Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Using Accelerometer Data and Machine Learning as Support for Clinical Studies

by

Andrea Fasciglione

Theses Series **DIBRIS-TH-2024-XX**

DIBRIS, Università di Genova
Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,

Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum

Using Accelerometer Data and Machine Learning as Support for Clinical Studies

by

Andrea Fasciglione

May, 2024

Dottorato di Ricerca in Informatica ed Ingegneria dei Sistemi
Indirizzo Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
http://www.dibris.unige.it/

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum
(S.S.D. INF/01)

Submitted by Andrea Fasciglione
DIBRIS, Univ. di Genova
Date of submission: May 2024

Title: Using Accelerometer Data and Machine Learning as Support for Clinical Studies

Advisor: Maurizio Leotta, Alessandro Verri

Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova
Ext. Reviewers: Giosuè Lo Bosco, Paolo Napoletano

Abstract

Recent improvements in technologies used in the medical & health field have significantly supported specialists in analyzing and studying diseases. Detecting daily life activities with the help of wearable devices might become a precise and objective method for remote patient monitoring.

The goal of this thesis is to investigate the use of data obtained from wearable devices and the application of Machine Learning methods as support for clinical studies. Our primary task is to classify activities of daily life focusing, in our case, on using accelerometer data.

Thanks to the study of the state-of-the-art, we noted three main areas in which to contribute: (1) public datasets, (2) methodology, and (3) reproducibility.

Regarding the first area, our contribution has been to record, prepare, and share two datasets of accelerometer data including daily life activities. The first dataset was recorded with healthy adults, the second involves children and adolescents suffering from chronic diseases, such as Juvenile Fibromyalgia, Juvenile Idiopathic Arthritis, and Juvenile Dermatomyositis.

Concerning the second area, our main effort has been in proposing a method that aids the classification of non-interesting activities for specialists, reducing the impact on the classification of known activities, and testing different approaches to improve this outcome.

With respect to the third area, we devised a review that analyzes the situation of related works, with a focus on sharing the used datasets and biometric information concerning the subjects included in the studies.

Table of Contents

Chapter 1

Introduction

In the medical & health field, in recent years there has been a widespread use of technologies to support specialists in analysis and prevention. Many successful studies have enabled achievements that were hardly feasible in the past. Practices, calculations, and analysis developed in recent years have not replaced the role of humans and specialists in decision-making, rather, they have become supportive tools to facilitate diagnoses, find objective measures, and standardize procedures.

Most of these developments have been made possible by factors in the field of technology and computer science that have been decisive in recent years. Among various factors that have allowed the results and testing of the new approaches described in this work, we can mention:

1. The miniaturization of hardware components, sensors, and consequently devices. Additionally, the reduction in the costs of these components has made it possible to produce wearable sensors with high precision capable of recording for different days data with high sensitivity, all within the size of a wristwatch;

2. The increased capacity and speed of computer processing, which has made it possible to analyze large amounts of data that were unmanageable until a few years ago. This, as a consequence, has enabled the development of data analysis and machine learning techniques, that allowed us to *learn* from data, also by finding features in these data that would be difficult to discover otherwise. All of this has also led the entire scientific community, among different study fields, to change approaches in problem-solving, starting from data to analyze the context and potential solutions, instead of starting from the study of the problem itself.

The entire work described here has originated and progressed thanks to the collaboration of the Software Engineering for Healthcare (SEH) Laboratory, at the University of Genova (Italy), involving two main partners. The first is the Janssen Italia company (previously Actelion Italia).

The second is a department of Istituto Giannina Gaslini, Italy, dedicated to the treatment of rheumatological diseases in children and adolescents. Specifically, the former dealt with treatments and care for individuals affected by diseases such as Pulmonary Hypertension (PAH), or Multiple Sclerosis (MS). The primary interest was to investigate the possibility of using wearable devices in a long-term scenario, to track the progression of a disease treatment. This was particularly relevant considering that the diseases under study led to easy fatigue and difficulties in daily life activities. In the latter case, there was a desire to leverage the precision and feasibility of wearable devices as support for studying rheumatological diseases. Even in this case, considering diseases like Juvenile Fibromyalgia, Juvenile Idiopathic Arthritis, and Juvenile Dermatomyositis, we planned to use wearable devices to better study the impact of symptoms on daily life activities.

The starting point and the ultimate goal of this work were, therefore, born from requests made by specialists in the medical and pharmaceutical field, and physicians themselves. The possibility of having close collaboration with both groups allowed for direct input from specialists and the opportunity to look for feedback on the work.

We can affirm that **the underlying final goal of this entire project was to investigate the use of data obtained from wearable devices and the application of machine learning methods as support for clinical studies**. To achieve this objective, we reasoned on which technologies and methodologies to study, choosing based on the problems submitted by our partners, and on the state of the art available at the beginning of the work. We opted to rely on the use of accelerometers as sensors and to employ machine learning techniques to recognize daily life activities and subsequently extract information useful to the specialists.

Practical tests and actual implementations were preceded by a study of the state of the art in the field of our primary task, that is to say, classifying activities (of daily life), a task known in the literature as *activity recognition* (AR), focusing, in our case, on using accelerometer data. The study of AR seemed, at first glance, to be a non innovative and non-exclusive topic even at the beginning of this work: literature already contained works on the problem claiming optimal results.

However, this fundamental initial phase allowed us to identify some issues in the works already presented in the literature and provided ideas for what we have subsequently done. We noted three main areas in which to contribute: (1) public datasets, (2) methodology, and (3) reproducibility.

1. **Datasets for Activity Recognition**
   Regarding the first area, about *datasets for activity recognition*, it was observed that generally publicly available datasets obtained from accelerometers for performing AR tasks were limited and included activities that were both too generic and not useful for our case study (e.g. limited to walking, running, standing, sitting, lying down). Taking into account our collaboration with Istituto Gaslini, dedicated to the cure of chronic diseases in children

and adolescents, we also considered the number of datasets where the population included non-adults; such datasets are extremely rare in the literature. Our contribution in this area was the recording, preparation, and sharing of two datasets. The first dataset was recorded with healthy adults using wearable devices with accelerometers. The second dataset, on the other hand, was recorded in collaboration with children and adolescents, including those affected by chronic diseases, such as Juvenile Fibromyalgia, Juvenile Idiopathic Arthritis, and Juvenile Dermatomyositis. The work carried out in this part of the thesis, has led to the publication of 2 articles: the first one, regarding healthy adults, has been presented during the *Pattern Recognition ICPR International Workshops and Challenges* [LFV21a]; the second one, regarding children and adolescents, is currently under review to be published in the *IEEE Journal of Biomedical and Health Informatics* [FLV$^+$].

2. **Activity Recognition: Methods & Improvements**
   Concerning the second area, namely *activity recognition: methods & improvements*, it was noted that in most works on AR with wearable devices data, approaches often provided to the scientific community were impractical. In this case, we specifically refer to the fact that the recognition phase was limited to only dealing with the learned activities. However, subjects usually perform a lot of different activities throughout the day. These approaches assumed the absence of an "open world", by thinking that only the known activities needed to be recognized by the classifier. In this part, our main contribution was to propose a method that aids the classification of non-interesting activities for specialists, reducing the impact on the classification of known activities. The outcomes achieved in this part of the thesis resulted in the publication of one article, published in the *IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises* [FLV21].

3. **Reproducibility in Activity Recognition**
   Finally, the third part, on *reproducibility in activity recognition*, is closely connected to the first two parts. In the initial stages of the state-of-the-art study, we noticed that a high percentage of the found works did not provide sufficient data to entirely replicate the results published in the works themselves. In this case, we refer to the lack of used datasets, code used for data processing or classification, etc. This problem obviously makes it difficult to replicate a published work and compare its results. Our contribution regarding reproducibility is a review that analyzes the situation in relation to sharing the used datasets and biometric information concerning the subjects included in the studies (e.g. weight, height, age, male/female, BMI). The activities conducted within this part of the thesis have culminated in the writing of one article, released in *IEEE International Conference on Systems, Man, and Cybernetics* [FLV22].

The structure of this thesis will then follow the three main topics addressed in this work: for each different covered area, there will be a part of the document. In order, after starting with

Part I describing background topics, there will follow Part II dedicated to datasets in activity recognition. Successively, Part III will concern methods and improvements achieved, and the final Part IV on reproducibility in activity recognition.

Part I

Background

Chapter 2

Background

Introduction

In this chapter, there will be listed a few concepts to have background knowledge to better understand topics treated in the thesis. To begin with, the working principles of accelerometer sensors will be described. Then, we will move to a description of *activity recognition*. Successively, there will be a brief description of the main classification methods considered in this thesis. The last topic will be linked to the *frequency* domain: Kalman Filter.

2.1 Accelerometer Sensors

Accelerometers are sensors commonly placed in different applications, such as smartphones, drones, fitness trackers, or, in general, wearable devices. Their general role is to measure the acceleration that they are subject to. The most common kind of these sensors is usually defined as *MEMS* accelerometers, standing for *Microelectromechanical Systems*, indicating the very small scale of these devices. Figure 2.1 represents an example of this sensor. In particular, there is a microscopic suspended mass (in green), attached to springs (in black), usually made in silicon. Near the suspended mass, there are fixed plates with electrodes (in blue).

When the accelerometer undergoes acceleration, in any direction, the mass resists the motion due to its inertia and the distance between the mass and the electrodes will change. Thus, the change in capacitance is measured with the electrodes and converted into data. When dealing with three-axial accelerometers, each axis has typically its own suspended mass and electrodes.

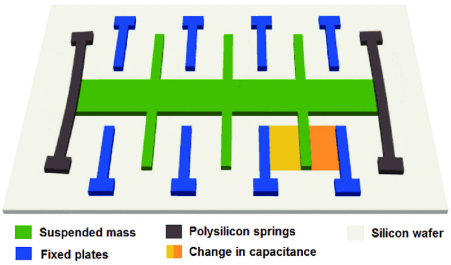

Figure 2.1: Example of a MEMS accelerometer.

2.2 Activity Recognition

Activity Recognition is the study of understanding, measuring, and classifying the physical actions that are performed by individuals, starting from data collected from various sensors.

The data sources can be diverse: ambient sensors (e.g., cameras, microphones, smart home devices) [GNO22, RMS17, SMVS01], or wearable sensors (e.g., accelerometers, gyroscopes, magnetometers, included in smartwatches, or fitness trackers) [TDF+18, JNJSS18].

In the most recent years, the availability of data from different sources made it possible to use activity recognition in different fields. For instance, in the surveillance field, by detecting suspicious activity patterns [PSW19]. The field of sports and fitness has been affected by this, for tracking workout types, or to analyze performances and obtain personalized coaching [HYCL18, NJK20, ABMP+10]. As a last example, and strictly correlated to the previous one, the healthcare field has tested many possible approaches based on activity recognition: different studies tested AR for fall detection, for monitoring patient mobility, or for rehabilitation purposes [TFM+18, GDC+16, TDF+18, JNJSS18].

One of the main challenges regarding activity recognition is privacy concerns. Indeed, it is difficult to balance AR benefits with ethical data usage, but it is a crucial aspect. In our work, we have chosen to rely on wearable devices, because, among the different positive aspects, these devices can perform well preserving, at the same time, final users' privacy.

2.3 Classification

The task of *classification* in the field of machine learning involves determining the category (i.e., also known as *class*) to which a new unseen sample belongs. To do this, the process needs a prior training set, containing labeled samples, for which the category is known. Generally, classification problem can be divided into two different cases: *binary classification*, where only two categories are involved, and *multiclass classification*, where the problem consists in assigning a sample to one of several (i.e., more than 2) classes).

Developed classification methods have been generally designed for binary classification, therefore, when dealing with multiclass cases(i.e., the number of classes is N), two approaches can be applied:

- *one-vs-one*: during training phase, N*(N-1)/2 binary classifiers are trained. Each classifier focuses on discriminating between a specific pair of classes, treating one class as positive (+1) and the other as negative (-1). Essentially, a unique classifier exists for each pair of possible classes.
  During the prediction phase, an unseen sample is presented to all OvO classifiers. The class associated with the classifier producing the highest number of positive (+1) labels becomes the predicted outcome

- *one-vs-all*: at training time, N individual classifiers are constructed, one for each class. For classifier C_i, instances belonging to the *i-th* class are labeled positive (+1), while all others are set as negative (-1).
  Classifiers, in this case, will typically produce real-valued confidence scores during predictions. For a new unseen sample, the class associated with the classifier returning the highest confidence score is chosen as the final prediction.

2.3.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification and regression purposes. For classification purposes, it can be simplified to the idea of having samples of two classes, to find the hyperplane that best divides these classes, as shown in Figure 2.2, where the line represents the hyperplane. The term *support vectors* refers to those points that are the nearest ones to the hyperplane; these points have a key role in the dataset and in the model, since moving (or removing) them would change the position of the hyperplane.

We have that the class of our data instances can be positive (+1) or negative (-1); our dataset can be described as a set of *n* pairs (x_i, y_i), where x_i is the data point and y_i is the associated label.

From this, we can define the hyperplane as

$$\{x : f(x) = w^T x + b = 0\}$$

where the term w is the *weight vector* (controlling the orientation of the hyperplane), while b is the *bias* term (controlling the distance of the hyperplane from the origin). We have that the sign of the discriminant function $f(x)$ for a given point will determine the side of the hyperplane on which that point will be [BHW10].

By having that:
- x_+ (or x_-) is the closest point to the hyperplane with the positive (or negative) label

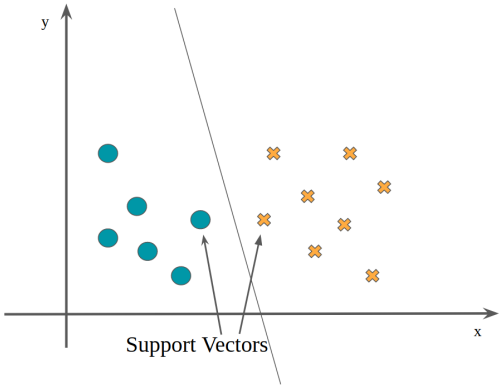

Figure 2.2: Example of a hyperplane dividing two classes

- $||w||$ is the *norm* of a vector - a *unit vector* $\hat{w}$ in the direction of w is defined as $w/||w||$ and has norm $||\hat{w}||$ = 1 and assuming that x_+ and x_- are equidistant from the hyperplane, setting the distance equal to 1 and dividing by $||w||$ we can define the margin *m* of a hyperplane *h* to be:

$$m(f) = \frac{1}{2}\hat{w}^T(x_+ - x_-) = \frac{1}{||w||}.$$

At this point, we can look for that hyperplane for which we have the maximum margin $1/||w||$ (or equivalently the minimum $||w||^2$). We can maximize the margin and consider also *outliers* (i.e., instances in the margin), also reducing the possibility of overfitting, by introducing in the constraint the *margin error* ϵ:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_i^n \epsilon_i$$

$$\text{subject to: } y_i(w^T x_i + b) \geq 1 - \epsilon_i; \epsilon_i \geq 0, i = 1, \ldots, n$$

where, by setting $0 \leq \epsilon_i \leq 1$ we allow instances to be in the margin, having $\epsilon_i \geq 1$ we allow to have misclassified instances, and the constant C > 0 controls the importance of maximizing the margin and minimizing the errors.

2.3.2 Decision Trees and Random Forests

Decision trees are a quite common and powerful classification algorithm, that uses a tree-like structure to predict labels of samples. The trees are structured such that each *internal node* represents a feature of the samples, and tests the feature value against a threshold; each *leaf node* represents a class prediction.

During the training phase, the goal is to create the tree starting from the entire dataset in a single root node. Then, the procedure searches for the best feature to split the dataset. The criterion

on which the dataset is split is the *information gain* measure, which measures the decrease in uncertainty about the *target class* after splitting the dataset on the chosen feature. The splitting step is then repeated until a stopping criterion is satisfied.

Concerning the prediction phase, each new sample is presented at nodes, starting from the root, and the feature value of the sample is tested versus the node threshold. Then, according to the test result, the procedure follows the branch corresponding to the test result. Once a leaf node is reached, the predicted class can be returned.

Random forest is a strategy based on decision trees and on the *bootstrapping* technique. Bootstrapping is a statistical technique that involves resampling a dataset with replacement. The procedure trains N decision tree models, using a bootstrapping technique to sample subsets from the entire dataset. Each subset works as training data for each individual tree. Since the random forest returns, for each new data point, a set of N predictions, the final one is chosen using a majority voting approach.

2.3.3 k-Nearest Neighbors

K-Nearest Neighbors (k-NN) is a *supervised learning algorithm* and is based on the idea of classifying new samples according to the similarity of its *k* nearest neighbors present in the training set. For each new unseen label, it is computed its *distance* with each training point. Then, among the training points, only the first *k* closest neighbors are considered. The final predicted label, assigned to the new sample will correspond to the most frequent class label in the *k* neighbors.

Different distance measures can be used in this approach, for instance: *Euclidean distance*, which calculates the straight-line distance between two given points, or the *Manhattan distance*, computed as the sum of absolute differences of coordinate values.

The value *k* is considered as a regularization parameter; having small *k* will generate a model that is more susceptible to noise and possible outliers, whilst a larger *k* will lead to smooth decision boundary, but possibly losing details.

2.3.4 Gaussian Naive Bayes

The Gaussian Naive Bayes classifier is considered a probabilistic approach for classification purposes, based on the idea that there is a strong independence between features. As the name suggests, it is based on Bayes' theorem. In fact, when classifying a new data point *x* belonging to class C, this approach estimates the probability of *x* belonging to each class C_i, using Bayes' theorem:

$$P(C_i|x) = P(x|C_i) * \frac{P(C_i)}{P(x)}.$$

where $P(C_i)$ is the prior probability of class C_i, and $P(x)$ is quite irrelevant in this particular case. For each class, $P(x|C_i)$ is estimated using *Gaussian distributions* for each feature, therefore assuming that features of a certain class follow a normal distribution. In the end, the class that obtained the highest $P(C_i|x)$ is returned as the prediction.

2.4 Kalman Filter

Kalman filter is a well-known algorithm that by observing a set of measures over time (e.g. statistical noise, uncertainty) produces estimates of unknown variables (i.e. states of a system). These estimates are generally more accurate than those based on a single measurement alone. In general, to understand how it works, we could distinguish two main steps: *prediction* and *update*.

In the *prediction* step, the filter uses the previous estimated state and its uncertainty to predict the next state of the system. In particular, a mathematical model is applied in order to project how the state is expected to change over time.

In the *update* step, instead, when a new measurement becomes available, the filter compares this new value with the predicted state to correct the estimated state. To do so, the algorithm considers the uncertainty of the measurements and the predicted state. This particular step is helpful since it is used to reduce the effects of noise or inaccuracies in the measurements.

The remarkable ability of the Kalman filter is that step-by-step it automatically adjusts the balance and the relevance between the predictions and measurements based on their respective uncertainties. To resume, it is a useful way to reduce noise in a signal, as in the case of data recorded with accelerometer sensors.

Part II

Datasets for Activity Recognition

Introduction to Datasets Part

In this part of the thesis, we are going to present the topic regarding *datasets*. Our contribution, in this scope, has been to record, prepare, and share two datasets. The first dataset has been recorded with healthy adults, performing daily living activities, using wearable devices equipped with accelerometers. The second dataset, instead, has been recorded cooperating with children and adolescents, involving subjects with three chronic diseases (i.e. Juvenile Fibromyalgia, Juvenile Idiopathic Arthritis, and Juvenile Dermatomyositis). The choice to create these datasets lies in the lack of publicly available content for our kind of applications (i.e. with accelerometer data, with quite complex activities, and/or with children/adolescents), as deduced from the state-of-the-art review.

To begin with, Chapter 3 introduces the devices used to record our datasets (i.e. Actigraph GT9X Link, Actigraph Centrepoint Insight, Empatica Embrace 2), with technical details and specifications. Chapter 4 presents the *Healthy Adults Dataset*, with its characteristics, recording procedures, and few results obtained by other researchers thanks to it. Lastly, Chapter 5 illustrates the *Gaslini Juvenile Dataset*, presenting a short survey about activities usually analyzed in youth-related AR studies, characteristics of the dataset, and the method used to aggregate clinical information of participants to create a score of diseases' impact.

Chapter 3

Used Devices

3.1 Actigraph GT9X Link

In this Section, we provide an overview of the first of three devices employed in this work. This is a medical *actigraphy* tool proposed by the ActiGraph corporation[1], one of the leading provider of medical-grade wearable activity and sleep monitoring solutions.

In the last years, ActiGraph corporation proposed several actigraphy tools (e.g., ActiGraph wGT3X-BT, Actigraph Centrepoint Insight Watch, Actigraph Leap). Here we focus on the ActiGraph GT9X Link device (GT9X from hereafter). In the next section, we will present the *Centrepoint Insight Watch*, that we have been using as well for our work.
Concerning this device, two pieces are fundamental to using the device and recording data, more in detail:

- GT9X, the device itself, which is an activity monitor combining various sensors like two accelerometers, a gyroscope, and a magnetometer to capture position and rotation data for advanced applications. It can be worn on the wrist (as a watch), or can be mounted on a clip support, to place it at the height of the waist.
- ActiLife, which is ActiGraph's data analysis software platform. This software, starting from data recorded with the device is able to evaluate different statistics about activities and sleep. It is necessary to use this software to download data from the GT9X.

GT9X Link (see Fig. 3.1) records high-resolution raw data, which is converted using ActiLife into a variety of objective activity and sleep measures using publicly available algorithms developed and validated by members of the academic research community.

[1] https://theactigraph.com/

Figure 3.1: ActiGraph GT9X Link

Available measures retrievable from ActiLife include raw data about acceleration, rotational velocity, and direction with respect to the earth's magnetic field. Other derived metrics are accessible, such as activity counts, energy expenditure, steps taken, physical activity intensity, and sleep efficiency.

GT9X houses different sensors, and according to the choice to use or not these sensors to record data, the battery could last from 14 days to 1 day. In the following the list of available sensors:

1. **Primary accelerometer**
   The *Primary accelerometer* is a 3-axis accelerometer. Fig. 3.2 shows the expected static acceleration for each device orientation with respect to gravity. The primary accelerometer has a dynamic range of ± 8 g (g = 9.81 m/s^2) and a sensitivity of 4 *mg/LSB* (least significant bit). Primary accelerometer data can be collected at a sample rate in the 30-100 Hz interval.

2. **Inertial Measurement Unit (IMU)**
   In addition to the Primary Accelerometer, the GT9X device houses an *IMU* unit (Inertial Measurement Unit), i.e. an electronic chip to capture position and rotation data for more advanced analyses. The IMU data are collected at a 100 Hz fixed sample rate. In detail, the IMU sensor contains an accelerometer, a magnetometer, a temperature sensor, and a gyroscope.

3. **Wear time sensor**
   The GT9X contains also a sensor used to detect if the device has been removed and is no longer worn on the wrist. This feature is commonly used for monitoring the usage of the device by patients and for data cleaning.

Once data have been recorded, it is possible to download raw data of the accelerometer by exporting .csv files (Fig. 3.3). These files are structured with these main contents:

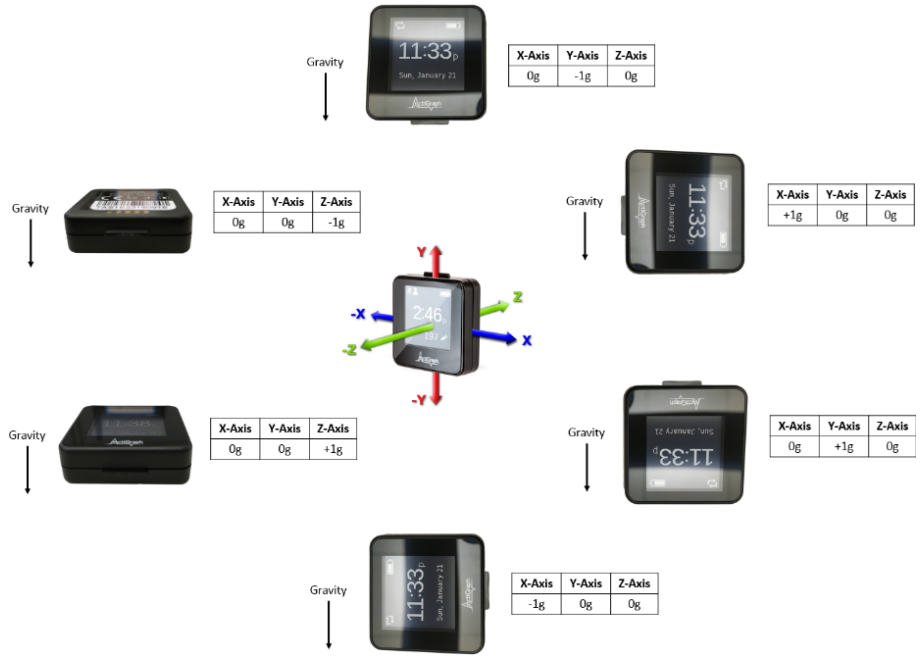

Figure 3.2: ActiGraph GT9X Link axis device orientation

Timestamp	Accelerometer			Temperature	Gyroscope			Magnetometer		
	X	Y	Z		X	Y	Z	X	Y	Z
2018-01-26T12:51:00.0000000	0,303711	-0,092285	0,933594	36,362267	-0,976563	1,770020	3,356934	4,980469	39,843748	33,691405
2018-01-26T12:51:00.0100000	0,303711	-0,099609	0,938477	36,362267	0,000000	2,136231	3,051758	4,980469	39,843748	33,691405
2018-01-26T12:51:00.0200000	0,302246	-0,100098	0,951660	36,356277	1,770020	2,380371	2,868653	4,980469	39,843748	33,691405
2018-01-26T12:51:00.0300000	0,297363	-0,087402	0,966797	36,359272	2,746582	2,441406	2,624512	4,980469	39,843748	33,691405
2018-01-26T12:51:00.0400000	0,296875	-0,090332	0,971680	36,371253	2,441406	2,075195	2,563477	4,980469	39,843748	33,691405
2018-01-26T12:51:00.0500000	0,293945	-0,100098	0,970215	36,359272	2,502442	1,342774	2,380371	4,980469	39,843748	33,691405

Figure 3.3: Example of IMU Raw data (recorded at 100 Hz)

- *Timestamp*. The timestamps shown in the exported .csv file are formatted to show the full date in DD/MM/YYYY format (i.e., date formatting is specific to the local settings of the used computer). The time is formatted as hh:mm:ss.sss where the last "sss" represents the fractional portion of the timestamp.
- *Accelerometer X, Y, Z*. Each accelerometer entry represents instantaneous acceleration for the axis indicated in units of gravity (g = 9.81 m/s^2).
- *Gyroscope X, Y, Z*. Gyroscope measurements from the IMU are presented in degrees/sec for each axis.
- *Magnetometer X, Y, Z*. The IMU magnetometer readings represent the magnetic field expe-

rienced by the Link device and are useful for discerning directional orientation (compass). These readings are in μT (micro Tesla)

- *IMU Temperature*. The IMU temperature reading indicates the temperature inside of the Link activity monitor (i.e., it is not suggested to consider it as the body temperature of the subject) and is indicated in degrees Celsius (C).

3.2 Actigraph Centrepoint Insight

This Section is about the second device considered for our study, offered by ActiGraph corporation: Centrepoint Insight Watch (CPIW), in Figure 3.4. This device is more up-to-date and innovative compared to GT9X. If in the case of GT9X, we needed ActiLife software to retrieve recorded data, in this case, the company chose to rely on an online platform: the *Centrepoint* platform [2].

Figure 3.4: ActiGraph Centrepoint Insight Watch

This platform is based on a cloud system and allows specialists to collect, process, and manage data from clinical investigations, with data organized by participants and made easily accessible. In our case, that was to mainly use the device to obtain precise raw data, we have been using the platform to initialize devices and to download data of interest. In particular, it was needed to set up the device (attached to a computer via USB cable) to let it start recording data. Once

[2]https://theactigraph.com/centrepoint

the recording phase was terminated, it was required to upload recorded data on the cloud system. As a final step, we could request data by querying the platform with a particular date range (i.e., the one in which data have been recorded), to get .csv files containing data structured as those described for GT9X. Data collection could be possible in a remote way as well by using additional instruments (i.e., CentrePoint Data Hub, a communication gateway), but it was not applicable in our particular scenario since these instruments were not available for us. Remote data collection, when used by physicians, reduces the number of visits in person and on-site, mitigating patients' stress and recording meaningful data in real-life scenarios.

We did not test other features of the CentrePoint platform, but it is useful for other purposes, such as accessing participant adherence to a study. It is possible, in fact, to check nearly in real-time wearing periods of each participant, remotely.

With respect to GT9X, the CPIW device houses only one accelerometer. This sensor can record data in an equal dynamic range (± 8 g), but the sample rate could be set from 32 Hz up to 256 Hz. According to ActiGraph, considering the battery life and the data storage, with this particular device we could record high-resolution data for up to 30 days continuously. Comparing CPIW with GT9X, with this newer device it was possible to wear it only on a wrist, with the help of a wristband.

3.3 Empatica Embrace 2

This section is dedicated to the third device involved in our study. Here, we will present this last instrument used to record data with accelerometers: Empatica Embrace 2[3] (in Figure 3.5). In this case, the components that allowed us to record data with this device are three, presented hereby.

- The *Embrace 2* itself, which is a device designed as a support tool for epilepsy. As a matter of fact, Embrace 2 is the only FDA-cleared wrist-worn wearable for this disorder. This device can recognize possible convulsive seizures, to immediately alert caregivers and allow users to get help when needed. The device is equipped with an accelerometer and gyroscope; other sensors can measure electrodermal activity (EDA) and temperature. It has been devised to be worn only on the wrist of subjects, making use of a wristband, even for juvenile subjects.
- The *Mate for Embrace Watch* smartphone app. This was fundamental to making the device work. In fact, Embrace 2 needs to be paired with a smartphone using this app, via Bluetooth. The Mate app, once logged in with the given credentials, can initialize the device to record data. Once in a while, the device sends recorded data to the Mate App; when

[3] https://www.empatica.com/en-eu/embrace2/

an internet connection is available, the Mate App transmits chunks of recorded data to an online portal.

- The Empatica Research online portal, where it is possible to generate credentials for users to be used with Mate App and to retrieve recorded data uploaded on the associated cloud system.

Moving to technical details, Embrace 2 has a battery that can record data for up to 48 hours, and, with 30 minutes of charging, it can get energy for an entire day of recordings. Regarding the cited sensors (i.e., accelerometer, gyroscope, electrodermal activity, and temperature) we could not configure the sampling frequency, nor which sensors to be activated or not. The accelerometer sampled data at 32 Hz, the sensor for electrodermal activity at 5 Hz, and the thermometer at 1 Hz. With our devices, we could not retrieve gyroscope data.

Figure 3.5: Empatica Embrace 2 - Seizure monitor

While testing the capabilities of this device during our studies, we noticed a weird behavior regarding the number of samples recorded per second. Among three different sessions of recording (each one was exactly 22 minutes long), we obtained three different numbers of samples from the accelerometer sensor. Knowing that the device was recording at a sampling frequency of 32 Hz for 22 minutes, we were expecting a total of 42240 samples (32 Hz x 22 min x 60 sec). With our three sessions, we have achieved 41931, 41924, and 42038 samples respectively; all of the three values were lower than the expected one (42240). This problem could be judged quite irrelevant since it was considered as losing only 10 seconds of recording over 22 minutes. However, this outcome is completely different with respect to used ActiGraph devices, where the expected number of samples was consistently equal to the number of samples recorded with the device.

Embrace technical support, which has been contacted for this issue, replied that “[...] Devices leverage an internal time reference to know when to sample the data. This internal time reference has its precision and stability. Moreover, the temperature can impact the stability of the clock”.

Chapter 4

Healthy Adults Dataset

4.1 Overview

In this Chapter, it will be described the first of two datasets that we have recorded and publicly shared. As aforementioned in the Introduction, in fact, we have recorded two datasets: one with healthy adults (i.e., the one illustrated in this chapter), and another one regarding juvenile subjects, including children and adolescents suffering from chronic diseases. We chose to give priority to recording this dataset with healthy adults in order to familiarize ourselves with the kind of data and with the recording protocol. To explain this choice, it is easy to imagine that registering data with adults is easier than dealing with children (potentially with mobility impairments). Moreover, adults perform movements during the activities of daily living in a more constant and controlled manner with respect to children. Thus, data related to adults were, possibly, easier to analyze for our purposes.

To better understand the aims of this first dataset, it is useful to remark that in the last years there has been an increasing interest in finding reliable methods for monitoring patients suffering from different diseases (or simply elderly), in particular using remote and non-intrusive methods (e.g. [KAMP14, PTP+06, SLES11]). Many diseases, in fact, strongly impact daily life activities because of their effects (e.g., Multiple Sclerosis, Pulmonary Arterial Hypertension, or Parkinson's). A possible way to assess the physical condition of a patient is based on assessing how much time is spent on specific activities and how the amount of time dedicated to each activity changes over time (e.g., a subject can decide to stop vacuuming because of the feeling of fatigue, or could need more time to eat because of hand tremors).

With the goal of working on a possible method to create an activity recognition classifier, starting from data recorded by wearable devices, we found the need to create a novel dataset regarding daily life activities. Indeed, by analyzing the state of the art concerning the publicly available datasets recorded with wearable devices, we noticed that despite the large interest in this topic,

there was a lack of datasets having, at the same time, the following characteristics: (1) containing data of numerous and different daily life activities, (2) containing data recorded using high-quality sensors (both concerning frequency and accuracy), and (3) containing data from different synchronized devices positioned on different parts of the body. From the point of view of a researcher, this lack could become an obstacle to performing more in-depth investigations and conceiving more advanced approaches to the problem of activity recognition using wearable devices.

For this reason, by creating this dataset, we have given our contribution, presenting and publicly releasing a novel dataset meeting the three characteristics mentioned above.

Concerning the first characteristic, our healthy adults' dataset includes 17 different daily life activities performed in real-life scenarios (e.g., eating, using a laptop, handwriting, vacuuming, walking, going upstairs/downstairs). On the contrary, datasets that were available in the literature often included a few activities (often 8 or less) or activities that are limited to a particular context (e.g., cooking, breakfast morning-routine, or walking at different speeds only) [CSC+13, DlTHB+09, AGO+13, SSH13].

Concerning the second characteristic, we recorded our dataset using professional devices produced by Actigraph. These devices are medical-grade activity monitors that thanks to their characteristics and to their reliability, have been broadly considered in different studies and for different purposes (e.g. [JF12], [RT12]). More in detail, we have used two Actigraph GT9X Link, with a sampling frequency of 100 Hz, and one Actigraph Centrepoint, with a sampling frequency of 256 Hz, both presented in detail in 3. This is an important characteristic of our dataset since, very often, the available datasets contain data recorded with low-cost (and precision) sensors like the ones included in Android smartphones, with non-certified accuracy of the values provided [MMN17, SS16, SSH13]. The high frequency of these devices can be useful for data scientists in order to perform, moreover, complete analysis by re-sampling data to different frequencies.

Another strength of our dataset (as a third characteristic) lies in the fact that we have placed the three aforementioned devices, synchronized together, on three different parts of the participants' bodies: the dominant wrist, right side of the hip, and right ankle (on the external side). Having synchronized data coming from different parts of the body, like in our dataset, would allow researchers to find methods based on the correlation of these data, thus creating more accurate activity analysis or recognition approaches. On the contrary, other available datasets were usually focused on only one part of the body (e.g. pockets or the hip) [MMN17, ZS12]. It is useful to reason that, regarding this topic, wearing more devices could be quite annoying for a subject in real settings: the trade-off between having more data and creating discomfort for the patient should be carefully analyzed when planning recording protocols.

4.2 Dataset Characteristics

The creation of the *Healthy Adults* dataset has been performed in three main phases: (1) Data Recording, (2) Data Extraction, and (3) Data Labeling and Cleaning. The final output is a labeled dataset containing the raw data of 17 daily living activities ready to be used by researchers for a variety of possible studies. In the following of this section, we describe the three phases in detail.

4.2.1 Data Recording

To record our dataset, we followed a precise protocol that we defined and that had been reviewed and approved by the ethical committee of the Department in which the data recordings took place. Apart from the operational details on the procedures to follow during data recordings, our protocol includes also a step where each participant is asked to sign an informed consent: this allows us to share, with the research community, all the data recorded and the physical biometric characteristic (e.g. age, height, weight, dominant hand) of each participant. The dataset is currently available online on Harvard Dataverse website[1]. Having our dataset hosted on such a dataset repository also allows it to be indexed by specific search engines (e.g. Google Dataset Search).

Participant inclusion/exclusion criteria. The former includes the following: (1) to be able to perform the requested actions, (2) age of 18 and over, and (3) to understand the purpose of the study and be willing to participate in the study. On the contrary, exclusion criteria include any planned surgery or procedures that would interfere with the conduct of the study and any major mobility difficulties.

Participants Characteristics. The dataset includes data from 8 volunteers: males aged between 23-37, with a weight between 52-90 kg and height between 172-186 cm. Regarding the dominant hand, two subjects out of 8 were left-handed, while the other ones (6 out of 8) were right-handed. Detailed information (i.e. age, height, weight, dominant hand) for each subject is reported within the dataset itself.

Activity Recording Procedure. During the data recording, we asked all the participants to perform the 17 different activities listed in Table 4.1. We split the list of activities into two different sets, as represented in the table: Set A and Set B; the former has been performed for a fixed time, while the latter has not. The differences between the two sets lie mainly in the fact that activities in Set B were constrained to a particular path or to a flight of stairs, while activities in Set A were quite stationary and did not require the subjects to move along a path. By depending on a fixed path, moreover, it is also possible to measure the different walking speeds (e.g. in terms of meters/second or of steps/second) of the subjects as a piece of additional information. More in detail, Set A activities have been performed for more than 120 seconds (in general for about

[1] https://doi.org/10.7910/DVN/G23QTS

Set A	Set A	Set B
1. Relaxing on a chair	7. Brushing Teeth	13. Walking
2. Keyboard Typing	8. Sweeping	14. Walking Fast
3. Using the Laptop	9. Vacuuming	15. Going Downstairs
4. Handwriting	10. Eating (a soup)	16. Going Upstairs
5. Washing Hands	11. Dusting a surface	17. Going Upstairs Fast
6. Washing Face	12. Rubbing a surface	

Table 4.1: List of Activities performed

150 seconds), and we included in our dataset the central 120 seconds of each execution to obtain cleaner data.

On the other hand, Set B includes:

- *Walking* performed for 160 meters (in at least 110 seconds);
- *Walking Fast* performed for 205 meters (in at least 110 seconds);
- *Going Downstairs, Going Upstairs*, and *Going Upstairs Fast* performed using a single flight of stairs with no intermediate floors between the steps for an average time of 40 seconds.

The participants have been followed and instructed during the data recording: they have been told what activity should have been performed and some details on it, but it has not been imposed to move exactly in a particular way[2]. We have shown a WHO video[3] on how to wash the hands, and we have better explained the difference between *dusting* and *rubbing*: respectively, dust a surface, or rub to clean a really dirty surface. Even if the chosen activities are characterized by relatively standard movements, as expected, we noticed that different subjects had their way of executing movements related to the activity. This is a positive characteristic since it can help to understand the natural differences that can occur when analyzing and comparing different subjects. The only constraint established during the data recording has been to always use the dominant hand (i.e. the one where the device was placed) when performing those actions that mostly involve single arm movements (e.g. handling the vacuum/broom with the dominant hand while vacuuming/sweeping). Otherwise, the signal recorded from the wrist would have had no information regarding the pattern of the dominant hand.

[2]Regarding the possible variation of the behavior of the subject while performing activities by knowing they were participating in an experiment (also known as Hawthorne effect), it is important to note that (1) there have not been judgments on how well subjects were performing activities, therefore they could behave in any way they preferred, (2) due to the short amount of time spent in recording data, possible variations in subjects behavior happened in the entire recording (i.e.the effect would be spread on train/test splits).

[3]"WHO: How to handwash? With soap and water", https://www.youtube.com/watch?v=3PmVJQUCm4E

Device	Sensor Type	Units of Measure
Both	3 axis primary accelerometer	g
GT9X	3 axis secondary accelerometer	g
GT9X	3 axis gyroscope	degrees/s
GT9X	3 axis magnetometer	microTesla (μT)
GT9X	temperature sensor	Celsius

Table 4.2: Characteristics of the devices sensors

Devices and their Positioning. In this work, we used two Actigraph GT9X Link and one Actigraph Centrepoint Insight Watch. Both devices have been described in detail in Chapter 3. As a recap, Table 4.2 shows the kinds of sensors available in these devices and the corresponding measurement units.

The three wearable devices were worn by the participants as follows and with the following settings:

- 1 Actigraph Centrepoint at the dominant wrist. Accelerometer recording at a sampling rate of 256 Hz.
- 1 Actigraph GT9X Link at the right hip at the height of the iliac crest (using the device belt clip). IMU (i.e., accelerometer, magnetometer, and gyroscope) recording at a sampling rate of 100 Hz.
- 1 Actigraph GT9X Link at the height of the right ankle placed, with the help of the belt clip, on the subject's right side of the shoe, over the malleolus. IMU recording at a sampling rate of 100 Hz.

Regarding the calibration of the devices, they have been precisely calibrated (using the automated procedure of the devices) at the beginning of each data recording session.

Ground Truth Definition. The ground truth annotation has been performed by two different persons, in parallel. By following the subjects while performing activities, with the help of a chronometer they were taking note of the starting and ending time of each activity. Moreover, while recording walking data researchers ensured that the subjects were following a specific walking path so that we could retrieve the average walking speed of the subjects for optional and additional tests.

4.2.2 Raw Data Extraction

After recording data with the subjects, we extracted the raw data from the devices using the proprietary software system developed for Actigraph devices. Then we exported the data as .csv

files. The two kinds of devices that we used were equipped with different sets of sensors, so the output of each kind of device will be different. The .csv produced for the Actigraph GT9X Link, will contain 11 columns:

- '*Timestamp*': timestamp of the sampled values
- *'Accelerometer X, 'Accelerometer Y', 'Accelerometer Z'*: instantaneous accelerations for each axis, measured in units of gravity (G)
- *'Temperature'*: IMU temperature, in Celsius degree
- *'Gyroscope X', 'Gyroscope Y', 'Gyroscope Z'*: the instantaneous measure of the gyroscope for each axis, measured in degrees/sec
- *'Magnetometer X', 'Magnetometer Y', 'Magnetometer Z'*: instantaneous measured magnetic field for each axis, measured in microTesla (mT)

For each row of the file, it is possible to find the sampled value at the specified timestamp from each of the sensors and axes. The .csv file produced using data recorded with the Actigraph Centrepoint, instead, will only have these columns: *'Timestamp', 'Accelerometer X', 'Accelerometer Y', 'Accelerometer Z'*.

4.2.3 Data Labelling and Cleaning

Thanks to the ground truth, we were able to label the data precisely. Labels were associated with each row of the recorded data indicating which activity was carried out in such an instant. A new column has been attached to data, where for each row we had a number corresponding to the activity performed in that instant (e.g., 1 is *Relaxing*, 2 is *Keyboard Typing*, ...). During this data processing step, we also used a label to identify data recorded in between two different actions. In this way, dataset users can decide whether to consider this kind of data or ignore it. This decision could lead to create a classifier able to recognize only activities of interest or to have a model able to deal also with activities not of interest for specialists.

4.3 Related Works

In this section, we will briefly analyze related works, considering publicly available datasets on activities recorded with wearable devices.

As briefly explained before, when looking for a publicly available dataset, we have focused our analysis on three main criteria: (1) the number and kind of recorded activities, (2) the reliability

of the recorded data according to the used device, and (3) which and how many parts of the body have been interested during data recording. To the best of our knowledge, a dataset satisfying the three aforementioned criteria is not currently available and this motivated our proposal.

About the first criterion, it is possible to find datasets focused on specific contexts of daily life: De la Torre et al. [DlTHB+09] presented a dataset on cooking activities, while Chavarriaga et al. [CSC+13] proposed a dataset on activities performed while preparing breakfast. On the other hand, it is also possible to find datasets related to a wider list of activities. Possible examples are the work of Anguita et al. [AGO+13], including more generic activities like *sitting, standing, walking, walking upstairs/downstairs* or the work of Leutheuser et al. [LSE13], including activities from a daily life scenario (e.g., *walking, vacuuming, washing dishes, lying, sitting*). We have noticed that many available datasets include quite similar activities such as *walking* but at different speeds, or in different directions, *sitting, standing* or *lying*. Micucci et al. [MMN17], in fact, with their brief literature review, have found out that the most frequent activities included in the daily life activities dataset are: *walking, standing* and *walking downstairs/upstairs*.

Regarding the second criterion, a large number of the datasets that we have analyzed used data recorded with an Android smartphone, with a requested sampling frequency of 50 Hz (e.g. [SS16, AGO+13, SSH13]). Nevertheless, according to the work of Micucci et al., Android OS does not guarantee the consistency between the requested and the effective frequency sampling rate, therefore, the acquisition rate actually fluctuates during the acquisition [MMN17]. This fact reduces, in our opinion, the reliability of the recorded data. On the contrary, some datasets use efficient devices with a high sampling frequency rate (i.e., higher than 100 Hz) as the work of Leutheuser et al. [LSE13] or the work of Zhang et al. [ZS12].

On the third criterion, during our investigation of existing datasets, we saw that some were focused only on one part of the body, particularly on pockets or the hip. Possible examples are the works of Micucci et al. [MMN17], Zhang et al. [ZS12] or Anguita et al. [AGO+13]. On the other hand, other available datasets include data from multiple sensors on different parts of our body, usually including waist, wrist, hip, and ankle data. This is the case of the works of Sztyler et al. [SS16], Shoaib et al. [SSH13] or Leutheuser et al. [LSE13]. In our opinion, having data retrieved from different parts of our body would allow us to achieve higher accuracy for activity recognition purposes.

4.4 Results Achieved by Other Researchers

Since this dataset has been publicly released, it has been used for approaches presented in the literature. This Section will list some of the works that used and cited our Healthy Adults dataset for studying proposed approaches.

Lattanzi et al, for instance, have been using our Healthy Adults dataset in two of their works. In

the first one [LCF22], the dataset has been used to validate a machine-learning approach to recognize when a subject is washing or rubbing its hands, starting from inertial signals collected from wearable devices. The devised approach was used later in a successive work [LC23], to evaluate the trade-off between energy consumption and classification accuracy of a machine learning-based handwashing recognition task, deployed on a real wearable device. In detail, comparing the accuracy and energy consumption of approaches based on Long Short-Term Memory networks and Support Vector Machine. The authors highlighted that our Healthy Adults dataset is one of the few public datasets available that contains hand-washing data sampled through inertial sensors.

Similarly, Mekruksavanich et al., in their article [MJHJ22], devised a Deep Learning model to recognize everyday living human activities starting from wearable inertial sensor data. Furthermore, Hinkle et al., have been using our dataset in two of their works. Their first article [HM22], deals with AR and proposes a method based on a Convolutional Neural Network. Their second work [HM23], has been used to create a framework inspired by the LLVM Compiler architecture that streamlines sensor data processing for machine learning applications.

Chapter 5

Gaslini Juvenile Dataset

5.1 Survey on Activities in Youth-Related AR Studies

With the prospect of a collaboration with Istituto Giannina Gaslini (IGG) to record data from both children affected by pediatric rheumatic diseases and a control group, we have started to investigate activities commonly related to studies regarding juveniles. In order to consider activities of interest for specialists of IGG, and to include also commonly considered activities, we have tried to understand which were the most frequently monitored activities in studies regarding Activity Recognition (AR) in Youth, within the age range of 2 - 18 years.

To do so, we decided to perform a brief survey on this topic. We have looked for studies by querying a scholarly literature web engine, using the following keywords:

- Actigraphy Activity Pediatric
- Activity Recognition Children
- Activity Recognition Pediatric
- Activity Recognition Juvenile

Firstly, we have selected the first 10 results obtained with each query. By looking at the title and abstract of the relevant works found, we have excluded the papers according to the following criteria: (1) the topic was not related to an AR system; (2) the authors were not specifying a list of selected activities; (3) selected activities were not part of the daily life, but were mostly activities limited to a particular context (e.g. sport) or were particular movements (i.e. segment of an activity) instead of activities. We have not limited our survey to works that were using accelerometer data (even if most of them were using this kind of data), since we were concerned

about the most monitored daily life activities for AR systems, and not on used methods, nor on the sensors positioning. By considering the criteria, we gathered a total of 14 works from the initial 40.

We have seen that there was an average number of monitored activities per work of 8 (minimum 3, maximum 11). Merging the lists of activities for each work, a total of 41 individual activities have been found. This quite big number should be read also taking into account that, for example, we have examined *walking, walking downhill, walking uphill, walking downstairs, walking upstairs*, and *walking on stairs* (upstairs and/or downstairs), as 6 different activities.

In Table 5.1, the most frequent activities in the considered studies with their associated number of presences.

Walking	13	Walking Upstairs	5
Running	10	Walking Downstairs	4
Sitting	9	Playing Video Games	3
Lying/Sleeping	9	Playing Basketball	3
Standing Still	8		

Table 5.1: Most frequent activities in the considered studies, with their associated number of presences

From the results shown in the table, it is possible to notice that most of the activities listed in the table involve somehow *walking*, or *staying still* at different positions.

Those activities that have been found in selected works but are not listed in the table, were present in less than 3 works and included sports (e.g. soccer, tennis, jumping, floor exercise, ...) or household activities (sweeping, wiping, laundry tasks, ...). Considering which activities have been considered in the works and their associated number of presences, we could reason about the fact that few studies among the analyzed ones considered genres of DLA other than walking, sitting, lying, or standing. This also suggested recording more complex activities.

5.2 Dataset Characteristics

In the context of a collaboration with a department of *Istituto Giannina Gaslini* [1] (IGG) in Genoa, Italy, designated to the cure of juvenile rheumatologic diseases, we had the desire to analyze the possibility of investigating three different chronic diseases using accelerometer data. We have considered, in particular, Juvenile Fibromyalgia (JF), Juvenile Dermatomyositis (JDM), and Juvenile Idiopathic Arthritis (JIA). To the best of our knowledge, there are currently no datasets

[1] https://www.gaslini.org/

related to accelerometer data recorded while performing activities of daily living (ADL), with subjects affected by these chronic diseases.

The main motivation behind this data recording is to investigate possible differences in movements of selected activities, between healthy subjects and patients suffering from the chosen diseases. For this reason, to record data we involved a population of patients suffering from either JF, JDM, or JIA, and others as a control group. This resulted in a total of 38 subjects, aged between 2 and 18 years, of which 4 for JF, 4 for JDM, 14 suffering from JIA, and 16 healthy.

Regarding the selected diseases we can say, in brief, that JIA implies joint inflammation. JD usually leads to issues such as muscle weakness, pain, or skin rashes as well. JF, instead, is associated with long-term widespread musculoskeletal pain, mixed with fatigue and problems related to attention, reasoning, memory, and sleep as well. To recap, all the considered diseases could, in any case, induce feelings of fatigue and, in general, difficulties in performing even simple activities of daily life. For this reason, with the help of physicians, we have chosen a list of activities to record (while carrying on the wearable sensors), whose movements could be influenced by pain due to chronic diseases. Therefore, we have determined a list of 15 activities of interest, that could be related to real-life scenarios (e.g. walking, eating, teeth brushing, wearing a jacket).

Considering the effects of selected diseases, with respect to chosen activities and our body, we have picked two critical points where we placed our devices: dominant wrist and (on the same side) ankle. We put one Actigraph Centrepoint [act] and one Empatica Embrace 2 on the dominant wrist. On the ankle, we put one Empatica Embrace 2 [emb].

In addition to this, we desired to associate the patient not only with the label of the related disease (for subjects under cure in the hospital) but also to indicate the level of severity of the disease for chosen body regions. To this purpose, but not only, additional data have been gathered by specialists. These data comprehended answers to standardized questionnaires that have been previously published and presented in the literature [FCS$^+$11, VFC$^+$13, KMP$^+$05, LLR$^+$99] , and are designed in order to assess the clinical situation of the subjects from the point of view of the patients (and of their parents). We devised, therefore, a way to gather this information with the goal of finding a value for the level of influence of the disease on body regions. This value can be interpreted as a ground truth score describing the condition of the patient and can be associated with accelerometer recordings of the performed activities.

To summarize, we are sharing a dataset containing accelerometer data with peculiar characteristics:

1. considering a population of non-adult subjects;
2. presenting a variety of subjects comprehending patients suffering from three chronic diseases and a control group;
3. containing data on numerous and different daily life activities;

4. comprehending data recorded using high-quality sensors and FDA-cleared devices;
5. including data from three different synchronized devices positioned on two different parts of the body associated with the selected activities
6. combining additional information giving an indication of the disease's level of influence on selected body regions, for subjects under cure
7. published and fully accessible on the Harvard Dataverse website.

5.3 Related Works

In this Section, we will briefly present some works available in the literature, with a focus on the usage of datasets recorded with accelerometers. We will also focus on recordings with subjects suffering from diseases and with juveniles in general.

As a recall, we are sharing a dataset of accelerometer data recorded while performing daily life activities, with the collaboration of a set of volunteers, some of whom suffer from JIA/JDM/JF, and others who were included as a control group. To the best of our knowledge, there is no dataset available in the current state-of-the-art, dedicated to our particular scenario (i.e., accelerometer data from juvenile subjects suffering from chronic diseases, including a control group, with a score of disease impact for body regions).

Different examples could be made if we focus on datasets comprehending accelerometer data recorded while performing ADL. In fact, in the last years, this kind of data has been widely tested to be used in the healthcare world as support. For example, Logacjov et al. [LBK+21] in 2021 presented HARTH, dedicated to human activities in free-living and labeled by experts, using two accelerometers and the help of 22 subjects that performed 9 different activities (e.g. walking, sitting, standing, running, cycling). A similar dataset was released in 2020, by Pires et al. [PGZL20]: in this case, authors recorded 5 ADLs, with the support of 25 subjects while wearing a waistband.

Concentrating on this type of work, we would like to cite a contribution that has helped to create this dataset. In the work from Leotta et al. [LFV21b], the dataset we described and shared was related to healthy adult subjects who performed daily life activities while wearing wearable devices. This previous work has helped the authors in devising the protocol, and it can be considered as a preliminary approach to move to this more complex scenario which includes juvenile subjects, some of them suffering from chronic diseases.

In the literature, other possible examples of datasets including accelerometer data and subjects with diseases have been examined. Data from this kind of sensor have been studied in relationship with Parkinson's and Alzheimer's diseases, for example. In 2021, Vergara-Diaz et al.

[VDDP+21] recorded data from accelerometers worn on the limb and trunk using wearable devices. The main goal of their study, in addition to sharing the dataset, was to assess if wearable sensor data could be useful to estimate the severity of limb-specific symptoms related to Parkinson's disease (PD). In addition, in 2011, Weiss et al. [WSP+11] recorded accelerometer data with sensors on the lower back of subjects while walking. In this case, the population included subjects suffering from PD and a control group. Analyzing frequency-based measures, authors found differences between the two groups and also in the case of anti-Parkinsonian medications. We have not been able to access the associated dataset.

Regarding Alzheimer's disease (AD), in 2017 Varma et al. [VW17] monitored physical activity with a hip-worn accelerometer, with the contribution of 92 subjects, including some suffering from AD and a control group. The authors could find differences in moderate-intensity physical activities between the control and the AD group. Previously, in 2016, Duque et al. [DNRMM16] studied the possibility of assessing the stage of disease (i.e. early, middle, or late) from accelerometer data and human movement patterns. Regarding these last two works, we could not find any related public dataset.

If we focus on work related to rheumatic diseases and using accelerometer data, we can cite two interesting studies. A dataset was published in 2023, by Belau et al. [BFB+23], where the authors recorded data from accelerometers to examine the physical activity of adults with Rheumatoid Arthritis (RA). According to this work, where 607 subjects were included, adults with RA were less physically active than adults without RA; at the same time, habitual physical activity levels did not differ during the day, compared to adults without RA. Another study regarding patients suffering from RA is the one from Hamy et al. [HGGP+20]. The main objective was to explore the usability of smartphone sensor data when performing wrist joint motion and walking tests. Their results showed differences in the range of motion of the wrist between participants with light-moderate versus severe pain, and in walking step times as well in subjects having slight versus moderate problems in walking. We could not retrieve the dataset used to generate the results of this last work.

Lastly, we would like to mention two contributions related to children and adolescents with mobility impairments or chronic diseases, associated with accelerometer data. The first one, released recently, in 2023, is a public dataset from Rast [Ras23]. This work comprises inertial sensor data from 31 children undergoing rehabilitation. Subjects were asked to perform an activity circuit at a rehabilitation center, including motions such as: watching a movie on the bed, playing cards, drinking from a glass, cycling, and walking. On this topic, it is useful to cite as well a systematic review and meta-analysis of 2017 from Elmesmari et al. [ERMP17]. The objective was to examine levels of moderate-to-vigorous physical activity and sedentary time, measured with accelerometers in children and adolescents with chronic disease. Comparing results with healthy peers, emerged that the physical activity levels of children with chronic diseases appear to be well below guideline recommendations, although comparable with the one of their healthy peers.

5.4 Dataset Creation

The dataset has been created with a process composed of three major phases: (1) Data Recording, (2) Data Extraction, and (3) Data Cleaning and Labeling. The result is the final dataset that includes raw data recorded with juvenile subjects (some affected by chronic diseases, others healthy) while performing a total of 15 daily life activities (DLA) and wearing wearable devices recording data.

The data recording process has been associated with a precise protocol defined with the help of specialists and approved by the ethical committee of Regione Liguria. Each parent (or their legal representative) of the subject was asked to accept informed consent, to have the authorization to share all data recorded and physical bio-metric characteristics, in an anonymized form, of each participant. The dataset and all related information are available on Harvard Dataverse website[2].

The published data will be ready for being used in new tests and studies by researchers. In the following, we will detail the phases that produced our dataset.

5.4.1 Data Recording

Hereby we will describe the details regarding data recording presenting: inclusion/exclusion criteria, population characteristics, activity recording procedure, device positioning, and definition of ground truth.

5.4.1.1 Inclusion/Exclusion Criteria

Participants of this study were recruited and selected according to the inclusion and exclusion criteria, described in the following. The following standards could be split into ”scientific” and ”organizational” ones.

Scientific inclusion criteria:

- able to understand the given instruction;
- able to perform the requested actions;
- aged 2-18 years;
- diagnosed with JIA/JDM/JF or being siblings of children with JIA/JDM/JF or other healthy children attending the clinic.

[2]https://dataverse.harvard.edu/... (Full link will be available before Ph.D. Thesis defense)

Organizational inclusion criteria:

- attending the department of IGG curing juvenile rheumatologic diseases;
- parents/legal representative having signed an informed consent/assent form indicating they understand the purpose of the study and are willing to participate in the study;
- willing and able to come to the study site.

Concerning the exclusion criteria, we can say that any potential candidate who did not satisfy all the inclusion criteria was not considered for the study. Moreover, subjects who were bound to a wheelchair, or bed or unable to walk for causes other than considered diseases were excluded. Lastly, if planned surgery or procedures could interfere with the conduct of the study, the subject could not be included.

Ultimately, subjects were aware that participation in the study was free, with no economic reward, and that it was possible to withdraw from the study at any point, without providing any reasons for doing so and with no consequences in the usual care of the child.

5.4.1.2 Participants Characteristics

Our dataset includes data from 38 volunteers: 4 for JF, 4 related to JDM, 14 associated with JIA, and 16 healthy ones. In Table 5.2, we present a distribution of the population with the health status and age of the involved subjects.

Age	JF	JD	JIA	Healthy
2-6 years	0	1	3	4
7-12 years	0	3	8	7
13-18 years	4	0	3	5

Table 5.2: Distribution of population according to health status and age

Few fields that are equal to zero in Table 5.2 are due to the characteristics of the disease itself (e.g. it is rare to have subjects between the age of 2-12 with fibromyalgia), or due to the lack of patients followed by the hospital with the requested characteristics (e.g. dermatomyositis in the 13-18 age range).

For what concerns body characteristics (e.g. weight, height), there is a high variance in these values considering the high rate of growth in the considered ages. Nevertheless, we can say that weights were between 10-90 kg (average 41 kg), and heights between 90 and 175 cm (average 138 cm). Regarding the dominant hand, 6 subjects out of 38 were left-handed, while the other ones (32 out of 38) were right-handed. Additional information in detail for each subject (i.e. age, height, weight, dominant hand) in an anonymized version, is reported in the dataset itself.

Set A	Set A	Set B
Relaxing	Eating	Tying shoes (3 times)
Using the Laptop	Brushing Teeth	Supine to Upright (5 times)
Handwriting	Walking	Wearing a jacket (5 times)
Washing Hands	Running	Going Downstairs
Washing Face	Jumping in place	Going Upstairs

Table 5.3: List of Activities performed

5.4.1.3 Activity Recording Procedure

To record data, we have been asking all participants to perform the 15 DLA listed in Table 7.1. As anticipated, the list of activities to perform has been chosen reasoning among physicians and data scientists in such a way that could satisfy needs from both the medical point of view (as the involved movements could be influenced by the considered diseases) and one of the data scientists (since the accelerometer data needed to be meaningful for the activity itself).

We can split the activities into two types: Set A and Set B. Activities belonging to Set A have been performed for a fixed amount of time (i.e. 90 seconds, or 60 seconds in the particular case of ”jumping in place”). DLAs belonging to Set B, instead, have been performed for a fixed amount of times (e.g. wearing a jacket 5 times) or were performed considering a particular path (e.g. when walking upstairs and downstairs, we made use of a singular flight of stairs). For these latter activities that were performed a specific number of times, we included also the movements required to return to an initial state (i.e. ”tying shoes” required to untie shoes as well).

Participants were followed and instructed by physicians during the data recording: before starting an activity subjects were told what activity should have been performed and described in brief what was the task. Subjects have never been asked to behave or to move in a particular way during data recording. When dealing with particularly young subjects, in particular cases specialists asked subjects to mimic movements, as a parent would do. Concerning this topic, during data recording, it has simply and kindly been asked the subjects to always use the dominant hand (i.e. the one where the device was placed) when performing those actions that mostly involve single arm movements (e.g. handwriting, eating, etc). If this had not happened, the signal recorded from the wrist would have included data not associated with the pattern of movements typical of the activity.

In a few cases, subjects did not manage to perform at all certain activities or could not execute a task for the total requested amount of time. This could happen because of the health status (e.g. due to fatigue, for activities such as walking, running,...) or because of the participant’s age. Few subjects were simply not able to do some tasks yet (e.g. handwriting, tying shoes); in other cases, it has been difficult to convince particularly young subjects to perform an activity without

interruptions for a fixed amount of time. Additional information presenting the availability of data according to subjects and activities will be presented in the dataset itself.

In order to help new studies in the task of Activity Recognition, dealing also with data that are not of interest to the physicians we also recorded (and kept in the dataset) data related to the moments in between two activities, since this kind of data is quite rarely shared in other available datasets [LFV21b, FLV21].

5.4.1.4 Ground Truth Definition

During data recording, two different individuals were assigned to instruct the subjects and guide the full procedure to respect the protocol. Normally, one of them was focused on supervising the subject and explaining the activities to be performed. The other individual was usually concentrated on annotating ground truth. In particular, it was needed to take note of the starting and ending time of each activity with the help of a chronometer, and if the activity had been performed for a shorter time or not at all.

5.5 Assigning Disease Impact Scores to Subjects

In this Section, we will describe how we started from clinical evaluation metrics data to obtain a score able to express the physical impact (e.g. pain, rigidity, mobility impairments) of the disease on their daily life.

As anticipated, parents (or their legal representative) of volunteers had to sign an informed consent/assent form indicating they understood the purpose of the study. Along with this, each parent of the subject (or, eventually, the subject itself, if mature enough to read and understand a written document) was asked to fill out a questionnaire. This questionnaire was an easy way to assess the clinical situation of the subjects from a point of view that is different from those of the specialists.

For each of the examined diseases, there is an associated questionnaire. Concerning Juvenile Arthritis, we took into consideration the *Juvenile Arthritis Multidimensional Assessment Report* (JAMAR c-JADAS71) [FCS+11]. Regarding Juvenile Dermatomyositis, the *Juvenile Dermatomyositis Multidimensional Assessment Report* (JDMAR) [VFC+13] has been used. For subjects suffering from Fibromyalgia, we adopted the *Juvenile Fibromyalgia Multidimensional Assessment Report (J-FiMAR)*, which is a questionnaire following the templates of JAMAR and JDMAR, revised for Fibromyalgia.

Additionally, for those subjects suffering from JDM, specialists considered the Kendall Manual Muscle Testing (MMT) [KMP+05] and the Childhood Myositis Assessment Scale (CMAS)

[LLR+99] as well as evaluation metrics. The first test is applied to a set of eight muscles that are evaluated according to the standard scores for Kendall MMT, with a scale from 0 to 10. The latter is a functional assessment tool to check muscle function with respect to strength and endurance across a wide range of abilities and ages, by considering a total of 14 movements.

The following are a few possible examples of questions that we can find in the forms.

> Please choose the answer that best describes your ability to perform the functional activities listed below, considering the *last four weeks* and indicating only the difficulties or limitations caused by the illness.
>
> - Climb 5 steps [No difficulties / Some difficulties / A lot of difficulties / Unable to do it / Not evaluable]
> - Squeeze an object with your hands [. . .]
> - Squat down on your knees [. . .]
> - ...
> - I had difficulties in taking care of myself (e.g. eating, dressing, washing) [Never / Sometimes / Often / Always / Not evaluable]
> - I had difficulties in carrying out activities that require a lot of energy such as running, playing football, dancing, etc. [. . .]

All of the questionnaires have a section to evaluate the functional abilities, and another one to measure the quality of life. As aforementioned, both the forms and the evaluation metrics have been used to better understand the clinical situation of the subjects, and this information is also definitely useful for tracking the success of cures and therapies.

In the following parts of this section, we will briefly explain the reasons that led us to aggregate information from clinical evaluations; then, we will describe the process used to obtain the score, and we will present how answers to questionnaires have been associated with body parts to obtain the final score.

5.5.1 Reasons for Aggregating Information of Clinical Evaluation

Each of the questionnaire and considered evaluation metrics has its own list of questions and it is focused on evaluating peculiarities of symptoms of the specific considered disease. Therefore, it is easy to imagine that we had a lot of different information for the entire list of volunteers. This data was, at first impact, unclear and difficult to understand. Moreover, even if we had a lot

of information regarding aspects of the subjects, we could not easily aggregate the whole data to infer the clinical status of the subjects. We were looking, in fact, to have a score able to express the physical impact (e.g. pain, rigidity, mobility impairments) of the disease on patients' daily lives: for example, it was not practical nor beneficial to have information regarding the singular fingers of the hands, if we were dealing with data recorded on the wrist. Additionally, if this data had been publicly shared, we could have caused a privacy issue since the given answers could have been considered as sensitive data.

For these reasons, we wanted to find a method to aggregate in a systematic way the clinical evaluations, collected via questionnaires and assessment tests, into a disease impact score by body regions. The score would have been a value from 0 to 100, where 0 means that the disease has no impact on the associated body region and 100 means that there is a high impact. In particular, a value of 100 would signify that for each of the questions in the form, the subject always answered with the option linked with higher difficulties due to the disease.

Considering that devices were placed on the dominant wrist and on the ankle, we have split the body into 3 main parts: upper limbs, lower limbs, and chest. In our hypothesis, we were hoping to notice differences in the movements of activity between subjects with a high disease impact score and healthy subjects. Moreover, by considering this partitioning of the body, we have the possibility to describe subjects not only by classifying them as "subject with JIA/JDM/JF", but to describe them as "subject with JIA/JDM/JF with impact on the upper limbs, lower limbs, and chest of *x,y*, and *z*, respectively", where *x,y,* and *z* are the obtained scores for the selected body region. This is useful for specialists because a subject could suffer from JIA having, for example, no impact from the disease on the upper limbs, but only on the lower limbs and thus having a different impact on the DLAs performed with such parts of the body.

5.5.2 Process to Obtain the Disease Impact Score

In the following, there will be presented the process followed to obtain the presented score. It is crucial to list the fundamental concepts that we have taken in mind when designing the process.

1. To simplify the task, we have considered the assumption that a score of 100 for a disease would be equivalent to a score of 100 for another disease. Relying on questionnaires made from the patient's point of view, where questions were different for each form but still similar, we believed that this assumption would be acceptable to make the interpretation of data easier. If any of the final users of this dataset wishes to change this assumption, it will be sufficient to customize the values of one disease with respect to others.

2. We have considered only questions/values associated with the physical impact of the disease, and not with the psychological one. For example, the question "I felt sad or depressed, during the last four weeks" was not kept in consideration for the score. On the

other hand, the point "I had difficulties in walking for at least 15 minutes or to walk upstairs" has been used for the score.

3. Each of the questions kept from the questionnaires and the evaluation tests, has given a contribution to the final score. The general idea is that the final score corresponds to the sum of contributions given from all questions. Except for certain particular cases, each question had equal weight with respect to the others. Once all of the contributions have been summed up, the total is normalized over 100.

4. We normalized the possible values of the answers to obtain a value between 0 and 1 for each question. As an example, the question "I had difficulties in walking for at least 15 minutes or to walk upstairs" had four possible answers: never, sometimes, often, always. These values have been encoded, respectively, as integer values: [0,1,2,3]. We normalized these values to have a real number between 0 and 1.

5. Each question could give a contribution to the final score of one or more body regions. For example, the question about evaluating difficulties of the functional ability "walking", supports the score of lower limbs score. The question about evaluating the task of "standing up from the floor", is responsible for the scores of lower limbs, upper limbs, and chest.

6. Focusing on the JAI disease, we had details (i.e., presence of tumefaction, pain, or functional limitations) for many of the possibly involved articulations. With the idea of being conservative to the availability of information regarding subjects, we wanted to allow these data to be part of the score. At the same time, for this peculiar circumstance, we did not want to give the same weights for the same articulations or parts of the considered body region. For example, looking at the upper limbs, the tumefaction of a singular phalanx is not comparable to the tumefaction of the elbow, due to the limitations this issue could cause. For this reason, we have given different weights for each of the different considered joints.

5.5.3 Association of Answers with Body Parts

In this part, details will be given according to the diseases since each questionnaire was different for each of the considered conditions.

5.5.3.1 Juvenile Fibromyalgia

For this disease, we have information regarding the functional ability and the evaluation of quality of life. In Table 5.4, we will show the list of information evaluated, and the body regions influenced by the answers. Each of the values has given an equal contribution to the final score.

All the considered information had answers coded in the range [0-3], then normalized in the range [0-1]. For the first set of questions, it was asked the subjects to answer reasoning about the preceding *three weeks*, while the latter ones refer to the preceding *three months*.

Information	Chest	Upper Limbs	Lower Limbs
Running for 50 meters			X
Walking for 15 minutes			X
Climbing 10 steps			X
Squeeze object with hand		X	
Write/draw with pencil		X	
Bring books with book bag	X		X
Taking care of myself	X	X	X
Energy-consuming DLAs	X	X	X
School activities or playing with friends	X	X	X

Table 5.4: Association of answers from Juvenile Fibromyalgia questionnaire and body parts

5.5.3.2 Juvenile Dermatomyositis

The score for this disease has been computed using four kinds of information: functional abilities, quality of life, Kendall Manual Muscle Testing (MMT), and Childhood Myositis Assessment Scale (CMAS) evaluation. Questions regarding functional abilities and quality of life had answers coded in the range [0-3]. On the other hand, the values returned from the MMT were in the range [0-10], while CMAS evaluation could give us values in different ranges according to the considered movement: [0-2],[0-3],[0-4],[0-5][0-6]. All of these ranges have been normalized in the range [0-1] in such a way that the maximum value would always correspond to the maximum impact on daily life due to the disease. In Table 5.5, the list of information regarding functional abilities and quality of life and the body regions influenced by the answers. For all these questions, it was asked the subjects to answer reasoning about the preceding *four weeks*.

As anticipated, in the case of JDM, the Kendall MMT evaluation has been considered as well. In Table 5.6, we will list the information obtainable from this test, performed at the time of the first study visit. In particular, specialists evaluate the capability of a list of muscles, presented in the Table.

Additional information was available for JDM since specialists evaluated the subjects with the Childhood Myositis Assessment Scale (CMAS) test as well. In Table 5.7, we will present the list of movements considered in the CMAS test, and the associated body part.

Information	Chest	Upper Limbs	Lower Limbs
Walking			X
Running for 10 meters			X
Tiptoe walking			X
Walking upstairs			X
Walking downstairs			X
Pick up object on the floor		X	X
Sitting on the floor	X	X	X
Getting up from the ground	X	X	X
Sitting up in bed from the lying position	X	X	X
Turn over in bed	X	X	X
Squeeze object with hand		X	
Open door with a handle		X	
Lift/Carry heavy objects		X	
Pick up heavy objects from a shelf higher than you		X	
Lift head off the pillow	X		
Taking care of myself	X	X	X
Walking for > 15 mins or walking upstairs			X
Energy-consuming DLAs	X	X	X
School activities or playing with friends	X	X	X
Feeling pain	X	X	X

Table 5.5: Association of answers from Juvenile Dermatomyositis questionnaire and body parts

5.5.3.3 Juvenile Arthritis

Data used to compute the score belonging to this disease has been retrieved from four different points: *functional abilities, quality of life, presence of pain/tumefaction*, and *presence of tumefaction/pain/functional limitations*. For the first three cases, the values were coming from questionnaire answers, therefore from the point of view of the subjects (or their parents). For the last case, we have referred to the analysis of specialists who have been visiting the subjects. The third and fourth points might look like each other, and have in fact possible redundant values, but have been assembled in completely different ways.
Questions regarding *functional abilities* and *quality of life* had answers coded in the range [0-3].

Muscle	Chest	Upper Limbs	Lower Limbs
Deltoids	X	X	
Biceps brachii		X	
Wrist extensors		X	
Quadriceps			X
Ankle dorsi-flexors			X
Neck flexors	X		
Gluteus medius			X
Gluteus maximus			X

Table 5.6: Association of measures from Kendall Manual Muscle Testing (MMT) and body parts

Exercise	Chest	Upper Limbs	Lower Limbs
Head lift	X		
Leg raise			X
Straight leg lift	X		X
Supine to prone	X	X	
Sit-ups	X		
Supine to sit	X	X	X
Arm raise [straighten]		X	
Arm raise [duration]		X	
Floor sit	X	X	X
All four manoeuvre	X	X	X
Floor rise	X	X	X
Chair rise	X	X	X
Stool step			X
Pick up object on the floor		X	X

Table 5.7: Association of movements evaluated in Childhood Myositis Assessment Scale (CMAS) and body parts

The other two points were just indicating the presence or not of *pain/tumefaction(/functional limitation)*, with 0 in case of absence and 1 in case of presence of pain. The values in the range [0-3] have been normalized in [0-1].

In Table 5.8, the list of information regarding functional abilities and quality of life and the body regions influenced by the answers. For all the questions, it was asked the subjects to answer reasoning about the preceding *four weeks*.

In Table 5.9, we will list the information obtainable from the questionnaires, regarding the pres-

Information	Chest	Upper Limbs	Lower Limbs
Walk for at least 10 meters			X
Climb 5 steps			X
Jump forward			X
Squat down on your knees			X
Pick up object on the floor		X	X
Activities involving fingers		X	
Opening/Closing fist		X	
Squeeze object with hand		X	
Open door with a handle		X	
Open and close a jar, or reopen jar previously open		X	
Extend arms completely		X	
Place hands behind head		X	
Turn head until you look over your shoulder	X		
Lean head back until you look at the ceiling	X		
Taking care of myself	X	X	X
Walking for ¿ 15 mins or walking upstairs			X
Energy-consuming DLAs	X	X	X
School activities or playing with friends	X	X	X
Feeling pain	X	X	X

Table 5.8: Association of answers from Juvenile Arthritis questionnaire and body parts

ence of pain or tumefaction in different joints of the body. The considered question asked was: *"Mark if you felt pain or swelling during the last 24 hours in the joints listed below"*. We will list the joints presented in the questionnaire and the associated body area for the score.

In the last table for JIA disease (Table 5.10), we will present the information we got from the clinical evaluation of the specialists. In this case, we have access to a list of articulations. Specialists have checked the presence of tumefaction, pain, and functional limitation for each articulation. To aggregate values, we have set 0 (absence) if none of the issues was present, or 1 (presence) if at least one issue was found.

For this case, we have given a weight (in the range [0-1]) to each of the articulations, considering

Joints	Chest	Upper Limbs	Lower Limbs
Fingers	X	X	
Elbow		X	
Wrist		X	
Knees			X
Ankle			X

Table 5.9: Association of subjective pain in selected joints and body parts

the impact of each articulation on our daily life. For example, we tried to give weights such that the pain of a singular phalanx of the foot finger was not considered as the presence of pain in the heel. The distribution of weights for the lower limbs is done in such a way that the sum of weights for each subregion of our body areas (i.e., foot, ankle, knee, hip) is 1. In the following, the list of articulations and the associated body area for our score and the given weight.

5.6 Dataset Characteristics

In this section, we provide additional details on the amount of recorded data and we will list the scores assigned to subjects following the procedure explained in the previous section. The first is to provide a hint of the volume of recorded data, in terms of minutes, megabytes, and number of samples. The latter is to discuss the values obtained with our approach made to evaluate the impact of diseases on the body regions.

To begin with, in Table 5.11 we would like, in the first place, to separate data according to the status of participants: belonging to the control group or as patients; we will list overall results as well, averaged over all participants. Then, we will provide, for each employed device (or associated body parts): the mean and standard deviation of minutes of recordings, megabytes of data, and number of samples.

To highlight the values presented in Table 5.11, considering the total population, we have recorded, on average, 38 minutes for each participant resulting in approximately 580000 samples, when using the device recording with the higher frequency, or approximately 72000 samples, when using the lower frequency device. It is interesting to notice that, considering the mean values, recordings for the patients have been slightly longer (i.e., considering the duration, patients' recordings are 2 minutes longer than the control group). This could have happened due to the higher difficulties in performing activities and in getting ready from one activity to the next one.

In Table 5.12 we will list the values obtained with our approach for assessing the impact of diseases on body regions. For each participant belonging to the *patients* group, we will list the scores of the individual for *chest*, *lower limbs*, and *lower limbs*. To highlight the characteristics of

Joint	Weight	Chest	Upper L.	Lower L.
Sternum-clavicular	1	X		
Acromio-clavicular	1		X	
Shoulder	1		X	
Elbow	1		X	
Wrist	1		X	
I/II/III/IV/V Metacarpo-phalangeal	1/14 each		X	
I/II/III/IV/V Inter-phalangeal (hand)	1/14 each		X	
II/III/IV/V Distal inter-phalangeal	1/14 each		X	
Hip	1	X		X
Knee	1			X
Tibia-tarsus	1			X
Subtalar	1/3			X
Sole	1/3			X
I/II/III/IV/V Metatarso-phalangeal	1/25 each			X
I/II/III/IV/V Inter-phalangeal (foot)	1/25 each			X
Cervical spine	1	X		X
Sacroiliac	1	X		X

Table 5.10: Association of pain/tumefaction/limitation in selected joints and body parts, from specialists point of view

the scores, we added, at the bottom of the table, a few measures for each region: mean, standard deviation, median, minimum, and maximum.

The scores and the measures show that considering the mean values, participants had a higher impact, on average, on lower limbs. The median values let us notice that in most cases the score was not high. By focusing on minima, we could say that there are patients for whom the impact on *chest* is equal to 0; at the same time, no patients have an impact equal to zero for *upper limbs* or *lower limbs*. Maxima's values, higher than 60 for all of the three regions, suggest that there are a few participants where the impact of the disease was quite high.

Group	Device	Minutes	# Samples	Megabytes
Control	Wrist - 256 Hz	37±4	561537±68260	41.0±4.8
	Wrist - 32 Hz		70114±8537	4.4±0.6
	Ankle - 32 Hz		69094±8777	4.6±0.6
Patients	Wrist - 256 Hz	39±4	592699±65093	43.1±4.8
	Wrist - 32 Hz		72713±9204	4.6±0.7
	Ankle - 32 Hz		73541±8647	4.8±0.7
Overall	Wrist - 256 Hz	38 ± 4	579578±67357	42.2±4.8
	Wrist - 32 Hz		71525±70982	4.5±0.7
	Ankle - 32 Hz		71654±70531	4.8±0.7

Table 5.11: Amount of data for each subject, in terms of megabytes, number of samples, and minutes of recording. Values are expressed as mean ± standard deviation

ID	Chest	Upper Limbs	Lower Limbs
1	3	23	7
2	44	26	55
3	7	4	11
4	33	46	55
9	22	11	58
10	21	11	19
13	11	5	31
14	71	56	58
15	0	13	33
16	11	5	21
17	7	1	13
18	11	10	5
19	3	17	15
20	3	11	9
21	55	60	77
22	9	4	7
24	0	5	11
25	18	9	52
28	14	35	15
31	3	1	21
32	3	1	17
38	22	64	31
Mean	16.9	19.0	26.9
Std. Dev.	18.7	20.2	21.9
Median	11	11	18
Min	0	1	3
Max	71	64	77

Table 5.12: Score of each subject suffering from either JF, JDM, or JIA, for each considered body region
In the bottom part: Mean, Standard Deviation, Median, Minimum and Maximum of the listed values for each region

Part III

Activity Recognition: Methods & Improvements

Introduction to Methodology Part

This part of the thesis, dedicated to *methodology*, has the objective of presenting the different techniques tested to perform Activity Recognition starting from data recorded with accelerometers in wearable devices. Concerning this area, it was noticed that most of the works regarding AR using wearable data were proposing approaches giving a minor advantage to practical applications. In fact, most approaches relying on a classifier to discern the activities were simply dealing with a set of known and learned activities. In real-life scenarios, however, subjects usually perform a plethora of different activities during the day. Thus, with the idea of using an AR system in a long-term scenario, the classifier should also be able to discern activities of interest for the specialists and known to the classifier, from the unknown activities that are irrelevant for studies.

Our contribution to this topic has been to devise a method that supports the classification of non-interesting activities, reducing the impact on the classification of activities of interest and known to the classifier. Then, we have been investigating the techniques to filter data and the impact this step could have on recognition and classification results. This has been also possible with the usage of Dictionary Learning techniques.

In the following, Chapter 6 introduces the baseline approach used to preprocess accelerometer data, extract features, and recognize activities performed, presenting achievable results. Successively, in Chapter 7 we describe the method used to deal with unknown activities, based on an ensemble technique and a voting system, with a system to remove transient misclassifications. Then, Chapter 8 is dedicated to a few attempts made to improve the accuracy of our approach and to improve the explainability of models and results. We will then move on Chapter 9 to introduce our analysis on filtering raw data and on the impact of this step on accuracy. Following to this, in Chapter 10 we will describe an approach based on Dictionary Learning and the achieved results.

Chapter 6

Baseline Method

Introduction

In this chapter, we will describe in detail the steps composing our baseline approach aimed at recognizing the activities performed by a subject when using a wearable device with an accelerometer (e.g. the devices presented in Chapter 3). This is the first method devised for our purpose, and the starting point to develop better techniques used for activity classification and for better analyzing the results. It is important to notice that for these preliminary tests, we have only limited the classifier to deal only with the known activities.

6.1 Approach

This first approach is mainly based on the usage of a Support Vector Machine (SVM) model. More in detail, starting from the labeled ***raw data*** (e.g., from our dataset), the first activity consists in a (1) *features extraction* phase, after which data will be split into ***training*** and ***test data***. Training data will be used for (2) *tuning the hyperparameter* and (3) *training the SVM model* with the correct parameters. At this point, the SVM model can be used for recognizing and therefore (4) *classifying activities* on novel unseen data. Thus, we use ***test data*** to evaluate the accuracy of the trained SVM model and, in general, of our approach.

In the following, we will describe in detail the first three steps (i.e., feature extraction, hyperparameter tuning, and model training). The classification phase and the discussion of the results are described in the following Section (Sec. 6.2). Our approach has been implemented using Python

and with the help of the Jupyter platform[1]; we relied on the *Scikit-learn* library[2], also known as *sklearn*, since it provides several instruments for data analysis that were useful in our study.

6.1.1 Features Extraction

This first step is one from which the final accuracy could depend the most. The goal is to take the raw labeled data and to extract relevant information from it. This step was needed, in our case, since we were starting from time series data, to elaborate it for training the model. Other ML methods and model could learn directly from chunks of time series as well, but it was not our case.

We decided to perform feature extraction as done in other similar approaches like the one in the work of Staudenmayer et al. [SHH+15]. In particular, we have used a *sliding window* approach to compute the features, using only the accelerometer data. This choice of using only these data lies in the fact that this sensor is the only one common to all of the devices considered in our work. However, the approach can be easily extended to include gyroscope and magnetometer data. In this phase, a sliding window passes over the data and for each axis (X, Y, Z) we extract some measures of the instant accelerations contained in the window: mean, variance, standard deviation, median absolute deviation, percentiles (10Th, 25Th, 75Th, 90Th). Having eight measures per axis allows us to compute 24 features for each window that composes the final feature set used.

Since the windows could go over a period of time corresponding to a transition between two different labels, the windows including different labels in the data have been discarded. Data have been also discarded when they were not enough for building a window, such as at the end of data recordings (i.e., if the remaining data covers less time than the length of the sliding window). This criteria will obviously lead to the loss of some data, but in doing so we can avoid labeling data in the wrong way, interfering with the performance of our solution.

Regarding the sliding window, its length represents an important parameter on which results could potentially highly depend. For this reason, we have performed some analysis to understand how the length of the window and the overlap between subsequent windows could affect accuracy. After these tests, we decided to use windows that were 2.0 seconds long, with 95% of overlap each other. These tests results are described in detail in Section 8.1. The value of window length is not only justified by our results, but is also motivated by the fact that typical human periodic movements have a period of no more than two seconds (e.g., each step during walking or hand movement during toothbrushing).

After the feature extraction step, data is ready to be used to train a Machine Learning model. Among the many different possible methods and techniques, for our approach, we have chosen

[1] https://jupyter.org/
[2] https://scikit-learn.org/

to rely on a Support Vector Machine (SVM). SVM, in fact, has been already used to estimate physical activity from accelerometers in the literature, showing good performances in this kind of task (e.g., [ZRMH12, HJ09]).

When using SVM, it is always considered as a good practice to *standardize* data in order to obtain better results. This is needed since SVM is based on the idea of finding the hyperplane that best divides different classes by maximizing the distance between the hyperplane and the data (i.e. Support Vectors), if one feature (i.e. one dimension) has larger values than the others, it will prevail on the others when computing distances. This will not be a problem if we standardize data: we did so by removing the mean and scaling to unit variance. In particular, each sample x will be calculated as $z = (x - u)/s$, where u is the mean of samples in training data, and s is the standard deviation of samples in training data.

Finally, to further prepare our data to feed the algorithm, we have also split our data into *training data* and *test data*: 75% and 25% of data of each activity, respectively. We paid attention in splitting windows in such a way that no raw data selected as training could be part of the test data, and vice-versa.

6.1.2 Hyperparameter Tuning

SVM needs some parameters to be tuned in order to achieve the best result: in combination with the different used *kernels* (*Radial Basis Function* and *Polynomial* kernels), the most important parameters to choose are the value for *C* (as a regularization parameter) and *gamma* (as a kernel coefficient). Focusing on the hyperparameter tuning, we know that while constructing a machine learning model, a general goal is to choose parameters such that we obtain a model that is able to learn, in the best way, all information from the training data, while, at the same time, it should be able to generalize well to new data. This problem of balancing these properties is known as the *Bias Variance Trade-off* problem [VLS11]. One possible way to find the best model is to use the cross-validation method [CT10].

Cross-validation is a frequently used procedure for evaluating a model. The basic idea is that training data are divided into complementary subsets; one subset is used to train the model and we validate the results using the other subset. To do so, we have decided to use the Grid Search method [BB12] for choosing the best parameters for the algorithm. For each parameter of the algorithm, a list of possible values is given in input to Grid Search. Each combination of the selected values generates a model that is then evaluated. The output of Grid Search is the list of chosen parameters that performed the best.

6.1.3 Training Model to Predict Data

After computing the best parameters for the Support Vector Machine model, the next step has been to train the model with the training data and using the previously found parameters, in order to conclude the process. Once the model had been created, it was ready to be fed with new unseen data in order to output its predictions. As explained before we have split, at the beginning, our whole processed data into training data and test data; the latter have been used for this last step to evaluate the accuracy of the created model. In the following Section 6.2 we will analyze the obtained results considering, separately, data of each body location.

6.2 Empirical Evaluation of the Baseline Approach

Hereby there will be presented the evaluation of the baseline approach. This evaluation took into consideration the standard approach of classifiers, where we have been trying to recognize only activities known to the classifier. Our *Healthy Adults Dataset* has been used for this evaluation procedure since data were more stable and therefore easier to process. As a reminder, this dataset involved 3 devices in three different body locations (i.e., wrist, hip, ankle), with the participation of 8 subjects. For the sake of simplicity, as a preliminary study, we independently consider the tree body locations. Thus, we will not test combinations of data coming from different body locations, but we will consider one device (e.g., one body location) at a time.

In particular, the focus of the first subsection is on the procedure used to evaluate our approach. The successive subsection will describe the achieved results.

6.2.1 Procedure

To evaluate our baseline approach we computed three confusion matrices for each subject in our dataset (one matrix for each of the three devices employed). More in detail, the values in each confusion matrix refer to the percentage of processed data of a specific class C_a that have been predicted to belong to the class C_b. More precisely, assuming to read the confusion matrix starting from the first row, representing the class C_a: each value in this row represents the percentage of data belonging to C_a that has been labeled as belonging to the class of the corresponding column. A flawless result would be represented as a matrix in which all the values on the diagonal are 100.0%, and the other values are 0.0% meaning that all the unseen data have been classified with the correct corresponding label.

We then averaged the confusion matrices obtained for each subject, creating a single confusion matrix for each considered body location. In this way, we could evaluate our approach by providing the average accuracy for each activity considered in our dataset. As a score that could

aggregate the results shown in each confusion matrix, we have been using the F1-score. This measure is computed as:

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where *TP* is the number of true positives, *FP* stands for false positives and *FN* is the value of false negatives.

6.2.2 Results

Starting from **wrist** data, in Figure 6.1 it is possible to see the confusion matrix obtained using data from this first device, averaged over all eight subjects' results. The overall mean F1 score obtained is 0.92 ± 0.03 (mean $\pm$ standard deviation). In general, we can say that the recognition of most of the activities achieved good results (values on the diagonal of the matrix are always greater than 0.74).

It is possible to make some considerations on the values in the confusion matrix. In this case, the most noticeable misclassifications are the ones regarding *keyboard typing* in *using laptop*, which are indeed very similar activities. The same analysis can be valid for the classifications of *sweeping* and *vacuuming*. The lower accuracy values are obtained in most of those activities that mostly involve leg movements (*walking, going downstairs/upstairs*): those are indeed quite similar activities if considering only wrist data.

In Figure 6.2 and Figure 6.3 we present the confusion matrices obtained using, respectively, **hip** and **ankle** data, averaged over all eight subjects' results, for which it is possible to perform similar considerations. The mean F1 score obtained with *hip* data is 0.81 ± 0.04, while the mean F1 score obtained with *ankle* data is 0.75 ± 0.06.

Analyzing these confusion matrices (Figures 6.2 and 6.3, we noticed both expected and unexpected results. In fact, as expected, since many performed activities mostly involve peculiar movements of the arms (e.g. *brushing teeth, washing hands/face, sweeping*), results obtained using hip and ankle data have lower mean accuracy than those obtained using wrist data. For the same reason, we were expecting to obtain low accuracy for the activities performed while sitting or while not walking (e.g. *using a laptop, relaxing, handwriting*) since the hip and ankles are not involved in any movements. On the contrary, we achieved quite high accuracies.

We further analyzed our data in order to explain these results. By plotting the accelerometer data, we noticed that there was a perceptible difference in the values between such different activities even in the ankle and hip data. We interpreted this as the fact that subjects, during data recording, unintentionally changed the orientation of the devices (e.g. by sitting in a different way, or by slightly moving a leg while sitting). These involuntary movements were leading to noticeable changes in the accelerometer values because of the variation in the orientation with respect to the earth's gravity g. We concluded that in some cases, the right classification of activities happens

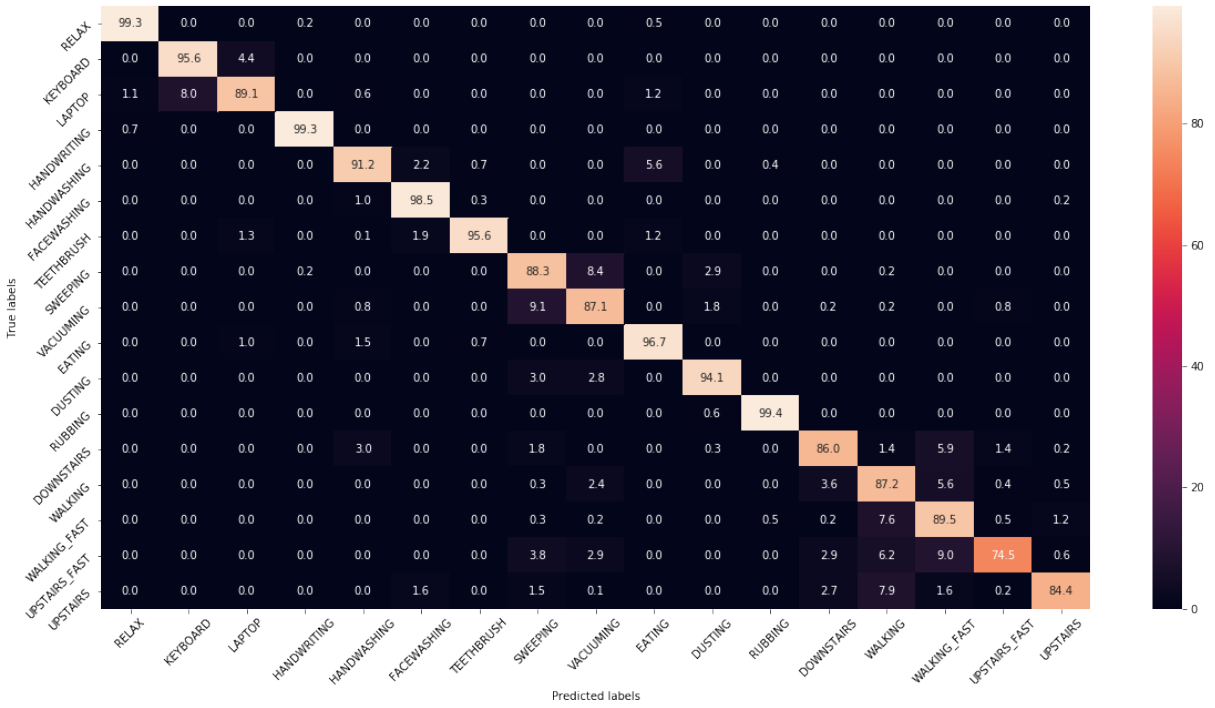

Figure 6.1: Average Confusion Matrix obtained with Wrist Data

not because of the peculiar characteristics of the activity movements, but because of the particular orientation of the device.

Therefore, in order to avoid this problem, when dealing with both hip and ankle data, we considered only activities that actively involve those parts of the body. In particular, we selected: *relaxing* (as a stationary activity), *sweeping, vacuuming, dusting, rubbing, going downstairs, walking, walking fast, going upstairs, going upstairs fast* and excluded *keyboard typing, using laptop, handwriting, hands washing, face washing, teeth brushing, eating*.

We show the confusion matrices obtained with the latest reduced activity set in Figure 6.4 (regarding hip data) and Figure 6.5 (regarding ankle data). In this case, the mean F1 score obtained with *hip* data is 0.48 ± 0.02 (mean $\pm$ standard deviation), and the mean F1 score obtained with *ankle* data is 0.47 ± 0.03.

In both confusion matrices (Figures 6.4 and 6.5) with a limited set of activities, it is clearly evident the scarce accuracy of the classifier in discerning from *sweeping, vacuuming, dusting* and *rubbing*. Indeed, all of the four listed activities have been performed by doing small and slow steps around the room when recording data. Regarding the overall mean F1 scores it is clearly a consequence of the wrong classification of the four activities previously listed. On the same topic, we should also consider that by having fewer activities to be recognized, any wrong

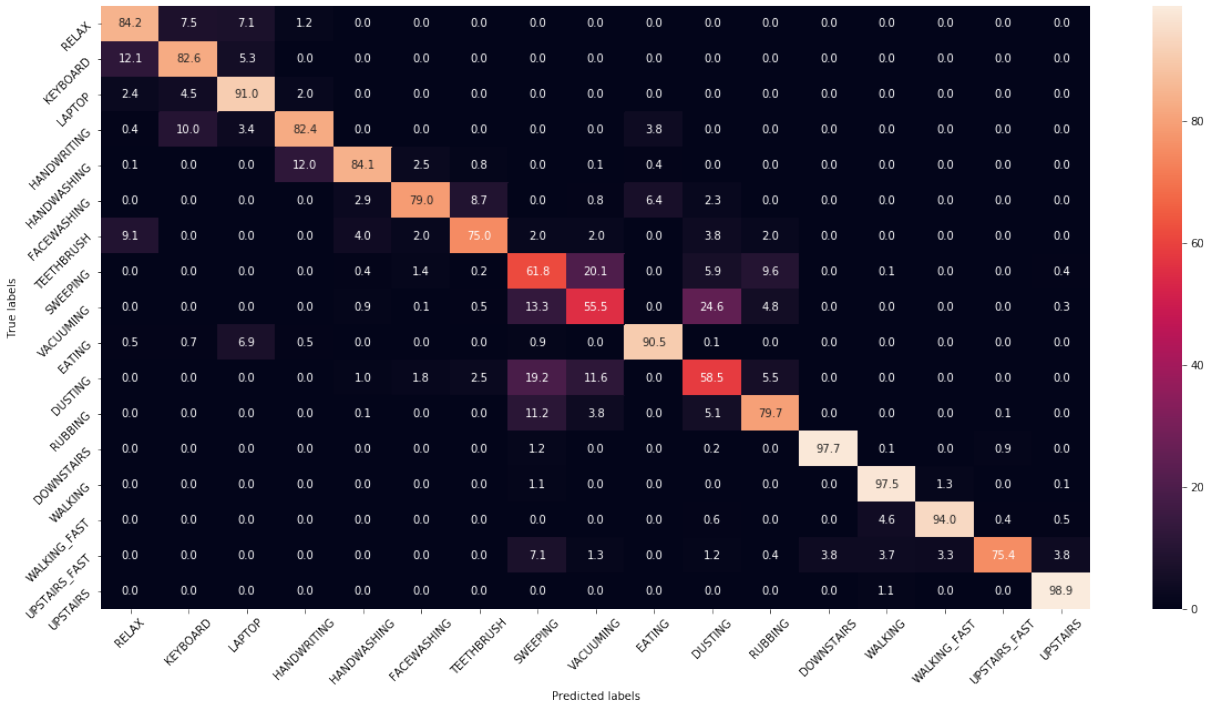

Figure 6.2: Average Confusion Matrix obtained with Hip Data

activity classification will have a significant impact on the F1 score.

On the other hand, the classifier is able to recognize with quite high accuracy (always over 78%) all the other activities and, in particular, the ones that required the subjects to walk and use stairs (*walking, going downstairs, upstairs, upstairs fast*), in which both ankle and hip are more involved.

6.3 Related Works

Different approaches for classifying daily life activities using Machine Learning algorithms have been proposed in the last years. Here we will consider three works that have similar scenarios to ours. Indeed, all of the considered methods deal with a triaxial accelerometer worn on the wrist by participants of the experiments while performing some activities. All the devices used in the considered experiments recorded accelerations at a frequency of 80-100 Hz.

The work of Zhang et al. [ZRMH12] tried to classify 4 main categories of activities: *sedentary* (lying, standing, PC working), *household* (window washing, sweeping, etc.), *walking* and *running* at different speeds. Mannini et al [MIR+13] tried to recognize as well 4 categories of

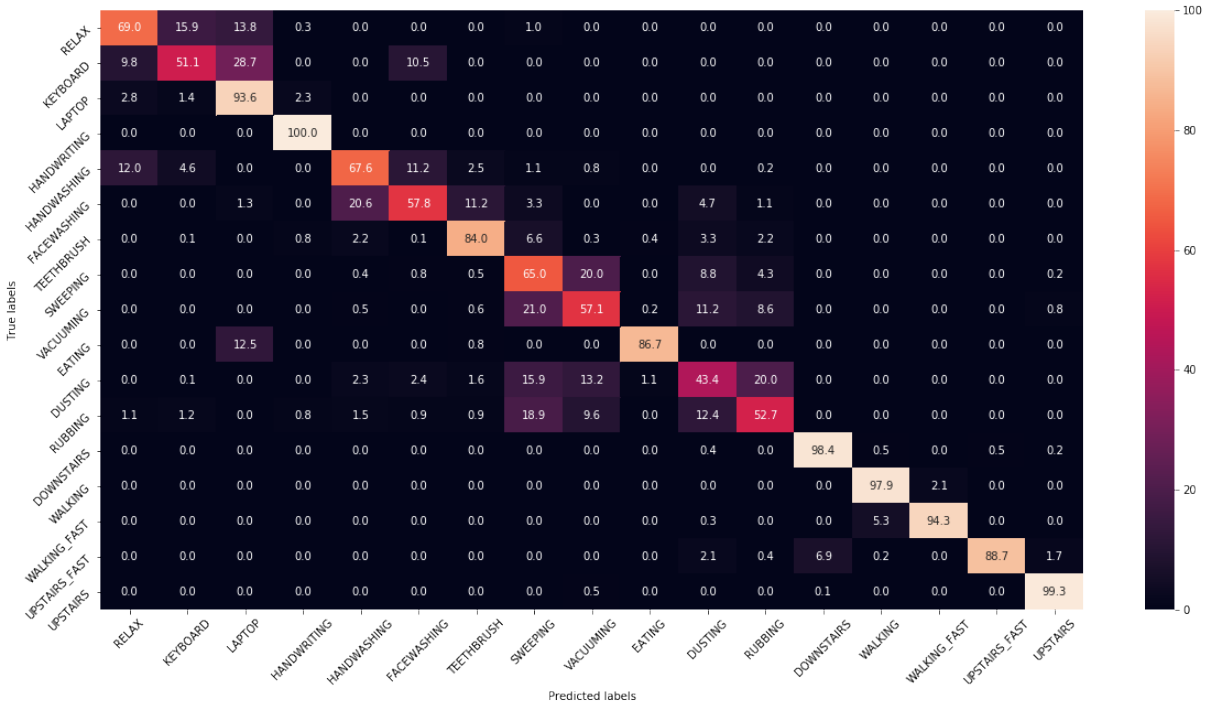

Figure 6.3: Average Confusion Matrix obtained with Ankle Data

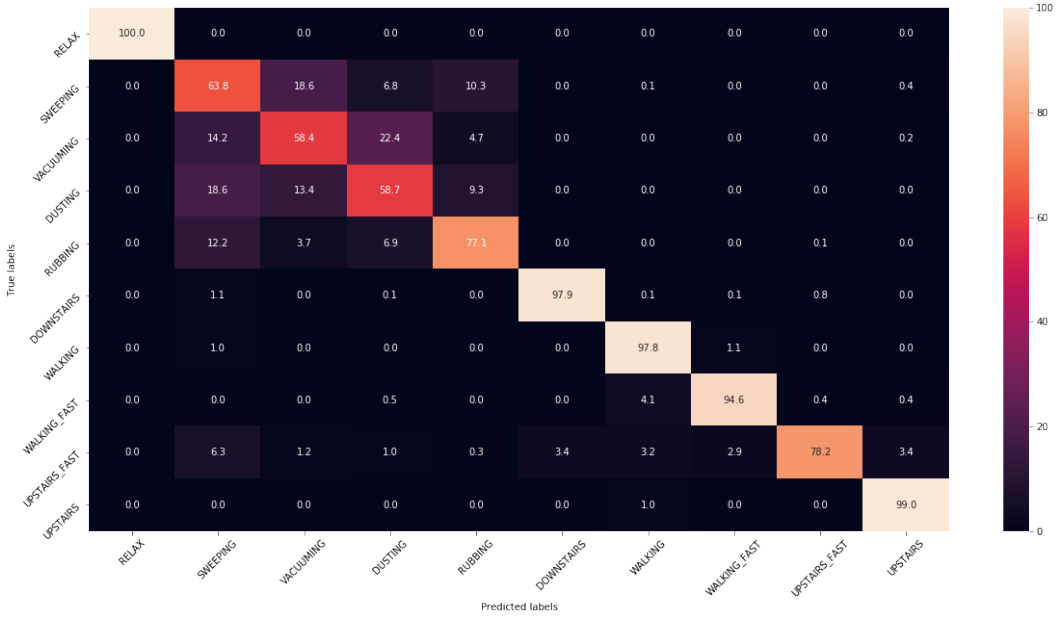

Figure 6.4: Average Confusion Matrix obtained with Hip Data, limited on hip-related activities

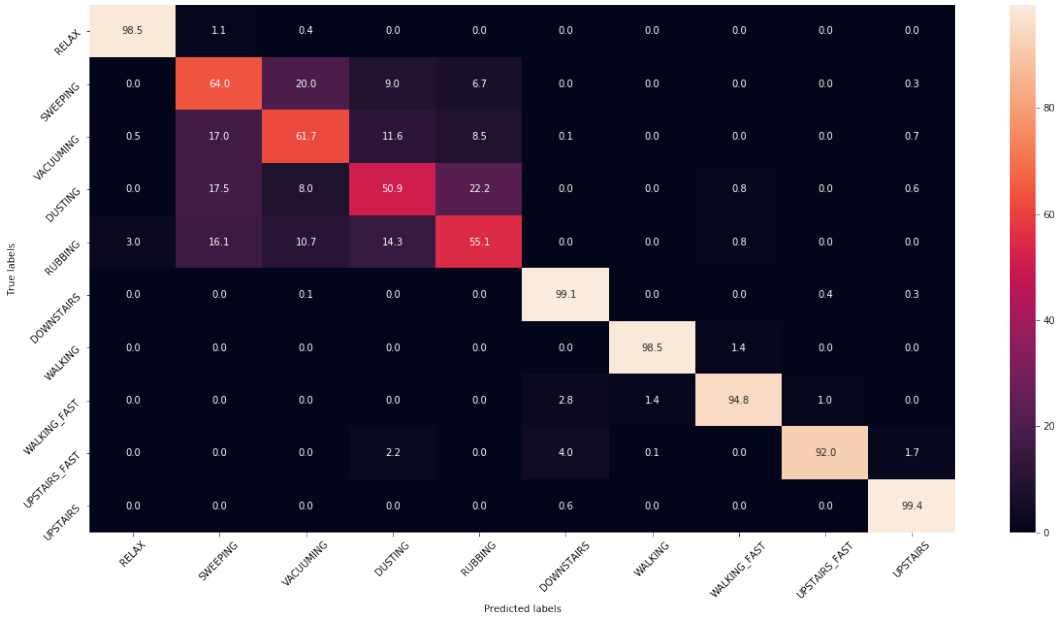

Figure 6.5: Average Confusion Matrix obtained with Ankle Data, limited on ankle-related activities

activities: *ambulation, cycling, sedentary and other*. Yang et al. [YCL+07] has categories of activities more similar to our scenario: *walking, running, scrubbing, standing, working at a computer, vacuuming, brushing teeth* and *sitting*. For what concerns the features, the ones used in all the experiments are based on three different aspects: (1) time (mean, standard deviation, mean absolute deviation, etc. of acceleration over time); (2) frequency spectrum (first dominant frequencies and their power in some particular ranges - e.g. [0.6, 2.5] Hz) and, (3) wavelet, based on the Discrete Wavelet Transform, therefore obtaining features linking both the frequency and the time domain.

Regarding the window length, in the aforementioned works, this parameter varied from 2.0 and 12.8 seconds, or, in terms of the number of samples, from 100 samples to 1152 (but taken at different frequencies). Those studies that compared the performances over the same data but using different lengths for the windows (e.g. [MIR+13]) have shown that longer windows would have meant higher performances, but also that windows with a length of 4.0 seconds were sufficient to obtain acceptable results.

Regarding the used ML algorithms, Zhang et al. [ZRMH12] tested performances using different algorithms (Decision Trees, Naive Bayes, Linear Regression, Support Vector Machine, Neural Networks) showing that all the algorithms had good performances (higher than 95.0%), with the DT and SVM being the ones with better results. Experiments in [MIR+13] used Support Vector Machine only, while in [YCL+07] a *neuro-fuzzy* classifier (classifiable as a Neural Network) has

been used. All of the analyzed algorithms in the documents obtained good and almost similar results, despite the used algorithms (overall accuracy always over 86%).

The major differences with respect to our baseline approach regard: (1) the length of the window (we adopted shorter windows of 2.0 seconds) and, (2) the number of activities to be classified (higher in our case).

Chapter 7

Reducing Misclassification of *Unknown* Activities

Introduction

In this chapter, it will be described in details the problem of misclassifications of *unknown* activities and how it has been tried to improve our baseline approach taking into consideration this issue. In particular, after introducing the problem, it will be presented a brief state of the art on the issue. Then, our proposed approach will be illustrated in detail, for each of its main parts (i.e., ensemble approach, correction of misclassification, and voting system). To conclude, the proposed approach will be evaluated on the Healthy Adults dataset (see Chapter 4), by comparing results with the baseline approach (see Chapter 6).

7.1 Overview

As briefly introduced before, different works are present in the literature dealing with the activity recognition problem and using wearable device data, but most are limited to recognizing a predefined set of activities. The problem of these kind of approaches is that the classifiers will frequently and hopefully obtain high accuracy when tested only with the predefined activities.

Indeed, typically, in a real-life context, those few activities that specialists are interested in monitoring in cases of a particular disease (e.g. walking, going upstairs when treating cardiological diseases) are actually performed in the middle of hours of recordings in which patients perform a great variety of ADLs. What happens in practice is that classifiers, when dealing with real-life scenarios, need to recognize an activity of interest overwhelmed by hours of recorded data

related to not labeled that are not easily recognizable. This scenario is strongly different from the typical approaches (and datasets) described in the literature, where the various activities are clearly defined. Moreover, with the perspect of using an AR system as a long-term instrument, this problem translates into a practical issue of obtaining false positives in classification.

As an example, we can consider the approach described previously in Chapter 6. The final model was trained on a limited set of activity, where, for instance, the *cooking* activity was not present. Supposing to obtain 24 hours of recording of a patient, the pieces of recording in which the patient was cooking, could be wrongly labelled as *washing face*, for example, that was one of the activities learnt from our classifier. This, on a large-scale, could be transformed in a huge complication.

In this chapter, therefore, we report a preliminary analysis of the literature regarding the problem of reducing the misclassification between activities of interest and unknown activities due to the presence of data of generic, not well defined, daily activities that act as a confounding factor for the typical classifiers. Afterward, we describe our proposed approach based on a combination of an ensemble method, a filtering technique, and a voting mechanism to improve the classification accuracy and mitigate the misclassification problem. Finally, we compare the results obtained with our approach with the ones from chapter 6.

In the following, Section 7.2 reports the actual state of the art and related works. In Section 7.3 we describe our proposed approach, while Section 7.4 presents the results quantifying the improvements due to the proposed approach. Conclusions and possible ways to improve this approach are given in Section 7.5, which concludes the paper.

7.2 Background and State of the Art

In this section, we will analyze the state of the art regarding activity recognition in real-life scenarios with a particular focus on one of the main limitations of the solutions currently proposed in the literature: the *misclassification* problem between activities of interest and unknown activities. Such a problem is due to the fact that models are usually trained on data recorded in controlled scenarios, where subjects consistently perform a series of activities, each for a determined period of time (e.g., 2 or 5 minutes). The same model is then tested on data that are recorded in the same way, including only data related to *known activities*: the predefined list of activities of interest performed while recording data. When dealing with real-life data, instead, possibly recorded for an entire day, subjects usually perform *other* activities that cannot be easily considered during the model training since they cover a huge variety of possible movements and they are neither well defined nor repeatable. The *misclassification* problem affects the actual accuracy achievable in real-life scenarios, since many *unknown activities* are classified as one of that of interest, and vice versa. This happens even if a generic label *unknown activity* has been included in the training data.

For this reason, we investigated how to improve the approach described in Chapter 6 with the aim of mitigating this problem. We started by analyzing the literature on this topic, and we found that the majority of works related to activity recognition systems simply do not consider the possibility of classifying data that are not related to activities of interest. Indeed they focus their attention only on the ability to recognize a predefined list of different activities using data related to only such activities. A few works propose solutions to this problem, for instance, by using a *threshold* approach on a *trust* measure, and classifying only data that had a trust measure higher than the threshold as an activity of interest [WGC+10, SVLS08].

We found few valuable works. One is by Nguyen et al. [NZTZ15], which focus on the topic of false positives in activity recognition systems, similarly to our case. Authors propose a novel possible solution based on an approach that takes advantage of both labeled training data and of additional unlabeled data. Another interesting work is the one of Garcia et al. [GCBCJG14] in which they present a dataset that includes labeled data of non-interesting activities, also showing the possible results of the classification of this additional class. Since we have not been able to find alternative hints in the literature on how to possibly reduce this classification problem, we have tried to find solutions by taking advantage of *ensemble learning*.

There are different available approaches where different classifiers are combined together, according to the *stacking* technique. For example, Alsheikh et al. [ASN+16] in their work wanted to perform human activity recognition using triaxial accelerometer data; to this end, a combination of Deep Learning and Hidden Markov Model (HMM) has been used with success. In 2014, Ha et al [HR14] have used different ensemble methods with different classifiers (e.g. Decision Tree, k-NN, and Random Forest) to improve the recognition of 12 activities using data from smartphone accelerometer.

Moving to related works that also include a voting system among different classifiers, we can cite Catal et al. [CTPK15], that have used a combination of J48 (Java implementation of Decision Tree), Logistic Regression and Multi-Layer Perceptron to increase the activity recognition accuracy. Similarly, Bayat et al. [BPT14] tested multiple combinations of classifiers for an analogous task, showing that the combination of Multi-Layer Perceptron, LogitBoost, and Support Vector Machine was the one obtaining the highest accuracy.

For what concerns the idea of filtering possible outliers in predicted labels by ”smoothing” the list (i.e., reducing the high-frequency variance in the list), we can refer to two works. In the first one, by Phan [Pha14], the author proposes a method based on pruning decision trees to identify spurious classifications. In a second work, by Shoaib et al. [SBI+16] authors propose a technique able to reduce the influence of random small pauses between actual motions; however, this method had not been tested in practice by authors.

In this context, it is quite difficult to actually compare the various approaches available in the literature to recognize activities. In particular, the lack of implementations of proposed solutions and of a benchmark dataset used for comparing results is significant. Moreover, as anticipated,

Table 7.1: List of Activities performed

1. Relaxing on a chair	7. Brushing Teeth	13. Downstairs
2. Keyboard Typing	8. Sweeping	14. Walking
3. Using the Laptop	9. Vacuuming	15. Walking Fast
4. Handwriting	10. Eating (a soup)	16. Upstairs Fast
5. Washing Hands	11. Dusting surface	17. Upstairs
6. Washing Face	12. Rubbing surface	

most of the works related to activity recognition do not consider the problem of classifying activities of interest performed between other unknown activities. In the following sections, we will describe our approach that tries to fulfill this gap.

7.3 Approach

In this section, we will describe in detail the proposed approach that allowed to reduce the number of misclassifications of unknown activities while, at the same time, increasing the overall accuracy of the classification of the activities of interest. The approach we propose in this paper is an improvement of the baseline approach we presented in Chapter 6. In the following, we first summarize the *baseline approach* and some details on the *dataset* employed, the one presented in Chapter 4. Then, we describe the multi-classifier extension and the two additional techniques that allowed us to improve the obtained results: the first one is based on a method aimed to *detect and remove outliers* and the second one is based on an ensemble method that implements a *voting system*.

7.3.1 Dataset and Baseline Approach

Concerning the *dataset* employed (described in detail in Chapter 4), hereby we are going to resume the main characteristics. The Healthy Adults

dataset contains the recordings of 17 ADL, listed in Table 7.1, performed by 8 different healthy subjects, while wearing three medical-grade wearable devices on their body: (1) one at the dominant wrist sampling at 256 Hz, (2) one at the right hip sampling at 100 Hz and (3) one at the right ankle sampling at 100 Hz. However, for these analyses, we have only considered data obtained from wrist devices. In addition to the 17 ADL, the raw data published online (available on Harvard Dataverse[1]) also contain unlabeled data recorded between the execution of the various activities: we relabeled all this data with the generic label Other.

[1]https://doi.org/10.7910/DVN/G23QTS

To resume the *baseline approach*, the first step concerns extracting the feature vectors and the associated labels, by using a *sliding window* approach, considering only the accelerometer data. During the features extraction, the sliding window passes over the data and, for each axis (X, Y, Z), different measures are extracted from the data contained in the window: mean, variance, standard deviation, median absolute deviation, percentiles (10Th, 25Th, 75Th, 90Th), obtaining a total of 24 features. While preprocessing our raw data to extract features, we created a sample approximately every 0.1 seconds of our recordings. This happened because we used windows that were 2.0 seconds long with a 95% overlap, therefore creating approximately 10 samples for each second of raw data.

After extracting the features, and splitting our data into training and test sets (75% and 25% of data, respectively), the approach requires training an SVM model with the data and then testing its accuracy to discern the 17 aforementioned activities.

Limitations. The focus of the baseline approach was limited to recognizing a predefined set of activities from a test set containing only well-defined activities. For this reason, in this study, we focus our attention on discerning our activities of interest also from ”other” data. A single SVM classifier, as described in our previous Baseline Approach, trained to classify also the ”other” data, has shown to be unable to give satisfying results In the following, we describe the extensions that improved the results.

7.3.2 Using Multiple Classifiers

Analyzing the results obtained by the Baseline Approach (SVM-based), we have reasoned about the fact that every Machine Learning algorithm can behave differently on different portions of the same dataset. Different classifiers can have strengths and weaknesses, and in the cases of weaknesses, the problem leads to erroneous classifications. At the same time, we had in mind the principle of ensemble learning, i.e., combining multiple classifiers to potentially increase the accuracy, efficiency, and robustness w.r.t. the single classifier [RG05].

Thus, our approach is based on the idea of ensemble learning and combines the results of different classifiers:

1. a Support Vector Machine model (SVM), as done in our previous Baseline Approach
2. a Decision Tree model (DT)
3. a Random Forest model (RF)
4. a k-Nearest Neighbor model (k-NN)
5. a Gaussian Naive Bayes model (GNB)

The choice of these particular algorithms, placed side by side with the baseline approach based on SVM, lies in the fact that all of them are among the most commonly used algorithms in activity recognition tasks [bANS+12, YWM+19, LL12]. Moreover, as previously mentioned, we have tried to combine algorithms based on different foundations and theories to avoid the pitfalls that a solution based on a single classifier can face. Thus, in conclusion, starting from an input dataset containing 24 features (the same described for the baseline approach), we executed five classifiers instead of only one and obtained as output five lists of candidate activities.

7.3.3 Detecting and Removing Transient Misclassifications

By analyzing the predicted labels obtained from every single classifier, we discovered that often many *transient misclassifications* appear in the list of predicted labels. In our case, we would call *transient misclassifications* those few samples that are predicted as label k in the middle of a long list of labels j. We often observed in the output provided by the various algorithms that in the middle of a long list of labels corresponding to a particular activity, few samples are predicted as a completely different activity. For example, while performing the *walking* activity for a minute, a subject would have started to *brush teeth* for 3 seconds only: this kind of behavior is due to short movements that are wrongly associated by the classifier to another activity (e.g., brush teeth).

By also considering a minimal reasonable length of an activity (in seconds) we decided to try to detect these misclassifications and correct them, by ”smoothing” the list of labels that we can obtain from each of the considered algorithms. To filter out these errors, we have used a method based on a sliding window, consisting of computing the mode over a set of labels. More practically, a window of size *window_size* passes over the labels and computes the mode, which is the most present value in the window. The extracted mode is *centered*, meaning that the computed label at position i will correspond to the mode of the original labels in the interval [i - *window_size/2*, i + *window_size/2*]. At the beginning and the end of the labels list, we have taken partial intervals to calculate the mode. In this way, we could obtain a new list of the same length as the original one, containing the labels filtered from misclassifications, ready to be used in the next step. For example, if an activity appears, erroneously, in the results of a classifier for 0.5 seconds (i.e., about 5 consecutive labels), considering a window of size 5 seconds (e.g., 50 labels), such activity will be totally corrected from the results after completing this step.

Thus, in conclusion, starting from the five lists of candidate activities computed with the five considered algorithms, our approach produces five lists where the transient misclassifications are removed.

7.3.4 Voting system

The final step of our approach takes as input the five lists of labels already filtered from misclassifications as described in the previous subsection and produces a final unique list of labels by relying on a *voting mechanism*. Since we have decided that each classifier has equal weight, we simply consider the majority of votes to determine the final label of each sample. More in detail, our five classifiers are trained to classify also data labeled as ”other”. Thus for each sample, we then set as the final label the label that is present more than three times in the classifiers' predictions for that sample; if the most present prediction has no majority, we set ”other” as the final label for that sample.

7.4 Evaluation of the Proposed Approach

The main *goal* of this section is to illustrate and discuss the results achieved by our approach and compare them with the results achieved with the simple SVM model presented in Chapter 6. Thus we have tried to understand if the proposed approach is actually able to improve the accuracy and efficiency with respect to the baseline approach.

When comparing the accuracy and efficiency of the proposed method, we will not only compare the obtained results by using the F1 score: we will also consider as measures (1) ”other” data correctly labeled as ”other” (from now on True Positives Other; TP Other for short) and (2) the average amount of data regarding activities of interest that are *wrongly* labeled as ”other” (from now on average False Positives Other; Avg FP Other for short). Considered metrics help us to measure the problem of misclassification and to make comparisons among the results achievable with different approaches.

It is useful to note that for these analyses, every model that we created during our tests was subject-dependent, meaning that we trained a model on the training data of a subject s and we tested that model on the test data of the same subject s. After every test, we computed the confusion matrix for each subject. We then averaged over all eight subject's data.

As a first step, we motivate the parameters' values used for defining the size of the windows used for computing the mode while filtering outliers.

7.4.1 Finding the proper window size for computing the mode

Hereby we report how we found an adequate window size for computing the mode used for filtering the outliers. To look for the best *window_size* to compute the mode, we have used our training data to estimate the results we could achieve in our final model, with a cross-validation approach. Keeping in mind that 10 samples are roughly associated to 1 second of raw data, we

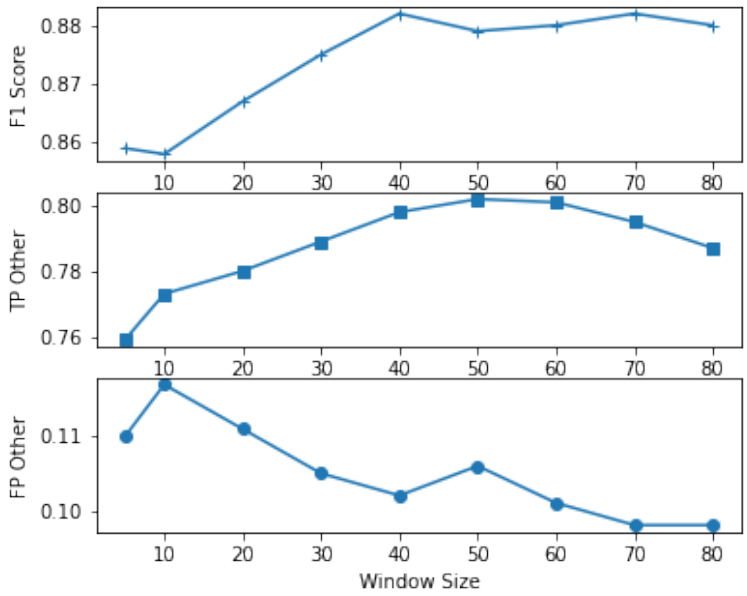

Figure 7.1: F1 Score, amount of TP Other and FP Other changing with different window sizes to compute mode while filtering outliers in training data.

have tested the values in this list for the window size: *[5, 10, 20, 30, 40, 50, 60, 70, 80]*. We could think of these values as being associated with windows that are from 0.5 seconds to 8 seconds long, thus including a reasonable minimal length for each activity.

We have measured the F1 Score, TP Other, and average FP Other as the window size was changing. In Figure 7.1, we can see how these measures vary over different sizes of windows.

The plots show that according to the tests we performed with training data, F1 Score and TP Other are maximized and FP Other is minimized when the window size is equal to 50. Therefore, in the following, we will refer to tests where the window size used to compute the mode during the filtering outliers step is set to 50 (i.e., about 5 seconds).

7.4.2 Baseline model

In the following, we will describe our starting point for the evaluation, which is the confusion matrix that we obtained by training an SVM model with the goal of being able to classify all activities of interest and also data labeled as ”other”. The confusion matrix shown in Figure 7.2 is averaged over all eight subjects’ results. In that confusion matrix, the labels in the interval [1, 17] correspond to activities of interest as listed in Table 7.1, while the label ”X” corresponds to the label ”other”.

In this scenario, the baseline approach is able to achieve an F1 Score of 0.86 $\pm$ 0.05 (mean $\pm$ standard deviation over different subjects), an amount of TP Other of 0.72 (meaning that 72% of data labeled as ”other” was correctly classified as ”other”) and an average amount of FP Other of 0.10 (meaning that on average, 10% of data of a random activity of interest would have been wrongly classified as ”other”).

In different attempts to reduce the amount of misclassified samples, we have been able to increase the TP Other, but as a side effect, we were also proportionally increasing the amount of FP Other, damaging the improvements. In the following, we will see how each step of the approach we propose in this work has changed the accuracy of our results.

7.4.3 Improvement due to Detecting and Removing Misclassifications

Here we quantify how the filtering phase has increased the ability of our classifier to recognize data labeled as ”other” correctly and to reinforce the overall accuracy without negatively influencing the number of misclassified samples.

In Table 7.2, it is possible to see the results of the singular classifiers, averaged over different subjects, before filtering labels. As in the previous analysis, to compare results, we will consider the number of True Positives ”other”, the average amount of False Positive ”other”, and the F1

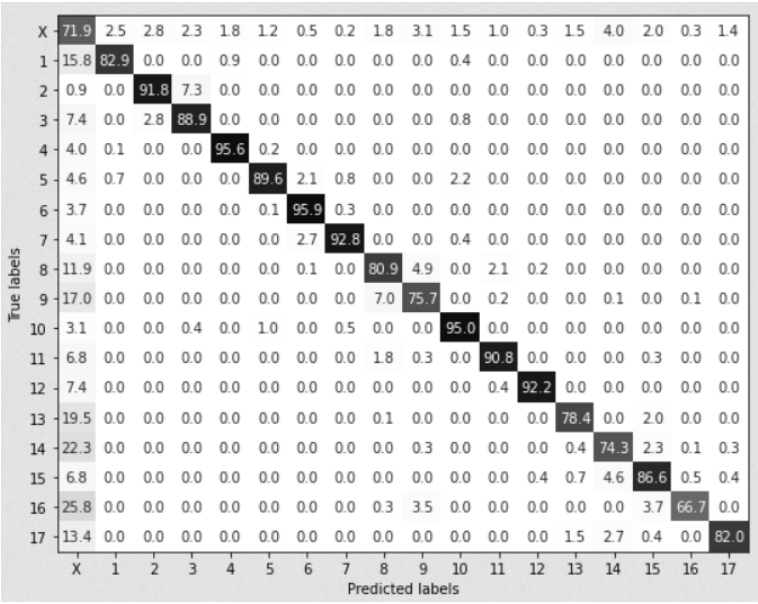

Figure 7.2: Confusion Matrix obtained using a simple SVM model, trained to classify also ”other” data, here labeled as ”X”. The cell color changes according to the accuracy in the matrix, using a scale from 0 to 100: the higher the value, the darkest the cell.

Score. We can see that regarding TP Other, the highest value is achieved with the Random Forest algorithm. The lowest average FP Other is obtained with the Naive Bayes algorithm, while we have the higher F1 Score when using the Support Vector Machine.

In Table 7.3, we can analyze how this step influences the results. We can notice that every measure has improved: any TP Other and F1 score is higher than in the respective previous cases and any average FP Other is lower than in the corresponding cases with no filtering. Averaging over the different classifiers, the TP Other has increased of 0.054, the F1 Score has augmented on average of 0.047 and the average FP Other has decreased by 0.010.

To resume, this step based on extracting the mode of an interval of labels, in order to reduce the transient misclassifications, has been able to improve the accuracy and efficiency of any model, in any of the measures we were interested in.

7.4.4 Improvements due to the Voting system

Hereby we will describe how the accuracy and efficiency of a classifier could change when introducing the subsequent voting step, as described in Section 7.3.4.

After having filtered our labels obtained by each of our five classifiers, as previously anticipated, for each sample, we take the label that has been predicted by the majority of classifiers; if there is no majority, we set ”other” as label.

In Figure 7.3, it is possible to see the confusion matrix averaged over all eight subjects results. As in the previous confusion matrix (Figure 7.2), the labels in the interval [1,17] correspond to activities of interest as listed in Table 7.1, while the label ”X” correspond to the label ”other”.

With this final phase, where we combine the filtering method and the voting system, we are able to achieve an F1 Score of 0.90 ± 0.05 (mean $\pm$ standard deviation over different subjects), an amount of TP Other of 0.80 (meaning that 80% of data labeled as ”other” was correctly classified as ”other”) and an average amount of FP Other of 0.09 (meaning that on average, 9% of data of a random activity of interest has been wrongly classified as ”other”).

To better understand how the accuracy in recognizing each singular activity changes from our base SVM model and the final approach, we have plotted the variation of results for each ac-

	SVM	**DT**	**RF**	**KNN**	**GNB**
TP Other	0.719	0.659	0.765	0.602	0.377
AVG FP Other	0.103	0.162	0.125	0.084	0.023
F1 Score	0.848	0.760	0.836	0.817	0.771

Table 7.2: Results achieved by each classifier without filtering labels

	SVM	DT	RF	KNN	GNB
TP Other	0.766	0.761	0.814	0.660	0.393
AVG FP Other	0.093	0.139	0.118	0.079	0.018
F1 Score	0.890	0.832	0.871	0.864	0.812

Table 7.3: Results achieved by each classifier after filtering labels step

tivity in Figure 7.4. It is easy to notice that every activity, except *going downstairs*, and barely, sweeping, has received an improvement in its recognition.

We started the evaluation of this approach with the goal of measuring if it is actually able to improve the accuracy and efficiency with respect to the baseline results. In conclusion, to better analyze the outcomes, we used a paired Wilcoxon test to compare the two approaches on the values of the metrics obtained for each subject. We decided as it is customary, to accept a probability of 5% of committing Type-I-error (α). For the F1 Score, TP Other, and FP Other we obtained respectively a p-value of <0.01, <0.01, and 0.08, meaning that the observed differences among the baseline approach and the one proposed in this work are statistically significant for the F1 Score ad for TP Other metrics.

To summarize, with respect to the previous SVM-based baseline approach, our current proposal is able to increase the F1 Score of 5.5%, the amount of TP Other has increased for a total of

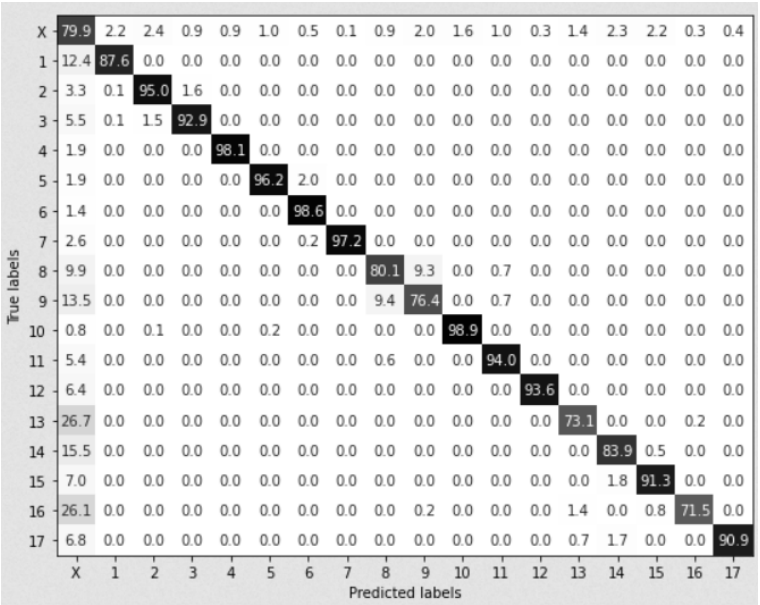

Figure 7.3: Confusion Matrix obtained after filtering outliers and by taking the most present label among five different classifiers' predictions.

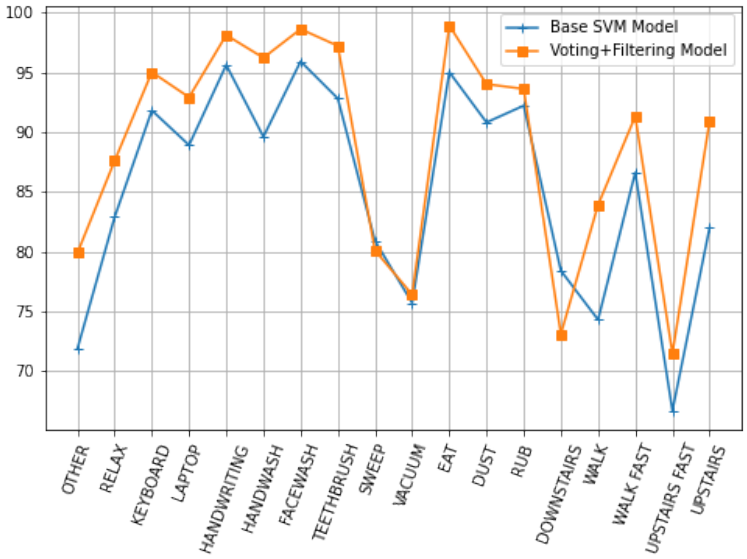

Figure 7.4: Variation of accuracy for each activity between base SVM model and final approach.

11.1% (equivalent to a reduction of 28.5% of misclassified data), and average FP Other has decreased of 15.5%. The observed differences for both F1 Score and TP Other are statistically significant.

7.5 Discussion

In this chapter we have considered the context of the activity recognition tasks, with the perspective of monitoring patients to track their physical activity and health status.

More particularly, in the context of activity recognition and about reducing the amount of misclassified samples when dealing with real-life scenarios and daily recordings, we have proposed a method able to increase accuracy and efficiency compared to a simple SVM-based classifier. The approach, in fact, based on an ensemble method with a voting system and on a technique able to filter misclassifications has been able not only to increase the amount of ”other” data correctly classified, but also to increase the overall accuracy measured as F1 Score and at the same time to reduce the amount of data that were wrongly classified as ”other”.

This approach could be easily extended by varying the combination of machine learning models used in the voting step, and possibly customized with additional steps using contextual information. For instance, if physicians were interested into considering only activities performed for more than 5 minutes, the discussed approach could be easily adjusted according to requests.

Chapter 8

Method Improvements and Explainability

8.1 About Lengths and Overlap of Sliding Windows

It might be useful to notice that for the previous analyses described in Chapters 6 and 7, we have always been using our Healthy Adults dataset (see Chapter 4) that includes data from 8 healthy subjects. In addition to this, we have always based our approach on *sliding windows* that were 2.0 seconds long, with 95% of overlap between two subsequent windows. We recall that these windows are used to perform feature extraction. The choice about the windows' length and overlap was made according to results achieved with tests that are described in the following.

The final accuracy and efficiency of a model highly depend on the feature extraction phases, as aforementioned, and in our case, this phase was subject to the windows' length and overlap. For this reason, we decided to check how the accuracy of the model would change, with respect to different lengths and overlaps of windows. For these tests, we considered the baseline approach only, in order to reduce the requested computational time.

We tested window of lengths equal to [1.0, 2.0, 3.0, 4.0, 5.0, 6.0] seconds, with an overlap of [20%, 40%, 60%, 80%, 90%, 95%]. Both the lower and upper bounds of the selected lengths were defined according to our particular scenario. Indeed, having windows with a length that was lower than 1.0 seconds would have no sufficient data to extract useful features. At the same time, having windows that were longer than 6.0 seconds could have produced (with certain values of overlap) too few samples to train a model.

Achieved results are summarized in Figure 8.1, where for each window length and for each percentage of overlap we report the obtained accuracy, averaged over different subjects, using only training and validation data.

Considering the lengths of the windows one by one, it is easy to see that by increasing the overlap the score always increases as well. This happens because by increasing the overlap we manage

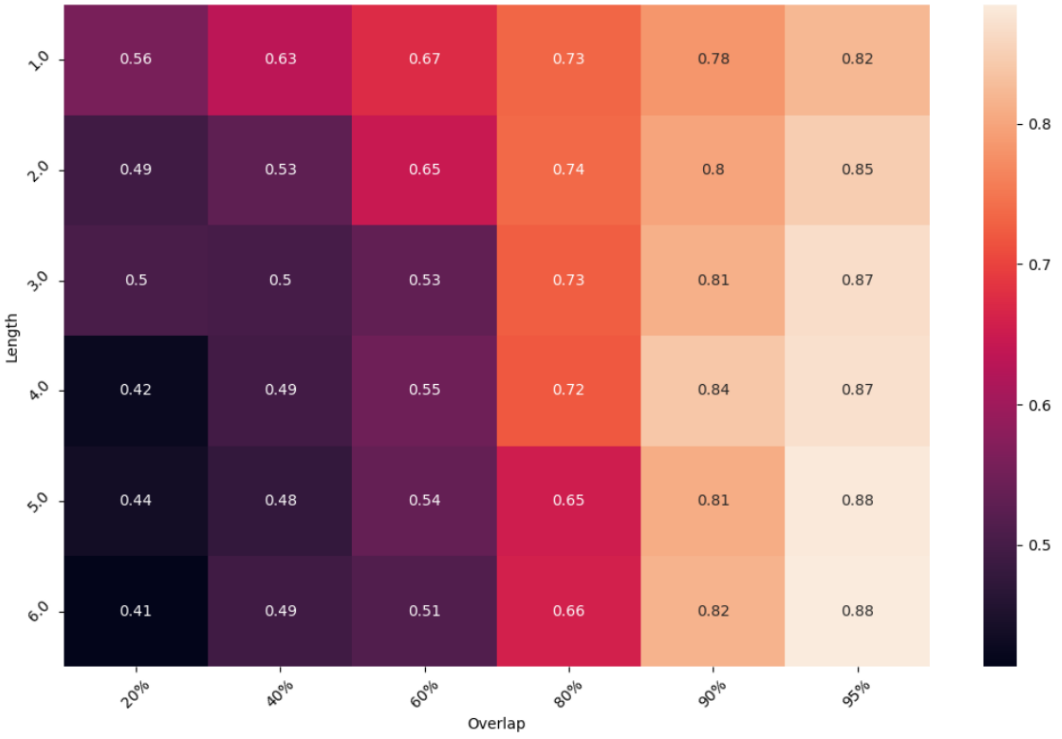

Figure 8.1: Heatmap representing the average accuracy according to different window lengths and overlaps.

to create more data during the feature extraction phase, and having more data to train the model can determine a huge increase in accuracy. We can not say the same thing when considering the overlap percentages, individually. In fact, for overlaps that go from 20% to 80%, the accuracy gets lower by increasing the length of the window; on the opposite, the accuracy gets higher by increasing the length of the windows for overlaps equal to 90% and 95%.

The higher scores, retrieved using only training data, could be achieved using windows that are 6.0 seconds long with 95% of overlap. Nevertheless, considering the possible idea of using the AR system to also predict the performed activity almost in real-time (for better monitoring the patients), we would accept having a slightly lower score in order to prefer shorter windows (i.e., 2.0 seconds long).

8.2 Distribution of Votes Among Different Classifiers

In Chapter 7, we discussed the approach devised to reduce misclassification of unknown activities based on three main concepts: (1) an ensemble of different classifiers, (2) a procedure to remove

	SVM	DT	RF	KNN	GNB
1st	8	0	21	3	0
2nd	5	4	5	17	1
3rd	1	12	4	11	4
4th	0	7	2	1	22
5th	18	9	0	0	5

Table 8.1: Ranking of classifiers according to their accuracy achieved

transient misclassifications, and (3) a final voting mechanism. In order to further analyze this approach and better understand its behavior at a low level, we tried to evaluate if there was a particular classifier that was prevailing over others and in general to check the strengths and weaknesses of each classifier. For instance, if a classifier achieves a higher accuracy over others most of the time, its weight could be increased during the *voting phase* of our approach. To analyze the behavior of classifiers, we have considered the distribution of votes over different subjects and runs. In particular, by considering 8 subjects and for each one the results of a 4-fold cross-validation, we could check the outcomes of 32 runs. These data have been performed using only training and validation data.

A first test was performed to understand the amount of agreement over the different classifiers for each of the test samples. In practice, for each sample, we wanted to measure if there was a common choice among only two/three/four classifiers or all of the five classifiers.

We have seen that averaged over 32 runs, we had these levels of agreement:

- in 2% of samples, there was agreement among 2 classifiers
- in 16% of samples, there was agreement among 3 classifiers
- in 46% of samples, there was agreement among 4 classifiers
- in 36% of samples, there was agreement among 5 classifiers.

These results have shown that for the majority of cases, 4 classifiers over 5 labeled a sample as the same activity. This means that there is quite often a consistency of results of different classifiers.

These results, by the way, were not helping us in determining if there is a prevalent classifier or not. For this reason, we ranked the classifiers, for each run, according to the achieved accuracy. In particular, for each run, we took note of the accuracy of each model in classifying unseen data and ranked their results. We gathered the outcomes in Table 8.1, showing for each classifier the number of times it was ranked as 1st, 2nd, 3rd, 4th, or 5th.

# PC	Variance	#PC	Variance
1	0.378	7	0.991
2	0.638	8	0.994
3	0.790	9	0.995
4	0.888	...	...
5	0.954	23	0.999
6	0.984	24	0.999

Table 8.2: Changing of cumulative variance with respect to the number of Principal Components, averaged over different subjects

From Table 8.1 we could see that, at first sight, the classifiers with better rankings were mostly Random Forest (RF) and k-Nearest Neighbours (KNN). At the same time, we also tried to cluster these results by subjects, noticing that different classifiers were performing better for different subjects. This let us realize that it would have been safer to keep each classifier with equal weight in the voting phase, to achieve good accuracy for different types of subjects that we could deal with.

8.3 Analysis of Principal Component Analysis

Principal component analysis (PCA) is a mathematical algorithm that allows the reduction of the dimensionality of data while keeping most of the variations present in the data set. Here we will describe the tests that we have performed to understand whether and how the application of PCA on our data could influence the accuracy and the recognition of data labeled as ”other”.

Starting from our 24 features, we have analyzed the variation of cumulative percentage of variance explained, according to the number of used Principal Components, averaged over all subjects. In Table 8.2, we list how the cumulative variance increases with respect to the number of used principal components. As it is evident from the values in the table, by using our training data, we have seen that by taking only the first 7 principal components we could obtain a variance higher than 0.990.

In the following subsections, we have analyzed how results were changing by using or not PCA. In particular, we have compared our approach with and without PCA on (1) activity recognition with our dataset *excluding* the ”other” activity, and (2) activity recognition with our dataset *including* the ”other” activity. The following tests have been made considering the baseline approach (see Chapter 6).

No PCA	With PCA
0.93±0.03	0.91±0.04

Table 8.3: Comparison of accuracy, excluding ”other” data, when using PCA or not

8.3.1 PCA and Activity Recognition excluding ”other” activity

The goal of this preliminary test was to understand how the addition of PCA could influence the performance of the classifier on the baseline approach considering, therefore, the easiest condition. Thus, ignoring the problem of false positives that occur when including data labeled as ”other” activity to be recognized.

By using our Healthy Adults dataset (filtering out data labeled as ”other”) and comparing the performances when using or not PCA, we have seen no improvements in accuracy. Rather, we have noticed a reduction of the F1-score of 0.02 when using PCA. In particular, we have reached an average F1-score of 0.93±0.03 (mean ± standard deviation) when not using PCA. In the case where we have applied PCA before training our classifier, we have reached an average F1-score of 0.91 ± 0.04. In Table 8.3 we report the resulting comparison. On the other hand, we have seen that PCA could doubtlessly help us reduce training time. Indeed, when using 7 principal components instead of our starting 24 features, we could reduce the training time by approximately 34%, with a slight impact on the accuracy.

8.3.2 PCA and Activity Recognition including ”other” activity

In this test, we have tried to understand if the application of PCA to our data in our approach could reduce the number of false positives. In our opinion, by transforming our data we could, possibly, alter it in such a way that was easier for the classifier to recognize ”other” activities from activities of interest.

In this particular case, we will not only compare the alternatives using the F1 score: we will also consider as measures (1) ”other” data correctly labeled as ”other” (from now on True Positives Other - TP Other -) and (2) the average amount of data regarding activities of interest that are *wrongly* labeled as ”other” (from now on average False Positives Other - Avg FP Other -).

Unfortunately, even in this test, we have seen no improvements by using PCA, neither on the overall F1-score nor on the measures regarding False Positives (TP Other and Avg FP). Table 8.4 resumes results achieved. Nevertheless, with this second test regarding PCA we could confirm the reduction of the amount of time requested for the training phase: when applying PCA the requested time has been reduced by 35%.

	No PCA	With PCA
F1 Score	0.86±0.05	0.83±0.05
TP Other	0.72±0.03	0.69±0.03
Avg FP Other	0.10±0.01	0.10±0.01

Table 8.4: Comparison of accuracy, averaged over different subjects, when applying or not PCA

8.4 Usage of Additional Features

It might be useful to note that our proposed approaches make use of features based on measures of accelerometer data over time but do not consider other possible domains or metrics. In our distinct scenario, we could additionally test the extraction of other additional features related to frequency analysis or to the correlation between axes. Concerning the features associated with the frequency, we relied on the Fast Fourier Transform (FFT) to analyze which frequencies characterized the signal in each window used to extract features.

Considering the literature, we decided also to compute these new features:

- **Amplitude of first and second peaks in FFT output (frequency-based)**
  Computing the FFT over a window of our signal, we took the amplitude of the first and second peaks for each axis (i.e., having 2 features and 3 axes, we produce 6 features);
- **Frequencies associated with first and second peaks in FFT output (frequency-based)**
  Considering the previous computed features, we also took the frequency associated with each peak, for each axis (i.e., having 2 features and 3 axes, we produce 6 features);
- **Energy (frequency-based)**
  Computed as the sum of the squared discrete FFT component magnitudes of a signal, the sum is then divided by the window length for normalization (i.e., having a single feature for 3 axes, we produce 3 features);
- **Correlation between axes (time-based)**
  Computed as the ratio of the covariance and the product of the standard deviations (i.e., having a single feature for 3 couple of axes, we produce 3 features)

For this particular test with these additional features, we needed to change the length of the sliding window used to extract features. Indeed, we were usually making use of 2.0 seconds-long windows, but there was little data to extract frequency-based features. For this reason, in this singular test, we moved to windows of 4.0 seconds length, with 0.1 sec of overlap. For this reason, we also needed to adjust our approach concerning the reduction of transient misclassification.

We trained a model based on our approach that reduces the misclassification of unknown activities, using a complete feature set, which comprehended features in the time domain (i.e. mean, variance, standard deviation, median absolute deviation, percentiles, and correlation between axes), and features of the frequency domain (i.e. energy, frequency peaks of the Fourier transform, peaks value of the Fourier transform), for a total of 42 feature. For comparison purposes, we will consider the F1 score, the amount of "other" data correctly labeled (as "TP Other"), and the average amount of "other" data wrongly classified as activities of interest (as "AVG FP Other").

With these settings, we could achieve an F1 score of 0.88 ± 0.04, TP Other 0.69 ± 0.05, and average FP Other of 0.16 ± 0.03. In Table 8.5 we report the comparison of these results with the ones achieved using only the temporal features.

	Temporal Feature Set	**Complete Feature Set**
F1 Score	0.89	0.88
TP Other	0.72	0.69
AVG FP Other	0.09	0.11

Table 8.5: Comparison of results between using a temporal features set or a complete one (comprehending temporal and frequency features)

From the achieved results, we can say that the usage of additional features did not increase the F1 score, nor the average "False Positives Other" metrics. There has been a minor increase in other data correctly labeled.
To conclude, these new additional features, when used altogether with other features previously used in our models, were not helping our classifiers achieve higher accuracy.

8.5 Feature Importance

We had the desire to understand if, among all the considered features, some had more *importance* in our model than others, meaning that their values were more helpful than others in classifying data. It is useful to remember that, currently, by using a sliding window approach, the larger considered feature set is composed of features in the time domain (i.e. mean, variance, standard deviation, median absolute deviation, percentiles, and correlation between axes), and features of the frequency domain (i.e. energy, frequency peaks of the Fourier transform, peaks value of the Fourier transform), for a total of 42 features. Indeed, it is not a large number of features, but the analysis could still give us results useful to understand the behavior of our approach and to leverage certain features to improve accuracy.
There are different possible ways to consider feature importance, and a possible method is by using Lasso Regression. In the following, the method will be introduced and results will be discussed. In conclusion, we present the validation of results using bootstrapping techniques.

8.5.1 Lasso Regression for Feature Importance

This technique is based on the idea of *Linear Regression*¸ and it tries to reduce a minimization function that involves weights associated with features. To simplify the concept behind it, when two features are linearly independent, it will empathize their weights, and reduce the weights of other features to zero. As a result, we could then take a look at the weights eventually associated with features to select the best features.

To understand the influence of the features on our dataset, we have performed two tests. For both tests, we compared the accuracy scores achieved with the *complete feature set* with the ones achieved using the *reduced feature set*. In particular, in the first test, the reduced feature set was not unique, meaning that for each subject we were considering a different reduced feature set, selected according to the results of Lasso Regression.
In the second test, the reduced feature set was fixed among different subjects, meaning that we have removed the features that were discarded the most in the first test. All of the tests made use of our approach based on the ensemble method and the voting system, discussed in Chapter 7, and considering the training data only.

Preliminary analyses were made to find the right value for the regularization parameter, which was then fixed for the following tests. Then we performed the first test with a non-fixed reduced feature set. Results, in Table 8.6, show that the accuracy is substantially similar to the one using the entire feature set. The average On average, over 42 starting features, only 17 were kept. This also suggests that a quite high number of features are correlated to each other.

	Non-Fixed Reduced Set	**Complete Set**
F1 Score	0.86	0.86
TP Other	0.81	0.81
AVG FP Other	0.18	0.19

Table 8.6: Feature Importance: Comparison of results using a non-fixed reduced feature set and the complete feature set

As mentioned before, during the first test we took note of the most discarded features, in order to obtain a *fixed* reduced feature set to use with all the subjects. We aggregated the reduced feature set over different subjects and decided to remove the features that were discarded the majority of times (i.e. for a number of subjects higher or equal to 6, over 8 subjects). The features mostly rejected were the temporal ones: most of the percentiles were discarded, and one-third of the features involving mean and standard deviations were discarded. No particular axes were discarded in a considerably higher measure than others. Indeed, we decided to use a *fixed* reduced feature set. In Table 8.7 we present the achieved results; even in this case, the accuracies were substantially the same as when using the complete feature set.

To conclude, the usage of a reduced feature set has proven that there is a high number of features

	Fixed Reduced Set	Complete Set
F1 Score	0.86	0.86
TP Other	0.80	0.81
AVG FP Other	0.17	0.19

Table 8.7: Feature Importance: Comparison of results using a fixed reduced feature set and the complete feature set

	Reduced Feature Set	Complete Feature Set
F1 Score	0.952 +- 2.32e-05	0.961 +- 1.70e-05

Table 8.8: Bootstrapping technique and Feature Importance. Comparison between using a reduced feature set and a complete one averaged over 30 runs for each subject

that are correlated to each other and that we could reduce the number of used features, without transforming our data (e.g. by using PCA) to achieve similar results.

8.5.2 Results confirmation about Feature Importance with Bootstrapping

In order to confirm the achieved results regarding Feature Importance, we decided to perform the Bootstrapping technique, to check the stability of the obtained outcomes. Bootstrapping is a technique for sampling data with replacement; the concept of "replacement" means that after sampling one data point, it can be included again (multiple times) in the resampled dataset. It is a common practice to create multiple resampled datasets, perform the desired analysis with data, and then consider the average and the variability of achieved results. Therefore, we have checked the feature importance and the variability of results over different runs.

In particular, some preliminary tests were made to find the right value for the regularization parameter, which was then fixed for the following tests. To begin with, we have completed 10 first runs for each of the 8 subjects to measure what were the features that were discarded the most. With this first test, the results were consistent with the previous ones (performed without bootstrapping). We have aggregated results obtained over different runs and different subjects and we noticed that the list of features that were discarded the most was *identical* to the one achieved when not using Bootstrapping, confirming also that the "less important" features, at least according to this technique, are the temporal ones.

We then wanted to check the stability of results regarding the accuracy achieved. For each subject, we have executed 30 runs for each of the 8 subjects. To compare the accuracy of the different feature sets, a first series of runs was made by using the *complete feature set*, and a second series of runs was made by using the *fixed reduced feature set*. We then averaged results over different results aggregating by each subject.

In Table 8.8 we present the results. A first consideration that we can make is that by using bootstrapping in general, we have achieved higher scores with respect to using the complete feature set. This is a common behavior, since by training our classifiers with more data, the model reduces overfitting. The results achieved with these different two feature sets are almost identical. The remarkably low variance suggests the stability of these numbers over different runs.

With our first tests, we have seen that there is a high number of features that are correlated to each other and that we could reduce the number of used features, without transforming our data to achieve similar results. Thanks to the usage of bootstrapping, we have demonstrated that previous results were not a fluke, but that over different runs the outcomes are stable.

8.6 Usage of Data Augmentation Techniques

In some cases, the possibility of having more training data might become useful to increase accuracy. In our case, it could have been helpful to better discern between activities of interest and non-interesting ones (also known as "other" activities). For this reason, we decided to perform some tests on the application of data augmentation. To increase the amount of data used to train our model, we took the existing pre-processed data and replicated it by adding some noise in order to increase the volume by *ten times*. The choice to consider preprocessed data instead of raw data lies in the fact that when using raw data, we would have probably lost the added noise, creating a sort of replica of the original data. The following Table 8.9 resumes the results (averaged over eight different subjects, with associated standard deviation) related to tests performed before and after data augmentation. As previously done in some of the preceding tests, we will consider as measures: (1) the overall F1 score, (2) the amount of "other" data correctly classified (TP Other), and (3) the average amount of a random activity of interest wrongly classified as "other" (Avg FP Other).

	Before Data Augmentation	**After Data Augmentation**
F1 Score	0.86 +- 0.03	0.87 +- 0.02
TP Other	0.81 +- 0.06	0.75 +- 0.05
Avg FP Other	0.13 +- 0.05	0.09 +- 0.02

Table 8.9: Comparison of results achieved before and after using Augmentation technique

It is easy to see from the table that: when using data augmentation the F1 score has a slight increase and the average FP Other is lower as well, but to the cost of having lower scores in TP Other and Avg FP Other. In general, in the confusion matrix produced using data augmentation, all of the activities of interest have equal or higher values on the diagonal. Moreover, when using data augmentation we have results that are more stable over different subjects, as noticeable from

the standard deviations. In conclusion, it might be useful to make use of data augmentation if we would like to obtain a slight improvement in accuracy and more stable results over different subjects, at the cost of reducing the TP Other metric.

8.7 Attempt to use Long Short-Term Memory Networks

Since most of the recent works regarding AR tasks, dealing with wearable device data, involve techniques based on Deep learning techniques [ZLZ+22, CLPW21], we decided to test similar approaches on our dataset as well. For this test, we rely on Long Short Term Memory networks, known as being able to perform well with our kind of data, and test its behavior with the healthy adults dataset.

We have tested two different structures for this test. Both tests involve *sequential models* and are commonly used for classifying or making predictions on temporal data. In the following, there will be a description of the structure of both considered models. Among the different values tested for the dimensions of the layers, here are the ones that returned the higher accuracy scores.

Structure of the first considered model:
- 2 consecutive *TimeDistributed* layers, since we deal with temporal data, applying a unidimensional convolution with 64 filters, a kernel size of 3, and *ReLU* as activation function
- *Dropout* layer, with a rate of 0.5, to prevent overfitting
- *MaxPooling* unidimensional layer, with a pooling size of 2
- *Flatten* layer, before passing data to LSTM
- *LSTM*, with 128 nodes
- *Dropout* layer, with a rate of 0.5
- *Dense* layer, with 36 nodes and *ReLU* as activation function
- a last *dense* layer to produce the result

The structure of the second considered model involves a convolutional LSTM: - *convolutional LSTM* layer, with 64 filters, a kernel size of (1, 3) and *ReLU* as activation function
- *Dropout* layer, with a rate of 0.5
- *Flatten* layer
- *Dense* layer, with 100 nodes and *ReLU* as activation function
- a last *dense* layer to produce the output

The first described structure could achieve an average accuracy of 0.74. Producing the confusion matrix associated with the predictions, it was possible to see good results for most of the activities, but low values (i.e., lower than 0.50) for the ones involving *walking* (i.e., walking, walking fast, upstairs, downstairs). The second described structure, instead, had more stable accuracy over different activities, and it could achieve a final average accuracy of 0.78. In both cases, *other* activity had on average low classification accuracy (i.e., lower than 0.55).

We tested the usage of neural networks and, in particular, of LSTM ones, to have a glance at achievable accuracy. Even if we performed non-exhaustive experiments on this topic, it is possible to affirm that the results were coherent with the ones achieved with our considered approaches. We preferred to keep on using approaches not based on neural networks to improve the explainability of our methods and comprehension of the obtained behaviors.

Chapter 9

Filtering Raw Data

Introduction

While struggling with making correct and reliable comparisons with other methods or datasets, we aimed our attention at studying our dataset and exploiting the results already achieved. We started focusing on the *frequency* aspect of our data.
Additionally, concerning approaches of activity recognition using wearable devices in the literature, we noticed that some works apply filters (e.g. Butterworth, band-pass, Kalman, ...) on raw data when preprocessing data. Other proposed approaches perform feature extraction directly from raw data. Therefore, there is not a standard procedure for the application of filters.

At the same time, devices usually involved in recording data have a sampling frequency certainly higher than peculiar frequencies associated with ADLs of interest (e.g., walking, going upstairs, brushing teeth). It is reasonable to think, for this reason, that it could be useful to filter raw data to exclude noise and frequencies of the signal not related to interesting activities.

Indeed, we have performed a few tests in analyzing the application of filters on raw data to understand if the accuracy of activity recognition could improve by removing possible noise from recorded data (e.g. by using low-pass filters to different frequencies). Those tests suggested that in our particular case, with our dataset, and using the selected feature set, the accuracy could even decrease.

For these particular results, we then decided to better analyze this topic and to try to understand if, in general, there could be an impact on the accuracy of activity recognition when using filters on raw data. In particular, we wanted to understand if by using filters we could remove details of our data that could be useful for the classifier to discern activities.

We started to investigate this topic by making use of the Kalman filter since it is a filter commonly used and it is a perfect candidate to understand what the impact of filters on our data could be. In

this chapter, therefore, there will be presented the results achieved by applying low-pass filters and using the Kalman filter on our data.

9.1 Low & High Pass Filters

To start analyzing which frequencies characterized our signals, we made use of the Fourier Transform. The Fourier Transform allows representing a function (a signal, in our case) in the *frequency* domain, obtaining the frequency spectrum of our signal. This can help us understand which frequencies characterized our recorded signals.

As a first preliminary test, we plot the output of the Fast Fourier Transform (FFT) applied to our entire raw data, and examine each subject separately. We could see that the significant frequencies in our signals were mostly those lower than 5 Hz; nevertheless, we could still see relevant values for frequencies in the range of 5-20 Hz. Considering that the entire raw data contained all of the activities performed while recording, which as human movements are characterized by frequencies in the range of 0.5-5 Hz, these results could confirm the quality of the signal in our dataset. At the same time, we could see room for improvement in our model accuracy, considering that performed ADLs are usually not characterized by frequencies higher than 5 Hz. This meant that we could try to apply a low-pass filter to remove noise from our signals and see if the accuracy of our model could increase.

During this analysis, we also wanted to further validate the reliability of recorded data, by checking if we could easily find peaks in FFT outputs when considering activities characterized by repeated and cyclic movements. We took the walking activity as an example: knowing that the average duration of one gait cycle for adults is around 1-second [MDK64], we were expecting peaks at particular values around 1 Hz when analyzing raw data of this activity. As expected, we could find peaks around 1 Hz for each subject, confirming the reliability of the recorded data.

Within these tests, we took into consideration also other activities characterized by repeated and cyclic movements (e.g. walking fast, walking downstairs). Results regarding these activities were consistent as well since for example the average period of "walking fast" activity was lower than the average period of "walking" for approximately 0.1 seconds. In fact, by walking fast, the period of the average step is lower, as expected.

The same applied to the comparison between *going upstairs* and *going upstairs fast* activities. Even in this case, over different subjects, the difference in the average period (computed considering the peaks in the frequency spectrum) was approximately 0.4 seconds between the two activities.

Moreover, when validating the quality of our dataset from the point of view of frequencies, we validated the usage of extracted features associated with the frequency domain. In fact, since our AR approach is based on a sliding window technique, we checked if the spectrum obtained using

the total signal of an activity ($\sim$ 90 seconds) was consistent with spectra obtained using singular windows of 4 seconds (i.e., as described in section 8.4). From our results, we would say that we have generally been able to see comparable results for both kinds of data (significant peaks, associated amplitude, and shapes of spectra were similar when compared).

9.1.1 Application of Low-Pass Filters

Encouraged by the results obtained in previous analysis about the frequencies, we have tried to see if we could improve the model accuracy by filtering our raw data with a low-pass filter, by considering each subject separately. This choice relied on the fact that previously we had seen that in the frequency spectra of recordings, the significant frequencies were mostly those lower than 5 Hz, but we could also still see relevant values for frequencies in the range of 5-20 Hz.

In these tests, we first selected 4 Hz as a threshold, according to other works that were filtering the signal with a low-pass filter before preprocessing data [FBB$^+$19]. To choose this threshold we have also considered the frequency that characterized movements of the peculiar activities performed in our dataset: when handwriting, typing at the keyboard, or when brushing our teeth, we could achieve frequencies of 3 Hz for these activities.

Therefore, after filtering our raw data, we preprocessed data again and created our model, following the approach described in Chapter 7, aimed to reduce the misclassification of unknown activities.

In Table 9.1 we present the results obtained with not filtered data and when using a low-pass filter of 4Hz. After applying the filter, we were able to achieve, with our model, an F1 score of 0.77, TP Other of 0.62, and average FP Other of 0.18. This means that all of the used metrics have revealed worse results. The activities for which there has been a higher difference between accuracies obtained in confusion matrices (considering our baseline model and the model where we used filtered data) were *relaxing* and *going upstairs*, with a difference higher than 0.20 in both cases.

	No Filter	**Low Pass 4 Hz**	**Low Pass 16 Hz**
F1 Score	0.86	0.77	0.77
TP Other	0.81	0.62	0.64
Avg FP Other	0.13	0.18	0.18

Table 9.1: Different results achieved without filtering raw data, and filtering raw data with low pass filter of, respectively, 4 Hz and 16 Hz

Considering these results, we have also supposed that with our filter we were unintentionally removing information on our data that was useful for our classifier, indeed. For this reason, we

then tried to raise the threshold of our filter to 16 Hz. Nevertheless, when filtering our data with a low-pass filter of 16 Hz we could obtain results analogous to those achieved with a 4 Hz filter: F1 Score of 0.77, TP Other 0.64, average FP Other of 0.18.

In Table 9.1 we have summarized the different results achieved without filtering raw data, and filtering raw data with low pass filter of, respectively, 4 Hz and 16 Hz. To conclude, the filtering approach did not help our classifiers to achieve higher accuracy.

9.2 Kalman Filter

In order to reduce the complexity of our tests, and to compare results achievable with non-filtered data and with filtered data, we decided to try to classify our data by using the baseline approach, described in Chapter 6. Our tests were made by considering each axis separately, such that our system, step by step, had to work with only one variable (the value of one axis of the accelerometer). We first performed a few preliminary analyses to work with some of the parameters associated with the Kalman filter, by considering the values that, by using only a portion of training data, were giving higher accuracy scores in the classification. Once data had been filtered, features have been extracted. Our tests for this analysis were to compare the results of classification achieved by using not filtered data and by using filtered data. In the following, the comparison between the F1 scores achieved with the raw data and the filtered data averaged over different subjects.

	Not Filtered Data	**Filtered Data**
F1 Score	0.86 ± 0.03	0.83 ± 0.04

Table 9.2: Application of Kalman Filter. Comparison between using filtered data and not filtered data averaged over different subjects. Values expressed as *mean* $\pm$ *standard deviation*

These results show no substantial differences in the accuracies achieved by using filtered and not filtered data. Rather, even in this case, the average accuracy is lower when applying filters to our data. We are aware that the usage of other parameters regarding the Kalman filter, different from the chosen one, could return different results. Nevertheless, from preliminary tests performed, we have seen that we would have risked having even lower accuracy, meaning that, probably, we were removing details of that signal useful for classification purposes.

Chapter 10

Dictionary Learning

Introduction

As aforementioned in Chapter 9, we obtained peculiar results regarding the application of filters on raw data. Therefore, we wished to understand if, in general, there could be an impact on the accuracy of activity recognition when using filters on raw data. In particular, to investigate if data useful for the classifier to discern activities are lost by using filters.
To do so, we decided to test the application of Dictionary Learning techniques, firstly as filters, and then as a classification method. In this Chapter, we are going to describe the results achieved with these techniques based on Dictionary Learning.

10.1 Dictionary Learning as Filters

As previously anticipated, we decided to study the usage of Dictionary Learning as a representation technique to investigate the topic of filtering data. In fact, a possible common usage for Dictionary Learning is denoising, even if mostly for images. From a quite abstract point of view, considering the signal of a particular activity, with this technique we would be able to *learn* what we would consider a *prototype* of the signal of that activity, and of the periodic movements characterizing it. Thus, we could reconstruct the signal by keeping only the most important characteristics associated with a singular activity, excluding noise or less important information. The main reason behind the decision to use Dictionary Learning is that nowadays, with a constant increase in the usage of Neural Networks and Deep Neural Networks for classification tasks, such as activity recognition, it is a common practice to delegate to the network itself the assignment to filter data, by learning to adapt to the raw input. Using Dictionary Learning, was the easiest way to move near to a similar process: a technique that reduces noise according to the type of

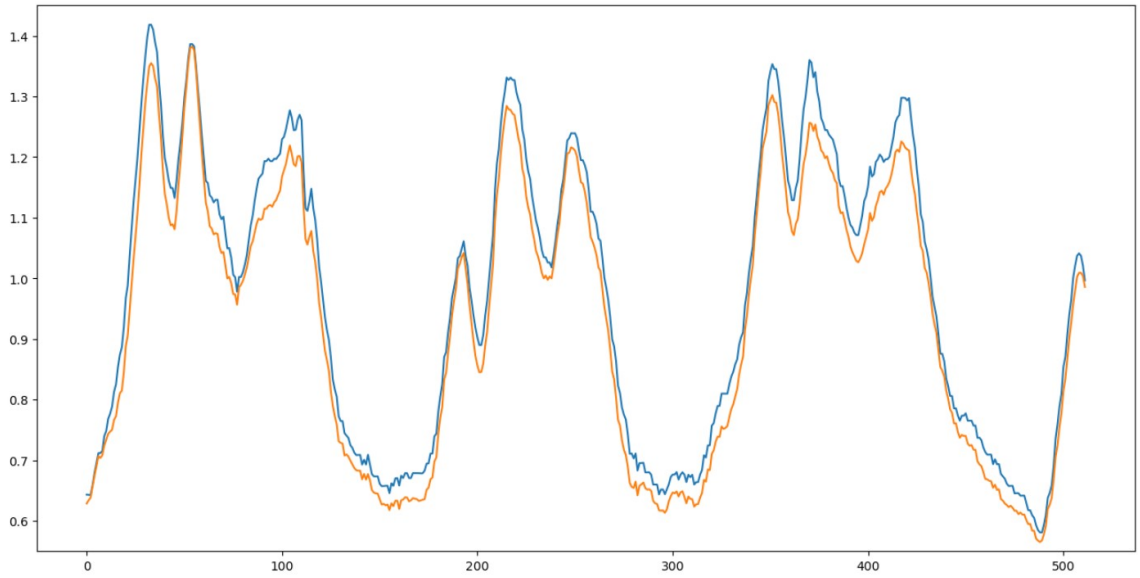

Figure 10.1: Dictionary learning. In blue is the plot of a signal contained in an original window of walking activity, and in orange is the reconstructed one, using one dictionary per activity

signal and keeps relevant information of starting data.

To describe, in brief, how Dictionary Learning works, we would say that the task is to find a set of basis vectors, called a *dictionary*, that can efficiently represent a given set of data samples. Each of these basis vectors is called an *atom*. The general goal is to find a linear combination of a "few" atoms from the dictionary such that it is close to the original set of data samples. More in detail, we would like to learn the dictionary and a sparse representation such that we will have a low reconstruction error.

To begin with, we have mostly performed preliminary tests to introduce ourselves to this technique and its potential. All of the tests have been made considering as samples the windows of our entire signal, from training data, keeping each axis of our accelerometer separated. With a first test, for example, we have tested the possibility of learning a single dictionary for the entire set of considered activities. Still, results show that this is not feasible: the reconstructed signal was far dissimilar from the original one. Instead, when learning one dictionary for each of the activities of interest, it is possible to reconstruct the raw data with good precision. In Figure 10.1, we plot in blue the signal contained in an original window of walking activity, and in orange the reconstructed signal, using one dictionary per activity.

Moreover, a positive thing about this technique is the possibility of checking learned atoms. For example, by considering walking activity, we can see that by plotting a few of the learned atoms, there is a similarity with the original signal itself, proving the ability of this technique to be trained in recognizing a possible prototype of treated activity. In Figure 10.2, the plot of a couple

of atoms that have been chosen at random.

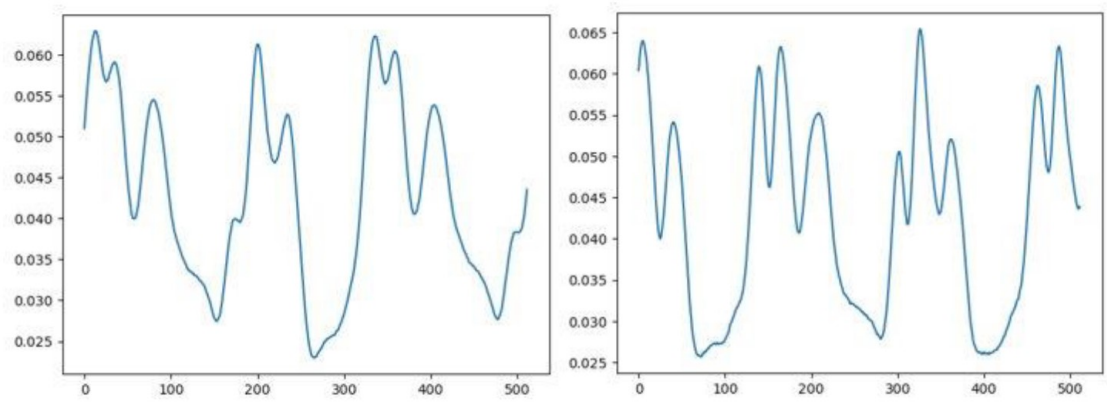

Figure 10.2: Dictionary Learning. Atoms learned from data of *walking* activity.

Within preliminary tests, we have seen that dictionaries had a good ability to reconstruct the signals of activities recorded with accelerometers. We, therefore, tested the possibility of analyzing what was the best-performing way to *train* dictionaries. We firstly focused, for a preliminary analysis, on a singular activity (i.e. walking, because of its peculiar pattern and since it was easy to work with it even from a visual point of view, by plotting signals and learned atoms). We then performed the same tests with more activities, to check the results and to see if previous outcomes were only related to that particular activity (i.e. walking). It might be helpful to say, at this point, that when dealing with Dictionary Learning we can recognize two main phases: *dictionary fitting* (or training) and *data transformation*. The first is the phase in which dictionaries are fitted over data used for training purposes, to be able to learn the atoms and the sparse representation giving the lowest reconstruction error. The latter phase, instead, is the phase in which we might ask the dictionaries to transform our data into a sparse representation. Then, by multiplying the sparse representation with the dictionary atoms, we could obtain the reconstructed data. The following lines of code and associated comments might help to better understand the concepts.

Create the object associated with the learner
`Dict_learner = DictionaryLearning(...)`
Train the dictionary using training data
`Dict_learner.fit(X_train)`
Using the trained dictionary, obtain the sparse representation of our test data
`X_sparse = Dict_learner.transform(X_test)`
Get the reconstructed version of the test data by multiplying (operator = matmul) the sparse representation (X_sparse) and the atoms of our dictionary (Dict_learner.components_)
`X_hat = X_sparse @ Dict_learner.components_`

We decided to compare three different methods and approaches to train our dictionaries, to check the best results achievable:

- **Case 0**: Train the dictionary using a portion (20%) of data of the *i-th* subject; consider then the reconstruction of signals of all *n-1* remaining subjects (as they were concatenated to each other). This was the most difficult situation in which reconstructing the signal using dictionaries.
- **Case 1**: Train the dictionary using a portion (10%) of data of all *n* subjects; consider then the reconstruction of the remaining data of all *n* subjects (for both training and reconstructing phases, data of different subjects have been concatenated to each other).
- **Case 2**: Leave-one-out approach: considering the *i-th* subject, fit the dictionary using a portion of data (10%) of all other subjects, and transform the entire data of the *i-th* subject.

For this task, since we wanted to check what could be the training method returning the lower reconstruction error, we decided to rely on the Mean Squared Error (MSE) measure, comparing our original samples (windowed signal with a window length of 2 seconds) and the sample reconstructed by the dictionary. We first perform preliminary analysis to work with some of the parameters associated with dictionaries and training; these parameters have then been fixed for the following tests. Since data of different axes are way different from each other, we have chosen to have one dictionary per axis, building, therefore, each dictionary considering each axis separately. By having different results for each axis and each subject, we have gathered results by taking mean MSE and averaging over all samples of the single activity. In the following Table 10.1, we can see the results for the *first test*, where we have only examined a single activity, and then in Table 10.2 we list results for the *second test*, where we have studied more activities.

	MSE
Case 0	0.008+- 0.005
Case 1	0.004+- 0.001
Case 2	0.004+- 0.002

Table 10.1: Results of different training methods for Dictionary Learning, with data of a single activity. Results are averaged over different subjects and axes, represented as *mean*$\pm$*standard deviation*

	MSE
Case 0	0.013 +- 0.016
Case 1	0.008 +- 0.010
Case 2	0.009 +- 0.012

Table 10.2: Results of different training methods for Dictionary Learning, using data of 9 different activities. Results are averaged over different subjects and axes, represented as *mean*$\pm$*standard deviation*

From the achieved results, we could say that the higher MSE, as expected, is achieved in Case 0 which was the most difficult situation in which reconstructing data using dictionaries. No substantial differences were found, instead, in MSEs achieved in Case 1 or Case 2 (Leave-one-out) approaches: both methods retrieved low MSEs. By comparing the results between the first and second tests, we can see that the results are constant over different cases. Moreover, the quite low reconstruction errors were not only associated with the walking activity but were quite stable over different activities and subjects.

10.2 Dictionaries as Classifiers

In this Section, there will be described the three different approaches that have been tested to use *dictionaries* for classification purposes. The first approach has been based on the idea of considering the mean squared error as a discriminant. Successively, the sparsity measure has been used as a discriminant, as done in other similar works. To conclude, we describe the approach based on using the inner product of dictionaries and samples as feature vectors, which returned the higher accuracy among the tested methods based on dictionary learning.

10.2.1 Mean Squared Error as discriminant

After having analyzed the potentialities of Dictionary Learning in reconstructing signals of activities and in dealing with our kind of signals in general, we decided to evaluate the possible usage of Dictionary Learning for classification purposes. Our first tests, to familiarize ourselves with this task, consisted of checking the achievable results in the case of binary classification. In our case, therefore, we did not consider the entire set of activities, but just a few couples of them.

The idea was to use two dictionaries, D_a and D_b: one for activity a and the other one for activity b. Each dictionary would have been trained on a percentage of data related to the associated activity. Then, each test sample would have been given to both dictionaries (D_a and D_b) to be reconstructed; the dictionary that would return a reconstructed sample with the lowest MSE should then be associated with the predicted activity, thus letting us obtain the label classifying that sample.

In our preliminary tests where we considered a binary classification, we took two couples of activities. The first task would have been quite straightforward, by deciding to classify *teeth brushing* and *walking*: two activities quite different from each other. The second task, instead, was quite more difficult, since we considered *walking* and *walking fast* activities: the patterns of these activities are quite similar to each other, but the signals have a different period and amplitude.

In Tables 10.3 and 10.4, the results obtained in classifying the activities, averaged over different

subjects.

Activity	F1 Score
Teeth brushing	0.905±0.108
Walking	0.959±0.059

Table 10.3: Results comparison in binary classification using Dictionary Learning and MSE as discriminant

Activity	F1 Score
Walking	0.936±0.106
Walking Fast	0.829±0.260

Table 10.4: Results comparison in binary classification using Dictionary Learning and MSE as discriminant

The results returned from the binary case were quite promising. As expected, the results obtained in the first test (*teeth brushing* vs *walking*) gave higher accuracy, due to the lower difficulty in recognizing the associated patterns. At the same time, lower accuracy was obtained in the second test when recognizing *walking fast*, since the two considered activities were quite similar and the goal task was, therefore, more challenging.

After observing the high accuracy obtained in the binary case, we moved to the multiclass task. In this case, the approach was analogous to the binary one but expanded for the multiclass task. By having one dictionary for each activity, we present each sample to each dictionary. The dictionary that will return the reconstructed sample with lower MSE (with respect to the original sample signal) will decide the label of that sample. The accuracy obtained in this test was definitely too low to consider using this particular approach in the future. Moreover, we decided to limit the tests of this approach to data to four subjects only, due to the low accuracy and the very large amount of time needed. In Table 10.5 we list the F1 scores achieved with data from four subjects.

Subject	F1 Score
# 0	0.255
# 1	0.301
# 2	0.192
# 3	0.350

Table 10.5: Results achieved using Dictionary Learning and MSE as discriminant, in multiclass classification

Considering the confusion matrices obtained from this approach, there were no general low values on the diagonal of the matrix, but instead, still considering the diagonal, there were a few

high values and then other values around 0. No particular behaviours in the confusion matrix have been observed constantly over different subjects, to let us make considerations.

Even if the results of this approach were not satisfying we have investigated the reasons behind them. Indeed, we noticed that even dictionaries of the "wrong" activities were good at reconstructing samples, therefore the lowest MSE was quite rarely the one associated with the right activity. To try to solve this problem, we have been tempted to use fewer atoms when creating our dictionaries. This way, we wanted to force the dictionaries to learn exclusively the important features of the signal related to selected activities, thus reducing this wrong behavior. Even this attempt returned low accuracy: when using definitely low numbers of atoms (i.e., 5, instead of 80) that behavior was still present. Using a smaller number of atoms (i.e., < 5), instead, caused the inability to reconstruct the signals.

10.2.2 Sparsity measure as discriminant

Considering other possible alternative methods to make use of the dictionaries associated with each class of our data, we tested another possible evaluation to be used as a discriminant for classifying our samples. If, in previous tests, we were deciding the predicted label according to the lowest MSE (confronting the original signal and the reconstructed one), we decided to use the sparsity of the reconstructed sample as a measure. We define the sparsity as the number of zero-valued elements divided by the total number of values in the considered sample. In particular, it is expected that the dictionary associated with the right class (i.e. activity) will return the reconstructed sample with higher sparsity. Results obtained with this approach were again quite discouraging. The most frequent behavior was that for the majority of samples, a single dictionary was the one returning reconstructed samples with the higher sparsity. This way, we were obtaining confusion matrices in which most of our samples were (wrongly) predicted as a unique activity.

10.2.3 Inner product of dictionaries and sample as feature vector

In spite of the poor results obtained in previous tests regarding dictionary learning, we did not stop to investigate the usage of this technique for our purposes. We reasoned about an additional possible way to leverage the information obtainable from dictionaries. We have thought, in fact, about using the atoms of each dictionary.

We started with the concept of having 17 dictionaries (one for each activity). Each dictionary is composed of a certain number of atoms (at the beginning 80 atoms were used, for this test we reduced this number to 30); each atom has dimensionality equal to the length of our sample (i.e., 512, in our case, since we are using windows of length 2.0 seconds and a sampling frequency of 256). We can say, for clarity, that the set of atoms of each dictionary created in our scenario has

dimensions equal to [30 x 512]. We decided to concatenate the set of atoms of *all the dictionaries* (i.e., all the activities), to create an "aggregated dictionary". We can consider this resulting matrix as a list of weights that relate both the data of our samples and the features of each particular activity. To find a way to relate each sample to the aggregated dictionary, we have thought about the inner product, which is, indeed, a possible way to associate a vector and a matrix. By making the inner product of the aggregated dictionary and each sample, we could obtain a new vector that considers the sample and the values of atoms, which can be considered as *weights* obtained from dictionaries. This new representation of data could be used as a *feature vector*. Thus, for the sake of simplicity, we relied on Support Vector Machine (SVM) to understand the capabilities of this approach.

The results of this approach were encouraging. As already done in past tests, we needed to consider the data of each axis separately. We then decided to aggregate results by considering a sort of *voting system*. Since for each class we had three predictions (one for each axis), we took the label that was most present; otherwise, we took the one of the y-axis, since, according to previous tests, it contained the most helpful data for classification purposes. Moreover, we have forced the positivity of atoms and of the sparse representation (meaning that both of them, when computed, had to have positive values). This is commonly done to reduce the time needed to learn dictionaries and to enhance the learning of the most important patterns present in data. This last aspect has made a little contribution to obtaining higher accuracy. In Figure 10.3, the obtained confusion matrix, obtained averaging results over different subjects.

The average F1 score obtained is 0.84 ± 0.05 (mean +- standard deviation). We can clearly see that we have obtained low accuracy in recognizing the Other activity (labeled as 'X' in the confusion matrix), but the overall results are quite satisfying, considering the results obtained in previous tests using Dictionary Learning.

To conclude, this approach has returned results that are comparable to other approaches, if considering only the accuracy in classifying activities of interest. The classification of data labeled as Other, with this method, is quite hard due to the inability to find a common pattern in data for this particular data.

10.3 Discussion

Considering the results discussed in Chapters 9 and 10, we can discuss the behavior of the application of filters on our data. By directly applying the low-pass filters (i.e., setting 4Hz and 16Hz as thresholds) to our raw data, we reduced the average F1 score of the classification from 0.86 to 0.77; considering additional metrics associated with "other" data (TP other, and average FP Other), both metrics got worse. The use of the Kalman filter did not show substantial differences using filtered raw data or not. In that test, in fact, the average accuracy achieved with *not filtered data* was an F1 score of 0.86, while using *filtered data* the model achieved an average F1 score

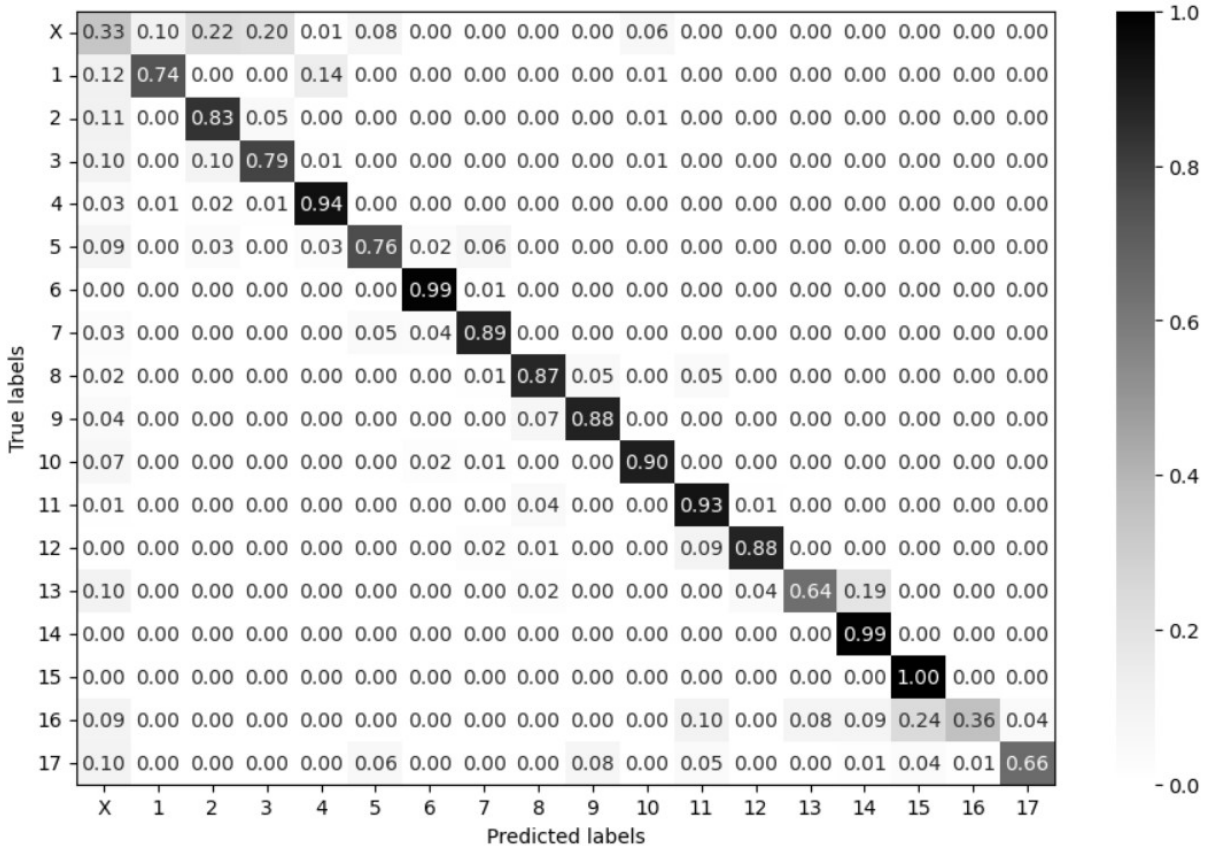

Figure 10.3: Confusion Matrix obtained using dictionary learning for classification, and the inner product of dictionaries and sample as feature vector

of 0.83.

For what concerns the techniques based on dictionary learning, we have first tried to understand which was the best approach to train dictionaries. Then, different methods have been tested for finding the procedure to follow to leverage dictionaries in classifying samples ignoring the possible noise of our signals. The procedure that returned the higher score had results comparable to other approaches if we focus on the recognition of *activities of interest* only. However, performed poorly in recognizing the data labeled as ”other”.

We could conclude, from our results, that the application of filters on raw data, before preprocessing, affect the final accuracy in a negative way for classification tasks. Tests on our data lead to the conclusion that applying filters can cut out parts of the signal useful for the classifiers. However, these tests and therefore the conclusions have some limitations. First, we tested only a few types of filters; other kinds of filters could return different results. Moreover, the number and the genres of considered extracted features could be more or less influenced by filters; this issue

could need additional tests to investigate on associated details. Additionally, we studied a single set of activities; future analysis could help in understanding the impact of filters on different activities.

Part IV

Reproducibility in Activity Recognition

Chapter 11

Reproducibility in Activity Recognition Based on Wearable Devices

Introduction

In this Chapter, there will be presented the topic of reproducibility in activity recognition based on wearable devices, with a focus on used datasets. The main topic is about a general lack of reproducibility in proposed approaches in the literature. We conducted a literature review to determine from a subjective point of view the significance of this issue. Starting from this review, our contributions regarding this part are mainly two: (1) finding the measure of works in the literature using public datasets or sharing the used one, and (2) studying the characteristics of used datasets.

Hereby, there will be given an overview of the presented issue, describing the reasons that led to this survey. Successively, there will be illustrated the empirical study behind the related research, presenting the considered procedure. The results of the research will be then discussed. Before the final discussion, there will be summarized the findings and remarks and then a list of related works.

11.1 Overview

As aforementioned in this thesis, automated recognition of daily living activities is an important task from many perspectives including medical research. An increasing interest emerged recently in activity recognition (AR) for monitoring people possibly using data obtained from wearable devices. As also demonstrated in our work, these devices in fact can be a help for monitoring

performed ADLs, to track the health status of patients, or elderly people, guaranteeing advantages for both patients and specialists, because of their portability and also due to their cheapness.

At the same time, many researchers have proposed approaches for the activity recognition task that were ML-based [TDF+18, JNJSS18]. It is also well known that *data* can have a large influence on ML approach results since data are used for both training models and testing their effectiveness and accuracy.

In Chapters 4 and 5 we have described the two datasets that we recorded and publicly released; in Chapters 6 and 7 we have presented approaches we used to perform activity recognition on our dataset with success, starting from wearable devices. With these works, during the analysis of the state of the art, we discovered a lack of well-recognized AR benchmark datasets to compare with, in contrast with other fields (e.g. Computer Vision or Text Mining) where there exist popular datasets used as benchmarks for proposed works. Additionally, we noticed that many of the existing approaches in AR are hardly reproducible (and therefore it is difficult to compare with novel proposals). This happens since researchers often do not make available approach implementation nor the datasets used for its validation. Few works are sharing used datasets when created on purpose for the associated article, and not many works are using already public datasets.

Moreover, even if usual AR works are commonly dedicated to elderly people's health and well-being, populations involved in recording datasets are often characterized by young healthy subjects. This is justifiable for different reasons: (1) researchers often do not want to bother elderly people by recording hours of data where undesirable events could happen, and (2) there are general difficulties in recruiting volunteers.

We think that replicability and reproducibility of proposed approaches are fundamental aspects of scientific research, not only for intensifying the confidence in published results and accuracy but in general in the entire field of *machine learning*. It is known that sometimes the enthusiasm for the achieved results can be transformed into doubts about the reproducibility of proposed works [Hut18, Gun20]. It is also a matter of fact that often ML models can overfit on seen data and therefore can generalize with difficulties on new data. Thus, to gain the trust of experts who are skeptical of results obtained with ML models, it is crucial to give all the possible details to reproduce proposed approaches.

For these reasons in this work, we want to explore the reproducibility issue and verify in what measure this is occurring. We started by analyzing the characteristics of datasets used in proposed approaches regarding AR involving data coming from inertial sensors placed in wearable devices.

To this end, we carried out a literature review by analyzing scientific papers proposing AR approaches using data from inertial sensors, to answer our questions regarding employed datasets. Therefore, regarding this part of the thesis, as a *first contribution*, we measure what percentage of works in the literature validate their approach using public datasets or sharing the ones created on purpose. As a *second contribution*, we have studied the characteristics of datasets used in

considered works, with a focus on the amount of data recorded, involved population, and studied activities.

The outcoming general objectives of this literature review and its results are to bring the attention of other researchers to these issues and to persuade future articles' authors to support the replicability and reproducibility of their works. Regarding these two latter terms (i.e. replicability and reproducibility) we are aware that there is a debate about their usage [Ple18, Dru09]. We would like to point out that our focus in this work will be on the *reproducibility* of proposed approaches. We will consider the approved definitions by the Association of Computer Machinery [Ass20]. According to these definitions, the recommendations are to use *replicability* when an independent group can obtain the same results using artifacts that they develop completely independently and to use *reproducibility* when an independent group can obtain the same result using the author's artifacts.

As aforementioned, we can say that ML is generally data-driven since created models are highly susceptible to used data. This fact indicates the importance of data related to a proposed work. When authors desire to propose a novel ML approach or technique, in any field, they could compare the effectiveness of their proposal with the existing ones only by recreating all the considered approaches and then testing these approaches on one or more datasets. Since setting up one or more existing approaches from scratch to completely reproduce an approach is not always an easy task, a quite common way in the scientific community to bypass this problem, for what concerns datasets, involves the usage of *benchmarks* datasets. They are reference datasets containing high-quality data. In this way, new proposals can report the achieved results on such benchmarks to obtain a comparison with the state of the art in a simple way.

This procedure is adopted in many contexts such as Computer Vision, Sentiment Analysis, or Intrusion Detection. For example, regarding Computer Vision, *ImageNet* and Google's *Open Images* are two remarkable datasets commonly adopted for different tasks related to this field. The first was released in 2009 and contains $\sim$14 million images, the latter has been released in 2016 and includes $\sim$9 million images to date [NHCB19]. For the Sentiment Analysis field, we would like to cite the *IMDB Large Movie Review* dataset consisting of sentences from movie reviews, released in 2004, and the *Blog* dataset, released by P. Melville in 2009 consisting of product reviews and political posts [YSZ17]. Regarding the field of intrusion detection, three relevant benchmark datasets are the KDD (released by DARPA in 1999), the UNM dataset (released in 2004), and the ADFA-LD (released by Creech and Hu in 2013) [ACMI15].

11.2 Empirical Study

Intending to investigate the problem of reproducibility in AR, we have defined the following research questions (RQ), regarding the availability and the characteristics of datasets associated with works related to AR based on inertial sensors data.

RQ1. *How many AR works offer public access to the used datasets to better understand and potentially reproduce their results?*

RQ2. *What are the characteristics of the datasets used in the considered works in relation to the amount of data available, involved population, and studied activities?*

The first research question deals with the fact that in case ML/AR specialists are interested in better understanding the results of a particular work, and reproducing it, authors should release to the public the used dataset. The same applies to the relevant cases in which specialists would like to compare the novel approach they are proposing to others present in the state-of-the-art literature. Such comparisons are an important requirement to empirically evaluate and assess the effectiveness of a novel approach (e.g., by analyzing the accuracy that can be obtained with the novel approach on the same datasets). For these reasons, the dataset used in a published work should be always made publicly available to let other authors reproduce the work or improve the performances associated with some tasks (e.g. classification of activities). For those datasets that have been released, we will also check their online availability. It is possible that after some years from the paper's publication, the reference could be broken, or the hosting servers could be down. Regarding the public access to the dataset, we will also analyze how many works were published as *Open Access*.

The second research question is about the characteristics of datasets used in selected works, focusing on the details regarding the amount of data recorded, the population involved in recording data, and the considered activities. In particular, looking at these characteristics allows us to understand:

1. which is the amount of data recorded in terms of (a) the number of subjects involved and (b) the quantity for each subject and each activity (note that several AR approaches need a minimum quantity required to correctly extract features, e.g. $>= 1$ minute per activity);

2. what are the characteristics of the specific population involved in the data recording: is there any portion of the "global population" that is less represented and thus studied? We will focus in particular on the following characteristics of the subjects involved: male/female proportions, age, height, weight (or BMI if height and weight are not available);

3. what is the number of activities that are usually recorded and thus that will be then recognized? Moreover, how many times does the number of activities correspond to the number of classes to be recognized by the classifier? Sometimes, the number of classes to be learned by the classifier is lower than the number of activities recorded, to simplify the work for the classifier.

11.2.1 Procedure

To try to answer our research questions, we have decided to rely on the Scopus digital library[1] as performed in other works also published in renowned venues (e.g. IEEE, ACM, Springer, Elsevier) [DLHFACQE18, dSSFL19, GCC+21].

In particular, we have looked for works in the context of AR using data recorded by wearable devices equipped with inertial sensors (i.e. IMUs [AGKK13], or simply accelerometers). Thus we defined an inclusive Scopus query able to select a large number of papers relevant to our study. From the output provided by Scopus, we then selected works according to defined *inclusion* and *exclusion criteria*. Such works have then been analyzed to extract data able to answer our research questions. In the following, we will further describe each of the performed steps.

11.2.1.1 Scopus Search

As anticipated, we were interested in finding works related to AR methods using data coming from wearable devices that are equipped with inertial sensors. Starting from this, we defined four key elements of our research (i.e. classifier, activities, wearable, inertial sensors) and then created query terms using keywords related to each of the key elements. The ”classifier” key element was needed to specify that we wanted to focus on classifying and recognizing something (i.e. activities) over a set of data. The ”activities” key term was used to specify the topic of classifiers; the key term ”wearable” was needed to indicate that data should have come from devices placed on subjects' bodies (no matter in which part of our body). The last key term ”inertial sensors” was used to point out that data should have been produced from sensors able to measure the forces at which they were undergoing.

We, therefore, defined the following search string:

```
("Daily Life Activities" OR "ADL" OR "daily living" OR "activity
recognition" OR "human activity recognition" OR "HAR" OR "Activity
classification") AND
("wearable") AND
("accelerometer" OR "IMU" OR "inertial sensor" OR "IMU sensor" OR
"inertial unit") AND
("machine learning" OR "ML" OR "classifier" OR "classification" OR
"deep learning" OR "neural network" OR "Hidden Markov Models" OR
"Feature extraction" OR "algorithm" OR "pattern recognition")
```

[1] https://www.scopus.com/

11.2.1.2 Document Selection

The query, submitted to Scopus on 23rd March 2022, returned a total of 1289 works, too many to be manually reviewed. For this reason, we have decided to reduce the number of results considering the number of citations per year (i.e., total citations of the paper, divided by the age of the paper in years; this can be considered as a proxy for measuring the relevance of the work). Indeed, if we had considered only the absolute number of citations, the most recent works could be excluded given the low number of citations. On the contrary, computing the citations per year allowed to include also recent works that have a reasonable number of citations (w.r.t. their age). In this step, in particular, we have selected only works that achieved a number of citations per year that was higher or equal to 5. Using this value as a threshold, we obtained a reasonable number to review; with this choice, we have gathered a total of 207 works.

We then defined the *inclusion* and *exclusion criteria* to select the works that had to be fully examined to answer our research questions. We have decided that we would have not considered (1) works that were surveys or reviews (because authors were not proposing methods requiring datasets for validation), (2) works that were not dealing with the general classification of activities (e.g. papers focusing only on specific tasks such as fall detection, gait analysis, freezing of gait, gestures recognition, tremors, were excluded), (3) works that were not dealing with inertial sensors placed on the body (i.e. inertial sensors, with other optional kinds of sensors, should have been always used to recognize activities).

By reading the titles and abstracts of the 207 selected works, we have chosen 146 articles according to the predefined criteria.

11.2.1.3 Data Extraction

In the data extraction step, we read and analyzed in detail the candidate documents, filling out a detailed form with the information gathered from each source [KC07]. The form consisted of 15 questions, elaborated to answer our research questions. Concerning RQ1, regarding the public access to datasets and works, we have defined the following questions:

1. Was a link (or access) to the used dataset provided at the time of paper publication?
2. Is the used dataset still available?
3. Does the associated paper respect open-access principles?

Concerning RQ2, regarding the amount of data available, involved population and studied activities, we have defined the following questions:

4. What is the average amount of recorded data per subject (measured in time)?

5. What is the average amount of recorded data per activity, per subject (measured in time)?
6. How many activities have been studied?
7. How many classes had to be recognized by the classifier?
8. How many subjects have been involved in data recording?
9. What percentage of male participants has been involved?
10. What percentage of female participants has been involved?
11. What is the average age of subjects?
12. What is the average weight of subjects?
13. What is the average height of subjects?
14. If weight and height were not available, what is the average BMI of subjects?

Each of these questions investigates valuable information regarding data used to validate an approach. We think, for example, that the average amount of recorded data per subject (i.e. questions 4 and 5) is useful to understand how much data are needed to train (and also test) a particular model to achieve a particular accuracy.

At the same time, we would need to know the number of activities (and/or classes) involved in the approach, and the number of subjects participating (i.e. question 6,7,8), to know basic information regarding the classification and the ability of the approach to generalize among different subjects.

To conclude, the latter questions regard several other characteristics of the subjects involved, since there could be differences in the way activities are performed within subjects with dissimilar ages or BMIs.

11.3 Results

After filling the extraction form with the information retrieved from each work, we analyzed the obtained results and here provide some considerations on the collected data. In this Section, we are first going to discuss a few initial considerations needed to understand the obtained results, and then we will present the outcomes of our analysis regarding RQ1 and RQ2.

11.3.1 Initial considerations

In general, researchers proposing a novel AR method can evaluate its effectiveness by using:

- one (or more) already public datasets
- one (or more) datasets specially made for that article
- a combination of already public and specially-made datasets

Out of the 146 considered articles, we have found out that:

- 38 were using already publicly available datasets
- 98 were using datasets specially made for associated works
- 10 were using a combination of already publicly available datasets and specially-made datasets.

From these first numbers, we can understand that the majority of works (more than 67%) evaluate the proposed AR approach using one or more datasets that have been recorded for that work. At the same time, only about a third of the works use already publicly available datasets to evaluate the results achievable with the proposed approach.

We have found a total of 110 datasets starting from those 108 (98+10) articles relying on datasets that were specially made for associated works (alone or in combination with already publicly available ones). In the same way, we have found 31 datasets starting from the 48 (38+10) articles that were using already publicly available datasets (alone or in combination with specially-made ones). In conclusion, by examining the 146 considered articles, we collected information about a total of 141 datasets.

11.3.2 RQ1

With respect to RQ1 (i.e. how many AR works offer public access to the used datasets in order to better understand and potentially reproduce their results?), we will consider the related questions that we have prepared to be filled in our form (as described in Section 11.2.1.3).

Among the total of 141 datasets mentioned in the works, we have found out that:

- 79 out of 141 datasets were not shared;

- 62 out of 141 datasets were shared, and currently:
  - 43 out of 62 datasets are actually accessible and downloadable from the web: for these datasets, we have verified the current data files' accessibility with success incurring no restrictions; two of these datasets were associated with links or servers that are nowadays not accessible, but updated links could be retrieved with a query on common web search engines;
  - 10 out of 62 datasets had restricted access since we could potentially access data after (a) contacting the authors or (b) registering to specific websites (e.g. IEEE DataPort[2], that require a subscription for a fee);
  - 9 out of 62 datasets appear to be no longer available online since associated links or servers are nowadays not accessible; even research using common web search engines could not help us to retrieve these datasets.

Table 11.1 provides a recap of datasets' availability and accessibility according to their provenance (i.e. already public or specially-made datasets).

For what concerns compliance with the Open Access principles of the paper where the considered datasets have been employed, we have gathered interesting stats as well. For this task, since we used Scopus as a digital library, we relied on the Scopus filter for Open Access. With the help of this filter, we could see if each of the works associated with the datasets reviewed was compliant with Open Access principles or not. Adhering to the Open Access principles for a paper proposing a dataset could be even more valuable since such a paper could contain relevant information about the dataset, thus being useful for the dataset's future users. For this reason, in particular, in this single analysis, we focused only on those papers associated with datasets that were specially made. According to the filter results, only 50 articles out of 108 ($\sim$46%) were compliant with Open Access principles.

	Free Access	**Restricted Access**	**Offline**	**Not Shared**
P	27/31 (87%)	2/31 (6%)	2/31 (6%)	0/31 (6%)
S	16/110 (15%)	8/110 (7%)	7/110 (6%)	79/110 (72%)
T	43/141 (31%)	10/141 (7%)	9/141 (6%)	79/141 (56%)

Table 11.1: Recap of datasets' availability and accessibility according to their provenance (i.e. already public or specially-made.
P represents Already Public datasets, **S** represents Specially-made datasets, **T** represents Total datasets (P+S)

[2]https://ieee-dataport.org

11.3.3 RQ2

Regarding RQ2 (i.e., what are the characteristics of datasets used in considered works with relation to the amount of data available, involved population, and studied activities?) we will again consider the related questions that we have provided to be filled in our form (as stated in Section 11.2.1.3).

A first interesting aspect that we would like to point out is related to gathering information about considered populations used to record data: quite often this information was not given. In fact, for 31 datasets out of 141 ($\sim$22%), we could find none of the basic data such as: *male/female distribution, ages, weights,* and *heights* or *BMIs* of people involved in data recording, by reading the associated paper or the description page on the related website to download data. In Table 11.2 there is an aggregated summary of the availability of this information among the studied datasets. In particular, the data that we focused on were: the number of subjects involved, male/female distribution, age, weight, and height or BMI distributions, number of activities recorded, amount of data per subject, and amount of data per activity for a single subject.

Some remarkable facts that we could extract from this table are:

- Almost all of the works (99%) specified the number of subjects involved and the number of studied activities since these are fundamental data related to studying AR approaches validity;
- Ages distribution was present for only 72% of the studied works: when present, this information was given as ranges or using the mean and standard deviation of ages;
- Body information such as weight and height or BMI were present in approximately only 33% of the works; when BMI was not present but we could retrieve it by using weight and height, we computed it by hand. These body data, when present, were given as ranges or using means and standard deviations as well;
- For one work out of three ($\sim$33%) we have not been able to retrieve statistics about the amount of data recorded per subject and/or per activity for a single subject.

We have also extracted other facts regarding the characteristics of available data, involved population, and studied activities regarding analyzed datasets. We have computed the median values of these features of datasets: the number of subjects, percentage of males and females over total subjects, age, weight, height, BMI, number of studied activities, amount of data per subject, and amount of data per activity for each singular subject.

For some of these datasets, it occurred that the entire set of subjects involved was described as two (or more) different populations (e.g. one to perform a set of activities, and the other to perform a second set of activities). In these cases, the characteristics of the subjects were

#Subjects	Male/Female	Age
140/141 (99%)	93/141 (66%)	102/141 (72%)
Weight	**Height**	**BMI**
48/141 (34%)	45/141 (32%)	47/141 (33%)
#Activities	**#Data per Subject**	**#Data per Activity**
140/141 (99%)	95/141 (67%)	95/141 (67%)

Table 11.2: Summary of availability of information about populations involved in recording studied datasets

# Subjects	% Males	% Females
10	61	39
Weight	**Height**	**BMI**
70 Kg	171 cm	24 kg/m^2
Age	**# Activities**	**# Classes**
29	8	8
# Data per Subject	**# Data per Activity**	
30 min	3 min	

Table 11.3: Median values of the considered features of datasets and populations

given for each population. For this reason, the extracted statistics will be related to considered populations and not in general for each dataset.

Moreover, since age, weight, height, and BMI have been represented in different ways in considered works (e.g. ranges or mean $\pm$ standard deviation), we have extracted a singular representative number for each population considered. When these data were presented as ranges we considered the mean value of the extremities of the range; when data were offered as mean $\pm$ standard deviation we took the mean; if both representations were used we preferred the second option for being more descriptive.

In Table 11.3 it is possible to see the median values for the considered features of datasets and populations. In Figure 11.1 there is a scatter plot associating representative heights and weights of considered populations. In Figures 11.2, 11.3, 11.4, 11.5, there are histograms representing the distributions of ages, BMIs, weights, and heights respectively.

We will now analyze in detail each of the features that we have considered in our analysis.

Subjects. The median number of subjects involved in datasets is 10. With such a number of subjects, it is reasonable to start to validate the ability of the approach to generalize among different subjects. On the contrary, an additional note can be made about the fact that 7 works

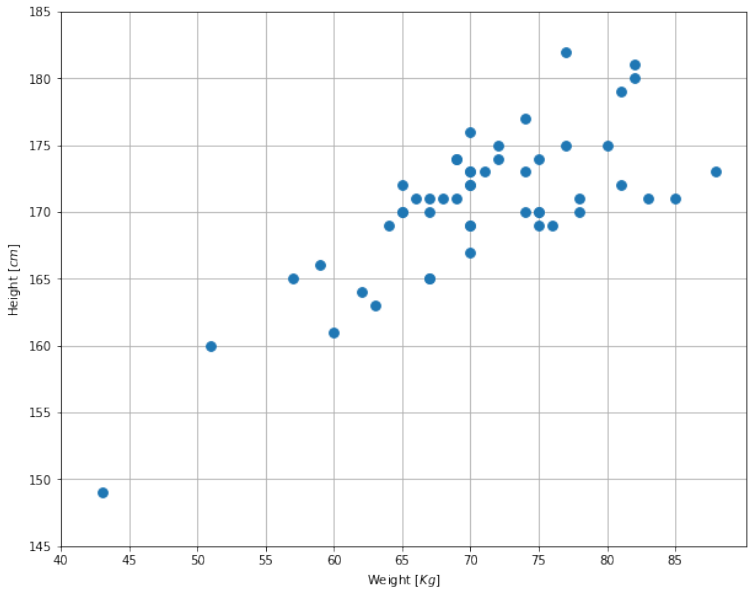

Figure 11.1: Distribution of weight and height among considered populations

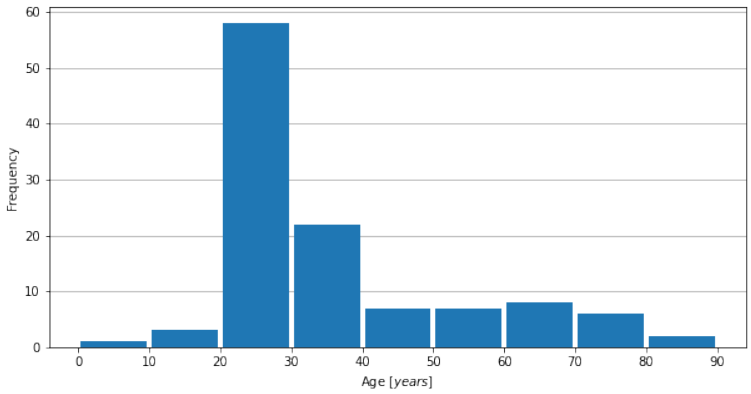

Figure 11.2: Distribution of ages among considered populations

out of 141 have used only 1 subject to validate their approach. This is, fortunately, a small number of works, but using data coming from only one subject would not prove the capacity of the approach to be efficient over different subjects.

Age. According to our analysis, we have a median value for the ages equal to 29 years. This means that populations involved in the datasets are usually young, even if the most common proposed applications for AR tasks concern elderly people (e.g., by detecting their particular activity patterns and acting accordingly). Considering the resulting histograms in Figure 11.2, we were expecting two possible optimal cases. In the first case, if all populations were equally

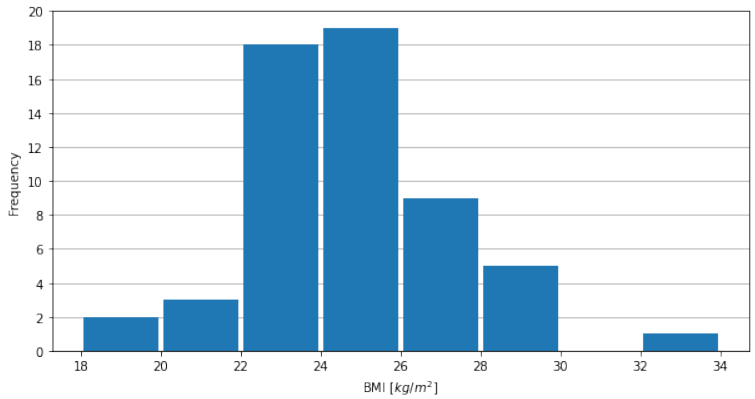

Figure 11.3: Distribution of BMIs among considered populations

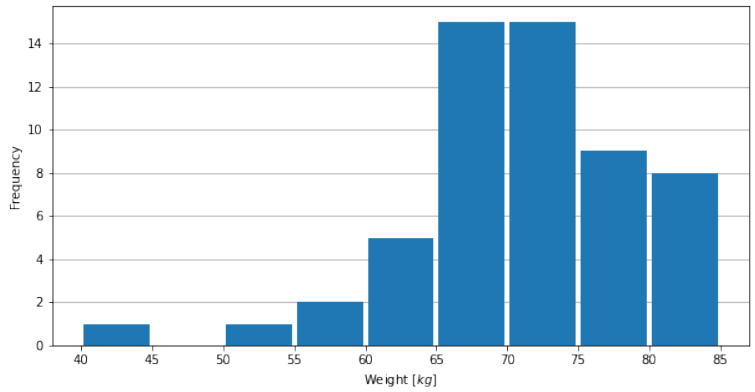

Figure 11.4: Distribution of weights among considered populations

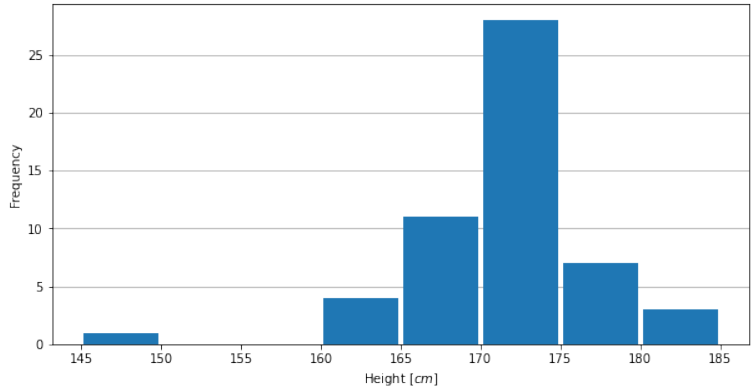

Figure 11.5: Distribution of heights among considered populations

distributed among adults of every age, we were expecting a peak around the middle-aged (40-50) people. In the second case, if most populations were involving elderly people, we were expecting to see peaks around higher ages. None of the optimal cases has taken place.

Male/Female Distribution The median percentages of male and female subjects over the total population are 61% and 39%, respectively. There is a bias in these values; nevertheless, we have seen populations completely unbalanced for both sides (only female or only male subjects). We appreciate the recent attention to this issue in order to have an equal distribution also of these aspects.

Weight, Height, BMI. For what concerns these characteristics, we have seen that except for a few datasets dedicated to children (where weight and height could appear as outliers in our statistics), there is a median value of around 70 kg, 171 cm, and 24 kg/m^2 for weight, height, and BMI respectively. These values represent people in the *normal* range of BMI, in agreement with the WHO (World Health Organization) classification of adults according to BMI [Org00]. Regarding these particular characteristics, as anticipated, it is useful to notice that we have also included datasets involving underage subjects, which can alter these results. To scale down this possible issue, we should notice as well that only four datasets concerned young persons, and that only two of them shared information regarding weight, height, and BMI.

Activities. The median value of considered activities is 8, as for the number of classes to be recognized by the proposed approaches. We should notice that this metric does not take into consideration the similarity between considered activities (i.e. similar activities would be more difficult to be classified). Furthermore, these results are coherent with the average number of ADLs found in analyzed datasets in a related work [CSRCG17].

Amount of data. According to these statistics, the median amount of data recorded for each subject corresponds to 30 minutes, where for each activity a subject has dedicated approximately 3 minutes. Considering the usual sampling frequency of commonly used inertial sensors used for these tasks (i.e., higher than 20 Hz), these durations would let specialists retrieve a good amount of data to extract features used to discern activities.

11.3.4 Threats to Validity

We are aware that our analysis could involve some possible issues or limitations. Firstly, for our research, we took into consideration only the output provided by the Scopus digital library, instead of using multiple sources and combining results. However, we should note that Scopus lists the papers published by all the major editors such as IEEE, ACM, Springer, Elsevier, and many others, so it covers almost all the peer-reviewed research published in relevant venues and journals. Moreover, even if we have examined only one source, we still think that the number of works that have been analyzed forms a good representation of the most relevant proposed approaches.

Secondly, we are sure that many of the authors of works that in our opinion were not sharing the datasets could accept to give access to use data if contacted. Similarly, for what concerns outdated links, we could reach authors in order to request data. The same applies to what concerns population and dataset information: we could possibly request these materials and we are sure that some authors would have accepted to collaborate. We still think that it would be useful, for reproducibility purposes, to clearly specify the possibility of requesting data if not openly shared online.

To conclude, regarding our extracted statistics, we are aware that when considering the distribution of ages, weights, heights, and BMIs, by considering only a single value (e.g. usually the mean values or the mean of range extremes) for each population we have removed useful information. Because of the different representations used by each work to describe these characteristics, it was quite difficult to find a method to aggregate all the data.

11.4 Finding and Remarks

In our empirical study, we have tried to answer the stated research questions seen in Section 11.2. Hereby, a recap of the consequent results.

Regarding RQ1 (i.e. how many AR works offer public access to the used datasets in order to better understand and potentially reproduce their results?) we have found out that:

- 44% of all the examined datasets were accessible at the moment of publication of the related article (i.e. this value is computed by considering that 56% of the considered datasets were not shared at all);
- more than 67% of works evaluated the associated approach using specially made datasets;
- among the total datasets (i.e. public ones or those that have been shared by authors) used to validate approaches that we have evaluated, about 56% of them have not been shared by authors;
- considering total datasets, 13% of these had restricted access (7%) or were offline (6%).

For what concerns RQ2 (i.e. What are the characteristics of the datasets used in the considered works with relation to the amount of data available, involved population, and studied activities?) we have discovered that:

- in 22% of works, we could find none of the basic data such as male/female distribution, ages, weights, and heights or BMIs;

- usually proposed works are evaluated trying to prove the capacity of models to be efficient over different subjects since the median number of subjects involved is 10;
- the median age of subjects engaged is around 29 years, meaning that the considered populations are usually young, even if typical applications of AR systems concern elderly people;
- the median amount of data recorded used to build the model is around 30 minutes for each subject and 3 minutes for each activity; these durations allow researchers to retrieve a sufficient amount of data to extract features.

In order to make this review fully transparent, we have published a CSV file containing the analyzed works and associated details online[3].

11.4.1 Implications

Starting from the aforementioned findings, here we suggest possible pieces of advice for future papers' authors who desire to record and publish new datasets or who wish to propose novel approaches with specially made datasets. These suggestions have the final goal of increasing the reproducibility of works in the literature (fundamental to adopting existing approaches and setting up in-depth comparisons).

1. Share used datasets and all the additional information that can be useful for readers to compare works and approaches;
2. Release articles with publishers adhering to Open Access principles, especially if presenting new datasets so that connected material can be openly accessible to any possible user;
3. Take care of the population involved in data recording in order to combine subjects with heterogeneous characteristics;
4. At the same time, in case there is a focus on recognizing activities or behavioral patterns that are typical of particular populations, take into consideration suitable subjects (e.g. studying movement behaviors in elderly people).

11.5 Related Works

The discussion regarding the lack of elements for reproducing and replicating proposed approaches is already known and considered in the literature. To the best of our knowledge, this

[3]https://sepl.dibris.unige.it/2022-Reproducibility-AR-Datasets.php

is the first work in the state of the art that focuses on reproducibility issues concerning AR approaches using data recorded by wearable devices equipped with inertial sensors.

To support the analysis of the discussed issue, there is the result of an internal study proposed by Samuel et al. [SLKR20]. It has been asked domain experts what are the challenges in reproducing published results of ML experiments. Among the cited challenges, is the unavailability of datasets used for training and evaluation.

McDermott et al. in their work [MWM+21] have evaluated the reproducibility of machine learning for health research, by reviewing 511 works across different subfields. Authors have found that when Machine Learning approaches are used for health purposes, there is a reduction in reproducibility metrics (e.g. datasets and code accessibility) with respect to other fields. For this reason, in their work, they also propose recommendations to focus on this problem. Their evaluations are quite consistent with our results since McDermott et al. declare that only $\sim$55% of papers used public datasets, compared to more than 90% of both computer vision and natural language processing papers. We would like to point out that in our analysis we have seen that $\sim$44% of the considered papers were using public datasets.

Similarly, also Wojtusiak, in a work regarding reproducibility and transparency of ML in health applications [Woj21] has proposed ten criteria to be used when presenting results. The goal of these works where authors propose recommendations and advice for illustrating results is not only to highlight the general issue but to propose possible solutions in order to guarantee reproducibility and also to achieve the trust of other specialists regarding achieved outcomes.

To conclude, regarding the same type of data (e.g. inertial sensors), we would like to cite also an analysis of public datasets for wearable detection systems [CSRCG17] by Casilari et al. Authors of this analysis have focused only on public datasets (whilst we started from approaches to reach used datasets), and concentrate on fall detection (whilst we selected articles that face the more general AR problem). At the same time, it is useful to know that many public datasets about fall detection contain also recordings of activities of daily living (usually for learning to discern falls from normal activities). For this reason, some of the datasets considered in Casilari's work have been reviewed in our analysis as well; in fact, the characteristics of experimental subjects presented in that analysis respect those presented in our study.

11.6 Discussion

In this chapter, we have considered the issue of reproducibility of works in activity recognition using data recorded with wearable devices, in particular working with inertial sensors. We have decided to study this topic by performing a literature review that has finally included 146 articles in order to answer our starting research questions.

We have measured what percentage of works in the literature validate their approach by using

public datasets or by sharing the ones created on purpose. The results have shown that only 33% of considered works (38+10 over 146) used public datasets to validate their results and that 28% (31 out of 110) specially made datasets were shared with the public at the moment of publication.

We have studied the characteristics of considered datasets, with a focus on the amount of data recorded, involved population, and studied activities. We have seen that the first factual problem lies in the absence of basic information regarding the involved population. We could find none of the properties regarding male/female distribution, ages, weights, heights, and BMIs in 22% of the considered datasets. According to our results, the median age of subjects involved is around 29 years, although the most frequently targeted people of AR works belong to the aged people.

On a positive note, with respect to the generalization of approaches over different subjects, and feature extraction, the median number of subjects involved is 10 and the median amount of data recorded is 30 minutes for each subject and 3 minutes for each activity. Within our previous experience on AR, described in Chapters 6 and 7 these numbers can be sufficient to perform a detailed analysis of the proposed approach.

Considering possible ways to improve this review, it would be reasonable to analyze the possibility of reproducing approaches in the literature by analyzing more aspects of methods that should be shared to guarantee reproducibility. Other possible details that should be shared in articles, for example, are implementation information regarding the approach. Moreover, this particular review could be advanced by including more works for example by lowering the threshold for the yearly number of citations.

This analysis could evolve by also checking if there is any trend (positive or negative) in the lack of reproducibility details over time, from the past up to now, for example by replicating our analyses by clustering the works per different years.

Part V

Conclusions

Chapter 12

Considerations and future works

12.1 Summary

In the work of this thesis, we could collaborate with two partners: a company dealing with treatments and care for patients, and a department of IGG, dedicated to the treatment of rheumatic diseases in children and adolescents. Both partners were interested in investigating the usage of accelerometers as an assistance for their studies. The first partner desired to use wearable device data to track the progression of a disease treatment. The latter wanted to use the precision of high-quality sensor data to study rheumatic diseases and the impact of symptoms in daily life activities.

Therefore, starting from the needs shared by our partners, we have studied the use of data obtained from wearable devices and the application of machine learning methods as support for clinical studies. In particular, we started by analyzing the state of the art regarding the usage of wearable data in AR tasks. We have considered different devices for our studies, made by two of the main companies involved in the production of high-quality devices, as described in Chapter 3. We could identify some issues in the work available in the literature, by detecting three areas where we have given our contribution described hereby.

1. As motivated in Chapter 4, publicly available datasets for AR tasks concerning accelerometer data were too generic and not useful for our tasks. For this reason, we have recorded and released to the public a first dataset with healthy adult data recorded with high-quality sensors, while volunteers performed a large set of daily life activities [LFV21a]. We have also seen that this dataset has been already successfully used by other researchers, for different purposes and published articles. Furthermore, we noticed a scarcity of datasets related to children and adolescents suffering from chronic diseases, involving accelerometer data. For this reason, as described in Chapter 5 we recorded and released a second

dataset involving patients of IGG suffering from rheumatic diseases and a control group [FLV$^+$].

2. We started to work on our task by implementing a baseline method for AR, using wearable device data. Our baseline method, described in Chapter 6, mainly consists of a sliding windows approach to extract simple features and the usage of the Support Vector Machine model for classification purposes. Later on, when studying the state of the art, it was noted that most works on AR with wearable devices data proposed impractical approaches. In particular, as illustrated in Chapter 7, the recognition phase was limited to deal *only* with learned activities, but, as previously seen, persons usually perform a plethora of different activities throughout the day. As a consequence, we have proposed a method that helps the classification of non-interesting activities (i.e., "other" activities), reducing the impact on the classification of known activities that are useful from physicians' point of view [FLV21].
   Along with this, we have performed different tests on our proposed approach based on an ensemble method, a technique to filter out transient misclassification, and a final voting mechanism, to try to improve it and for explainability purposes; we listed these tests and experiments in Chapter 8. We have studied, in our *particular scenario*, the behavior of windows' length and overlap, the distribution of votes among different classifiers, the impact of PCA techniques, and the usage of additional features with their importance. Concerning PCA, we have seen that it is possible to reduce the time requested for the training phase by accepting a small loss in accuracy measured as F1-score. The usage of a more complete feature set, including features mostly based on the frequency domain, has not returned higher accuracy with respect to our baseline feature set. This has been justified also by feature importance tests, that have proven a high correlation between different features.
   Additionally, as presented in Chapter 9 we have studied the application of filters, trying to understand if by filtering data there could be an impact on final accuracy, due to the removal of data details useful for the classifier. To do so, we have tested the application of Kalman filter and dictionary learning techniques on our data, as seen in Chapter 10. The results about these last topics have shown that the application of filters on raw data, before preprocessing, affect the final accuracy in a negative way for classification tasks. This can possibly happen due to the loss of signal parts useful for the classifiers.

3. Starting from our perception that a high percentage of proposed approaches related to AR and wearable data were not providing sufficient data to replicate published results, we investigated this issue with a dedicated review. With this work, presented in Chapter 11, we have analyzed the current situation concerning sharing used datasets and biometric information (e.g., weight, height, age, male/female, BMI) concerning populations included in the studies. This revealed that only a third of considered works used public datasets to validate results and that there was an absence of basic population information involved in a quite high percentage of works [FLV22].

12.2 Future Directions

As anticipated in the Introduction of this thesis, the task of *activity recognition* starting from data recorded with wearable devices data could seem a non-innovative and non-exclusive topic since literature already contains works on this problem, which claim excellent results. However, the work devised in this thesis has shown that this topic still requires contributions from the scientific community. Most importantly, we have seen that a good percentage of articles in the literature propose impractical methods or do not share sufficient details to replicate the proposed approach.

The activities completed in this thesis could proceed, in the future, on different topics and in different directions. Considering the part of the thesis related to datasets, at first glance, it looks like we are surrounded by data to learn about and to be used to train ML models. We have observed, instead, that datasets related to specific scenarios are lacking in the literature and are certainly useful. Keeping in mind the difficulties and issues associated with sharing sensitive data (e.g., biometric data, treatment data), the collaboration of the involved parties (i.e., patients, physicians, and specialists) is becoming increasingly essential. These collaborations can lead to the development of technologies to support multiple end users and consumers.

Regarding the part of this thesis dedicated to methods and improvements, we are certain that multiple approaches should consider the possibility of having to classify activities unknown to the classifier, in order to move towards real-life scenarios. Moreover, since the considered topic is closely related to the healthcare field, it is highly necessary for the proposed approaches to dedicate more space to the *explainability* of methods, to better understand the behavior of implemented models. The treated topic regarding the use of filters on raw data before preprocessing would certainly need further studies. Our work has led to preliminary conclusions, but additional tests with different datasets, various considered activities, and the use of different techniques would be helpful. It would be interesting to further explore the behavior of techniques based on neural networks and how they directly relate to raw data and the use of any filter.

Our dataset that deals with people affected by chronic diseases does not include a large number of individuals. Our initial expectations and the timelines required for this work have changed due to the Covid-19 pandemic, coinciding with our studies. However, with the assumption of obtaining data from a larger number of subjects, we can consider producing a model that is more accurate and adaptable to the characteristics of an individual's body (e.g., height, weight). The same reasoning can be extended to different sets of activities to be considered, based on the subject's age and health status.

Moreover, from a high-level point of view, an interesting task regarding our studied topic would be to create a system that is not only able to determine *what* activity is being performed, but also able to measure *how well* each activity is executed. With the recording of our dataset in collaboration with IGG we have taken a first step toward this goal, generating a score that describes the impact of the disease on the execution of daily activities.

Bibliography

[ABMP+10] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, well-being and sports applications: A survey. In *23th International conference on architecture of computing systems 2010*, pages 1–10. VDE, 2010.

[ACMI15] Adamu I Abubakar, Haruna Chiroma, Sanah Abdullahi Muaz, and Libabatu Baballe Ila. A review of the advances in cyber security benchmark datasets for evaluating data-driven based intrusion detection systems. *Procedia Computer Science*, 62:221–227, 2015.

[act] Actigraph centrepoint® insight watch.

[AGKK13] Norhafizan Ahmad, Raja Ariffin Raja Ghazilla, Nazirah M Khairi, and Vijayabaskar Kasi. Reviews on various inertial measurement unit (imu) sensor applications. *International Journal of Signal Processing Systems*, 1(2):256–262, 2013.

[AGO+13] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3, 2013.

[ASN+16] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. In *Workshops at the 30th AAAI Conference on AI*, 2016.

[Ass20] Association for Computing Machinery. Artifact review and badging, 2020.

[bANS+12] Mohd Fikri Azli bin Abdullah, Ali Fahmi Perwira Negara, Md Shohel Sayeed, Deok-Jai Choi, and Kalaiarasi Sonai Muthu. Classification algorithms in human activity recognition using smartphones. *International Journal of Computer and Information Engineering*, 6(77-84):106, 2012.

[BB12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[BFB+23] MH Belau, F Flaßkamp, H Becher, A Hajek, HH König, and L Baumbach. Physical activity in adults with and without rheumatoid arthritis: cross-sectional results from the survey of health, ageing and retirement in europe (share). *Scandinavian Journal of Rheumatology*, pages 1–6, 2023.

[BHW10] Asa Ben-Hur and Jason Weston. A user's guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.

[BPT14] Akram Bayat, Marc Pomplun, and Duc A Tran. A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34:450–457, 2014.

[CLPW21] Ling Chen, Xiaoze Liu, Liangying Peng, and Menghan Wu. Deep learning based multimodal complex human activity recognition using wearable devices. *Applied Intelligence*, 51:4029–4042, 2021.

[CSC+13] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[CSRCG17] Eduardo Casilari, José-Antonio Santoyo-Ramón, and José-Manuel Cano-García. Analysis of public datasets for wearable fall detection systems. *Sensors*, 17(7):1513, 2017.

[CT10] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[CTPK15] Cagatay Catal, Selin Tufekci, Elif Pirmit, and Guner Kocabag. On the use of ensemble of classifiers for accelerometer-based activity recognition. *Applied Soft Computing*, 37:1018–1022, 2015.

[DLHFACQE18] Emiro De-La-Hoz-Franco, Paola Ariza-Colpas, Javier Medina Quero, and Macarena Espinilla. Sensor-based datasets for human activity recognition–a systematic review of literature. *IEEE Access*, 6:59192–59210, 2018.

[DlTHB+09] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009.

[DNRMM16] Rafael Duque, Alicia Nieto-Reyes, Carlos Martínez, and José Luis Montaña. Detecting human movement patterns through data provided by accelerometers. a case study regarding alzheimer's disease. In *Ubiquitous Computing and Ambient Intelligence: 10th International Conference, UCAmI 2016, San Bartolomé de Tirajana, Gran Canaria, Spain, November 29–December 2, 2016, Proceedings, Part I 10*, pages 56–66. Springer, 2016.

[Dru09] Chris Drummond. Replicability is not reproducibility: nor is it good science. 2009.

[dSSFL19] Bruno Samways dos Santos, Maria Teresinha Arns Steiner, Amanda Trojan Fenerich, and Rafael Henrique Palma Lima. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, 138:106120, 2019.

[emb] Empatica embrace2 watch.

[ERMP17] Rabha Elmesmari, John J Reilly, Anne Martin, and James Y Paton. Accelerometer measured levels of moderate-to-vigorous intensity physical activity and sedentary time in children and adolescents with chronic disease: A systematic review and meta-analysis. *PloS one*, 12(6):e0179429, 2017.

[FBB+19] Jonatan Fridolfsson, Mats Börjesson, Christoph Buck, Örjan Ekblom, Elin Ekblom-Bak, Monica Hunsberger, Lauren Lissner, and Daniel Arvidsson. Effects of frequency filtering on intensity and noise in accelerometer-based physical activity measurements. *Sensors*, 19(9):2186, 2019.

[FCS+11] Giovanni Filocamo, Alessandro Consolaro, Benedetta Schiappapietra, Sara Dalprà, Bianca Lattanzi, Silvia Magni-Manzoni, Nicolino Ruperto, Angela Pistorio, Silvia Pederzoli, Adele Civino, et al. A new approach to clinical care of juvenile idiopathic arthritis: the juvenile arthritis multidimensional assessment report. *The Journal of rheumatology*, 38(5):938–953, 2011.

[FLV+] Andrea Fasciglione, Maurizio Leotta, Alessandro Verri, Clara Malattia, and Nicola Ruperto. Daily living activity dataset of juvenile rheumatic patients from wearables data. *IEEE Journal of Biomedical and Health Informatics*, (under review).

[FLV21] Andrea Fasciglione, Maurizio Leotta, and Alessandro Verri. Improving activity recognition while reducing misclassification of unknown activities. In *2021 IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 153–158. IEEE, 2021.

[FLV22] Andrea Fasciglione, Maurizio Leotta, and Alessandro Verri. Reproducibility in activity recognition based on wearable devices: a focus on used datasets. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3178–3185. IEEE, 2022.

[GCBCJG14] Enrique Garcia-Ceja, Ramon F Brena, Jose C Carrasco-Jimenez, and Leonardo Garrido. Long-term activity recognition from wristwatch accelerometer data. *Sensors*, 14(12):22500–22524, 2014.

[GCC+21] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8):1–34, 2021.

[GDC+16] Catarina Godinho, Josefa Domingos, Guilherme Cunha, Ana T Santos, Ricardo M Fernandes, Daisy Abreu, Nilza Gonçalves, Helen Matthews, Tom Isaacs, Joy Duffen, et al. A systematic review of the characteristics and validity of monitoring technologies to assess parkinson's disease. *Journal of neuroengineering and rehabilitation*, 13(1):1–10, 2016.

[GNO22] Gaurvi Goyal, Nicoletta Noceti, and Francesca Odone. Cross-view action recognition with small-scale datasets. *Image and Vision Computing*, 120:104403, 2022.

[Gun20] Odd Erik Gundersen. The reproducibility crisis is real. *AI Mag.*, 41(3):103–106, 2020.

[HGGP+20] Valentin Hamy, Luis Garcia-Gancedo, Andrew Pollard, Anniek Myatt, Jingshu Liu, Andrew Howland, Philip Beineke, Emilia Quattrocchi, Rachel Williams, and Michelle Crouthamel. Developing smartphone-based objective assessments of physical function in rheumatoid arthritis patients: the parade study. *Digital biomarkers*, 4(1):26–44, 2020.

[HJ09] Zhenyu He and Lianwen Jin. Activity recognition from acceleration data based on discrete consine transform and svm. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 5041–5044. IEEE, 2009.

[HM22] Lee B Hinkle and Vangelis Metsis. Individual convolution of ankle, hip, and wrist data for activities-of-daily-living classification. In *2022 18th International Conference on Intelligent Environments (IE)*, pages 1–4. IEEE, 2022.

[HM23] Lee B Hinkle and Vangelis Metsis. An llvm-inspired framework for unified processing of multimodal time-series data. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 91–94, 2023.

[HR14] Eu Tteum Ha and Kwang Ryel Ryu. Activity recognition by smartphone accelerometer data using ensemble learning methods. *Int. J. Electr. Comput. Eng*, 8:480–483, 2014.

[Hut18] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[HYCL18] Yu-Liang Hsu, Shih-Chin Yang, Hsing-Cheng Chang, and Hung-Che Lai. Human daily and sport activity recognition using a wearable inertial sensor network. *IEEE Access*, 6:31715–31728, 2018.

[JF12] Dinesh John and Patty Freedson. Actigraph and actical physical activity monitors: a peek under the hood. *Medicine and science in sports and exercise*, 44(1 Suppl 1):S86, 2012.

[JNJSS18] Artur Jordao, Antonio C Nazare Jr, Jessica Sena, and William Robson Schwartz. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*, 2018.

[KAMP14] Eliasz Kantoch, Piotr Augustyniak, M Markiewicz, and D Prusak. Monitoring activities of daily living based on wearable wireless body sensor network. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 586–589. IEEE, 2014.

[KC07] B. Kitchenham and S Charters. Guidelines for performing systematic literature reviews in software engineering, 2007.

[KMP+05] Florence Peterson Kendall, Elizabeth Kendall McCreary, Patricia Geise Provance, Mary McIntyre Rodgers, William Anthony Romani, et al. *Muscles: testing and function with posture and pain*, volume 5. Lippincott Williams & Wilkins Baltimore, MD, 2005.

[LBK+21] Aleksej Logacjov, Kerstin Bach, Atle Kongsvold, Hilde Bremseth Bårdstu, and Paul Jarle Mork. Harth: a human activity recognition dataset for machine learning. *Sensors*, 21(23):7853, 2021.

[LC23] Emanuele Lattanzi and Lorenzo Calisti. Energy-aware tiny machine learning for sensor-based hand-washing recognition. In *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*, pages 15–22, 2023.

[LCF22] Emanuele Lattanzi, Lorenzo Calisti, and Valerio Freschi. Unstructured handwashing recognition using smartwatch to reduce contact transmission of pathogens. *Ieee Access*, 10:83111–83124, 2022.

[LFV21a] Maurizio Leotta, Andrea Fasciglione, and Alessandro Verri. Daily living activity recognition using wearable devices: A features-rich dataset and a novel approach. In A. Del Bimbo et al., editor, *Proceedings of 25th International Conference on Pattern Recognition Workshops (ICPR 2021 Workshops)*, volume 12662 of *LNCS*. Springer, 2021.

[LFV21b] Maurizio Leotta, Andrea Fasciglione, and Alessandro Verri. Daily living activity recognition using wearable devices: A features-rich dataset and a novel approach. In *International Conference on Pattern Recognition*, pages 171–187. Springer, 2021.

[LL12] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

[LLR+99] Daniel J Lovell, Carol B Lindsley, Robert M Rennebohm, Susan H Ballinger, Suzanne L Bowyer, Edward H Giannini, Jeanne E Hicks, Joseph E Levinson, Richard Mier, Lauren M Pachman, et al. Development of validated disease activity and damage indices for the juvenile idiopathic inflammatory myopathies: Ii. the childhood myositis assessment scale (cmas): a quantitative tool for the evaluation of muscle function. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 42(10):2213–2219, 1999.

[LSE13] Heike Leutheuser, Dominik Schuldhaus, and Bjoern M Eskofier. Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset. *PloS one*, 8(10):e75196, 2013.

[MDK64] M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *JBJS*, 46(2):335–360, 1964.

[MIR+13] Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise*, 45(11):2193, 2013.

[MJHJ22] Sakorn Mekruksavanich, Ponnipa Jantawong, Narit Hnoohom, and Anuchit Jitpattanakul. Deep learning models for daily living activity recognition based on wearable inertial sensors. In *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2022.

[MMN17] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 7(10):1101, 2017.

[MWM+21] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

[NHCB19] James A Nichols, Hsien W Herbert Chan, and Matthew AB Baker. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical reviews*, 11(1):111–118, 2019.

[NJK20] Amir Nadeem, Ahmad Jalal, and Kibum Kim. Accurate physical activity recognition using multidimensional features and markov model for smart health fitness. *Symmetry*, 12(11):1766, 2020.

[NZTZ15] Le T Nguyen, Ming Zeng, Patrick Tague, and Joy Zhang. I did not smoke 100 cigarettes today! avoiding false positives in real-world activity recognition. In *Proceedings of the 2015 ACM Int. Joint Conference on Pervasive and Ubiquitous Computing*, pages 1053–1063, 2015.

[Org00] World Health Organization. Obesity: preventing and managing the global epidemic. 2000.

[PGZL20] Ivan Miguel Pires, Nuno M Garcia, Eftim Zdravevski, and Petre Lameski. Activities of daily living with motion: A dataset with accelerometer, magnetometer and gyroscope data from mobile devices. *Data in brief*, 33:106628, 2020.

[Pha14] Thomas Phan. Improving activity recognition via automatic decision tree pruning. In *Proceedings of the 2014 ACM Int. J. Conf. on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 827–832, 2014.

[Ple18] Hans E Plesser. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76, 2018.

[PSW19] Andrea Prati, Caifeng Shan, and Kevin I-Kai Wang. Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. *Journal of Ambient Intelligence and Smart Environments*, 11(1):5–22, 2019.

[PTP+06] F Pitta, Thierry Troosters, VS Probst, MA Spruit, Marc Decramer, and Rik Gosselink. Quantifying physical activity in daily life with questionnaires and motion sensors in copd. *European respiratory journal*, 27(5):1040–1055, 2006.

[Ras23] Fabian Rast. Labeled inertial sensor data of children with mobility impairments for activity recognition purposes, 2023. V1.

[RG05] Dymitr Ruta and Bogdan Gabrys. Classifier selection for majority voting. *Information fusion*, 6(1):63–81, 2005.

[RMS17] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):667–681, 2017.

[RT12] Kristi M Robusto and Stewart G Trost. Comparison of three generations of actigraph™ activity monitors in children and adolescents. *Journal of sports sciences*, 30(13):1429–1435, 2012.

[SBI+16] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, 16(4):426, 2016.

[SHH+15] John Staudenmayer, Shai He, Amanda Hickey, Jeffer Sasaki, and Patty Freedson. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *Journal of applied physiology*, 119(4):396–403, 2015.

[SLES11] Maja Stikic, Diane Larlus, Sandra Ebert, and Bernt Schiele. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2521–2537, 2011.

[SLKR20] Sheeba Samuel, Frank Löffler, and Birgitta König-Ries. Machine learning pipelines: provenance, reproducibility and fair data principles. In *Provenance and Annotation of Data and Processes*, pages 226–230. Springer, 2020.

[SMVS01] Tanveer Syeda-Mahmood, A Vasilescu, and Saratendu Sethi. Recognizing action events from multiple viewpoints. In *Proceedings IEEE Workshop on Detection and Recognition of Events in Video*, pages 64–72. IEEE, 2001.

[SS16] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2016.

[SSH13] Muhammad Shoaib, Hans Scholten, and Paul JM Havinga. Towards physical activity recognition using smartphone sensors. In *2013 IEEE 10th international conference on ubiquitous intelligence and computing and 2013 IEEE 10th international conference on autonomic and trusted computing*, pages 80–87. IEEE, 2013.

[SVLS08] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. Exploring semi-supervised and active learning for activity recognition. In *2008 12th IEEE Int. Symp. on Wearable Computers*, pages 81–88. IEEE, 2008.

[TDF+18] Niall Twomey, Tom Diethe, Xenofon Fafoutis, Atis Elsts, Ryan McConville, Peter Flach, and Ian Craddock. A comprehensive study of activity recognition using accelerometers. In *Informatics*, volume 5, page 27. Multidisciplinary Digital Publishing Institute, 2018.

[TFM+18] Emanuele Torti, Alessandro Fontanella, Mirto Musci, Nicola Blago, Danilo Pau, Francesco Leporati, and Marco Piastra. Embedded real-time fall detection with deep learning on wearable devices. In *2018 21st euromicro conference on digital system design (DSD)*, pages 405–412. IEEE, 2018.

[VDDP+21] Gloria Vergara-Diaz, Jean-Francois Daneault, Federico Parisi, Chen Admati, Christina Alfonso, Matilde Bertoli, Edoardo Bonizzoni, Gabriela Ferreira Carvalho, Gianluca Costante, Eric Eduardo Fabara, et al. Limb and trunk accelerometer data collected with wearable sensors from subjects with parkinson's disease. *Scientific Data*, 8(1):47, 2021.

[VFC+13] GC Varnier, C Ferrari, A Consolaro, D Marafon, C Pilkington, S Maillard, M Jelusic Drazic, S Dalpra, A Civino, A Martini, et al. Pres-final-2012: introducing a new approach to clinical care of juvenile dermatomyositis: the juvenile dermatomyositis multidimensional assessment report. *Pediatric Rheumatology*, 11(2):1–2, 2013.

[VLS11] Ulrike Von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier, 2011.

[VW17] Vijay R Varma and Amber Watts. Daily physical activity patterns during the early stage of alzheimer's disease. *Journal of Alzheimer's Disease*, 55(2):659–667, 2017.

[WGC+10] Liang Wang, Tao Gu, Hanhua Chen, Xianping Tao, and Jian Lu. Real-time activity recognition in wireless body sensor networks: From simple gestures to complex activities. In *2010 IEEE 16th Int. Conf. on Embedded and Real-Time Computing Systems and Applications*, pages 43–52. IEEE, 2010.

[Woj21] Janusz Wojtusiak. Reproducibility, transparency and evaluation of machine learning in health applications. In *HEALTHINF*, pages 685–692, 2021.

[WSP+11] Aner Weiss, Sarvi Sharifi, Meir Plotnik, Jeroen PP van Vugt, Nir Giladi, and Jeffrey M Hausdorff. Toward automated, at-home assessment of mobility among patients with parkinson disease, using a body-worn accelerometer. *Neurorehabilitation and neural repair*, 25(9):810–818, 2011.

[YCL+07] Jhun-Ying Yang, Yen-Ping Chen, Gwo-Yun Lee, Shun-Nan Liou, and Jeen-Shing Wang. Activity recognition using one triaxial accelerometer: A neuro-fuzzy classifier with feature reduction. In *International conference on entertainment computing*, pages 395–400. Springer, 2007.

[YSZ17] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[YWM+19] Guan Yuan, Zhaohui Wang, Fanrong Meng, Qiuyan Yan, and Shixiong Xia. An overview of human activity recognition based on smartphone. *Sensor Review*, 2019.

[ZLZ+22] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476, 2022.

[ZRMH12] Shaoyan Zhang, Alex V Rowlands, Peter Murray, and Tina L Hurst. Physical activity classification using the genea wrist-worn accelerometer. *Medicine and science in sports and exercise*, 44(4):742–748, 2012.

[ZS12] Mi Zhang and Alexander A Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 1036–1043, 2012.