Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer and vital Signals

Mahsa Sadat Afzali Arani

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

January 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mahsa Sadat Afzali Arani**

Entitled: **Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer and vital Signals**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Jun Cai

_____ Supervisor
Dr. Emad Shihab

Approved by _____
Lata Narayanan, Chair
Department of Computer Science and Software Engineering

_____ 2022

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer and vital Signals

Mahsa Sadat Afzali Arani

Inertial sensors (IMU) are widely used in the field of human activity recognition (HAR), since this source of information is the most informative time series among non-visual datasets. HAR researchers are actively exploring other approaches such as different feature extraction methods, machine learning models and classifiers, different sources of signals and sensor positioning to improve the performance of HAR systems.

Human physical activities have a significant impact on human body, specifically heart activity and oxygen delivery, thus, we explore heart activity related bio-signals to check if these signals are advantageous in the field of HAR research. In this thesis, we investigate the impact of combining bio-signals with dataset acquired from accelerometer on recognizing human daily activities. To achieve this aim, we used PPG-DaLiA dataset consisting of 3D-accelerometer (3D-ACC), electrocardiogram (ECG), photoplethysmogram (PPG) signals acquired from 15 individuals while performing daily activities. We extracted hand-crafted time and frequency domain features, then we applied correlation-based feature selection approach to reduce feature-set dimensionality. After introducing early fusion scenarios, we trained and tested random forest models with subject-dependent and subject-independent setups. Our results indicate that combining features extracted from 3D-ACC signal with ECG signal improve the classifier's performance F1-scores by 2.72% and 3.00% (from 94.07% to 96.80%, and 83.16% to 86.17%) for subject-dependent and subject-independent approaches, respectively.

Related Publication

The following publication is related to the current thesis:

- **M.S. Afzali Arani**, D.E. Costa, E. Shihab. Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals. *Sensors. 2021 Jan;21(21):6997.*

Acknowledgments

Above all, I would like to say a special thank you to my supervisor, Dr. Emad Shihab, for trusting me and letting me to be part of the Data-driven Analysis of Software (DAS) team and pursue my master's study. I really appreciate his continuous support and advise during my study and research. He also helped me experience the life after graduation by referring me to Motsai research company as an intern. Furthermore, I would like to thank Dr. Diego Elias Costa for his help and insightful comments and suggestions in editing my thesis (and our paper).

Last but not least, I want to thank my family for supporting me emotionally during my though times, even though we are apart by hundreds of kilometers; and my partner, who was always supportive and inspiring specially during the hard time of Covid19 pandemic and quarantine period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Problem Explanation

With the recent increase in the use of smart phones and wearable devices, we can record and access a plethora of raw data and information from built-in inexpensive sensors. Human activity recognition (HAR) refers to analyzing this data to extract meaningful information about human daily habits and physical activity patterns Chung, Lim, Noh, Kim, and Jeong (2019). Breakthroughs in HAR research has led to various applications in health care and rehabilitation, elderly fall detection, fitness trackers, assisted living and smart homes Y. Wang, Cang, and Yu (2019). One of the most frequently used sources of data for activity recognition purposes is the inertial sensor Aguileta, Brena, Mayora, Molino-Minero-Re, and Trejo (2019); Demrozi, Pravadelli, Bihorac, and Rashidi (2020); Lara and Labrador (2012). Inertial measurement unit (IMU) contains triaxial accelerometer (3D-ACC), gyroscope, and magnetometer to measure velocity and acceleration, rotation, and strength of a magnetic field, respectively. Based on previous studies, the 3D-ACC sensor outperforms gyroscope and magnetometer; interestingly, combining 3D-ACC with gyroscope performs better in classifying activities A. Wang, Chen, Yang, Zhao, and Chang (2016). This suggests sensor combination has the potential to offer better classification power for a HAR system. To improve a HAR system, researchers investigate different sources of signals as well as different analyzing approaches. In many cases, researchers evaluate the significance of one signal individually or IMU signals combination, thus, combined contribution of other types of signals (e.g, bio-signals) require

more exploration Ravi, Wong, Lo, and Yang (2016); Shoaib, Bosch, Incel, Scholten, and Havinga (2016).

1.2 Motivation

In this thesis, we aim to investigate the impact of using bio-signals on the performance of a HAR system. Bio-signal refers to any signal generated by a living creature and can be recorded continuously Hadjileontiadis (2006). Bio-signals sensors have been shown to be quite accurate in capturing the bio-signals Athavale and Krishnan (2017), but they have not yet been extensively explored in the context of HAR systems. Given that heart rate is sensitive to physically demanding activities Strath et al. (2013), can we rely on bio-signals to complement 3D-ACC sensors in recognizing certain types of activities?

To compare the performance of different signals, we analyse the data acquired from 3D-ACC, electrocardiogram (ECG) and, photoplethysmogram (PPG) sensors individually. Moreover, we use fusion methods to combine data from mentioned signals to examine their contribution level in the system's output. Fusion methods refer to any integration of information from different sources, at the sensor, feature or classifier levels Aguileta et al. (2019). To analyse the signals' contribution in HAR systems, we segment the signals, using a sliding window method to extract time and frequency domain features. Finally, we train random forest classifier models for subject-dependent and subject-independent setups.

We evaluate the bio-signals significance in HAR using two types of models: subject-specific and cross-subject models. Both models are commonly used in HAR systems and research, and more importantly, each has its advantages and disadvantages Lee, Khan, and Kim (2011); Micucci, Mobilio, and Napoletano (2017). Subject-specific models are personalized models, trained and evaluated using the data of a single user. Hence, subject-specific are usually more accurate than cross-subject models, at a cost of requiring training data from the target user. A cross-subject model, on the other hand, is trained on multiple users and attempts to recognize the activity of a previously untrained user. This model tends to be more generic and is commonly used in practice, since cross-subject models are cheaper to train and easier to deploy Ferrari, Micucci, Mobilio, and

Napoletano (2021); Lara, Pérez, Labrador, and Posada (2012).

1.3 Contributions

In relation to the aforementioned points, more investigation about the bio-signals' contribution, fusion approaches and different evaluation setups, we formulate our research questions to cover both subject-specific and cross-subject models. Thus, in our thesis, we focus on answering two research questions.

RQ1: What is the contribution level of signals under study in subject-specific HAR systems? With this research question we investigate the impact of bio-signals in subject-specific setup, meaning how ECG and PPG signals recorded from the same subject would help a HAR system for its future classifications.

RQ2: What is the contribution level of signals under study in cross-subject HAR systems? Since cross-subject models are more complicated, we want to further investigate the impact of bio-signals in addition to the 3D-ACC signal for potential added values.

In addition to answering both of the research questions, we explore and discuss other aspect of fusing the signals, such as per activity contribution level or the impact of physical characteristics on contribution level of bio-signals in HAR systems. To summarize, we provide an overview of our contributions:

- To the best of our knowledge, we are the first study to compare the combined performance of the 3D-ACC, ECG and PPG signals recorded simultaneously from subjects performing the same set of activities. We use hand-crafted features to evaluate the performance of classifiers for HAR.
- We investigate the significance of bio-signals and compare the usefulness of ECG and PPG signals in HAR.
- We investigate the impact of combining 3D-ACC signal with ECG signal in recognizing some specific activities in detail. For instance, the importance of ECG signal in distinguishing walking activity from ascending/descending stairs.

- For further analyzes on cross-subject models, we study the impact of fitness level, height, and weight, etc. on the contribution level of ECG signal
- We also explore and compare the late fusion approach with the early fusion method

1.4 Overview

The rest of this thesis is organized as follows; we describe the related work in Chapter 2. In Chapter 3, we elaborate on the characteristics of the sensors and signals we evaluate then we review detailed information about the dataset that we use. Next, in Chapter 4, we explain our methodology and workflow, from data pre-processing to feature extraction and selection. We allocate Chapter 5 to explain the fusion scenarios as well as classification and evaluation methods. Chapter 6 describes the results and findings of our work and we answer our research questions. In Chapter 7, we discuss the impact of ECG signal in HAR system's performance per activity and with respect to the subjects physical characteristics. Moreover, we explore and compare the late fusion method with the early fusion approach and provide its results. Lastly, in our discussion we point out the most important features for the best fusion scenario. We conclude this thesis by mentioning limitation of our work as well as the future work, and finally, the overall conclusion of this thesis in Chapter 8.

Chapter 2

Related Work

In this chapter, we discuss the results and insights from previous studies. HAR research leverages multiple approaches to advance the accuracy and performance of systems, such as different sensor positioning Chung et al. (2019), varying feature extraction approaches Machado, Gomes, Gamboa, Paixão, and Costa (2015) and exploring several classification methods Catal, Tufekci, Pirmit, and Kocabag (2015); Chen, Zhong, Zhang, Sun, and Zhao (2016) to extract more informative knowledge from dataset and enhance the HAR systems performance.

In terms of different sensor positioning using a single sensor, Bayat et al., conducted a study to evaluate the effectiveness of builtin 3D-ACC in a smartphone in recognizing human activities Bayat, Pomplun, and Tran (2014). Participants were instructed to perform tasks first using the smartphone in their hands then perform the same tasks with the phone in their pockets. Based on the reported result, smart phone positioning in hand and in the pocket yield similar results in the HAR models. Moreover, they applied the average of probabilities method to combined classifiers and they claimed this method outperforms using only one classifier. Casale et al., used data acquired from only one 3D-ACC sensor and proposed a new set of feature extraction methods to be used with a random forest classifier in recognizing five different daily activities Casale, Pujol, and Radeva (2011). They observed that the new feature set is more informative compared to the commonly used feature sets, thus they reported model's performance enhancements after using the new feature set.

To overcome the limitations of using a single sensor, researchers have explored the idea of applying fusion methods to enhance a HAR system Aguileta et al. (2019). Fusion methods can be

categorized into early fusion and late fusion methods Vrigkas, Nikou, and Kakadiaris (2015).

To elaborate, early fusion refers to combining raw data or extracted features, then feed the new combined dataset or feature set to a classifier Mendes Jr, Vieira, Pires, and Stevan Jr (2016). Thus, early fusion can be categorized accordingly: 1. data-level or sensor-level fusion Mendes Jr et al. (2016), 2. feature-level fusion. Data-level fusion is an approach in which data acquired from multiple sensors of same type (homogeneous sensor), or varying type of sensors (heterogeneous sensors Nweke, Teh, Mujtaba, and Al-Garadi (2019)) positioning on different body parts. Feature-level fusion, on the other hand, refers to combining extracted features from varying modalities into a new feature matrix with larger dimensions. It is relevant to mention that combining features extracted with different approaches, i.e. time domain and frequency domain features, is also considered as early fusion Nweke et al. (2019).

Another combination method for a HAR system is called late fusion, or decision-level fusion with the purpose of providing stronger classifier models. As the name implies, in this method the models are trained separately on each signal, but their predictions are combined into an ensemble model (e.g., voting system), that predicts the final classifications Vrigkas et al. (2015). In this work, we apply the early fusion method for cross-subject and within-subject setups, moreover, we explore and discuss the late fusion method for the cross-subject setup.

2.1 Early Fusion with IMU

Early fusion methods fuse extracted features from different signal sources into a combined dataset, which serve as input for human-activity classifier. Several works have applied this technique to improve the performance of classifier models. Concerning IMU signal fusion, Chung et al. applied the sensor fusion approach by placing eight IMU sensors on different body parts of five right-handed individuals Chung et al. (2019). They trained a Long Short-Term Memory (LSTM) network model to classify nine activities. Based on their results, to get a reasonable classification performance one sensor should be placed on upper half of the body and one on lower half; particularly, on right wrist and right ankle. Regarding signal fusion, the authors stated that 3D-ACC sensor

combined with gyroscope performed better (with accuracy 93.07%) than its combination with magnetometer. Moreover, Shoaib et al. followed the same data-level fusion approach and generated their own data set by placing one smart phone in subject's pocket and another one on his dominant wrist and recording 3D-ACC, gyroscope, linear acceleration signals Shoaib et al. (2016). They tried different scenarios such as combination of 3D-ACC and gyroscope signals which they claim leads to more accurate results, particularly for "stairs" and "walking" activities. Moreover, they claim that the combination of signals captured from both packet and wrist improve the performance specially for complex activities.

2.2 Early Fusion with Bio-Signals

Regarding using bio-signal combination approach, Park et al. experiment was to extract Heart-Rate Variability (HRV) parameters from recorded ECG data and combine it with 3D-ACC signal for HAR researches Park, Dong, Lee, and Youn (2017). They employed feature-level fusion approach by fusing features extracted from HRV and 3D-ACC signals. They classified five activities by examining three different scenarios. First, using only four features extracted from 3D-ACC (83.08%); second, considering 31 more features extracted from ECG signal along with ones used in the first scenario (94.81%). Finally, using features extracted from 3D-ACC and only some selected features from ECG signal which outperformed previous scenarios by achieving 96.35% accuracy. Park et al. conclude that ECG signal performs as a complementary source of information along with 3D-ACC for HAR researches. Tapia et al. applied early fusion by recording acceleration signal obtained from five 3D-ACC in addition to heart rate (HR) information Tapia et al. (2007). The authors, applied C4.5 decision tree and Naïve Bayes classifier to classify 30 gymnastic activities with different levels of intensity. They claim adding HR to 3D-ACC can improve the models' performance by 1.20% and 2.10% for subject-dependent and subject-independent approaches, respectively. Based on Tapia et al. (2007), for subject-independent approach, different fitness level and variation in heartbeat rate during non-resting activities are potential reasons for this minor recognition improvement.

As for PPG signal fusion, Biagetti et al. investigated the level of contribution of PPG signal in addition to 3D-ACC signal toward accurately detecting human activities Biagetti, Crippa,

Falaschetti, Orcioni, and Turchetti (2017). The authors proposed feature extraction technique based on singular value decomposition (SVD) in addition to Karhunen-Loeve transform (KLT) method for feature reduction. According to the authors, employing only PPG signal is not enough for physical activity recognition. Thus, they compared applying only 3D-ACC signal with a combination of PPG and 3D-ACC, consequently, they conclude that signal fusion incremented the overall accuracy by 12.30% to 78.00%. In another study, Mehrange et al. used a single PPG-ACC wrist-worn sensor placed on the dominant wrist of 25 male subjects to evaluate fused HAR system power in classifying indoor activities with different intensity Mehrang, Pietilä, and Korhonen (2018). They extracted time and frequency domain features and feed it to a random forest classifier. In terms of contribution level of PPG-based HR related features in classifying activities, their results suggests a very slight overall improvement. Regarding per activity performance, HR addition did not help the classifier to indicate most of the activities except for intensive stationary cycling with 7% improvement in accuracy.

2.3 Late fusion

An empirical study was conducted by Sun el al. Sun, Kamel, and Wong (2005) to compare different late fusion strategies, namely, classifier-, class- and sample-based weighting schemes. In terms of fusing different type of classifiers, Catal et al. proposed an ensemble classifier model to detect six different activities present in the WISDM dataset Catal et al. (2015). They trained three classifiers, particularly, a decision tree algorithm named J48, logistic regression, and a multi-layer perceptron neural network. Then as their proposed model, they combined the mentioned classifiers using voting strategies to get the final label for each class. The authors claimed their model outperformed using each classifier separately. The authors also tested different voting strategies, stating that the average of probabilities yielded the best performance.

On the other hand, class-based weighting late fusion approach was applied by Chowdhury et al. Chowdhury, Tjondronegoro, Chandran, and Trost (2017) on two datasets, both recording acceleration data while 3D-ACC sensors were placed on three different body parts. As part of the aim of their study, they evaluate and compare the sensor positioning using fusion methods on data obtained

from sensors positioned on different body parts (ankle, chest, and wrist). As weighting metric, they used F1-score obtained after applying the 10-fold cross validation method on training data, then they adjusted the obtained weights and calculated the final prediction on test data. Tsanousa et al. Tsanousa, Meditskos, Vrochidis, and Kompatsiaris (2019), also investigated the class-based weighted late fusion approach, however, modified the weighting strategy. Instead of F1-score, they proposed using class weights based on detection rate. Detection rate is the ratio of true positive to all the instances. They used this method to late-fuse models trained with 3D-ACC and gyroscope data.

This thesis differs and complements the former studies in the following ways. First, thanks to the dataset that we used, we have 3D-ACC, ECG and PPG signals all recorded simultaneously and related to same group of subjects, thus, beside evaluating the added value of bio-signals to 3D-ACC, we can also compare the significance of each of the mentioned bio-signals. Moreover, we investigate the impact of bio-signal, not only on the overall performance of the HAR models, but also per single activity, to assess the impact of bio-signals on each set of activities. We use the class-based weighting late fusion approach for the LOSO setup to evaluate the contribution level of 3D-ACC, ECG and PPG signals using both mentioned weighting schemas. Also, we compare the performance of the late fusion approach to the early fusion one.

Chapter 3

Studied Dataset

In this chapter, first, we describe the characteristics of each signal under study (Section 3.1), which is relevant to our methodology. Then, we describe the publicly available dataset we use in this thesis (Section 3.2).

3.1 Sensors and Signals Overview

In our thesis, we consider three sources of signals, 3D-ACC, ECG, and PPG, for each of which, we outline a brief explanation. Inertial Measurement Unit, known as IMU sensor, is a set of measurement units placed together in one device to capture information about kinetic status of a device. This measurement tool may include a 3D-ACC, gyroscope and magnetometer sensors. 3D-ACC is a source of information frequently used in HAR researches and applications. This sensor is an mechanical device converting mechanical forces to electrical signals. Thus, 3D-ACC is capable of measuring constant forces caused by gravity and rotation along three axes, in addition to dynamic forces such as acceleration and vibration (Andrejašic, 2008). Having this knowledge is critical for the feature extraction phase. Figure 3.1 depicts 3D-ACC signal while subject number 1 was performing "Sitting" and "Cycling" activities. Notice the 3D-ACC signal fluctuations in cycling activity, compared to the one related to the sitting activity.

Bio-signals are capable of capturing meaningful information about human body. ECG is one of the bio-signals and generated by electrical activity of the heart. In order to record this electrical

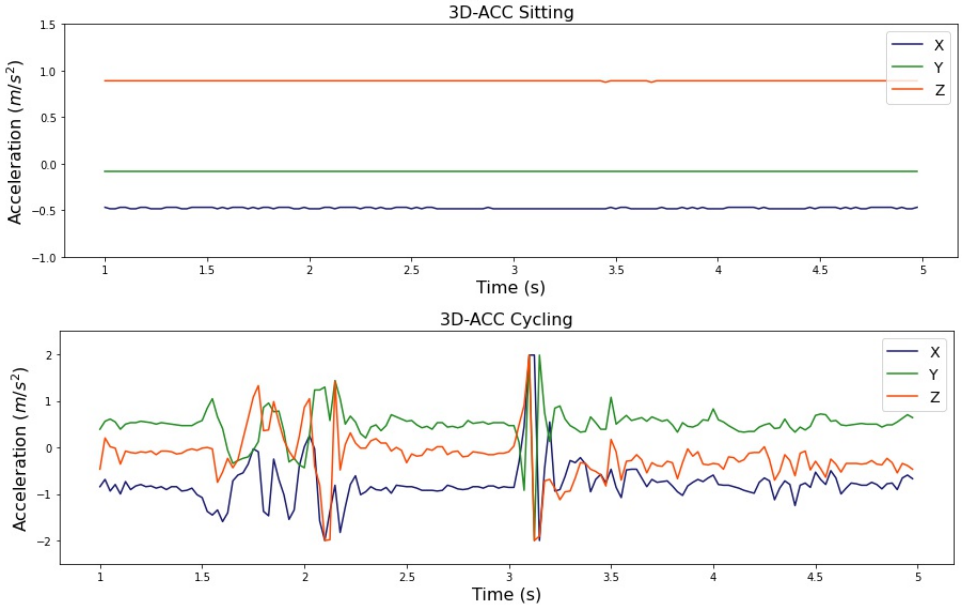


Figure 3.1: Five seconds window of 3D-ACC signal related to "Sitting" and "Cycling" activities performed by subject number 1.

activity, a certain number of electrodes must be placed on a person's chest; these electrodes record changes in voltage during each phase of cardiac cycle, then the recorded voltage is plotted against time based on the sampling rate frequency. ECG signal has a specific pattern and a complete ECG period is made up of different intervals corresponding to a specific phase. The most well-known and obvious peak in one period of ECG signal is called the R peak which represents one heartbeat (McSharry, Clifford, Tarassenko, & Smith, 2003). Counting R peaks in a fixed time interval is equivalent to the number of heartbeats during that specific time interval, thus, this information is capable of illustrating how fast the heart is beating, which may be a source of information in HAR researches. In figure 3.2, we compare two 5 seconds time windows of ECG signal related to "sitting" and "cycling" activities.

Recently, another source of information has been used for HAR research called PPG signal. This signal is generated when infrared light passes through a human finger, wrist or earlobe, then captured by a light-detector after crossing the body part. During this process some of the light is absorbed by the skin, bones and especially by hemoglobin protein in the red blood cells, the rest of the light is recorded. The resulting signal is called the PPG signal (Castaneda, Esparza, Ghamari,

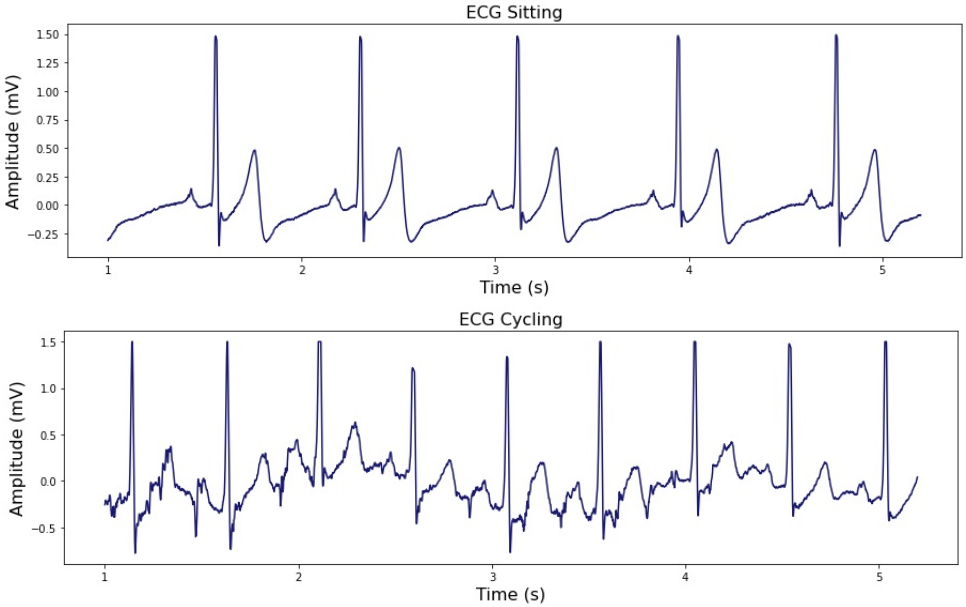


Figure 3.2: Five seconds window of ECG signal related to "Sitting" and "Cycling" activities performed by one subject

Soltanpur, & Nazeran, 2018). Every time the heart pumps blood throughout the body, more lights gets absorbed by the hemoglobin protein and the PPG signal can represent this as heart beats. An advantage of using PPG signal over ECG is that PPG can be recorded via a wrist-worn device, which is a more convenient solution for a subject to wear, than ECG signal that needs a chest-worn device. However, PPG signals suffer from motion artifacts. Motion artifacts refer to any sort of voluntary or involuntary movements that are recorded by the PPG sensor but are actually noise associated with the signal. Luckily, there are ways to eliminate motion artifacts or even use them; as Boukhechba et al., (Boukhechba, Cai, Wu, & Barnes, 2019) analyzed the PPG signal and decomposed this signal to cardiac and respiratory signals; moreover, they took advantage of motion artifact noise associated with PPG signal to recognize five type of human daily activities. Figure 3.3 presents two 5 seconds sample of PPG signal, again for "Sitting" and "Cycling" activities, respectively.

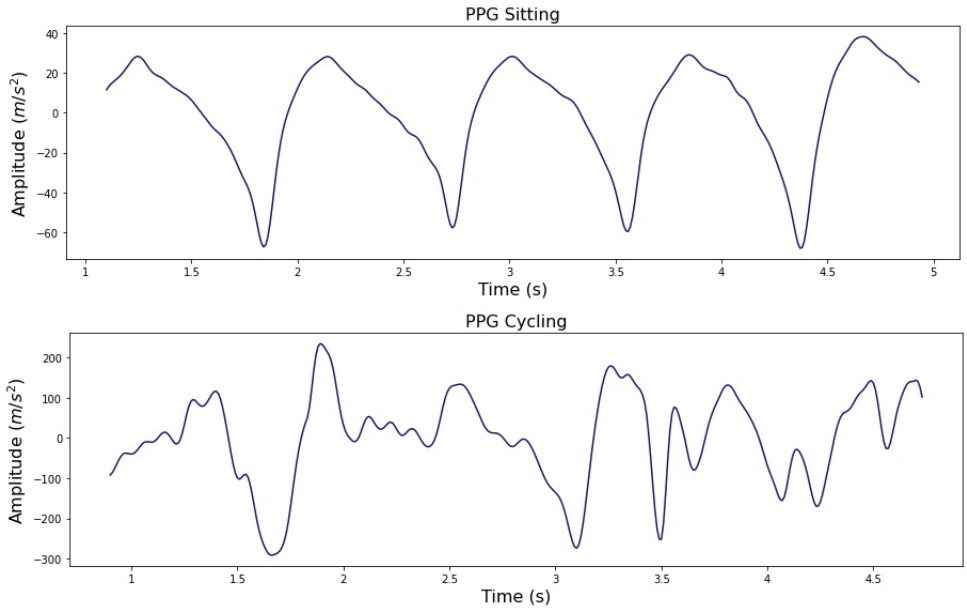


Figure 3.3: Five seconds window of PPG signal related to "Sitting" and "Cycling" activities performed by subject number 1.

3.2 Used Dataset

We use the PPG-DaLiA dataset (Reiss et al., 2019) in our thesis. This dataset was collected by Reiss et al. (Reiss et al., 2019), in which 15 volunteers, 7 male and 8 female, took part. Table 3.2 presents physical characteristics about the participants. Reiss et al. used two different devices to acquire desired signals. A chest-worn device, RespiBAN (*BiosignalsPLUX. RespiBAN Professional. 2019. Available online:*, 2019) was placed on each subjects' chest to record ECG signals, their respiration, and 3D-ACC at a 700 Hz sampling rate. In addition, subjects wore a device on their non-dominant wrist called Empatica E4 (*Empatica. E4Wristband. 2018. Available online:*, 2018) to collect 3D-ACC at a 32Hz sampling rate, blood volume pulse (BVP) signal which contains PPG signal at 64Hz, electrodermal activity (EDA) and body temperature both at 4Hz. After placing the mentioned devices on volunteers' chest and wrist, Reiss et al. asked them to perform some daily activities: sitting, ascending and descending stairs, playing table soccer, outdoor cycling, driving a car, being on lunch break, walking and working. Beside the mentioned activities, the authors also recorded the transient activities in between each of the aforementioned activities.

A noteworthy point about the PPG-DaLiA dataset is that the initial motivation for its collection

Table 3.1: Physical characteristics of the participants.

Characteristics	Average	Min	Max
Age (years)	30.14	21	55
Height (cm)	175.21	164	195
Weight (kg)	68.93	56	105
Fitness level	4.79	2	6

was not for HAR research, but for extracting heart rate estimations using the PPG signal. This is why the chest-worn device sampling rate is relatively high compared to wrist-worn recorded signals in this dataset. We select this dataset for the following reasons; first and foremost, the experiments were performed indoors and outdoors which makes it more realistic and closer to daily life experiences compared to in-lab signal recording. While the researchers provided some protocols to instruct participants, they were free to execute each task in their own natural way. In addition to the recorded dataset, labels were provided for each activity, along with other significant information about each individual's age, height, weight, fitness level, gender and skin type. Each subject performed all the activities for a total duration of 2.5 hours. Table 3.2 provides detailed information about each activity completion protocol.

Table 3.2: Type of activities and detailed protocol based on the study of Reiss et al. (Reiss et al., 2019)

ID	Activity	Protocol
1	Sitting	Sitting while reading
2	Ascending/ descending stairs	Climbing six floors up and going down, repeated two times
3	Play table soccer	Playing table soccer, 1 vs. 1
4	Cycling	Cycling 2km outdoors cycling with gravel and paved road condition
5	Driving a car	Driving on a defined road for 15 minutes
6	Lunch break	Includes queuing and fetching food, eating, and talking at the table
7	Walking	Walking back from the canteen to the office, with some detour
8	Working	Subjects' work mainly consisted of working on a computer.
0	Transient periods	Each transition between activities

We analyze only five human activities from the dataset: sitting, ascending/descending stairs, playing table soccer, outdoor cycling, and walking. It is important to highlight that this dataset is an

imbalanced dataset. That is, more than 50% of all the instances are related to walking and sitting activities, 27% and 24% respectively; and the smallest category is playing table soccer which make up 13% of the entire dataset. At first glance, it may seems that this is a problem that we must address, however, this is a real life challenge which reflected in this dataset.

We disregard the remaining recorded activities, such as driving a car, lunch break, and working, because these activities are clasified as complex activities (Sakr, Abu-Elkheir, Atwan, & SOLIMAN, 2018). Complex activities are categorized as sequential, concurrent or interleaved human activities, and analysing them is beyond the scope of this experiment.

Sequential human activities are those occurring in sequence, for example, in case of lunch break, one may first heat the food, then bring the food to the table and start eating. Moreover, concurrent activities are those in which two different activities accruing at the same time, for instance, people may start a conversation at the table or a phone call while eating food. Third category, interleaved human activities are basically one long-term task interrupted by other short-term tasks, for example, consider the working activity in which the subject may leave their desk to access the printer.

Among all the recorded signals, we only consider the wrist-worn 3D-ACC, PPG and chest-worn ECG signals for our thesis. We disregard the chest-worn 3D-ACC data, as we already gather 3D-ACC data from the wrist device, which provides better quality data for a HAR system (Chung et al., 2019; Shoaib et al., 2016). Finally, we also disregard the data related to one of the subjects due to hardware issues during data recording.

In this chapter, we explained the type of the data and the dataset that we used in detailed. Next, we will discuss the initial part of our methodology including, signal pre-processing, windowing and feature extraction and selection.

Chapter 4

Feature Extraction and Selection

In this chapter, we describe the methodology used in our thesis to evaluate the importance of the three different signals in HAR. Figure 4.1 presents an overview of the steps we take to segment the data, extract and select the most relevant features to feed it to ML model.

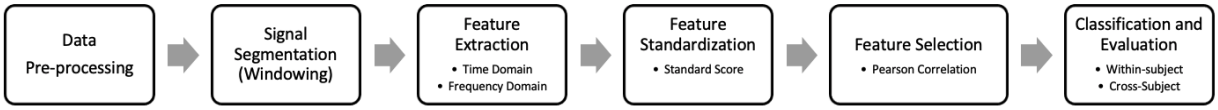


Figure 4.1: Human activity recognition workflow

4.1 Data Pre-processing

As described in Section 3.2, the dataset we use in this thesis contains signals with different sampling rates, as 3D-ACC was captured at 32 Hz while PPG was recorded at 64 Hz. To analyze the 3D-ACC and PPG signals captured from the wrist worn device, we up-sample the 3D-ACC signal from 32 Hz to 64 samples per seconds. We decide not to down-sample the PPG signal, as this would mean eliminating half of the PPG dataset which was not appropriate. Thus, we up-sample the 3D-ACC signal by interpolating two-consecutive data points with their average value. We use the original chest-worn ECG signal, which was recorded at 700 Hz sampling rate.

4.2 Windowing

Before we start extracting features from the data, we segment the entire signal into small sequences of fixed size (same number of data points). This approach is known as sliding window. The main intuition for windowing technique is to retrieve meaningful information from the time series. Each single datapoint in the time-series is not representative of any specific activity, however, a group of consecutive datapoints (a slice in the time series) is capable of providing insightful information about the human activity.

The size of the window is an important parameter in sliding window techniques. Each window must be wide enough to capture enough information for further signal processing and analyzing. However, the window size should not be too large, since larger windows may delay the real-time signal processing and the eventual activity recognition. The reason being that the model has to wait for the entire duration of the window to be able to start recognizing the next activity. Thus, there is a trade-off between capturing the appropriate amount of information and the speed of recognition. There is no standard fixed window size that researchers can utilize, as the appropriate window size depends highly on the characteristics of the signal. For instance, if we have a periodic signal, an adequate window size may be the one that is wide enough to cover at least one period of the signal in each segment.

Researchers have attempted different strategies to select an appropriate window size. One strategy is to use adaptive window sizes in which a feedback system is employed to calculate the likeliness of a signal belonging to an activity, then the desired window size is selected based on the probabilities (Noor, Salcic, Kevin, & Wang, 2017). In most cases, however, researchers opt for fixed window sizes, Banos et al. (Banos, Galvez, Damas, Pomares, & Rojas, 2014), studied the impact of different window sizes on human activity recognition accuracy. They observed that many researchers have applied varying fixed window sizes from 0.1 to 12.8 seconds, although, near 50% of the considered studies have used 0.1 to 3 seconds window sizes.

In this work, we examine different fixed window sizes from 0.5 to 15 seconds on all three sources signals, and we select window size of seven seconds. As depicted in Figure 4.2, larger window sizes, provide only a slight improvement in the performance of the 3D-ACC signal. This

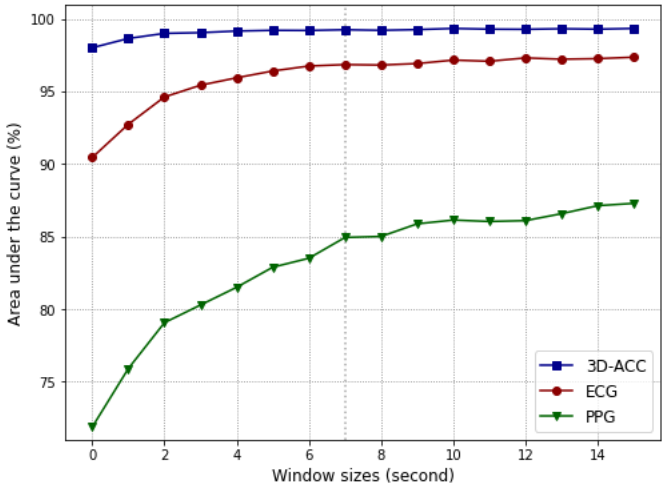


Figure 4.2: Comparison between different window sizes for 3D-ACC, PPG, ECG signals. X-axis: Window sizes represents in seconds. Y-axis: Area under the receiver operating characteristic curve after train and test random forest models.

means that smaller window sizes are still capable of capturing enough information out of the 3D-ACC signal and at the same time preserve a reasonable speed in recognizing an activity. Contrasting with 3D-ACC signals, larger window sizes are more informative for the bio-signals (ECG and PPG). Because having a larger window means capturing more than one period of cardiac activity in one window, thus, the heartbeat rate can also be taken into account.

Based on the purpose of this research, which is comparing the performance of mentioned signals, we must select equal window sizes, in terms of time duration, to have a reasonable comparison. Thus, we need to keep the balance between selecting smaller window sizes for 3D-ACC and larger ones for PPG and ECG. As stated, according to aforementioned explanations and figure, we decided to select a window size of seven seconds to start our next step, feature extraction phase.

Since, each window of the segmented signal is not completely independent and identical from its neighboring windows, we applied non-overlapping sliding windows. Based on the results obtained from Dehghani et al. (Dehghani, Sarbishei, Glatard, & Shihab, 2019), such signals are not independent and identically distributed (i.i.d.), so that overlapping would lead to classification model overfitting.

4.3 Feature extraction

After segmenting the signals in windows of seven seconds, we extract two types of features from each window: hand-crafted time- and frequency-domain features. In the following we provide more detailed information about these two categories of features.

4.3.1 Time-domain features

Time-domain features are the statistical measurements calculated and extracted from each window in time series. As formerly described, we segmented five raw signals 3D-ACC, PPG and ECG with sampling rate of 64, 64 and 700 hertz, respectively. In total, we extract seven statistical features from each of these windows. Table 4.1 presents the type of the features and their respective description. Features that we mention in the following table are easy to understand and are not computationally expensive, moreover, are capable of providing relevant information for HAR systems. Therefore, these features are frequently used in the filed of human activity recognition (Attal et al., 2015; Koskimaki, Siirtola, et al., 2016; Shoaib et al., 2016).

Table 4.1: Hand-crafted time domain features and descriptions. Each of these features calculated over data points within each window.

Hand-crafted time-domain feature	**Description**
mean	average value of the data-points
min	smallest value
max	largest value
median	the value at the 50% percentile
standard deviation	measures how scatter are the data-points from the average value
zero-crossing rate	counts the number of times that the time-series crosses the line $y = 0$
mean-crossing rate	counts the number of times that the time-series crosses the line $y = mean$

4.3.2 Frequency-domain features

Transferring time domain signals to the frequency domain provides insights from a new perspective of the signal. This approach is widely used in signal processing researches as well as human activity recognition field (Attal et al., 2015; Hassan, Uddin, Mohamed, & Almogren, 2018; Koskimaki et al., 2016).

In the first step to extract frequency domain features, we segment the raw time-domain signals into fixed window sizes. Then, we transfer each segmented signal into the frequency domain using Fast Fourier Transform (FFT) method (Cochran et al., 1967). It is important to perform these two steps in the aforementioned order, otherwise, each window would not contain all the frequency information. That is, low frequency information would appear in the early windows and then the high frequency components would be placed in the last windows. By contrast, the correct way is that each window must have all the frequency components. After obtaining frequency components from each window, we extract eight statistical and frequency-related features. Table 4.2 presents different extracted features and a brief description for each of them.

Table 4.2: Hand-crafted frequency-domain features and descriptions. Each of these features calculated over frequency component within each window.

Hand-crafted frequency-domain feature	**Description**
mean	average value of the data-points inside one window
min	smallest value
max	largest value (only for bio-signals, PPG and ECG)
second-max	second largest value (only for 3D-ACC signal)
DC component	zero frequency component
standard deviation	measures how scatter are the data-points from the average
median	is the middle value after sorting data-points from smallest value to the largest one
dominant frequency	is the frequency correspond to maximum energy (amplitude)
mean-crossing	counts the number of times that the time-series crosses the line $y = mean$

From the frequency-domain features presented in Table 4.2, the DC component and dominant frequency are the less intuitive ones. Thus, next we explain these two features in more detail. The "DC component" is the frequency-domain amplitude value which occurs at zero frequency. In other words, DC component is the average value of signal in time-domain over one period. Regarding 3D-ACC signal, its DC component corresponds to gravitational accelerations (He & Jin, 2009). To be more precise, in the absence of device acceleration, the 3D-ACC output is equivalent to device rotation along axes (Pedley, 2013). This explains the reason why its DC component value is relatively larger than the rest of the frequency coefficients in the same window (Figo, Diniz, Ferreira, & Cardoso, 2010). As for bio-signals, however, DC component is not highly greater than other frequency coefficients for which the reason is that bio-signals such as PPG and ECG are dynamic signals. Figure 4.3 represents PPG and X-axis of accelerometer signals related to a specific

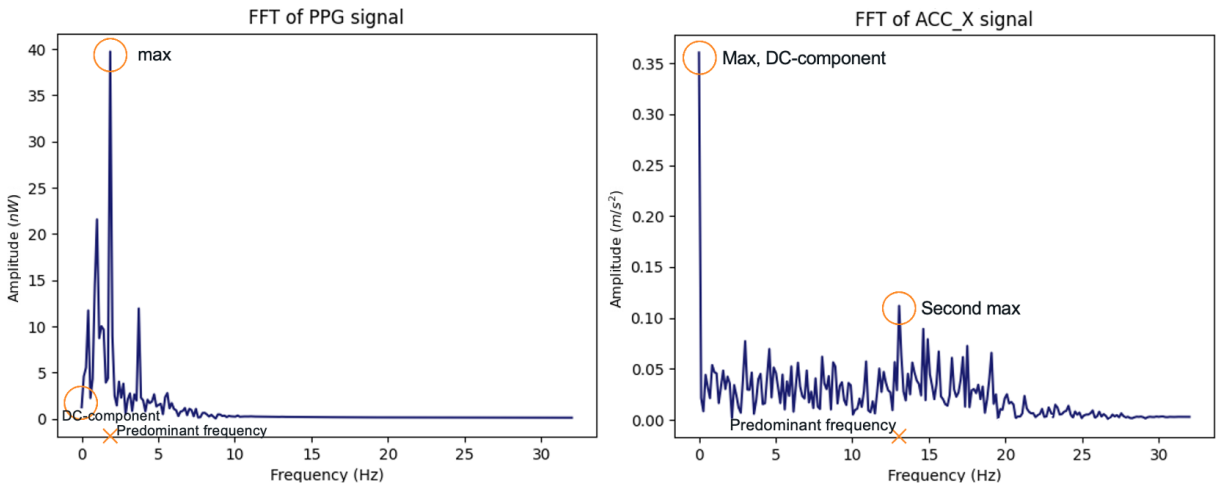


Figure 4.3: A seven seconds window of PPG and X-axis of accelerometer signals transformed to frequency domain related to the "cycling" activity. Notice the difference between DC components, moreover, maximum amplitude and its correspond frequency (predominant frequency) for different sources of signals.

time span of seven seconds of cycling activity in frequency domain. The difference between DC component values between these two signals is clear.

Regarding 3D-ACC signal, since DC component is also the maximum amplitude, we decide to introduce another feature, namely, "second-max" to avoid feature redundancy and differentiate between mentioned features. However, for bio-signals, we only consider the maximum amplitude.

Dominant frequency is the frequency at which the highest amplitude occurs (Telgarsky, 2013). Based on this definition, again for 3D-ACC signal, we disregard the amplitude corresponding to the zero frequency (DC component); instead, we consider the frequency correspond to the second largest amplitude value (second max) as the "dominant frequency". However, for the bio-signals, namely ECG and PPG signals, "dominant frequency" is based on the first maximum amplitude value.

4.4 Feature standardization

Since we are examining different sources of signals with different characteristics, feature values will inevitably have different ranges, thus, we need to standardize the features before the model

classification. For instance, the scale difference between X-axis of the ACC and the PPG signals have clearly distinct standard deviation of 0.18 and 64.2, respectively. This shows a huge difference in the amplitude of these signals which must be addressed. To standardize the extracted features, we calculate the standard score for each feature using formula 1, this approach used in some previous HAR studies Park et al. (2017); Ronao and Cho (2016).

$$z_i = \frac{f_i - \mu}{\sigma}, \qquad 0 \leq i \ < \ m \tag{1}$$

where for a given feature vector of size m, f_i represents the i_{th} element in the feature vector, μ and σ are the mean and standard deviation for the same vector, respectively. The resulted value, z_i, is the scaled version of the original feature value, f_i. Using this method, we reinforce each feature vector to have zero mean and unit variance. However, the mentioned transformation retains the original distribution of the feature vector. Note that we split the dataset into train and test set before the standardization step. It is necessary to standardize the train set and the test set separately; because we do not want the test set data to influence the μ and σ of the training set, which would create an undesired dependency between the sets Buitinck et al. (2013).

4.5 Feature Selection

In total, we extract 77 features out of all sources of signals. Following the scaling phase, we remove the features which were not sufficiently informative. Omitting redundant features helps reducing the feature table dimensionality, hence, decreasing the computational complexity and training time. To perform feature selection, we apply the Correlation based Feature Selection (CFS) method and calculate the pairwise Spearman rank correlation coefficient (Saeys, Abeel, & Van de Peer, 2008) for all features. Correlation coefficient has a value in $[-1, +1]$ interval, for which zero indicates having no correlation, $+1$ or -1 refer to a situation in which two features are strongly correlated in a direct and inverse manner, respectively. In this thesis, we set the correlation coefficient threshold to 0.85, moreover, among two recognized correlated features, we omit the one which was less correlated to the target vector. Finally, we select 45 features from all signals. It is worth mentioning that the DC component feature related to 3D-ACC signal was among the dropped features,

however, for bio-signals this feature was taken into count.

In the current chapter, we described our methodology up to and including creating desired feature tables to be able to proceed to the next steps. In the following chapter, we will explain how we use the feature tables to feed and run the classification models considering two different setups, namely, subject-specific and cross-subject models.

Chapter 5

Classifiers and Experiment Setup

In this chapter, we present our designed early fusion scenarios, then we discuss the classifier that we use and its related parameters. Finally, we explain the models evaluation setups, namely, subject-specific and cross-subject models.

5.1 Proposed scenarios

To evaluate the level of contribution for each of the 3D-ACC, ECG and PPG signals, we take advantage of the early fusion technique and introduce seven scenarios presented in Table 5.1. Subsequently, we feed the classifier with feature matrices constructed based on each of these scenarios. We use the Python Scikit-learn library for our implementation (Pedregosa et al., 2011).

Table 5.1: Different proposed scenarios to evaluate the level of contribution for each of the 3D-ACC, ECG and PPG signals and their combination, in addition to the total number of applied features (time and frequency domains) after feature selection.

Scenario ID	Considered signals	Total number of features
1	3D-ACC	24
2	ECG	12
3	PPG	9
4	3D-ACC + ECG	36
5	3D-ACC + PPG	33
6	ECG + PPG	21
7	3D-ACC + ECG + PPG	45

5.2 Classifier

In our work, we examine three different machine learning models, namely, multinomial logistic regression, k-nearest neighbors, and random forest. Based on our initial observations, the random forest classifier outperformed the other models in recognizing different activities. Thus, we conduct the rest of our experiment using only the random forest classifier.

Random forest is an ensemble model consisting of a set of decision trees each of which votes for specific class, which in this case is the activity-ID, then the random forest decides on the final label of an instance (Ho, 1995). In this thesis, we set the total number of trees to 300, and to prevent the classifier from being overfitted, we assign a maximum depth of each of these trees to 25. One advantage about using random forest as a classifier is that this model provides extra information about feature importance, which is useful in recognizing the most essential features.

5.3 Performance Evaluation

In our thesis, we evaluate two types of models, the subject-specific model and the cross-subject model. In the following we provide detailed explanation about these two models and evaluation techniques.

5.3.1 Subject-specific model

Subject-specific models are the most accurate types of models, as they train and test using the data belonging to same user. Hence, it is important we evaluate if bio-signals can be useful to make such models even better.

To evaluate the performance of our subject-specific model, we employ a k-fold cross validation technique (Kohavi et al., 1995). K-fold cross validation is a widely-used method for performance evaluation and consists in randomly segmenting the dataset into k parts (folds). The machine learning model is trained on $k - 1$ partitions and is tested on the remaining partition; this procedure repeats k times, always testing the model on a different fold. For each of the k runs, the evaluation procedure is done based on the scoring parameter. Finally, the average value of obtained scores is reported as the overall performance of the classifier.

As stated in Section 3.2, we have an imbalanced dataset, therefore, it is important to specify how to split the dataset into folds. We use the stratified k-fold method to preserve the proportion of each class label in each fold to be similar to the proportion of each class label in the entire set. Regarding scoring parameters, we evaluate our models with two metrics, namely, F1-score and area under the receiver operating characteristic (ROC) curve (Elamvazuthi, Izhar, Capi, et al., 2018; Fawcett, 2006). Since our study is a multi-class classification problem, we aggregate mentioned scores using average weighted by support.

In our case, to evaluate the subject-specific model, we consider one feature set related to only one subject and split it into a train set (80%) and a test set (20%). Subsequently, we apply the $10 - fold$ CV technique on the training set and store the resulting F1-score and AUC measurements per fold. Finally, we apply the trained model on the test set, then we record its classification performance in terms of F1-Score and AUC. Our objective of evaluating the model's performance on the train set, and then on test set, was to confirm that the model is not overfitting the data. An overfitted model fits perfectly on the train set, but has poor performance on the test set (Domingos, 2012).

We repeat described procedure 14 times, as many as the number of subjects. Eventually, we calculate the average F1-Score and AUC, over all subjects' results and will report its performance in Chapter 6.

Within-subject evaluation approach is a subject-dependent technique, since we train the model on features related to one subject and then test the model using the remaining features belonging to the same subject; also known as "personal model" in the study of Weiss et al. (Weiss, Timko, Gallagher, Yoneda, & Schreiber, 2016).

5.3.2 Cross-subject model

Cross-subject models are not as accurate as the subject-specific models, however, since such models are cheaper, in practice these are more commonly used. Cross-subject models are cheaper because these do not require the user's personal data, instead, need data from other individuals. Therefore, knowing that bio-signals can contribute to this type of model is important to improve the generalization of the model.

To evaluate the performance of our cross-subject model we use the Leave-One-Subject-Out

(LOSO) evaluation method (Hastie, Tibshirani, & Friedman, 2009; Saeb, Lonini, Jayaraman, Mohr, & Kording, 2017). This method consists in training models on a group of subjects and testing the model on an unseen individual data (test set). Similar to the k-fold, in a group with n subjects, the model is trained in $n - 1$ subjects and tested on the remaining subject's data. This process repeats n times, to cover all subjects in the test set, always using the other subjects' data for training.

In our case, to evaluate the cross-specific model, we create a larger feature table consisting of 13 feature sets related to all subjects but one. We use this table as an input matrix to train the random forest model. Then we test our model by the feature table of the remaining subject. Again, we aggregate F1-score and AUC measurement using average weighted by support.

We rerun this process fourteen times. Then, we calculate the average F1-Score and AUC, over all subjects' results and will report its performance in Chapter 6.

The aforementioned evaluation method, minimizes the risk of overfitting, moreover, it is subject independent; also called as "impersonal model" in the study of Wiess et al. (Weiss et al., 2016). Hence, if a classification model performs well given the LOSO evaluation method, then this model is generalizable to other subjects.

In this chapter, we had a detailed explanation about our fusion scenarios, as well as the applied classification algorithm with two different setups. In the next chapter, we will provide the resulting information with respect to the scenarios and setups and present related figures.

Chapter 6

Early fusion methods results

In this chapter, we report the results obtained from aforementioned evaluation strategies. We answer our initial questions about how informative are each of the sources of signals and whether sensor fusion improves the performance of a HAR classifier. We organize the result section based on the scenarios presented in Table 5.1, that is, using only one source of signal (scenario 1, 2 and 3), considering a combination of two signals (scenario 4, 5, and 6), and scenario 7 which is related to 3D-ACC, ECG and PPG signals fusion. We conclude each subsection by reporting our observations regarding per activity performance of the HAR models.

6.1 RQ1: What is the contribution level of signals under study in subject-specific HAR systems?

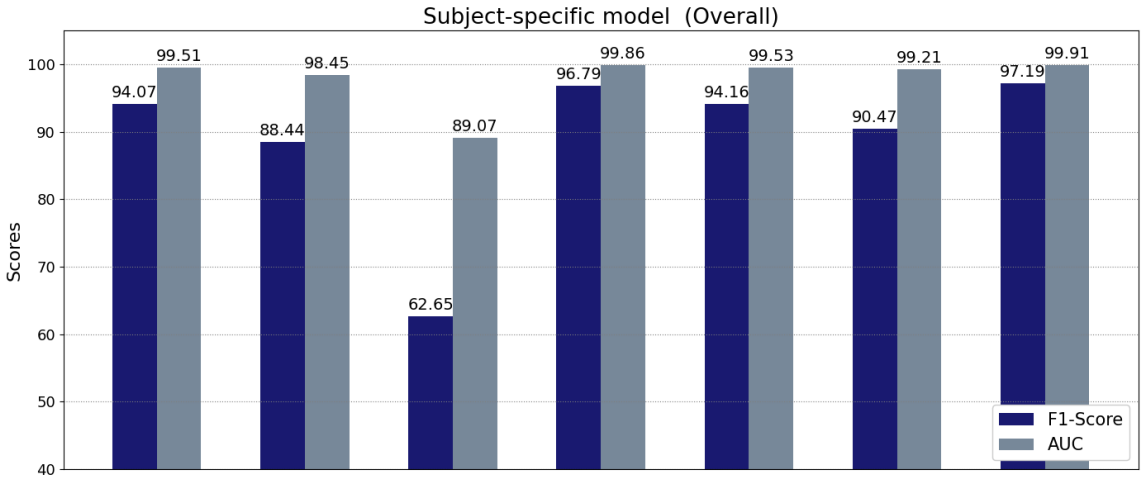
We explained in Section 5.3.1 that we train and test subject-specific models and we are interested in the contribution level of each source of signals in these models. We report the results of the subject-specific model evaluation in Figure 6.1.

One type of signal. Considering only one source of signal, 3D-ACC signal (Scenario 1) outperforms the other two bio-signals in recognizing human activities (Scenarios 1 - 3). Interestingly, a model using exclusively the ECG signal (Scenario 2) performs relatively satisfactory, with a comparable AUC performance to a 3D-ACC trained model, yielding a much better performance than the model trained solely with PPG signal (Scenario 3).

Two signals fusion. When combining the 3D-ACC with the ECG signal (Scenario 4), the performance of the model surpasses the model using only 3D-ACC (Scenario 1). As shown in Figure 6.1, adding ECG to 3D-ACC improves the human activity recognition by 2.72% in terms of F1-Score. Our results suggest that including the ECG signal in HAR systems that are based solely on the 3D-ACC can slightly improve the model's performance. However, adding PPG signal to the 3D-ACC signal (Scenario 5) does not provide any significant enhancement for our HAR models. Combining solely bio-signals (PPG and ECG in Scenario 6) does not yield a model with superior performance than just a 3D-ACC model, even if they outperform models trained with one bio-signal (Scenarios 2 and 3).

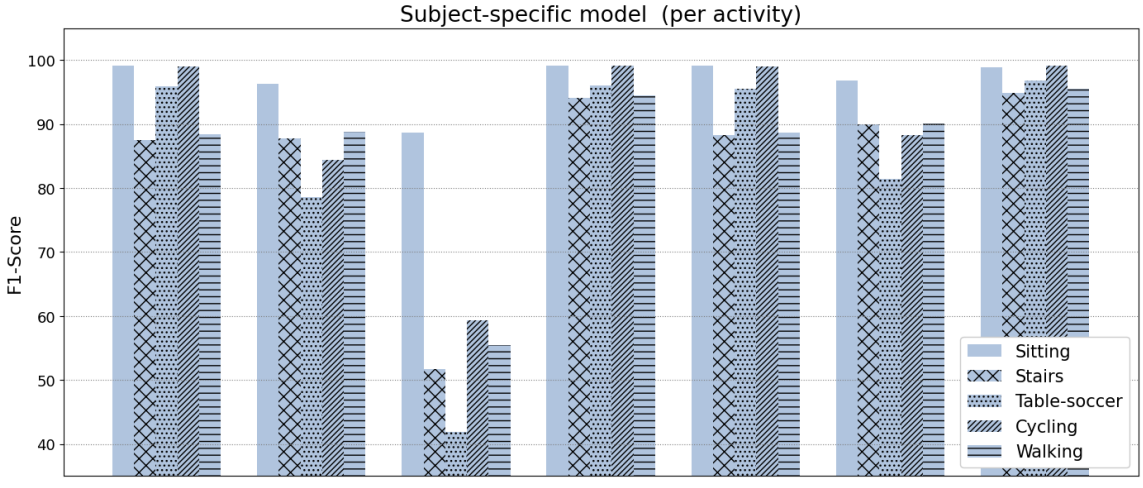
Three signals fusion. Regarding Scenario 7, when we consider all three sources of signals, we realize that human activity recognition performance remain almost the same compared to Scenario 4 when we only took 3D-ACC and ECG signals into account. Therefore, we conclude that PPG signal fusion did not add any strength to the classifiers in our evaluation. In addition, it is obvious from Figure 6.1 that the PPG signal is not very informative, not exclusively nor in combination with other sources of signals for subject-dependent HAR systems.

Per activity performance. Figure 6.2 represents results of the subject-specific model per activity. It is noticeable that "ascending/descending stairs" and "walking" are the two activities that our models have difficulty distinguishing when using only the 3D-ACC signal. However, feeding the model with features extracted from both 3D-ACC and ECG signals (Scenario 4), improves "stairs" and "walking" distinction significantly by 6.54% and 6.05% F1-score, respectively. An important takeaway from Figure 6.2 is that bio-signals have reliable power in distinguishing stationary activities such as "sitting" from non-stationary ones such as "walking" and "cycling". When comparing the contribution of the PPG signal to the 3D-ACC per activity, we note that the combination did not yield any improvement in distinguishing mentioned activities. In fact, model's performance when combining 3D-ACC and PPG is highly similar to its performance when we apply only 3D-ACC signal (as in Scenario 1), which indicates an inadequacy of the PPG signal features in distinguishing any information not already captured by the 3D-ACC.



Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 6.1: Subject-specific random forest model results



Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 6.2: Subject-specific random forest model results per activity

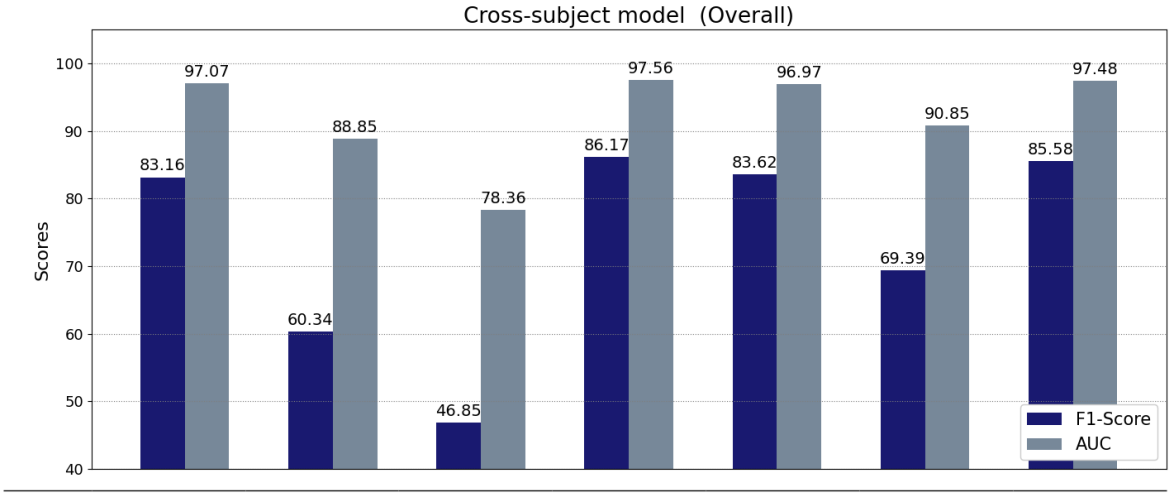
6.2 RQ2: What is the contribution level of signals under study in cross-subject HAR systems?

In Section 5.3.2, we mentioned the cross-subject models and the fact that these models tend to perform worse than subject-specific models, given that cross-subject are more general. Therefore, for the current evaluation setup, we observe a more significant contribution from bio-signals. Figure 6.3 shows the overall performance of signals under study in terms of F1-score and AUC measurements (aggregated as stated above). As expected, cross-subject models' performance have an overall lower performance than the subject-specific models. This decrease in performance is expected as cross-subject models are trained on other individual's data, and individuals perform activities differently. Moreover, other factors such as subject's height, weights, gender and level of fitness may contribute in this variation.

One type of signal. According to the Figure 6.3, 3D-ACC signal provides the most informative data to our cross-subject models, yielding a performance with a F1-score of 83.16 (Scenario 1). Contrasting with the results observed for subject-specific models, a model trained only with ECG (Scenario 2) did not yield comparable performance with a 3D-ACC model (Scenario 1). Still, cross-subject models trained using only ECG signal (Scenario 2) outperforms the models trained exclusively with the PPG signal (Scenario 3), by 13.49% in terms of F1-score.

Two signals fusion. The combination of 3D-ACC and ECG signals (Scenario 4) has shown to improve the performance of our cross-subject model by 3% (F1-score), compared to using exclusively the 3D-ACC signal. Once again, a fusion of PPG and 3D-ACC signals has shown to not yield performance improvements (Scenario 5) to trained 3D-ACC models. Interestingly, combining both ECG and PPG (Scenario 6) yields a model with better performance than the models trained exclusively with ECG (Scenario 2) and PPG (Scenario 3), even if it still underperforms against pure 3D-ACC trained model (Scenario 1). In the end, we conclude that ECG signal can be complement well 3D-ACC signal in HAR systems, while PPG did not provide informative data to our cross-subject models.

Three signals fusion. Once we fuse all three sources of signals, we observe a decrease in the model's performance compared to just combining ACC-3D and ECG signals (Scenario 4). This

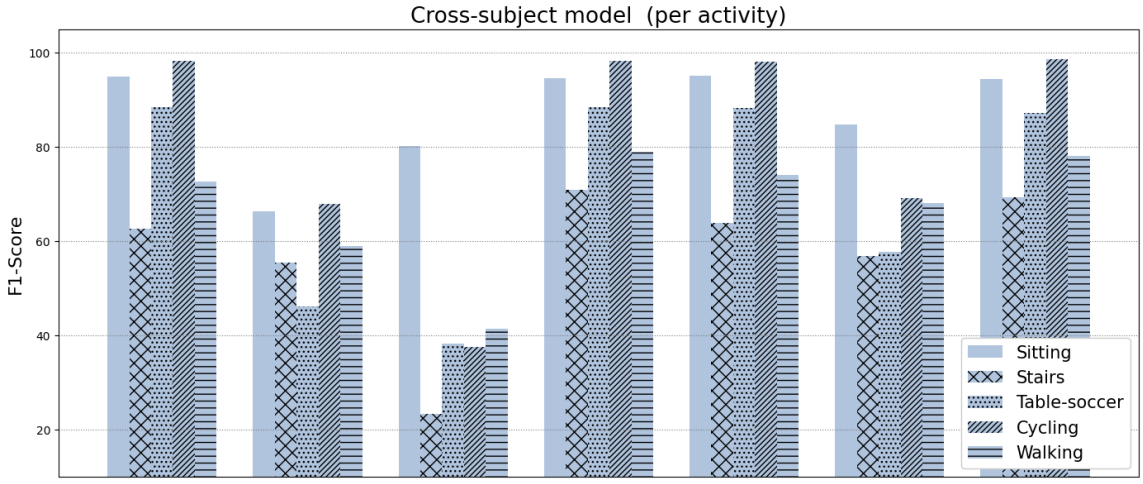


Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 6.3: Cross-subject random forest model results

further corroborates with the notion that adding PPG signal to the combination of 3D-ACC and ECG signals is disadvantageous.

Per activity performance. Figure 6.4 shows the performance of the models broken down per activity. First, we note that "cycling", "table-soccer" and "sitting" activities remain rather stable in all models trained with 3D-ACC, exclusively or in combination. Second, cross-subject models often miss-classifies "stairs" and "walk" activities. Once we fuse both 3D-ACC and ECG signal, the model is better able to distinguish between the two activities, which explains the gain in the overall performance of the model. Third, models trained exclusively with bio-signals have very distinct performance profiles per activity. Note that PPG is reasonably good at distinguishing the "sitting" activity, as this is the least physically demanding activity in our dataset (lower heart rate). ECG models, on the other hand, outperforms PPG models in all other activities, further corroborating that ECG signal is, on average, more informative for cross-subject HAR models than the PPG signal.



Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 6.4: Cross-subject random forest model results per activity

The present chapter was allocated to the resulting insight about the contribution level of the bio-signals in a HAR system. In the subsequent chapter, we will dive into the detailed impact of ECG signal on HAR systems performance in terms of per activity impact and try to answer “Why ECG is not always helpful”. Moreover, we will discuss about late fusion approach and the resulting information. Last but not least, we present the top 20 important features.

Chapter 7

Discussion

In this chapter we elaborate more on the impact of ECG signal (scenario 4) on models' per activity performance. We check confusion matrices related to all subjects in both subject-specific and cross-subject models, and explore the subjects characteristics to understand why ECG signal provide more performance improvement for some subjects compared to the others. Additionally, we investigate the impact of the fusion method on the performance of a HAR system. We apply a class-based weighting late fusion approach to fuse the 3D-ACC, ECG and PPG signals and compare its performance with the early fusion method. At the end of this chapter we review the most important features in the best model which is the combination of 3D-ACC and ECG signal.

7.1 Detailed impact of ECG signal addition

In this section we explore the detailed per activity impact of adding ECG signal to the 3D-ACC signal and investigate the impact of subjects' fitness level on the contribution of the ECG signal.

7.1.1 ECG signal on Subject-specific model

As our results suggested in Chapter 6, considering only 3D-ACC signals, models already reach a high recognition performance of 94.07% F1-score. Thus, most of the instances in confusion matrices are labeled accurately. However, the activities of using stair and walking were frequently confused with each other. Therefore, we contrast the confusion matrices of the model which includes

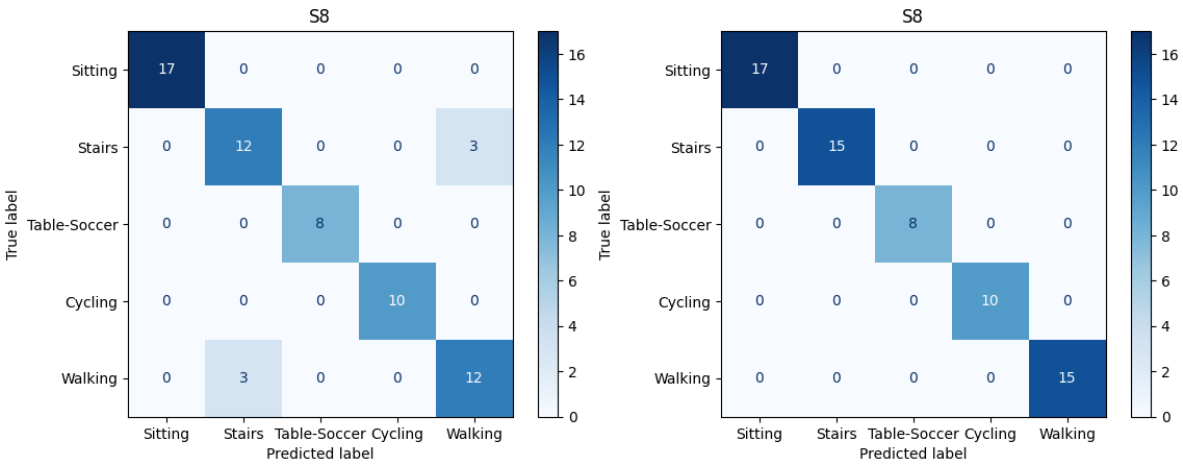


Figure 7.1: Comparison between confusion matrices in subject-specific models. On the left: the model performance when considering only 3D-accelerometer. On the right: the model performance when combining 3D-accelerometer with ECG signal.

only 3D-ACC (Scenario 1) against the models including both 3D-ACC and ECG signals (Scenario 4). We observe significant improvement in distinguishing mentioned activities after adding ECG signal. Figure 7.1 presents confusion matrices related to subject number 8 in subject-specific model. On the left side of Figure 7.1, we can observe the model performance when considering only 3D-ACC. Note the instances which are miss-classified and confused between "Stairs" and "Walking" activities. On the right side of Figure 7.1, however, it is clear that after adding ECG signal, mentioned confusions is solved.

It is important to mention that for 10 out of 14 subjects we observe "Stairs - Walking" improvement after adding ECG signal to 3D-ACC, however, in 3 out of 14 cases adding ECG signal does not improve the "Stairs - Walking" classification. Moreover, in 1 case, the model perfectly distinguishes between "Stairs - Walking" by just using the 3D-ACC, leaving no space for improvement for the 3D-ACC and ECG fusion model.

7.1.2 ECG signal on Cross-subject model

Analyzing the ECG signal contribution in cross-subject models yielded a more insightful analysis, since these models miss-classify activities more often, compared to subject-specific models. As depicted in Figure 6.3, using only the 3D-ACC signal we obtained F1-score of 83.16% which is

relatively lower than model's performance in subject-specific setup. After a detailed investigation in confusion matrices of the 3D-ACC trained model, we once again identify that the activities"stairs" and "walking" are miss-labeled. In addition to the mentioned pair of activities, another pair is miss classified in cross-subject models, namely, "sitting" and "playing table soccer".

We once again compare the confusion matrices related to the models trained with 3D-ACC (Scenario 1) signal versus the model trained with both 3D-ACC and ECG signals (Scenario 4). We observe that ECG signal significantly helps the model recognize "Stairs - Walking", however, it does not add any value when it comes to distinguishing "Sitting - Table-Soccer" pair. Figure 7.2 depicts both confusion matrices related to subject number 7 in cross-subject model. The left side of Figure 7.2 is related to the model performance when considering only 3D-ACC, note the huge portion of "Walking" instances which are miss-classified as "Stairs". However, on the right side of Figure 7.2, it is obvious that after adding ECG signal "Stairs - Walking" detection enhances noticeably.

It is worth noting that for 9 out of 14 subjects, we observe "Stairs - Walking" improvement after adding ECG signal to a pure 3D-ACC model. In 3 out of 14 cases, adding ECG signal yield no significant impact; and, in 2 out of 14 cases ECG signal addition resulted in a decline in the "Stairs - Walking" classification.

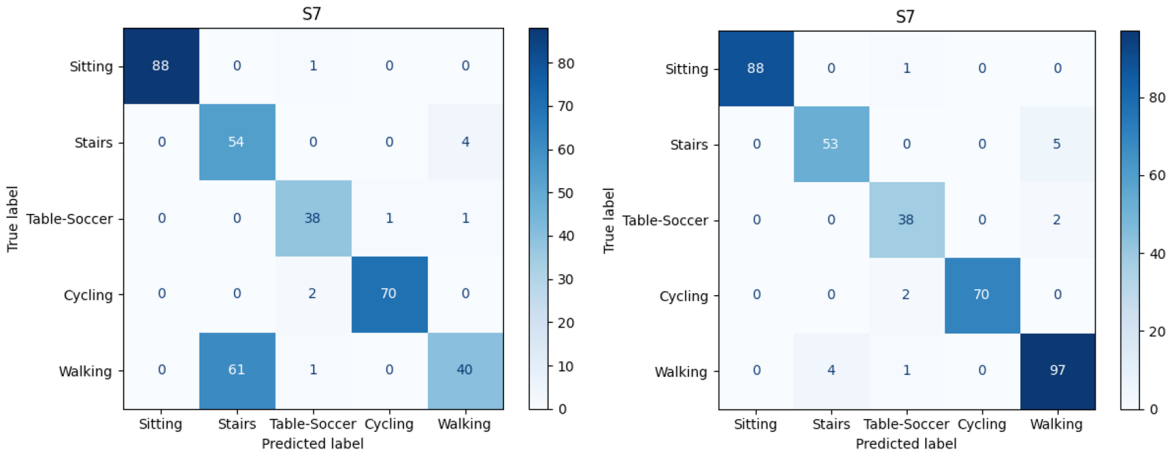


Figure 7.2: Comparison between confusion matrices in cross-subject models. On the left: the model performance when considering only 3D-accelerometer. On the right: the model performance when combining 3D-accelerometer with ECG signal.

7.1.3 Why ECG is not always helpful?

In the past two sections we stated that the ECG and 3D-ACC signal fusion does not always yield a performance improvement in the cross-subject HAR system; more precisely, we observed decline in the model's performance in 2 out of 14 cases. Since the used dataset (Section 3.2) provided the information related to the subjects characteristics (e.g., height, weight, etc.), we investigate these factors on the cross-subject models for the scenario number 1 and 4 to answer the question "Why ECG is not always helpful?". Table 7.1 shows the detailed information about the subjects. The last three columns in the table present per subject resulting F1-scores obtained from models trained with 3D-ACC exclusively (scenario 1), the models' performance after ECG fusion (scenario 4), and the last column provides the F1-score improvement after ECG addition. We sort the table based on the performance improvement.

First looking at the subjects' characteristics and the exclusive 3D-ACC models performance. In the cross-subject models, where we train on a group of people and test the model on an unseen subject, we observe that the more the unseen subject's physical characteristics matches or is close to the average of the people in the train set, the better the performance that the 3D-ACC will yield. As an example, the best performance for scenario 1, is when we test the model on subject number 14, whose physical characteristics factors are the closest to median measurement of the factors related to all subjects; the test F1-score for this model is equal to 91.37%. However, a potential reason for a 3D-ACC based model not performing as great on the unseen subject, could be that the subject is an outlier in terms of the physical characteristics presented in the Table 7.1, such as in the case of subjects 7 and 10 who are the outliers in fitness level and age, respectively; also, for subjects 2 and 12 who are outliers in both height and weight, for whom, we observe the least 3D-ACC performance.

Elaborating on the contribution level of the ECG signal after fusing it with 3D-ACC signal, we observe the greatest impact of this bio-signal on subject 7 who has the lowest fitness level. Based on a study by (Bjørnstad, Storstein, Meen, & Hals, 1993), increasing in fitness level results in lower heart rate. Thus, for people who are less fit, their heart reacts more significantly to physical activities. Moreover, we observe that for subject 12 who is the tallest person among the subjects,

ECG signal adds value to the model's performance. On the other hand, we observe that, for subject 14, where the exclusive 3D-ACC had the best performance, the signal fusion caused a decline in model's performance.

Table 7.1: Detailed information about subjects characteristics and resulted F1-score for scenario 1 and 4. Last column is the models improvement/ decline after fusing the ECG signal with 3D-ACC.

Subject ID	**Gender**	**Age**	**Height**	**Weight**	**Fitness level**	**ACC**	**ACC + ECG**	**Improvement**
7	F	21	168	58	2	78.36	95.59	17.22
12	M	43	195	105	5	75.88	86.06	10.17
4	M	25	168	57	5	84.33	89.95	5.62
3	M	25	170	60	5	85.70	91.10	5.40
1	M	34	182	78	6	86.69	90.87	4.17
13	F	21	170	63	6	85.29	89.23	3.93
5	F	21	180	70	4	78.81	82.62	3.81
8	M	43	179	70	5	82.63	85.19	2.55
2	M	28	189	80	5	70.27	72.36	2.08
11	F	24	168	62	5	88.69	90.45	1.75
15	M	28	183	79	5	86.72	88.29	1.57
9	F	28	167	60	5	86.92	88.40	1.47
14	F	26	170	67	4	91.37	84.28	-7.08
10	F	55	164	56	5	82.57	71.86	-10.70
Median	-	27	170	65	5	84.81	88.35	3.18

We proceed with studying the characteristics of the subjects by adding this information to the feature matrix in addition to the time and frequency features, to see its effect on the model's performance. To that end, we add six more columns to the feature table, namely, height, weight, age, gender, and fitness level as well as height-weight ratio which we calculate. We train and test the cross-subject models using the new feature table for Scenario 1 in which we consider subjects characteristics in addition to 3D-ACC features and we observe 1.53% F1-score improvement compared to using only 3D-ACC features. Regarding Scenario 4, in which we use features extracted from 3D-ACC and ECG signals, we observe 0.15% performance decline after we add the subjects' characteristics as features. It is worth noting that in both cases, all the subjects' characteristic features were considered as less important features in comparison to the 3D-ACC and ECG related features. However, out of these characteristic features, height and weight were the most informative ones, respectively. Overall, we summarize this section by noting that the subjects' fitness level has a

significant impact of the ECG signal performance after we fuse this signal with the 3D-ACC signal. However, since we only have one subject with low fitness level, more experiment is required. Moreover, in some cases ECG signal fusion is capable of mitigating the impact of outlier height and weight.

7.2 Class-based weighting late fusion

In Chapter 2, in addition to the early fusion approach, we mentioned the late fusion methods in which fusion happens at the decision level. More precisely, in class-based weighting late fusion approach, during the train phase, multiple models are trained and weights are calculated for each model and each class, then, in the test phase, these weights are applied on the corresponding models' output. Aggregating the weights and models' outputs result in late fusion. As weighting technique, Chowdhury et al. suggested F1-score calculation for each class (Chowdhury et al., 2017). Later, Tsanousa et al. proposed a class-based weighting late fusion algorithm which uses the following equations, referred to as detection rate (DR) (Tsanousa et al., 2019).

$$DR = \frac{TP}{TP + FP + TN + FN} \tag{2}$$

Where DR is basically the ratio of true positive (TP) instances, meaning the instances which are correctly labeled to the total number of instances. Now to calculate the weight vector based on the DR, Tsanousa et al. proposed using Equation 3, by which we can give more weights to the classes which are harder to classify (Tsanousa et al., 2019).

$$w_{ij} = 1 - DR_{ij} \tag{3}$$

Where i refer to the ith model and j refers to the jth class. After weight calculation, we proceed to the test phase in which we apply the trained models on the test instances. That is, for each sources of signals, we use its specific model and get the outputs as probability matrix. The dimensionality of this matrix is equivalent to total number of test instances (observations) multiply by total number of the existing classes (activity labels). As the final step, we need to apply the weight vector on the

probability matrix using the following equation.

$$PM \;=\; aw_i \;+\; (1-a)\, P_i \tag{4}$$

$$final\ label \;= arg\ max(PM) \tag{5}$$

Where PM is the final probability matrix calculated for fusing all three models. Equation 4 shows the approach to aggregate the weights with the probably matrix by calculating the matrix entry wise multiplication and then summation. The adaptation parameter a is a adjustment parameter to perform the summation. Based the Tsanousa et al. study, a is a value between 0 and 1, we set a to be a vector representing the population of each class in percentage, with this approach we can adjust the weights in case of having imbalanced dataset. Moreover, in Equation 5 we apply the $arg\ max$ function on the final probability matrix to get the final labels. For each row of the final probability matrix, $arg\ max$ function returns the index of the maximum probability which represents an activity label. As for the F1-score based weighting method, we calculate per class F1-score during 10-fold cross validation on train instances, thus, $w_{ij} = F1 - score_{ij}$ (related to ith model and jth class). Then, we follow the same steps, that is using Equations 4 and 5.

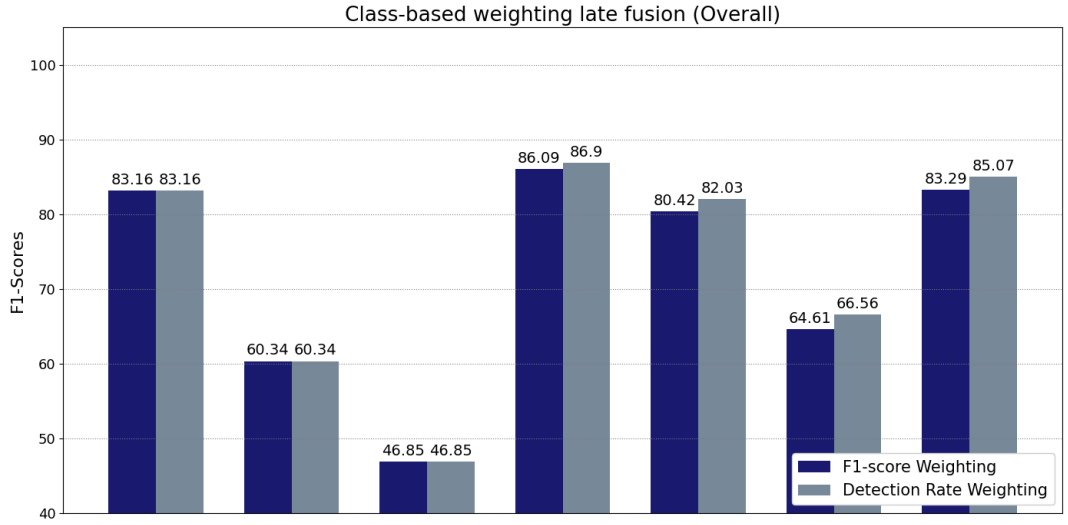
In the current thesis, we only perform and evaluate the late fusion method for the cross-subject (LOSO) setup, since cross-subject models are more complicated. To calculate the weights, we train random forest models for each of the three signals under study. For each subject, we train and evaluate the models using the 10-fold cross-validation approach (only on train instances) in order to calculate an average weight for each class. Then, we use Equation 6 which is derived from Equation 3 to fuse models for different scenarios. Equation 6, is related to scenario 7 where we fuse all three sources of signals. We apply and compare both F1-score and DR-based weighting approaches to perform the late fusion on 3D-ACC, ECG and PPG signals. Table 7.2 summarizes the results we obtained for the late fusion method with both of the weighting approaches and compares its results with the early fusion and the basic models with no fusion.

$$PM \;=\; (aW_{ACC} + (1-a)\, P_{ACC}) + (aW_{ECG} + (1-a)\, P_{ECG}) + (aW_{PPG} + (1-a)\, P_{PPG}) \tag{6}$$

Table 7.2: Resulted F1-score for LOSO set up performing early and late fusion method as well as the base lines with no fusion

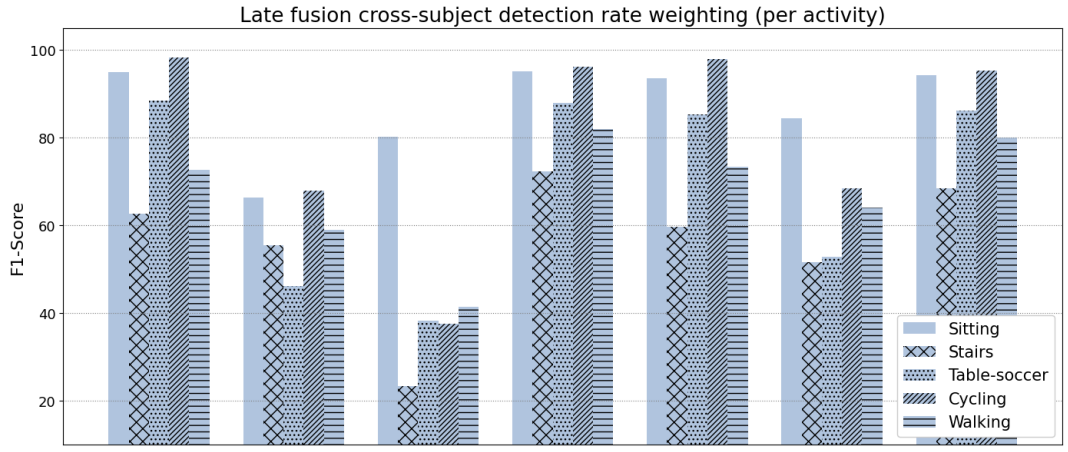
LOSO F1-score	**ACC**	**ECG**	**PPG**	**ACC + ECG**	**ACC + PPG**	**ECG + PPG**	**All**
No fusion (baseline)	83.16	60.34	46.85	-	-	-	-
Early fusion	-	-	-	86.17	83.62	69.39	85.58
Late fusion (f1-score)	-	-	-	86.09	80.42	64.61	83.29
Late fusion (detection rate)	-	-	-	86.90	82.03	66.56	85.07

As the results in Table 7.2 indicate both early and late fusion approaches yield almost similar results, meaning that the combination of the 3D-ACC and ECG signals (scenario 4) is providing the best performance. It is worth mentioning that the late fusion approach does not necessarily perform better than the early fusion. In the Table 7.2, we observe that for 3D-ACC and ECG signals combination, DR-based late fusion performs slightly better than the early fusion, however, for the rest of the fusion scenarios, early fusion performs better. We provide the resulting figures for per activity and overall models' performance in late fusion approach. Figure 7.3 shows a comparison between overall performance of the models weighted DR-based and F1-score based approaches. We can observe that the DR-based weighting method performed slightly better than the F1-score based weighting method. Moreover, Figures 7.4 and 7.5 presents the per activity performance of the detection rate and F1-score weighting late fusion approaches, respectively.



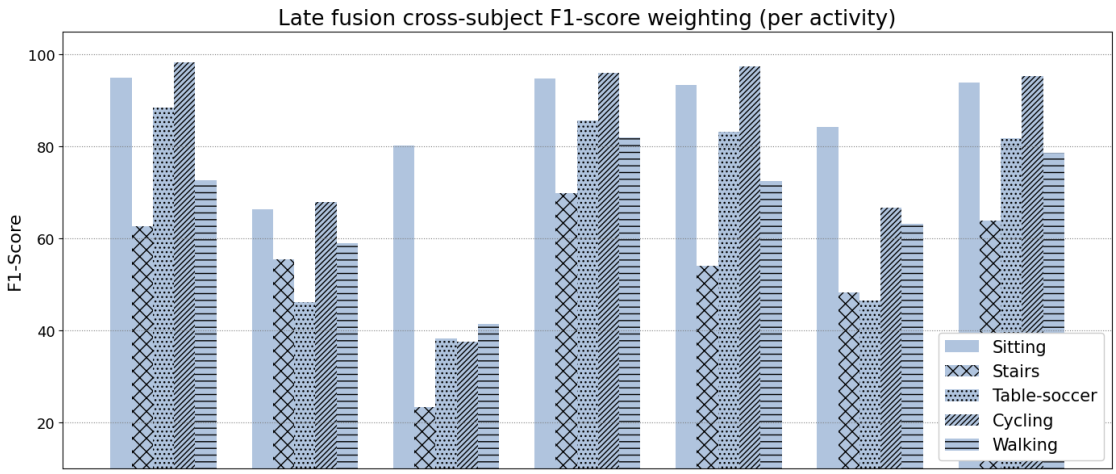
Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 7.3: Resulting F1-score for overall performance for a cross-subject late fusion model and a comparison between per class detection rate and per class F1-score weighting approach.



Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 7.4: Resulting F1-score performance for a cross-subject late fusion model using per class detection rate weighting approach.



Scenario	1	2	3	4	5	6	7
3D-ACC	X			X	X		X
ECG		X		X		X	X
PPG			X		X	X	X

Figure 7.5: Resulting F1-score performance for a cross-subject late fusion model using per class F1-score weighting approach.

7.3 Feature importance

We have shown that fusing 3D-ACC and ECG signals yielded the best performance in classifying human activities in our thesis. However, which features from both signals were the most relevant to our model? In this section we present the feature importance ranking of the model that combines 3D-ACC and ECG (Scenario 4) using the cross-subject model, as we want to investigate the best features across multiple subjects. We calculate the feature importance using the Mean Decrease in Impurity (MDI) of our random forest model Louppe (2014). To aggregate the importance score for each model evaluated on a single subject, we calculate the average score for each feature over all the subjects and rank their importance score. As Table 7.3 shows, out of top 20 features, 16 features are related to the 3D-ACC signal and 4 of them to the ECG signal. Naturally, as 3D-ACC provides the best signal of the individual signal models (scenario 1), we expect to see a dominance of 3D-ACC features in the top-20 ranking. Interestingly, frequency domain features rank higher than time domain ones, confirming that frequency-based feature provides informative data to the our classifiers.

Another pattern that emerges is that, for 3D-ACC axis Y and Z are relatively more important than X-axis of the accelerometer sensor. As the original authors of the PPG-DaLiA dataset collected the accelerometer using a wristband, when the subject's hand is positioned alongside the body, the X-axis captures movement in the forwards-backwards direction, the Y-axis captures up-down movements, and finally, the Z-axis captures left-right movements *Empatica. E4Wristband. 2018. Available online:* (2018).

Table 7.3: Feature importance for 3D-ACC + ECG model (scenario 4) in the cross subject setup.

Rank	**Feature domain**	**Signal**	**Feature name**	**Importance score**
1	Frequency	ACC-Y	median	0.0990
2	Time	ACC-Z	median	0.0812
3	Frequency	ACC-Y	second-max	0.0799
4	Frequency	ACC-Z	standard deviation	0.0633
5	Frequency	ACC-Y	standard deviation	0.0626
6	Frequency	ACC-X	second-max	0.0579
7	Frequency	ACC-Y	mean-crossing	0.0535
8	Time	ACC-X	max	0.0475
9	Frequency	ACC-Z	mean-crossing	0.0454
10	Time	ACC-Y	max	0.0385
11	Time	ACC-X	median	0.0365
12	Time	ACC-Y	median	0.0361
13	Time	ACC-Y	mean	0.0343
14	Time	ECG	mean-crossing	0.0318
15	Time	ECG	zero-crossing	0.0279
16	Time	ECG	standard deviation	0.0276
17	Frequency	ACC-Z	predominant frequency	0.0275
18	Frequency	ECG	mean-crossing	0.0203
19	Time	ACC-Z	min	0.0191
20	Frequency	ACC-Y	predominant frequency	0.0188

In this chapter, we provided the complementary insight about the overall topic of this thesis. More precisely, we discussed about the detailed contribution level of ECG signal in HAR systems, a comparison between the late fusion and the early fusion and the most important features. In the chapter that follows, we will wrap up this thesis by the conclusion and the potential future works.

Chapter 8

Conclusion and Future work

8.1 Limitation and future work

In this section we discuss the limitation and the potential future works. We explained in Chapter 3 that a variety of recorded signals are provided in PPG-DaLiA dataset, however, out of IMU sensors (3D-ACC, gyroscope, and magnetometer), the available signals are limited to the 3D-ACC. This limitation prevent us from further investigating the contribution level of bio-signals when combined with gyroscope, and magnetometer signals. Another limitation of our work is the significant difference between different sources of signals sampling rates, to review, 3D-ACC (32 Hz), PPG (64 Hz), and ECG (700 Hz). We stated in Chapter 4 that we up-sample the 3D-ACC signal, however, we keep the ECG signal as its original sampling rate. We decide not to down-sample the ECG signal because we do not want to eliminate some of the ECG information. To this end, as our future work, we plan to record our dataset including all the IMU sensors information, as well as PPG and ECG signals all recorded with comparable sampling rates to reduce the impact of other factors (sampling rate and information volume) on the comparison and contribution level evaluation of different signals. Furthermore, we plan to have more examples of subjects with similar characteristics, in terms of fitness level, height and weight.

Regarding our methodology, we suggest applying other feature extraction approaches such as wavelet transformation or even applying the deep learning approaches to avoid the limitation of conventional machine learning models. A potential future work is to perform the same investigation

considering other types of activities, the transition activities (meaning the activity performed in between two certain activities, e.g. the duration between sitting and standing up) or even the complex activities (such as sequential, concurrent or interleaved human activities which explained in Chapter 3) to evaluate the contribution level of PPG and ECG signals. Even though, our main aim in this thesis is to evaluate the contribution level of heart related data, namely ECG and PPG (indirectly related to heart activity), we suggest including and evaluating other type of bio-signals fusion such as electromyogram (EMG) in HAR systems.

8.2 Conclusion

In this thesis, we perform a comparative study to evaluate the contribution level of different sources of signals in HAR systems. We use a dataset consisting of 3D-ACC, PPG, and ECG signals recorded while 14 subjects were performing daily activities.

After examining different window sizes, we identify that window size of seven seconds is the best segment duration to obtain sufficient information out of the mentioned signals. Afterwards, we extract time and frequency domain features from each segment, then standardize features so that all the features have zero mean and unit variance. Next, we omit features with more than 85% correlation and select 45 features out of total 77 extracted features. We introduce seven scenarios of evaluation, including models trained solely by one source of signal and models trained by a combination of signal features. Finally, we evaluate two types of HAR models using Random Forest classifiers, a subject-specific model and a cross-subject model.

We conclude that in both subject-specific and cross-subject models, 3D-ACC signal is the most informative signal if the HAR system designer's purpose is to record and use only one source of signal. However, our results suggest that 3D-ACC and ECG signal combination improves recognizing activities such as walking and ascending/descending stairs. It is worth mentioning that not in all cases ECG addition would improve the HAR performance, it is highly correlated with subject's fitness level. Moreover, we experimentally assess that features extracted from the PPG signal are not informative for HAR system, not exclusively, nor when applying signal fusion. Although, both bio-signals yield a satisfactory performance in distinguishing stationary activities (e.g. sitting)

from non-stationary activities (e.g. walking, cycling). Lastly, the early fusion and the late fusion approaches yield almost similar results in terms of F1-score and different signals contribution. Overall, our results indicate that it might be beneficial to combine features from the ECG signal in scenarios in which pure 3D-ACC models struggle to distinguish between activities that have similar motion (walking vs walking up/down the stairs) but differ significantly in their heart rate signature.

Table 8.1, represents the abbreviations we used in this thesis:

Table 8.1: Table of abbreviations

Abbreviation	**Definition**
HAR	Human Activity Recognition
IMU	Inertial Measurement Unit
3D-ACC	Three dimensional accelerometer signal
ECG	Electrocardiogram signal
PPG	Photoplethysmogram signal
FFT	Fast Fourier Transform
CFS	Correlation based Feature Selection
AUC	Area Under Curve
ROC	Receiver Operating Characteristic
LOSO	Leave-One-Subject-Out cross validation

References

Aguileta, A. A., Brena, R. F., Mayora, O., Molino-Minero-Re, E., & Trejo, L. A. (2019). Multi-sensor fusion for activity recognition—a survey. *Sensors*, *19*(17), 3808.

Andrejašic, M. (2008). Mems accelerometers. In *University of ljubljana. faculty for mathematics and physics, department of physics, seminar.*

Athavale, Y., & Krishnan, S. (2017). Biosignal monitoring using wearables: Observations and opportunities. *Biomedical Signal Processing and Control*, *38*, 22–33.

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors*, *15*(12), 31314–31338.

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window size impact in human activity recognition. *Sensors*, *14*(4), 6474–6499.

Bayat, A., Pomplun, M., & Tran, D. A. (2014). A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, *34*, 450–457.

Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., & Turchetti, C. (2017). Human activity recognition using accelerometer and photoplethysmographic signals. In *International conference on intelligent decision technologies* (pp. 53–62).

Biosignalsplux. respiban professional. 2019. available online:. (2019). https://biosignalsplux.com/products/wearables/respiban-pro.html. Retrieved from https://biosignalsplux.com/products/wearables/respiban-pro.html (accessed on 21 May 2019.)

Bjørnstad, H., Storstein, L., Meen, H. D., & Hals, O. (1993). Electrocardiographs findings according to level of fitness and sport activity. *Cardiology*, *83*(4), 268–279.

Boukhechba, M., Cai, L., Wu, C., & Barnes, L. E. (2019). Actippg: using deep neural networks for activity recognition from wrist-worn photoplethysmography (ppg) sensors. *Smart Health*, *14*, 100082.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... others (2013). Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Casale, P., Pujol, O., & Radeva, P. (2011). Human activity recognition from accelerometer data using a wearable device. In *Iberian conference on pattern recognition and image analysis* (pp. 289–296).

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, *4*(4), 195.

Catal, C., Tufekci, S., Pirmit, E., & Kocabag, G. (2015). On the use of ensemble of classifiers for accelerometer-based activity recognition. *Applied Soft Computing*, *37*, 1018–1022.

Chen, Y., Zhong, K., Zhang, J., Sun, Q., & Zhao, X. (2016). Lstm networks for mobile human activity recognition. In *2016 international conference on artificial intelligence: technologies and applications* (pp. 50–53).

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data. *IEEE journal of biomedical and health informatics*, *22*(3), 678–685.

Chung, S., Lim, J., Noh, K. J., Kim, G., & Jeong, H. (2019). Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors*, *19*(7), 1716.

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., ... Welch, P. D. (1967). What is the fast fourier transform? *Proceedings of the IEEE*, *55*(10), 1664–1674.

Dehghani, A., Sarbishei, O., Glatard, T., & Shihab, E. (2019). A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors*, *19*(22), 5026.

Demrozi, F., Pravadelli, G., Bihorac, A., & Rashidi, P. (2020). Human activity recognition using

inertial, physiological and environmental sensors: a comprehensive survey. *IEEE Access*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Elamvazuthi, I., Izhar, L. I., Capi, G., et al. (2018). Classification of human daily activities using ensemble methods based on smartphone inertial sensors. *Sensors*, *18*(12), 4132.

Empatica. e4wristband. 2018. available online:. (2018). https://www.empatica.com/research/e4/. Retrieved from https://www.empatica.com/research/e4/ (accessed on 21 May 2019.)

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Ferrari, A., Micucci, D., Mobilio, M., & Napoletano, P. (2021). Trends in human activity recognition using smartphones. *Journal of Reliable Intelligent Environments*, 1–25.

Figo, D., Diniz, P. C., Ferreira, D. R., & Cardoso, J. M. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, *14*(7), 645–662.

Hadjileontiadis, L. J. (2006). Biosignals and compression standards. In *M-health* (pp. 277–292). Springer.

Hassan, M. M., Uddin, M. Z., Mohamed, A., & Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, *81*, 307–313.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, Z., & Jin, L. (2009). Activity recognition from acceleration data based on discrete consine transform and svm. In *2009 ieee international conference on systems, man and cybernetics* (pp. 5041–5044).

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, p. 278-282 vol.1). doi: 10.1109/ICDAR.1995.598994

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Koskimaki, H., Siirtola, P., et al. (2016). Accelerometer vs. electromyogram in activity recognition.

Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, *15*(3), 1192–1209.

Lara, O. D., Pérez, A. J., Labrador, M. A., & Posada, J. D. (2012). Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and mobile computing*, *8*(5), 717–729.

Lee, M.-W., Khan, A. M., & Kim, T.-S. (2011). A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. *Personal and Ubiquitous Computing*, *15*(8), 887–898.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Machado, I. P., Gomes, A. L., Gamboa, H., Paixão, V., & Costa, R. M. (2015). Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing & Management*, *51*(2), 204–214.

McSharry, P. E., Clifford, G. D., Tarassenko, L., & Smith, L. A. (2003). A dynamical model for generating synthetic electrocardiogram signals. *IEEE transactions on biomedical engineering*, *50*(3), 289–294.

Mehrang, S., Pietilä, J., & Korhonen, I. (2018). An activity recognition framework deploying the random forest classifier and a single optical heart rate monitoring and triaxial accelerometer wrist-band. *Sensors*, *18*(2), 613.

Mendes Jr, J. J. A., Vieira, M. E. M., Pires, M. B., & Stevan Jr, S. L. (2016). Sensor fusion and smart sensor in sports and biomedical applications. *Sensors*, *16*(10), 1569.

Micucci, D., Mobilio, M., & Napoletano, P. (2017). Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, *7*(10), 1101.

Noor, M. H. M., Salcic, Z., Kevin, I., & Wang, K. (2017). Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer. *Pervasive and Mobile Computing*, *38*, 41–59.

Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research

directions. *Information Fusion*, *46*, 147–170.

Park, H., Dong, S.-Y., Lee, M., & Youn, I. (2017). The role of heart-rate variability parameters in activity recognition and energy-expenditure estimation using wearable sensors. *Sensors*, *17*(7), 1698.

Pedley, M. (2013). Tilt sensing using a three-axis accelerometer. *Freescale semiconductor application note*, *1*, 2012–2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ravi, D., Wong, C., Lo, B., & Yang, G.-Z. (2016). A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics*, *21*(1), 56–64.

Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep ppg: large-scale heart rate estimation with convolutional neural networks. *Sensors*, *19*(14), 3079.

Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, *59*, 235–244.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience*, *6*(5), gix019.

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 313–325).

Sakr, N. A., Abu-Elkheir, M., Atwan, A., & SOLIMAN, H. (2018). Current trends in complex human activity recognition. *Journal of Theoretical & Applied Information Technology*, *96*(14).

Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., & Havinga, P. J. (2016). Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, *16*(4), 426.

Strath, S. J., Kaminsky, L. A., Ainsworth, B. E., Ekelund, U., Freedson, P. S., Gary, R. A., ... Swartz, A. M. (2013). Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the american heart association. *Circulation*, *128*(20), 2259–2279.

Sun, Y., Kamel, M. S., & Wong, A. K. (2005). Empirical study on weighted voting multiple classifiers. In *International conference on pattern recognition and image analysis* (pp. 335–344).

Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., & Friedman, R. (2007). Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *2007 11th ieee international symposium on wearable computers* (pp. 37–40).

Telgarsky, R. (2013). Dominant frequency extraction. *arXiv preprint arXiv:1306.0103*.

Tsanousa, A., Meditskos, G., Vrochidis, S., & Kompatsiaris, I. (2019). A weighted late fusion framework for recognizing human activity from wearable sensors. In *2019 10th international conference on information, intelligence, systems and applications (iisa)* (pp. 1–8).

Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics and AI*, *2*, 28.

Wang, A., Chen, G., Yang, J., Zhao, S., & Chang, C.-Y. (2016). A comparative study on human activity recognition using inertial sensors in a smartphone. *IEEE Sensors Journal*, *16*(11), 4566–4578.

Wang, Y., Cang, S., & Yu, H. (2019). A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, *137*, 167–190.

Weiss, G. M., Timko, J. L., Gallagher, C. M., Yoneda, K., & Schreiber, A. J. (2016). Smartwatch-based activity recognition: A machine learning approach. In *2016 ieee-embs international conference on biomedical and health informatics (bhi)* (pp. 426–429).