Master thesis : Prediction of the absolute and relative intensity of daily life activities

Auteur : Di Matteo, Alyssa
Promoteur(s) : Schwartz, Cédric
Faculté : Faculté des Sciences appliquées
Diplôme : Master en ingénieur civil biomédical, à finalité spécialisée
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24791

Prediction of the absolute and relative intensity of daily life activities

Di Matteo Alyssa

Thesis presented to obtain the degree of :
Master of Science in Biomedical Engineering

Thesis supervisor :
Prof. Cédric Schwartz

Academic year: **2024 - 2025**

Acknowledgments

I would first like to express my gratitude to Professor Cédric Schwartz, whose valuable advice, availability and ideas have contributed to the development of my work. I am especially grateful for having been given access to the Human Movement Analysis Laboratory and for the confidence he showed in allowing me to use the required equipment.

I would also like to extend my sincere thanks to the members of the jury, Professor Olivier Brüls and Professor Didier Maquet, for dedicating their time to reviewing this work.

I wish to express my appreciation to all the participants who volunteered for this study and for the time they dedicated to it. Without their contribution, this work would not have been possible.

I am would also like to acknowledge Hassan Ashraf, a PhD candidate at the University of Liège, for his guidance, and help in the field of machine learning.

I am especially thankful to my sister, Mélina, for reading my work and sharing her opinions. I am also deeply grateful to my parents, who have always supported me throughout my studies and the completion of this thesis.

Finally, I would like to sincerely thank all those close to me for their support and encouragement during my studies. Their guidance and support throughout the past five years have been invaluable.

Abstract

Physical activity plays a fundamental role in health, contributing to the prevention of chronic diseases and the promotion of overall well-being. Despite its benefits, a significant portion of the population remains inactive, highlighting the need for accurate and individualized methods to assess physical activity intensity. Intensity, which influences health outcomes, can be measured in absolute terms or in relative terms. While absolute intensity is independent of an individual's fitness level, relative intensity offers a more individualized perspective, which is important for personalized assessment of physical activity.

This study explores the feasibility of classifying relative intensity using wearable inertial sensors combined with machine learning techniques, and compares its performance with absolute intensity classification. Data were collected from twenty participants performing seven daily activities, using inertial measurement units placed on the wrist and the foot to record movements. After data segmentation, feature extraction, and the construction of the final datasets, three supervised machine learning models including Random Forest, Support Vector Machine, and Multilayer Perceptron, were developed and evaluated. Both multiclass (sedentary, light, moderate and vigorous) and binary (sedentary-light vs. moderate-vigorous) classification were explored to assess model performance across different tasks and sensor placements. A feature importance analysis was also conducted to identify the most discriminative features in the classification of relative intensity using the wrist sensor.

The results indicated that classifying relative intensity from inertial sensor data remains challenging. Multiclass classification often resulted in confusion between intensity levels, whereas binary classification improved overall performance. Sensor placement affected model outcomes, with performance varying across algorithms and tasks. Absolute intensity classification achieved higher overall performance in both multiclass and binary tasks and for both sensor placements, although some confusion still occurred. These results show that absolute intensity is generally easier to classify than relative intensity, which reflects the additional complexity of personalizing intensity assessments. Among the tested models, Random Forest consistently achieved the highest performance across all tasks and sensor placements.

These findings provide insight into the challenges of using wearable sensors and machine learning to assess the relative intensity of physical activity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

From walking to work, climbing stairs, or even doing household chores, everyday activities shape our health. On a broader scale, physical activity plays a fundamental role in promoting health and preventing chronic diseases, which represent a major public health challenge worldwide [1]. Despite its well-established benefits, many people remain inactive, making physical inactivity a major contributor to premature death worldwide [2]. Given this, it is crucial to understand not only if people are active or not, but also how they engage in physical activity. It can be performed in various ways, differing not only in type, frequency or duration but also in intensity [3]. This intensity is a key factor that influences health outcomes [1]. However, accurately measuring it remains challenging and requires reliable methods.

Different approaches exist to assess activity intensity, including absolute and relative measures of intensity, each with its own advantages and limitations [3]. Absolute intensity is independent of an individual's fitness level, while relative intensity provides a more individualized perspective by evaluating how demanding an activity is for a specific person [3]. While more difficult to measure, relative intensity is particularly important for personalized assessment, prescription, and monitoring of physical activity [3].

Methods for assessing intensity range from subjective evaluations, based on perceived exertion, to objective techniques that rely on physiological or biomechanical data [4]. In recent years, technological advancements have expanded the use of wearable sensors as objective tools for continuous monitoring and analyzing physical activity [4]. Furthermore, advances in data processing methods, and particularly in machine learning, provide new opportunities for more accurate and personalized approaches to assess physical activity [4].

This study focuses on exploring how wearable sensors combined with machine learning techniques can be used to predict relative intensity, aiming to contribute to better individualized physical activity assessment.

This thesis is organized into six chapters, beginning with an introduction to the study in this first chapter. Chapter 2 provides a comprehensive literature review on the role of physical activity in health and the methods used to assess physical activity intensity. It covers both absolute and relative intensity and subjective and objective assessment techniques. The chapter also outlines the research objectives. Chapter 3 details the methodology, including the experimental protocol, data collection, and the steps of data processing. It also presents the machine learning models used, hyperparameter tuning, evaluation metrics, and a feature importance analysis. Chapter 4 presents the results and their analysis, focusing on model performance across different tasks and sensor placements, comparing relative and absolute intensity classification, and exploring predictive features. Finally, chapter 5 discusses the limitations of the study, while Chapter 6 concludes by summarizing the main findings.

Chapter 2

State of the art

This chapter provides a comprehensive review of the existing literature related to the role of physical activity in health, with a particular focus on the methods used to assess physical activity, particularly intensity.

First, this chapter highlights the crucial role of physical activity not only in preventing chronic diseases but also in promoting health for all. Then, it introduces the concepts of absolute and relative intensity, which correspond to different approaches of evaluating how demanding an activity is for an individual. Subsequently, the chapter reviews the different approaches used to assess the intensity of physical activity. It distinguishes between subjective methods, based on individual perception and objective methods, which rely on physiological or biomechanical data, such as wearable sensors, as well as standardized estimates like the Metabolic Equivalent of Tasks. Particular attention is given to the use of wearable sensors, particularly accelerometers, which are a widely adopted tool in this field. Different approaches for interpreting accelerometer data are discussed, with a focus on the cut-point method as well as machine learning techniques. Finally, the chapter explores the potential of inertial measurement units.

2.1 Prevalence and impact of chronic diseases

Chronic diseases represent a growing burden for global health, contributing considerably to both morbidity and mortality [5]. According to the World Health Organization (WHO), a chronic disease is a condition which is not passed from person to person, of long duration and generally slow progression. The main types include cardiovascular diseases, cancer, diabetes and chronic respiratory diseases [6]. These types of conditions have both physical and psychological impacts, negatively affecting quality of life [5].

According to a WHO press release, chronic diseases accounted for 4 of the 10 leading causes of death worldwide in 2000, a number that rose to 7 out of 10 in 2019 [7]. More recent data from the WHO's European Region indicates that, in 2024, chronic diseases are responsible for 8.2 million deaths annually [8]. These figures highlight the importance of prevention in mitigating the impact of chronic diseases.

2.2 Importance of physical activity

A well-established, non-invasive prevention method is physical activity, which has been proven to be particularly effective in reducing the risk and effects of chronic disease [1]. As stated by WHO, physical activity is defined as "any bodily movement produced by skeletal muscles that requires energy expenditure" [1]. A sedentary lifestyle is closely associated with an increased risk of chronic disease [2]. On the other hand, an active lifestyle, even involving light-intensity activity, can provide health benefits [1]. Physical activity can induce substantial health benefits for adults and older adults suffering from chronic diseases such as cancer, hypertension and type 2 diabetes [1]. Being active may help to reduce all-cause mortality, as well as mortality specifically linked to cardiovascular disease in cases of hypertension and type 2 diabetes [1]

These findings underline the crucial role that physical activity plays in the management of chronic health conditions. While essential for individuals with chronic diseases, physical activity remains just as important for the general population. Regular physical activity helps to prevent numerous health issues and promotes both physical and mental health [1, 2]. For example, it has been associated with improving cognitive function, reducing symptoms of anxiety, and enhancing quality of life [1, 2]. Moreover, for older adults, it helps preserve functional capacity, maintain independence and reduce the risk of falls [1, 2].

According to WHO recommendations on physical activity and sedentary behavior, adults aged over 18, whether in good health or living with a chronic disease, should engage in 150 to 300 minutes of moderate-intensity aerobic physical activity per week or at least 75 to 150 minutes of vigorous-intensity aerobic activity [1]. This can also be achieved through an equivalent combination of both intensities throughout the week. Furthermore, they should include at least two weekly sessions of muscle-strengthening activities involving all major muscle groups at a moderate or greater intensity to gain additional health benefits [1]. These recommendations highlight the need to differentiate between the terms: aerobic physical activity and muscle strengthening activities. Aerobic activity involves rhythmic movements of large muscle groups performed for a continuous period that improves cardiorespiratory fitness [7]. Muscle-strengthening activities, on the other hand, improve skeletal muscle strength, endurance and power [7]. Physical activity can be performed in a variety of ways and at different intensities daily, including occupational tasks, transportation, household chores, or leisure activities [3].

Accurately measuring physical activity is fundamental for understanding the relationship between physical activity and health [3]. To comprehensively assess physical activity, four parameters are commonly considered: frequency, duration, intensity and type of activity [3]. Frequency refers to how often physical activity is performed and duration to the time of the activity. The type characterizes the nature of the activity performed and intensity describes the level of effort required to perform an activity. Although all these parameters are important, this work will focus specifically on intensity, which is particularly crucial due to its influence on health outcomes [1].

2.3 Absolute and relative intensity of physical activities

The intensity of physical activity can be measured in different ways, either in an absolute or in a relative manner [3]. The absolute measure consists of quantifying the rate of energy expenditure required to perform a certain activity, without taking into account the individual's capacity [3]. One of the most common methods of measuring absolute intensity is the use of Metabolic Equivalent of Task (MET), where one MET is equivalent to the amount of energy expended by a person sitting at rest [1]. This baseline energy expenditure reflects the resting metabolic rate, which represents the energy required to maintain essential body functions in an awake individual in a thermo-neutral environment and at rest [9].

Absolute measures of energy expenditure are commonly divided into four categories: sedentary, light, moderate and vigorous [1]. Sedentary behavior is characterized by an energy expenditure of less than or equal to 1.5 METs. Above this threshold, a light-intensity activity is described by an activity performed between 1.6 METs and less than 3 METs. An activity is considered moderate intensity when its metabolic equivalent is at or above 3 METs and below 6 METs, when the METs are equal or above 6.0, the activity is perceived as vigorous [1]. To facilitate the classification of intensity and enhance comparability between studies, METs estimates for various physical activities have been published in Compendiums of Physical Activity [10]. There is a Youth Compendium for children of 5-18 years old, an Adults Compendium for adults aged from 19 to 59 years, and an Older Adults Compendiums for adults of $\geq$60 years old [10]. As the resting metabolic rate varies throughout life, different compendiums are needed for different populations [10].

On the other hand, the relative intensity is determined through the level of effort made by an individual relative to their maximum capacity [3]. It can be assessed using both objective and subjective methods. Objective measures are based on physiological parameters typically expressed as a percentage of individual's maximal capacity, such as the percentage of maximal oxygen consumption (%$\dot{V}O_{2\text{max}}$), percentage of maximal heart rate (%HR_{max}) and percentage of maximal heart rate reserve (%HRR) [2]. Subjective methods, such as the Rate of Perceived Exertion (RPE), evaluate how a person feels when exercising [3].

The variation in cardiorespiratory fitness between individuals highlights the importance of not using absolute intensity and relative intensity interchangeably [11]. Indeed, for older adults, individuals with low levels of fitness or, on the contrary, people with a high physical ability, the activities measured in absolute and relative intensity will not be categorized into the same class. Someone with a low level of physical ability will find a specific activity more difficult than someone with a higher level of fitness [3]. The use of absolute intensity, which takes no account of individual differences, leads to a misclassification of intensity levels. For instance, if physical activity is measured using predefined METs, intensity could be overestimated, which can result in individuals receiving a lower return on their investment than expected [12]. Conversely, if the intensity of an activity is too intense for an individual's ability, it may decrease the maintenance and adherence to the exercise routine, reducing the potential health benefits [11]. Measuring relative intensity has gained interest due to its potential for more accurate individualization according to each person's fitness level [12]. A summary comparing absolute and relative intensity of physical activity is presented in Table 2.1.

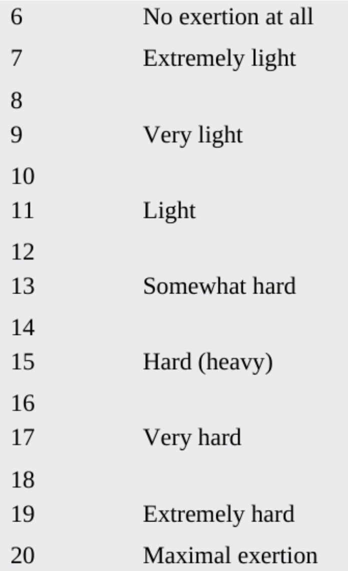
Parameter	**Absolute Intensity**	**Relative Intensity**
Definition	Measures energy expenditure required to perform an activity, regardless of individual capacity.	Measures the effort relative to an individual's maximum capacity.
Unit of measurement	e.g., METs	e.g., %$\dot{V}O_{2\text{max}}$, %HR_{max}, %HRR , RPE
Advantages	- Standardized and easy to apply - Facilitates comparison across studies - Available in compendiums	- Accounts for individual differences - More accurate for health-related recommendations
Limitations	- Ignores inter-individual variability - Risk of misclassification	- Requires personalized assessment - Less standardized across populations

Table 2.1: Comparison between absolute and relative intensity of physical activity

2.4 Assessment of intensity using subjective methods

Subjective methods frequently mentioned in the literature are based on self-assessment by participants and include questionnaires and activity diaries [13]. Although they are inexpensive and accessible, allowing them to be used in studies with large numbers of participants, they present significant limitations in terms of reliability and validity [4, 13]. Indeed, these methods are subject to errors, which can lead to both overestimating and underestimating physical activity [14]. For instance, questionnaires require individuals to remember the intensity of their physical activity over a given period, making them susceptible to memory bias [13]. Activity diaries, on the other hand, consist of a detailed record of one's physical activity and, although they reduce these biases when completed immediately after exercise, they represent a greater load for participants [13].

Another subjective approach, allowing a real-time assessment, is based on the RPE scale. Perceived exertion refers to an individual's assessment of the effort their body is exerting during physical activity. This perception is influenced by various physiological sensations, such as increased heart rate, respiration or breathing rate during exercise [15]. Unlike the questionnaires or activity diaries mentioned above, RPE is an instantaneous measure, enabling individuals to self-assess the intensity of their effort during or immediately after exercise. Among the various RPE scales, the most widely used is the Borg scale, which assesses perceived intensity on a scale ranging from 6 to 20 [2]. This scale is visually represented in Figure 2.1, where values are associated with descriptive labels, ranging from "no exertion at all" at 6, to "maximal exertion" at 20. This gradation has been chosen to correspond approximately to the heart rate multiplied by 10 in healthy adults (an effort perceived at 12 corresponds to around 120 beats/min) [15]. A modified version of Borg's scale, called Category Ratio 10 Scale, can also be used [15]. This version ranges from 0 to 10, where 0 represents no effort and 10 maximum effort [15]. Although the RPE scale remains influenced by factors such as health status or even the test environment, this scale is highly correlated with physiological markers such as heart rate, or blood lactate concentration, making it a relevant tool for assessing the relative intensity [2].



6	No exertion at all
7	Extremely light
8	
9	Very light
10	
11	Light
12	
13	Somewhat hard
14	
15	Hard (heavy)
16	
17	Very hard
18	
19	Extremely hard
20	Maximal exertion

Figure 2.1: The Borg rating of perceived exertion scale [2]

2.5 Assessment of intensity using objective methods

Despite the advantages of the self-assessment methods reported above, objective methods have been prioritized and developed to provide more accurate measurements [14]. Objective methods include using METs or wearable sensors, which encompass heart rate monitors, accelerometers, inertial sensors or multi-sensor approaches combining, for example, accelerometers and heart rate monitoring [9, 14, 16]. Nonetheless, the gold standard for assessing the intensity of physical activity remains doubly labeled water [9]. This technique involves administering stable isotopes of deuterium and oxygen orally, then tracking their elimination rates from the body [9]. The difference between the elimination rates corresponds to the rate of carbon dioxide production, which can be converted to energy expenditure [9]. However, this method is expensive and difficult to use, requiring trained personnel [9]. Consequently, more affordable and practical methods, particularly wearable sensors, have gained interest for the objective assessment of physical activity [14].

2.5.1 Metabolic Equivalent of Tasks

As previously mentioned, a MET represents the amount of energy expended by a person sitting at rest. The Compendium of Physical Activity is widely used in studies as a criterion measure for defining activity intensity levels by associating observed activities with their corresponding MET values [10]. This classification allows for an estimation of energy expenditure without the need for direct measurement [17]. Although specific compendiums exist for different populations (children, adults, older adults) applying these MET values could lead to significant misclassification of intensities [10]. The aim of the Compendium is to facilitate epidemiological research by assigning MET values to activities and improving comparisons between different studies [10]. Nonetheless, as Herrmann et al. (2024) highlighted, many users of the compendiums have used it to estimate individual energy expenditure and to design exercise programs, even though this approach is not fully accurate [10]. Indeed, as many factors influence the resting metabolic rate and the energy expenditure related to physical activity, the compendium does not provide an accurate measure of energy expenditure specific to each individual but offers a starting point for classifying and prescribing physical activities [10].

Moreover, 1 MET is also defined as 3.5 mL $kg^{-1}min^{-1}$, which represents the average value for a healthy standard 70kg person of 40 years old [18]. By knowing the MET value of an activity, it is possible to approximate the oxygen consumption required to perform it. Although commonly used in the literature, this approach has been shown to overestimate the intensity of activities in different populations, particularly in older adults [12, 18]. Using predefined formulas or METs without measuring physiological parameters provides an estimation of the absolute intensity. Certain measures of the oxygen consumption allow the assessment of relative intensity to individualize and better classify the intensity of different activities within a population. These measures include the percentage of maximal oxygen consumption and the percentage of oxygen consumption reserve, which is defined as the difference between an individual's maximal oxygen uptake and their resting oxygen uptake [2]. However, measuring oxygen consumption requires expensive equipment and trained professionals to perform the test and interpret the data, which makes it difficult to use in studies with a large number of participants [2].

2.5.2 Heart rate monitoring

Heart rate monitors are low-cost, non-invasive tools to measure heart rate [13]. They can be used to indirectly estimate intensity based on the principle that heart rate varies linearly with activity intensity and oxygen consumption during moderate and vigorous activities [13]. However, the correlation is weak for sedentary and light activities, making differentiation difficult [13]. To improve accuracy, the relationship between heart rate and energy expenditure can be individually calibrated by directly measuring oxygen consumption at different levels of effort, although this approach is difficult to apply in large-scale studies [2, 19]. To overcome this constraint, generalized predictive models have been developed using equations derived from group data to estimate energy expenditure from heart rate [19]. Although these models provide acceptable estimates for some population groups, they remain less accurate than personalized approaches [19].

In addition, heart rate can be used to assess relative intensity through different approaches, such as the percentage of a person's heart rate reserve or the percentage of a person's maximal heart rate [2]. Heart rate reserve is defined as the difference between an individual's maximum heart rate and their resting heart rate [2]. Maximal heart rate can be estimated using age-predicted formulas or measured individually through progressive stress tests, such as maximal graded exercise tests [16].

2.5.3 Accelerometers

Accelerometers are cheap, practical, and portable motion sensors increasingly used to assess physical activity intensity [4]. They measure the acceleration of body movements, where acceleration refers to the rate of change of velocity over a specified time. By analyzing acceleration data, the frequency, intensity and duration of physical activity can be assessed [20]. When measuring intensity, the recorded data allow for estimating energy expenditure to quantify the absolute intensity of physical activities [20]. In general, the most common types of accelerometers currently used for evaluating physical activities are uniaxial and triaxial [20]. Uniaxial accelerometers measure acceleration in a single plane, generally the vertical plane (up-down direction). Triaxial accelerometers determine acceleration along three axes: vertical, horizontal (forward-backward direction) and mediolateral (side-to-side movements). The combination of the three planes can be expressed as a vector magnitude (VM) using the formula,

$$VM = \sqrt{X^2 + Y^2 + Z^2}$$

where X, Y and Z represent the sum of counts from the x-axis, y-axis and z-axis, respectively [20].

Output of accelerometers

Accelerometers record raw acceleration data, which represent the direction and magnitude of acceleration from each axis. The format of the output can vary depending on the sensor used [4]. One of the methods commonly reported in the literature is output in the form of count [20, 21]. A count corresponds to an arbitrary measure that aggregates the intensity and the frequency of an activity over a specific time period [21]. When accelerometers return counts, the raw acceleration data is directly transformed into counts using proprietary algorithms, making the raw data inaccessible [20]. These algorithms are developed by the manufacturers of these sensors and involve different filters, amplifiers or sampling frequencies [21]. The main limitation is that, even if the same signals are measured, variations in the brand and the algorithms used result in different counts, which complicates the comparison between data from different accelerometers [20]. Furthermore, these algorithms are specific to the age group as well as the sensor placements, which further complicate comparison across different studies [4, 22].

An alternative approach progressively used to improve the limitations of the non-standardization of outputs in the form of counts, is to measure and store the raw data in gravitational units [21]. This process, however, requires data processing, which represents an additional challenge. Analyzing raw signals involves a number of difficulties such as managing the huge amount of data generated and the need for suitable mathematical and statistical tools to analyze the data accurately and make valid interpretations from it [21]. Among the methods used to process raw acceleration data from triaxial accelerometers, the most common statistical tools in the literature are the Euclidean Norm Minus One (ENMO) and the Mean Amplitude Deviation (MAD) methods [4, 21]. The two metrics are computed from the Euclidean norm of the three acceleration signals. MAD represents the mean value of the dynamic acceleration. The Euclidean norm includes two components: a dynamic component related to changes in velocity and a static component due to gravity. The MAD value is then obtained by removing the effect of gravity, leaving only the dynamic part. The dynamic component is then analyzed to extract its average value. On the other hand, in the ENMO metric, gravity is adjusted by subtracting one gravitational unit from the Euclidean norm of the three raw acceleration signals [21].

Cut-point method

Once the raw acceleration has been converted into counts or gravitational units, a value calibration step can be carried out. The cut-point method is the most widely used approach for calibration to estimate physical activity intensity, and it is commonly applied to estimate or predict absolute intensity [4, 20]. This method relies on defining thresholds to categorize movement into sedentary, light, moderate or vigorous intensity levels. The first step involves transforming the results of data processing into a more intuitive measure of physical activity. This measure can be an estimate of intensity, generally expressed as energy expenditure [4]. This is typically done by converting acceleration signals into MET values using statistical analysis techniques such as linear regression or the receiver operating characteristic curve (ROC) [23]. Linear regression is used to predict energy expenditure from acceleration data. ROC curves, on the other hand, establish optimal thresholds that offer the best combination of sensitivity and specificity for distinguishing between different intensity categories [23]. Sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives, while specificity represents the ratio of true negatives to the total of true negatives and false positives. Once cut points are established, accelerometer data can be categorized into intensity levels, allowing researchers to quantify time spent in different activity intensities.

When relying solely on accelerometer data, the cut-point method is primarily used to estimate absolute intensity levels. However, studies such as those by Siddique et al. (2020) and Miller et al. (2010), have explored the possibility of adapting this method to the prediction of relative intensity by integrating physiological measures specific to each individual, such as heart rate or aerobic capacity, or by incorporating RPE [16, 24]. These findings suggest that integrating such data could allow the definition of personalized thresholds [16, 24].

The cut-point method presents several important limitations. Intensity thresholds are defined based on the activities selected in the validation studies, as well as the characteristics of the population being studied [20]. If other researchers wish to reuse these thresholds to analyze accelerometer data, they must follow the same data collection and processing protocols as those used in the initial study [20]. Differences in cut points lead to the same activities being classified differently, which complicates standardization and comparability across studies [20].

In addition to variability across accelerometer brands and populations, sensor placement on the body significantly influences recorded acceleration [20, 22]. The location of the sensor affects both the raw signal and the resulting intensity classification. Furthermore, certain types of physical activity may be more accurately captured by specific placements than others [25]. As a result, cut points established and validated for a specific placement cannot be directly applied to another placement without a loss of accuracy [20].

In the literature, the most common placements are the hip and wrist, although other placements, such as the thigh or ankle, have also been studied [4, 26]. The hip, which is the conventional attachment site for accelerometers due to its proximity to the center of mass, provides an accurate measurement of intensity [4]. Despite the accuracy of hip placement, the wrist has gained significant interest in the literature, notably due to its better acceptability, comfort, and compliance [25]. This interest is reflected in its use in large-scale studies such as the National Health and Nutrition Examination Survey [25]. However, a study by Montoye et al. (2020) shows that the cut points developed for wrist placement in free-living settings are less accurate for physical activity assessment than hip-worn accelerometers [25]. Similarly, Trost et al. (2022) found that cut points derived from wrist-worn accelerometers perform poorly and are not a reliable indicator of physical activity intensity, highlighting the need for alternative approaches for processing wrist accelerometer data [23].

Machine learning

The cut-point method, while simple to use, has significant limitations, as mentioned above, which led researchers to turn to other approaches, notably the use of machine learning to process accelerometer data [4, 17, 20]. Despite its potential, the adoption of machine learning methods remains relatively limited [27]. This is partly due to the technical challenges involved in processing large volumes of raw accelerometer data, which often requires specialized software and expertise [27]. Machine learning is a branch of artificial intelligence capable of learning automatically and improving through experience, while not being explicitly programmed [28]. This ability makes machine learning well suited to handle complex and high-dimensional datasets, like accelerometer signals [28].

Two main approaches are generally used: supervised learning and unsupervised learning. Supervised learning methods aim to identify the relationship between input features and a target variable based on a set of labeled data. They rely on regression tasks, where the aim is to predict continuous outcomes, and classification, where the goal is to assign data to predefined categories [28]. In this context, regression refers to estimating energy expenditure, typically expressed as METs, and classification refers to assigning intensity levels [29]. Among supervised methods, activity intensity classification can be accomplished indirectly or directly. In the indirect approach, energy expenditure is first predicted, and intensity levels are then derived by applying threshold values [17]. However, this approach has shown bias and high errors [17]. In contrast, studies show that direct classification of activity intensity into categories may improve measurement accuracy [17, 30]. While supervised learning requires labeled data to predict outcomes or classify intensities, unsupervised learning uses unlabeled data to uncover hidden patterns, structures or relationships in the data [28].

Machine learning models development

To implement machine learning models effectively, feature engineering is an essential step [28]. It is divided into two main steps: feature extraction and feature selection or reduction. Feature extraction is the process of converting raw data into meaningful features, while feature selection or reduction is the process of choosing discriminative features for machine learning modeling. Both feature extraction and feature selection have been shown to impact the performance of classification models [28, 31]. Indeed, the features must capture relevant characteristics to allow an accurate classification of activity intensity levels.

Feature extraction from accelerometer data usually involves computing features from both the time domain and the frequency domain [32]. Time domain features are extracted directly from the raw signal and are widely used in research, due to their ease of computation [31]. Moreover, time domain characteristics have been shown to be effective in predicting activity intensity [30]. In contrast, frequency domain features are obtained by transforming the signal into the frequency domain [31]. Even if they are frequently employed, their usefulness remains controversial in research [30, 31, 33]. For instance, Montoye et al. (2018), using a wrist-worn accelerometer to classify physical activities, found no significant improvement in model performance when frequency-domain features were included [30]. Conversely, Ellis et al. (2016), reported that, particularly for wrist sensors, incorporating these features contributed to better classification performance [34].

Regarding feature selection, it remains inconsistent across studies [31, 32]. Some use data-driven methods to identify the most relevant features for optimal performance, while others rely on predefined features already used in previous studies, without algorithmic optimization [31, 32].

Once features have been extracted and selected if needed, the development of machine learning algorithms is the next key step in optimizing model performance [28]. A wide range of machine learning algorithms have been applied to physical activity data, but to date, there is no consensus regarding which method performs best [17, 27, 35]. The performance of the models varies depending on factors such as input features, sensor placement, and population characteristics [17, 27]. Many algorithms have been used for both regression and classification. Among these approaches, random forest (RF) and artificial neural networks (ANN) have shown promising performance in several studies.

For instance, regarding RF, Ahmadi and Trost (2022) compared machine learning classification models to traditional cut-point methods in preschool-aged children [27]. They showed that RF trained on accelerometer data achieved significantly higher agreement with the ground truth, particularly in classifying intensities into three categories (sedentary, light and moderate-to-vigorous) [27]. Another relevant study by Montoye et al. (2018) evaluated several machine learning models, including RF, to predict activity intensity from wrist-worn accelerometer data in adults [30]. Their results showed that RF outperformed other methods and provided reliable predictions in various real-life activity settings [30].

Regarding ANN, a study by Trost et al.(2012) used these networks to classify physical activity types and estimate energy expenditure in youth using accelerometer data [36]. Their results showed that the ANN model outperformed traditional regression-based methods and achieved classification accuracies up to 88.4% [36]. Moreover, Montoye et al. (2016) demonstrated the relevance of ANN models for predicting energy expenditure in adults from raw tri-axial accelerometer data, even when recorded at different body locations [37]. Their ANN models achieved comparable or superior accuracy to traditional methods, including regression and cut-point methods, highlighting the robustness and generalizability of ANN approaches across populations and sensor placements [37].

The support vector machine (SVM) model is also a commonly used machine learning algorithm and has shown promising results [31, 36]. For example, Trost et al. (2018) developed and tested SVM classifiers on accelerometer data collected from preschool children wearing devices on the hip and wrist [38]. They found that SVM outperformed traditional cut-point approaches to classify physical activity intensity [38].

Other supervised algorithms, such as k-nearest neighbors, decision trees, or gradient boosting, have also been used in some studies [29, 30, 32, 35].

Although most algorithms commonly used for physical activity classification are supervised algorithms, studies have demonstrated the potential of unsupervised methods for classifying intensity levels [31, 39, 40]. For example, Li et al. (2020) showed that using a clustering algorithm to define intensity thresholds results in better agreement with gold standard references compared to other supervised approaches [39].

Factors influencing model performance

Sensor placement remains a critical factor, as observed in cut-point methods. With the cut-point method, wrist-worn accelerometers often show lower accuracy than hip-worn accelerometers, limiting their reliability [25]. However, machine learning techniques have improved the accuracy of wrist-based measurements to a level comparable to hip placement [4]. Additionally, machine learning methods have demonstrated that other placements, such as the thigh and ankle, can outperform the commonly used placements: hip and wrist [32, 37].

The context of data collection also affects machine learning model performance. Models developed in laboratory settings reduce the variability in the types of activities and the manner in which participants can perform them [30]. These controlled conditions help improve model performance [41]. However, they often fail to capture the natural variability and complexity of everyday movements. As a result, models developed in laboratory settings tend to show reduced accuracy when applied to free-living environment [41]. Nevertheless, Ahmadi and Trost (2022) have trained and validated models in free-living conditions and reported good performance [27]. They showed that machine learning approaches such as RF can handle the variability of free-living data and outperform traditional cut-point methods when developed and tested in free-living contexts [27].

Machine learning applied to relative intensity

While machine learning has been widely applied to classify absolute intensity levels from accelerometer data, its application to relative intensity remains limited [35, 42]. Indeed, instead of taking individual variability into account, most studies continue to rely on fixed thresholds [35]. This gap highlights the need for further research exploring the potential of machine learning to classify relative intensity, which could allow for more personalized assessments of physical activity [12].

A few recent studies have addressed relative intensity [29, 42, 43]. However, they do not rely solely on accelerometer data as input. Indeed, additional physiological or demographic features are used. For example, Chowdhury et al. (2019) explored the use of supervised machine learning models to classify relative intensity levels using physiological signals such as heart rate, electrodermal activity, and skin temperature, with the RPE as ground truth [42]. While promising, this approach focuses solely on physiological data and did not integrate biomechanical inputs, such as IMU sensors. In contrast, Nnamoko et al. (2021) introduced a machine learning regression model designed to predict personalized acceleration cut-points for sedentary behavior and moderate-to-vigorous physical activity in older adults [43]. Rather than applying fixed thresholds, their model uses individual features such as age, gender, and weight to generate continuous estimates of cut-points tailored to each person [43]. Results are promising, showing that personalized cut-points are not only feasible but also superior to generic thresholds in capturing older adults' movements [43]. Such findings underscore the potential of individualized methods to more accurately reflect physical activity patterns across diverse populations.

2.5.4 Inertial measurement units

Although accelerometer data are widely used in the literature, they capture only linear acceleration. Other aspects of movement, like rotation, can be measured using additional sensors, including gyroscopes [44]. Gyroscopes measure angular velocity, indicating how fast and in what direction the sensor rotates. Accelerometers and gyroscopes are often integrated into a single device know as an inertial measurement unit.

Despite their potential, gyroscopes have been less frequently used in studies on physical activity intensity [44, 45]. However, they have proven effective in predicting energy expenditure [44]. For instance, Hibbing et al. (2018) demonstrated that using gyroscope data in addition to accelerometer data reduces the errors when predicting energy expenditure [44]. Moreover, their study showed that the gyroscope outperformed the accelerometer in classifying sedentary behavior [44]. The use of inertial measurement units is therefore promising for improving the classification and prediction of physical activity intensity.

2.6 Objectives

Although several methods exist to assess physical activity intensity, some limitations remain. In particular, the classification of relative intensity has not yet been fully explored using machine learning algorithms applied to data from wearable inertial measurements units.

The main objective of this thesis is to assess the feasibility of classifying the relative intensity of daily life activities using data collected from wearable inertial measurement units with machine learning algorithms. Estimating the relative intensity allows for a more personalized assessment of physical activity. Such personalization is particularly relevant in the context of prescribing and monitoring physical activity, since absolute and relative intensity measures can lead to different interpretations of the same effort, depending on an individual's fitness level and condition.

This study aims to evaluate whether relative activity intensity can be classified into four categories: sedentary, light, moderate and vigorous. In addition, a simplified classification into two broader categories will also be evaluated, grouping sedentary and light classes together and moderate with vigorous. This grouping is relevant in light of current public health guidelines, which emphasize moderate to vigorous physical activity [1]. The performance of the relative intensity classification will be compared to the ability of the same models to predict absolute intensity, which will also be analyzed in detail, since absolute measures are more commonly used in the literature [16]. Furthermore, this thesis will identify the most suitable machine learning model among commonly used algorithms in physical activity intensity assessment and investigate the impact of sensor placement on classification performance.

As a secondary objective, feature importance will be analyzed to identify which input variables contribute the most to the classification of relative intensity. This analysis will help to better understand which factors have the greatest impact on the model's predictions.

Chapter 3

Materials and methods

This third chapter provides a comprehensive overview of the procedures and methods used throughout this study, from data collection to model development. It begins with a description of the participants' characteristics, followed by the equipment used for data acquisition, and a detailed explanation of the experimental protocol. The data processing steps are subsequently described, including segmentation, feature extraction, and the construction of the final datasets. Next, the three supervised machine learning models selected for the classification tasks are introduced, along with a specific section on the choice and tuning of hyperparameters. In addition, the chapter presents the evaluation metrics used to assess and compare the model performance. Finally, feature importance will be explored to identify the inputs that most influence model classifications.

3.1 Ethical approval

Ethical approval was obtained on March 17, 2025, from the Hospital-Faculty Ethics Committee of the University of Liège.

3.2 Study participants

The study participants consisted of twenty healthy adults, including ten females and ten males between 18 and 25 years old. Initially, we planned to recruit an additional group of participants over 60 years old, in order to provide a clear contrast between age groups. However, due to time constraints, recruitment was limited to the younger age group only. Exclusion criteria were the presence of pain, elevated blood pressure, a history of neurological disorders and any contraindications to physical activity or the specific exercises mentioned in the study. A pretest assessment was conducted to ensure the absence of these exclusion criteria. The following information was also collected: age, height and weight and is summarized in Table 3.1. Most of the participants reported engaging in more than two hours of sport per week. In addition, written informed consent was obtained from the participants.

	Women (n = 10)	**Men (n = 10)**	**All (n = 20)**
Mean age ± std (years)	22.5 ± 1.65	23.00 ± 2.00	22.75 ± 1.81
Mean height ± std (cm)	166.10 ± 4.82	178.85 ± 5.89	172.28 ± 8.22
Mean weight ± std (kg)	62.40 ± 10.26	75.45 ± 8.14	68.93 ± 11.27

Table 3.1: Descriptive statistics of participants by sex

3.3 Equipment

Two types of sensors were used: Trigno Delsys Avanti inertial measurement units (IMU) and the Polar Verity Sense optical heart rate monitor. The Trigno Avanti sensors are equipped with a 9-degree-of-freedom IMU, which provides measurements of linear acceleration, angular velocity, and magnetic field. Only the linear acceleration and angular velocity signals were used in this study, providing three-dimensional acceleration and rotational data, with sensor axes oriented as illustrated in Figure 3.1. The Polar Verity Sense uses photoplethysmography to continuously record heart rate, which refers to the measurement of blood volume changes using light [46].

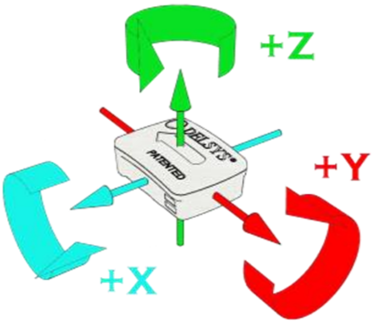


Figure 3.1: Axes of the accelerometer and the gyroscope in the Trigno Delsys Avanti sensor [47]

Each participant wore two Trigno Avanti sensors: one placed on the non-dominant wrist and the other on the non-dominant foot, to record acceleration and angular velocity data. The first sensor was placed on the posterior surface of the forearm, with its proximal edge located 5 cm proximal to the ulnar styloid, with the sensor arrow pointing distally. The second sensor was placed on the shoe at the laces, with the arrow pointing distally. The heart rate monitor was worn as an armband on the non-dominant forearm, just distal to the elbow, to measure heart rate continuously. The correct sensor placements are illustrated in Figure 3.2.

The non-dominant side was chosen based on findings from Montoye et al. (2016), indicating that the non-dominant side may be more effective than the dominant one for assessing physical activity intensity and sedentary behavior [37]. Indeed, one possible explanation is the greater variability of movement in the dominant hand during everyday activities, which may affect the accuracy of activity classification [37].

The placement of the sensor on the wrist was chosen due to its high compliance and its potential for use in free-living environments. In addition, combining wrist placement with machine learning techniques has shown promising levels of accuracy [37]. However, this placement has certain limitations. While it performs well in detecting upper-body movements, it may be less effective for low-impact physical activities such as cycling or for movements where the arms are not involved, for example when walking while carrying bags [26, 48]. To better understand the impact of sensor location, a second sensor was also placed on the foot. The aim was not to combine data from both sensors, but to compare wrist and foot placements independently in terms of classification performance. While previous studies have used placements on the ankle or thigh as lower-limb placement, the shoe was chosen here as a more practical and comfortable alternative, particularly in the context of future applications in free-living settings [37, 45].

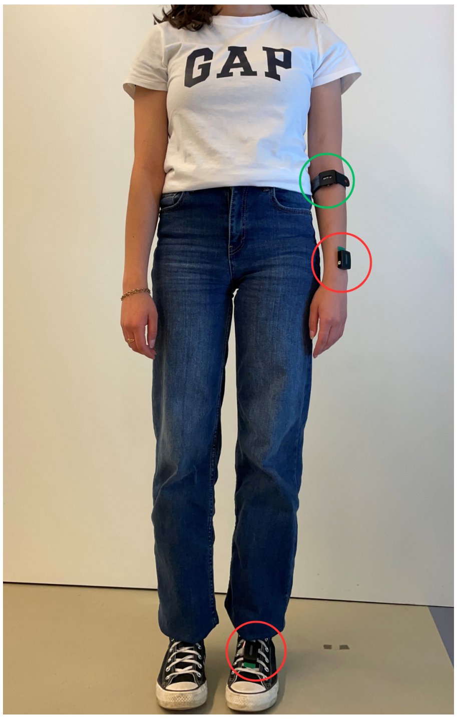


Figure 3.2: Correct sensor placements with IMUs circled in red, and the heart rate sensor in green.

3.4 Data collection

The experience took place at the Laboratory of Human Motion Analysis from the University of Liège, at Sart-Tilman campus. The data collection phase consists of three different stages: resting heart rate measurement, daily life activities and a maximum graded treadmill test.

3.4.1 Resting heart rate measurement

The first step involved measuring heart rate at rest. Participants sat at rest, without doing any other activity. The measurement was taken once the heart rate had stabilized, which indicated that it had reached a resting state.

3.4.2 Daily life activities

The second step consisted of daily life activities performed by the participants. The activities were performed from the least intense to the most intense, each lasting 3 minutes. Between each activity, rest periods were included to allow participants' heart rates to return to resting levels. During the activities, accelerometer data, gyroscope data and heart rate were recorded.

The activities were initially selected using the Physical Activity Compendium for both adults and older adults, as well as existing literature in order to ensure they represent daily life activities and cover different intensity levels [10, 49]. Since only young adults were recruited for the study, the MET values for the selected activities correspond to those listed in the Physical Activity Compendium for adults [10, 50]. The different activities and their descriptions are summarized in Table 3.2. The sedentary activity "sitting and looking at a phone" is not exactly listed in the compendium, therefore we used the MET value for "sitting while reading a book", which is also identical to "sitting watching TV", as these are the more similar activities.

Intensity	Activity	Description	METs	Compendium code
Sedentary *(METs≤1.5)*	Looking at a smartphone	Sitting quietly and looking at a phone for 3 minutes.	1	09030
Light *(1.6≤METs<3)*	Household walking	Walking calmly around the laboratory simulating everyday walk at home during 3 minutes. The route the participants had to follow is detailed in Appendix A: Walking route plan of the household walking activity.	2.3	17150
	Cooking	Consists of four steps: • the movement of beating eggs (30 sec) • cutting clay (1 min) • washing dishes (45 sec) • arranging dishes by moving plates up and down one at a time (45 sec)	2	05050
Moderate *(3≤ METs<6)*	Treadmill walking	Walking at 4.5 km/h on a treadmill for 3 minutes.	3.5	17352
	Sweeping	Fast sweeping using a broom for 3 minutes.	3.8	05012
Vigorous *(METs≥6)*	Stairs with 10 kg load	Repeated climbing and descending backwards 3 steps while wearing 10 kg backpack for 3 minutes. Participants holds the handrails with both hands.	6	17025
	Treadmill running	Running at 7.5 km/h on a treadmill for 3 minutes.	7.8	12029

Table 3.2: Summary and description of all physical activities performed during the experiment.

3.4.3 Treadmill graded test

The third stage consists of performing the maximal effort test on a treadmill as described in the study by Siddique et al. (2020), with the purpose of measuring the real maximal heart rate of each participant [16]. The choice of this maximal treadmill exercise test is based on several considerations, particularly the inclusion of elderly participants in the initial protocol. Although the final study population consisted solely of healthy young adults, the initial protocol was maintained. This type of test allows to gradually reach their maximum heart rate by adjusting the intensity, mainly by increasing the gradient, with the speed increasing slightly or not at all during the first stages. This approach is suited to older adults or populations with lower physical capacity, as it ensures a progressive increase in effort while minimizing the physical impact.

The test consists of nine different stages, each lasting two minutes, with a progressive increase in speed and/or gradient at each stage. The various stage parameters are listed in Table 3.3.

During the test, heart rate, as well as accelerometer and gyroscope data are recorded. At the end of each stage, heart rate and RPE, assessed using the Borg scale from 1 to 10, are measured. The participants continued the test until they reached their maximal heart rate and are unable to continue, which corresponds to number 10 on the Borg scale.

The theoretical maximum heart rate has been computed using the Tanaka formula $208 - 0.7 \times \text{age}$. This formula is based on a large meta-analysis conducted on healthy adults' populations and has been demonstrated to show promising accuracy compared to the commonly used formula $220 - \text{age}$, proposed by Fox et al. [51]. Moreover, the latter has been shown to overestimate HR_{max} in young adults and underestimate HR_{max} in older adults [51]. Computing the theoretical HR_{max} provided a reference value, which made it possible to verify that the maximal heart rates reached by participants were consistent with the values expected for their age [16]. For additional verification, we applied the criterion of reaching at least 85% of the theoretical HR_{max}, as reported in previous studies [16].

Step	Speed (km/h)	Incline (%)	Estimated METs
1	4.8	2	4.1
2	5.47	6	6.4
3	5.47	10	8.3
4	5.47	14	10.1
5	5.47	18	12
6	5.47	22	13.8
7	6.76	22	15.7
8	6.76	25	17.1
9	9	25	19

Table 3.3: Table of speed, incline and estimated METs for each step [16]

3.5 Data acquisition

Sensors collecting heart rate and movement data were connected to the Trigno Discover software. This platform allowed real-time monitoring and recording of signals from each sensor. For the IMU sensors, the sampling frequency was set to 307.37 Hz and data were recorded in three axes: vertical, mediolateral and anteroposterior. The accelerometers were set to measure movements with a sensitivity of 16g to record the acceleration and the gyroscopes were set to 2000 degrees per second to record the angular velocity. The heart rate was measured at a sampling frequency of approximately 4 to 5 Hz.

Each daily life physical activity was recorded for a duration of exactly 180.9 seconds. The data collection process began only when the participant had started the activity, in order to avoid capturing transitional phases, such as incorrect speed or sedentary behavior at the beginning. No data were recorded during resting periods between activities. During the treadmill graded test, data were recorded for 120.9

seconds at each step, with a new recording starting at each new step.

The collected raw data were then exported from Trigno Discover in .csv format for processing and analysis. A .csv file was generated for each daily life activity as well as for each step of the treadmill test. Each file contained accelerometer (x, y, z-axes) and gyroscope (x, y, z-axes) data from the wrist and foot placement, as well as heart rate. A unique identifier was assigned to each participant, and a specific identifier was used for each activity to facilitate data processing.

The performance of each IMU sensor placement was evaluated individually during the analysis stage to determine their effectiveness for activity intensity classification. Although two IMU sensors are used, their data are not combined, as using multiple sensors simultaneously could increase participant burden and reduce comfort [52]. Therefore, the study focused on examining each placement separately, with the aim of identifying the most effective single-sensor location.

3.6 Data processing

In order to structure the data, the .csv files corresponding to each daily life activity for each participant were analyzed to separate the raw data from the wrist sensor, the foot sensor and the heart rate sensor. The raw signals were then segmented into fixed-length time windows and features were subsequently extracted from each segment. As heart rate data were not used as an input variable for the machine learning models, they were excluded from these feature engineering steps. Finally, different datasets were created for classification tasks based on absolute and relative physical activity intensities. All data processing and feature extraction were performed using Python. The following subsections describe each step of this data processing.

3.6.1 Segmentation

Data segmentation consists of dividing raw signals into time segments to facilitate data management, analysis and interpretation [4]. Dividing the data into shorter, manageable time windows helps to approximate stationarity within each segment, allowing for the extraction of stable and representative features. This process also facilitates the training of machine learning models by providing more instances for learning. These segmented windows can be either overlapping or non-overlapping. In their review, Farrahi et al. (2018), reported that majority of the studies they analyzed for estimating energy expenditure use fixed-size, non-overlapping windows, typically ranging from 4 to 60 seconds for accelerometer data[32].

To explore the impact of segment sizes on model performance, we evaluated six window sizes: 10, 15, 20, 30 and 60 seconds using raw data from the wrist and the foot. These values were selected to cover a broad range of temporal resolutions, from short to longer segments [32]. This analysis was performed separately for each sensor, as wearable sensor placement has a direct impact on the recorded signal, due to different movement dynamics [52].

3.6.2 Feature extraction

Once the raw sensor data had been separated and segmented, each sensor was processed separately. To transform the raw data from the IMU sensors into more meaningful features, a feature extraction step was performed [28]. Indeed, this reduced the complexity and dimensionality of the raw data and extracted relevant information, making it easier to use in classification algorithms. The data were not filtered prior to feature extraction to avoid any loss of potentially useful information.

The temporal features extracted from IMU sensors were selected based on recent trends observed in the literature, which have shown promising results in classifying physical activity intensity and predicting energy expenditure [27, 31, 33]. These include features reflecting signal variation around the mean (such as standard deviation and variance), as well as measures describing the relationship between axes (such as cross-correlation) and metrics reflecting how the signal changes (such as zero crossings and peak-to-peak amplitude). Similarly, the selection of frequency features was based on those most commonly reported in the literature [31, 33]. Frequency features were extracted using the Fast Fourier Transform and were computed across the entire frequency spectrum [33].

As accelerometers are more widely used, the most frequently used accelerometer features have also been applied to the gyroscope data. For each segment, feature extraction was performed on the x, y, and z axes of the accelerometer and the gyroscope. The extracted features are listed in Tables 3.4 and 3.5.

Time features	Definitions
Sum	Total sum of all signal values.
Mean	The sum of all values of the signal divided by the number of values.
Standard deviation	Measures how the signal values vary around the mean
Variance	The square of the standard deviation.
Coefficient of variation	The ratio between the standard deviation and the mean.
10^{th} percentile	Represents the value below which 10% of the signal values fall.
25^{th} percentile	Represents the value below which 25% of the signal values fall.
50^{th} percentile	Represents the value that divides the signal into two equal halves.
75^{th} percentile	Represents the value below which 75% of the signal values fall.
90^{th} percentile	Represents the value below which 90% of the signal values fall.
Skewness	Quantifies the asymmetry of the signal probability distribution.
Kurtosis	Measures the tailedness of a distribution. It characterizes the presence of outliers in the signal's distribution.
Maximum	Largest value observed in the signal over the analysis window.
Minimum	Smallest value observed in the signal over the analysis window.
Peak to peak amplitude	Represents the difference between the maximum and minimum values.
Sum of signal power	Represents the sum of the squared values of the signal.
Lag 1 autocorrelation	Measures the correlation between each value in the signal and the value immediately preceding it.
Logarithmic energy	Represents the logarithm of the sum of the squared signal values.
Inter quartile range	Represents the difference between the 75^{th} percentile and the 25^{th} percentile.
Zero-crossing	The number of times the signal changes from positive to negative or from negative to positive in the analysis window.
Cross-correlation XY	Measures the linear relationship between the signals on the x and y axes.
Cross-correlation XZ	Measures the linear relationship between the signals on the x and z axes.
Cross-correlation YZ	Measures the linear relationship between the signals on the y and z axes.

Table 3.4: List of features extracted in the time domain [33]

Frequency features	Definitions
Dominant frequency	The frequency component with the highest power in the signal's spectrum.
Power of dominant frequency	Represents the power associated with the dominant frequency component.
Spectral entropy	Measures the peakiness of the spectrum.

Table 3.5: List of features extracted in the frequency domain [33, 34]

3.6.3 Datasets creation

Both absolute and relative intensity classification were studied. For both classification tasks, datasets were created to allow testing two classification schemes: multiclass classification and binary classification.

The multiclass classification refers to classifying the activity intensity into the original four classes (sedentary, light, moderate and vigorous). This approach allows for a detailed representation of physical activity intensity. The binary classification groups sedentary and light intensities into one class and moderate and vigorous into another. This distinction aligns with the WHO guidelines on physical activity and sedentary behavior, which emphasize moderate-to-vigorous physical activity as the primary contributor to health benefits [1, 3]. Indeed, most WHO recommendations focus on reaching a minimum amount of moderate-to-vigorous activity for health gains [1].

In total, eight datasets were created and analyzed, corresponding to the four classification schemes applied separately to wrist and foot sensor data. All analyses were performed individually for each dataset.

Datasets to classify absolute intensity

The extracted features were saved in new .csv files for classification. Wrist and foot data were separated into two different files for each participant. This resulted in one .csv file per location, containing the participant codes, activity codes and the extracted features.

For the multiclass classification, physical activity intensities were assigned by associating each activity with an intensity category (sedentary, light, moderate or vigorous) based on the classification presented in Subsection 3.4.2. As the aim was to classify activity intensity into different categories, each activity was linked to an intensity class, not to a precise MET value.

Additionally, for binary classification, separate datasets were created combining the sedentary and light activities into one class, and the moderate and vigorous into another.

Although participant and activity codes are included for dataset organization and labeling, they are not used as input features in the machine learning models, which rely solely on features extracted from IMU sensors.

Datasets to classify relative intensity

As with the absolute intensity datasets, the extracted features were saved in .csv files for classification, and wrist and foot data were separated into two different files for each participant. However, the relative intensity files required more detailed processing, as they had to be assessed individually for each participant. To achieve this, the percentage of maximal heart rate, which is a well-established method for measuring relative intensity, was used [2, 3, 16]. It was calculated for each participant using the following formula

$$\%\text{HR}_{\text{max}} = \frac{\text{HR}_{\text{measured}}}{\text{HR}_{\text{max}}} \times 100$$

The HR_{max} is the maximal heart rate and refers to the final heart rate recorded at the moment the participant stopped the treadmill graded test. The measured heart rate, $\text{HR}_{\text{measured}}$, corresponds to the average heart rate over the last 15 seconds of each activity. This time window was selected to ensure that heart measurements reflected a stabilized physiological response to the activity. Once the $\%\text{HR}_{\text{max}}$ had been calculated for each participant, it was used exclusively as a method for labeling the relative intensity of each activity. The thresholds based of $\%\text{HR}_{\text{max}}$ defining sedentary, light, moderate and vigorous activity were based on the threshold values proposed by Norton et al. (2010), summarized in Table 3.6 [49]. This provides the ground truth for the classification task. Both binary and multiclass datasets were created for analysis.

Intensity category	**HR_{max} thresholds**
Sedentary	$\%\text{HR}_{\text{max}} < 40\%$
Light	$40\% \leq \%\text{HR}_{\text{max}} \leq 55\%$
Moderate	$55\% < \%\text{HR}_{\text{max}} \leq 70\%$
Vigorous	$\%\text{HR}_{\text{max}} > 70\%$

Table 3.6: HR_{max} thresholds used to defined intensity categories [49]

The extracted features intended to be used as input data for classification models are exclusively from IMU sensors. While informative, these signals alone may not fully account for inter-individual differences in the perception of effort. This is suggested by previous studies which, when using IMU signals to predict relative intensity, also incorporated individual characteristics such as age, height or weight or used relative intensity measures like $\%\text{HR}_{\text{max}}$ or RPE to determine personalized thresholds [16, 24, 40].

To be able to predict relative intensity, we introduced an individual-specific reference into the datasets. For each participant, features extracted from all segments of their *household walking* activity were aggregated by computing the mean. The *household walking* activity has been chosen as the reference because, although the path was the same for all participants, each individual completed it at their own pace. We chose to compute the mean because using the extracted features from a single segment of *household walking* activity could potentially introduce bias if it contained irregular or poorly performed activity. The aggregated mean features were then appended to each activity performed by that participant, including the *household walking* activity itself.

The aim of this approach is to implicitly provide the model with a point of comparison. By including this reference sample, which is representative of an effort specific to each participant, the model may have an additional element to interpret the other IMU signals and potentially determine relative intensity more accurately.

This resulted in one .csv file per location, containing the participant codes, activity codes, the extracted features, the individual-specific reference features, and the corresponding intensity labels. Additionally, mean heart rate over the last 15 seconds, the HR_{max} and the calculated $\%\text{HR}_{\text{max}}$ for each activity were included. These variables, along with participant codes and activity codes, were not used as input features in the machine learning models.

3.7 Preparation of the data and validation techniques

3.7.1 Separation and validation

Our datasets are composed of segmented signals from twenty participants, each performing seven activities. When splitting the data into a training and testing set, it is crucial to ensure that the same participant is not included in both sets. Indeed, the signals may reflect individual characteristics and if a participant is included in both sets, the models may learn characteristics specific to that person rather than patterns that can be generalized to the whole population [31].

In supervised learning, model evaluation requires testing on data that were not used during training to ensure generalization to unseen data. A basic approach is to divide the dataset into two parts: one for training and the other for testing [28]. However, performance can be highly dependent on the specific split, which can potentially lead to high variance in results [28]. To address this, cross-validation techniques perform multiple evaluations across the dataset using a resampling procedure. These approaches provide a more robust and reliable performance estimation by reducing variability caused by data partitioning [28]. Among these methods, k-fold and Leave-One-Out (LOO) cross-validation are commonly used. K-fold cross-validation consists of dividing the data randomly into k parts of equal size, where k is an arbitrary number. At each iteration, the model is trained on k-1 folds and evaluated on the remaining fold. In this study, we used LOO, which is a variation of k-fold where k equals the number of samples. It consists of iteratively using one sample as the test set, while the model is trained on the remaining samples, ensuring that each sample serves exactly once as the test set (as illustrated in Figure 3.3). In our case, k is the number of participants, so each iteration tests the model on a single participant while training on the remaining participants.

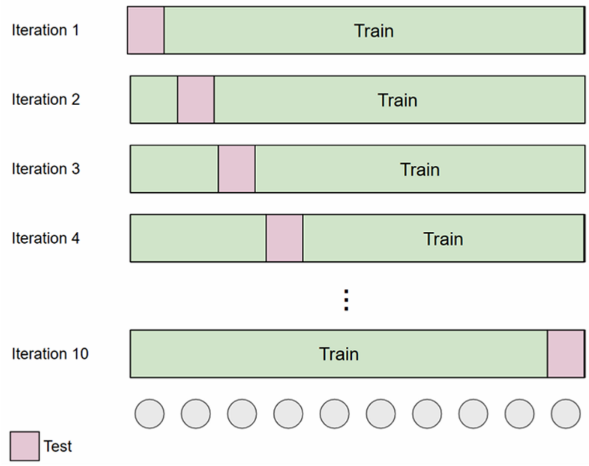


Figure 3.3: Representation of a LOO cross-validation [53]

Prior to model development, multiple datasets were generated using different segmentation window sizes. Each dataset was initially split into a training set (70%) and a testing set (30%). The test set remained untouched throughout the entire process to ensure an unbiased final evaluation. To determine the optimal window size, only the training set was used, applying LOO cross-validation. Once the window size yielding the best performance was identified, the corresponding dataset was retained for subsequent steps. Next, hyperparameter optimization was conducted exclusively on the same 70% training set. A more detailed explanation of the hyperparameter optimization process is provided in Subsection 3.9.5. Finally, after selecting the best model configuration, the unseen 30% test set was used to evaluate the final models' performance.

The datasets were split into training and testing using the `train_test_split` function from Scikit-learn, with `random_state` fixed at 42, to ensure reproducibility and that the same sets are obtained every time. The `LeaveOneGroupOut` cross-validation method of Scikit-learn library was implemented. It allows us to split the data while ensuring that the same individual never appears in the training and test sets simultaneously.

3.7.2 Standardization

The accelerometer and gyroscope datasets were standardized using the Z-score method, which involved transforming each value so that it has a mean of 0 and a standard deviation of 1, according to the formula:

$$z = \frac{(x - \mu)}{\sigma}$$

where μ represents the mean and σ the standard deviation [54].

For implementation, `StandardScaler()` from the Scikit-learn library was used. Standardization was applied after splitting the data into training and testing sets to prevent data leakage.

This transformation is essential in certain machine learning algorithms, as it prevents features with larger numerical ranges from dominating the model [53, 54]. It ensures that all features contribute more equally to the learning process. More detailed explanations on the importance of standardization in algorithms used in this study are provided in Section 3.8.

3.8 Choice of machine learning model

Our goal was to directly classify activities according to their intensity category, and therefore supervised classification algorithms were used. Three different models were developed. Choosing several approaches allows us to compare their respective performances, with the aim of identifying the one that offered the best results. The selected models are Random Forest, Support Vector Machine and Artificial Neural Network, which are supervised algorithms commonly used in for the classification of physical activity intensities [31, 48].

3.8.1 Random Forest model

Random Forest (RF) is an ensemble algorithm widely used for physical activity intensity classification or prediction [27, 31, 55]. It trains an ensemble of decision tree models, where a decision tree is a predictive model that uses a tree structure to represent the data. A decision tree recursively divides the data into subgroups according to the values of the input variables, with the aim of forming groups that are as homogeneous as possible with respect to the target. In a random forest, each tree is trained independently on a random sample drawn with replacement (called a bootstrap sample), and at each node split, only a random subset of features is considered [28]. The final model prediction is obtained by majority vote across the trees, which reduces overfitting and improves generalization capability compared with a single tree [28].

In this study, the `RandomForestClassifier` implementation from the Scikit-learn library in Python was used.

3.8.2 Support Vector Machine model

The principle of Support Vector Machine (SVM) is to construct decision boundaries, which are called hyperplanes, that optimally separate samples into predefined class categories. These hyperplanes are positioned to maximize the distance between classes by maximizing the distance between the closest data points of each class, which are known as support vectors. The support vectors define the decision boundary and ensure maximal separation between classes. When a new observation is introduced, it is projected into the multidimensional feature space and its class is predicted based on which side of the hyperplane it lies [28, 38]. SVM can be particularly effective when dealing with high-dimensional data, which is often the case in the analysis physical activity behaviors [28].

As explained in Subsection 3.7.2, the input features were standardized before training. This step is particularly important for SVM, as the objective function assumes that features are centered and have a similar variance [54]. Without standardization, features with larger variance could dominate the model and impact its learning performance.

The SVM model was implemented using `SVC` classifier from the Scikit-learn library.

3.8.3 Artificial Neural Network model

Artificial neural networks are inspired by biological neural networks and is used to model complex relationships between inputs and outputs. They are composed of interconnected neurons organized into several layers: an input layer, one or more hidden layers and an output layer. The input layer receives the data, which are then processed through the network. Each neuron in the hidden layers receives multiple inputs, multiplies them by corresponding weights that reflect their importance, and computes a weighted sum. The sum is passed through an activation function, which enables the neuron to produce an output that is transmitted to the next layer.

In this study, we used a feedforward neural network, which has demonstrated strong performance in similar tasks [17, 29, 36, 56]. In the feedforward neural network, information flows in a single direction, from the input layer through the hidden layers to the output layer [36]. Each neuron receives signals only from the immediately preceding layer. Specifically, we implemented a multilayer perceptron (MLP), which is a type of feedforward neural network composed of fully connected neurons with non-linear activation functions [57]. Figure 3.4 illustrates the architecture of an MLP. The MLP was selected mainly for its simplicity and ease of implementation in Python using standard libraries.

As in the other models, standardized input features were used. Indeed, in the case of neural networks, input scaling directly affects the scaling of the weights in the first layer. Standardizing facilitates weight optimization and leads to more stable and efficient regularization during learning [58].

We implemented an MLP using the `MLPClassifier` from the Scikit-learn library.

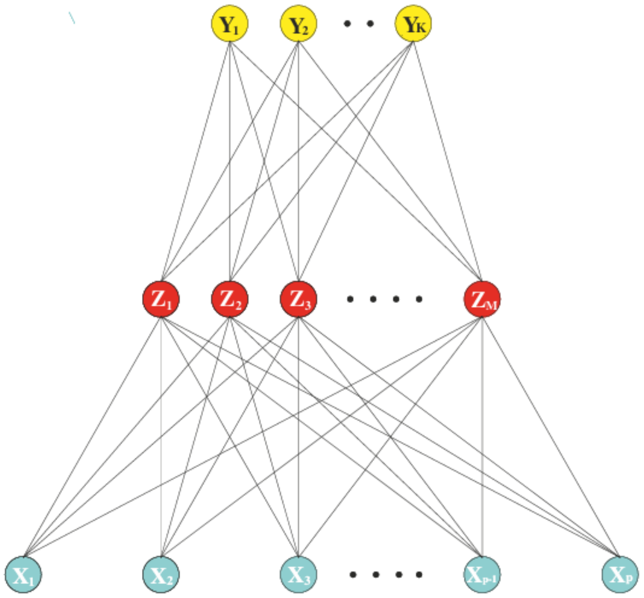


Figure 3.4: Representation of an MLP [58]

3.9 Hyperparameters and optimization

Machine learning algorithms require the definition of several hyperparameters, which are parameters that directly influence model behavior and performance [28]. Proper adjustment of the hyperparameters aims to maximize performance while avoiding overfitting [53]. Overfitting is a common issue in machine learning and occurs when the model learns the training data too well, leading to a lack of generalization. Underfitting can also be observed and occurs when the model is too simple to capture important patterns, which also lead to poor generalization.

In the following subsections, the main hyperparameters of the three classifiers are presented and the method used for their optimization is described [29, 53].

3.9.1 Random Forest hyperparameters

In the RandomForestClassifier, the main hyperparameters include [29, 53] :

- `n_estimators`: This specifies the number of trees in the forest. A higher number of trees typically improves the model's performance but increase the computational time.
- `max_depth`: This represents the maximum tree depth. Lower values prevent overfitting, but values that are too low may lead to underfitting problems. A high value requires more computational time.
- `min_samples_split`: This specifies the minimum number of samples required to split a node. The higher the value, the more effectively overfitting is prevented.
- `min_samples_leaf`: This defines the minimum number of samples in a leaf. A higher value typically prevents overfitting.

3.9.2 Support Vector Machine hyperparameters

In the SVC model, the most important hyperparameters include [53]:

- `kernel`: This function transforms the input data into a higher-dimensional space, enabling the model to find a linear decision boundary to separate the classes, even if the data are not linearly separable in the original space.
- `C`: This is the regularization parameter, which controls the trade-off between achieving a low training error and keeping the model simple. A high value of C makes the model focus on correctly classifying all training examples. In contrast, a low C allows the model to make some errors in exchange for better generalization to unseen data and reducing the risk of overfitting.
- `gamma`: : The kernel coefficient for non-linear kernels. It defines the influence of a single training example. A low value means wider influence, while a high value limits the influence to nearby points.

3.9.3 Artificial Neural Network hyperparameters

The MLPClassier also requires the definition of several hyperparameters, which include [53]:

- `learning_rate_init`: This controls the step size the optimizer takes to adjust the model during training. A smaller learning rate will lead to higher chances of finding an optimum solution, however, it increase computational time.
- `hidden_layer_sizes`: This controls the number of layers and the number of nodes within each layer. The higher the number of layers or the number of nodes, the higher the complexity of the model, and therefore the higher the chance of overfitting.
- `activation`: This refers to the non-linear transformation function applied to a neuron's output. It determines how much signal is transmitted to the next layer based on the input it received.
- `alpha`: This hyperparameter helps to prevent overfitting.

3.9.4 Common hyperparameters across models

In addition to hyperparameters specific to each model, the parameter `random_state` was fixed at 42 for each classifier. It represents the random seed number to ensure that the model produces the same samples every time, which is crucial for reproducibility [53]. Moreover, for the RF and SVM models, the `class_weight` hyperparameter was set to `'balanced'`, which manages class imbalances by automatically assigning a weight to each class according to its frequency in the input data [53]. Higher weights are given to classes that have fewer samples.

Class imbalance is a relevant aspect to consider when examining the datasets. Indeed, in the dataset used for determining absolute intensity, the sedentary class is underrepresented compared to the other classes, which occur twice as often per participant. Using `class_weight = ‘balanced’` ensures the model accounts for the minority class during training. Moreover, the datasets created to classify the relative intensity may also be affected by class imbalance.

3.9.5 Hyperparameters optimization

For each model, hyperparameters were initially set to fixed values, either by using default settings or simple choices to serve as a baseline. These settings were applied during the process of determining the optimal window size across the datasets. The parameters are provided in Appendix B: Initial hyperparameters fixed for each model.

Once the optimal window size found, hyperparameters optimization was performed for each classification model and tasks using a grid search approach. For each hyperparameter, a list of values was defined, and the grid search procedure evaluated all possible combinations of hyperparameters within this defined space [53]. Each combination was then evaluated using cross-validation on the training data, in order to estimate its generalization performance. The cross-validation method applied during grid search was LOO cross-validation, which ensures that each participant is used once as a validation set. The combination of hyperparameters that achieved the best performance during cross-validation was selected. This combination was then used to train the final models on the full training set [53]. Their final performance was then evaluated on the 30% unseen test set.

The complete list of tested hyperparameter values for each classification model is provided in Appendix C: Grid search spaces for each machine learning model.

This approach was implemented using the `GridSearchCV` class from the Scikit-Learn library using the `LeaveOneGroupeOut` from Scikit-learn as cross-validation approach.

3.10 Evaluation metrics

Once the model has been trained, it must be evaluated using evaluation metrics to ensure its reliability. In our approach, we selected the macro F1-score as our primary evaluation metric for both determining the optimal window size and comparing different sets of hyperparameters. This metric is particularly well-suited for classification tasks involving imbalanced classes, as it assigns equal importance to each class, regardless of its frequency in the dataset.

The macro F1-score is based on the F1-score, which is the harmonic mean of precision and recall, and is calculated as follows

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Precision measures the proportion of correct predictions among all instances that the model has identified as belonging to the target classes, and is defined by

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive + False positive}}$$

Recall assesses the ability to correctly identify all actual instances of the target class:

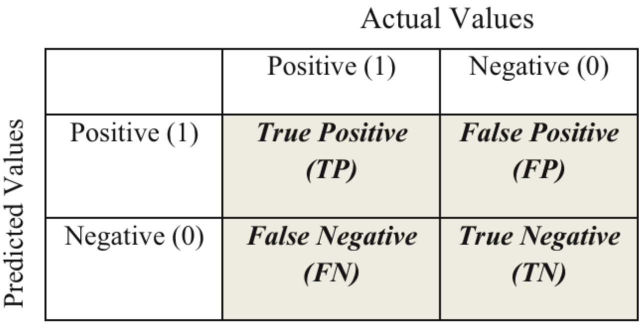
$$\text{Recall} = \frac{\text{True positive}}{\text{True positive + False negative}}$$

To obtain the macro F1-score, the F1-score is first computed individually for each class. The final score is then derived by taking the unweighted average across all classes:

$$F1_{\text{macro}} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

where N is the total number of classes and $F1_i$ is the F1-score for class i.

The final supervised models' performance will be evaluated using normalized confusion matrices, which compare model predictions with actual values, and is illustrated in Figure 3.5. For the multiclass classification, the intensity classes are abbreviated as follows: *sedentary = sed, light = light, moderate = mod, vigorous = vig* . For binary classification, the classes are abbreviated as: *sedentary-light = sed-light and moderate-vigorous = mod-vig*



Actual Values

Predicted Values

	Positive (1)	Negative (0)
Positive (1)	***True Positive (TP)***	***False Positive (FP)***
Negative (0)	***False Negative (FN)***	***True Negative (TN)***

Figure 3.5: Representation of a confusion matrix [59]

In addition to the confusion matrices, three metrics are employed, the macro F1-score, the balanced accuracy and the Cohen's Kappa.

While accuracy is a commonly used metric in physical activity intensity classification or prediction studies, it can be misleading when dealing with imbalanced classes. Accuracy measures the overall proportion of correctly classified observations, but in the presence of class imbalance, it may overestimate performance. For example, in a dataset where one class is much more represented than others, a model that always predicts this class would achieve high accuracy without performing well on the minority classes. To address this limitation, we used balanced accuracy, which provides a more reliable evaluation in the presence of class imbalance. It is computed as the average of the recall obtained for each class, ensuring equal consideration of all classes.

$$\text{Balanced accuracy} = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}$$

where N is the total number of classes and TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

Cohen's Kappa is commonly used in the context of physical activity intensity classification or prediction [17, 27, 31, 38]. This metric is particularly relevant when dealing with imbalanced datasets. Indeed, Cohen's Kappa measures how closely the predictions of a machine learning algorithm align with the ground truth labels, beyond what would be expected from random guessing based on class distribution. In other words, it ensures that the overall prediction accuracy is not obtained only by chance. The formula for Cohen's Kappa is defined by

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

where P_0 is the observed outcome class, and P_e is the expected outcome class [59].

For interpreting the Cohen's Kappa values, we followed the ratings used in other studies predicting the intensity of physical activity, which are based on the guidelines proposed by Landis and Koch (1979) [17, 27, 38]. The categorization of Cohen's Kappa values is summarized in Table 3.7

Cohen's Kappa	**Strength of agreement**
< 0.00	poor
0.00-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect

Table 3.7: Interpretation of Cohen's Kappa values

3.11 Features importance analysis in a best-performing case

As part of an exploratory approach, feature importance was assessed using the RF model, as it provides intrinsic measures of feature relevance. This analysis focused on a well-performing case for relative intensity classification, which represents the main focus of our study. This evaluation aimed to better understand which features contributed most significantly to the model's performance.

This analysis was performed using the `feature_importances_` attribute of the RF implementation from Scikit-Learn library.

Chapter 4

Results and discussions

This fourth chapter presents and interprets the main findings of this study. It begins with an overview of the relative intensity datasets, focusing on how activities are distributed across the four intensity levels. Next, the impact of window size on model performance is assessed by segmenting the datasets into fixed-length windows of 10, 15, 20, 30, and 60 seconds. This analysis is performed exclusively on the training set using a leave-one-out cross-validation approach, with the macro F1-score used as the evaluation metric. Following the selection of the optimal window size, hyperparameter optimization is performed for each model, sensor location, and classification task using a grid search approach with a leave-one-out cross-validation, again based solely on the training data and evaluated using the macro F1-score. Once the models are tuned, their final performances are evaluated on the unseen test set. For each classification task and each sensor placement, the models are evaluated and interpreted, starting with relative classification, followed by absolute intensity classification. The performances of wrist and foot data are then compared within each task to assess the influence of the sensor placement. Finally, a comparison between relative and absolute intensity classification results is conducted to highlight the differences between performances. In addition to evaluating model performance across tasks and sensor placement, an exploratory feature importance analysis is conducted of one of the best-performing models for relative intensity classification.

4.1 Characteristics of the relative intensity datasets

4.1.1 Distribution of relative intensity across activities

Each participant's activities were classified into relative intensity categories based on established $\%HR_{\text{max}}$ thresholds [49]. Since intensity is relative to each individual, it is expected that the same activity may be experienced differently from one person to another. This variability is clearly reflected in the distribution patterns shown in Table 4.1, which summarizes how the various activities are distributed across the four intensity levels. Indeed, activities are not consistently classified within the same intensity categories across participants. These findings are consistent with established guidelines and existing literature, which highlight that the perceived intensity of a given activity can vary considerably between individuals [2, 3, 12].

The activity *looking at a smartphone* was predominantly classified as sedentary, with the rest classified as light intensity. This outcome aligns with expectations, as the activity involves minimal physical effort.

The activities, *household walking*, *cooking* and *treadmill walking* were primarily classified as light intensity. *Sweeping*, although mostly categorized as moderate, was classified as light intensity for eight participants. This variability may be attributed to the less structured nature of the task. The participants were instructed to sweep quickly, but each interpreted and performed this instruction at their own pace. In addition, movement patterns were influenced by individual interpretation.

Finally, the activity *stairs* was mainly classified as moderate intensity, although it was perceived as light or vigorous for some participants. Similarly, *treadmill running* appeared in both the moderate and vigorous intensity categories.

Activity	**Sedentary**	**Light**	**Moderate**	**Vigorous**
Looking at a smartphone	13	7	0	0
Household walking	1	16	2	0
Cooking	2	17	1	0
Treadmill walking	0	16	4	0
Sweeping	1	8	11	0
Stairs	0	2	13	5
Treadmill running	0	0	6	14

Table 4.1: Distribution of relative intensity levels per activity

4.1.2 Distribution of relative intensity across participants

When examining the distribution of intensity labels across participants, we observed that not every participant covered all intensity categories. Among the twenty individuals, fourteen were missing one class. In most cases, the missing category was either sedentary or vigorous, and only for one participant the moderate category was absent.

The vigorous category was missing among participants who were more physically active. This is consistent with the results of the treadmill graded test, where these individuals all reached stage 7, suggesting a generally lower perceived exertion for activities that are objectively intense. Conversely, the absence of the sedentary category was more variable and seemed to depend on each participant.

4.1.3 Influence of participant characteristics

The distribution observed in Table 4.1 can also be partly explained by the characteristics of the study population. While the protocol was initially designed to include older participants, only young adults were recruited due to time constraints. The activities had been selected so that both age groups would perform exactly the same ones. We chose activities that fell into the same intensity category for both age groups, even though the associated MET values differ depending on age. Despite this, the activities selected for both populations were retained. However, for young adults, the METs for these activities generally fall toward the lower end of the ranges corresponding to the intensity categories, especially for the moderate and vigorous categories.

Additionally, the participants were healthy, and most reported engaging in regular physical activity during the week. It is therefore expected that the activities felt less intense for them than they would have for an older or more sedentary population. For example, this likely led to a greater number of activities being classified as light intensity.

4.2 Optimization of window size

4.2.1 Selection of window size

Given that the classification of relative intensity is our main focus, our decision is primarily guided by the results presented in Figure 4.1 for the multiclass classification of relative intensity and Figure 4.2 for the binary classification of relative intensity. Since we use leave-one-out cross-validation, we rely on the average macro F1-score across folds to identify window sizes. It is important to note that, because the folds involve different participants, macro F1-scores vary between folds, reflecting expected inter-individual differences. The goal is to determine a window size that provides the best performance trade-off for each sensor on these tasks, while also maintaining strong robustness on absolute intensity tasks.

For the wrist sensor, a window of 30 seconds is selected. In the multiclass relative intensity task shown in Figure 4.1, this window size yields the highest performance with the SVM model. Although the RF model achieves the best macro F1-score of 0.603 with a 15-second window, the 30-second configuration yields a macro F1-score of 0.546, which is close to optimal. For the MLP model, while the peak performance is observed at 10 seconds, the 30-second window remains within a competitive performance range. In the binary classification of relative intensity shown in Figure 4.2, the window of 30 seconds delivers the best results with the RF model, performs nearly optimally with the SVM model, and is among the best configurations for the MLP model.

For the absolute intensity classification tasks, Figures 4.3 and 4.4 show that all window sizes produce consistently strong results. Specifically, in multiclass classification, all macro F1-scores exceed 0.89, and for the binary classification task, all scores are above 0.9. Selecting the 30-second window appears to be a reliable choice, since it consistently delivers results that are close to the highest observed scores.

A similar analysis is conducted for the foot sensor, leading to the selection of a 30-second window. In the multiclass relative intensity task shown in Figure 4.1, this window size provides solid performance. Although it is not the maximum macro F1-score, it remains very close to the maximum for all models, particularly for RF and MLP. For the SVM model, even though the optimal score is 0.459 at 60 seconds, the performance at 30 seconds still achieves 0.409, which is the second-best score. This choice also represents a solid trade-off in the binary classification task shown in Figure 4.2, consistently yielding near-optimal scores across the RF and MLP models. As with the multiclass relative intensity task, the SVM model does not reach its peak performance at 30 seconds, but still remains close to its best.

Regarding absolute classification, the 30-second window also demonstrates strong robustness. In the multiclass classification shown in Figure 4.3, it remains very close to the best performance for SVM and MLP. While the RF model performs slightly better with a 60-second window, its performance at 30 seconds still falls within a reasonably high range. In the binary classification shown in Figure 4.4, the 30-second window achieves the best score with the SVM model, shows nearly identical results to the optimal score with the RF model, and remains competitive with the MLP model as well.

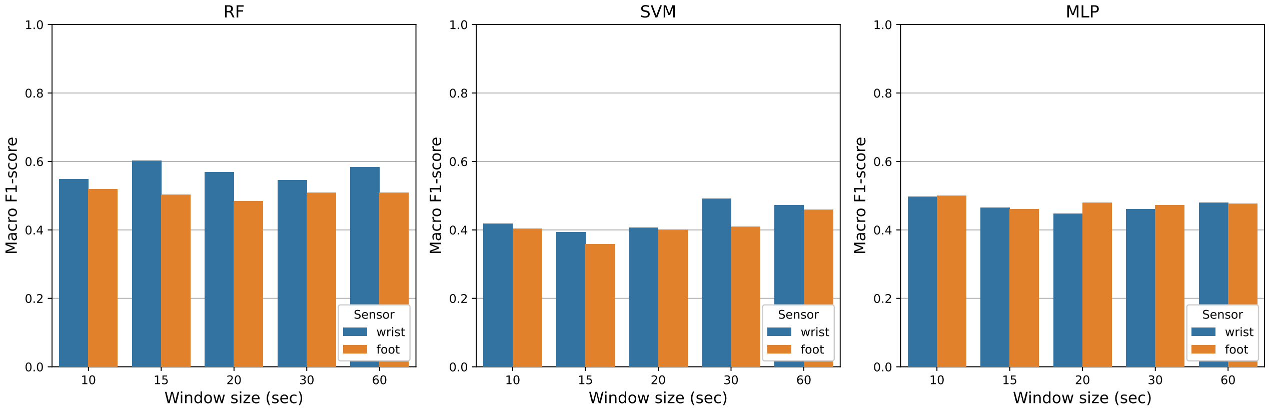


Figure 4.1: Macro F1-scores for the RF, SVM and MLP models for different window sizes - multiclass classification of relative intensity

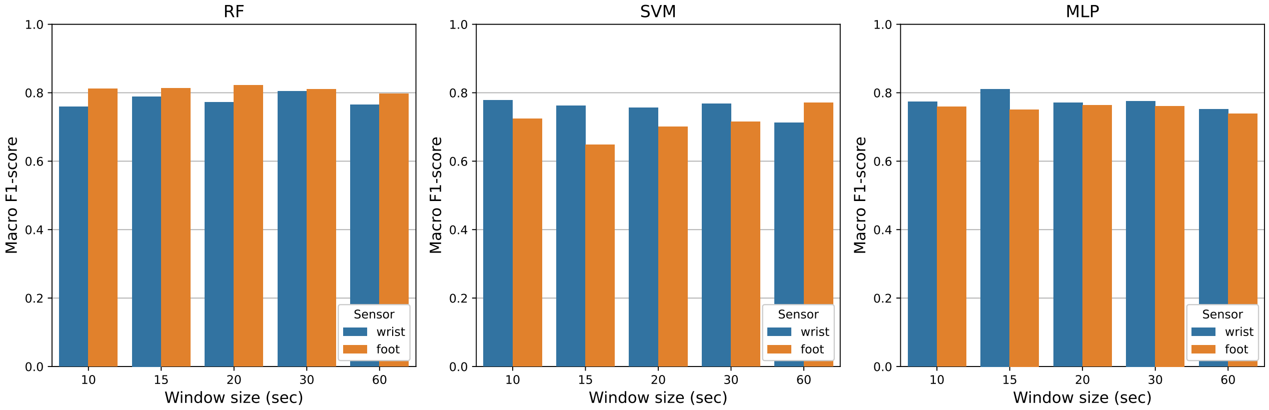


Figure 4.2: Macro F1-scores for the RF, SVM and MLP models for different window sizes - binary classification of relative intensity

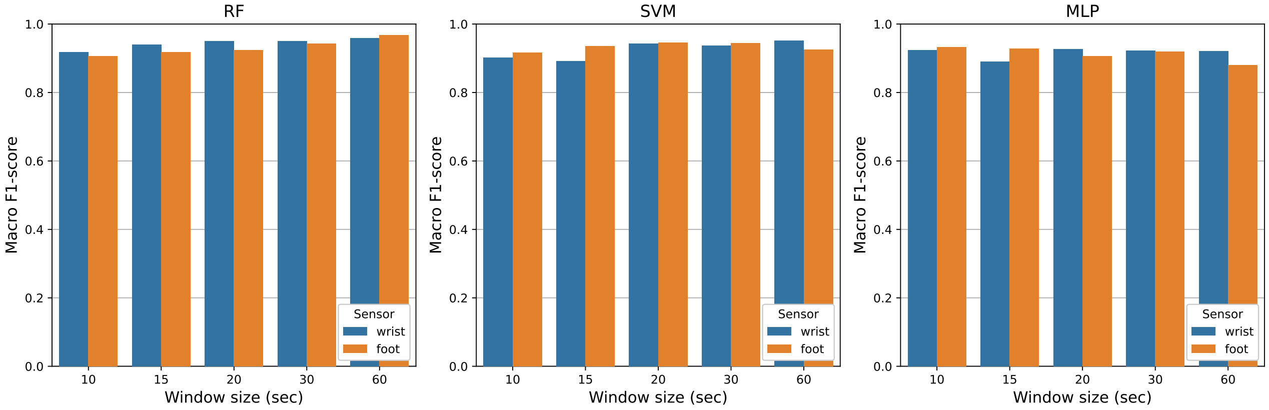


Figure 4.3: Macro F1-scores for the RF, SVM and MLP models for different window sizes - multiclass classification of absolute intensity

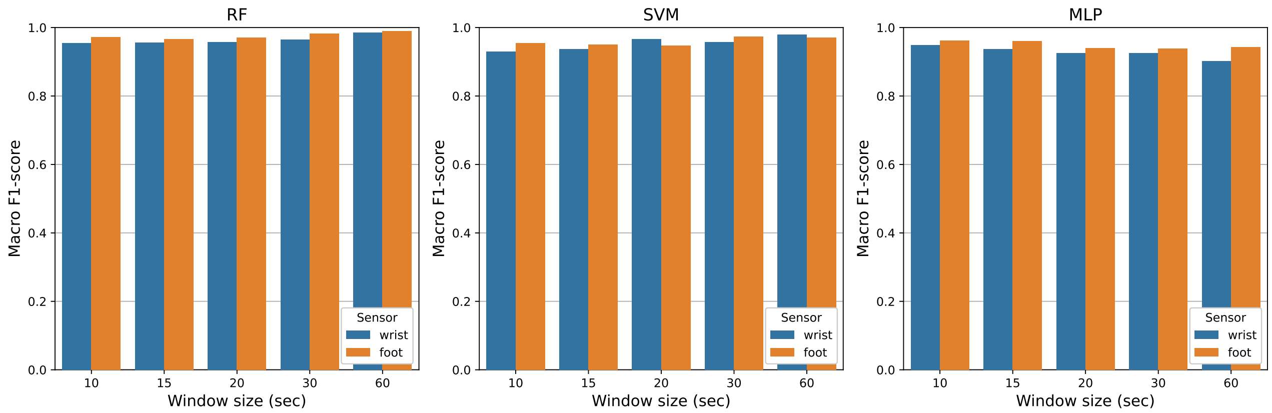


Figure 4.4: Macro F1-scores for the RF, SVM and MLP models for different window sizes - binary classification of absolute intensity

In general, for both sensors, the differences between window sizes remain limited, suggesting a relative stability in performance regardless of the chosen window size. Using the same window size for both sensors also facilitates direct comparison across their performance, as this ensures an equal number of segments for each sensor. Our choice to use 30-second windows for both wrist and foot sensors is consistent with segmentation practices in physical activity intensity research using accelerometers [32, 37]. In particular, Montoye et al. (2016) demonstrated that accelerometers worn on the wrist and thigh achieved high classification accuracy across physical activity intensity categories using a 30-second window, with sensitivities and specificities exceeding 90% [37].

The detailed macro F1-scores values for each model and window size are provided in Appendix D: Detailed macro F1-scores for window size optimization.

4.2.2 Influence of segmentation on datasets

The activity classes in the absolute intensity datasets were defined during dataset creation. As a result, an equal number of samples were expected for the light, moderate, and vigorous classes and half that number for the sedentary class.

The recordings were segmented into fixed non-overlapping windows of different sizes. Each activity was intended to last approximately 180.9 seconds, which should have produced the expected number of samples per class. However, one participant's light activity (*household walking*) was slightly shorter (179.9 seconds) due to an unintended recording error. As a result, the last window of that activity was incomplete and excluded, leaving one fewer sample in the light class for all window sizes, as illustrated in Table 4.2. This minor difference does not significantly impact the overall class balance, as the counts for the light, moderate, and vigorous classes remain roughly equal.

Datasets for binary classification were also created, containing 359 samples in the sedentary-light class and 480 samples in the moderate-vigorous class for a 30-second window size.

Intensity level	Number of samples
Sedentary	120
Light	239
Moderate	240
Vigorous	240

Table 4.2: Class distribution of absolute activity intensity classes for a 30-second window size

The relative intensity datasets were derived from the same segmented windows. Consequently, the missing segment from the *household walking* activity also results in one fewer instance being included in the relative intensity datasets. The number of instances per class is summarized in Table 4.3. The binary datasets are slightly more balanced but still show a disparity. For example, for the 30-second window size, the sedentary-light class has 503 samples, while the moderate-vigorous class has 336 samples

Intensity level	**Number of samples**
Sedentary	102
Light	401
Moderate	222
Vigorous	114

Table 4.3: Class distribution of relative activity intensity classes for a 30-second window size

4.3 Hyperparameters optimization

In this section, we present the optimal configurations selected for each model, according to the sensor location, task, and intensity type. Tables 4.4, 4.5, 4.6, and 4.7 summarize the optimized hyperparameters selected for each case.

Sensor	**Model**	**Hyperparameters**

Wrist	RF	`max_depth = 5; min_samples_leaf = 5;`
		`min_samples_split = 2; n_estimators = 100`

	SVM	`kernel = linear; C = 10; gamma = scale`

	MLP	`activation = relu; alpha = 0.0001;`
		`hidden_layer_sizes = (100, 50);`
		`learning_rate_init = 0.001`

Foot	RF	`max_depth = None; min_samples_leaf = 1;`
		`min_samples_split = 5; n_estimators = 100`

	SVM	`kernel = rbf; C = 10; gamma = scale`

	MLP	`activation = relu; alpha = 0.001;`
		`hidden_layer_sizes = (100, 50);`
		`learning_rate_init = 0.001`

Table 4.4: Optimized hyperparameters for relative intensity multiclass classification

Sensor	**Model**	**Hyperparameters**
Wrist	RF	`max_depth = None; min_samples_leaf = 1; min_samples_split = 10; n_estimators = 300`
	SVM	`kernel = rbf; C = 10; gamma = 0.001`
	MLP	`activation= relu; alpha =` 10^{-5}`; hidden_layer_sizes = (100,70); learning_rate_init = 0.001`
Foot	RF	`max_depth = None; min_samples_leaf = 2; min_samples_split = 5; n_estimators = 100`
	SVM	`kernel = rbf ; C = 1; gamma = scale`
	MLP	`activation = relu; alpha = 0.001; hidden_layer_sizes = (100, 50); learning_rate_init = 0.001`

Table 4.5: Optimized hyperparameters for relative intensity binary classification

Sensor	**Model**	**Hyperparameters**
Wrist	RF	`max_depth = None; min_samples_leaf = 1; min_samples_split = 2; n_estimators = 700`
	SVM	`kernel = rbf; C = 10; gamma = 0.001`
	MLP	`activation = tanh; alpha =` 10^{-5}`; hidden_layer_sizes = (100, 70); learning_rate_init = 0.001`
Foot	RF	`max_depth = None; min_samples_leaf = 2; min_samples_split = 2; n_estimators = 100`
	SVM	`kernel = rbf; C = 10; gamma = scale`
	MLP	`activation = relu; alpha =` 10^{-5}`; hidden_layer_sizes = (100,); learning_rate_init = 0.001`

Table 4.6: Optimized hyperparameters for absolute intensity multiclass classification

Sensor	Model	Hyperparameters

Wrist	RF	`max_depth = None; min_samples_leaf = 1;`
		`min_samples_split = 2; n_estimators = 700`

	SVM	`kernel = rbf; C = 10; gamma = 0.001`

	MLP	`activation = relu; alpha = 0.001;`
		`hidden_layer_sizes = (100, 50);`
		`learning_rate_init = 0.001`

Foot	RF	`max_depth = 5; min_samples_leaf = 1;`
		`min_samples_split = 10; n_estimators = 700`

	SVM	`kernel = rbf; C = 10; gamma = scale`

	MLP	`activation = relu; alpha =` 10^{-5}`;`
		`hidden_layer_sizes = (100, 70);`
		`learning_rate_init = 0.001`

Table 4.7: Optimized hyperparameters for absolute intensity binary classification

The process of hyperparameter optimization revealed that optimal model configurations were highly dependent on the context, varying significantly across sensor locations and classification tasks.

For RF models, a `max_depth` of `None`, which corresponds to an unlimited depth, was frequently selected, indicating that deeper trees were often beneficial. Regarding the number of trees (`n_estimators`), it varies considerably depending on the sensor and the task. For relative classifications, values range from 100 to 300, whereas for absolute tasks, higher values such as 700 were selected, particularly for the wrist sensor. Finally, parameters such as `min_samples_leaf` and `min_samples_split` varied depending on the configuration, without following any clear trend.

Regarding SVM, the `rbf` kernel was predominantly selected across most tasks suggesting that the decision boundaries for these tasks are non-linear. The `linear` kernel was chosen only for the relative intensity classification of the wrist sensor. Additionally, the `C` parameter was generally 10, except for one specific task and sensor where it was set to 1, reflecting differing levels of tolerance for misclassification. The `gamma` parameter is only taken into account when using `rbf` kernel and varies between 0.001 or `scale`.

For MLP, the optimized `hidden_layer_sizes` generally consisted of two hidden layers, with sizes varying depending on the task and sensor data. For example, the notation (100, 70) indicates two layers containing 100 and 70 neurons, respectively. The `ReLU` function was predominant as the optimal activation function. It activates a neuron if its input is positive, for negative inputs, the output is zero. In one case, `tanh` was chosen instead. Unlike `ReLu`, `tanh` produces values ranging from -1 to 1, with outputs centered around zero. The regularization parameter `alpha` showed different values configurations, and all models employed the same initial learning rate.

It is important to note that all configurations were obtained through a grid search, which means that only combinations of values explicitly defined within the search space were evaluated [53]. As such, a trade-off had to be made between the number of combinations explored and the computational cost [53].

4.4 Classification of relative intensity

4.4.1 Dataset characteristics

As previously explained, our datasets were highly imbalanced for the relative intensity multiclass classification tasks. Indeed, nearly three quarters of participants did not have samples for all activity classes. An analysis of the test set confirmed that while each class was represented, there was a strong imbalance in the number of samples per class. Notably, the sedentary class was significantly underrepresented compared to other classes, which is a critical factor to consider during model evaluation.

In this context, support, defined as the number of samples per class in the test set, highlights the class distribution imbalance. For this task, the support was 12 samples for sedentary, 132 for light, 72 for moderate, and 36 for vigorous activities. This class imbalance was taken into account to ensure a fair interpretation of model performance.

For the binary classification, grouping the classes into two broader categories helped to reduce this imbalance. The support is 144 samples for the sedentary-light class and 108 samples for the moderate-vigorous class.

4.4.2 Multiclass classification

Performance of the wrist sensor

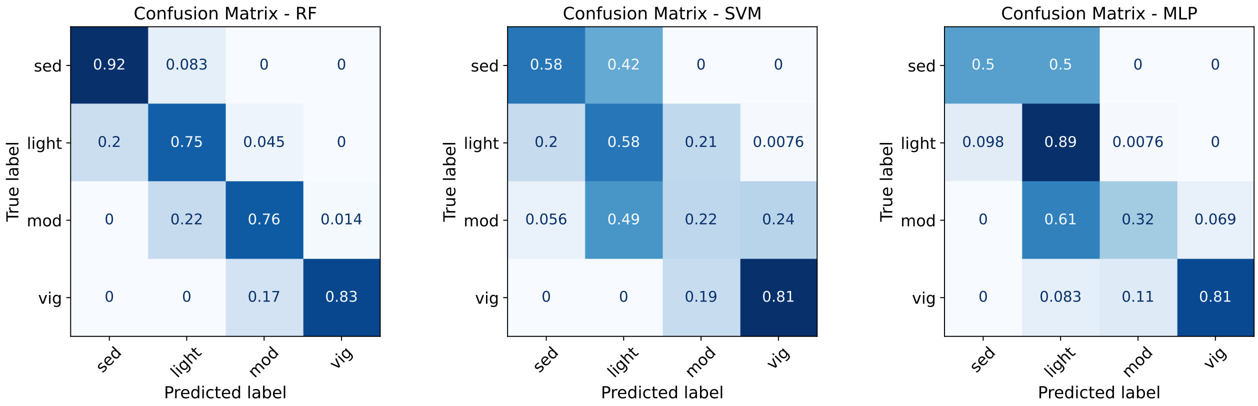
The results for the wrist sensor show varying levels of classification performance across models and intensity levels.

The confusion matrix in Figure 4.5a demonstrates that the RF model achieves particularly strong results for the sedentary class, with 92% of samples correctly classified, and achieves slightly lower performance for the vigorous class. However, its effectiveness is more limited for the light and moderate intensity classes, exhibiting confusion especially between adjacent classes.

On the other hand, Figure 4.5b shows that the SVM model recognizes the vigorous class relatively well, with 81% of the samples correctly classified. However, it shows significant confusion among the other three classes, particularly the moderate class, with only 22% of the instances correctly classified.

Figure 4.5c shows that the MLP model achieves strong performance for the light class, with 89% of samples correctly identified. It also classifies the vigorous intensity class reasonably well with 81% of instances correctly classified, which is comparable to the RF and SVM models. However, it struggles with the sedentary and moderate classes.

The overall difficulties of SVM and MLP models are further reflected in the lower balanced accuracy, macro F1-score, and Cohen's Kappa coefficients presented in Table 4.8, showing limited reliability. The Cohen's Kappa shows a fair agreement for SVM and a moderate agreement for MLP between predictions and actual labels. In contrast, the RF model achieves a fairly good average recall across all classes, and the macro F1-score suggests a good balance between precision and recall across all classes. The Cohen's Kappa coefficient shows a substantial agreement between predictions and ground truth.



(a) Random Forest (b) SVM (c) MLP

Figure 4.5: Confusion matrices for RF, SVM and MLP models for multiclass relative intensity classification using wrist sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.816	0.731	0.656
SVM	0.548	0.466	0.266
MLP	0.630	0.614	0.489

Table 4.8: Evaluation metrics for multiclass relative intensity classification using wrist sensor

Performance of the foot sensor

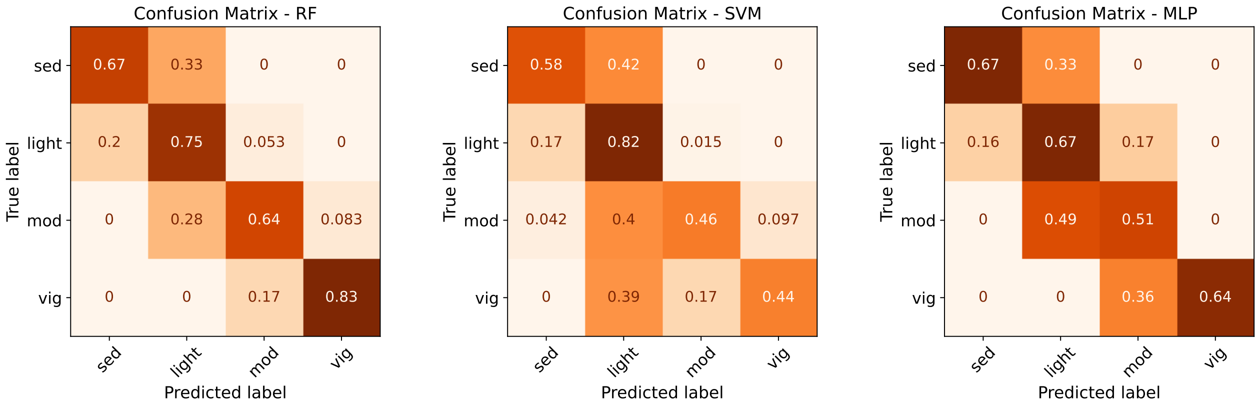
The classification results for the foot sensor reveal differences in classification performance across the three models.

As shown in Figure 4.6a, the RF model achieves balanced results across classes. It performs relatively well for the vigorous class, with 83% of the instances correctly classified. Light activity is slightly less well recognized, while sedentary and moderate show similar but lower performance. Misclassifications occur mostly between adjacent levels, indicating that the model tends to confuse adjacent intensity levels.

On the other hand, Figure 4.6b shows that the SVM model tends to favor the light class, correctly classifying 82% of the samples. However, it struggles with the other classes, which are frequently misclassified across other intensity levels.

The confusion matrix for the MLP model, presented in Figure 4.6c, shows that MLP achieves comparable performance for sedentary, light, and vigorous classes, but performs less well for the moderate class, with only 51% of instances correctly classified. Nonetheless, misclassifications only occur between adjacent classes.

Evaluation metrics in Table 4.9 reflect these trends. The RF model obtains the highest balanced accuracy, macro F1-score, and Cohen's Kappa, followed by SVM and then MLP. Both the RF and SVM models show a moderate level of agreement between predictions and ground truth labels, while MLP is slightly lower, indicating only a fair agreement.



(a) Random Forest (b) SVM (c) MLP

Figure 4.6: Confusion matrices for RF, SVM and MLP models for multiclass relative intensity classification using foot sensor

Model	Balanced Accuracy	Macro F1	Kappa
RF	0.722	0.665	0.579
SVM	0.576	0.549	0.428
MLP	0.622	0.590	0.400

Table 4.9: Evaluation metrics for multiclass relative intensity classification using foot sensor

Comparison between sensor placements

When comparing wrist and foot sensors, the RF model demonstrates better overall performance across both sensors, with superior results using wrist data compared to foot data. While classifications are similar for light and vigorous activities, the RF model performs better for the sedentary and moderate classes using wrist data. The SVM model shows stronger performance with the foot sensor. Both sensors achieve comparable results for the sedentary class, but they follow different trends for the other intensity levels. The MLP model performs better with the wrist sensor, especially for light and vigorous intensities. However, the performance for sedentary and moderate intensities is better with the foot sensor.

4.4.3 Binary classification

Performance of the wrist sensor

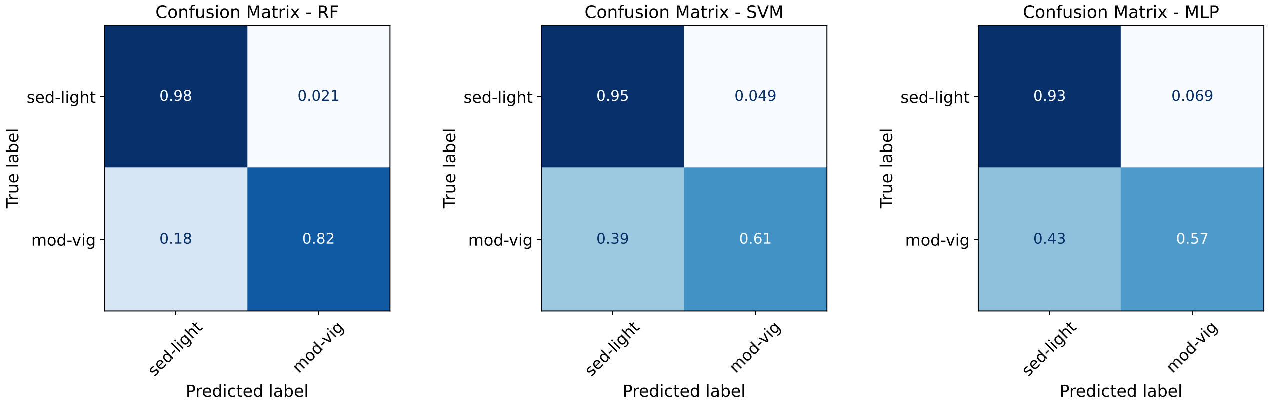
The classification results for the wrist sensor show clear differences in performance across models.

The confusion matrix in Figure 4.7a shows that the RF model classifies sedentary-light category with 98% of samples correctly classified and performs relatively well for the moderate-vigorous category with 82% samples correctly classified.

Figure 4.7b shows that the SVM model demonstrates a strong ability to identify the sedentary-light class, with 95% of instances correctly classified. However, its ability to detect the moderate-vigorous class remains limited, with 61% of samples correctly classified.

Finally, Figure 4.7c illustrates that MLP correctly identifies 93% of sedentary-light samples, indicating similar performance compared to other models for this class. Its ability to classify moderate-vigorous category is noticeably lower, with only 57% of instances correctly classified.

These results are reflected in the evaluation metrics summarized in Table 4.10. The RF model exhibits the highest balanced accuracy, macro F1-score and Kappa coefficient, showing a strong level of agreement between predictions and true labels. On the other hand, the SVM and MLP models show lower metrics, with their respective Cohen's Kappa coefficients highlighting only a moderate level of agreement.



(a) Random Forest (b) SVM (c) MLP

Figure 4.7: Confusion matrices for RF, SVM and MLP models for binary relative intensity classification using wrist sensor

Model	Balanced Accuracy	Macro F1	Kappa
RF	0.902	0.909	0.818
SVM	0.781	0.789	0.586
MLP	0.752	0.758	0.527

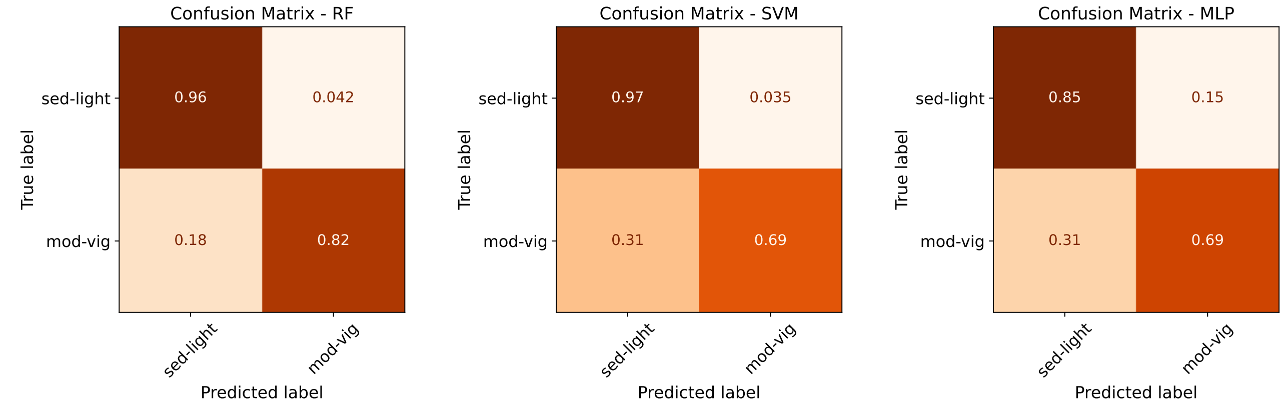
Table 4.10: Evaluation metrics for binary relative intensity classification using wrist sensor

Performance of the foot sensor

The confusion matrices, represented in Figures 4.8a, 4.8b, and 4.8c show the classification performances of the three models using foot sensor data.

The RF model correctly classifies 96% of sedentary-light instances and 82% of moderate-vigorous instances. The SVM model correctly identifies 97% of the sedentary-light samples, while identifying only 69% of the moderate-vigorous samples. In a similar trend, the MLP model classifies 85% of sedentary-light instances while showing limited performance for the moderate-vigorous class, with 69% of instances correctly classified.

The evaluation metrics for each model are summarized in Table 4.11. The RF model achieves the highest balanced accuracy and macro F1-score, with a Cohen's Kappa coefficient indicating a substantial agreement between the predicted and true labels. The SVM model performs slightly worse, with a Cohen's Kappa still indicating a substantial level of agreement, while the MLP model shows lower performance.



(a) Random Forest (b) SVM (c) MLP

Figure 4.8: Confusion matrices for RF, SVM and MLP models for binary relative intensity classification using foot sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.891	0.897	0.794
SVM	0.825	0.834	0.673
MLP	0.774	0.778	0.556

Table 4.11: Evaluation metrics for binary relative intensity classification using foot sensor.

Comparison between sensor placements

When comparing model performance, RF consistently outperforms SVM and MLP models across both sensor placements. Predictions from the RF model are nearly identical between wrist and foot sensors. The SVM model also performs similarly across both placements. In contrast, MLP shows different behavior depending on the sensor placement. It performs better at distinguishing the sedentary-light class with the wrist sensor, while it is more effective at classifying the moderate-vigorous class with the foot sensor.

4.4.4 Discussion and related work

Challenges in multiclass classification

Multiclass classification of relative intensity presents several challenges. A primary limitation is the individualized nature of relative intensity. As observed in Table 4.1, the same activity can be experienced differently across participants. To address this variability, we introduced a personalized reference line based on each individual's *household walking* activity. However, despite this approach, confusion between classes remains frequent. This is likely due to the homogeneity of our study population, which may have limited the differentiating power of the *household walking* reference. In this context, the features of this activity may have been too similar across participants.

Another significant challenge is the strong class imbalance, which leads to low support for certain classes, especially sedentary and vigorous activities. While good performance can be observed with those classes, it should be interpreted with caution, as the small number of examples may not provide enough evidence to confirm the model's ability to generalize. More examples would be needed to validate the consistency of classification for these underrepresented classes. Furthermore, this imbalance also likely affects the model's ability to learn representative patterns across intensity levels.

Model performance and related work

Few studies have focused on classifying relative intensity from IMU sensors data in different intensity classes. For instance, Chowdury et al. (2019) explored a similar task but used physiological data rather than inertial measurements [42]. Despite the difference in input features, the same three machine learning models were applied in both studies. In their study, intensity levels were categorized as low, moderate and high. Compared to our classification, their moderate and high categories correspond to our moderate and vigorous categories, while their low class includes both our sedentary and light activities. Despite methodological differences, both our study and Chowdury et al. show a common pattern, where misclassifications are more frequent between adjacent intensity levels (for instance, light vs. moderate) than between extremes (sedentary vs. vigorous) [42]. This can be explained by the more distinct movement patterns exhibited at the extremes of intensities. Specifically, sedentary activities involve minimal movement, while vigorous activities involve larger movements, particularly in the running activity. Such differentiation in movement patterns facilitates more accurate classification by the models.

Interestingly, Chowdury et al. (2019) reported that the SVM model showed slightly better performance than other models when trained on physiological data [42]. In contrast, our results consistently show that SVM tends to perform worse than RF. The performance of the MLP model relative to SVM depends on the sensor: for the wrist, SVM appears weaker, while for the foot, it achieves slightly better metrics, but the differences remain small.

Additionally, our results do not reveal a clear trend regarding which sensor placement performs best, which is consistent with literature highlighting that the best sensor placements vary across studies [26, 32]. In our study, sensor locations and datasets were kept consistent across models and classification tasks. However, which placement yields better performance varies depending on the machine learning model, and on whether the classification task is multiclass or binary. This indicates that the choice of algorithm and classification approach can influence which sensor placement yields better results.

Binary classification vs. multiclass classification

Binary classification performs well across both sensors using RF models and consistently outperforms the multiclass approach across all three classifiers. To compare binary and multiclass classification more appropriately, we rely on the Cohen's Kappa coefficient, which accounts for the agreement expected by chance and provides a more objective evaluation of actual model performance.

These results suggest that binary classification enhances model effectiveness in this case. This improvement may be partly explained by a reduction in class imbalance. In the multiclass datasets, the imbalance is strong, particularly with very few samples in the sedentary class. By grouping sedentary and light activities together and moderate and vigorous into another group, we were able to reduce the imbalance, even though a difference between class sizes remains. This likely contributed to the better model performance observed in the binary scenario. Additionally, the simplification likely allowed for clearer decision boundaries for the models, which is relevant in relative intensity, where the changes between intensity levels are not sharply defined and can vary between one person to another. While binary classification does not provide the same level of detail as multiclass classification, it still offers valuable information. As mentioned above, moderate-to-vigorous activity is considered a key focus in physical activity guidelines, which set the minimum threshold of moderate-to-vigorous activity for significant health benefits [1].

The binary classification offers practical advantages. It yielded better performance, reduced computational complexity and aligned with health goals. However, grouping sedentary and light activities together limits the ability to detect sedentary behavior, which can be a disadvantage depending on the goals. Indeed, sedentary behavior has been linked to negative health outcomes, while light intensity has been shown to confer health benefits [1, 2]. Therefore, the inability to distinguish between these two intensity levels may reduce the utility of binary classification for interventions focused on minimizing sedentary behavior or promoting light intensity physical activity. In contrast, multiclass classification, while being more complex and showing lower performance in our study, provides a more detailed differentiation between intensity levels. However, this approach would require further improvement to enhance its reliability.

4.5 Classification of absolute intensity

4.5.1 Dataset characteristics

Compared to relative intensity, datasets for absolute intensity classification present a more balanced distribution across intensity classes. For the multiclass tasks, the support includes 36 sedentary, 72 light, 72 moderate, and 72 vigorous instances. In the binary scenario, the support is 108 instances of sedentary-light and 144 instances of moderate-vigorous.

4.5.2 Multiclass classification

Performance of the wrist sensor

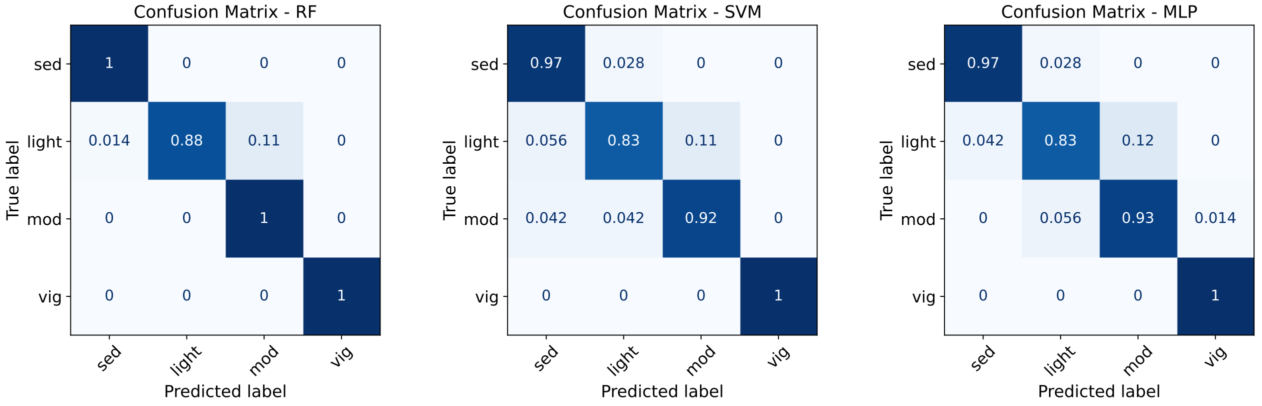
The results for the wrist sensor show high classification performance across all models.

In Figure 4.9c, we observe that the RF model achieves near-perfect classification, with sedentary, moderate, and vigorous activity classes perfectly identified. The light activity class, however, is more challenging, with 88% of instances correctly classified and the remaining misclassified as adjacent classes.

Figure 4.9b indicates that the SVM model performs slightly worse than RF, but still achieves strong results, with over 83% of instances correctly classified across all classes. Similar to the RF, light intensity remains the most difficult class to identify correctly. In this case, misclassifications are not limited to adjacent classes, some moderate instances are also labeled as sedentary.

Figure 4.9c shows that the MLP model follows a similar trend to SVM, with 97% of the instances of the sedentary class correctly classified and perfect classification for the vigorous class. As with the other models, light activity remains the most difficult to classify.

Evaluation metrics summarized in Table 4.12 confirm strong overall performance for all three models, with RF consistently achieving the highest scores. All models achieve balanced accuracy and macro F1-scores above 0.92, and Cohen's Kappa values fall within the almost perfect agreement range between predicted labels and ground truth.



(a) Random Forest

(b) SVM

(c) MLP

Figure 4.9: Confusion matrices for RF, SVM and MLP models for multiclass absolute intensity classification using wrist sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.969	0.967	0.951
SVM	0.931	0.921	0.898
MLP	0.934	0.930	0.903

Table 4.12: Evaluation metrics for multiclass absolute intensity classification using wrist sensor

Performance of the foot sensor

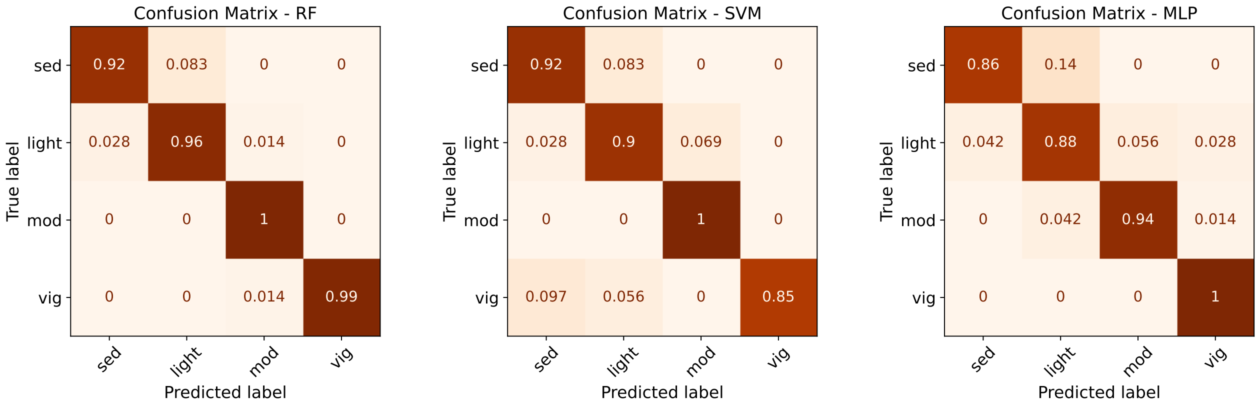
The results for the foot sensor indicate generally strong classification performance across the three models.

As observed in Figure 4.10a, the confusion matrix for the RF model shows that each activity intensity is correctly identified with at least 92% of instances correctly classified, and perfect classification for the moderate category. Minor misclassifications are observed primarily between adjacent categories.

The performance of the SVM model, illustrated in Figure 4.10b, indicates overall robust performance, with perfect classification for the moderate category. Misclassifications occur more frequently in the vigorous activity class, where 85% of instances are correctly identified. Some vigorous samples are misclassified as sedentary and light, indicating the greater difficulty of the SVM model in identifying vigorous activities.

As shown in the confusion matrix in Figure 4.10c, the MLP model exhibits lower overall classification performance compared to the two other models, except for the vigorous class, which is perfectly classified. Similar to the RF and SVM models, the moderate class is well recognized. The light class is also relatively well identified, however, some misclassifications occur with all other intensities.

These classification patterns are reflected with the evaluation metrics summarized in Table 4.13. The RF model demonstrates high performance, with a high balanced accuracy, macro F1-score, and Cohen's Kappa, indicating strong classification across all classes. The SVM and MLP achieves lower metrics, which corresponds with the increased confusion observed in the predictions.



(a) Random Forest

(b) SVM

(c) MLP

Figure 4.10: Confusion matrices for RF, SVM and MLP models for multiclass absolute intensity classification using foot sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.965	0.967	0.962
SVM	0.917	0.909	0.887
MLP	0.920	0.923	0.903

Table 4.13: Evaluation metrics for multiclass absolute intensity classification using foot sensor

Comparison between sensor placements

The RF model shows similar performance between the wrist and the foot sensors, although the wrist has more difficulty classifying the light intensity compared to the foot. For the SVM model, comparable metrics are observed across both placements. However, the foot again performs better at classifying light intensity, while the wrist achieves better results for vigorous intensity. Finally, the MLP model follows a similar trends across sensors, with the wrist performing better at identifying sedentary intensities.

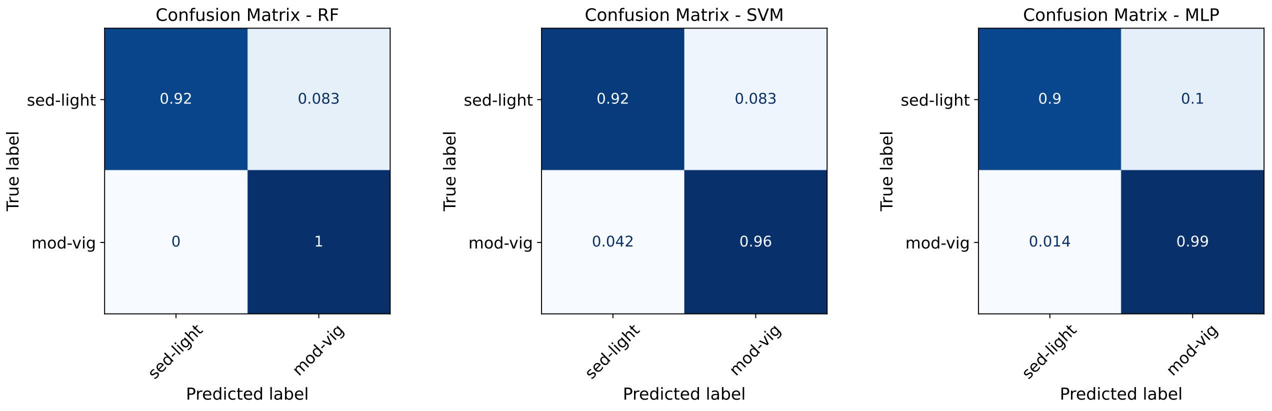
4.5.3 Binary classification

Performance of the wrist sensor

All three models achieve strong results using the wrist data, as shown in Figures 4.11a, 4.11b, and 4.11c.

We observe that each model correctly classifies at least 90% of instances in both intensity classes. RF achieves a perfect classification for the moderate-vigorous class, while MLP and SVM also demonstrate strong performance, with only a small proportion of misclassified instances. In all cases, the sedentary-light class is classified similarly, but with slightly more misclassifications compared to the other class.

The overall trends are reflected in the evaluation metrics summarized in Table 4.14, with all models reaching high balanced accuracy, macro F1-scores and Cohen's Kappa values, indicating consistent performance across classes.



(a) Random Forest (b) SVM (c) MLP

Figure 4.11: Confusion matrices for RF, SVM and MLP models for binary absolute intensity classification using wrist sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.959	0.963	0.926
SVM	0.938	0.939	0.878
MLP	0.942	0.947	0.894

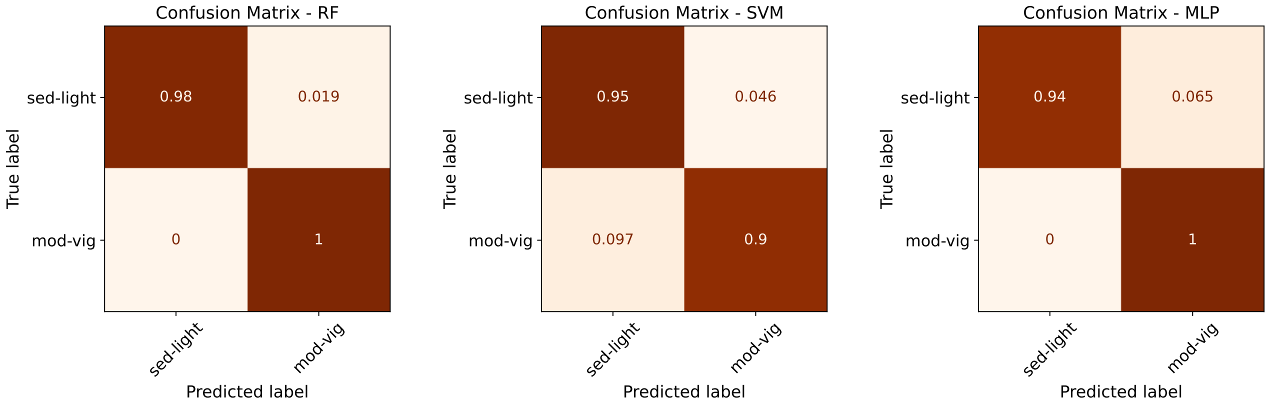
Table 4.14: Evaluation metrics for binary absolute intensity classification using wrist sensor

Performance of the foot sensor

For the foot sensor, all models achieve strong results, as represented in Figures 4.12a, 4.12b, and 4.12c.

The RF model achieves a nearly perfect classification, with perfect classification of the moderate-vigorous class and 98% of instances correctly classified for the other class. The SVM model also demonstrates robust performance, with a strong classification of the sedentary-light class and only 9.7% of moderate-vigorous instances misclassified. Conversely, the MLP model achieves perfect classification on the moderate-vigorous class, with a slightly higher error rate for the sedentary-light class.

Overall, the performance metrics in Table 4.15 confirm a high level of agreement with the ground truth across all three models, particularly for the RF and MLP models.



(a) Random Forest (b) SVM (c) MLP

Figure 4.12: Confusion matrices for RF, SVM and MLP models for binary absolute intensity classification using foot sensor

Model	Balanced accuracy	Macro F1	Kappa
RF	0.991	0.992	0.984
SVM	0.928	0.924	0.848
MLP	0.968	0.971	0.943

Table 4.15: Evaluation metrics for binary absolute intensity classification using foot sensor

Comparison between sensor placements

When comparing model performance across the two sensor placements, the RF model shows higher performance than the SVM and MLP models. Its classification of the moderate-vigorous class remains stable across sensors, while performance for the sedentary-light improves using the foot sensor. The

MLP model follows a similar pattern.

In contrast, the SVM model's performance varies with sensor placement. It performs better for the moderate-vigorous class with the wrist sensor, whereas classification of the sedentary-light is slightly improved with the foot sensor.

4.5.4 Discussion and related work

Challenges in multiclass classification

In multiclass classification of absolute intensity, each activity is assigned to a single intensity level for all individuals. Consequently, one of the main challenges is the nature of the movements and activities themselves.

In our case, walking activities can represent a notable source of confusion. Indeed, both light and moderate activities includes walking components. Depending on the participant, movement characteristics such as walking speed or gait patterns during the light intensity task (*household walking*) could closely resemble those of moderate intensity task (*treadmill walking*), which could have contributed to some misclassifications. However, unlike *treadmill walking*, the light activity followed a predefined path that involved sitting, bending to take a book, and avoiding obstacles. These additional movement variations likely introduced distinct motion patterns, which may still have helped the algorithm differentiate between the two activity intensities.

This challenge was also observed with the wrist sensor. Indeed, across the three models light-intensity activities were generally more difficult to classify. This could be explained by the nature of the tasks and how they were performed. For example, during the *cooking* task, participants used their dominant hand for most movements, which may have resulted in less detectable motion on the non-dominant wrist, where the sensor was placed. While some parts of the activity (arranging the dishes) involved both hands in most cases, the overall motion detected on the wrist could have been limited, leading to confusion between light and sedentary intensity.

Model performance and related work

The results demonstrate strong classification performance for both wrist and foot sensor placement for multiclass classification. Our results align with existing literature, which emphasizes that the wrist sensor enables accurate prediction of absolute activity intensity levels using machine learning [37]. Furthermore, the foot sensor also performed well overall. These observations are consistent with Lazar et al. (2020), who highlighted the relevance of ankle-worn accelerometers for physical activity intensity classification, and Montoye et al. (2016), who demonstrated the potential of thigh placements, both supporting the relevance of lower-limb placements for physical intensity assessment [37, 45].

We observe overall stability in performance between wrist and foot sensors, even though they capture different aspects of movements. It can be explained by the structured design of the protocol. All participants performed the same predefined tasks in a similar manner, within clear absolute intensity categories. This produces relatively uniform signal patterns across both sensors, which benefits the classification models. These overall observations regarding sensor placement are consistent with literature, which confirms that models trained and tested in controlled laboratory environments generally achieve high accuracy, while performance drops when applied to free-living environments [27, 41].

Despite the generally good results in multiclass classification, some unexpected misclassification patterns are observed, particularly for the foot sensor, for reasons that remain unclear. For instance, SVM produces misclassifications between extreme intensity levels, classifying a small percentage of vigorous intensity as sedentary. This is a notable finding given the distinct differences in movement patterns between these classes. Similarly, the MLP frequently confuses multiple intensity classes with one another.

Binary classification vs. multiclass classification

When comparing multiclass to binary classification, results show that this approach generally does not provide significant improvement in performance, particularly for the wrist sensor. Indeed, the Cohen's Kappa metric, which adjusts for chance-level agreement, shows that performance remains comparable or lower. This suggests that simplifying the problem by merging classes into two categories is not broadly beneficial in this case. The only exceptions are the RF and MLP models using the foot sensor data, where binary classification provides modest improvements.

The limited change in performance between multiclass and binary classification can be partly explained by the variability within the grouped classes. For example, the sedentary-light class combines activities with different movement patterns, such as sitting and walking. This results in signal heterogeneity, making it harder for models to learn a clear decision boundary. This is expected in the context of absolute intensity classification, where each activity is assigned a fixed intensity level for all participants.

4.6 Absolute intensity vs. relative intensity classification

In both binary and multiclass tasks, absolute intensity classification consistently outperformed relative intensity, highlighting a clear difference in model performance. This result is not unexpected given the nature of absolute intensity. Unlike relative intensity, absolute intensity relies on standardized thresholds, such as those based on MET values, making it more straightforward for machine learning models to learn and generalize patterns. Previous research has already demonstrated promising results for absolute intensity classification, particularly with accelerometer data [27, 30, 36, 37]. In one study, these results were further supported by the inclusion of gyroscope data [44].

In contrast, relative intensity is an individualized measure, which introduces inter-individual variability, making the classification task more complex. The same activity falls into different intensity categories depending on the participant, making it harder for models to generalize, even if a reference line is used. Moreover, given the homogeneity of our study population, movement patterns tend to be quite similar across individuals, which further complicates the model's ability to distinguish between classes. Nevertheless, while absolute intensity classification achieves higher performance, predicting relative intensity remains a crucial goal, as it allows for a more accurate and personalized assessment of physical activity.

Despite some differences in performance, RF remains the most effective model for classification. These results align with other studies that have found RF effective for absolute intensity classification [27, 30, 55]. However, when it comes to relative intensity, RF achieves lower performance compared to its results on absolute intensity, indicating that classification of relative intensity remains a challenging task, particularly for multiclass classification.

4.7 Feature importance analysis in a best-performing case

In this section, we present the results of the feature importance analysis performed on the binary classification of relative intensity, using wrist sensor data and an RF model. This analysis was conducted using the best scenario, focusing on the binary classification task as it yielded better balanced accuracy, macro F1-score, and Cohen's Kappa than in multiclass classification. Since performance for wrist and foot sensors were similar for RF, we chose to focus on wrist placement, as it has already been validated as an acceptable and compliant placement. While the specific top features may differ between sensor locations, comparable performance suggests that both placements capture relevant movement information.

This analysis was performed using the RF model, which provides intrinsic measures of feature relevance. These importance scores are therefore specific to RF and may differ from those of the other algorithms evaluated in this study.

Figure 4.13 displays the 30 most discriminative features. The prefix "GYRO" indicates features derived from the gyroscope, "ACC" from the accelerometer, and "relative" refers to features from the added reference line. The letters X, Y, and Z correspond to the sensor axes, and each feature name is followed by its specific measure. The importance values observed are relatively low (all below 0.022), which is expected given that the model's feature importance are normalized to sum to 1. Given the large number of features in our dataset, even the most relevant ones receive a relatively low individual score.

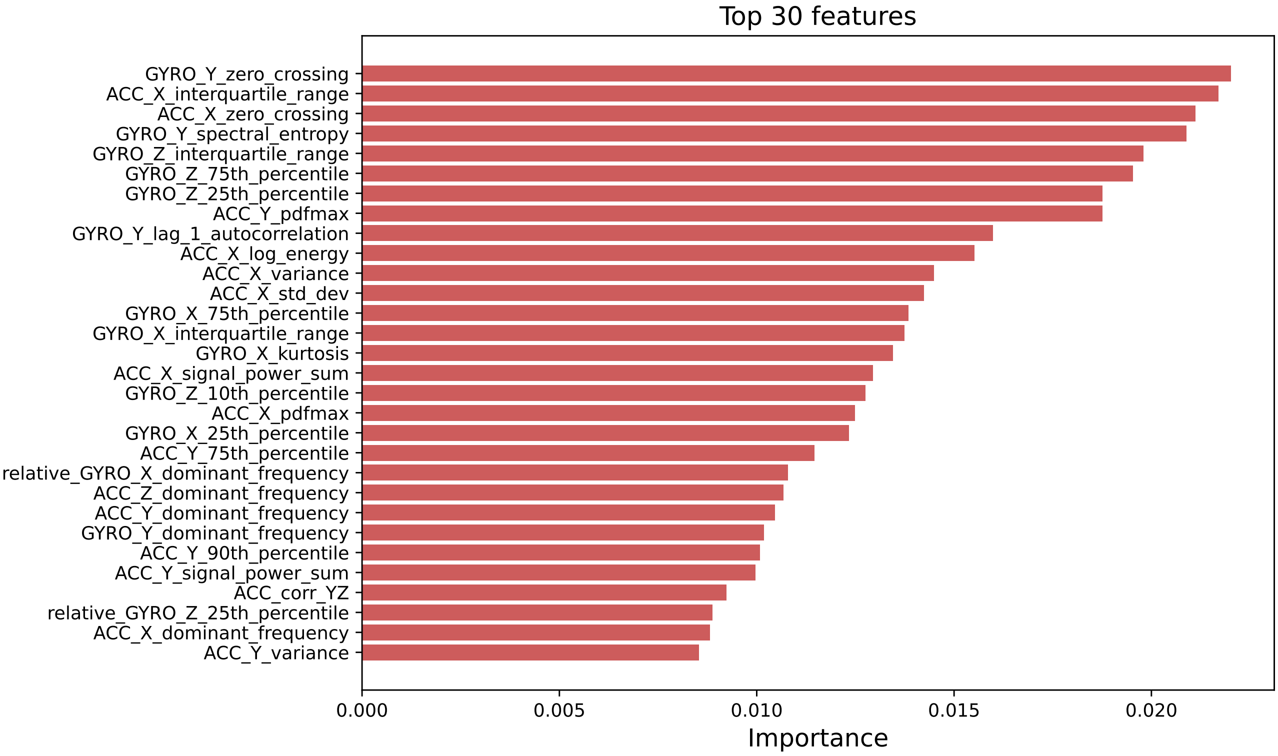


Figure 4.13: Top 30 most important features in RF model for binary classification of relative intensity from wrist sensor data

By examining Figure 4.13, we observe that gyroscope and accelerometer features are present in roughly equal proportions among the top 30 features. This is interesting because most studies aiming to predict or classify physical activity intensity tend to focus exclusively on accelerometer data [44, 45]. Hibbing et al. (2018) demonstrated that gyroscope data can enhance the prediction of absolute intensity, and our results extend this observation by demonstrating that gyroscope features are also valuable when predicting relative intensity [44].

Among the most discriminative features, we observe the number of zero crossings for the x-axis of the accelerometer and y-axis of the gyroscope. Zero crossings quantify how many times the signal changes sign (from positive to negative or vice versa) within a given time window. This feature indicates how often the movement changes direction, linearly for the accelerometer and rotationally for the gyroscope. A high number of zero crossings reflects rapid, repetitive or oscillatory movements, which are typical for vigorous activities such as the activity *treadmill running*, while lower counts are associated with slower or sustained motions.

Beyond zero crossing, statistical measures of signal variability, such as the inter quartile range, also appear highly discriminative. The inter quartile range, which is defined as the difference between the 75$^{\text{th}}$ and 25$^{\text{th}}$ percentiles, measures the dispersion around the median. Higher values indicate larger or more variable movements, which can help distinguish activities of different intensities.

Frequency features such as spectral entropy, power of the dominant frequency (referred to as *pdfmax*) and dominant frequency were also among the most discriminative. The spectral entropy of the y-axis of the gyroscope is particularly discriminative. Uniform movements such as in *treadmill walking* or *treadmill running* produce a low spectral entropy due to a high frequency concentration [33]. In contrast, more irregular activities like *cooking*, *sweeping* or *household walking* correspond to a higher spectral entropy [33]. Previous work by Montoye et al. (2018) reported that adding frequency features did not significantly improve the prediction of absolute intensity from wrist accelerometer data [30]. In contrast, our findings indicate that frequency features contribute meaningfully to the classification of relative intensity.

We observe other repeated occurrences of statistical dispersion measures, such as various percentiles, which capture the amplitude and variability of both acceleration and angular velocity. Similarly, features related to signal strength, such as the sum of the signal power (*signal_power_sum*) appear multiple times. Overall, the most discriminative features include both time and frequency domains, cover all three axes and include features from both accelerometers and gyroscopes. This indicates that relative intensity classification in this case benefits from a diverse combination of signal characteristics.

Although an individual-specific reference was included into the dataset to provide the model with an individualized comparison point, only two of these variables (*relative_GYRO_X_dominant_frequency* and *relative_GYRO_Z_25th_percentile*) appears among the 30 most important features. This suggests that these features were not dominant in contributing to model performance, but they still contributed to distinguishing between sedentary-light and moderate-vigorous intensities. As previously discussed, their limited impact may partly be due to the homogeneity of our participants, but features that do not appear among the top ones can still contribute indirectly by interacting with other features within the model.

Chapter 5

Limitations

This fifth chapter examines the main limitations encountered during this study, highlighting them as opportunities for future improvement.

5.1 Participant characteristics and datasets challenges

A major limitation of this study is the small number of participants. Although we initially planned to recruit 40 participants, including both young and older adults, only 20 young adults were enrolled due to time constraints. This resulted in a limited amount of data and a very homogeneous sample, as all participants were healthy and aged between 18 and 25. Most of them were students, since they were the most available to attend the laboratory sessions, which further contributed to the similarity within the group. Nonetheless, despite the overall similarity in profiles, some participants were noticeably more physically active than others. Including a more heterogeneous population, especially with older adults as originally planned, would have allowed us to better explore differences in relative intensity between groups. It would also have allowed us to assess whether our algorithms could generalize to more complex tasks. Furthermore, when training on a limited dataset, models may overfit, resulting in poor generalization to new data.

To mitigate these issues and improve generalizability within our limited sample, we applied a leave-one-out cross-validation when possible. This approach maximizes the use of each data and provides a better estimation of the model's ability to generalize, as in each iteration all participants except one are used for training, and the remaining participant is used solely for testing.

Another important limitation concerns the strong class imbalance within the relative intensity datasets, particularly for certain intensity levels. The small test sets combined with the fact that not all participants experienced every intensity level, further limited model training and evaluation. Including a wider variety of activities that cover more intensity levels, along with involving more participants, would help to improve model robustness and ensure more reliable classification across all intensity classes.

A future perspective would therefore be to increase the number of participants and ensure greater heterogeneity, both in participant profiles and activity types, which would allow for a more thorough evaluation of the models.

5.2 Laboratory setting

Another limitation of this study is related to its laboratory setting, where all activities were structured and performed uniformly by all participants, with only the *sweeping* and *household walking* tasks allowing for slightly more natural movement.

Several studies have shown that algorithms developed and tested only in laboratory settings tend to underperform when applied to free-living data [27, 30, 41]. This highlights the importance of developing and validating models directly in real-world environments to ensure their applicability. While the final goal is to use these algorithms in free-living contexts, starting with controlled laboratory experiments provides a stable environment that makes it easier to collect baseline data and to train the algorithms before progressing to more complex settings.

To address this limitation, we included everyday activities such as cooking, walking around the house, sweeping and climbing stairs, which are activities that adults may commonly perform in daily life. This approach aligns with previous research indicating that adult movement patterns rarely consist solely of walking or running, and thus recommends incorporating a variety of activities [26].

Future research should focus on testing and validating these algorithms in semi-structured or fully free-living environments, to better capture the variability and complexity of everyday movements.

Chapter 6

Conclusions

Accurately measuring the intensity of physical activity remains a challenge, particularly when the goal is to provide individualized assessments. This thesis addressed this challenge by evaluating how wearable inertial sensors, combined with machine learning techniques, can classify the relative intensity of daily life activities. Two classification schemes were explored: a multiclass approach with four classes (sedentary, light, moderate and vigorous) and a simplified binary classification grouping sedentary and light together, and moderate and vigorous together.

Beyond testing the feasibility of this approach, this study compared its performance with absolute intensity classification on the same tasks, assessed the influence of two sensor placements, and highlighted the most discriminative variables in an optimal case.

Classifying relative intensity solely from inertial measurement unit data proved to be challenging. Despite incorporating a personalized reference based on the *household walking* activity, confusion between intensity classes occurred. In the multiclass tasks, misclassifications varied across models, occurring either only between adjacent classes or spanning multiple classes. By contrast, binary classification consistently outperformed multiclass approaches across all models and sensor placements. This approach helped to reduce the class imbalance and improve overall performance, while still aligning with public health guidelines that emphasize moderate-to-vigorous physical activity. However, this simplification reduces the level of detail, particularly when it comes to distinguishing sedentary behavior, which is also a significant health factor. Consequently, the choice of classification method should be guided by the specific objectives of the application. Moreover, in both classification tasks, the best sensor placement depends on the model, as different algorithms may exploit the signal characteristics differently.

Regarding absolute intensity, classification results were generally better than for relative intensity. This was expected, as each activity is assigned a fixed intensity level for all individuals, without considering personal differences. Additionally, the structured nature of the activities helped produce uniform movement patterns, which likely benefited classification performance. Nevertheless, some confusion still occurred between intensities. In contrast to relative intensity, binary classification did not significantly improve performance for absolute intensity. Combining classes increased heterogeneity within groups, which likely made it harder for the model to learn clear decision boundaries.

Overall, three machine learning models (RF, SVM, and MLP) were evaluated, and RF consistently outperformed the two other models in classifying physical activity intensity across all tasks and sensor placements.

To better understand the features affecting classification performance, a feature importance analysis was conducted on the binary classification task using wrist sensor data with the RF model. The results revealed that both the accelerometer and gyroscope features contributed significantly, highlighting the relevance of combining these data. Interestingly, the personalized reference features had limited impact, possibly due to participant homogeneity. Further analysis across other classification tasks and sensor placements would be valuable to confirm these observations.

The main limitations of this study include a small, homogeneous sample of young adults and a controlled laboratory setting, which restrict the generalizability of the models. Additionally, class imbalance and the limited variety of activities further challenge classification performance. Future work should involve larger and more diverse populations, as well as developing models in semi-structured or free-living environments to better capture real-world movement variability.

In conclusion, this study provides valuable insights into the challenges of classifying relative physical activity using wearable inertial sensors and machine learning, in comparison with absolute classification. Despite exploring two classification schemes and sensor placements, the results underscore the difficulty of achieving accurate classification. While binary classification achieved relatively good performance with RF, results from other models and multiclass classification were more limited, reflecting the complexity of classifying relative intensity. These findings highlight the importance of selecting models based on the specific application. Further research is needed to improve model performance and generalizability, thereby enabling more reliable use of wearable sensors for assessing the relative intensity of physical activity.

Use of Artificial Intelligence

This master's thesis was developed with the occasional support of artificial intelligence tools, such as ChatGPT and Perplexity. Their use was strictly limited to specific tasks, including rephrasing of passages I had already written and the translation or refinement of certain terms. All ideas, analyses, and interpretations are the result of my own work, informed by scientific literature and discussions with my supervisor.

In some cases, artificial intelligence was used to help identify coding errors and suggest possible debugging approaches. It was also employed to produce visually clear graphs. These tools served only as assistants, while implementations and every final decision were made by me, based on scientific literature and with guidance from my supervisor and a doctoral researcher.

Appendices

Appendix A: Walking route plan of the household walking activity

Figures 6.1 and 6.2 illustrate the route followed by the participants during the *household walking* activity introduced in Section 3.4.2.

The order of movements is indicated by numbers on the two diagrams represented in Figures 6.1 and 6.2, which illustrate the outward and return paths, respectively. The path itself is shown with green dotted arrows. The participants started at point 1 and walked to point 2, where they sat on a chair for 10 seconds. Then, they moved around an obstacle represented as point 3 and passed through a narrow passage formed by two chairs, simulating movement in a confined space, at point 4. Next, they walked to point 5 and picked up a book placed on a chair, which they carried to another chair located at point 6. The book is illustrated in red. Afterwards, participants retraced their steps by passing again through the narrow passage and moving around the obstacle, represented by point 7 and point 8, respectively. Finally, they walked to point 9, the finishing point, before starting the route again. This cycle was repeated continuously for 3 minutes.

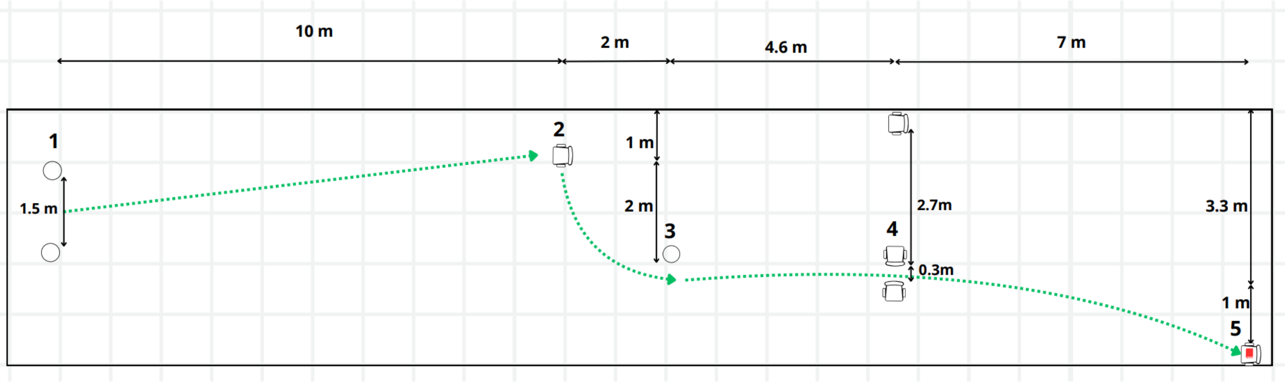


Figure 6.1: Outward path for the household walking activity

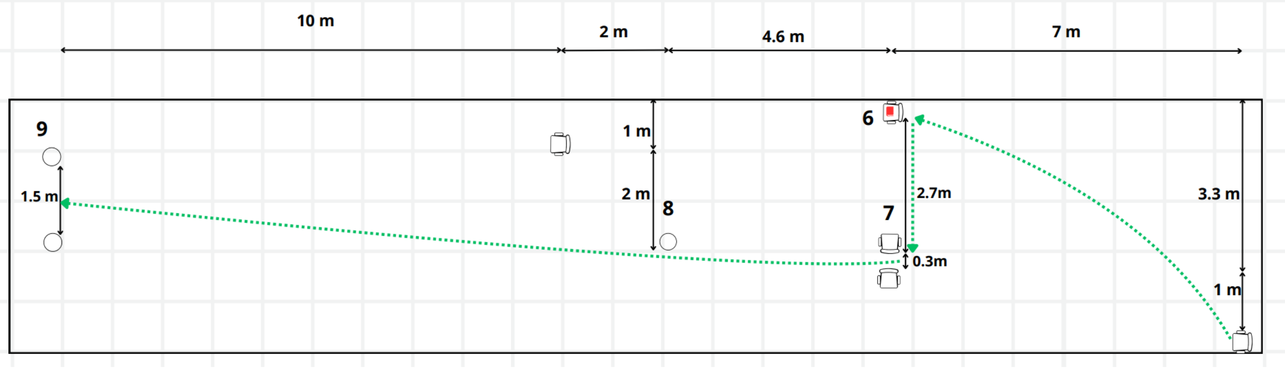


Figure 6.2: Return path for the household walking activity

Appendix B: Initial hyperparameters fixed for each model

Models	**Fixed hyperparameters**
Random Forest	`random_state` = 42 `class_weight = 'balanced'`

Support Vector Machine	`kernel = 'linear'`
	`random_state` = 42
	`class_weight = 'balanced'`

Multilayer Perceptron	`max_iter` = 10 000 `early_stopping = True` `random_state` = 42

Table 6.1: Initial hyperparameters fixed for each model.

Appendix C: Grid search spaces for each machine learning model

Parameter	Values
n_estimators	100, 300, 500, 700
max_depth	None, 5, 10
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 5

Table 6.2: Grid search space for Random Forest

Parameter	Values
kernel	linear, rbf
C	0.01, 0.1, 1, 10
gamma	scale, 0.001, 0.01, 0.1

Table 6.3: Grid search space for Support Vector Machine

Parameter	Values
hidden_layer_sizes	(20,), (50,), (100,), (70,50), (100,50), (100,70)
activation	relu, tanh
alpha	0.00001, 0.0001, 0.001
learning_rate_init	0.0001, 0.001

Table 6.4: Grid search space for Multi-Layer Perceptron.

The notation used for hidden_layer_sizes hyperparameter represents the number and size of the hidden layers in the neural network. For instance, (20,) corresponds to a single hidden layer with 20 neurons, whereas (100, 50) indicates a network with two hidden layers containing 100 and 50 neurons, respectively.

Appendix D: Detailed macro F1-scores for window size optimization

Multiclass classification of relative intensity

Model	10s	15s	20s	30s	60s
RF	0.549	**0.603**	0.570	0.546	0.583
SVM	0.419	0.394	0.407	**0.492**	0.473
MLP	**0.497**	0.465	0.448	0.461	0.479

Table 6.5: Macro F1-scores for multiclass classification of relative intensity - wrist sensor

Model	10s	15s	20s	30s	60s
RF	**0.519**	0.503	0.484	0.509	0.509
SVM	0.404	0.359	0.401	0.409	**0.459**
MLP	**0.500**	0.461	0.480	0.473	0.477

Table 6.6: Macro F1-scores for multiclass classification of relative intensity - foot sensor

Binary classification of relative intensity

Model	10s	15s	20s	30s	60s
RF	0.759	0.789	0.773	**0.804**	0.765
SVM	**0.778**	0.762	0.756	0.768	0.712
MLP	0.773	**0.811**	0.770	0.776	0.752

Table 6.7: Macro F1-scores binary classification of relative intensity - wrist sensor

Model	10s	15s	20s	30s	60s
RF	0.812	0.813	**0.822**	0.811	0.797
SVM	0.724	0.649	0.701	0.715	**0.771**
MLP	0.759	0.751	**0.764**	0.761	0.739

Table 6.8: Macro F1-scores binary classification of relative intensity - foot sensor

Multiclass classification of absolute intensity

Model	10s	15s	20s	30s	60s
RF	0.918	0.940	0.950	0.950	**0.959**
SVM	0.902	0.892	0.943	0.937	**0.952**
MLP	**0.924**	0.890	0.928	0.923	0.921

Table 6.9: Macro F1-scores for multiclass classification of absolute intensity - wrist sensor

Model	10s	15s	20s	30s	60s
RF	0.907	0.918	0.925	0.943	**0.969**
SVM	0.916	0.935	**0.946**	0.944	0.926
MLP	**0.933**	0.928	0.907	0.920	0.880

Table 6.10: Macro F1-scores for multiclass classification of absolute intensity - foot sensor

Binary classification of absolute intensity

Model	10s	15s	20s	30s	60s
RF	0.955	0.956	0.958	0.965	**0.986**
SVM	0.930	0.937	0.966	0.958	**0.979**
MLP	**0.949**	0.937	0.926	0.925	0.901

Table 6.11: Macro F1-scores for binary classification of absolute intensity - wrist sensor

Model	10s	15s	20s	30s	60s
RF	0.973	0.967	0.971	0.983	**0.990**
SVM	0.954	0.951	0.947	**0.974**	0.971
MLP	**0.962**	0.961	0.940	0.938	0.943

Table 6.12: Macro F1-scores for binary classification of absolute intensity - foot sensor

Bibliography

[1] World Health Organization, *Who guidelines on physical activity and sedentary behaviour*, Geneva, 2020.

[2] American College of Sports Medicine, *ACSM's Guidelines for Exercise Testing and Prescription*, 11th. Wolters Kluwer, 2021.

[3] 2018 Physical Activity Guidelines Advisory Committee, "2018 physical activity guidelines advisory committee scientific report," U.S. Department of Health and Human Services, Tech. Rep., 2018.

[4] D. Arvidsson, J. Fridolfsson, and M. Börjesson, "Measurement of physical activity in clinical practice using accelerometers," *Journal of Internal Medicine*, vol. 286, no. 2, pp. 137–153, 2019. doi: 10.1111/joim.12908.

[5] A. Marques *et al.*, "The association between physical activity and chronic diseases in European adults," *European Journal of Sport Science*, vol. 18, no. 1, pp. 140–149, 2018. doi: 10.1080/17461391.2017.1400109.

[6] World Health Organization, *Maladies non transmissibles*, Accessed: 17 August 2025. [Online]. Available: https://www.who.int/fr/news-room/fact-sheets/detail/noncommunicable-diseases.

[7] World Health Organization, *Who reveals leading causes of death and disability worldwide 2000–2019*, Accessed: 17 August 2025. [Online]. Available: https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.

[8] World Health Organization, *Data for a healthier future: How countries can protect people from noncommunicable diseases*, Accessed: 17 August 2025. [Online]. Available: https://www.who.int/europe/news/item/10-09-2024-data-for-a-healthier-future--how-countries-can-protect-people-from-noncommunicable-diseases.

[9] A. P. Hills, N. Mokhtar, and N. M. Byrne, "Assessment of physical activity and energy expenditure: An overview of objective measures," *Frontiers in Nutrition*, vol. 1, no. 5, pp. 1–16, 2014. doi: 10.3389/fnut.2014.00005.

[10] S. Herrmann *et al.*, "Adult compendium of physical activities: A third update of the energy costs of human activities," *Journal of Sport and Health Science*, vol. 13, no. 1, pp. 6–12, 2024. doi: 10.1016/j.jshs.2023.10.010.

[11] J. Fridolfsson *et al.*, "Accelerometer-measured absolute versus relative physical activity intensity: Cross-sectional associations with cardiometabolic health in midlife," *BMC Public Health*, vol. 23, no. 1, p. 2322, 2023. doi: 10.1186/s12889-023-17281-4.

[12] A. Warner *et al.*, "Agreement and relationship between measures of absolute and relative intensity during walking: A systematic review with meta-regression," *PLoS One*, vol. 17, no. 11, e0277031, 2022. doi: 10.1371/journal.pone.0277031.

[13] D. Ndahimana and E.-K. Kim, "Measurement methods for physical activity and energy expenditure: A review," *Clinical Nutrition Research*, vol. 6, no. 2, pp. 68–80, 2017. doi: 10.7762/cnr.2017.6.2.68.

[14] B. Ainsworth *et al.*, "The current state of physical activity assessment tools," *Progress in Cardiovascular Diseases*, vol. 57, no. 4, pp. 387–395, 2015. doi: 10.1016/j.pcad.2014.10.005.

[15] N. Williams, "The borg rating of perceived exertion (rpe) scale," *Occupational Medicine*, vol. 67, no. 5, pp. 404–405, 2017. doi: 10.1093/occmed/kqx063.

[16] J. Siddique *et al.*, "Individualized relative-intensity physical activity accelerometer cut points," *Medicine and Science in Sports and Exercise*, vol. 52, no. 2, pp. 398–407, 2020. doi: 10.1249/MSS.0000000000002153.

[17] V. Farrahi *et al.*, "Evaluating and enhancing the generalization performance of machine learning models for physical activity intensity prediction from raw acceleration data," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 27–38, 2020. doi: 10.1109/JBHI.2019.2917565.

[18] J. Leal-Martín *et al.*, "Resting oxygen uptake value of 1 metabolic equivalent of task in older adults: A systematic review and descriptive analysis," *Sports Medicine*, vol. 52, no. 2, pp. 331–348, 2022. doi: 10.1007/s40279-021-01539-1.

[19] S. J. Strath *et al.*, "Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the american heart association," *Circulation*, vol. 128, no. 20, pp. 2259–2279, 2013. doi: 10.1161/01.cir.0000435708.67487.da.

[20] Z. Gao *et al.*, "The dilemma of analyzing physical activity and sedentary behavior with wrist accelerometer data: Challenges and opportunities," *Journal of Clinical Medicine*, vol. 10, no. 24, p. 5951, Dec. 2021. doi: 10.3390/jcm10245951.

[21] K. Bakrania *et al.*, "Intensity thresholds on raw acceleration data: Euclidean norm minus one (enmo) and mean amplitude deviation (mad) approaches," *PLOS ONE*, vol. 11, no. 10, e0164045, 2016. doi: 10.1371/journal.pone.0164045.

[22] G. J. Sanders *et al.*, "Evaluation of wrist and hip sedentary behaviour and moderate-to-vigorous physical activity raw acceleration cutpoints in older adults," *Journal of Sports Sciences*, vol. 37, no. 11, pp. 1270–1279, 2019. doi: 10.1080/02640414.2018.1555904.

[23] S. G. Trost, D. S. K. Brookes, and M. N. Ahmadi, "Evaluation of wrist accelerometer cut-points for classifying physical activity intensity in youth," *Frontiers in Digital Health*, vol. 4, p. 884 307, 2022. doi: 10.3389/fdgth.2022.884307.

[24] N. E. Miller *et al.*, "Estimating absolute and relative physical activity intensity across age via accelerometry in adults," *Journal of Aging and Physical Activity*, vol. 18, no. 2, pp. 158–170, 2010. doi: 10.1123/japa.18.2.158.

[25] A. H. K. Montoye *et al.*, "Development of cut-points for determining activity intensity from a wrist-worn actigraph accelerometer in free-living adults," *Journal of Sports Sciences*, vol. 38, no. 22, pp. 2569–2578, 2020. doi: 10.1080/02640414.2020.1794244.

[26] M. J. Duncan *et al.*, "Using accelerometry to classify physical activity intensity in older adults: What is the optimal wear-site?" *European Journal of Sport Science*, vol. 20, no. 8, pp. 1131–1139, 2020. doi: 10.1080/17461391.2019.1694078.

[27] M. N. Ahmadi and S. G. Trost, "Device-based measurement of physical activity in pre-schoolers: Comparison of machine learning and cut point methods," *PLoS One*, vol. 17, no. 4, 2022. doi: 10.1371/journal.pone.0266970.

[28] V. Farrahi and M. Rostami, "Machine learning in physical activity, sedentary, and sleep behavior research," *Journal of Activity, Sedentary and Sleep Behaviors*, vol. 3, no. 1, p. 5, 2024. doi: 10.1186/s44167-024-00045-9.

[29] R. O'Driscoll *et al.*, "Comparison of the validity and generalizability of machine learning algorithms for the prediction of energy expenditure: Validation study," *JMIR mHealth and uHealth*, vol. 9, no. 8, Aug. 2021. doi: 10.2196/23938.

[30] A. H. K. Montoye *et al.*, "Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer," *Journal of Applied Physiology*, vol. 124, no. 5, pp. 1284–1293, 2018. doi: 10.1152/japplphysiol.00760.2017.

[31] J. Chong *et al.*, "Machine-learning models for activity class prediction: A comparative study of feature selection and classification algorithms," *Gait & Posture*, vol. 89, pp. 45–53, 2021. doi: 10.1016/j.gaitpost.2021.06.017.

[32] V. Farrahi *et al.*, "Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches," *Gait Posture*, vol. 68, pp. 285–299, 2019. doi: 10.1016/j.gaitpost.2018.12.003.

[33] S. Liu, R. X. Gao, and P. S. Freedson, "Computational methods for estimating energy expenditure in human physical activities," *Medicine and Science in Sports and Exercise*, vol. 44, no. 11, pp. 2138–2146, 2012. doi: 10.1249/MSS.0b013e31825e825a.

[34] K. Ellis *et al.*, "Hip and wrist accelerometer algorithms for free-living behavior classification," *Medicine and Science in Sports and Exercise*, vol. 48, no. 5, pp. 933–940, 2016. doi: 10.1249/MSS.0000000000000840.

[35] M. de Almeida Mendes *et al.*, "Calibration of raw accelerometer data to measure physical activity: A systematic review," *Gait Posture*, vol. 61, pp. 98–110, 2018. doi: 10.1016/j.gaitpost.2017.12.028.

[36] S. G. Trost *et al.*, "Artificial neural networks to predict activity type and energy expenditure in youth," *Medicine and Science in Sports and Exercise*, vol. 44, no. 9, pp. 1801–1809, 2012. doi: 10.1249/MSS.0b013e318258ac11.

[37] A. H. K. Montoye *et al.*, "Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior," *AIMS Public Health*, vol. 3, no. 2, pp. 298–312, 2016. doi: 10.3934/publichealth.2016.2.298.

[38] S. G. Trost *et al.*, "Sensor-enabled activity class recognition in preschoolers: Hip versus wrist data," *Medicine and Science in Sports and Exercise*, vol. 50, no. 3, pp. 634–641, 2018. doi: 10.1249/MSS.0000000000001460.

[39] S. Li *et al.*, "Calibrating wrist-worn accelerometers for physical activity assessment in preschoolers: Machine learning approaches," *JMIR Formative Research*, vol. 4, no. 8, e16727, 2020. doi: 10.2196/16727.

[40] C. B. Thornton, N. Kolehmainen, and K. Nazarpour, "Using unsupervised machine learning to quantify physical activity from accelerometry in a diverse and rapidly changing population," *PLOS Digital Health*, vol. 2, no. 4, e0000220, 2023. doi: 10.1371/journal.pdig.0000220.

[41] J. E. Sasaki *et al.*, "Performance of activity classification algorithms in free-living older adults," *Medicine and Science in Sports and Exercise*, vol. 48, no. 5, pp. 941–950, 2016. doi: 10.1249/MSS.0000000000000844.

[42] A. K. Chowdhury *et al.*, "Prediction of relative physical activity intensity using multimodal sensing of physiological data," *Sensors*, vol. 19, no. 20, 2019. doi: 10.3390/s19204509.

[43] N. Nnamoko *et al.*, "Personalised accelerometer cut-point prediction for older adults' movement behaviours using a machine learning approach," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106 165, 2021. doi: 10.1016/j.cmpb.2021.106165.

[44] P. R. Hibbing *et al.*, "Estimating energy expenditure with actigraph gt9x inertial measurement unit," *Medicine and Science in Sports and Exercise*, vol. 50, no. 5, pp. 1093–1102, 2018. doi: 10.1249/MSS.0000000000001532.

[45] D. M. Lazar *et al.*, "Statistical learning methods to predict activity intensity from body-worn accelerometers," *Journal of Biomedical Analytics*, vol. 3, no. 1, pp. 27–50, 2020.

[46] J.-M. Redouté, *An introduction to ppg, spo, ipg and bis (and continuous bp measurements): Bioelectronics*, Course material, [Lecture notes,University of Liège], 2023.

[47] Delsys Incorporated, *Trigno wireless biofeedback system user's guide*, 2021.

[48] F. Liu, A. A. Wanigatunga, and J. A. Schrack, "Assessment of physical activity in adults using wrist accelerometers," *Epidemiologic Reviews*, vol. 43, no. 1, pp. 65–93, 2022. doi: 10.1093/epirev/mxab004.

[49] K. Norton, L. Norton, and D. Sadgrove, “Position statement on physical activity and exercise intensity terminology,” *Journal of Science and Medicine in Sport*, vol. 13, no. 5, pp. 496–502, 2010. doi: 10.1016/j.jsams.2009.09.008.

[50] *Compendium of physical activities 2024*, Accessed: 17 August 2025. [Online]. Available: https://pacompendium.com..

[51] H. Tanaka, K. D. Monahan, and D. R. Seals, “Age-predicted maximal heart rate revisited,” *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153–156, 2001. doi: 10.1016/s0735-1097(00)01054-8.

[52] F. Attal *et al.*, “Physical human activity recognition using wearable sensors,” *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015. doi: 10.3390/s151229858.

[53] L. Owen, *Hyperparameter tuning with Python : boost your machine learning model’s performance via hyperparameter tuning*. Packt Publishing, 2022, isbn: 9781803241944.

[54] scikit-learn developers, *Sklearn.preprocessing.standardscaler*, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, Accessed: 24 June 2025.

[55] T. Pavey *et al.*, “Field evaluation of a random forest activity classifier for wrist-worn accelerometer data,” *Journal of Science and Medicine in Sport*, vol. 20, no. 1, pp. 75–80, 2017. doi: 10.1016/j.jsams.2016.06.003.

[56] A. H. K. Montoye *et al.*, “Wrist-independent energy expenditure prediction models from raw accelerometer data,” *Physiological Measurement*, vol. 37, no. 10, pp. 1770–1784, 2016. doi: 10.1088/0967-3334/37/10/1770.

[57] scikit-learn developers, *Supervised Neural Networks - Scikit-learn Documentation*, https://scikit-learn.org/stable/modules/neural_networks_supervised.html, Accessed June 5, 2025.

[58] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition* (Springer Series in Statistics), 2nd ed. 2009. Springer New York, 2009, isbn: 9780387848587.

[59] G. Naidu, T. Zuva, and E. M. Sibanda, “A review of evaluation metrics in machine learning algorithms,” in *Artificial Intelligence Application in Networks and Systems*, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2023, pp. 15–25, isbn: 978-3-031-35314-7.