InVivo Muscle
Analysis of Functionality-related Patterns

Pedro A. Chagas
Instituto Superior Técnico, Lisboa, Portugal
pedro.chagas@tecnico.ulisboa.pt

Ana L. N. Fred
Instituto de Telecomunicações and Instituto Superior Técnico, Lisboa, Portugal

Mamede A. Carvalho
Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal

***Abstract*—InVivo Muscle project aimed to characterize the Portuguese elderly population. Two datasets were created, one resultant from answers to sociodemographic, health and physical activity questionnaires and results from functionality tests and other from biomechanical signals from Walking, Stair Ascend and Stair Descend trials. The goal of this work was to apply different Pattern Recognition techniques to those datasets in order to discriminate Low and High-Functional populations and determine the most relevant features for that discrimination.**

For the first dataset, using k-NN and NB Classifiers (Supervised Learning), error rates of 6.9% and 8.7% were respectively obtained. Using Wrapper Feature selection (FS) those results improved: 4% and 6.4%. Using Ward's and K-means clustering (Unsupervised Learning) errors of 8.8% and 8.1% were obtained. With Filters and Wrappers FS those results improved to 8.2%, 6.6% (Filters), 7.6% and 4.6% (Wrappers). For the Biomechanical dataset, the error for k-NN was 4.9% and 0% (with and without FS), clearly overfitted results. NB presented errors of 16.0% and 12.3%, respectively. Ward's obtained errors of 29.6%, 13.6% and 18.5% (no FS, Filter and Wrapper FS) and for K-means, 37%, 14.8% and 18.5% of error were determined for the same respective cases. Ward's lifetime achieved consistency levels of 0.889.

The most relevant features non-directly related to the calculation of Functionality scores were, for the first dataset, gender, body mass index, physical activity, age and living alone. For the Biomechanical dataset joint angles and moments, especially from the Pelvis, Hip and Ankle, were the most selected features.

***Keywords*—Pattern Recognition, Functionality, Biomechanical signals, Elderly, Supervised and Unsupervised Learning, Feature Selection.**

I. Introduction

A. Problem under Study and Objectives

Throughout the last decades, the world population, especially in industrialized countries, increased significantly its elderly population rate. According to the World Health Organization, the population over 60 years was estimated to be 688 million in 2006 and it is expected to reach 2 billion by 2050 [1]. The main reasons for this rise in the elderly population are related to the increase in life expectancy, the decline in fertility rates and some demographic factors such as migration, with consequences on the economic balance of the countries where this trend occurs, due to the growing need for care and support structures to the elderly population [2]. The rise of the elderly population rates led to the need of thinking about strategies that would increase their quality of life, in particular their functionality, because elders' quality of life is strongly related to their ability to perform everyday activities. Inactivity and sedentary lifestyle in the elderly contribute to the difficulty in performing this type of activity, resulting in the loss of function and independence as well as increasing the risk of falls [3] [4].

Given the need of improving the elder populations' life quality and Functionality levels, the project InVivo Muscle was created. It was developed by research teams from Faculdade de Motricidade Humana (FMH), Fundação da Faculdade de Ciências e Tecnologia (FFCT) and Instituto de Telecomunicações (IT) from Insituto Superior Técnico. The beginning of the project dates back to 2009 and intended to study and characterize the Portuguese senior population, regarding some sociodemographic parameters, health, physical activity and functionality, as well as determining the decisive factors to this population falling tendency. This study was conducted with a test group of about 1500 Portuguese elders aged over 65 years, randomly selected in Lisbon and Tagus Valley area. They had to answer to a number of questionnaires regarding sociodemographic, health, functionality and falls-related variables, as well as physical activity evaluation, by performing some functional tests [5]. From the original group, a subsample of 55 individuals was selected, in order to assess their biomechanical performance. These tests were based on three different locomotor tasks: Walking, Stair Ascend and Stair Descend. Each task was performed over a force plate, in a Biomechanical Laboratory, and resulted in the recording of 36 relevant signals for each task and each individual, corresponded to the angles, moments and power values of the ankle, knee and hip joints in respect to the three major axes (sagittal, frontal and horizontal), the trunk and foot absolute angles in relation to these same three axes and the foot ground reaction forces of the subjects. In addition, 16 spatiotemporal global variables were recorded, which included the time, speed and other global features for each trial [6].

The aim of this study concerns the discrimination of elderly populations, based on Functionality levels, by analysing and processing both sets of signals resultant from the questionnaires and functional fitness tests performed and the biomechanical trials, using Machine Learning approaches, and the determination of the most discriminant feature with respect to Functionality. After applying the proposed Learning methods to the datasets, the goal is to find agreement between the results and two classes of labels, corresponding to the level of Functionality of each individual, dichotomized in terms of Low and High Functionality, according to the Functional Fitness tests results. Feature selection techniques will be used to study the features with largest discriminant potential in terms of Functionality levels.

II. State of the art

A. Gait, ageing and Functionality

The word Gait describes a particular manner of walking or movement of the limbs over solid ground typical of animals. It is characterized by a sequential pattern and most species have several (but restricted) types of gait movement according to their different needs. Even among individuals of the same species, the gait pattern is different for every single one, since it depends on factors such as the body segments size, the muscleskeletar system, health concerts, among others. Human gait is a bipedal pattern characteristic of human movement and encompasses walks, jogs, runs, sprints and other kinds of human natural locomotion. A full gait cycle can be divided in the phases of Heel Strike, Foot Flat, Mid-Stance, Heel-Off, Toe-Off and Mid-Swing , always relative to the movement of one limb. The phase that ranges from Heel Strike to Toe-Off is named Stance phase, while the rest of the cycle is named Swing Phase (Figure 1). Parameters like the time and length of the full cycle, the stance and swing phases, cadence and speed are important for the correct analysis of Gait.

Ageing can lead to several social, physical and physiological changes and all those changes can become a contributing factor to adaptations of elders' gait cycle. Decrease in muscle strength, loss of

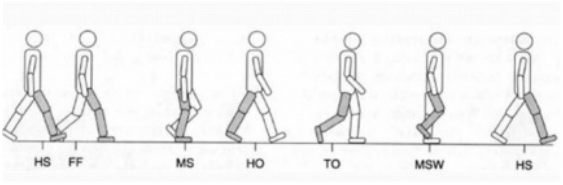


Fig. 1: A full gait cycle: HS: heel strike; FF: foot flat ; MIS: mid-stance; HO: heel-off; TO: toe-off: MSW: mid-swing [7].

motor neurons, muscle fibers and aerobic capacity, joint wear, loss of range of motion, sedentary lifestyle, decrease in physical activity and functionality are among the main factors that contribute to the gait adaptations observed in elderly subjects [8].

Several studies have been conducted in order to analyse human gait patterns and variables acquired in biomechanical trials. Gait patterns have been related to the study of Hemiplegia [9] and Cerebral Palsy [10], to analyze the implantation of prosthetics [11] or even to individual recognition [12] [13]. Regarding the analysis of elderly gait patterns, statistical [14] [15] and different Machine Learning [16] [17] [18] approaches have been proposed. Regarding the InVivo Muscle project, studies of falls with statistical analysis of the Questionnaires dataset [5] [19], of Functionality with Machine Learning approaches to the Biomechanical dataset [6] and of joint kinematics and kinetics with the same dataset [20] have been presented so far.

B. Mathematical tools and Algorithms

1) Pattern Recognition: Pattern recognition is the act of collecting and analyzing raw data, processing the information and comparing it to previous experiences and is a crucial ability in everyday aspects of human life. Machines can also develop this kind of learning through previous examples and are essential when human processing capacities reach is limit. Machine Learning and Pattern Recognition have a wide spectrum of applications, such as supporting medical decisions and computer-aided diagnosis, biometrics, speech and handwriting recognition, text classification, computer vision or image analysis and even in the area of psychology and emotion modeling.

A basic pattern recognition system involves the tasks of sensing, processing and classification. Sensing is the process through which raw data are gathered and transduced from the environment physical objects, normally interpreted by the computer as signals, which are processed and transformed a pattern set, defined as a n x d matrix X, such that $X = [x_1, x_2, \ldots, x_n]^T$. Matrix X has n, d-dimensional patterns $x_j = [x_{j1}, \ldots, x_{jd}]$, for $1 < j < d$. Each pattern, representing an individual, data point, sample, example, or observation, corresponds to a single row in matrix X and its respective columns represent the values measured for each feature. The group of all features $F = [f_1, \ldots, f_d]$ is called a feature set. The dimensionality of the learning space is then defined by the number of features d and each pattern set is composed by n patterns an d features. Features can either be discrete (categorical) or continuous (numeric) [21] [22]. Each pattern has an associated class y_i that belongs to a Class Set $Y = [y_1, \ldots, y_c]$ of c different classes. The number of classes depends on the problem itself. Using the pattern set X as training input, the purpose of the classifier is to accurately decide to which class some unknown patterns should be assigned. In order to improve the performance of the Classifier, and thus the performance of the whole Pattern Recognition system, some extra steps can be added, as shown in Figure 2. These include data pre-processing steps (normalization, for example) and feature extraction or feature selection (for dimensionality reduction purposes).

2) Feature Definition and Feature Selection: In order to reduce the computational cost of Machine Learning techniques, data redundancy, to improve the accuracy of classifiers and to avoid overfitted results, it is common to describe datasets in terms of their most relevant features. This can be a quite subjective process that normally requires

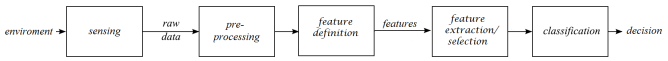


Fig. 2: Pattern Recognition System.

human expertise and lots of experience because one has to be able to capture the most relevant information from the original dataset in order to correctly discriminate the existing classes, keeping in mind that too few features can lead to loss of important information and too many features increase the probability of overfitting.

The term Feature Extraction (FE) can either be applied to the definition of new features from a dataset or to the remapping of data into low dimensionality spaces. This methods include Principal Component Analysis, Multidimensional scaling and Auto Encoders and can either be applied to a raw dataset or to a predefined feature set, with the disadvantage of losing the original characteristics of data. To avoid confusing the two terms, from here on the method of extracting relevant information from raw datasets using visual analysis and measures over the data will be named Feature Definition (FD), while the process of mapping data into lower-dimensional spaces will be named FE.

The process of Feature Selection (FS) is useful to select only the subset of features that ensures the best possible classification results, thus reducing the set's dimensionality and potential redundant information. FS methods can be divided in Wrapper, Filter and Embedded methods. Wrappers test a given feature subset based on a criterion J, that can either be the performance of classifier, trained for the specific subset, in the case of supervised learning, or an evaluation measure of a clusters result, in unsupervised learning. To avoid exhaustive search on all the possible feature subspaces, heuristic approaches are often applied, in particular Greedy Search, a class of methods that, at each step of the algorithm, considers all the possible hypothesis and chooses the best one, without the possibility of later revoking that choice. Sequential Forward Search (SFS) is a greedy search method that starts with an empty set and adds, at each step, the feature whose inclusion maximizes J. The algorithm continues until all features are added or some stopping criterion is met. Other Wrapper approaches include Sequential Backward Search, Sequential Floating Forward Search and Branch and Bound methods. Filters, on the other hand, perform tests independently of the learning machine or classifiers, the subset choice is only based on its own characteristics. The Filter FS approach consists in ranking features, either individually or in groups, based on some measure of information, consistency or separability of the feature subset. Finally, Embedded approaches perform feature selection and subset evaluation within the learning machine, hence the name embedded.

3) Learning Methods: Machine Learning methods are divided in Supervised and Unsupervised learning, according to the use of labeled data in the learning process.

Given a dataset X of d samples for n features and a respective output y, supervised learning aims to find a relationship between the dataset X and its output y. This process is achieved by splitting the data into a training set X_L and a test set X_t. This splitting can be achieved by cross-validation, a method that separates the data into k equal-sized disjoint subsets, called folds, randomly chosen from the original set, using one set for testing and the rest for training, repeating the process so that all folds are used as test sets. Leave-one-out cross validation is a special case of k-fold, where k equals the number of observations. The training set X_L, along with its respective output y_L, will train a classifier that will try to map the input to the output. Later, the test set X_t is tested with the trained classifier in order to predict its output and this predicted result is compared to the test set's true label y_t. Decision Trees (DT), Artificial Neural Networks (ANN), Statistical algorithms, Instance-based learning algorithms and Support Vector Machines (SVM) are the most commonly used classifiers. Naive Bayes (NB) is a statistical

classifier based on the Bayes' Theorem that assumes that the features of the dataset being classified are independent from each other but still performs with feasible results when this assumption is wrong. The algorithm chooses the most probable hypothesis to class prediction, using a maximum a posteriori (MAP) decision rule [23]. The classes' probabilities can be assumed to be equal for all the classes or estimated from the training set (Maximum likelihood class estimate). The distribution for the features probabilities is obtained by assuming a model or distribution, like Gaussian models. An output predicted class y^* is chosen according to the class c_j that maximizes:

$$y^* = \underset{c}{\operatorname{argmax}} P(C = c_j) \prod_{f_i \in F} P(f_i | C = c_j). \tag{1}$$

The most known instance-based classifier is the k-Nearest Neighbours (k-NN) algorithm, based on the assumption that similar patterns, possibly with the same class labeling, should be spatially close to each other. A new pattern being tested is compared to the k nearest patterns and its class is chosen by majority voting, the most frequent class of the k nearest neighbours is the class assigned to the new pattern. This assignment can also be done through weighted voting, for example by associating a weight proportionally inverse to the distance to the new pattern. The most common metrics used in k-NN are Euclidean, Manhattan, Mahalanobis, Minkowski, Chebychev or Hamming distances. The choice of the value of k is determinant for the result of the classification process. While higher values of k can reduce the influence of potentially noisy data, lower values of k are better if the class defining regions are small. Ideally, this value should be tested and chosen through optimization [23].

Unsupervised learning aim to find structures or patterns in unlabeled data. Some examples include Self-Organizing Maps and Clustering techniques. The main goal of clustering is to find natural groups within a set of data using only its intrinsic characteristics. The clustering problem can be formulated as finding K groups in a dataset of n objects based on similarity measures, so that objects of the same group have high similarity with each other and low similarity with objects clustered in other groups [24]. Hierarchical clustering algorithms construct clusters by repartitioning the set in each step of the algorithm, either staring with one cluster to which all objects belong and dividing it until each object has its single cluster, the divisive (top down) approach, or by starting with a cluster per object and merging the clusters until there is only one, the agglomerative (bottom up) approach. Data are merged or split hierarchically based on a proximity matrix between their elements and the clustering result is often displayed as a tree-like structure called dendrogram. The final clustering result comes from cutting the dendrogram at different levels, either to obtain a fixed number of clusters or according to some chosen criteria. Ward's method [25] is an agglomerative clustering method that merges cluster as a variance problem, instead of using distance metrics or measures of association, because it does not combine the two most similar objects or clusters, but instead combines the ones whose merging will minimize the increase of the overall within-cluster variance. That variance is measured by the error of the sum of squares between two clusters, given by:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \sum_{k=1}^{p} (\bar{x}_A^k - \bar{x}_B^k)^2 \tag{2}$$

where $\bar{x}_X^k$ represents the mean of the examples for attribute k in cluster X, hence the center of the cluster, and n_X the number of examples in X. The value of Δ starts as zero, for the case where all the elements belong to its own cluster and grows with the subsequent merging of clusters. The two clusters merged at each step are the ones whose merging minimizes Δ. The concept of k-cluster lifetime is defined as "the range of threshold values on the dendrogram that lead to the identification of k clusters" and it was proposed by Fred and Jain [26] as a way of finding data partitions that recover the data's natural clusters.

K-means clustering [27] [28] is a partitional square-error based algorithm that assigns data examples to K clusters according to a proximity criterion to K centroids, the means of the points belonging to that cluster. The algorithm works by minimizing the squared-error between the each cluster's centroid and its elements. In each step, the algorithm assigns each element to the cluster corresponding to its nearest centroid and recalculates the centroids for the K clusters. K-means is a greedy algorithm, that converges to local minima and requires the user to specify the number of means and the distance metric, usually Euclidean distance.

4) Evaluation Criteria: FS schemes use different measures to evaluate the feature subsets and find the best possible one. While Wrappers measure performance with cluster validation methods, Information and Correlation-based criteria are amongst the most recurrent measures for feature ranking in Filter FS. Mutual Information (MI) reflects the dependency or shared information and measures the amount of information obtained about a variable X by measuring the other variable Y. MI is a non-negative measure and it is equal to zero if and only if the two measured variables are independent from each other. It is a symmetric measure and can be expressed in terms of the entropy and the joint entropy of the variables. MI is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

$$I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X,Y) \tag{4}$$

where $H(X)$ and $H(X,Y)$ represent Entropy and Joint Entropy.

A filter FS method using MI was proposed [29] based on the concept that the features in a good feature subset should have maximal relevance with the class labels and minimal redundancy with each other (mRMR). The best combinations of these two criteria are the Mutual information difference (MID) and Mutual information quotient (MIQ) schemes, shown respectively in Equation (5) and Equation (6) [30].

$$max_{x_i \in S^*} [I(x_i; c) - \frac{1}{|S|} \sum_{x_j \in S^*} I(x_i; x_j)] \tag{5}$$

$$max_{x_i \in S^*} [I(x_i; c) / [\frac{1}{|S|} \sum_{x_j \in S^*} I(x_i; x_j)]] \tag{6}$$

where $|S|$ its cardinality of the feature set.

Correlation-based criteria measure the dependence (correlation) between two variables X and Y. Pearson product-moment correlation coefficient, the most popular, measures a linear correlation between the variables that can take any value between -1 and 1, taking the extreme values when variables have perfectly linear relationship (negative or positive).

In supervised learning, the performance of the classifier is measured in terms of error or accuracy between the predicted outcome and the true label. Error is expressed in terms of a Probability of Error P_e, that is related to the number of incorrectly classified instances and is defined as the number of misclassified elements divided by the total number of elements. Accuracy ACC, on the other hand, takes into account the number of correctly classified elements so, naturally, $ACC = 1 - P_e$. Unsupervised learning assesses the quality and consistency of the determined cluster based on internal or external methods depending, respectively, on the usage or not of external information to validate the clusters. Internal criteria evaluate the characteristics of the dataset and the result is typically considered a good one if the cluster presents good intra-cluster compactness and inter-cluster separation, assessed by measures such as the Davies-Bouldin index [31], the Dunn's index [32], the Silhouette index [33] or the S Dbw index [34]. External criteria assume the existence of external information, like the true class labels of the clustered dataset or some pre-classified objects, to compare with the resultant clusters found by algorithm. Jaccard Index (JI) [35] is a statistical measure

used to assess the similarity of finite sets, defined as the coefficient between the size of the intersection between two sets C_1 and C_2 and the size of their union: $|C_1 \cap C_2|/|C_1 \cup C_2|$ and bound between 0 and 1. For the binary classification case, where only two clusters/classes exist, the index can be written as:

$$J = \frac{n_{11}}{n_{01} + n_{10} + n_{11}} \tag{7}$$

where 1 and 0 represent the two clusters and n_{ij} represents the number of objects clustered in the Cluster i with true class label j.

Adjusted Mutual Information (AMI) [36] is a variation of the MI measure, proposed specifically for clustering results comparison, corrected for the randomness of MI between two clusters. It is based on the correction proposed for the Adjusted Rand Index [37]: Adjusted Index= (Index−Expected Index)/(Max Index−Expected Index). Therefore, AMI is given by:

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\max\{H(U), H(V)\} - E\{MI(U,V)\}} \tag{8}$$

where U and V represent two distinct cluster partitions (the result and the true label, for example), H represents Entropy, MI represents the clusters' Mutual Information and $E\{MI(U,V)\}$ represents the expected value.

In 2001 Fred [38] proposed a method to test results of different clustering algorithms and a criterion of evaluation performance. The clusters' consistency is expressed in terms of a Consistency Index (CI), a ratio between shared objects in matching clusters in different partitions (the clustering result and the true label, for example) and the total number of objects. This index, pc_idx is given by:

$$pc\_idx = \frac{1}{n} \sum i = 1^{min(nc_1, nc_2)} nshared_i \tag{9}$$

where nc_j represents the number of clusters in partition j. Given that the final clustering algorithm result finds a number of clusters corresponding to the number of classes of the true label, the value of the CI corresponds to the aforementioned concept of Accuracy ACC. If and only if this is verified, it can be stated that $CI = 1 - P_e$.

III. Methodology

For every situation, data was tested in their raw version, as well as with two normalization methods, the Z-score (Equation (10)) and the Feature Scaling (Equation (11)). The three testing methods will be further named n_0, n_1 and n_2, respectively.

$$X' = \frac{X - \mu}{\sigma} \tag{10}$$

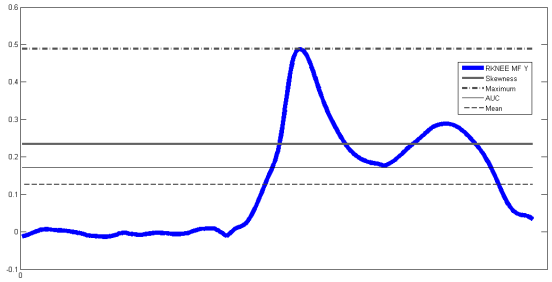
$$X' = \frac{X - min(X)}{max(X) - min(X)} \tag{11}$$

where X, μ and σ represent the set of data, its mean and standard deviation, respectively.

A. Feature Definition

As the Questionnaires and Functional Fitness tests dataset was the result of answers or test's scores, it does not make sense to further process or calculate specific characteristics of the set, nor to apply dimensionality reduction techniques, as potentially discriminative and relevant information would be lost.

For the Biomechanical dataset, taking as basic descriptor the variables provided by FMH team, new features were defined based on statistical analysis or computations from the above, in particular those corresponding to time series (Cyclic Variables). Based on a previous study for functionality-related pattern recognition with this same dataset [6], the same features were manually extracted from the biomechanical data. Three feature sets were created, one for each of the 3 tasks, and the defined features consisted on means, standard deviations, maxima, minima, areas under the curve and skewness of



(a)

Fig. 3: Stair Ascend illustrative signal and features defined.

gait signals. The final feature sets have 33, 31 and 37 features, for the W, SD and SA tasks, respectively. Figure 3 illustrates one of the biomechanical acquired signals for a High-functionality and a Low-functionality subject, the respective features defined for that signal

B. Stratification

To overcome for missing or incorrectly annotated data, the Questionnaires and Functional Fitness tests datasetas tested using different groups. g_1 is the original set of variables. g_2 is the groups for each the missing or incorrect values were replaced by the feature mean. g_3 is the groups for each only features with valid answer to every individual were selected. An extra group was tested, g_4, for each only dichotomized variables were chosen. The dimensionality of the four groups is 63 features for the first two and 19 and 32 features for the last two, respectively. Every group has samples for 1364 individuals.

C. Feature Selection

Both feature sets were tested in their full form and some models of FS were applied in order to achieve better results, by removing some redundant features. For the Supervised Learning proposed methodology, the FS process starts by partitioning the data in 2 equal sized random parts, using 2-fold cross validation. One of this partitions, the validation set, was used to assess the final performance of the trained classifier. The remaining partition was tested using the proposed methodology. First, k-fold cross validation will be performed to partition data into a training and a test set. Then, a Wrapper FS method will be used, performing Sequential Forward Search (SFS) in order to find the number of features that ensures the best classification result (largest consistency with the true labels) for the test set, using NB and k-NN as chosen classifiers. The best subset of features found by the selection algorithm trained a new classifier, whose performance was accessed by the validation set. Two FS strategies were used with Unsupervised Learning: Wrapper and Filter FS. Wrappers started by selecting the one best feature (using 3 different measures: AMI, JI and CI) for the K-means, Ward's and Ward's Lifetime clustering cases. Then, it used a SFS approach to select the remaining features, using the 3 measures to determine the features order of importance. 1 run of the Ward's clustering and 25 run of the K-means method was used. Two different results were recorded for the K-means algorithm, the feature subset that was more frequently selected in the 25 runs and the feature set that originated the best absolute result, independently of the number of times it was selected. Filter algorithms elaborate a rank of features, based on a given criterion. Six different criterion were tested: Correlation of the individual feature with the remaining features (corr2), Correlation of the individual feature with the true label (corr1), Mutual information between the individual feature and the remaining features (mi1), Mutual Information between the individual feature with the true label (mi2), Self-Information of each individual feature (si) and two feature rankings based on the mRMR concept [29], Mutual Information Difference (MID) and Mutual Information Quotient (MIQ), both in

relation to the true label. The next step consists on clustering the first feature of the ranking, via K-means and Ward's clustering, and obtain the respective consistency with the true label. The process continue by adding each next feature of the ranking at each step, and calculating the respective consistencies. The final feature subset is the one that presents the highest consistency score.

D. Learning Methods

The basic setup proposed for the Supervised Learning consists on performing k-fold cross validation over the feature set, using one fold as a test set to evaluate the performance of a classifier trained with the remaining folds. All the folds are used as test sets, in different runs. Two different classifiers are used: the NB, assuming a Gaussian distribution, and the K-NN. The cross validation is performed with different values of k, in order to find the partition of data that gives better results. The k-NN classifier will also be tested for different values of k, in order to find the number of neighbours that ensures the best classification result. Each test was repeated 10 times. The error probability is calculated as the number of subjects from the test set misclassified by the algorithm divided by the total number of subjects.

As for Unsupervised techniques, two different clustering techniques were applied to both datasets. Since the true labels were known, the result could be compared, using external validation techniques. The two clustering approaches are the K-means algorithm using squared euclidean as distance measure and two imposed centroids and Ward's Hierarchical clustering, assuming euclidean as intra-cluster distance and inner squared (minimum variance) as intra-cluster distance. Ward's was used in two different approaches: setting the number of cluster to 2, in order to match the number of true class labels, and Ward's Lifetime, in order to discover the number of clusters corresponding to the largest lifetime. K-means first centroids are randomly determined, and this fact can alter the final clustering result and so 25 runs were performed for each number of features. The error probability was calculated as 1 minus the number of shared elements in a given cluster compared to the true label, for the K-means and 2-cluster Ward's. In the case of Ward's Lifetime, results are expressed in terms of the consistency index proposed by Fred [38].

IV. Results and Discussion

A. Questionnaires and Functional Fitness tests dataset

For this dataset, using Supervised Learning, data was partitioned in in 3, 5, 8, 10, 15, 25, 30 and 100 folds. Leave-one-out cross validation was also tested. Values of k=1, 2, 3, 4, 5, 7, 10. 12, 15, 20. 30. 50 and 100 neighbours were chosen for the k-NN Classifier. With FS, due to the high computational complexity and running time of this method, only 5, 8, 10, 15, 20, 25, 30 folds and Leave-one out were tested.

The best obtained results for the Questionnaires and Functional Fitness tests dataset are summarized in Table (I) and can be considered very relevant in the context of the aim of this study, the discrimination of High and Low-Functionality populations. Overall, the result for populations recognition was 4% or error rate, in the best possible situation. Considering only the best result for each test performed, 22.7% of error rate was the worst performance. In general, Supervised Learning approaches achieved better results than Unsupervised Learning approaches. This realization is natural and expected, since data fed to the Supervised Learning algorithms is trained along with the respective class labels, enhancing the learning process and the establishment of relationships between the input and the desired output. Therefore, when a test sample is input to the classifier, it is normal that the recognition rates exceed those of Unsupervised Learning approaches, where the class labels were only used to assess the final result or, at most, optimizing the number of features used in the learning mechanism.

TABLE I: Comparison of the best results (in terms of percentage of error) for each test with the Questionnaires and Functional Fitness tests dataset. noFS - Results without Feature Selection.

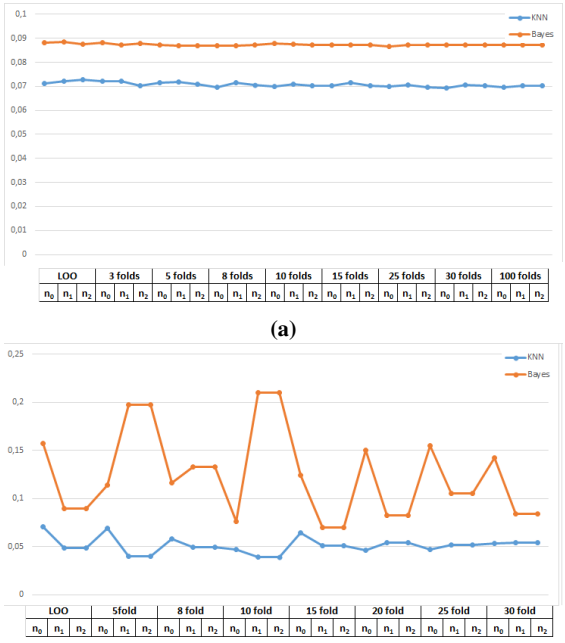
	g_1		g_2	
USL	Ward's	K-means	Ward's	K-means
noFS	22.7%	10%	11.2%	7.8%
filters	8.2%	6.6%	8.4%	6.7%
wrappers	7.6%	4.6%	7.6%	4.9%
SL	k-NN	NB	k-NN	NB
noFS	8.7%	8.8%	7%	8.8%
wrappers	6.4%	9.5%	4%	7%
	g_3		g_4	
USL	Ward's	K-means	Ward's	K-means
noFS	9.8%	13.5%	8.8%	8.1%
filters	8.6%	6.6%	8.5%	7.6%
wrappers	8.3%	6.5%	8.4%	6.7%
SL	k-NN	NB	k-NN	NB
noFS	6.9%	8.7%	7.3%	9.4%
wrappers	5%	6.4%		

Without FS, Supervised Learning approaches achieved very good results, from 7% to 8.8% of error rate. Considering the best results for each test performed, k-NN Classifier ensured better results than the NB Classifier. This results can be explained because of the assumptions inherent to the use of NB, feature independence and Gaussian distributed populations, since both these assumptions are not verified. Many features are dependent from each other, mainly because of the simultaneous presence of scores and sum of the same scores, features in their original and dichotomized form and features that were calculated based on the values of other features (like the Body Mass Index). This effect could have been minimized by using only independent features and that situation was also tested, with the g_4 group, where only dichotomized features were used. The results, however, show that the performance of the Naive Bayes Classifier is the worst for this case, which can be explained because of the second assumption, since Gaussianity is only defined for continuous variables, and even if the algorithm tried to interpret discrete data as Gaussian-distributed, dichotomized variables would not present Gaussianity at all and thus the worsening of the g_4 results is only natural. Even so, considering that both the assumptions are not verified, Naive Bayes classifier ensured a error rate of 8.7%, which is a good performance. The outperformance of the k-NN Classifier (7%) could also be related to situations of overfitting for lower values of k [39]. However, the values of k corresponding to the best results were never the lowest tested. Nonetheless, since the largest number of k tested was 100 and the dataset has almost 1500 individuals, largest values of k could have been tested to verify this situation.

The use of FS, in general, improved the results of Supervised Learning, achieving 4% and 6.4% of error rates, for the k-NN and the NB classifiers, respectively. This results are very good and could possibly improve, if other partitions than 50:50 split were tested for the validation set, since the classifiers were always trained with less patterns than those of the validation set. An alternative partition, one third of data for validation and the rest for FS for example, would probably ensure even better results.

Figure 4 illustrates the performance of the g_2 and g_3 groups for the different partitions tested with and without FS. Figure 5 shows the distribution of the parameters tested for the Supervised Learning with this dataset, regarding the best results, in particular the values of k chosen for the k-NN classifier and the best partitions for the cross-validation.

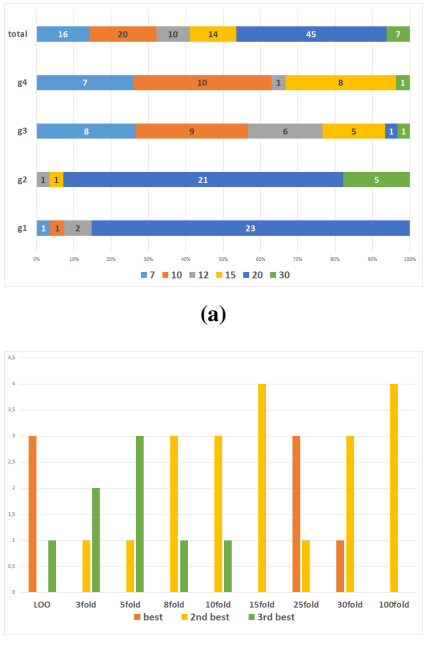
Regarding Unsupervised Learning approaches, since the class labels are not used for designing clusters, it is normal to verify that results without performing any FS were, in general, not so good in terms of Functionality discrimination. Even so, some surprisingly good results were found in these tests. The group g_1, since it contains the whole non-processed dataset, with missing or incorrect values,



(a)

(b)

Fig. 4: Scores for the Supervised tests with different partitions: (a) g_3. (b) g_2 with FS.



(a)

(b)

Fig. 5: Distribution of the best parameters in Supervised tests with Questionnaires dataset. (a) Best k values for the k-NN Classifier. (b) Best partitions for the k-folds cross validations.

(a)

(b)

Fig. 6: Scores for the Unsupervised tests with FS for the g_2 group: (a) Wrappers. (b) Filters.

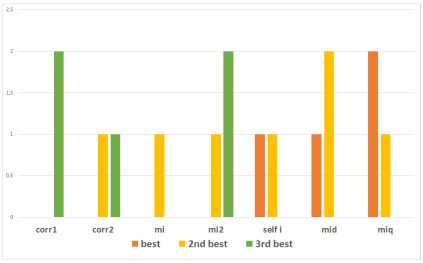
presented the worst result of 27% of error for the Ward's method. By replacing the incorrect information (g_2), selecting only features with correct information for all the individuals (g_3) or selecting only dichotomized features (g_4), Ward's results improved to 11.2%, 9.8% and 8.8%, respectively. For the K-means clustering, the best results for the four groups were 10%, 7.8%, 13.5% and 8.1%, respectively.

Filter FS methods were applied to the dataset in order to rank the best features according to several criteria. By selecting only the most relevant features, results improved in every situation. The best results were 6.6% or error rate for the K-means clustering (g_1 and g_2) and 8.2% or error rate for the Ward's method (g_1). Considering that this method does not take the class labels in account to perform FS, but only the inherent characteristics of data, these results can be considered very good. Mutual Information Quotient in relation to the true label was the ranking criteria that ensured best results more often. Since Wrappers use the class labels, not to cluster design but to perform cluster evaluation in order to select the best set of features, it is only normal to assume that this method ensured the best results. In every test, the best results improved in relation to the previous ones. For the K-means clustering, the best result was 4.6% of error rate (g_1), almost as good as the best result with Supervised Learning. For the Ward's method, the best result achieved was 7.6%. Several clustering validation criteria were used by the algorithms and JI was the one that more often ensured the best result. Ward's Lifetime was also used in Wrappers but K-means clustering ensured the best result for every possible situation. Ward's with 2 imposed clusters almost always achieved the second best results, relegating the Lifetime approach to the worst method in almost every test. The best performances of the K-means could be explained by the nature of the data representation, meaning that it is the most robust method for clustering datasets with categorical data.

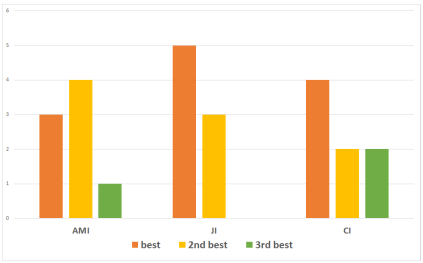
Figure 6 illustrates the performance of the g_2 and g_3 groups for the different partitions tested with and without FS. Figure 7 shows the distribution of the parameters tested for the Unsupervised Learning with this dataset, regarding the best results, in particular the best criteria for Filter feature ranking and Wrapper evaluation criteria.

B. Biomechanical dataset

For the Biomechanical dataset, since the number of individuals is considerably smaller, the data was only partitioned in 3, 5, 8 and 10



(a)



(b)

Fig. 7: Distribution of the best parameters in Supervised tests with Questionnaires dataset. (a) Best Filter rank criteria. (b) Best Wrapper evaluation criteria.

folds, as well as using Leave-one-out cross validation. Additionally, because 3 trials for each individual are present in the dataset, which could overfit the system in some tests, two other approaches were used. In the first case, LIO1, the all the data corresponding to one individual (from the 3 repetitions per trial) was removed and used as testing set against all the other individuals. In the second approach, LIO2, two trials of one individual were used as test set against a training set containing the third trial of the same individual, as well as all the remaining data. Values of k=1, 2, 3, 4, 5, 7, 10. 12, 15 and 20 were tested for the k-NN. The parameters chosen for the tests with FS were the same, with the exception that LIO partitions were not tested.

The best obtained results for the Biomechanical dataset are summarized in Table (II). These results were clearly worst in terms of High and Low-Functionality populations discrimination, when compared to the previous dataset. Although overall, the best result for populations recognition was 0% or error rate, this result is biased and the justifications will be presented. Considering only the best result for each test performed, 40.7% of error rate was the worst performance. As expected, Supervised Learning methods presented better results in general than the Unsupervised Learning ones. Like it was previously explained, this is a normal and expected outcome, due to the characteristics of each learning method.

Considering Supervised Learning with the whole dataset, it is important to start by discussing the (unrealistically) good results presented by the k-NN Classifier. For every task, Walking, Stair Ascend and Stair Descend, the best result with the k-NN was 0% of error rate, meaning that every sample in the test set was correctly classified, but results are clearly overfitted, because of two major reasons. First, since each set of data has three signals for each subject, it is only natural that once a pattern is being classified, the algorithm recognizes patterns from the same subject has the closest ones. Furthermore, k-NN classifier is said to overfit for low values of k [39] and the best results found were always corresponding to values of k between 1 and 4. To overcome this problem, the LIO1 case was tested and the results proved that overfitting was, in fact, present in the other tests, since the best results for the k-NN Classifier in these tests were 40.7%, 32.1% and 34.6% for the W, SD and SA trials, respectively. In this situation, the values of k had much more

TABLE II: Comparison of the best results (in terms of percentage of error) for each test with the Biomechanical dataset. noFS - Results without Feature Selection.

	W		SA	
USL	Ward's	K-means	Ward's	K-means
noFS	37.0%	37.0%	40.7%	37.0%
filters	18.5%	24.7%	13.6%	14.8%
wrappers	23.5%	22.2%	18.5%	18.5%
SL	k-NN	NB	k-NN	NB
noFS	0.0%	24.7%	0.0%	12.3%
wrappers	4.9%	24.7%	7.4%	16.0%

	SD	
USL	Ward's	K-means
noFS	29.6%	38.3%
filters	16.0%	21.0%
wrappers	19.8%	19.8%
SL	k-NN	NB
noFS	0.0%	15.2%
wrappers	8.6%	18.5%

variation and were always higher. For LIO2, as there was always still one pattern from that individual in the training set, the algorithm chose k values of 1 or 2 and also presented error rates of zero or near-zero. Even considering the other partitions tested, since finding patterns from the same individual was always a possibility, the algorithm chose low values of k in every situation, and the results ranges from 0% to 10% of error, although we can consider that overfitting was partially avoided in the latter case. As for the NB Classifier, the results are worse, in theory, but the overfitting problem is probably avoided, so they can in fact be considered more precise. The best result for the NB classifier was achieved for the SA trials, an error rate of 12.3%. The best results for W and SD trials presented error rates of 24.7% and 15.2% of error rate, respectively. These results, once again, can be explained because of the failure to verify the assumptions of the NB Classifier, feature independence and Gaussian distribution. While the first assumption is probably not true, because of features determined from the same original signal or features from related signals (joint power depends on joint moment, for example), the histograms for these features proved that, while some features present an acceptable Gaussian shaped distribution, some of them do not present Gaussianity at all.

Due to the small size of the data, results were worse for the FS tests for almost all tests. k-NN best results ranged from 4.9% to 8.6% of error rate, and always for values of k equal to 1 or 2. Every time that a test's best result correspond to a different value of k, the results were always worse (10% to 30%) and yet, the maximum value of k chosen considering the best results was k=5. LIO tests were not performed in this case, so the conclusions regarding overfitting are limited to those presented for the k-NN Classifier. Considering the results for the NB Classifier with FS, they were a bit worse than those for the whole set, ranging from 16% to 24.7%. This worsening can be explained by the reasons stated above, regarding the small size of the dataset. Probably a larger dataset or a different partition of data for FS optimization, as previously referred, would achieve best classification results with FS.

Figure 8 illustrates the performance of the SA trial for the different partitions tested with and without FS. Figure 9 shows the distribution of the parameters tested for the Supervised Learning with this dataset, regarding the best results, in particular the values of k chosen for the k-NN classifier and the best partitions for the cross-validation.

The results for the Unsupervised Learning were not as good as the Supervised Learning results (considering, for fair comparison purposes, the NB results, which were not overfitted). Without performing any FS, the Ward's clustering achieved error rates of 29.6% at best and the K-means clustering 37%. Although Ward's Lifetime results are expressed in terms of CI and not P_e, if we consider that $CI = 1 - P_e$ for the Ward's and K-means clustering, Ward's Lifetime performed best in this case, achieving a CI value of 0.741 for 3 clusters in the Walking task, which is best than the best result of 0.630 achieved

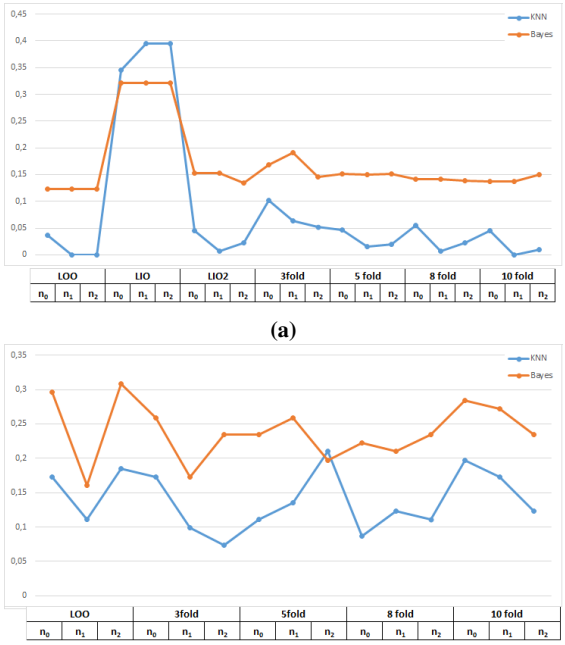


Fig. 8: Scores for the Supervised tests with different partitions: (a) SA. (b) SA with FS.

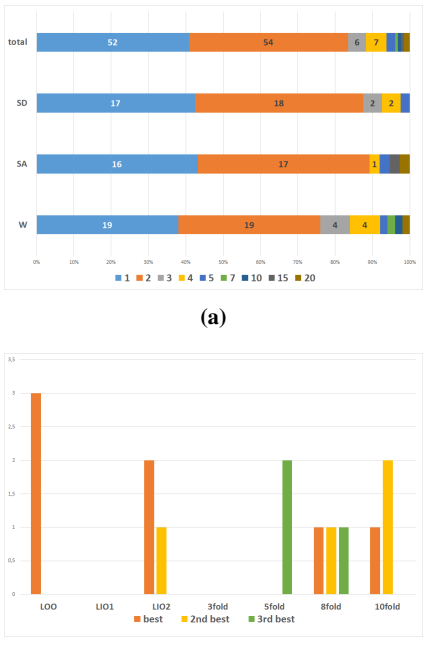


Fig. 9: Distribution of the best parameters in Supervised tests with Biomechanical dataset. (a) Best k values for the k-NN Classifier. (b) Best partitions for the k-folds cross validations.

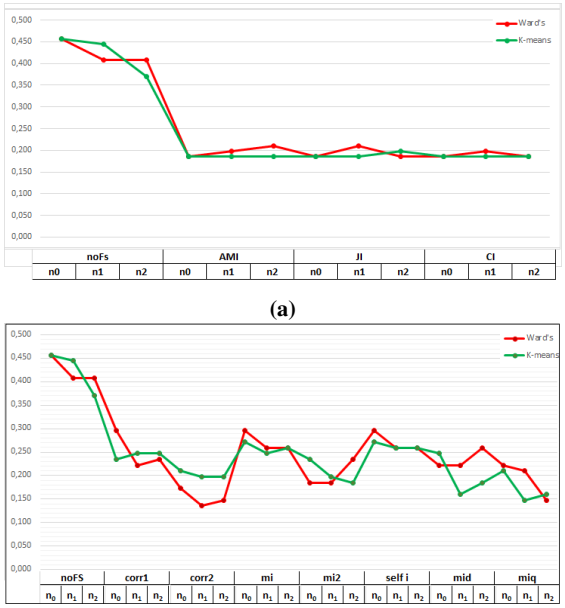


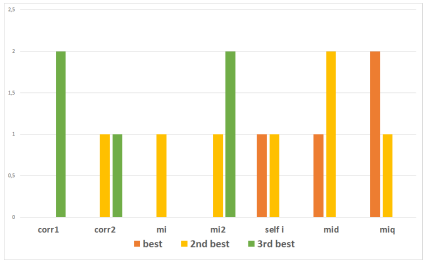
Fig. 10: Scores for the Unsupervised tests with FS for the SA trial: (a) Wrappers. (b) Filters.

by the 2 clusters Ward's for the same trial. The best result for the 2 clusters Ward's (CI =0.704), however, was corresponding to the Stair Descend trial, for which Ward's Lifetime presented the same result.

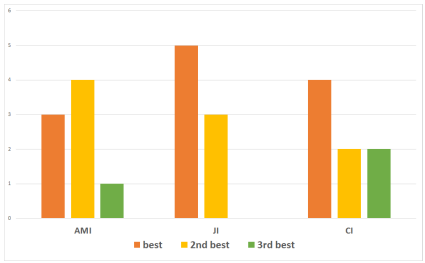
The results for Filter FS methods improved the results for the Biomechanical dataset, as expected, in every test. The best result arose from the Ward's method for the SA trial, 13.6% of error rate, almost as good as the best NB result for the same trial, 12.3%. The best result for the K-means clustering with Filter FS was 14.8% of error rate, for the same trial. For these tests, Ward's always performed better than K-means and Ward's Lifetime was not tested. Correlation with the true label (corr2) was always chosen has the ranking criteria with best result. For this dataset, Wrappers best results were outperformed by the Filters, for both the Ward's method with 2 clusters and the K-means clustering, 18.5% of error rate for both methods. Ward's Lifetime aproach outperformed the Filters, achieving a CI of 0.889 for 3 or 4 clusters (depending on the normalization) for the SA trial (the best Filter result corresponds to a CI of 0.864, for the same trial). For the remaining trials, the results for the Ward's Lifetime aproach were always better than every other Wrapper result and also better than Filters with K-means clustering. This outperformance of Ward's method could be explained by the fact that, in this dataset, all data are numeric, and Ward's has good results when only numeric data are clustered. Due to the small dimensions of the dataset and since the learning proccess to select the features in Wrapper methods uses a set of data that is not representative of the full structure, while in Filter methods, the full set of data is used hence, even without evaluating the classifier result to select the best features, the methods had more information regarding the data to perform selection, achieving better results. No clustering validation criteria was clearly best, but CI was the one that more often ensured a best result.

Figure 10 illustrates the performance of the SA trials for the different partitions tested with and without FS. Figure 11 shows the distribution of the parameters tested for the Unsupervised Learning with this dataset, regarding the best results, in particular the best criteria for Filter feature ranking and Wrapper evaluation criteria.

An analysis on the most frequently selected features was made from the results of the several FS tests. Regarding the Questionnaires and Functional Fitness tests dataset, it is important to note that the Functionality scores were calculated as a sum of the scores for the



(a)



(b)

Fig. 11: Distribution of the best parameters in Unsupervised tests with Biomechanical dataset. (a) Best Filter rank criteria. (b) Best Wrapper evaluation criteria.

Functional Fitness tests, hence it was expected that the variables from each of the tests were selected as more discriminant. In fact, 11 of the 20 most frequently selected features by FS methods with this dataset were used to calculate or related to the variables used to calculate the score of functionality. Nevertheless, remaining variables: gender, age, body mass index, fear of falling, two physical activity indicators (standing and walking), the variable related to living alone, education level and amount of medication intake were considered relevant and were more frequently selected, regarding Functional Fitness discrimination. This results are in accordance with the conclusions presented by Moniz-Pereira et al. [5]. Even though this study was related to falls, it also concluded that Functionality levels and falls tendency are highly correlated. One of the major conclusions of this study was that age was not directly related to falls, but this work's results proved that age and Functionality were closely correlated, since age-related variables were among the 20 most selected features and also among the 20 features with highest individual scores. Other interesting fact is that the most selected features were not always the ones with best individual scores: the feature related to the basic level of education was one of the most selected and had the worst individual score. As for the 20 features with highest scores, only 8 of them are contained in the top-20 of selected features. Most of this features had also determined to be good for fall-related populations distinction. Since the Functional Fitness tests were originally proposed for the task of falls discrimination, it is only natural to verify this. Student's *t*-test was also performed in order to analyze the significance of the population's mean comparison test. Accounting for the 15 features most frequently selected for every group, only two features (in g_4) were considered statistically non-significant.

As for the Biomechanical dataset, means of joint and absolute angles and joint moments were in general the most selected features. Other measures such as maxima, minima and standard deviations, as well as Center of Mass displacement means were also among the top-15 selected features for the 3 Biomechanical trials, although less frequently. Student *t*-test results were less coherent for the features of the Biomechanical trials. On the Walking trial, only 6 of the 15 most selected features were considered to have significance. The number of significant features, considering the best 15, was 12 and 10 for the Stair Ascend and Stair Descend trials, respectively. This realization supports the use of Machine Learning approaches when compared to purely statistical analysis on data, a methodology frequently used in Gait analysis studies. Combing the results obtained with FS and the significance decisions from the *t*-tests, some signals for the Biomechanical dataset arise as the most distinct in High and Low-Functionality populations. For every trial, the Hip and Pelvis angles mean in relation to the X axis were larger for the Low-Functionality population. Ankle and Hip's joint moment was diminished in the Walking trial and in both Stair trails, respectively, for the Low-Functionality subjects. This results show an adaptation Gait patterns from Low-Functionality subjects, as explained in previous studies [6] [8] [40] [41] [42]. The individual scores for the Biomechanical features did not differ from each other significantly, thus it is not very interesting to try to compare the best individual scores with the more often selected features, as it was for the previous dataset.

V. Conclusions

The several methods proposed and tested for Functionality discrimination achieved overall good recognition results. In particular, for the Questionnaires and Functional Fitness tests dataset both supervised and unsupervised learning approaches produced very small errors. 6.9% and 8.7% were the error rates obtained for the k-NN and NB Classifiers, respectively. With FS, those rates improved to 4% and 6.4%. Ward's and K-means clustering achieved error rates of 8.8% and 8.1%, which were reduced to 8.2% and 6.6% with Filter FS and to 7.6% and 4.6% with Wrappers. Regarding the Biomechanical dataset, k-NN found overfitted results of 0% and 4.9% with and without FS. NB Classifier achieved worse but more realistic results of 12.3% and 16.0% for the same situations. The increase of the error with FS was due to the small size of the dataset. With Unsupervised methods, Ward's obtained 29.6%, 13.6% and 18.5% of error rate for the cases of no FS, Filter FS and Wrapper FS and K-means clustering achieved errors of 37%, 14.8% and 18.5%, respectively. Ward's Lifetime approach achieved a better result, a consistency level of 0.889.

The most relevant features non-directly related to the calculation of Functionality scores were, for the first dataset, gender, body mass index, fear of falling, physical activity, age, living alone, education level and medication intake. For the Biomechanical dataset joint angles and moments, especially from the Pelvis, Hip and Ankle, were the most selected features.

Some other approaches would be worth to investigate. First of all, Global variables from the Biomechanical tests, related to the spatiotemporal information of the trials, were not used with Machine Learning approaches. In fact, Student's *t*-test results for these variables only found significance in 5 of them (from a total of 16 variables) for the Walking task and one of them from the Stair Ascend task. However, as stated, the results of this statistical test do not always reflect the relevance of the features in terms of Functionality discrimination, so a future study with this variables (which are directly measured in the trials) could be interesting in order to verify their discriminant potential. Furthermore, other Machine Learning techniques could be tested in order to compare with the ones presented in this study. An increase on the size of the data set would probably be a good improvement for the results for the Biomechanical dataset. Feature Extraction techniques could be applied to datasets to access if the results would improve when compared to the Feature Selection methods tested. Other preprocessing or Feature Selection methods can also be tested for optimization of results. Regarding Learning methods, two approaches are considered to be promising. The use of Gaussian and Multinomial Mixture Models (like the one proposed by Figueiredo and Jain [43]) which was initially considered as an approach for this study, could have been interesting, in particular because the algorithm automatically selects the number of components, a good indicator to compare with the two Functionality levels. Deep Learning networks would also be a possible approach,

since this method allows the automation of Feature Extraction and Selection, reducing the subjectivity level imposed by human expertise. Deep Learning has been proved to produce good results in several Machine Learning areas.

REFERENCES

[1] W. H. O. Ageing and L. C. Unit, *WHO global report on falls prevention in older age*. World Health Organization, 2008.

[2] G. F. Anderson and P. S. Hussey, "Population aging: a comparison among industrialized countries," *Health affairs*, vol. 19, no. 3, pp. 191–203, 2000.

[3] A. Ozturk, T. T. Simsek, E. T. Yumin, M. Sertel, and M. Yumin, "The relationship between physical, functional capacity and quality of life (qol) among elderly people with a chronic disease," *Archives of Gerontology and Geriatrics*, vol. 53, no. 3, pp. 278 – 283, 2011.

[4] K. Collins, B. L. Rooney, K. J. Smalley, and S. Havens, "Functional fitness, disease and independence in community-dwelling older adults in western wisconsin," *WMJ-MADISON-*, vol. 103, no. 1, pp. 42–48, 2004.

[5] V. Moniz-Pereira, F. Carnide, M. Machado, H. Andre, and A. P. Veloso, "Falls in portuguese older people: procedures and preliminary results of the study biomechanics of locomotion in the elderly," *Acta Reumatol Port*, vol. 37, pp. 324–332, 2012.

[6] M. S. Santos, V. Moniz-Pereira, A. Lourenço, A. L. N. Fred, and A. Veloso, "Relevant elderly gait features for functional fitness level grouping." in *PhyCS*, 2014, pp. 153–160.

[7] P. Bergwandelen, "Loopcyclus," https://eduweb.hhs.nl/ bergwandelen/onderzoek.htm, 2006, [Online; accessed 01-November-2014].

[8] F. Prince, H. Corriveau, R. Hébert, and D. A. Winter, "Gait in the elderly," *Gait & Posture*, vol. 5, no. 2, pp. 128–135, 1997.

[9] M. Brandstater, H. de Bruin, C. Gowland, and B. Clark, "Hemiplegic gait: analysis of temporal variables." *Archives of physical medicine and rehabilitation*, vol. 64, no. 12, p. 583, 1983.

[10] M. A. Wong, S. Simon, and R. A. Olshen, "Statistical analysis of gait patterns of persons with cerebral palsy," *Statistics in medicine*, vol. 2, no. 3, pp. 345–354, 1983.

[11] S. A. Wilson, P. D. McCann, R. S. Gotlin, H. Ramakrishnan, M. E. Wootten, and J. N. Insall, "Comprehensive gait analysis in posterior-stabilized knee arthroplasty," *The Journal of arthroplasty*, vol. 11, no. 4, pp. 359–367, 1996.

[12] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto, "Identifying users of portable devices from gait pattern with accelerometers," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2. IEEE, 2005, pp. ii–973.

[13] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[14] D. C. Kerrigan, M. K. Todd, U. Della Croce, L. A. Lipsitz, and J. J. Collins, "Biomechanical gait alterations independent of speed in the healthy elderly: evidence for specific limiting impairments," *Archives of physical medicine and rehabilitation*, vol. 79, no. 3, pp. 317–322, 1998.

[15] J. M. Hausdorff, D. A. Rios, and H. K. Edelberg, "Gait variability and fall risk in community-living older adults: a 1-year prospective study," *Archives of physical medicine and rehabilitation*, vol. 82, no. 8, pp. 1050–1056, 2001.

[16] G. Gioftsos and D. Grieve, "The use of neural networks to recognize patterns of human movement: gait patterns," *Clinical Biomechanics*, vol. 10, no. 4, pp. 179–183, 1995.

[17] S. Mulroy, J. Gronley, W. Weiss, C. Newsam, and J. Perry, "Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke," *Gait & posture*, vol. 18, no. 1, pp. 114–125, 2003.

[18] R. K. Begg, M. Palaniswami, and B. Owen, "Support vector machines for automated gait classification," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 5, pp. 828–838, 2005.

[19] V. Moniz-Pereira, F. Carnide, F. Ramalho, H. Andre, M. Machado, R. Santos-Rocha, and A. P. Veloso, "Using a multifactorial approach to determine fall risk profiles in portuguese older adults," *Acta reumatologica portuguesa*, vol. 38, no. 4, pp. 263–272, 2013.

[20] V. Moniz-Pereira, S. Cabral, F. Carnide, and A. P. Veloso, "Sensitivity of joint kinematics and kinetics to different pose estimation algorithms and joint constraints in the elderly," *Journal of applied biomechanics*, vol. 30, no. 3, pp. 446–460, June 2014.

[21] H. H. P. da Silva, "Feature selection in pattern recognition systems," Master's thesis, Instituto Superior Tecnico, 2007.

[22] A. J. Ferreira, "Feature selection and discretization for high-dimensional data," Ph.D. dissertation, Instituto Superior Tecnico, 2013.

[23] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.

[24] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[25] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[26] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 835–850, 2005.

[27] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci*, vol. 1, pp. 801–804, 1956.

[28] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. California, USA, 1967, pp. 281–297.

[29] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[30] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *J Bioinform Comput Biol*, 2003, pp. 523–529.

[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, Feb. 1979.

[32] J. C. Dunn, "Well separated clusters and optimal fuzzy-partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.

[33] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

[34] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a dataset," 2001.

[35] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Sociète Vaudense des Sciences Naturelles*, vol. 44, pp. 223–270, 1908.

[36] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010.

[37] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[38] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, J. Kittler and F. Roli, Eds. Springer Berlin Heidelberg, 2001, vol. 2096, pp. 309–318.

[39] D. T. Larose, "k-nearest neighbor algorithm," *Discovering Knowledge in Data: An Introduction to Data Mining*, pp. 90–106, 2005.

[40] P. DeVita and T. Hortobagyi, "Age causes a redistribution of joint torques and powers during gait," *Journal of applied physiology*, vol. 88, no. 5, pp. 1804–1811, 2000.

[41] J. M. Hausdorff, S. L. Mitchell, R. Firtion, C. Peng, M. E. Cudkowicz, J. Y. Wei, and A. L. Goldberger, "Altered fractal dynamics of gait: reduced stride-interval correlations with aging and huntington's disease," *Journal of applied physiology*, vol. 82, no. 1, pp. 262–269, 1997.

[42] A. Novak and B. Brouwer, "Sagittal and frontal lower limb joint moments during stair ascent and descent in young and older adults," *Gait & posture*, vol. 33, no. 1, pp. 54–60, 2011.

[43] M. A. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, March 2002.