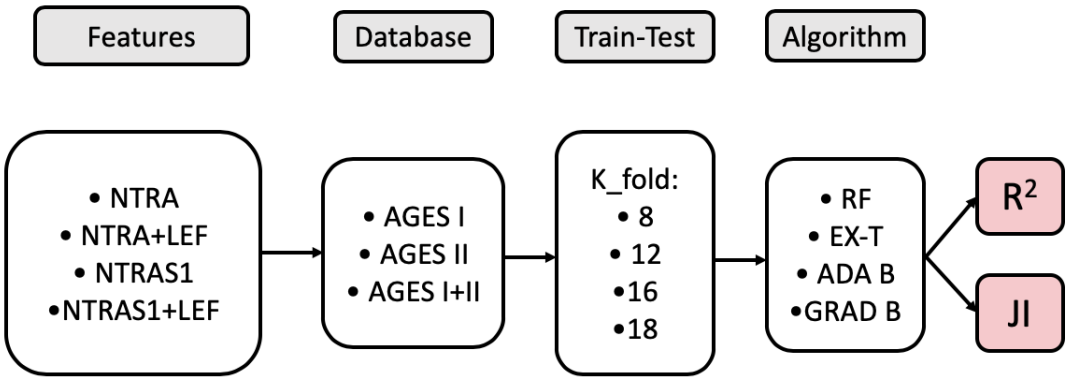

Figure 4.10: Prediction scheme for BMI with all the feature selections

4.2.1 Features management - BMI

The features selections used for BMI are based on the results of fig. 4.9 which shows the feature importance mean using only the 11 NTRA parameters. The mean is obtained with the combination of all k-fold divisions (8-12-16-18) and all the 4 algorithms applied on AGES I+II: practically all the R^2 obtained from the scheme of fig. 4.1.

The different selections of features used in addition to the 11 NTRA, are the following:

- NTRA + LEF= 11 NTRA, CHOL and LEF measurements (TUG, ISO, Ngait and Fgait),
- NTRAS1= a selection of the 11 NTRA based on the results of the mean feature importance of fig. 4.9 (All the N and the μ),
- NTRAS1 + LEF= NTRAS1, CHOL and LEF measurements.

Fig. 4.10 is illustrates the prediction scheme used for both regression and classification for the BMI with all the several possible combinations.

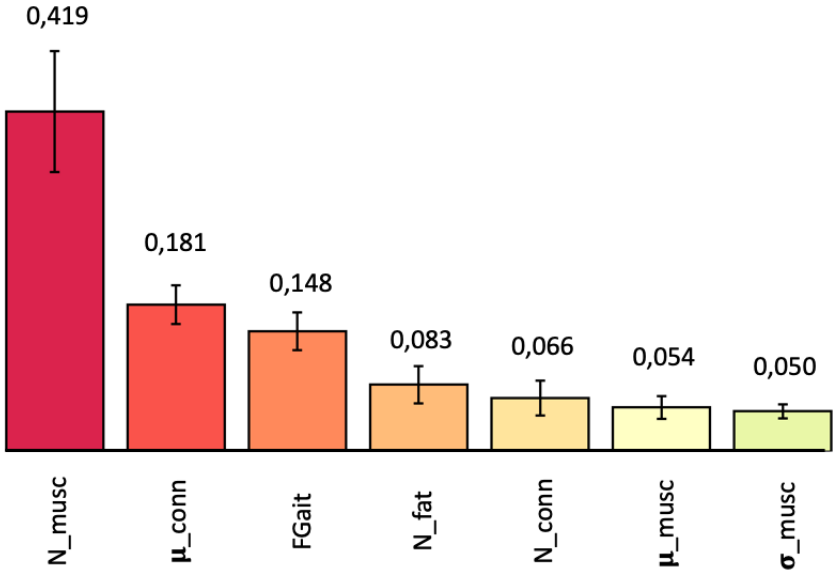

Figure 4.11: ISO prediction - Feature Importance mean (with relative std) from 0 to 1 using only NTRAS2 and FGait as regressors.

4.2.2 Features management - ISO

As already mentioned, the prediction results for ISO are satisfactory enough to deserve further analysis on the feature selection. After numerous combinations attempts and observing the results of the feature importance obtained with the 11 NTRA paramenters, it can be deduced that among the LEF measurements only Fgait contributes to improve the results and only if combined with a selection of the 11 NTRA (fig. 4.11). Therefore, to avoid an excessive accumulation of useless results, only 2 further feature selection are considered:

- NTRAS2= a selection of the 11 NTRA based on the results of the mean feature importance of fig. 4.11 (All the N, μ_musc, μ_conn and σ_musc),
- NTRAS2 + Fgait

In fig. 4.12 is illustrated the prediction scheme used for both regression and classification for the ISO with all the several possible combinations combining the features

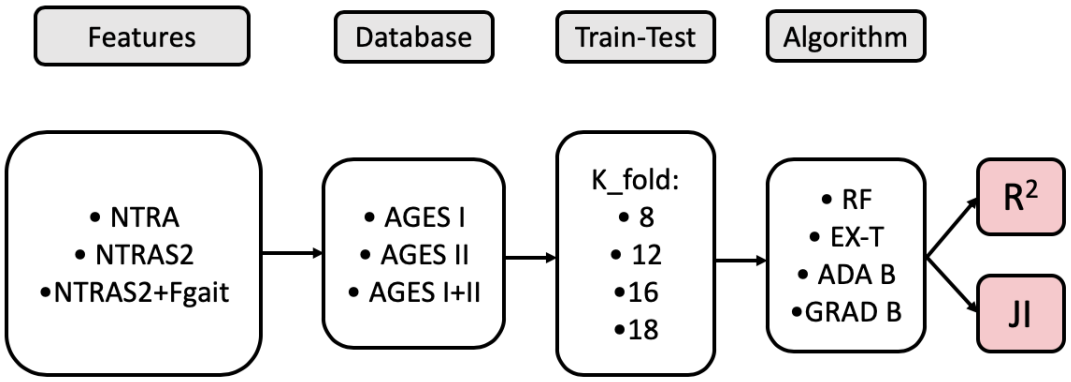

Figure 4.12: Prediction scheme for ISO with all the feature selections

selections, the databases, all the 4 k-fold divisions and the 4 tree based algorithms. As already mentioned previously the best results, which will be explained in detail in the next chapter, are obtained using only the AGES I database, so with a smaller number of patients.

Chapter 5

Regression and Classification Results

In this chapter all the results of the predictions are presented: greater emphasis will be given to the positive ones, but the negative ones will not be excluded. The negative results have no significant value in terms of prediction, but they allow us to understand that there is no relationship between the parameter itself and the starting 11 NTRA features of the mid-femur CT scan.
The best results are achieved with the test parameter BMI and with ISO while the R^2 and JI values for CHOL, TUG and NGait are quite negative.
Regarding classification, as it is conceivable [22] [23], the 3 classes prediction gives always better results than the 5 classes prediction. If the R^2 values for regression are good, also the JI values are satisfactory, and vice-versa.
From an algorithmic point of view GRAD-B gives always the best R^2 and JI in any occasion: even if the results are negative, the best ones are from GRAD-B. ADA-B, RF and EX-T rarely give better results than GRAD-B. They work better or worse depending on the parameter taken into account for the prediction. In general, the ADA-B algorithm is the less effective for classification.
For what concerns the number of k-folds the best values of R^2 and JI are achieved with k=16,18 if we consider the maximum, while in general the mean value is higher with k=8,12 and the latter also give a more meaningful result from the statistical point of view as the size of the training and testing set is neither excessively large nor excessively small.

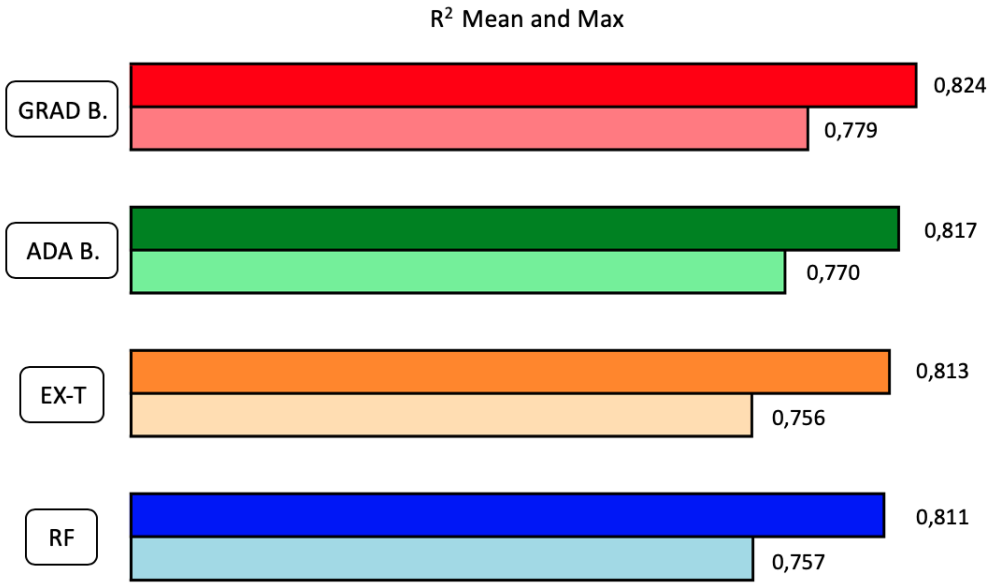

Figure 5.1: BMI - Mean and Max values of R2 for the 4 algorithms (with default values) obtained combining all feature selections and all the k-fold divisions.

5.1 BMI

In this section the results related to the BMI regression and classification are discussed: some of them were already anticipated in the previous chapter to explain the prediction system.
All the BMI results are obtained using as Database AGES I+II.

5.1.1 BMI - Regression

Fig. 5.1 shows the mean and the max R^2 obtained for each of the 4 algorithms with all the combination of features selections and k-fold divisions. Those are obtained using the default values of the functions in SL [25]. Clearly, GRAD-B gives the best results followed by the ADA-B algorithm. The others have comparable results.
The maximum obtained without any modification of the default values is R^2=0,824 (fig. 5.1) using the 11 NTRA parameters as regressors, k-fold=16 and GRAD-B algorithm. Modifying the default value of this combination setting n_estimators=200 the highest possible R^2 of 0,831 can be achieved.

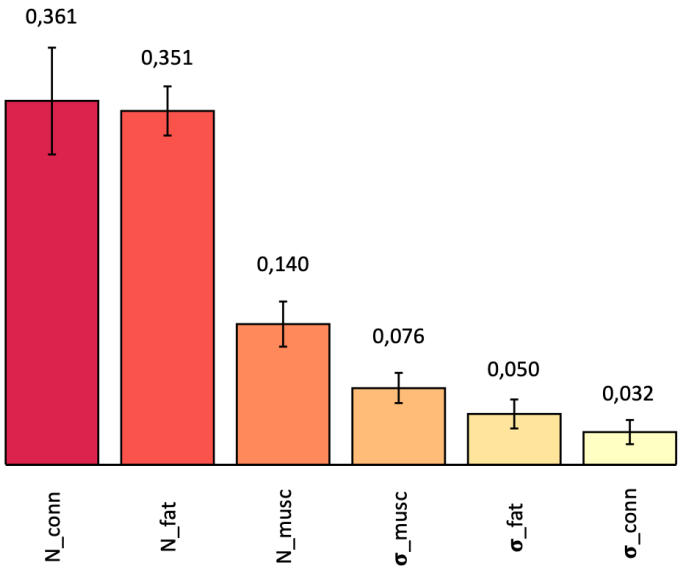

Figure 5.2: BMI prediction - Feature Importance mean (with relative std) from 0 to 1 using NTRAS1 as regressors.

Using the NTRA selection of features, the R^2 results are higher: selecting only some of the 11 NTRA parameters (NTRAS1) the prediction does not improve. The same happens if LEF and Chol are added to NTRA or NTRAS1.
The most important features are N_conn and N_fat. They in each possible combination always cover more than 50% of the total importance. The importance of Chol and LEF measurements is always less than 0,016 (from 0 to 1). For GRAD-B and RF the most important feature is N_conn while for EX-T and ADA-B the most relevant is N_fat (fig. 4.5, 4.6, 4.7, 4.8) . The high importance of a connective tissue parameter is an unexpected result: the muscles and fat are usually analyzed in a CT scan and little or even no importance is given to the connective tissue. These results that link the amplitude of the connective tissue extracted from a CT scan to the BMI, in such an important way, suggest that even for future further applications, such as comorbidity classification, the connective tissue may have a great importance in predictive process. DM and DM2 or vascular and cardiac diseases like CHD, CVD, CHF can be considered as comorbidities.

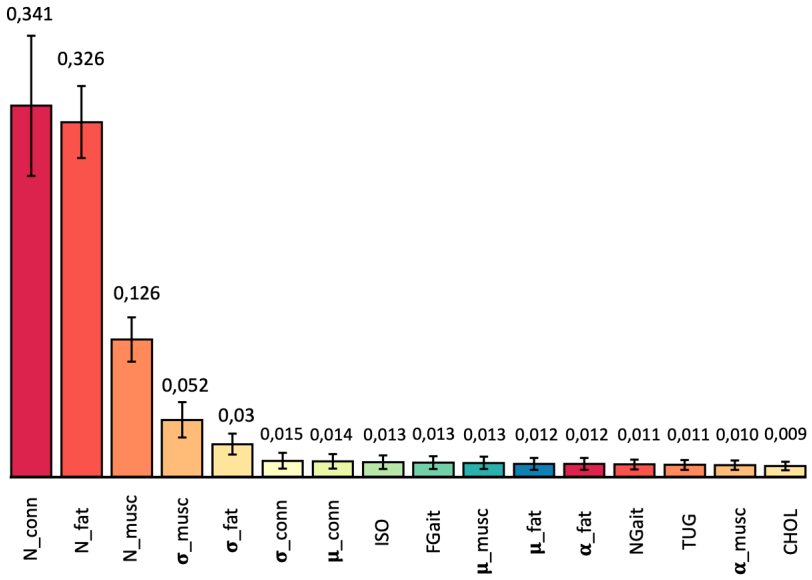

Figure 5.3: BMI prediction - Feature Importance mean (with relative std) from 0 to 1 using NTRA + LEF as regressors.

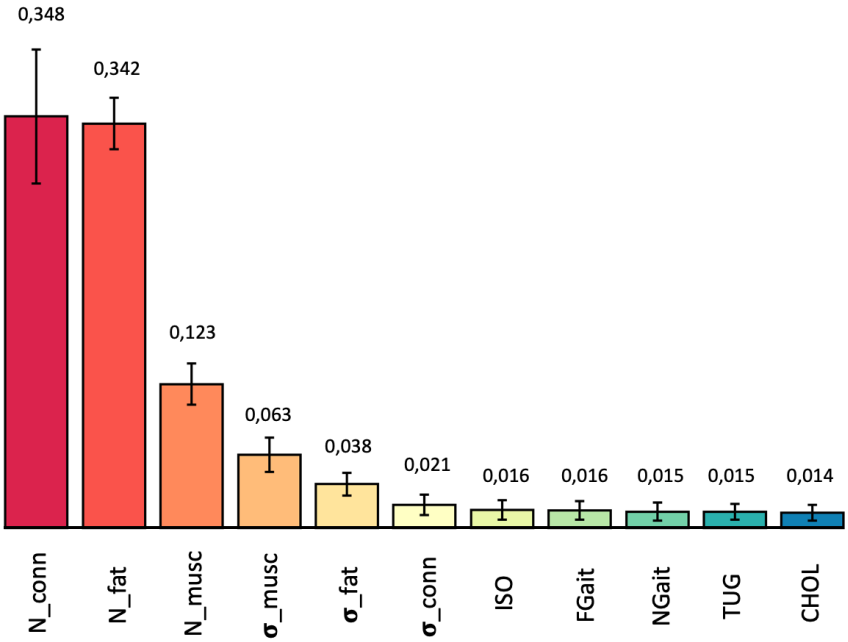

Figure 5.4: BMI prediction - Feature Importance mean (with relative std) from 0 to 1 using NTRAS1 + LEF as regressors.

ML Algorithm	JI (3 Classes) • Mean • Max	JI (5 Classes) • Mean • Max
Random Forest	• 0,701 • 0,756	• 0,642 • 0,712
EXTRA Tree	• 0,696 • 0,766	• 0,638 • 0,715
ADA Boosting	• 0,631 • 0,714	• 0,565 • 0,618
Gradient Boosting	• 0,732 • 0,797	• 0,679 • 0,741

Table 5.1: BMI Classification - JI Mean and Max for the 4 algorithms obtained combining all feature selections and all the k-fold divisions

In fig. 4.9, 5.2, 5.3, ,5.4, it is possible to see what explained above: the mean of the feature importances using all the k-fold divisions and all the 4 algorithms for the different selections of initial features has always as the most important ones N_conn and N_fat while CHOL and LEF parameters are not useful for the prediction.
The amplitude N and the width σ for fat, muscle and connective tissue are always at the first 6 positions while the skewness α in particular does not contribute significantly to the prediction.

5.1.2 BMI - Classification

Table 5.1 shows the results of the classification using JI as an indicator of the accuracy. The results, as expected, are better with 3 classes, but the ones with 5 classes are not far from the previous ones, so both classifications can be considered reliable. The maximum JI (0,797) is obtained using GRAD-B with 3 classes, combined with NTRAS1 and k-fold=18. The maximum JI with 5 classes (0,741) is obtained with the

same combination. From the results of table 5.1 is also possible to understand that also for classification GRAD-B is the best algorithm while ADA-B does not work in a good way.

In the violin plots of fig. 5.5 is possible to see the distribution on the values of JI with k-fold=8 combining all the feature selections for the 3 and 5 classes classification. In the 3 classes classification some folds obtain a very low JI and this causes to lower the final average value. In the 5 classes classification distribution it is easy to notice the big difference between ADA and the other predictive algorithms.

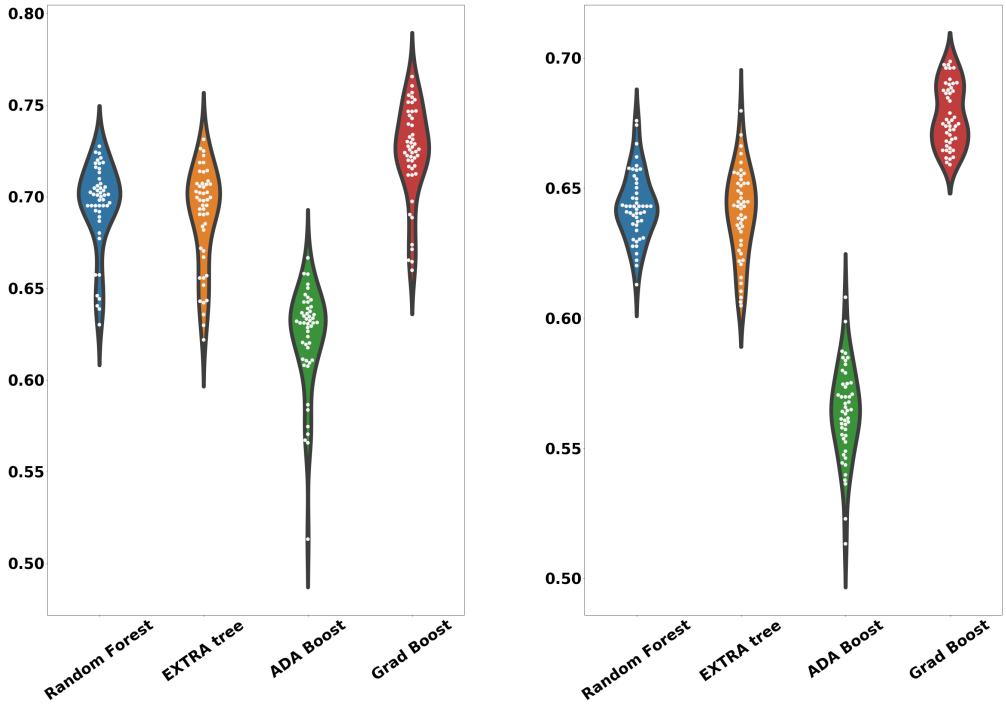

Figure 5.5: Violin Plot for JI distribution in the 3 (left), 5 (right) classes classification of BMI with K-fold=8

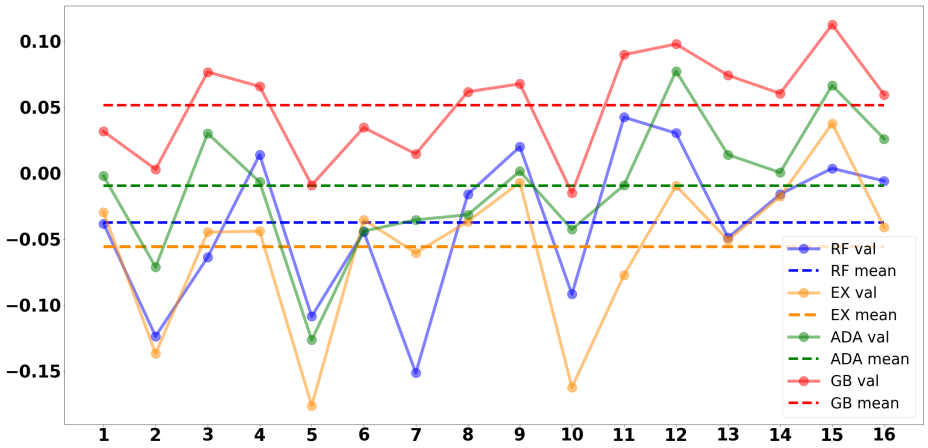

Figure 5.6: R^2 CHOL - Kfold=16 - features=NTRA

5.2 LEF: CHOL, TUG & NGait

In this section the results related to the LEF parameters, in particular CHOL, TUG and NGait, both for regression and classification, are discussed. No prediction analysis was done for FGait: applying the prediction algorithms already to the NGait it is useless to repeat it for a very similar parameter, which is also not significant for patients as old as those present in AGES.
Next section is dedicated only to the ISO as its results are so much better than those presented here. For these three measurements the prediction results both for regression and for classification are not particularly positive: this, however, allows to deduce that the parameters of muscles, fat and connective tissue of a mid-femur CT scan are not in any way predictive for these analyzed data.
The features used for the prediction are always only the 11 NTRA parameters: no feature selection has been applied due to the bad results.

5.2.1 LEF - Regression: CHOL, TUG & NGait

In fig. 5.6, 5.7, 5.8 it is possible to see the values of R^2 for CHOL, TUG and NGait with k-fold=16. All of them are really low, even below zero (this is possible due to the

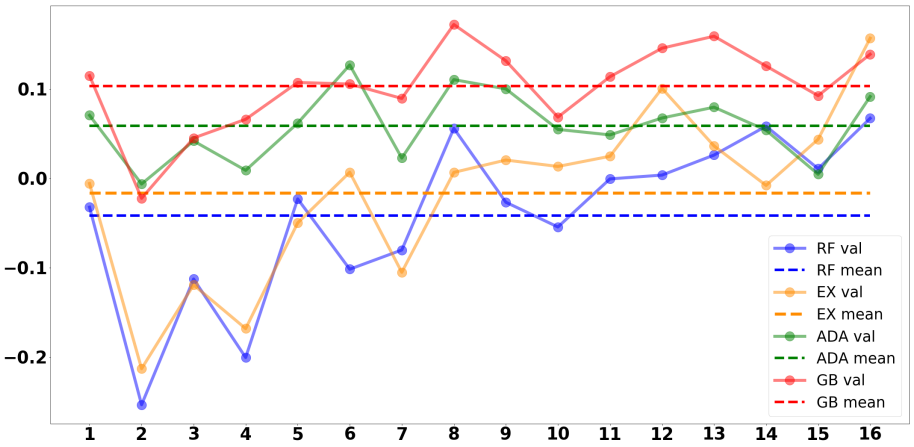

Figure 5.7: R^2 TUG - Kfold=16 - features=NTRA

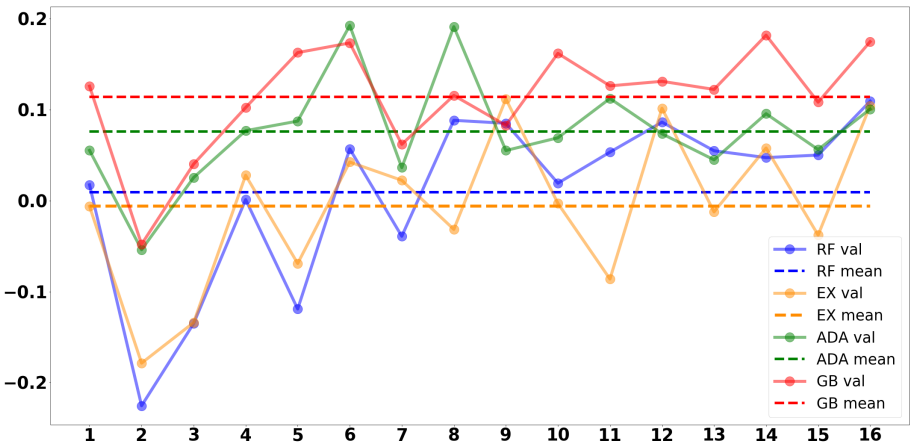

Figure 5.8: R^2 NGait - Kfold=16 - features=NTRA

definition of R^2 itself [21]). They do not change with other values of k in the k-fold division: the mean R^2 is never more than 0.1. In this case try to use a selection of the 11 NTRA or add other parameters as regressor is completely useless as well as the extraction of the feature importance.

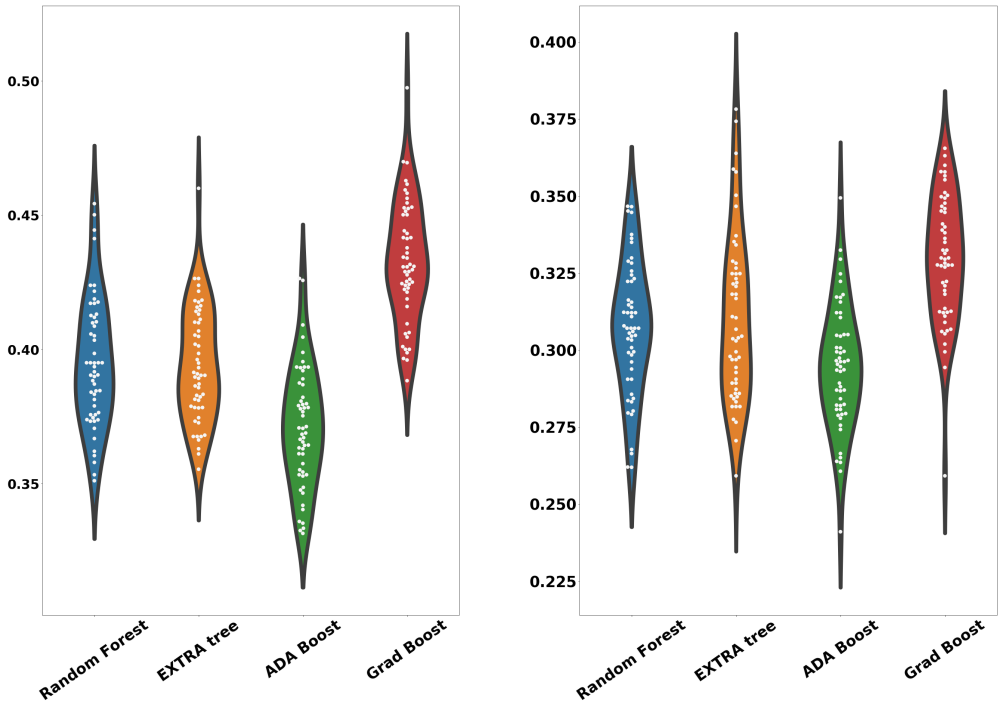

Figure 5.9: Violin Plot for JI distribution in the 3 (left), 5 (right) classes classification of CHOL with K-fold=16

5.2.2 LEF - Classification: CHOL, TUG & NGait

As it is possible to see in the violin plot of fig. 5.9 and in tables 5.2, 5.3, also the classification results are completely unsatisfactory: most of the JI values obtained do not exceed 0.5 and this means that less of half of the classes are predicted correctly. For the 5-class classification, no value of JI ever exceeds 0.5, reaching the lowest values even below 0.25.

ML Algorithm	CHOL (3 classes) • Mean • Max	TUG (3 classes) • Mean • Max	NGait (3 classes) • Mean • Max
Random Forest	• 0,392 • 0,454	• 0,421 • 0,467	0,449 0,506
EXTRA Tree	• 0,397 • 0,460	• 0,414 • 0,490	0,444 0,491
ADA Boosting	• 0,367 • 0,410	• 0,395 • 0,427	0,411 0,491
Gradient Boosting	• 0,432 • 0,467	• 0,464 • 0,508	0,493 0,563

Table 5.2: LEF 3 classes Classification - JI Mean and Max for the 4 algorithms obtained combining all feature selections and all the k-fold divisions

As in the BMI classification also for these three parameters the values in the prediction with 3 classes are greater than the ones achieved with 5 classes and the best algorithm is still GRAD-B while the worst is always ADA-B. In any case these comments, given the low JI values , are not significant: the results are simply negative for every type of possible combination.

5.3 ISO

In this section the results related to the ISO regression and classification are discussed. The R^2 and JI obtained can be considered satisfactory even if they do not reach the values achieved with the BMI: it can be said that there is a link, even if not extremely strong, between the values extracted from the CT scan and the strength of the leg in elderly people. Moreover, the most important features are different from those of the BMI, and this confirms the quality of the prediction.
All the following ISO results are reached through all the combination of the four k-fold

ML Algorithm	CHOL (5 classes) • Mean • Max	TUG (5 classes) • Mean • Max	NGait (5 classes) • Mean • Max
Random Forest	• 0,306 • 0,346	• 0,339 • 0,385	0,375 0,415
EXTRA Tree	• 0,306 • 0,378	• 0,333 • 0,396	0,372 0,418
ADA Boosting	• 0,294 • 0,349	• 0,306 • 0,381	0,330 0,371
Gradient Boosting	• 0,328 • 0,363	• 0,367 • 0,400	0,422 0,462

Table 5.3: LEF 5 classes Classification - JI Mean and Max for the 4 algorithms obtained combining all feature selections and all the k-fold divisions

divisions, the four algorithms and the three different feature selections as shown in fig. 4.12 considering as database AGES-I.

5.3.1 ISO - Regression

Fig. 5.10 presents the R^2 values using the 11 NTRA as initial features and k-fold=16: the maximum R^2 mean values is reached with GRAD-B and it is near to 0,55 with a maximum which exceeds 0,61.
Fig. 5.11 shows the maximum, the mean and the minimum R^2 obtained with all the k fold divisions and the three different features selections (NTRA, NTRAS2 and NTRAS2+FGait). The maximum mean values is still the one obtained with GRAD-B but the maximum value (0,614) comes from the EX-T algorithm combined with NTRAS2 + FGait and k-fold=16. With EX-T it is possible to reach the maximum but also the lowest value of R^2=0.305. The low values derive from the cross validation with k-fold=16,18: they contribute to lowering the general average which would be slightly higher if only k=8,12 were supposed to be used.

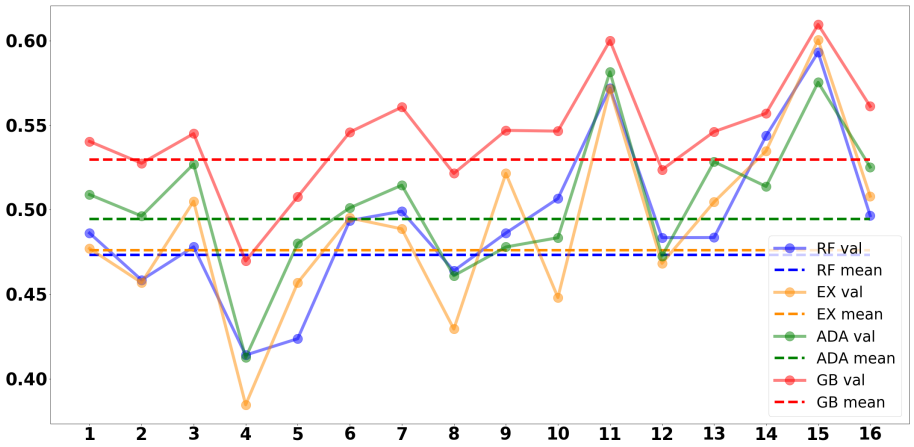

Figure 5.10: R^2 ISO - Kfold=16 - features=NTRA

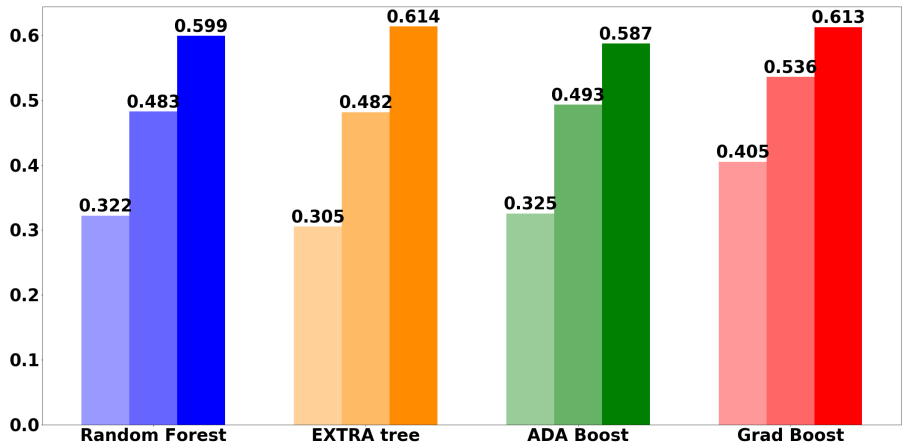

Figure 5.11: R^2 Min-Mean-Max - ISO - All K-fold and Features Selection combinations

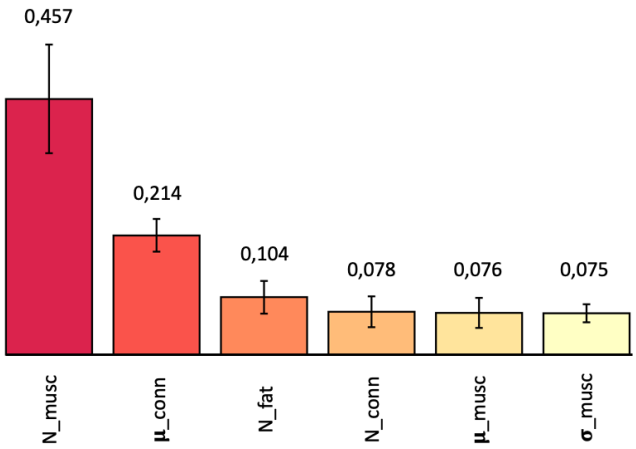

Figure 5.12: ISO prediction - Feature Importance mean (with relative std) from 0 to 1 using NTRAS2 as regressors.

As it is possible to see in fig. 4.11, 5.12, 5.13, N_musc covers almost the 50% of the feature importance while the connective tissue, expecially the μ value, has again a high importance in the prediction. This strengthens the results obtained with the BMI: connective tissue is again very relevant and it must be considered as one of the main prediction factors for the the eventual future applications. The order of importance of the features is always the same in the case of NTRA and NTRAS2 while in the case of adding FGait, which allows to reach the best R^2 values, it is the third most important feature. All the other measurements such as LEF and CHOL, following several experiments, do not affect the final results, only FGait allows improving the results if added to NTRAS2.
All the results achieved for ISO are achieved using the default values of the algorithms function of SL [25].

5.3.2 ISO - Classification

Table 5.4 shows the results of 3 and 5 classes classification about ISO. As for the BMI, the JI values are so much better with 3 classes, and, opposite to the BMI, the JI results for the 5 classes classification are absolutely not reliable and not comparable with the good ones obtained for BMI. This big difference between the 3 and 5 classes

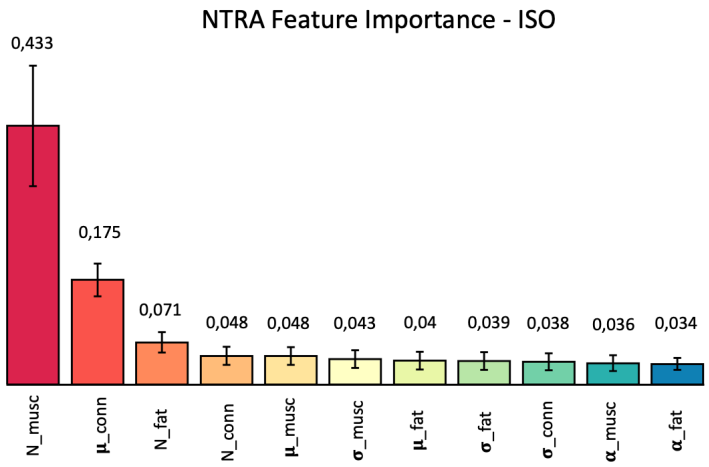

Figure 5.13: ISO prediction - Feature Importance mean (with relative std) from 0 to 1 using only NTRA as regressors.

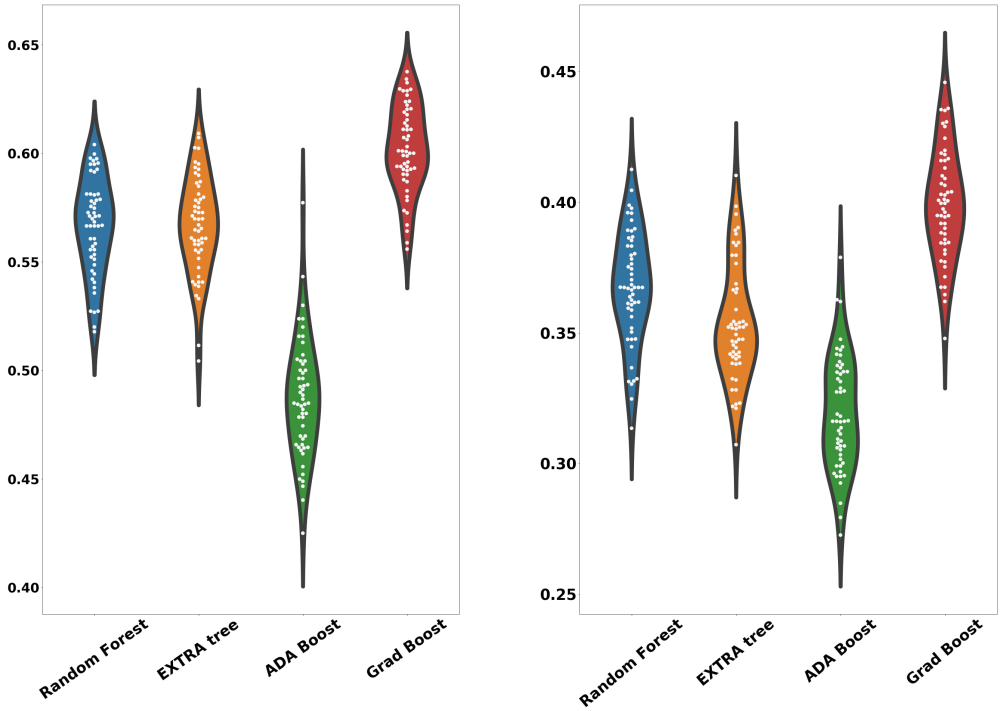

Figure 5.14: Violin Plot for JI distribution in the 3 (left), 5 (right) classes classification of ISO

ML Algorithm	JI (3 Classes) • Mean • Max	JI (5 Classes) • Mean • Max
Random Forest	• 0,566 • 0,604	• 0,366 • 0,412
EXTRA Tree	• 0,566 • 0,609	• 0,355 • 0,410
ADA Boosting	• 0,486 • 0,543	• 0,319 • 0,362
Gradient Boosting	• 0,603 • 0,637	• 0,400 • 0,445

Table 5.4: ISO Classification - JI Mean and Max for the 4 algorithms obtained combining all feature selections and all the k-fold divisions

is probably due to a wrong previous classification of the ISO itself, but a maximum for the 5 classes classification of JI=0,445 is even worse than the results obtained for CHOL, TUG and NGait. On the other hand JI results for the 3 classes classification are not good as the ones obtained with the BMI but at least good enough: the maximum of JI= 0,637 is obtained with GRAD-B (which is again the best of the 4 ML algorithms) combined with NTRAS2 + FGait and k-fold=18. ADA is again the worst of the 4 algorithms while RF and EX-T are comparable: in the case of classification, EX-T algorithm is not as good as for the regression.
Violin plots of fig. 5.14 shows the big difference between 3 and 5 classes classification on the values of the y-axis and that the worst results are reached with ADA algorithm.

Chapter 6

Conclusions and Possible Developments

The thesis work was focused on the prediction of some physiological measurements using as initial features 11 parameters called NTRA extracted from a Transaxial Mid-femur CT scan using the AGES-Reykjavik dataset of 6314 patients provided by Icelandic Hearth Association. This was the first investigation of this dataset through ML methodologies.
Usually the ML technologies applied to medical images work on the image itself (pixels) in order to create masks or to do segmentations: in this thesis, we try to link a CT scan to physiological measurements that are not relative to the image itself and which apparently may have not connections with a mid-femur CT scan.
The obtained results, especially for BMI, but also for ISO, allow us to say that the 11 NTRA parameters, combined sometimes with other measurements, can have a very significant predictive value. To confirm this we can also comment on the negative results that reinforce the predictive value of the 11 NTRA parameters: they are not predictive for all the parameters but only for some of them, in fact with the used ML tree based algorithms they cannot predict at all the cholesterol, the gait analysis measurements and the Time up and Go, but with the same methodology they can predict very well the weight height ratio and the leg strength. The tree based algorithms give good results, but, eventually, a future exploration of other ML algorithms can be done in order to improve, or at least confirm, the achieved results.
The results of the features importance for the BMI and for the ISO are particularly rel-

evant: the very hight ones obtained from the 3 connective tissue parameters deserves a special mention. Much importance is usually given to muscles and fat in the CT scan analysis, but these results confirm that in the predictive process the connective tissue has an importance that is absolutely not negligible, in some cases also primary.
The database contains also other parameters like comorbidities relative to cardiac, lung or diabete diseases. The same methodology can be applied to classify these comorbidities: the classification would be binary, so good results could be achieved. In case of positive results they would further strengthen the predictive value of the 11 NTRA parameters. For this possible future implementation we could work choosing different numbers for the k-fold, avoiding k=16,18 and maybe considering k=10.
Finally this thesis can be considered as a starting point for a more in-depth analysis of the AGES-Reykjavik database: the prediction with high values of JI of any comorbidities would be an excellent result.

Bibliography

[1] Jiang F, Jiang Y, Zhi H, *Artificial intelligence in healthcare: past, present and future*, Stroke and Vascular Neurology 2017, June 2017

[2] Holzinger A, *ML for Health Informatics*, LNAI 9605 ,pp.1-24, 2016

[3] Nishita M, Anil P, *Concurrence of big data analytics and healthcare: a systematic review*, International Journal of Medical Informatics, 114, 57-65, March 2018

[4] Koh HC, Tan G, *Data mining applications in healthcare*, Journal of healthcare information management 19(2):65, 2011

[5] Romeo V, Maurea S, Cuocolo R, Petretta M, Mainenti PP, Verde F, *Characterization of Adrenal Lesions on Un-enhanced MRI Using Texture Analysis: A Machine-Learning Approach*, Journal of Magnetic Resonance Imaging, 2018

[6] Stanzione, A, Cuocolo R, Cocozza S, Romeo V, Persico F, Fusco F, Imbriaco M, *Detection of Extraprostatic Extension of Cancer on Bioparametric MRI Combining Texture Analysis and Machine Learning: Preliminary Results*, Academic radiology, 2019

[7] Johannesdottir F, Aspelund T, Siggeirsdottir K, Jonsson B, Mogensen B, Sigurdsson S, et al. *Age, Gene/Environment Susceptibility – Reykjavik Study: Multidisciplinary Applied Phenomics*, Am J Epidemiol 1076-1087, May 2007

[8] Chaco S, *"Machine Learning - Teaching Computers to Think"*, The NIH Cathalist vol. 26, Issue 4, July - August 2018

[9] Harris TB, Launer JL, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, Thorgeirsson G, et al. *Mid-thigh cortical bone structural parameters, muscle mass and strength, and association with lower limb fractures in older men and women (AGES-Reykjavik Study)*, Calcif Tissue Int. 354-364, 2012

[10] Edmunds KJ, Gislason MK, Sigurdsson S, Gudnason V, Harris TB, Carraro U and Gargiulo P, *Advanced quantitative methods in correlating sarcopenic muscle degeneration with lower extremity function biometrics and comorbidities*, PLoS ONE 13(3), March 2018

[11] Edmunds KJ, Arnadottir I, Gislason MK, Carraro U and Gargiulo P, *Nonlinear Trimodal Regression Analysis of Radiodensitometric Distributions to Quantify Sarcopenic and Sequelae Muscle Degeneration*, Comput Math Methods Med, December 2016

[12] Mah P, Reeves TE, McDavid WD, *Deriving Hounsfield units using grey levels in cone beam computed tomography*, Dento Maxillo Facial Radiology 39(6): 323–35, 2010

[13] Petursson T., Edmunds KJ, Gislason MK, Magnusson B, Magnusdottir G, Halldorsson G, and Gargiulo P, *Bone Mineral Density and Fracture Risk Assessment to Optimize Prosthesis Selection in Total Hip Replacement*, Computational and Mathematical Methods in Medicine Article, ID 162481, 2015

[14] Podsiadlo D, Richardson S. *The timed "Up and Go": a test of basic functional mobility for frail elderly persons*, J Am Geriatr Soc. 39(2):142–148, 1991

[15] Cesari M, Kritchevsky SB, Penninx BW, Nicklas B, Simonsick E, Newman A, *Prognostic value of usual gait speed in well-functioning older people - results from the Health, Aging and Body Composition Study*, J Am Geriatr Soc. 53(10):1675–1680, 2005

[16] Lang T, Cauley JA, Tylavsky F, Bauer D, Cummings S, Harris T, *Computed tomographic measurements of thigh muscle cross-sectional area and attenuation coefficient predict hip fracture: the health, aging, and body composition study*, Mineral Research 25(3), 513–9, 2010

[17] Raschka S, *Python Machine Learning*, Packt Publishing, Birmingham - Mumbai, 2015

[18] Russle SJ and Norving P, *Artificial intelligence: A modern approach*, Pearson Education Limited, Harlow - England, 2016

[19] Deo RC, *Machine Learning in Medicine*, Circulation; 132(20): 1920–1930, November 2015

[20] Visintin M, *ICT for Health - Master Degree Course*, ICT for Smart Societies - Politecnico di Torino, Academic Year 2016/2017

[21] Asuero AG, Sayago A, Gonzalez AG , *The correlation coefficient: an overview*, Critical Reviews in Analytical Chemistry, 36:41–59, 2006

[22] Zadrozny B, *Reducing multiclass to binary by coupling probability estimates*, Advances in Neural Information Processing System, 2002

[23] Aly M, *Survey on Multiclass Classification Methods* , Technical report, California Institute of Technology, 2005

[24] Jaccard P, *Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines*, Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272, 1901

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research. 12: 2825–2830, 2011

[26] Breiman L, Friedman J, Olshen R and Stone C, *Classification and Regression Trees*, Belmont, California: Wadsworth, 1984

[27] Ho Tin Kam *Random Decision Forests*, Proceedings of the 3rdInternational Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282, August 1995

[28] Ho Tin Kam *The Random Subspace Method for Constructing Decision Forests*, IEEE Transaction on Pattern Analysis and Machine Intelligence, VOL.20, NO.8, August 1998

[29] P. Guerts, D. Ernst, L. Wehenkel *Extremely Randomized Trees*, Machine Learning 63: 3-42, 2006

[30] Y. Freund and R.E. Shapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Science 55, 119-139 Article no. SS971504, 1997

[31] Drucker, Harris, *Improving Regressors using Boosting Techniques*, ICML, 1997

[32] J.H. Friedman *Greedy Function Approximation: a Gradient Boosting Machine*, Technical report, Dept. Of Statistics, Standford University, 1999

[33] J.H. Friedman *Stochastic Gradient Boosting*, Computational Statistics and Data Analysis, Vol 38, Issue 4, Pages 367- 378, https://doi.org/10.1016/S0167-9473(01)00065-2, 2002

[34] Singh H, *Understanding Gradient Boosting Machines*, Towards Data Science, November 2018

[35] Shaikh R, *Cross Validation Explained: Evaluating estimator performance*, Towards Data Science, November 2018

[36] Kohavi R, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, International Joint Conference on Artificial Intelligence, 1995

[37] G. Louppe, L. Wehenkel, A. Sutera and P. Geurts *Understanding variable importances in forests of randomized trees*, Conference Paper in Advances in neural information processing systems 26, December 2013