NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

FÁBIO BARBOSA MENDES

Bachelor of Science in Biomedical Engineering

TIME-SERIES ANALYSIS BASED ON MACHINE LEARNING FOR OCCUPATIONAL RISK EVALUATION IN PUBLIC ADMINISTRATION

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

TIME-SERIES ANALYSIS BASED ON MACHINE LEARNING FOR OCCUPATIONAL RISK EVALUATION IN PUBLIC ADMINISTRATION

FÁBIO BARBOSA MENDES

Bachelor of Science in Biomedical Engineering

Adviser: Doctor Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

Co-adviser: Doctor Cátia Cepeda de Vela Bastos
Post-Doctoral Research Fellow, NOVA University Lisbon

Examination Committee

Chair: Doctor Ricardo Nuno Pereira Verga e Afonso Vigário
Associate Professor, NOVA University Lisbon

Rapporteur: Doctor David José da Silva Aresta Belo
Researcher, Fraunhofer Portugal

Adviser: Doctor Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

Time-series analysis based on machine learning for occupational risk evaluation in public administration

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.15) [1].

Acknowledgements

After these five years studying to become a Biomedical Engineer, I could not be more proud of my choice and my beautiful journey.

Firstly, I would like to thank my adviser, Professor Hugo Gamboa, for the opportunity to enter the world of biosignals and to join a project of such impact as PrevOccupAI, which allowed me to grow as a student and as a person. His advice and the will to always do better were very important for me while carrying out this work.

I would also like to express my deepest gratitude to my co-adviser, Cátia Cepeda, whose support, advice, motivation, and patience were vital for me. Thank you for always being available to answer my questions and discuss new ideas.

I have to extend my sincere thanks to all the LIBPhys team, especially to Phillip Probst. Their warm welcome and constant willingness to share their expertise were crucial to my enormous learning while developing this thesis.

I am also very grateful for all the friendships I made during my journey at FCT NOVA. In particular, special thanks to Carolina Nascimento and Carolina Moraes, who were always with me in good and bad times. They have always believed in my capabilities and without them this would have not been possible.

Lastly, I would be remiss in not mentioning my family, especially my mother, my father, my brother, and my grandparents. All of them have always trusted my potential to achieve great things. A huge thank you for the support throughout my life.

Abstract

Occupational diseases are currently a concerning problem for office workers, who spend long periods of time seated in static positions. Musculoskeletal disorders, specifically, have the highest prevalence among workers, contributing negatively by 17% to the Years Lived with Disability worldwide.

This work is part of the PrevOccupAI project, which monitors office workers through wearable sensors and questionnaires, in order to provide them reports that bring to their attention some risk factors that can potentiate occupational diseases. During this work, a study with 40 subjects working in a real environment was carried out.

After data pre-processing and synchronization, as it was only intended to analyze sitting data, the periods in which the participants were not seated were removed from the acquired signals. For this purpose, a machine learning model was developed, which uses features from the smartphone's accelerometer signal to distinguish between sitting and walking. The best model reached an accuracy of 100.0%.

Additionally, a model capable of partially predicting the participants' answers to daily pain questionnaires was developed. Using the electromyography signals and personal information gathered from other questionnaires, it was possible to train a model that predicts if the subject reported pain or not, both at the beginning and end of the working day. Using the Random Forest algorithm, it was possible to achieve a mean accuracy of 86.3%.

For each acquisition performed by the 40 participants, a relative ergonomic occupational risk was assigned through variables that characterize postural variability. Using machine learning algorithms, models were trained to attempt to predict the modelled risk. A mean accuracy of 65.7% was achieved for the classification model, and a mean absolute error of 0.84 for the regression model.

Keywords: Occupational risk, Office work, Machine learning, Ergonomics, Pain prediction, Inertial sensors, Electromyography

Resumo

As doenças ocupacionais são, atualmente, um problema preocupante em trabalhadores de escritório, que passam muito tempo sentados em posições estáticas. As doenças musculoesqueléticas, especificamente, são as que têm maior prevalência entre os trabalhadores, contribuindo negativamente em 17% para os Anos Vividos com Incapacidade.

Esta dissertação é parte do projeto PrevOccupAI, que monitoriza trabalhadores de escritório através de sensores e questionários, de forma a fornecer-lhes relatórios que chamem à sua atenção alguns dos fatores de risco que podem potenciar doenças ocupacionais. Durante este trabalho, foi realizado um estudo em 40 sujeitos a trabalhar em contexto real.

Depois de pré-processamento e sincronização dos dados, como só se pretendia analisar dados de trabalhadores sentados, os períodos em que os participantes não estiveram sentados foram retirados dos sinais adquiridos. Para isso, foi desenvolvido um modelo de aprendizagem automática, que usa características do sinal do acelerómetro do telemóvel para distinguir entre sentado e a andar. O melhor modelo atingiu uma exatidão de 100,0%.

Adicionalmente, foi desenvolvido um modelo capaz de prever parcialmente as respostas dos participantes a questionários diários de dor. Através dos sinais de eletromiografia e informação pessoal retirada de outros questionários, foi possível treinar um modelo que prevê se o sujeito reportou dor ou não, tanto no início como no fim do dia de trabalho. Utilizando o algoritmo de Floresta Aleatória, foi possível atingir uma exatidão média de 86,3%.

A cada aquisição realizada pelos 40 participantes foi atribuído um risco ocupacional ergonómico relativo, através de variáveis que caracterizam a variabilidade postural. Utilizando algoritmos de aprendizagem automática, foram treinados modelos para tentar prever o risco modelado. Para o modelo de classificação, atingiu-se uma exatidão média de 65,7%, enquanto que para o modelo de regressão se conseguiu que o erro médio absoluto não ultrapassasse 0,84.

Palavras-chave: Risco ocupacional, Trabalho de escritório, Aprendizagem automática, Ergonomia, Previsão de dor, Sensores inerciais, Eletromiografia

Contents

List of Figures **xv**

List of Tables **xvii**

Acronyms **xix**

1 Introduction **1**
1.1 Context and motivation 1
1.2 Objectives 2
1.3 Structure 3

2 Theoretical concepts **5**
2.1 Musculoskeletal disorders 5
2.2 Sensors 6
2.2.1 Inertial sensors 6
2.2.1.1 Accelerometer 7
2.2.1.2 Gyroscope 8
2.2.2 Magnetometer 9
2.2.3 Electromyography 10
2.3 Machine learning 11
2.3.1 Random forest 13
2.3.2 Support vector machine 13
2.3.3 Logistic regression 14
2.3.4 Linear regression 14

3 State of the art **15**
3.1 Sensors 15
3.1.1 Human activity recognition 15
3.2 Occupational risk assessment 16
3.2.1 Machine learning in occupational risk assessment 17

4 Data acquisition **19**
4.1 Main acquisitions . . . 19
4.2 Complementary acquisitions . . . 24
4.2.1 Sit-stand transitions detection . . . 25
4.2.2 Sitting and walking distinction . . . 25

5 Methods **27**
5.1 Computational methods . . . 27
5.2 Data synchronization . . . 27
5.3 Sitting detection . . . 29
5.3.1 Sit-stand transitions detection . . . 29
5.3.2 Sitting and walking distinction . . . 31
5.3.2.1 Data pre-processing . . . 31
5.3.2.2 Dataset characterization . . . 31
5.3.2.3 Classification models . . . 32
5.3.2.4 Feature extraction and selection . . . 32
5.3.2.5 Model optimization . . . 34
5.4 Pain prediction . . . 35
5.4.1 Electromyography pre-processing . . . 35
5.4.2 Resting and moving distinction . . . 36
5.4.3 Dataset characterization . . . 38
5.4.4 Classification models . . . 39
5.4.5 Label generation . . . 39
5.4.6 Feature extraction and selection . . . 39
5.4.7 Model optimization . . . 41
5.5 Occupational risk prediction . . . 41
5.5.1 Risk modelling using postural variables . . . 42
5.5.2 Risk classification and regression models . . . 44
5.5.3 Dataset characterization . . . 45
5.5.4 Label generation . . . 45
5.5.5 Feature extraction and selection . . . 47
5.5.6 Model optimization . . . 48

6 Results and discussion **51**
6.1 Sitting detection . . . 51
6.1.1 Sit-stand transitions detection . . . 51
6.1.1.1 Transition patterns . . . 51
6.1.1.2 Transition detection with z-axis patterns of different subjects . . . 51
6.1.1.3 Transition detection with multi-axis patterns of a single subject . . . 53

6.1.1.4 Transition detection with multi-axis patterns of different subjects . . . 53
6.1.1.5 Discussion . . . 53
6.1.2 Sitting and walking distinction . . . 55
6.1.2.1 Model based on accelerometer and gyroscope data . . . 55
6.1.2.2 Model based on accelerometer data . . . 56
6.1.2.3 Discussion . . . 57
6.2 Pain prediction . . . 59
6.2.1 Model based on electromyography data . . . 60
6.2.2 Model based on electromyography and personal data . . . 60
6.2.3 Discussion . . . 62
6.3 Occupational risk prediction . . . 64
6.3.1 Risk modelling . . . 64
6.3.2 Classification models . . . 64
6.3.3 Regression models . . . 69
6.3.4 Discussion . . . 72

7 Conclusion **75**
7.1 Main conclusions . . . 75
7.2 Future work . . . 76

Bibliography **77**

List of Figures

2.1 Typical accelerometer signal of a seated person. 7
2.2 Beam-type triaxial accelerometer. 8
2.3 Typical gyroscope signal of a seated person. 8
2.4 Typical magnetometer signal of a seated person. 9
2.5 Motor unit action potential measured by EMG. 10
2.6 Example of a confusion matrix of a binary classifier, where *0* corresponds to the negative class and *1* to the positive class. 12

4.1 Screenshot of the PrevOccupAI homepage (web version). 20
4.2 Human body representations in the frontal plane, used in the daily pain questionnaire. 21
4.3 Color scale used in the daily pain questionnaire. 21
4.4 Smartphone placement for data acquisition. 22
4.5 Smartwatch placement for data acquisition. 22
4.6 MuscleBAN placement for data acquisition. 22
4.7 Screenshot of the PrevOccupAI acquisition interface (mobile version). 24

5.1 Example of signal cropping from the time axis of different smartphone sensors. 28
5.2 Example of an acquisition log file. 36
5.3 Example of an EMG signal before processing (top), after filtering (middle), and after normalization (bottom). 37
5.4 Example of an acceleration magnitude signal, where the red line marks the threshold for moving classification, defined as 0.5 m/s^2. 37
5.5 Considered body parts (painted in blue) for pain prediction. 40
5.6 Coordinate system of the smartphone. 43
5.7 Division of the xz-plane to define the possible postures. 44
5.8 Example of a CDF, where the risk zones are marked in red, when $a > 0.02$. 46
5.9 Example of a CDF, where the risk zones are marked in red, when $a \leq 0.02$. 46

6.1 Acquired patterns of the sit-to-stand (left) and stand-to-sit (right) transitions. 52

6.2 Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's z-axis and three different subjects. 52
6.3 Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's three axes. 54
6.4 Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's three axes and three different subjects. 54
6.5 Distribution of the variable referring to the mean velocity of the subject's third most common posture. 65
6.6 Distributions of the 12 selected variables for risk modelling. 66
6.7 Distributions of the 12 selected variables for risk modelling (cont.). 67
6.8 Predicted and real values for the linear regression model with lower MAE. . 71
6.9 Predicted and real values for the RF regression model with lower MAE. . . 71
6.10 Predicted and real values for the support vector regression model with lower MAE. 72

List of Tables

2.1 List of the most common MSD. 6

4.1 List of questionnaires answered by the participants of the main study. 20

5.1 Number of samples per class for sitting/walking classification. 32
5.2 Extracted features of statistical domain. 33
5.3 Extracted features of temporal domain. 33
5.4 Characterization of the classification algorithms used for RFE. 34
5.5 Values used for each hyperparameter when performing grid search. 35
5.6 Statistical analysis of the demographic data from the participants of the main study. 38
5.7 Number of samples per class by period of day for pain classification without personal data. 38
5.8 Number of samples per class by period of day for pain classification with personal data. 38
5.9 Extracted features of spectral domain. 40
5.10 Number of samples per class for the occupational risk regressor. 45
5.11 Number of samples per class for the occupational risk classifier. 45
5.12 Characterization of the regression algorithms used for RFE. 48
5.13 Values used for the regressors' hyperparameters when performing grid search. 48

6.1 Number of selected features by RFE for sitting/walking classification with accelerometer and gyroscope. 55
6.2 Characterization of the models used for sitting/walking classification with accelerometer and gyroscope. 56
6.3 Statistical analysis of the sitting/walking classifier's accuracies with accelerometer and gyroscope. 56
6.4 Number of selected features by RFE for sitting/walking classification using only accelerometer. 57

6.5 Characterization of the models used for sitting/walking classification using only accelerometer. 58
6.6 Statistical analysis of the sitting/walking classifier's accuracies using only accelerometer. 58
6.7 Number of selected EMG features by RFE for pain prediction. 60
6.8 Characterization of the models used for pain prediction without personal data. 61
6.9 Statistical analysis of the pain classifier's accuracies without personal data. . 61
6.10 Number of selected features by RFE for pain prediction, considering the personal data. 61
6.11 Characterization of the models used for pain prediction, considering the personal data. 62
6.12 Statistical analysis of the pain classifier's accuracies when considering the personal data. 62
6.13 Statistical analysis of the extracted postural variables. 65
6.14 Definition of risk zones for each of the selected variables, where x represents the value of the variable. 67
6.15 Number of selected features by RFE for the occupational risk classifier. . . . 68
6.16 Characterization of the classification models used for occupational risk prediction. 68
6.17 Statistical analysis of the occupational risk classifier's accuracies. 69
6.18 Statistical analysis of the occupational risk classifier's precisions. 69
6.19 Statistical analysis of the occupational risk classifier's recalls. 69
6.20 Statistical analysis of the occupational risk classifier's F_1 scores. 70
6.21 Number of selected features by RFE for the occupational risk regressor. . . . 70
6.22 Characterization of the regression models used for occupational risk prediction. 70
6.23 Statistical analysis of the occupational risk regressor's MAEs. 70
6.24 Results of the statistical tests performed to verify the fulfillment of the assumptions for linear regression. 71

Acronyms

3D Three-dimensional *(pp. 8, 17, 42)*
4D Four-dimensional *(p. 42)*

AT Autoridade Tributária e Aduaneira *(pp. 2, 19)*

BMI Body Mass Index *(pp. 40, 41, 60)*

CDF Cumulative Distribution Function *(pp. 46, 47, 64)*

DGS Direção-Geral da Saúde *(p. 2)*

EMG Electromyography *(pp. 5, 6, 10, 11, 15, 17, 18, 21, 29, 35, 36, 39, 40, 47, 59–64, 68, 70, 72, 73, 75, 76)*

GDP Gross Domestic Product *(p. 1)*
GPS Global Positioning System *(p. 15)*

HAR Human Activity Recognition *(pp. 15, 16)*

ILO International Labour Organization *(p. 1)*
IPAQ International Physical Activity Questionnaire *(pp. 40, 41, 60)*

LSTM Long Short-Term Memory *(p. 76)*

MAC Media Access Control *(pp. 35, 36)*
MAE Mean Absolute Error *(pp. 13, 48, 69, 73, 74, 76)*
MEMS Microelectromechanical Systems *(pp. 8–10, 15)*
ML Machine Learning *(pp. 2, 3, 5, 11, 12, 15–18, 31, 32, 34, 35, 39, 41, 42, 44, 47, 48, 55, 74, 75)*

MSD Musculoskeletal Disorders *(pp. 1, 3, 5, 16, 17, 75)*
MVC Maximum Voluntary Contraction *(p. 36)*

RF Random Forest *(pp. 3, 13, 16, 17, 32, 34, 35, 39, 44, 47, 48, 55–63, 68–70, 72, 73, 76)*
RFE Recursive Feature Elimination *(pp. 32, 41, 47, 55, 57, 60, 68)*
RNN Recurrent Neural Networks *(pp. 74, 76)*

SVM Support Vector Machine *(pp. 3, 13, 14, 16–18, 32, 34, 35, 39, 44, 47, 55–63, 68–70, 72, 73)*

TSFEL Time Series Feature Extraction Library *(pp. 27, 32, 39)*

VIF Variance Inflation Factor *(pp. 47, 48)*

WHO World Health Organization *(p. 1)*
WRD Work-related Disorder *(pp. 1, 2, 19)*

YLD Years Lived with Disability *(p. 5)*

1 Introduction

1.1 Context and motivation

Most people spend a lot of their time at work. Therefore, it is essential that good working conditions are ensured, to provide workers a safe working environment in which they feel confident and where occupational risks are kept at a minimum. This contributes to people's well-being, as well as productivity and economic development [2].

An occupational disease or Work-related Disorder (WRD) can be seen as any health condition mainly caused by exposure to risk factors at the work environment [3]. These risks may include heat, noise, chemicals, bad postures, and dangerous machines, and can directly cause diseases or aggravate some health conditions. Additionally, factors like stress, hierarchy, social relations at work, and inadequate wages can also contribute negatively to the health of workers [2].

The occupational health condition with the highest incidence in Europe is Musculoskeletal Disorders (MSD) [4]. Other common problems include chronic respiratory diseases, hearing loss, depression, and occupational cancer [2].

In Portugal, the most common work-related health problems reported by workers are back pain, muscle pain, joint pain, and general fatigue. Moreover, shift and night workers are quite prone to depression, anxiety, and insomnia. The physical risk factors that most affect Portuguese workers are repetitive upper limb movements (83.2%), standing for long periods (71.1%), inadequate positions (46.8%), working with computers/screens (46.5%), and sitting for long periods (42.5%). From a psychosocial point of view, 68.4% of the Portuguese workers report an accelerated work rhythm, 60% report strict deadlines, and 31.9% work under stress [5].

According to the World Health Organization (WHO), 70% of all workers do not have monetary compensation in case of an occupational disease [2]. Furthermore, work-related conditions can decrease the Gross Domestic Product (GDP) of most countries by 4 to 6% [2], [4]. Additionally, the economic burden for the European Union countries caused by work injuries is estimated to be between approximately 2500 and 4500 € per employed person [6]. The International Labour Organization (ILO) estimates that WRDs cause more

than 2 million deaths every year, all around the world [4].

WRDs have consequences not only at an individual level, but also at the employer and society levels. These consequences include the reduction of the financial status, less career prospects, social isolation, early retirement, and loss of production [4]. However, nowadays, only a small part of workers around the world have access to specialized occupational health services [2].

Therefore, it is essential to monitor and prevent WRDs, creating a safe environment and promoting health at work. Strategies such as the removal of pollution sources, noise reduction, better ventilation, the improvement of the workplace conception, and the education on keeping a proper posture can be very effective to reduce the risk of one of the aforementioned diseases [2], [5].

This thesis is part of the PrevOccupAI project [7], which has the support of the Portuguese Autoridade Tributária e Aduaneira (AT) and Direção-Geral da Saúde (DGS). This project aims to prevent occupational diseases in the office context, through the identification of risk factors, in order to promote occupational health. As part of the project, a cross-platform application was developed [8], which takes into account three main domains of risk factors for WRDs: psychosocial, biomechanical, and environmental. It is composed of various mini apps that collect data through questionnaires, sensors from mobile devices, such as smartphones and smartwatches, and wearable biosensors. The application is used as a means to monitor risk factors and provide individual reports to office workers, in order to educate them and reduce the risks of WRDs.

1.2 Objectives

The main goal of this thesis is to assess occupational risks, based on biosignals and Machine Learning (ML) models, in order to reduce the incidence of occupational diseases in office workers. These diseases are a growing problem that could benefit from the computational tools available nowadays, namely ML techniques.

Particularly, the major objectives are the following:

- Data collection in both laboratory and non-laboratory settings;
- Development of tools to process inertial and physiological signals;
- Feature extraction from the collected signals and training of ML models;
- Development of an algorithm to detect non-sitting periods;
- Development of a model capable of automatically assessing whether a person has pain or not;
- Modelling the occupational risk of public administration workers through postural variables;

- Development of a model that predicts the modelled occupational risk.

To meet the aforementioned objectives, different ML algorithms will be used, such as logistic regression, Random Forest (RF), and Support Vector Machine (SVM). Furthermore, several pre-processing and cross-validation techniques will be performed.

This thesis will help the development of the PrevOccupAI project, which takes into account various aspects of occupational diseases. This work will essentially focus on MSD, which are the occupational diseases with the highest incidence, especially in the office context.

1.3 Structure

This document is divided into seven chapters, which are subdivided into different sections. The present chapter (Chapter 1) presents the context and motivation that led to the development of this work, as well as its main objectives. Chapter 2 describes the main theoretical concepts essential to the understanding of this thesis. Chapter 3, on the other hand, presents a literature review.

Chapter 4 describes the data acquisitions that were performed to carry out this work, while Chapter 5 introduces and describes the methodology used to accomplish the objectives.

Finally, the results obtained using these methods are presented in Chapter 6 and the main conclusions are highlighted in Chapter 7.

2

Theoretical concepts

This chapter introduces the main theoretical concepts essential to the understanding of this work. These concepts include the Musculoskeletal Disorders (MSD), which are the occupational diseases that this work focuses on, the main sensors used for data acquisition, namely inertial sensors and Electromyography (EMG) sensor, and Machine Learning (ML), which is the basis of the models to develop.

2.1 Musculoskeletal disorders

MSD include a wide range of conditions that affect muscles, tendons, ligaments, joints, nerves, and blood vessels [9]. There are more than 150 MSD, some of them presented in Table 2.1. The most common ones are low back pain, neck pain, osteoarthritis, rheumatoid arthritis, and gout [10].

These diseases are currently the most common cause of disability, particularly lower back pain [11]. It is estimated that MSD contribute negatively by 17% to the Years Lived with Disability (YLD) worldwide, with approximately 149 million YLD [11].

MSD went unnoticed for a long time, because they were considered a normal part of ageing [12]. These type of conditions are commonly difficult to diagnose, because some of its symptoms are subjective (e.g., pain and stiffness) and the knowledge of experts varies. Its diagnosis is usually done by trial and error, which delays the process and often prevents an early diagnosis [9].

MSD are the most common occupational diseases, but they are not necessarily work-related and can be caused by other factors. They are, in fact, one of the leading causes to work absence [10].

In the office context, there is a high incidence of MSD, essentially due to prolonged periods of sitting, inappropriate postures, and repetitive movements of the upper limbs. Additionally, factors such as age, mental health, pressure at work, and job dissatisfaction can also be associated with MSD [13].

Table 2.1: List of the most common MSD. From [11].

Type	Musculoskeletal disorder
Joints	Osteoarthritis
	Rheumatoid arthritis
	Psoriatic arthritis
	Gout
	Ankylosing spondylitis
Bone	Osteoporosis
	Osteopenia (and associated fractures)
	Traumatic fractures
Muscles	Sarcopenia
Spine	Back pain
	Neck pain
Multiple areas	Regional and widespread pain disorders
	Connective tissue diseases
	Lupus erythematosus

2.2 Sensors

In this thesis, time-series of wearable sensors, such as accelerometer, gyroscope, magnetometer, and EMG sensor are analyzed. Therefore, these sensors are described in the following sections.

2.2.1 Inertial sensors

Inertial sensors are sensors capable of providing data regarding linear and angular motion, about one or more axes [14]. The most common ones are the accelerometer and the gyroscope.

A conventional inertial sensor is based on a mass-spring system, where a mass is suspended on a mechanical frame. There is an input force representative of the physical quantity to be measured, which causes a displacement on the mass, and that displacement is measured. Depending on the type of sensor, there are numerous transduction methods that transform a physical quantity into a force acting on the mass [15].

Along with the force that represents the quantity to be measured, there are also some undesired forces acting on the mass, such as forces that represent other types of motion and damping forces. A high-quality sensor must be capable of rejecting those unwanted forces [15].

In order to use a sensor for body motion measurement, it is imperative to define a coordinate system. If we have a moving coordinate system, the inertial force (F) is given by

$$F = -mA_0 + 2mu' \times \omega + m\omega \times (r' \times \omega) + mr' \times \frac{\mathrm{d}\omega}{\mathrm{d}t} \tag{2.1}$$

where A_0 and ω are the acceleration and angular velocity of the moving coordinate system,

and r' and u' are the position vector and the velocity of a mass m. The terms from the right side of the equation represent the acceleration corresponding to the linear inertial force and apparent forces, such as the Coriolis and centrifugal forces [16].

2.2.1.1 Accelerometer

According to Newton's second law, the force (F) that acts on a mass (m) is given by

$$F = m\alpha \tag{2.2}$$

where α is the acceleration of linear motion.

An accelerometer is an inertial sensor capable of measuring linear acceleration, usually in g or m/s^2 units, where 1 g equals approximately 9.81 m/s^2 and corresponds to earth's gravity. Therefore, a multi-axis accelerometer can measure the orientation of gravity and the linear acceleration due to motion [15]. Figure 2.1 shows an example of a triaxial accelerometer's signal. If the accelerometer data is integrated with respect to time, it is possible to obtain velocity and position. Therefore, in the opposite process, it is also possibly to obtain the acceleration, but that process is very susceptible to noise [16].

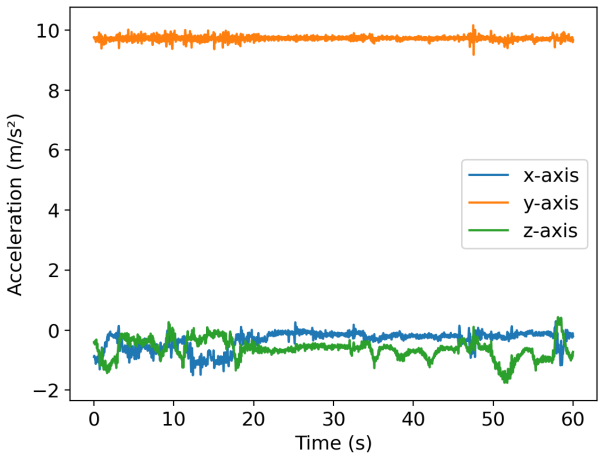

Figure 2.1: Typical accelerometer signal of a seated person.

The most common accelerometer is the beam-type one (illustrated in Figure 2.2), which is composed of an elastic beam fixed to a base at one end, with a seismic mass attached to the other end. The beam can be replaced by other elastic material, such as a spring or a diaphragm. Similarly to what was described in Section 2.2.1, the beam bends elastically in response to the force that accelerates the seismic mass. The displacement of the mass can be measured with different techniques, namely by capacitance or using a piezoresistive material [16]. A differential-capacitance accelerometer relies on the change of capacitance with the applied acceleration, providing very high sensitivity [14]. In a piezoresistive accelerometer, there is a cantilever beam that incorporates a piezoresistive

element. This way, the resistance changes along with the bending of the beam, in response to the acceleration [16].

Nowadays, most of the accelerometers are based on the Microelectromechanical Systems (MEMS) technology. A smartphone, for example, typically features a triaxial accelerometer based on this technology, which provides a small and sensitive system, capable of determining the amplitude and direction of the acceleration in a Three-dimensional (3D) space [16].

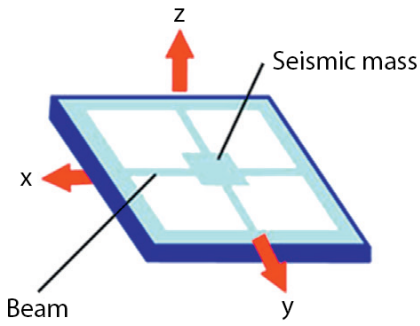

Figure 2.2: Beam-type triaxial accelerometer. Adapted from [16].

2.2.1.2 Gyroscope

A gyroscope is an inertial sensor capable of measuring the angular motion, usually in deg/s or rad/s units, as exemplified in Figure 2.3. If the gyroscope data is integrated, it is possible to obtain angle measurements.

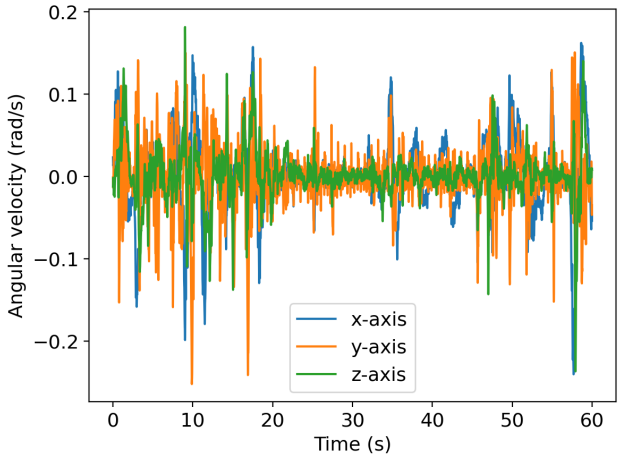

Figure 2.3: Typical gyroscope signal of a seated person.

A typical gyroscope consists of a spinning wheel mounted on a movable frame. When the axis direction is changed, a torque proportional to the rotation rate of the axis emerges. This principle can be used to measure the angular velocity [16].

However, a smartphone usually possesses a vibratory rate gyroscope, based on the MEMS technology. In this type of gyroscopes, there is a vibrating proof mass. When there is a rotation along an axis orthogonal to the vibratory axis, it causes a proportional Coriolis acceleration, which is measured [15]. The Coriolis acceleration (a_{cor}) is given by

$$a_{\text{cor}} = 2v_{\text{pm}} \times \Omega \tag{2.3}$$

where v_{pm} is the velocity of the proof mass and Ω is the rate of rotation [16].

This way, in order to calculate the rate of rotation, it is mandatory that the proof mass velocity is known. Hence, a vibratory rate gyroscope electronically forces the proof mass to oscillate in a direction parallel to the chip surface, ensuring a stable and known velocity. When there is rotation in the orthogonal direction, a Coriolis force causes a deflection of the mass in this direction. This deflection can produce a voltage that is measured, and the angular velocity can be determined [15], [16].

2.2.2 Magnetometer

A magnetometer is a sensor capable of measuring magnetic fields, such as the earth's magnetic field and the magnetic field generated in the presence of a magnetic material [17]. Nevertheless, magnetometers are often used together with accelerometers and gyroscopes to measure body movement and orientation, which is the principle behind the smartphone's rotation vector sensor [16], [17]. If we assume that the sensor does not travel a long distance compared to the size of the earth, the earth's magnetic field can be considered constant. This way, if there is not a magnetic material near the sensor, the magnetometer can give information about the orientation, together with the accelerometer/gyroscope data [17]. Figure 2.4 shows an example of a magnetometer signal.

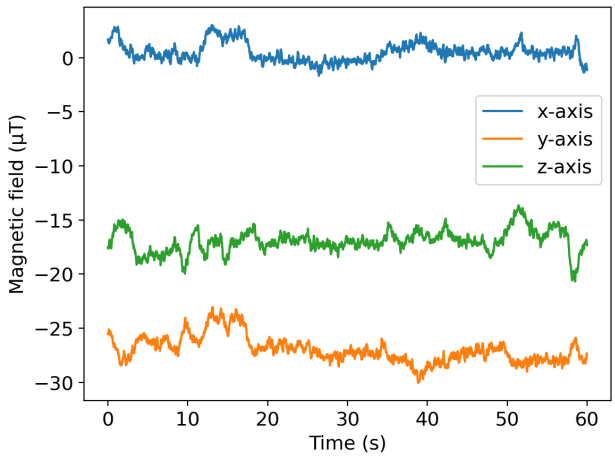

Figure 2.4: Typical magnetometer signal of a seated person.

Typical magnetometers are based on the Hall effect, which underlies the interaction of moving electrons with an external magnetic field. This interaction generates magnetic

impedance and resistance. An electron moving through a magnetic field is subjected to a force (F) given by

$$F = qvB \tag{2.4}$$

where q is the electronic charge, v is the electron speed and B is the magnetic field [16].

As well as the inertial sensors, the magnetometers present in smartphones are based on the MEMS technology. Usually, there are three orthogonal magnetometers in a smartphone or similar device [18].

2.2.3 Electromyography

The nervous system orchestrates muscle recruitment through electrical impulses. When an action potential travels down the axon of a neuron and reaches the neuromuscular synapse, it releases acetylcholine and sends a signal to the muscle. This phenomenon generates the motor unit action potential, which enables muscle contraction. The system composed of a motor neuron and its muscle fibers is called a motor unit [19], [20]. The number of motor units active and the discharge rate influence muscle force, but the contribution of these parameters varies with contraction type and between different muscles [19].

The motor unit action potential (electrical potential) is the source of the EMG signal, which basically measures the energy generated by a muscle, as illustrated in Figure 2.5. The electromyograph (instrument that provides the EMG signal) is considered a very sensitive voltmeter, as the muscle energy is usually in the microvolt (µV) range [20].

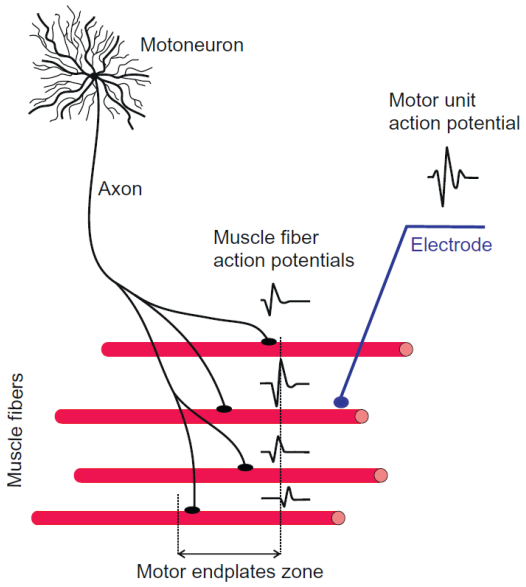

Figure 2.5: Motor unit action potential measured by EMG. From [21].

In intramuscular EMG, the needle-like electrodes are directly inserted in the muscle tissue, giving a more precise and strong signal. On the other hand, in surface EMG, the electrodes are positioned on the skin surface, above the muscle tissue. The latter is a more widely used technique, essentially due to its non-invasiveness [19].

Surface EMG is a safe, easy, and non-invasive method that allows to quantify the muscle energy at rest and its changes during movement [20]. The electrical potentials on the skin surface have very small amplitudes and typical frequencies between 15 and 400 Hz. This technique can give information about the intensity of muscle contraction, muscle fatigue, and recruitment of motor units [19].

The EMG signal can be acquired in a monopolar or bipolar configuration. In a monopolar EMG, the potential is measured in respect to a reference electrode at a bony region. However, this configuration can lead to electromagnetic interference or to the crosstalk phenomenon [19]. Crosstalk consists of the contamination of the record field of a muscle with energy from another distant muscle. This way, it may become impossible to isolate the signal from a certain muscle. This issue can be improved with certain electrode placements but, to this day, there is not an adopted standard for electrode placement in EMG [20]. A bipolar EMG results from the difference between two monopolar configurations and minimizes those interferences, but attenuates the contribution of deep motor units [19].

Surface EMG has some inconveniences, such as the limitation to the measurement of only a few muscles. Additionally, it is important to understand that the EMG signal is not a direct measure of muscle strength, but rather of electrical activity in the muscle. Finally, it is necessary to consider the impedance of the skin at the electrode site, that must be as low as possible [20]. For this purpose, the skin can be cleaned with alcohol to remove old skin particles and grease/dirt.

2.3 Machine learning

ML is a branch of artificial intelligence and corresponds to a set of algorithms that are able to learn from and find patterns in data. Currently, it has diverse fields of application, such as medicine, finance, entertainment, engineering, computer vision, and medicine [22].

One of the most well-accepted definitions for ML was presented by Tom Mitchell in 1997 as "A computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T, as measured by P, improves with experience E" [23].

ML algorithms are not explicitly programmed to produce a certain output but, in order to become better at performing a certain task, they automatically adapt their architecture with experience. This is called training and it consists of providing the algorithm with new samples of input data. The learning process does not have to be limited in time, it may occur during the algorithm lifetime [22].

According to its degree of supervision, ML can be divided into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, each training example fed to the algorithm is labeled with a corresponding class. Two of the most common supervised learning tasks are

classification (predict a discrete class) and regression (predict a continuous numeric value) [24].

In unsupervised learning, the training data does not have labels. Therefore, the algorithm has the freedom to find patterns in the presented data, because there is not a specific intended outcome [22]. The most popular unsupervised learning algorithms are the clustering algorithms [24].

In semi-supervised learning, data is partially labeled. While some examples are labeled, there are others that are not. This way, the labeled part can help the learning process with the unlabeled part [22]. Most of the semi-supervised algorithms are combinations of supervised and unsupervised algorithms [24].

Finally, in reinforcement learning, the learning system observes the input data and performs a certain action, receiving either a reward or a penalty according to the generated output. Its objective is to find a strategy that maximizes the rewards received [24].

In order to evaluate the performance of a ML algorithm and its ability to generalize to examples it has never seen, there are many ways to divide the dataset into training and testing sets. If the dataset is large and has all cases well represented, it is possible to use the holdout method, in which there is a random sampling between a training and a testing group (usually 70% for training and 30% for testing) [25].

When the dataset is not so large, the most used method is *k*-fold cross-validation. In this method, the dataset is first divided into *k* subsets. Then, *k*-1 subsets are used for training and the other one is used for testing. The process is repeated *k* times, where in each of the iterations a different subset is used for testing, with the remaining ones being used for training. If the dataset is too small, it is common to set *k* as the size of the dataset, and the *k*-fold cross-validation turns into the leave-one-out method [25].

To evaluate the performance of a classifier, it is important to compute and visualize its confusion matrix. This matrix allows the comparison between the classes predicted by the classifier and the actual targets (labels). Figure 2.6 presents an example of a confusion matrix of a binary classifier.

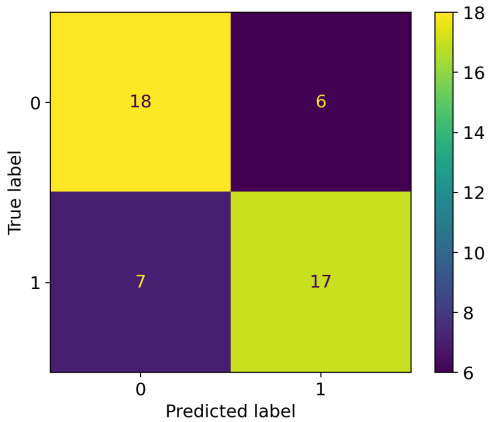

Figure 2.6: Example of a confusion matrix of a binary classifier, where *0* corresponds to the negative class and *1* to the positive class.

The confusion matrix can be used to calculate some performance measures of the classifier, such as

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.5}$$

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$recall = \frac{TP}{TP + FN} \tag{2.7}$$

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{FN+FP}{2}} \tag{2.8}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively [24]. In the example given in Figure 2.6, the accuracy is 73%, the precision is 74%, the recall is 71%, and the F_1 score is 72%.

Regarding regression tasks, it is not possible to build a confusion matrix, because the predicted values are continuous. Therefore, a regressor cannot be evaluated by the same perfomance measures as a classifier. For regression, the Mean Absolute Error (MAE) is one of the typical performance measures and it is given by

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |p_i - r_i| \tag{2.9}$$

where m is the number of samples, and p_i and r_i are the predicted and real values of each sample, respectively [24].

2.3.1 Random forest

The Random Forest (RF) algorithm involves a large number of individual and independent decision trees. A decision tree is a rule-based and tree-like model that recursively splits data into subdomains, using nodes, with each leaf node representing a classification [26], [27]. Each decision tree makes its own classification, and the RF's final classification is usually the class assigned by most trees [24].

The RF regressor is very similar the RF classifier, but it is used to predict a continuous numeric value. It is also composed of numerous individual decision trees, but its final output is the average of the outputs of each decision tree. Additionally, the node separation criterion is also different, usually the mean squared residual [28].

2.3.2 Support vector machine

A Support Vector Machine (SVM) uses a line or hyperplane to separate the different classes, ensuring that it stays as far away as possible from the closest training instances. If the dataset is not linearly separable, other kernels can be used to separate it, such as polynomial, sigmoid, and radial basis function. When classifying a new instance, the SVM assigns the class corresponding to the region where the instance fits [24].

Support vector regression works the opposite way to SVM and predicts a continuous numeric value. Instead of using the hyperplane to ensure the instances of each class are as far away as possible, it tries to maintain the instances of each class as close to the hyperplane as possible [24].

2.3.3 Logistic regression

The logistic regression, despite being a regression, is usually employed for classification tasks, estimating the probability of a sample belonging to the positive class. This probability is called the logistic (sigmoid function) of a weighted sum of the input features. If it is greater than 50%, the model predicts that that sample belongs to the positive class [24].

2.3.4 Linear regression

Linear regression tries to find a linear relationship between the independent and dependent variables in the input data. It consists of a weighted sum of the input features plus a constant term (bias) [24]. If there is only one independent variable, it is called simple linear regression. On the other hand, if there are multiple independent variables, it is called multivariate linear regression. When this model is used, it is assumed that there is in fact a linear relationship in the input data [29].

3

State of the art

This chapter presents a review of the available literature regarding inertial sensors, Electromyography (EMG), Machine Learning (ML) and occupational risk assessment, which are essential for the work to be carried out.

3.1 Sensors

Historically, inertial sensors have been used exclusively for Global Positioning System (GPS) applications. However, the development of sensors based on the Microelectromechanical Systems (MEMS) technology allowed to attach them to the human body, resulting in a wide range of new applications [30]. Nowadays, inertial sensors have been widely used for fall detection [31], and human movement and activity recognition [32], [33] in devices such as smartphones and smartwatches. Furthermore, applications in which inertial sensor data has been used as input to ML algorithms has become common [34]–[36].

Taking into account the EMG sensor, besides being used on the clinical environment to help diagnose neuromuscular diseases and other disorders, it is also used in a vast number of other applications. These include prosthesis control [37], evaluation of movements [38], and evaluation of muscular fatigue [39]. As with inertial sensors, EMG data is also frequently used to feed ML algorithms, for diverse types of applications [40]–[42].

3.1.1 Human activity recognition

Human Activity Recognition (HAR) is a well-developed field of research which uses sensors from mobile devices (e.g., smartphone and smartwatch), together with ML techniques, in order to identify the activities that a subject is performing [43].

In this context, in 2015, Filios et al. [43] developed an Android application to detect different activities in real-time, including sitting, lying, walking, and running. The authors tested different protocols and ML algorithms, such as *k*-nearest neighbors, naive Bayes, and decision tree. Using both accelerometer and gyroscope, they achieved a maximum accuracy of 95.7%, while using only accelerometer they achieved a maximum accuracy of 99.4%.

In 2016, Capela et al. [44] built a decision-tree classifier for HAR (standing, sitting, lying, walking, climbing stairs, and small movements), testing its performance on able-bodied and stroke participants. The authors found that, although it is possible to use the system for both populations, the HAR performance was lower for participants who suffered a stroke.

In 2018, San-Segundo et al. [45] used the accelerometer data from smartphones and smartwatches to HAR, testing the effects of different approaches on the performance of the classifier. The authors used the Heterogeneity Human Activity Recognition Database, which includes sitting, standing, biking, walking, walking upstairs, and walking downstairs. Different devices, features, feature selection methods, filters, normalization techniques, and ML algorithms were tested, in order to obtain a model as robust as possible. Using smartphone data, a maximum accuracy of 99.3% was obtained for the Random Forest (RF) model.

In 2018, Mekruksavanich et al. [46] used the accelerometer and gyroscope data from a smartwatch to distinguish between sitting, standing, lying, walking, walking upstairs, and walking downstairs. The authors built their own database of 600 samples and tested decision trees, Support Vector Machine (SVM)s, neural networks, and different stacking techniques. Regarding specifically sitting detection, the SVM model achieved 86.9% accuracy when using accelerometer and gyroscope, and 85.9% when using only accelerometer. On the other hand, the decision tree model achieved accuracies of 93.2% (accelerometer and gyroscope) and 91.9% (accelerometer).

3.2 Occupational risk assessment

In the occupational context, there are some publications on the use of observational and data mining techniques to risk assessment. Coggon et al. [13], using questionnaire and statistical data of 18 countries, linked the burden of Musculoskeletal Disorders (MSD) with age, physical workload, work pressure, and some adverse beliefs about the influence of work on MSD. Liao and Perng [47] used associated rule mining to find the most prone conditions to occupational accidents in the construction sector in Taiwan.

Regarding sensors in the occupational context, there are some non-ML techniques reported. Frigerio and Virga [48] tested a smartphone app that monitors noise in the workplace and proposed a system to alert workers when the noise exceeds safety levels. Nath et al. [49] presented an approach which uses smartphone sensors to monitor construction workers' posture and identify work-related ergonomic risks.

Caputo et al. [50] evaluated angles and postures during movement in industrial workers by using inertial sensors in different body segments. Cerqueira et al. [51] developed a smart vest that performs ergonomic risk assessment and alerts the worker through biofeedback.

Carnide et al. [52] assessed possible causes for MSD through questionnaires and clinical exams, including EMG. Lenzi et al. [53] developed a toolbox to support expert video analysis of manual handling of low loads at high frequency, through the use of inertial sensors.

3.2.1 Machine learning in occupational risk assessment

There is also a wide range of ML applications in the occupational area, even without the use of wearable sensors. Sarkar et al. [54] applied clustering techniques to text descriptions of occupational accidents and used neural networks to predict its risk in steel plants. Dalal and Bassu [55] prototyped a system that fuses the aforementioned technique with cameras, audio, and tablets to produce risk maps through deep learning and identify occupational threats in real time.

Recently, Liu et al. [56] started to develop a system that uses deep learning to analyze video frames, building a Three-dimensional (3D) model of the worker and detecting his posture, in order to determine the risk of MSD. There are also quite a few publications on the use of neural networks to predict the incidence of pneumoconiosis in the mining industry [57]–[59].

Di Noia et al. [60], [61] used different ML techniques, together with genetic algorithms, to assess the risk of different occupational diseases, based on age, gender, type of job, and work time. Cheng et al. [62], [63] applied decision trees to occupational accidents databases, in order to find the most prone conditions to certain types of accidents in the petrochemical and construction industries.

There are also some publications on the fusion of neural networks and fuzzy logic. The latter corresponds to a mathematical theory that models the human ability of making decisions, allowing to deal with the concepts of uncertainty, imprecision and partial truth [64]. Ciarapica and Giacchetta [65] presented a neuro-fuzzy network that calculates the frequency and severity of occupational lesions, based on databases containing the history of occupational accidents. Martiniano et al. [64] developed a similar network to predict how many hours of health-related absenteeism a company would face per month.

Regarding the use of wearable sensors with ML, a variety of applications can be found in the literature. Bota et al. [66] developed a semi-automatic technique to annotate human activity recognition signals. Sánchez and Rodriguez [67] started developing a smart personal protective equipment (including helmet and belt) capable of responding in real-time to occupational risks, through the use of RF and neural networks.

Nath et al. [68] continued their previous work (mentioned in Section 3.2) and used a SVM classifier for heavy and repetitive activities, which can be used to assess ergonomic risk levels. Conforti et al. [69] applied ML to identify incorrect postures in material handling tasks, using inertial sensors. Manjarres et al. [70] developed a system composed of motion and heart-rate sensors, which uses RF to predict activities and assess workload in physically demanding jobs.

There are also some developed classifiers using neuro-fuzzy networks [71] and a fuzzy SVM [72] to help diagnose neuromuscular disorders, based on EMG signals.

However, to the best of our knowledge, there are no studies regarding inertial sensors, EMG, and different ML techniques to assess occupational risk in office workers.

4

Data acquisition

This chapter presents the different data sources used for this work. This includes the main data source to be analyzed, acquired in a real working environment, and the parallel data acquisitions, performed in a laboratory environment and used to develop algorithms to detect if a subject is seated or not.

4.1 Main acquisitions

As aforementioned, PrevOccupAI is a project that aims to prevent occupational diseases in the office context. For that purpose, its data acquisition involves different types of data, in order to take into account the psychosocial, biomechanical, and environmental risk factors for a Work-related Disorder (WRD). The biomechanical domain was assessed through wearable sensors and questionnaires, while the psychosocial and environmental domains were assessed with questionnaires.

These data acquisitions were performed on office workers from Autoridade Tributária e Aduaneira (AT), working in a real-world scenario, in their own workplace. Therefore, data was not collected in a controlled environment. The acquisition sessions took place in four different AT divisions and in four different weeks. There were a total of 40 participants, 10 for each AT division, and each participant was monitored while working for five consecutive days.

Before the five consecutive acquisition days, the participants had a meeting with the experts, where they were fully informed about the acquisition process, that all data would be handled anonymously, and that they could drop out at any moment. The participants also signed the informed consent and were asked to fill out a set of questionnaires that cover three domains of risk factors for WRDs, as well as their personal data. These questionnaires were all integrated in the PrevOccupAI application [8], which web version is shown in Figure 4.1. The questionnaires are listed in Table 4.1, grouped by domain.

Additionally, in each acquisition day, each participant filled out a daily pain questionnaire, twice. This questionnaire is the only one that was done daily and consists of a representation of the human body in the frontal plane (anterior and posterior), presented

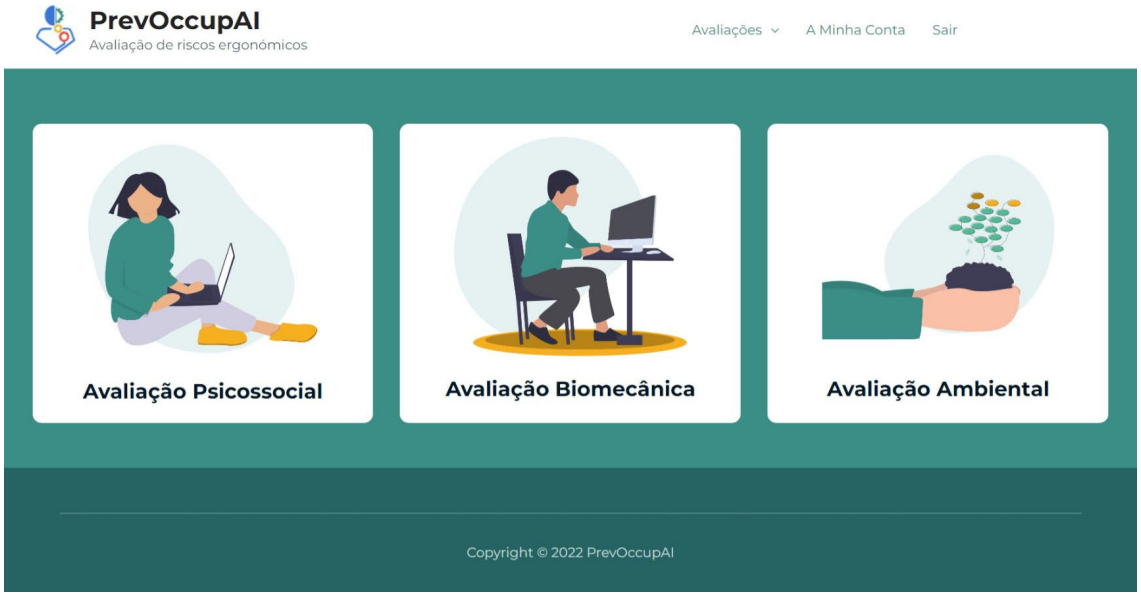

Figure 4.1: Screenshot of the PrevOccupAI homepage (web version). From [8].

in Figure 4.2. Using the color scale show in Figure 4.3, where *1* (green) corresponds to light pain and *10* (red) to the worst possible pain, the participants had to paint the body areas corresponding to their pain at that particular moment. The daily pain questionnaire was filled out at the beginning (morning) and end (afternoon) of each of the five acquisition days.

The postural assessment is part of the biomechanical evaluation and takes advantage of wearable sensors. In addition to the daily pain questionnaire, this was the assessment performed daily throughout the five days.

Table 4.1: List of questionnaires answered by the participants of the main study.

Domain	**Questionnaire**
Personal	Demographic data
	Lifestyle
	Physical activity
Psychosocial	Work demands
	Work organization and content
	Social relations and leadership
	General well-being
	Values in the workplace
Biomechanical	Office design
	Equipment
	Disability and suffering associated with pain
Environmental	Office environment
	Office organization
	Office privacy

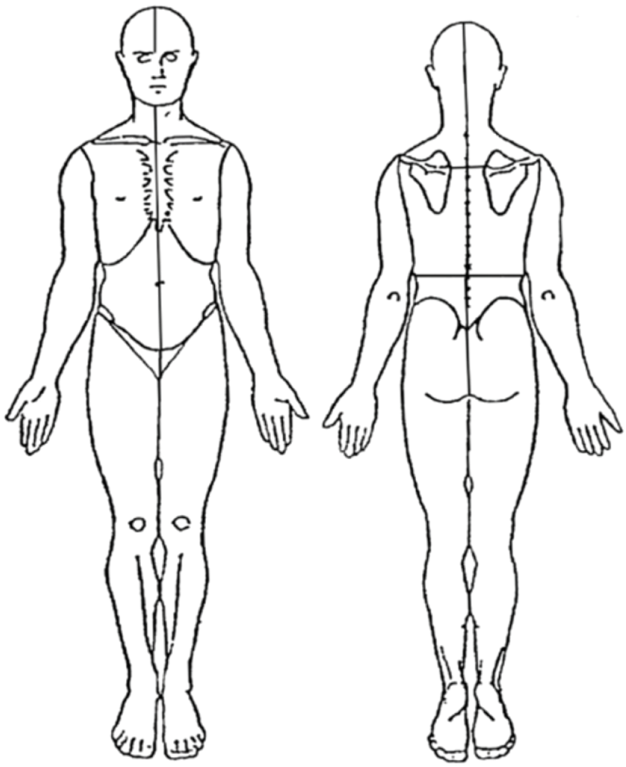

Figure 4.2: Human body representations in the frontal plane, used in the daily pain questionnaire. Image provided by courtesy of the PrevOccupAI project.

Figure 4.3: Color scale used in the daily pain questionnaire.

The participants were equipped with several devices. These include a smartphone, a smartwatch, and two muscleBANs, which are essentially specialized wireless Electromyography (EMG) sensors combined with inertial sensors. The smartphone was fixed to the participant's chest, using a special strap around the neck and torso (see Figure 4.4), and the smartwatch was placed on the wrist of the participant's dominant hand (see Figure 4.5). The muscleBANs were placed on the left and right upper trapezius muscles (see Figure 4.6), using disposable electrodes that were directly hooked up to the devices.

The smartphones used are Xiaomi Redmi Note 9 models (Xiaomi Inc., www.mi.com), which include a variety of sensors, such as accelerometer, gyroscope, magnetometer, rotation vector, and microphone. The microphone was used to develop a digital noise sensor. On the other hand, the smartwatches are OPPO 46 mm watches (Guangdong Oppo Mobile Telecommunications Corp., Ltd, www.oppo.com), and include the same sensors as the smartphones (except the microphone), as well as a heart rate sensor. The muscleBANs are produced by PLUX Biosignals (Plux - Wireless Biosignals, S.A, www.pluxbiosignals.com) and include a single-channel EMG sensor, accelerometer, and magnetometer.

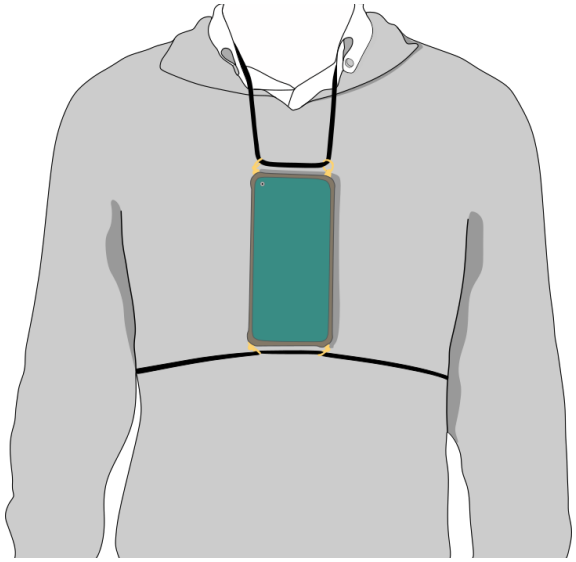

Figure 4.4: Smartphone placement for data acquisition. Image provided by courtesy of the PrevOccupAI project.

Figure 4.5: Smartwatch placement for data acquisition. Image provided by courtesy of the PrevOccupAI project.

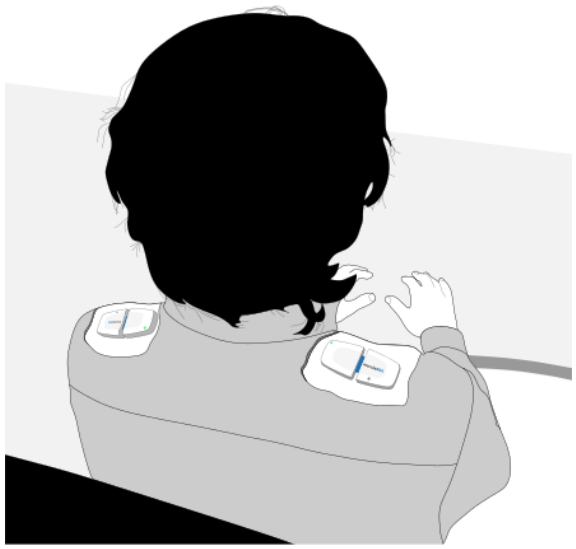

Figure 4.6: MuscleBAN placement for data acquisition. Image provided by courtesy of the PrevOccupAI project.

The sampling rate for acquisition was set to 100 Hz for the smartphone and the smartwatch, except for the noise and heart rate sensors, which were set to 1 Hz. However, the heart rate sensor only acquired for 1 minute, every 3 minutes, so that its battery could last the entire working day. On the other hand, the muscleBANs acquired at a sampling rate of 1000 Hz.

All the devices communicate with the PrevOccupAI cross-platform application [8] (designed for web and Android), using bluetooth technology. This application allows to schedule different acquisitions that can be performed with all devices simultaneously or only with some of them. In this case, for the first group (first week), four 20-minute acquisitions including all devices were set: two in the morning and two in the afternoon. For the remaining groups, the 20-minute acquisitions were set for the smartwatch and muscleBANs, while the smartphone sensors acquired for the whole working day. Thus, while participants were working, the acquisitions started and ended automatically at the scheduled times, being noticeable through smartphone vibrations. Each acquisition was stored in the smartphone's internal memory as *.txt* files (one for each sensor), which were automatically uploaded to Google Drive. The acquisition interface of the PrevOccupAI mobile application is presented in Figure 4.7.

On each of the five acquisition days, the acquisition process included the following steps:

1. The participant was again informed of all steps of the study, and that he could withdraw at any time.

2. The devices were all disinfected with alcohol.

3. The participant logged into his account in the PrevOccupAI mobile and web applications.

4. The participant used a computer to answer the daily pain questionnaire.

5. The acquisitions were scheduled in the PrevOccupAI mobile application, according to the subject's preferences ensuring that the subject would be executing work while being seated.

6. The smartphone was placed on the participant's chest, using a strap around the neck/shoulders and torso.

7. The participant's skin was properly cleaned with alcohol for electrode placement.

8. The two muscleBANs, with the respective electrodes, were placed on the left and right upper trapezius muscles.

9. The smartwatch was attached to the participant's wrist, on the side of his dominant hand.

10. The participant left the room and went to his workplace. The four scheduled acquisitions started and ended automatically throughout the day, with the smartphone indicating the start and end of the acquisition by executing a vibration pattern.
11. At the end of the working day, the participant returned.
12. The devices were all removed and the participant's skin was cleaned with alcohol.
13. The participant answered the daily pain questionnaire again.
14. The devices were again disinfected and recharged for the acquisitions of the following day.

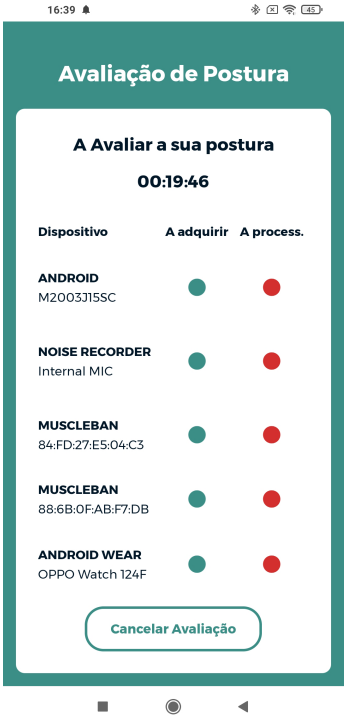

Figure 4.7: Screenshot of the PrevOccupAI acquisition interface (mobile version). From [8].

4.2 Complementary acquisitions

In order to analyze only the sitting data, we had to be sure when the participants were seated during the acquisitions, as they were carried out in an uncontrolled environment. For that, we had to acquire additional data in a controlled environment, to allow the possible development of an algorithm capable of assessing whether a person is sitting or not, based on smartphone signals.

Giving the fact that this data had to be collected under the same conditions as the main study, the smartphone was placed on the subject's chest with a strap, similarly to

what was described in Section 4.1 and illustrated in Figure 4.4. Since the smartphone data would be enough for sitting detection, the smartwatch and muscleBANs were not used within the scope of these acquisitions.

4.2.1 Sit-stand transitions detection

For the first approach to sitting detection, we performed acquisitions of stand-to-sit and sit-to-stand transitions on three different subjects, collecting the smartphone's accelerometer at 100 Hz. Transition movements were performed for approximately 3 seconds and in between transitions a pause of 3 seconds was executed. The acquisition included the following steps:

1. The participant was informed of all steps of the study.
2. The smartphone was disinfected with alcohol.
3. The smartphone was placed on the participant's chest, using a strap around the neck/shoulders and torso.
4. The participant was asked to sat down on a chair and to wait 3 seconds.
5. The participant was asked to stand up and to wait 3 seconds.
6. Steps 4 and 5 were repeated five times.
7. The smartphone was removed and again disinfected with alcohol.

4.2.2 Sitting and walking distinction

In this case, a single subject participated in data acquisition. The acquired smartphone sensors were the accelerometer, gyroscope, magnetometer, rotation vector, and noisemeter, with a sampling rate of 100 Hz, except for the noisemeter (approximately 1 Hz).

Data was collected when the participant was both sitting and walking, to perform sitting detection. The acquisition included the following steps:

1. The participant was informed of all steps of the study.
2. The smartphone was disinfected with alcohol.
3. The participant logged into his account in the PrevOccupAI mobile application.
4. Two 30-minute acquisitions were scheduled in the PrevOccupAI mobile application.
5. The smartphone was placed on the participant's chest, using a strap around the neck/shoulders and torso.
6. The participant sat down on a chair and remained seated for 30 minutes, while the smartphone was acquiring data.

7. The participant was asked to stand up and to walk continuously for 30 minutes, while the smartphone was acquiring data.

8. The smartphone was removed and again disinfected with alcohol.

5

Methods

This chapter presents and describes the methods used to meet the objectives of this thesis. This includes the computational tools, the data synchronization process, and the development of the sitting, pain, and occupational risk models.

5.1 Computational methods

All developed code within the scope of this thesis was performed with the Python programming language, in the PyCharm environment. Several Python libraries were used, the most important being NumPy [73], SciPy [74], pandas [75], scikit-learn [76], and Time Series Feature Extraction Library (TSFEL) [77].

5.2 Data synchronization

As mentioned in Section 4.1, the devices used for data collection include a smartphone, a smartwatch, and two muscleBANs. The ideal would be to perform synchronization of all four devices, which would allow its simultaneous analysis. This synchronization requires a very distinctive event, which can be a person jumping up and down for a couple of times, for example. In that case, the jumps would certainly be easily recognized in the accelerometer of each device and could serve as a synchronization event.

However, since the acquisitions were performed in a real-world scenario and were not continuous throughout the day, the participants were not asked to perform that distinctive movement. Therefore, the devices could not be accurately synchronized between them and had to be processed and analyzed separately.

In addition to the inter-device synchronization, there is also the intra-device synchronization of the sensors, particularly on the smartphone and smartwatch. These devices run the Android operating system, which is designed to prioritize battery saving. That can lead to a loss of priority of the acquisition, which results in the sensors not sampling at the same time and in a non-equidistant sampling procedure. Hence, the acquired sensors of a device become misaligned in time and it is necessary to synchronize them.

The first step is to define the starting and stopping points, and crop or pad the signals according to that. For this work, we chose the starting point as the initial timestamp of the last sensor that started acquiring, and the stopping point as the final timestamp of the first sensor that stopped acquiring. For example, assuming the smartphone sensors started and stopped acquiring in the following order:

Start: *Accelerometer - Gyroscope - Magnetometer - Noisemeter - Rotation vector*

Stop: *Magnetometer - Gyroscope - Accelerometer - Noisemeter - Rotation vector*

In this case, we would choose the first sample of the rotation vector as the starting point and the last sample of the magnetometer as the stopping point, as represented in Figure 5.1. Thus, the beginning of the accelerometer, gyroscope, magnetometer, and noisemeter signals has to be cropped, as well as the end of the accelerometer, gyroscope, noisemeter, and rotation vector signals. It is also important to mention that the differences between the starting and stopping times of the sensors are not large, usually not exceeding a couple of seconds.

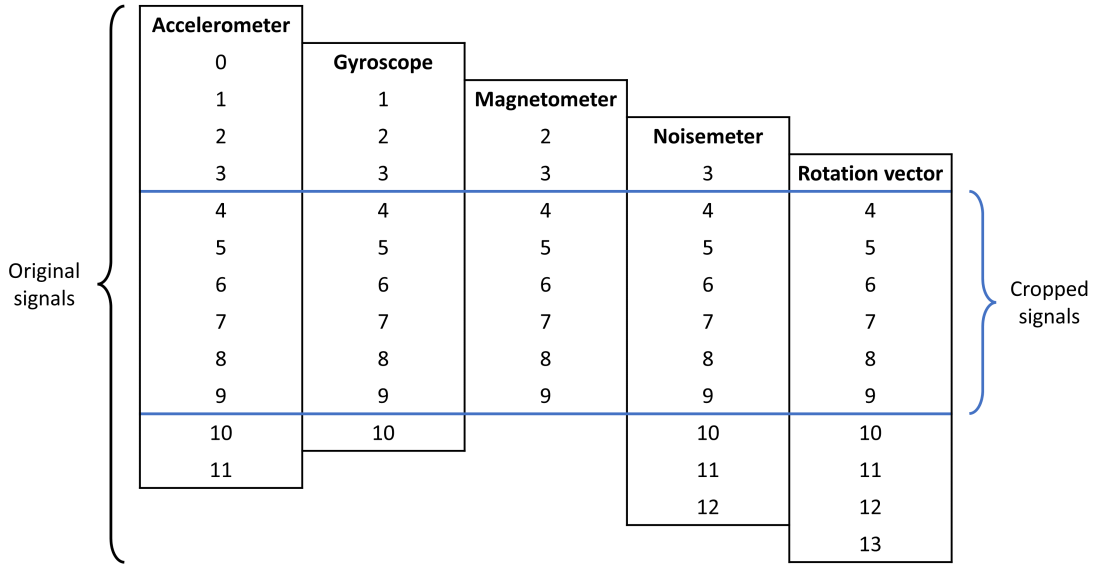

Figure 5.1: Example of signal cropping from the time axis of different smartphone sensors.

After cropping, each signal has to be individually resampled, to assure that the sampling frequency is constant and the same for all sensors. Using the SciPy [74] library, a new time axis with constant intervals was generated, beginning and ending at the defined starting and stopping points, respectively. Then, each component of each signal was individually interpolated, using the new time axis. This way, the sensors of each device were finally synchronized.

The process defined above was applied to the five sensors of the smartphone (accelerometer, gyroscope, magnetometer, noisemeter, and rotation vector). The new sampling rate was set to 100 Hz and a linear interpolation was performed, except for the noisemeter, where an interpolation of the type *previous* was preferred, which assigns the value of the

previous point to the new point. This type of interpolation is more appropriate for the noisemeter, because its original sampling rate of approximately 1 Hz is low and the noise does not exactly vary in a linear way.

For the smartwatch, the same process was applied to four of its five sensors (accelerometer, gyroscope, magnetometer, and rotation vector), with a new sampling rate of 100 Hz and linear interpolation. The heart rate signal was processed separately and it was not synchronized, because this sensor does not acquire in a continuous way (only for one minute, every three minutes). Each one-minute segment was resampled to a lower sampling rate of 1 Hz.

The muscleBAN devices are professional equipment for the acquisition of Electromyography (EMG) signals (as well as accelerometer and magnetometer) and therefore their sampling scheme is equidistant at 1000 Hz and its sensors are originally synchronized. So, it is not necessary to apply the intra-device synchronization process to these devices. However, the muscleBAN's EMG signal sometimes exhibits some erratic zeros. To eliminate this behaviour, a process of linear interpolation was applied to the signals of the EMG sensor.

5.3 Sitting detection

The objectives of this work comprise the analysis of sensor data of people working seated in an office setting. However, the acquisitions were not performed in a controlled setting, but in a real-world scenario, so it is possible that the participants did not remain seated for the entire acquisition period.

To ensure the validity of the analysis to be performed, it was important to develop an algorithm to detect the periods when a participant was not seated and remove them from the data to analyze. For this purpose, two different approaches were tested, namely the detection of the sit-to-stand and stand-to-sit transitions, and the classification of the signals between sitting and walking.

5.3.1 Sit-stand transitions detection

One of the possible ways to detect whether a person is sitting or not based on the acquired signals is to find the stand-to-sit and sit-to-stand transitions using distance metrics between the pattern (in this case the transition between sitting and standing) and the recorded signal. In this approach, the pattern is slid over the signal to be analyzed and the distance measure is calculated for each sample within the overlapping section of the pattern and the signal. From the resulting values a mean is calculated. The pattern is slit over the signal until it reaches the end of the signal and thus the resulting output time-series has a length corresponding to the number of windows. The implementation of the detection was realized using the x-, y-, and z-axis of the smartphone's accelerometer sensor.

To obtain the transition patterns, smartphone data of three people performing the transition movements (stand-to-sit and sit-to-stand) was recorded, as described in Section 4.2.1. The participants were asked to do 5 transitions in either direction, giving a total of 30 patterns (10 for each subject). The obtained patterns were pre-processed by smoothing them using a Hanning window with a sample size of 10 and applying a Z-score normalization. This type of normalization is the process of subtracting the pattern's mean and dividing it by its standard deviation [78]. This way, each normalized pattern has zero mean and unit standard deviation, making possible to compare it to other signals.

Signals are pre-processed in the same manner as the patterns. Additionally, the signals are segmented into windows of three seconds, to match the length of the patterns. The hop (sliding factor) between windows was set to 1 sample. Thus, the number of windows (n) is given by

$$n = \frac{k - w}{h} + 1, \quad h \leq w \tag{5.1}$$

where k is the number of samples in the signal, w is the window length (in samples), and h is the hop between windows.

Finally, the comparison between each signal window and the pattern is performed by calculating the euclidean distance (l) between each point of the signal window and the corresponding point of the pattern, given by

$$l = \sqrt{(p - s)^2} \tag{5.2}$$

where p and s are the values of the pattern and the signal window, respectively.

Then, the mean of those distances is calculated and is inserted as a point of a new time-series. The pattern is slid across the signal and the process is repeated until the sliding window reaches the end of the signal to analyze.

The newly generated time-series, where each of its points represents the mean of the distances between each point of the signal window and the corresponding point of the pattern, can then be used to detect when stand-to-sit or sit-to-stand transitions occurred, depending on the pattern used (sitting or standing, respectively). These transitions can be identified by the possible prominent local minima present in this time-series. Ideally, a transition would occur where the mean distance is close to zero.

We tested a range of different variations of this algorithm, to evaluate its efficiency. In a first step, z-axis patterns of both transitions from three different subjects were used for the analysis. From each subject one sit-to-stand and stand-to-sit transitions were extracted and these patterns were then slid over the signal to be analyzed. This means that for each transition type, three patterns were used to calculated distance metrics. The mean of the resulting output time-series was subsequently calculated, respectively for either of the two transition types, and this was then considered to be the final output.

In another approach, the three axes of the accelerometer were utilized. This time, only one pattern was used and each axis of the signal window was compared to the

corresponding axis of the pattern. The final distance was considered as the mean value of the distances of the three axes.

Finally, we tried to combine the three axes of one pattern of each of the three acquired subjects. For that, nine different Euclidean distances were calculated, one for each axis (x, y, and z) and for each of the three patterns. Then, the value inserted in the new time-series was the mean of those nine distances.

The signal that was used to test these approaches consisted of a recording containing various sit-to-stand and stand-to-sit transitions at intervals of approximately 3 seconds.

5.3.2 Sitting and walking distinction

Another approach to detect if a person is not seated is to build a Machine Learning (ML) classifier, to automatically distinguish between different classes.

For this purpose, it was decided to try to classify the signals between sitting and walking. A standing class was not considered necessary, because it was assumed that the participants would not be just standing in a static way, but somehow moving. To achieve this goal, smartphone data (accelerometer and gyroscope sensors) was acquired as described in Section 4.2.2.

5.3.2.1 Data pre-processing

First of all, the *.txt* files corresponding to the accelerometer and gyroscope signals of each class (sitting and walking) were manually verified, to ensure there was no corrupt or missing data. Then, the files were loaded, with the first 2 seconds of each signal being ignored, which could contain some undesired movements related to the start of each acquisition.

Next, a process of synchronization similar to the one in Section 5.2 was applied to the accelerometer and gyroscope data. Additionally, a smoothing filter with a Hanning window of 30 samples was applied to the signals, which essentially consists of a low-pass filter without delay.

After that, the signals were divided into windows, to allow the extraction of features. Two different types of windows were considered, to compare its final results: 3 and 5 seconds, both without overlapping.

5.3.2.2 Dataset characterization

As stated in Section 4.2.2, for sitting/walking distinction, data was collected from only one subject. The participant was a healthy 22-year-old male, with a height of 175 cm and a weight of 80 kg.

Signals were segmented into windows of either 3 or 5 seconds, resulting in sample sizes of 1250 and 748, respectively, which correspond to approximately 30 minutes of acquisition for each of the two classes. Table 5.1 presents the distribution of both classes, sitting and walking, among the generated dataset.

Table 5.1: Number of samples per class for sitting/walking classification.

Window size (s)	Label	Class	Number of samples
3	0	Sitting	619
	1	Walking	631
5	0	Sitting	370
	1	Walking	378

5.3.2.3 Classification models

It was decided to use three different ML algorithms that would potentially provide good results for sitting/walking classification, namely Random Forest (RF), Support Vector Machine (SVM), and logistic regression, which were described in Sections 2.3.1, 2.3.2, and 2.3.3, respectively.

Furthermore, two different versions of each model were developed, which were fed with different input data. Firstly, the three different models were trained using features from both smartphone's accelerometer and gyroscope signals. Then, in a second version, the three models were trained using only accelerometer features, to verify what influence the removal of the gyroscope features would have on the classification. Additionally, for each of the two model versions, 3- and 5-second windows were tested for feature extraction, to verify its impact on the classification.

5.3.2.4 Feature extraction and selection

To perform the extraction of features from the signals, the TSFEL [77] library was used, which allows for an easy extraction of features in three different domains: statistical, temporal, and spectral. For this classifier, it was decided to only extract features from the statistical and temporal domains. The extraction of features from the spectral domain is more computationally expensive and it is not essential for this type of classification. The extracted features are listed in Tables 5.2 and 5.3, for the statistical and temporal domains, respectively. There is a total of 12 features from the statistical domain and 18 from the temporal domain. Taking into account that all the aforementioned features were extracted from the three axes of both accelerometer and gyroscope signals, we have a total of 180 features for each sample.

Furthermore, three methods of feature selection were applied to the set of features, to understand which features were the most informative. Firstly, using the TSFEL [77] library, the highly correlated features were removed. Then, the features that were constant across all samples (zero variance) were also removed. Finally, the features that would provide a more accurate classification were chosen, using Recursive Feature Elimination (RFE) with stratified k-fold ($k = 10$) as the cross-validation method. This last method was performed differently for each algorithm to test, using models with its default parameters from the scikit-learn [76] library. The different models used for RFE are characterized in Table 5.4.

Table 5.2: Extracted features of statistical domain. Adapted from [79].

Feature	**Description**
Interquartile range	Computes interquartile range of the signal.
Kurtosis	Computes kurtosis of the signal.
Maximum	Computes the maximum value of the signal.
Mean	Computes mean value of the signal.
Mean absolute deviation	Computes mean absolute deviation of the signal.
Median	Computes median of the signal.
Median absolute deviation	Computes median absolute deviation of the signal.
Minimum	Computes the minimum value of the signal.
Root mean square	Computes root mean square of the signal.
Skewness	Computes skewness of the signal.
Standard deviation	Computes standard deviation of the signal.
Variance	Computes variance of the signal.

Table 5.3: Extracted features of temporal domain. Adapted from [79].

Feature	**Description**
Absolute energy	Computes the absolute energy of the signal.
Area under the curve	Computes the area under the curve of the signal computed with trapezoid rule.
Autocorrelation	Computes autocorrelation of the signal.
Centroid	Computes the centroid along the time axis.
Entropy	Computes the entropy of the signal using the Shannon Entropy.
Mean absolute differences	Computes mean absolute differences of the signal.
Mean difference	Computes mean of differences of the signal.
Median absolute differences	Computes median absolute differences of the signal.
Median difference	Computes median of differences of the signal.
Negative turning points	Computes number of negative turning points of the signal.
Neighbourhood peaks	Computes the number of peaks from a defined neighbourhood of the signal.
Peak to peak distance	Computes the peak to peak distance.
Positive turning points	Computes number of positive turning points of the signal.
Signal distance	Computes signal traveled distance.
Slope	Computes the slope of the signal by fitting a linear equation to the observed data.
Sum absolute difference	Computes sum of absolute differences of the signal.
Total energy	Computes the total energy of the signal.
Zero-crossing rate	Computes zero-crossing rate of the signal.

Table 5.4: Characterization of the classification algorithms used for RFE.

Algorithm	Parameters
Logistic regression	C: 1.0 max_iter: 1000 penalty: l2 solver: lbfgs
RF	criterion: gini n_estimators: 100
SVM	C: 1.0 kernel: linear

5.3.2.5 Model optimization

Before using a ML algorithm, it is necessary to decide which data will be used to train and test the model. There are several approaches but, in our case, we decided to split the dataset into 70% for training and 30% for testing. This separation was performed randomly but in a stratified way, to ensure the two subsets had the same class distribution as the complete dataset, even though there was no significant class imbalance, as can be seen in Table 5.1.

Furthermore, the vast majority of the ML algorithms have its own hyperparameters that have to be manually selected to obtain the maximum performance when the model is fitted on the training data. There are some different ways to automatically achieve this optimal selection of hyperparameters, to obtain the best results possible without losing too much time. One of the most popular methods is grid search, which is essentially an exhaustive search over a manually selected subset (grid) of all possible hyperparameters. This algorithm tries all possible combinations of the defined hyperparameter grid, fitting each model on the training data and obtaining some performance metric, typically the accuracy. Additionally, the performance evaluation can be done with cross-validation techniques, to avoid overfitting [80]. For sitting/walking classification, grid search was performed with the scikit-learn library [76] on the RF and SVM algorithms, using stratified k-fold ($k = 10$) as the cross-validation method. On the other hand, logistic regression does not allow grid search, because its scikit-learn [76] implementation does not have definable hyperparameters. The possible values searched for each hyperparameter are presented in Table 5.5.

In order to evaluate the performance of the models, each of them was trained 5 times, using different seeds to perform the splitting into training and testing sets. The seeds used were the following: 52374, 52347, 52734, 52743, and 52473. In each of the iterations, the accuracy was calculated by Equation 2.5. Thus, each model is associated to 5 accuracies, which allows a statistical analysis.

Table 5.5: Values used for each hyperparameter when performing grid search.

Algorithm	Hyperparameter	Values
RF	n_estimators	50, 100, 250, 500
	max_depth	5, 6, 7, 8, 9, 10
	criterion	gini, entropy
SVM	C	0.1, 1, 10, 100, 1000
	gamma	1, 0.1, 0.01, 0.001, 0.0001
	degree	2, 3, 4, 5
	kernel	linear, poly, rbf, sigmoid

5.4 Pain prediction

Another goal of this thesis was to predict the pain that is reported by the participants in the morning and in the afternoon, as stated in Section 4.1. Therefore, different ML algorithms were tested in order to build a classifier capable of assessing whether a person has pain or not, based on their EMG signals and possibly questionnaires' answers. If successful, this classifier can reduce the need for the participants to fill out the daily pain questionnaire, which is done twice a day.

5.4.1 Electromyography pre-processing

First of all, in order to develop a classifier based on EMG signals, it is necessary to apply some pre-processing steps to these signals.

As we only wanted to analyze the periods in which people were seated, we applied the accelerometer-only version of the sitting/walking classifier (see Section 5.3.2) to the muscleBAN signals, because these devices do not have a gyroscope. To perform this, the muscleBAN signals were first loaded and they were only considered for analysis if they were more than 50% of the original duration of the acquisitions (more than 10 minutes, to be precise). Additionally, if one of the muscleBAN signals (left or right) was missing from a certain acquisition, that acquisition was also not considered.

Then, both muscleBANs were synchronized, using a log file that was available for 30 of the 40 participants. This is due to the fact that this feature was developed after acquisitions with the first 10 subjects. This log file contains the absolute time each device and sensor started to acquire, making a fairly accurate synchronization possible. Figure 5.2 shows an example of an acquisition log file. In the illustrated example, the two muscleBANs, which are identified by its Media Access Control (MAC) addresses (*C4:14:A2:04:3D:EE* and *EA:31:04:5F:58:DC*), started acquiring at 10:20:07.721 and 10:20:11.124. This means that the second muscleBAN started to acquire 3.403 seconds later, and this value can be used to align the time axis of both muscleBANs. There is no need for interpolation, because the sampling rate is fixed at 1000 Hz. For the acquisitions not containing the log file, the times present in each filename (e.g., *opensignals_C414A2043DEE_2022-04-01_10-20-07.txt*) were considered to synchronize the muscleBANs.

```
# LOG FILE HEADER
# {"smartphone":LIBPhys #001,"Date_Time":2022-04-01_10-20-00,"column":["LOG_TIME","LOG"]}
# EndOfHeader
10:20:04.770    WEAR_COMMUNICATION: first data from smartwatch received
10:20:04.779    SENSOR_DATA: received first data from WEAR_ACCELEROMETER
10:20:05.195    SENSOR_DATA: received first data from WEAR_GYROSCOPE
10:20:05.313    SENSOR_DATA: received first data from WEAR_ROTATION_VECTOR
10:20:05.451    SENSOR_DATA: received first data from WEAR_MAGNETOMETER
10:20:07.452    OS SERVER: OS device state changed for: C4:14:A2:04:3D:EE state: CONNECTED
10:20:07.632    OS SERVER: OS device state changed for: C4:14:A2:04:3D:EE state: ACQUISITION_TRYING
10:20:07.704    OS SERVER: OS device state changed for: C4:14:A2:04:3D:EE state: ACQUISITION_OK
10:20:07.721    SENSOR_DATA: received first data from C4:14:A2:04:3D:EE
10:20:10.923    OS SERVER: OS device state changed for: EA:31:04:5F:58:DC state: CONNECTED
10:20:11.088    OS SERVER: OS device state changed for: EA:31:04:5F:58:DC state: ACQUISITION_TRYING
10:20:11.114    OS SERVER: OS device state changed for: EA:31:04:5F:58:DC state: ACQUISITION_OK
10:20:11.124    SENSOR_DATA: received first data from EA:31:04:5F:58:DC
10:20:51.303    SENSOR_DATA: received first data from WEAR_HEART_RATE
```

Figure 5.2: Example of an acquisition log file.

After synchronization, it was necessary to temporarily perform downsampling of the muscleBAN data from 1000 to 100 Hz. This is due to the fact that the sitting/walking classifier was trained with smartphone data, which is collected at 100 Hz. Moreover, the smartphone has a different coordinate system than muscleBANs, so the axes of the muscleBANs had to be swapped, in order to extract features for sitting/walking classification: $x_{phone} \leftrightarrow y_{mBAN}$; $y_{phone} \leftrightarrow z_{mBAN}$; $z_{phone} \leftrightarrow x_{mBAN}$. Then, the features for sitting detection could be extracted and the classification was performed. Next, the windows classified as *walking* were removed from the signals and the downsampling was reverted. Additionally, each muscleBAN was identified as left or right, using its MAC addresses, which were manually labeled and stored in a *.txt* file.

At this point, we had the original muscleBAN signals, but without the walking periods. Next, the EMG signals had to be filtered, to allow their analysis. Firstly, a butterworth band-pass filter was applied with cutoff frequencies of 10 and 499 Hz. This is the frequency band that contains the essential information in the EMG signals [81]. Then, the signal was rectified, because we were only interested in its magnitude. In order to smooth the signal, a moving average filter with a window size of 100 samples was also applied. Finally, each EMG signal was normalized, to allow comparison with the signals of other people. The ideal would be to normalize it with each person's Maximum Voluntary Contraction (MVC) value, but that value was not available, because the acquisitions were performed in an uncontrolled environment, as already mentioned. Instead, the maximum EMG value for each person and for each day was used to normalize each 20-minute acquisition. An example of an EMG signal before and after pre-processing is presented in Figure 5.3.

5.4.2 Resting and moving distinction

Before extracting the EMG features for pain prediction, we considered important to divide the EMG signals between resting and moving, because these can potentially contribute differently to detect the pain each person has. The resting windows correspond to the periods in which a person is almost static, while the moving windows correspond to the person moving their torso and/or shoulders while seated. Thus, a computational inexpensive algorithm was developed to achieve this distinction.

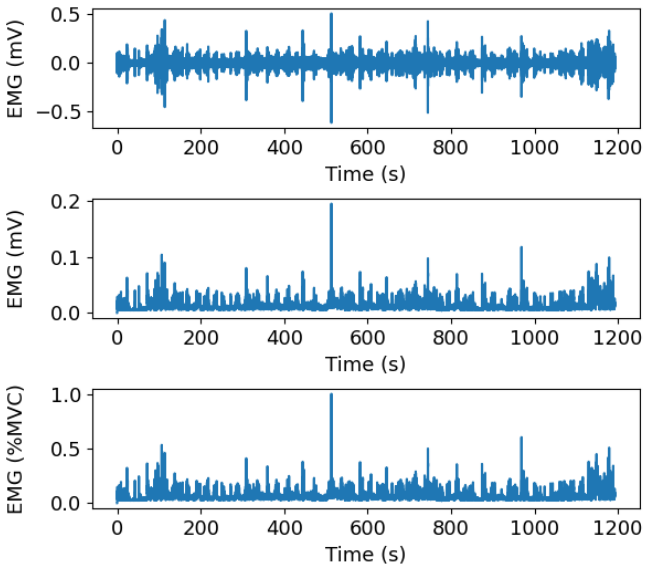

Figure 5.3: Example of an EMG signal before processing (top), after filtering (middle), and after normalization (bottom).

To detect movement in the muscleBAN signals, we had to use its accelerometer. So, the signals of the accelerometer's three axes were first filtered with a smooth filter, using a Hanning window of 30 samples. Then, the acceleration magnitude (a) was calculated for each sample by the expression

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{5.3}$$

where a_x, a_y, and a_z are the values of the acceleration for the x-, y-, and z-axis, respectively.

Then, the mean value of the acceleration magnitude was removed and the signal was rectified, because we were only interested in the absolute value of the magnitude. Finally, the rectified signal was processed in windows. For each 1-second window (corresponding to 100 samples), if it surpasses a certain threshold, it is classified as *moving*. Otherwise, it is classified as *resting*. This threshold was experimentally set to 0.5 m/s^2, as exemplified in Figure 5.4 by the red horizontal line.

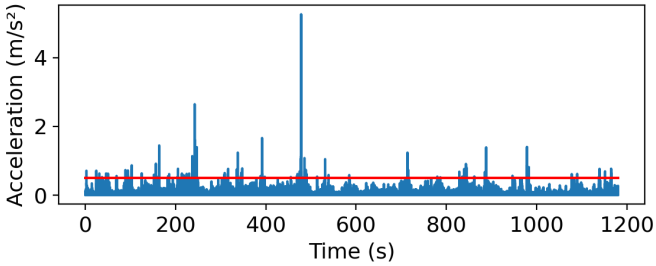

Figure 5.4: Example of an acceleration magnitude signal, where the red line marks the threshold for moving classification, defined as 0.5 m/s^2.

5.4.3 Dataset characterization

For the pain prediction classifier, we analyzed the main source of acquired data (see Section 4.1). Data was collected from 40 different subjects, who have their demographic information displayed in Table 5.6. Furthermore, 29 out of the 40 individuals were female and 11 were male. This information was obtained through the questionnaires answered by the participants.

Table 5.6: Statistical analysis of the demographic data from the participants of the main study.

	Minimum	**Maximum**	**Mean**	**Standard deviation**
Age (years)	41	63	51	6
Height (cm)	150	184	166	8
Weight (kg)	47	108	70	14

The participants often performed four acquisitions a day. However, for pain prediction, we only considered the first acquisition in the morning and the last in the afternoon. Thus, considering each 20-minute acquisition as a sample, we would have an ideal number of samples of 400, which results from 2 acquisitions per day, for 5 days, for 40 participants. However, there was a considerable data loss, especially in the first week of acquisitions, because some devices did not work as expected. Additionally, there was a couple of times in which the participants did not fill out the daily pain questionnaire. Hence, we had a total of 286 samples, which distribution by class and time of day is presented in Table 5.7.

Table 5.7: Number of samples per class by period of day for pain classification without personal data.

Label	**Class**	**Time**	**Number of samples**	**Total samples**
0	No pain	Morning	93	152
		Afternoon	59	
1	Pain	Morning	68	134
		Afternoon	66	

If the personal information from the questionnaires is also added to the dataset, the number of samples is further reduced, because one of the participants did not respond to all questionnaires. So, considering the personal information, the new dataset had a total of 274 samples, distributed by class and time of day as presented in Table 5.8.

Table 5.8: Number of samples per class by period of day for pain classification with personal data.

Label	**Class**	**Time**	**Number of samples**	**Total samples**
0	No pain	Morning	89	148
		Afternoon	59	
1	Pain	Morning	64	126
		Afternoon	62	

5.4.4 Classification models

For pain classification, it was decided to use the same ML algorithms as for the sitting/walking classification (see Section 5.3.2.3), namely logistic regression, RF, and SVM.

In a first approach, only the features extracted from the EMG signals were used as input to the three different models. Then, in a second approach, the personal information extracted from the questionnaires was also used as features to feed the models, which would potentially allow to improve the classification results.

5.4.5 Label generation

To allow the training of a ML model to predict pain, it is imperative to assign a class to each sample, the so-called process of labelling.

The pain information is contained in the daily pain questionnaire that people fill out twice a day, in the morning and in the afternoon. This questionnaire consists of a picture that is painted by the participants, indicating where they have pain. This information must be converted to a numeric value, to serve as label for the ML models.

Two classes were then defined: *0*, corresponding to not having pain, and *1*, corresponding to having pain. Since the muscleBANs were placed on the upper trapezius muscle and we were trying to predict pain using the EMG sensor, only the pain that could potentially be manifested in that specific EMG signal was considered. Thus, each painting was manually codified to a numeric value (*0* or *1*), taking into account that only the blue regions in Figure 5.5 were considered. These regions include neck, shoulders, and upper back. So, if a participant reported shoulder pain, that was codified as *1*, but if a participant reported leg pain, that was codified as *0*.

Given that acquisitions were done in the morning and in the afternoon, and that the pain questionnaires were also filled out during the morning and the afternoon, the generated labels were matched according to their respective times. For this classification, we only used the first acquisition in the morning and the last in the afternoon, as aforementioned, because they were considered to be the most informative to predict the pain reported by the participants. Thus, the first acquisition in the morning was matched with the label generated from the morning pain questionnaire and the last acquisition in the afternoon was matched with the label generated from the afternoon pain questionnaire.

5.4.6 Feature extraction and selection

Once again, the TSFEL [77] library was used to extract features from the EMG signals. In addition to the statistical and temporal features, listed in Tables 5.2 and 5.3, some spectral features were also extracted, namely the ones listed in Table 5.9.

For this classifier, the signal was not split into windows, because we only had one pain label for each 20-minute acquisition. It is also very important to mention that each individual EMG feature was extracted 6 times for each sample: two for the left and right

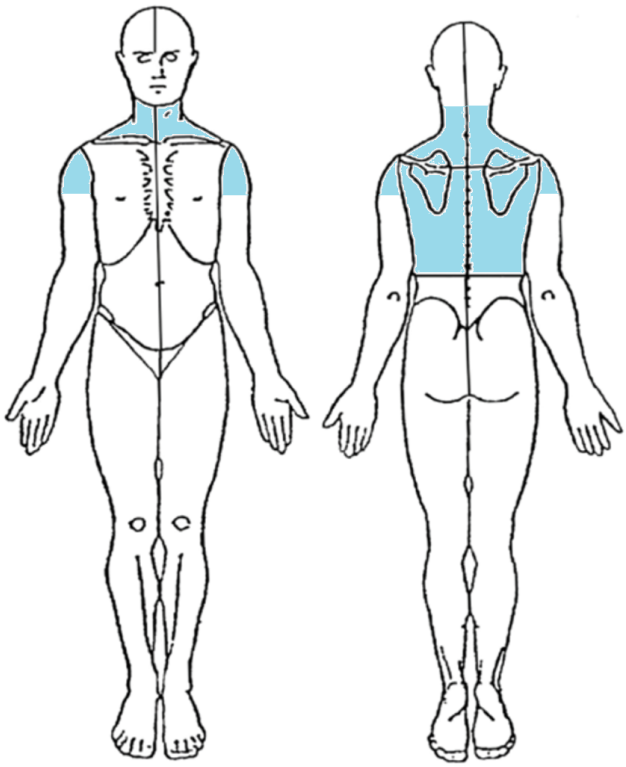

Figure 5.5: Considered body parts (painted in blue) for pain prediction.

Table 5.9: Extracted features of spectral domain. Adapted from [79].

Feature	**Description**
Fundamental frequency	Computes fundamental frequency of the signal.
Maximum frequency	Computes maximum frequency of the signal.
Median frequency	Computes median frequency of the signal.
Power bandwidth	Computes power spectrum density bandwidth of the signal.

entire 20-minute EMG signals, two for the left and right resting portions of each signal, and two for the left and right moving portions of each signal.

Beyond the EMG features, some relevant questionnaires' answers were also used as features and considered to build the pain classifier. These data include:

- Age;
- Gender;
- Height;
- Weight;
- Body Mass Index (BMI);
- International Physical Activity Questionnaire (IPAQ) score;

- Number of working hours per week;
- Number of hours sitting per working days;
- Number of hours sitting per weekend.

It is worth noting that the BMI was calculated using the following expression

$$BMI = \frac{w}{h^2} \tag{5.4}$$

where w is the weight in kg and h is the height in m.

The categorical variables, namely gender and IPAQ score, had to be pre-processed before they could be used as features in ML models. This pre-processing is called feature encoding and it is essential to transform categorical data into numerical data.

The gender, as it is a nominal variable, without a defined order, had to be encoded using two columns, one for male and one for female. A male participant received a *1* in the male column and a *0* in the female column, while a female participant received a *0* in the male column and a *1* in the female column.

The IPAQ score, on the other hand, is an ordinal variable, because its possible values have a defined order (*low* < *moderate* < *high*). Therefore, only one column was needed to encode it, with three possible values. So, *low* was defined as *0*, *moderate* as *1*, and *high* as *2*.

In the end, the feature vector was composed of 213 features for each sample (20-minute acquisition): 102 from each muscleBAN (12 statistical, 18 temporal, and 4 spectral for the entire signal, resting portions, and moving portions) and 9 from the questionnaires.

Furthermore, the three methods of feature selection already described in Section 5.3.2.4 were applied to the set of features. These comprise the removal of the highly correlated and constant features, and the RFE with stratified k-fold ($k = 10$) as the cross-validation method. The different models used for RFE are the same, characterized in Table 5.4.

5.4.7 Model optimization

In order to build the ML models, the dataset was divided into 70% for training and 30% for testing, in a stratified way, to maintain the distribution between the two classes.

The grid search process used to find the optimal hyperparameters is the same as explained in Section 5.3.2.5. The selected grid of hyperparameters is also the same and is shown in Table 5.5.

To evaluate the performance of the models, they were trained 5 times using different seeds (52374, 52347, 52734, 52743, and 52473) to split the dataset, giving rise to an accuracy for each of the iterations, as already described in Section 5.3.2.5.

5.5 Occupational risk prediction

The main objective of this work was to model the occupational risk for our sample of approximately 40 individuals and, then, try to predict that risk using different types of

features. As the risk was modelled using postural variables, its prediction was performed using variables non-related to posture. It is indispensable to note that this occupational risk refers to a relative risk, which represents a deviation from the distribution of specific variables from this 40-subject population.

5.5.1 Risk modelling using postural variables

In order to develop a model capable of assessing occupational risk, we first had to define a measure for risk, which would serve as label for the ML model. This risk model is a model that considers risk as a function of postural variability of a person. Postural variability corresponds to the adjustments each person makes to their sitting posture. In the risk model it is considered that with higher postural variability the ergonomic risk decreases.

To calculate the ergonomic occupational risk, we started by defining a set of variables that we considered representative of postural variability. It is important to keep in mind that posture is defined as the position where we keep our body when we are seated. The considered variables include:

- Number of changes in posture;
- Number of different postures;
- Mean time of transition between postures;
- Time spent in each of the subject's three most common postures;
- Time spent in the remaining postures;
- Total distance covered;
- Distance covered in each of the subject's three most common postures;
- Distance covered in the remaining postures;
- Variance in each of the subject's three most common postures;
- Mean variance in the remaining postures;
- Mean velocity;
- Mean velocity in each of the subject's three most common postures;
- Mean velocity in the remaining postures.

To extract these variables from the available data, we used the smartphone's rotation vector, which allows the calculation of the subject's trunk position at each moment. The rotation vector sensor merges accelerometer, gyroscope, and magnetometer data, and it is based on the mathematical concept of quaternions, which is the description of Three-dimensional (3D) orientation using a Four-dimensional (4D) complex number system

[82]. Thus, the smartphone's rotation vector returns four values that describe the phone's orientation relative to the phone's base coordinate system. The smartphone's reference axes are illustrated in Figure 5.6.

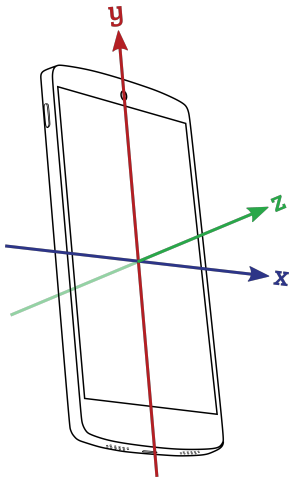

Figure 5.6: Coordinate system of the smartphone. From [83].

Firstly, for the subjects who performed smartphone whole-day acquisitions (30 out of 40), these were trimmed to the 20-minute periods in which the other devices also acquired. Then, we applied the sitting/walking classifier (see Section 5.3.2) to the 20-minute smartphone synchronized data, to remove the periods in which each person was not seated. Next, the quaternions provided by the rotation vector were converted to Euler angles, using the SciPy [74] library. After that, the median of each Euler angle was removed and that was considered as the reference point. The Euler angles were finally transformed into positions in the xz-plane (according to Figure 5.6), which corresponds to the horizontal plane when a person has the phone placed on his chest (according to Figure 4.4). Through the x and z coordinates, it is possible to determine the inclination of the trunk in that plane, at a given moment, which defines the different postures of each person. To allow the comparison of postures between subjects, these coordinates were normalized by the height of each individual.

Furthermore, it was also necessary to categorize the different possible postures into finite ranges. Taking into account that the obtained x and z coordinates ranged from -1 to 1, the xz-plane was divided into a grid of equal squares, as represented in Figure 5.7. The grid was equally distributed in both directions, consisting of 49 different possible postures. However, it is important to keep in mind that some of those 49 intervals are humanly impossible to reach.

Regarding the extracted variables, it is important to note that some additional pre-processing was applied to extract some of them, namely the variables not involving variability within the same posture (number of changes in posture, number of different postures, mean time of transition between postures, time spent in each of the subject's three most common postures, time spent in the remaining postures). To extract these

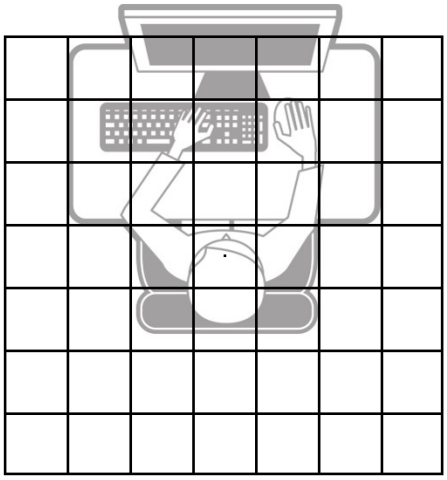

Figure 5.7: Division of the xz-plane to define the possible postures.

variables, the postural sway behaviour was removed. This behaviour corresponds to the small and unconscious movements around the body's center of mass needed to maintain balance while standing or sitting [84]. This postural sway, if not removed, can cause oscillations between the limits of two of the defined postures, affecting the extracted variables. These small oscillations are unconscious and do not involve changes in posture, as they are only adjustments that each individual makes to their posture. For this reason, to extract some of the variables, the higher frequencies were removed from the signals by applying a low-pass filter with a cut-off frequency of 0.3 Hz [85] to the Euler angles.

5.5.2 Risk classification and regression models

For the assessment of the relative occupational risk, we tried both classification and regression ML algorithms. For classification, we considered the same algorithms as before, namely logistic regression, RF, and SVM. For regression, we tested the linear regression, RF regression, and support vector regression algorithms.

For the classification approach, the ideal would be to have at least three classes, corresponding to low, medium, and high occupational risks. However, there was a significant class imbalance in the dataset, which would lead to few samples for the *medium* class and no samples for the *high* class. For this reason, we only considered two classes for classification, namely *0*, which corresponds to *no risk*, and *1*, which corresponds to *risk*.

For the regression approach, the ideal would be to do regression between 0 and 12, which corresponds to the maximum possible risk for our population, according to the number of postural variables considered. However, there were no samples with a risk higher than 7 in the dataset and, for that reason, the regression was performed between 0 and 7. Since we consider risk to be an integer, the outputs of the regression models were rounded to the nearest integer.

5.5.3 Dataset characterization

For defining and predicting the occupational risk, we used the same data as the pain prediction model (see Section 5.4.3), corresponding to the 40 participants of the main acquisitions (29 female and 11 male). Their demographic information is presented in Table 5.6.

However, for these risk models, we considered all four daily acquisitions that every participant performed. Thus, considering each 20-minute acquisition as a sample, we would have an ideal number of samples of 800, which results from 4 acquisitions per day, for 5 days, for 40 participants. As previously mentioned, there were acquisitions where one or more devices did not acquire, which resulted in a considerable data loss. Hence, we had a total of 430 samples, which distribution by class (risk score) is presented in Table 5.10. This distribution was used to train the regression models. On the other hand, for the classification models, a new categorization had to be made, as shown in Table 5.11.

Table 5.10: Number of samples per class for the occupational risk regressor.

Risk score	**Number of samples**
0	234
1	96
2	35
3	29
4	19
5	10
6	5
7	2

Table 5.11: Number of samples per class for the occupational risk classifier.

Label	**Class**	**Number of samples**
0	No risk	234
1	Risk	196

5.5.4 Label generation

Having extracted the variables representative of postural variability for each 20-minute acquisition, the next step was to define, for each variable, which values correspond to the existence or not of occupational risk, bearing in mind that this is a risk relative to the 40-people population.

Firstly, each variable was individually normalized, so that the variables could be compared with each other by their standard deviation. Then, only the variables with a standard deviation greater than 0.10 were considered for risk calculation. This is due to the fact that variables with low standard deviation do not allow an effective distinction between the relative risks of each acquisition.

Moreover, the individual values of each variable were visualized through histograms. Then, a Weibull distribution curve was fitted to each histogram, in order to obtain a parameterizable curve representative of each variable. The scikit-learn [76] implementation of the Weibull distribution returns three parameters, namely *c*, *loc*, and *scale*. These parameters were then used to plot the Cumulative Distribution Function (CDF) corresponding to each distribution. The initial and final 5% of the CDF were considered as risk zones, as marked in red in Figure 5.8, where the first 5% are delimited by a. However, when a was very small (below 0.02), we did not consider the initial 5%, but only the last 10% of the CDF, as shown in Figure 5.9. As mentioned before, these relative risk zones represent the areas that are further away from the areas with higher probability density.

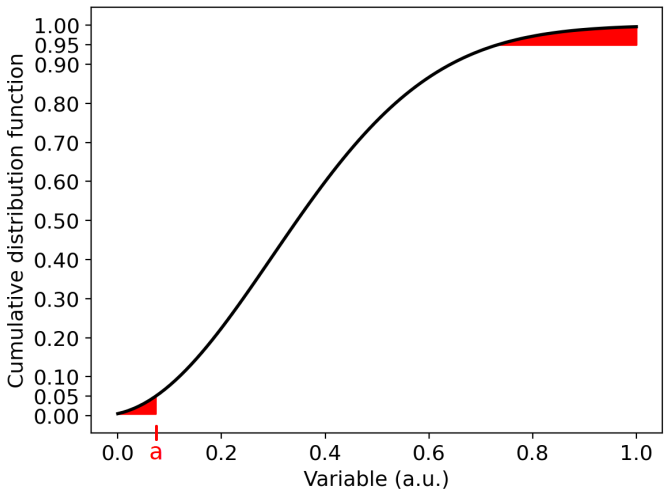

Figure 5.8: Example of a CDF, where the risk zones are marked in red, when $a > 0.02$

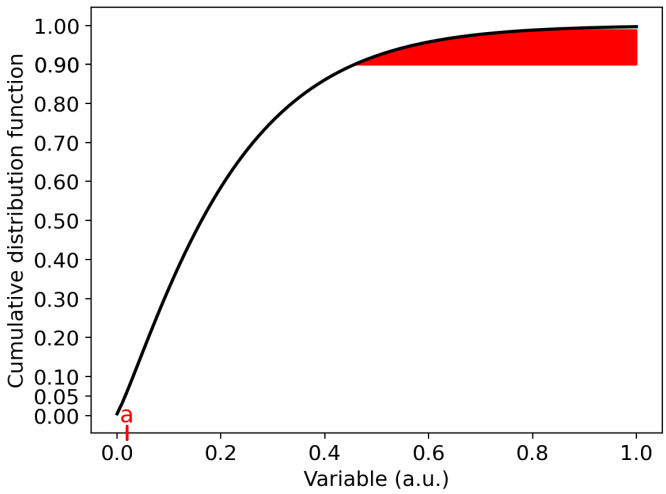

Figure 5.9: Example of a CDF, where the risk zones are marked in red, when $a \leq 0.02$

Based on the variables, a global risk score was calculated. The global risk score consists of the sum of all subrisks. Subrisks are either 0 or 1 based on whether the score for a variable of a person lied within the risk zone of the corresponding CDF or not. If it lies within the risk zone, a 1 is assigned, otherwise a 0 is assigned. To exemplify, if a certain sample (corresponding to a 20-minute acquisition) had 5 variables within the respective risk zones, the global risk score was defined as 5.

Finally, each sample was labeled differently for classification and regression algorithms. For the classification algorithms, a global risk score of 0 was defined as *0* (no risk), while all the other global risk scores were defined as *1* (risk). On the other hand, for the regression algorithms, the samples were directly labeled with the calculated global risk score.

5.5.5 Feature extraction and selection

For occupational risk prediction, we considered an extended feature set, which includes EMG features, heart rate features, noise features and information gathered from the questionnaires. Since this ergonomic risk was modelled by postural variables, we used features unrelated to posture to try to predict it.

The process of feature extraction from the EMG signals was exactly the same as described in Section 5.4.6, giving rise to 204 features, 102 from each muscleBAN (12 statistical, 18 temporal, and 4 spectral for the entire EMG signal, resting portions, and moving portions).

The heart rate features were directly extracted from each of the pre-processed (resampled) 20-minute heart rate signals. A total of 3 features were considered, namely the minimum, maximum, and mean values.

To extract the noise features, the smartphone synchronized files of the subjects who performed whole-day acquisitions had to be trimmed to 20-minute periods, corresponding to the periods in which the smartwatch and muscleBANs also acquired. The extracted features were the same as for the heart rate (minimum, maximum, and mean values).

Finally, the answers to the questionnaires were also used as features. The complete list of questionnaires is available in Table 4.1. In addition to the personal information (presented in Section 5.4.6), the psychosocial, biomechanical, and environmental questionnaires were also used. The answers to the questionnaires corresponding to these three last categories were converted into scores, ranging from 0 to 1. The questionnaires provided a total of 51 additional features (9 personal, 31 psychosocial, 5 biomechanical, and 6 environmental) to feed the ML models.

In the end, there was a total of 261 features available to predict the relative occupational risk. The features were then selected through different methods, in the same process as explained in Section 5.3.2.4. The algorithms used for RFE are characterized in Tables 5.4 and 5.12. In addition to the same classification algorithms (logistic regression, RF, and SVM), we also used regression algorithms, namely linear regression, RF regression, and support vector regression. For linear regression, the features with a Variance

Inflation Factor (VIF) greater than 10 were also removed, to avoid the multicollinearity phenomenon. The VIF is a measure for the correlation between multiple independent variables in a multivariate linear regression. A high correlation between the variables would cause multicollinearity, which leads to an increase in the standard error of the regression coefficients [86].

Table 5.12: Characterization of the regression algorithms used for RFE.

Algorithm	Parameters
Linear regression	-
RF regression	criterion: squared_error n_estimators: 100
Support vector regression	C: 1.0 kernel: linear

5.5.6 Model optimization

As previously said, we used classification and regression ML techniques to predict the relative ergonomic occupational risk. Using the classification algorithms, we tried to distinguish between *risk* and *no risk*. On the other hand, using the regression algorithms, we tried to predict a numeric value for the risk score.

To develop the ML models, for both classification and regression, the dataset was divided into 70% for training and 30% for testing in a stratified way. We tested a total of six different algorithms, three classifiers and three regressors, as aforementioned.

Again, we used grid search to choose the hyperparameters for the algorithms (as in Section 5.3.2.5). The grids used for hyperparameter selection are presented in Tables 5.5 and 5.13. Like logistic regression, linear regression also does not allow grid search, for the same reasons.

Table 5.13: Values used for the regressors' hyperparameters when performing grid search.

Algorithm	Hyperparameter	Values
RF regression	n_estimators	50, 100, 250, 500
	max_depth	5, 6, 7, 8, 9, 10
	criterion	squared_error, absolute_error, poisson
Support vector regression	C	0.1, 1, 10, 100, 1000
	gamma	1, 0.1, 0.01, 0.001, 0.0001
	degree	2, 3, 4, 5
	kernel	linear, poly, rbf, sigmoid

Each model was trained 5 times, using different seeds (52374, 52347, 52734, 52743, and 52473). Since we considered occupational risk to be an integer, it is important to emphasize that the values returned by the regression models have been rounded to the nearest integer. Then, the Mean Absolute Error (MAE) was calculated for each iteration

and each regressor, to evaluate their performance. The performance of the classifiers was evaluated through their accuracies, precisions, recalls, and F_1 scores.

Additionally, to verify the validity of the linear regression model, some statistical tests were carried out to check the fulfillment of certain assumptions that are essential to perform linear regression. These assumptions are all related to the residuals, which correspond to the individual differences between each of the real values and the estimated values.

Firstly, the autocorrelation of the residuals was studied through the Durbin-Watson test, where a test statistic of 2 indicates no autocorrelation. The heteroskedasticity, corresponding to the presence of non-constant variance in the residuals, was assessed through a Goldfeld-Quandt test. Finally, the normality of the residuals was verified using a Jarque-Bera test.

6

Results and discussion

This chapter presents the results obtained by applying the methods described in Chapter 5, and their respective discussion. This includes the sitting detection algorithms, the pain prediction model, and the occupational risk model.

6.1 Sitting detection

6.1.1 Sit-stand transitions detection

Our first approach for trying to assess when the participants of the main study were seated was to develop an algorithm capable of automatically detecting the sit-to-stand and stand-to-sit transitions, and distinguishing between both of them. For that, we used the accelerometer of the smartphone, which was placed on the subjects' chest.

After collecting data of people sitting down and standing up with the smartphone on their chest, we performed some tests to understand if this solution would be viable for the intended purpose. The results for these tests are presented and discussed in the following sections.

6.1.1.1 Transition patterns

The acquired patterns of sit-to-stand and stand-to-sit transitions are shown superimposed, for the three axes of the accelerometer, in Figure 6.1. Analyzing the figure, it seems that both types of transitions present relatively well-defined patterns for the accelerometer's y- and z-axis. However, when comparing all three axes with each other, the z-axis shows the least variations for both transitions between subjects. Thus, this axis should allow for a better detection of sit-to-stand and stand-to-sit transitions.

6.1.1.2 Transition detection with z-axis patterns of different subjects

In this first approach, one z-axis pattern of each type of transition was extracted from the signals of three different subjects. The final output obtained after comparison between the patterns and the signal is presented in Figure 6.2.

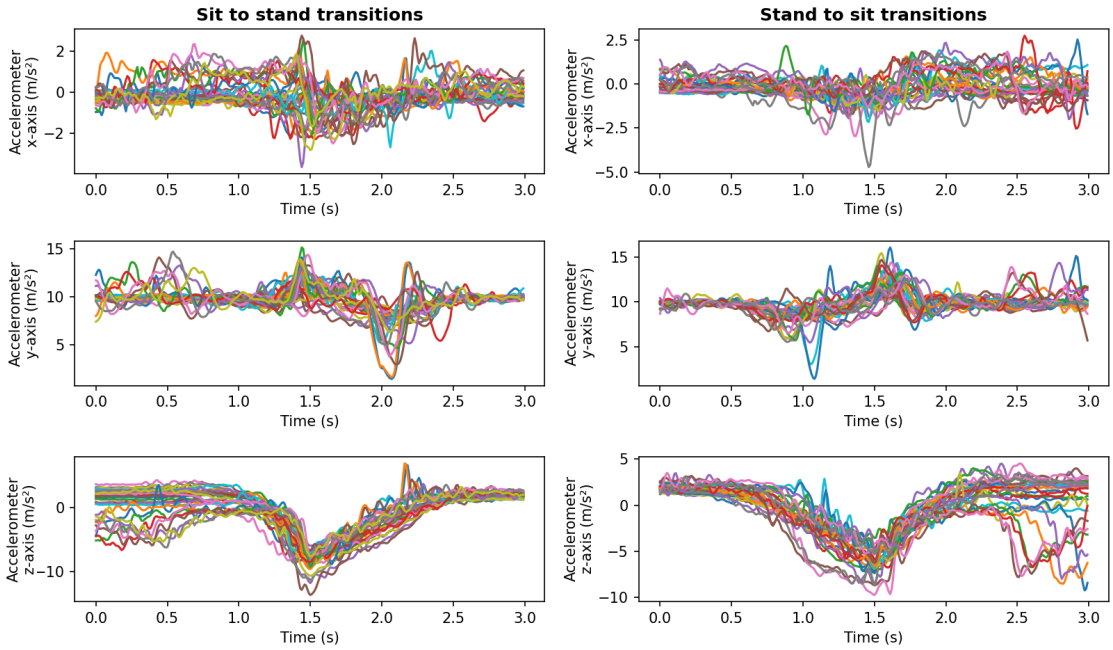

Figure 6.1: Acquired patterns of the sit-to-stand (left) and stand-to-sit (right) transitions.

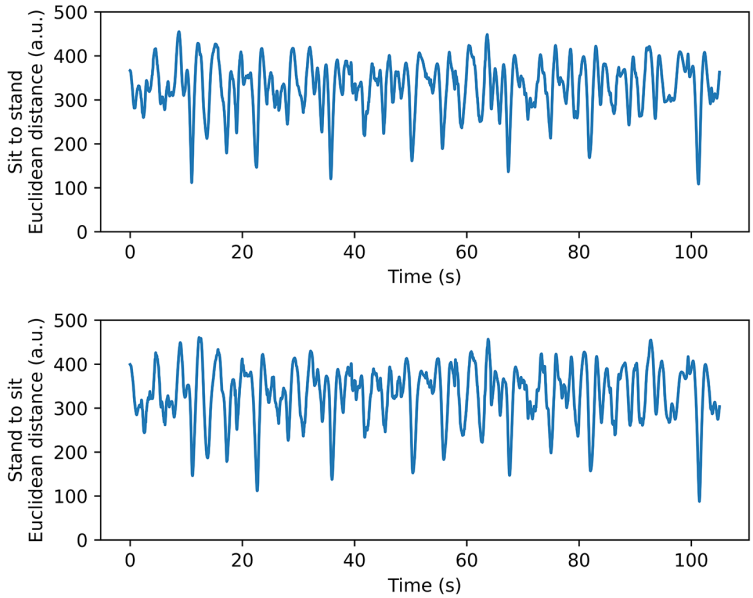

Figure 6.2: Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's z-axis and three different subjects.

The figure shows the Euclidean distances for both the sit-to-stand transition (top) and the stand-to-sit transition (bottom). Ideally, each transition would be recognized by an Euclidean distance of zero, because that would mean the normalized 3-second signal window and the 3-second pattern were equal. However, in real context there are some variations such as movement execution and speed of the transition that make a full identity between the pattern and recorded signal impossible. Instead, the transitions should be identifiable by prominent minima. Nevertheless, if we compare the top chart with the bottom chart of Figure 6.2, there are several minima that appear in the same position in both charts, making it difficult to identify whether it is a sit-to-stand or stand-to-sit transition.

6.1.1.3 Transition detection with multi-axis patterns of a single subject

In another approach, the accelerometer's three axes were used. For each type of transition, only one pattern (for each axis) was extracted from one of the acquired subjects. The final output, which represents the mean distance measure, is shown in Figure 6.3 for each type of transition. As can be seen in the figure, the use of a combination of the three axes of the accelerometer introduced more noise to the time-series, when comparing with Figure 6.2. Thus, the identification of the prominent minima has become more difficult to achieve.

6.1.1.4 Transition detection with multi-axis patterns of different subjects

In a final test, we tried to combine three patterns of the individual axes of the three acquired subjects, giving rise to 9 different distance measures. The mean of these 9 patterns is represented in Figure 6.4. The figure shows that this last combination introduced even more noise to the time-series, when comparing with Figures 6.2 and 6.3. Additionally, the distances are even further away from zero. This way, the identification of the prominent peaks and its corresponding type of transition was even more difficult.

6.1.1.5 Discussion

As can be seen in Figures 6.2, 6.3, and 6.4, none of the three approaches presented satisfactory results. The use of multi-axis patterns made the detection even more difficult, probably due to the x-axis being quite noisy (see Figure 6.1). Thus, the introduction of this axis into the distance metrics made the final time-series also noisy. Additionally, there seem to be too many differences between subjects, which also makes the identification hard when using patterns of different subjects.

In order to potentially make a reliable identification, it would be necessary to extract many patterns from each transition of several people. Nevertheless, the processes of acquiring data, processing, and verifying the quality of the results would be very time-consuming.

Furthermore, even within the same subject, sitting down and standing up patterns may vary from day to day. For example, a person experiencing back pain would likely

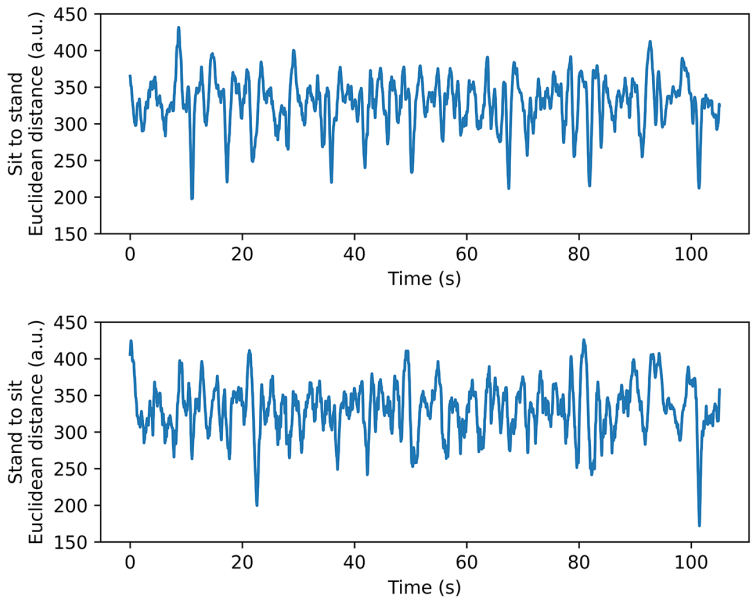

Figure 6.3: Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's three axes.

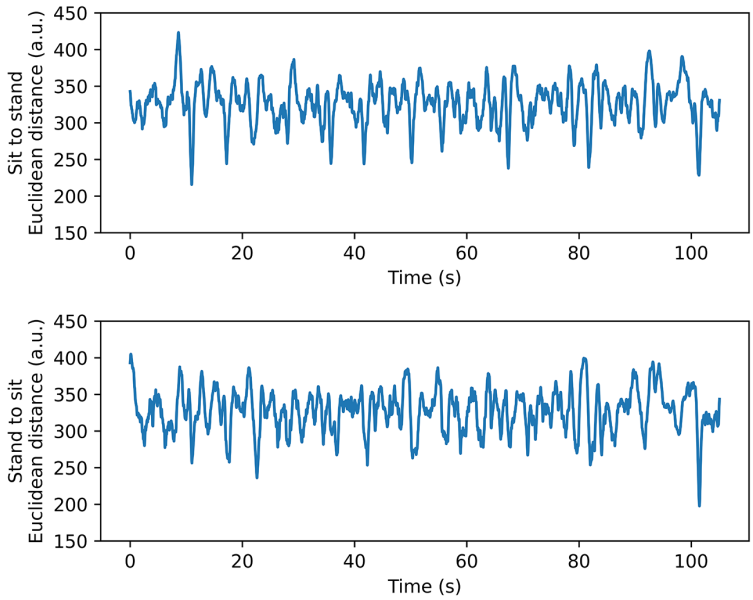

Figure 6.4: Euclidean distances between the signal and the pattern of the sit-to-stand (top) and stand-to-sit (bottom) transitions, considering the accelerometer's three axes and three different subjects.

make these transitions in a different way, which would lead to them possibly not being identified by the algorithm.

For the reasons mentioned above, the use of an algorithm of this type would not be possible in a real-world and/or commercial context. While this approach would be computationally inexpensive, it was not able to provide a reliable identification of the transitions. Therefore, we decided to explore a Machine Learning (ML) path for removing the periods in which the individuals were not seated from the main acquisitions.

6.1.2 Sitting and walking distinction

As the non-ML path did not give the intended results for sitting detection, we opted for a ML approach, described in the following sections. For this purpose, we developed two different models to distinguish between sitting and walking. The first one receives accelerometer and gyroscope features as input, while the second one only receives accelerometer features. For each model, different classification algorithms and window sizes were tested.

6.1.2.1 Model based on accelerometer and gyroscope data

For the first version of the sitting/walking classifier, 30 features (12 statistical and 18 temporal features, described in Tables 5.2 and 5.3) were extracted for each axis of both accelerometer and gyroscope signals, which corresponds to a feature vector composed of a total of 180 features. After feature extraction, three different processes of feature selection were performed. Firstly, the highly correlated and zero variance features were removed from the dataset. For the 3-second windows, a total of 87 features were removed by correlation and 7 by variance. On the other hand, for the 5-second windows, a total of 90 features were removed by correlation and 7 by variance. Then, features were selected through Recursive Feature Elimination (RFE). The number of selected features for each algorithm and window size is presented in Table 6.1.

Table 6.1: Number of selected features by RFE for sitting/walking classification with accelerometer and gyroscope.

Algorithm	Window size (s)	Number of selected features
Logistic regression	3	3
	5	10
RF	3	4
	5	19
SVM	3	12
	5	7

The next step was to perform grid search on the Random Forest (RF) and Support Vector Machine (SVM) algorithms. As the logistic regression does not allow grid search, the hyperparameters were manually selected. The final selection of hyperparameters for each algorithm is shown in Table 6.2.

Having defined the most informative features and the hyperparameters to use for each algorithm, all the models were fitted on the training data. The statistical analysis for each model using accelerometer and gyroscope is presented in Table 6.3.

Table 6.2: Characterization of the models used for sitting/walking classification with accelerometer and gyroscope.

Algorithm	Window size (s)	Hyperparameters
Logistic regression	3	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
	5	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
RF	3	criterion: gini max_depth: 5 n_estimators: 250
	5	criterion: gini max_depth: 5 n_estimators: 100
SVM	3	C: 0.1 gamma: 0.1 degree: - kernel: rbf
	5	C: 0.1 gamma: 1 degree: 2 kernel: poly

Table 6.3: Statistical analysis of the sitting/walking classifier's accuracies with accelerometer and gyroscope.

Model	Windows (s)	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	3	98.9	100.0	99.7	0.4
	5	99.1	100.0	99.7	0.4
RF	3	99.7	100.0	99.9	0.1
	5	99.6	100.0	99.8	0.2
SVM	3	99.5	100.0	99.9	0.2
	5	99.6	100.0	99.9	0.2

6.1.2.2 Model based on accelerometer data

To assess whether similar classification results could be achieved with only the accelerometer as input to the model, the models were re-trained and re-configured. Not using the gyroscope, the new total number of extracted features is 90 (12 statistical and 18 temporal

features for each of the accelerometer's three axes). The new number of selected features by RFE for each algorithm and window size is shown in Table 6.4.

Table 6.4: Number of selected features by RFE for sitting/walking classification using only accelerometer.

Algorithm	Window size (s)	Number of selected features
Logistic regression	3	3
	5	10
RF	3	3
	5	8
SVM	3	4
	5	3

The choice of hyperparameters has also changed, as presented in Table 6.5. The hyperparameters for logistic regression were the same as before, because they were manually selected, unlike RF and SVM, where the hyperparameters were found via grid search.

The models were again fitted on the training data. The statistical analysis for each model using only the accelerometer is presented in Table 6.6, considering both 3- and 5-second windows.

6.1.2.3 Discussion

Regarding the accelerometer and gyroscope models, Table 6.1 shows that, for the logistic regression and RF algorithms, an increase in window size leads to an increase in the number of selected features. On the other hand, for the SVM algorithm, the number of features decreases with an increase in window size.

However, observing Table 6.3, it can be seen that window size and respective number of selected features have a very small impact on the quality of the classification, taking into account its accuracy. Moreover, all models present excellent and similar accuracies, using either 3- or 5-second windows. All models reached a maximum accuracy of 100%, with the mean accuracies ranging from 99.7% to 99.9%. The highest mean accuracies are achieved by the 3-second RF, 3-second SVM and 5-second SVM, reaching mean accuracies of 99.9%. The logistic regression models have the lowest minimum and mean accuracies, as well as the highest standard deviations.

There are some differences if the results from the accelerometer and gyroscope models are compared with previous studies (see Section 3.1.1). Regarding study [43], our maximum accuracy is higher (100% vs 95.7%), but its authors used more classes and different algorithms. Furthermore, comparing the SVM results with sitting detection in study [46], it can be seen that our maximum accuracy was also higher (100% vs 86.9%).

If Tables 6.1 and 6.4 are compared, it is possible to verify that the number of features needed for classification has reduced for the RF and SVM algorithms, but remained

Table 6.5: Characterization of the models used for sitting/walking classification using only accelerometer.

Algorithm	Window size (s)	Hyperparameters
Logistic regression	3	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
	5	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
RF	3	criterion: gini max_depth: 5 n_estimators: 50
	5	criterion: gini max_depth: 5 n_estimators: 50
SVM	3	C: 0.1 gamma: - degree: - kernel: linear
	5	C: 0.1 gamma: 1 degree: 2 kernel: poly

Table 6.6: Statistical analysis of the sitting/walking classifier's accuracies using only accelerometer.

Model	Windows (s)	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	3	98.9	100.0	99.7	0.4
	5	98.7	100.0	99.2	0.5
RF	3	99.7	100.0	99.9	0.1
	5	100.0	100.0	100.0	0.0
SVM	3	99.5	100.0	99.8	0.2
	5	99.1	100.0	99.6	0.3

unchanged for the logistic regression algorithm. This is due to the fact that the gyroscope features were never selected for the logistic regression model, even when they were available for selection.

Analyzing Table 6.6, it can be seen that the accuracies are still very high, even using only the accelerometer signals. The maximum accuracy, particularly, still stands at 100% for all models. The minimum accuracies of the 3-second models also did not change. The minimum accuracies of the 5-second logistic regression and SVM models slightly decreased by 0.4% and 0.5%, respectively. Regarding the mean accuracies, the 3-second RF model did not experience variations, while the 5-second one has even increased by 0.2%. The 5-second RF model, particularly, has reached an outstanding accuracy of 100% for all five seeds.

Comparing the accelerometer-only models with state of the art studies, our results are still higher. Regarding study [43], which distinguishes between more classes and uses other algorithms, our maximum accuracy is higher (100% vs 99.4%). Regarding the RF model, we also achieved a higher maximum accuracy when comparing with study [45] (100% vs 99.3%) but, again, its authors considered more classes. Regarding the SVM model, we surpassed the accelerometer-only sitting detection maximum accuracy of study [46] (100% vs 85.9%).

Thus, it can be deduced that the gyroscope is not necessary to make the distinction between sitting and walking. This sensor slightly improved the performance of almost all tested models but, in certain situations, it can be preferable to neglect it. Therefore, not using the gyroscope can make the sitting detection algorithm more lightweight and with less need for computational capacity.

High accuracies were already expected, given the fact that sitting and walking are very distinct activities. Regarding linear acceleration, sitting gives rise to more constant accelerometer signals, containing less evident variations. In opposition, walking has a very distinctive and repetitive pattern, with higher standard deviation, which is visible in its accelerometer signals.

Given its high accuracy and less need for computational power, the 5-second accelerometer RF model was chosen to remove the periods in which the participants of the main study were not seated during the acquisitions.

6.2 Pain prediction

To predict the pain reported by the participants of the main study in the morning and in the afternoon, we developed two different models. The first one only uses Electromyography (EMG) features to predict pain, while the second one adds the personal information extracted from the questionnaires to the feature vector.

6.2.1 Model based on electromyography data

Firstly, we trained different models to predict pain based only on features from the EMG signals. These features were extracted from the complete signals, but also simultaneously from its resting and moving portions (see Sections 5.4.2 and 5.4.6).

The EMG signals provided a total of 204 features, with 102 belonging to each muscleBAN (12 statistical, 18 temporal, and 4 spectral features from the entire signal, resting portions, and moving portions). The highly correlated and zero variance features were removed from the dataset, with a total of 72 features being removed by high correlation and 2 by low variance. The remaining features were then selected through RFE. The number of selected features of each signal portion for each algorithm is presented in Table 6.7.

Table 6.7: Number of selected EMG features by RFE for pain prediction.

Algorithm	Signal portion	Selected features
Logistic regression	Resting	1
	Moving	0
	Complete	0
RF	Resting	14
	Moving	25
	Complete	26
SVM	Resting	7
	Moving	11
	Complete	12

After selecting the features, grid search was performed to find the best hyperparameters for the RF and SVM algorithms, while the hyperparameters for logistic regression were manually selected. The hyperparameters used for the three models are presented in Table 6.8.

After feature and hyperparameter selection, the models were trained, using five different seeds. The statistical analysis of the accuracies of each model is presented in Table 6.9.

6.2.2 Model based on electromyography and personal data

To build the EMG and personal data model, the complete dataset was used. The new feature vector extracted from this data consisted of the EMG features and personal information, namely the age, gender, height, weight, Body Mass Index (BMI), International Physical Activity Questionnaire (IPAQ) score, number of working hours per week, number of hours sitting per working days, and number of hours sitting per weekend.

Using this dataset, which slightly reduced the number of samples (see Section 5.4.3), we performed feature selection. This time, 69 features were removed by high correlation and 2 by low variance. The new sets of selected features for each model are presented in Table 6.10.

Table 6.8: Characterization of the models used for pain prediction without personal data.

Algorithm	Hyperparameters
Logistic regression	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
RF	criterion: gini max_depth: 7 n_estimators: 50
SVM	C: 10 gamma: 0.1 degree: - kernel: rbf

Table 6.9: Statistical analysis of the pain classifier's accuracies without personal data.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	57.0	61.6	59.3	1.6
RF	53.5	62.8	57.0	3.5
SVM	53.5	65.1	59.5	4.1

Table 6.10: Number of selected features by RFE for pain prediction, considering the personal data.

Algorithm	Feature source	Selected features
Logistic regression	Resting EMG	19
	Moving EMG	24
	Complete EMG	25
	Personal information	9
RF	Resting EMG	2
	Moving EMG	3
	Complete EMG	2
	Personal information	5
SVM	Resting EMG	8
	Moving EMG	11
	Complete EMG	11
	Personal information	6

The process of hyperparameter selection by grid search was repeated as well. The manually selected hyperparameters for logistic regression remained the same. The new combination of hyperparameters for each model is listed in Table 6.11.

The models were trained, using both the EMG features and personal data. Table 6.12 presents the statistical analysis of the pain classifier's accuracies, when using the complete set of features.

Table 6.11: Characterization of the models used for pain prediction, considering the personal data.

Algorithm	Hyperparameters
Logistic regression	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
RF	criterion: entropy max_depth: 8 n_estimators: 100
SVM	C: 10 gamma: 1 degree: - kernel: rbf

Table 6.12: Statistical analysis of the pain classifier's accuracies when considering the personal data.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	51.8	72.3	58.6	7.3
RF	81.9	89.2	86.3	2.7
SVM	71.1	79.5	74.9	2.8

6.2.3 Discussion

Regarding the EMG-only model, Table 6.7 shows that the signal division into resting and moving periods had high importance for the RF and SVM algorithms, with a lot of the selected features coming from both these portions of the EMG signals. On the other hand, the logistic regression model achieved its best classification using only one EMG feature from the signals' resting sections. This means that, although we can provide more features to the logistic regression model, it will not improve the accuracy of its classification. In fact, the logistic regression algorithm only needs one feature to compute the sigmoid function that models the probabilities used to assign either *0* or *1* to the output. Adding more features could modify this function slightly, but apparently not enough to significantly alter its decision boundary and final predictions. The SVM algorithm, similarly, also is capable of defining the hyperplane using only one feature but, in this case, it needed more

features to provide more accurate classifications. On the other hand, the RF algorithm, although it can be used with only one feature, benefits from the use of more features, which allow the definition of more nodes in its trees.

Table 6.9 shows that the SVM model has the highest mean accuracy (59.5%), as well as the highest maximum accuracy (65.1%). The logistic regression model presents a similar mean accuracy (59.3%), but its minimum accuracy is higher (57.0%) and its maximum accuracy is lower (61.6%). None of the models was able to surpass a mean accuracy of 60% for pain prediction using only EMG features.

Given that only accuracies of maximum 65.1% were achieved, the models are not able to properly classify the pain given the input data. This means that the decision space is either of higher complexity or the models are not adequate for the problem. To further investigate this matter the feature space was increased by adding features derived from the participants personal information.

With regard to the EMG and personal data models, and analyzing Table 6.10, it should be noted that, of the 9 features from the questionnaires, at least 5 were selected by all models. This suggests that personal information has major importance when predicting pain. Additionally, the number of EMG features selected by RF drastically reduced from 65 to 12, when comparing to the dataset that did not contain personal information. This appears to indicate that the RF algorithm found a strong relation between personal information and pain, giving primacy to features from the questionnaires and reducing the need for EMG features. However, EMG features and their splitting between resting and moving still seem to be important for the models. On the other hand, the logistic regression model, interestingly, greatly increased the number of selected features. This indicates a strong need of this model to combine both types of features (EMG and questionnaires) to achieve accurate results.

Taking a look at the results present in Table 6.12, the logistic regression model seems to be the least affected by the questionnaires' features. Its mean accuracy remained almost unchanged (59.3% vs 58.6%), but its maximum accuracy benefited from the questionnaires' features and improved a lot (61.6% vs 72.3%). However, its minimum accuracy decreased by 5.2% and its standard deviation increased by 5.7%. On the other hand, the RF and SVM models were greatly affected positively by personal information. The new RF model, particularly, had a drastic improvement in terms of accuracy, with its mean accuracy increasing from 57.0% to 86.3%. Thus, through the addition of questionnaire features, the classification accuracy of RF improved drastically, exceeding the accuracies of logistic regression and SVM. This may be due to the fact that random forest is an ensemble learning method, taking into account the outputs of multiple decision trees. Although some of the decision trees that compose this algorithm possibly return wrong predictions, which can be considered noise, this noise is not manifested in the final prediction.

Therefore, we were able to make predictions on the presence or absence of pain reported by the participants in the morning and afternoon questionnaires, with relatively high accuracies. Although classification based on EMG features alone did not give satisfactory

results, the personal/demographic data extracted from the questionnaires proved to be of extreme relevance when combined with EMG features. Furthermore, the process of feature selection also benefited from the splitting that was applied to the 20-minute EMG signals, dividing them between the periods where the subject was more static (rested) and the periods where he/she was moving (while seated).

The resting periods can contain the subject's relaxation patterns, assuming that he/she is able to relax while not moving. In these periods, the muscle can recover (at least partially), which corresponds to relaxation patterns. On the other hand, the complete EMG signal turns out to be a mean of the fluctuations present in its resting and moving portions. When the EMG signal has higher mean, that corresponds to higher strain on the muscles, and potentially more pain.

6.3 Occupational risk prediction

6.3.1 Risk modelling

After the extraction of variables related to postural variability, we had a total of 21 variables to model the relative risk of the 40-subject population. The variables, after normalization, are characterized in Table 6.13 by its mean value and standard deviation.

As can be seen in the aforementioned table, the variables with a standard deviation greater than 0.10 were selected to model the population's occupational risk. The variable referring to the mean velocity of the subject's third most common posture was also excluded from the analysis, despite having a standard deviation of less than 0.10, since it was not possible to achieve a proper fit using the Weibull distribution, as can be seen in Figure 6.5.

For each of the 12 selected variables, a Weibull distribution curve was fitted, as shown in Figures 6.6 and 6.7. Then, using the Cumulative Distribution Function (CDF), the risk zones were defined for each variable, resulting in the conditions presented in Table 6.14. These conditions limit the risk zone(s) for each variable.

6.3.2 Classification models

In a first approach, we trained different classification models to predict the presence of relative occupational risk, based on EMG signals, heart rate signals, noise signals, and questionnaires, which are not directly related to posture, used to model the risk.

The EMG signals provided a total of 204 features, with 102 belonging to each muscleBAN (12 statistical, 18 temporal, and 4 spectral features from the entire signal, resting portions, and moving portions). The heart rate and noise signals provided 3 additional features each. The questionnaires provided a total of 51 additional features (9 personal, 31 psychosocial, 5 biomechanical, and 6 environmental). Then, 76 highly correlated and 2 zero variance features were removed from the dataset. The remaining features were

Table 6.13: Statistical analysis of the extracted postural variables.

Variable	Mean (a.u.)	Standard deviation (a.u.)	Selected
Number of changes in posture	0.17	0.14	✓
Number of different postures	0.37	0.20	✓
Mean time of transition between postures	0.04	0.07	
Time spent in the most common posture	0.59	0.24	✓
Time spent in the second most common posture	0.34	0.20	✓
Time spent in the third most common posture	0.23	0.21	✓
Time spent in the remaining postures	0.17	0.17	✓
Total distance covered	0.42	0.17	✓
Distance covered in the most common posture	0.40	0.19	✓
Distance covered in the second most common posture	0.17	0.17	✓
Distance covered in the third most common posture	0.06	0.12	✓
Distance covered in the remaining postures	0.21	0.17	✓
Variance in the most common posture	0.05	0.08	
Variance in the second most common posture	0.04	0.09	
Variance in the third most common posture	0.04	0.09	
Mean variance in the remaining postures	0.07	0.09	
Mean velocity	0.05	0.11	✓
Mean velocity in the most common posture	0.05	0.10	
Mean velocity in the second most common posture	0.02	0.05	
Mean velocity in the third most common posture	0.09	0.15	
Mean velocity in the remaining postures	0.07	0.09	

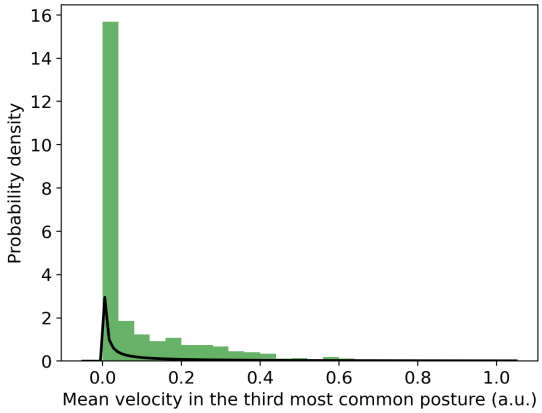

Figure 6.5: Distribution of the variable referring to the mean velocity of the subject's third most common posture.

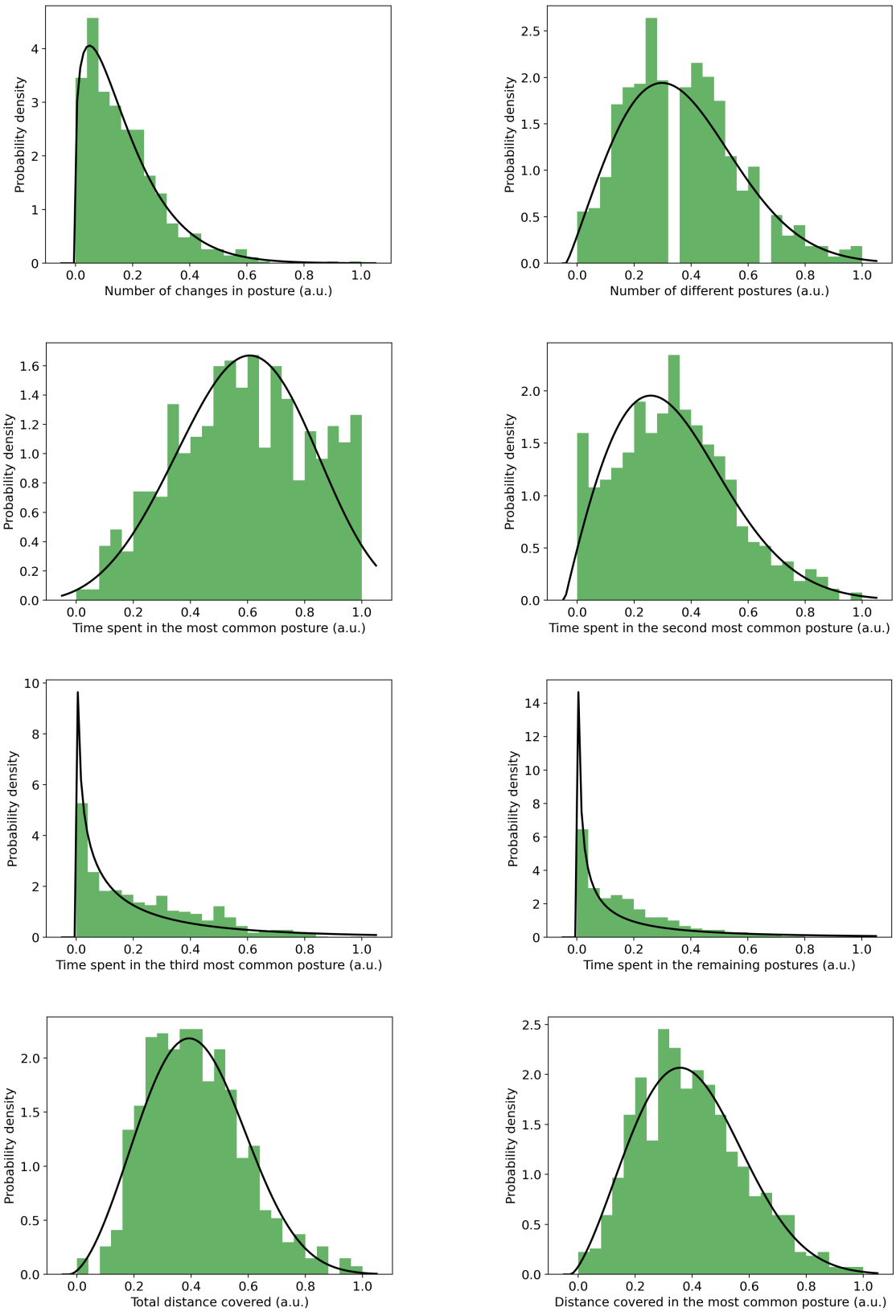

Figure 6.6: Distributions of the 12 selected variables for risk modelling.

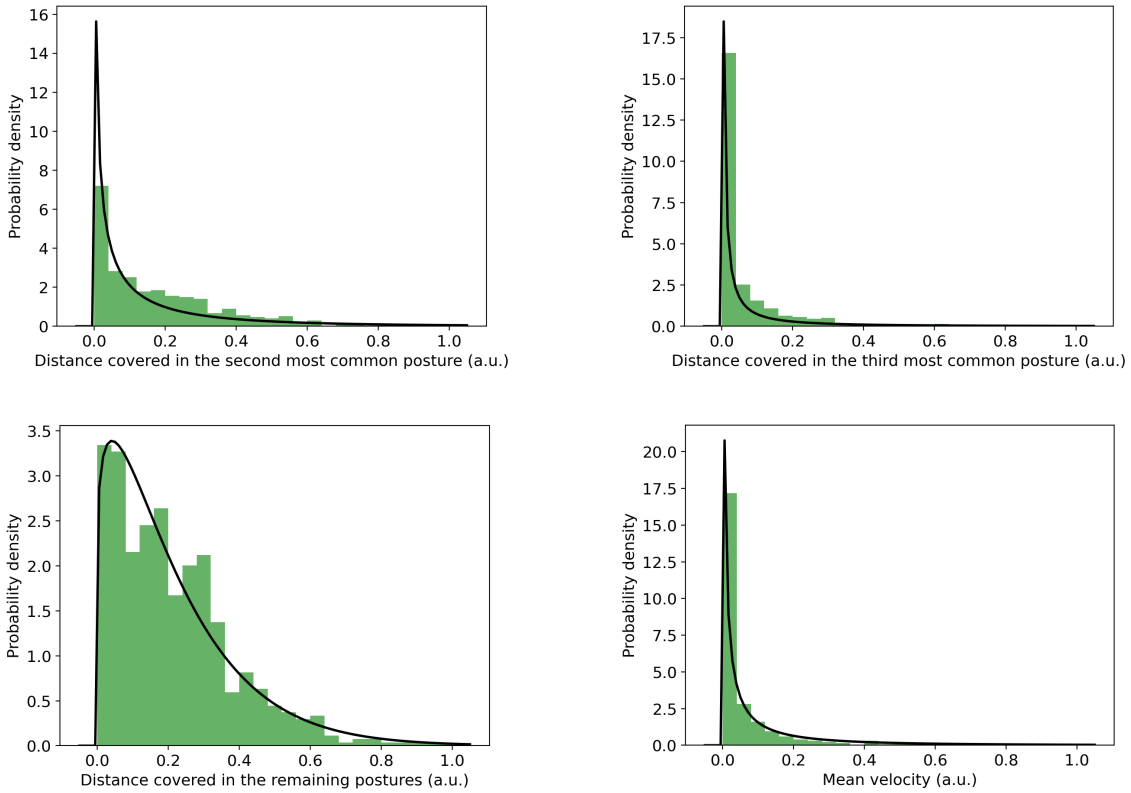

Figure 6.7: Distributions of the 12 selected variables for risk modelling (cont.).

Table 6.14: Definition of risk zones for each of the selected variables, where x represents the value of the variable.

Variable	**Risk zone(s) condition**
Number of changes in posture	$x < 0.35$
Number of different postures	$0.07 < x < 0.73$
Time spent in the most common posture	$0.20 < x < 0.96$
Time spent in the second most common posture	$0.05 < x < 0.72$
Time spent in the third most common posture	$x < 0.68$
Time spent in the remaining postures	$x < 0.67$
Total distance covered	$0.14 < x < 0.72$
Distance covered in the most common posture	$0.11 < x < 0.72$
Distance covered in the second most common posture	$x < 0.47$
Distance covered in the third most common posture	$x < 0.15$
Distance covered in the remaining postures	$x < 0.46$
Mean velocity	$x < 0.32$

selected through RFE. The number of selected features for each source and for each algorithm is presented in Table 6.15.

The hyperparameters used for the three models are presented in Table 6.16, taking into account that grid search was performed for the RF and SVM algorithms.

After feature and hyperparameter selection, the models were trained, using five different seeds. The statistical analyzes of specific evaluation metrics, namely accuracy, precision, recall, and F_1 score, are presented for each model in Tables 6.17, 6.18, 6.19, and 6.20, respectively.

Table 6.15: Number of selected features by RFE for the occupational risk classifier.

Algorithm	Feature source	Selected features
Logistic regression	Resting EMG	0
	Moving EMG	1
	Complete EMG	1
	Heart rate	0
	Noise	0
	Questionnaires	0
RF	Resting EMG	18
	Moving EMG	34
	Complete EMG	30
	Heart rate	3
	Noise	3
	Questionnaires	45
SVM	Resting EMG	0
	Moving EMG	0
	Complete EMG	1
	Heart rate	0
	Noise	0
	Questionnaires	0

Table 6.16: Characterization of the classification models used for occupational risk prediction.

Algorithm	Hyperparameters
Logistic regression	C: 1.0 max_iter: 1000 penalty: none solver: lbfgs
RF	criterion: entropy max_depth: 5 n_estimators: 50
SVM	C: 1000 gamma: 0.01 degree: - kernel: rbf

6.3.3 Regression models

In a second approach, three different regression algorithms were tested to predict the relative occupational risk of the population of 40 subjects.

Using a feature selection process similar to the classification models, 76 features were removed by high correlation and 2 by low variance. The new sets of selected features of each source and for each model are presented in Table 6.21.

The process of hyperparameter selection by grid search was also similarly used for the RF regression and support vector regression algorithms. The new combination of hyperparameters for each model is listed in Table 6.22.

The regression models were fitted on the training data, using five different seeds. The statistical analysis of their Mean Absolute Error (MAE) is shown in Table 6.23, taking into account that the outputs were rounded to the nearest integer. Figures 6.8, 6.9, and 6.10 present the real values and respective values predicted by the regression models with the lowest MAEs.

Regarding the statistical tests performed to verify the assumptions for linear regression, their results are presented in Table 6.24, with the respective test statistics and p-values.

Table 6.17: Statistical analysis of the occupational risk classifier's accuracies.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	62.8	71.3	65.7	3.0
RF	55.8	69.0	63.3	5.5
SVM	60.5	65.9	62.0	2.1

Table 6.18: Statistical analysis of the occupational risk classifier's precisions.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	62.8	75.0	69.3	3.9
RF	52.9	71.1	63.4	7.5
SVM	56.5	77.8	68.3	7.9

Table 6.19: Statistical analysis of the occupational risk classifier's recalls.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	35.6	55.9	45.1	6.5
RF	30.5	54.2	44.1	10.4
SVM	27.1	59.3	35.9	12.1

Table 6.20: Statistical analysis of the occupational risk classifier's F_1 scores.

Model	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Logistic regression	47.2	64.1	54.4	5.4
RF	38.7	61.5	51.8	9.9
SVM	38.6	57.9	45.3	7.2

Table 6.21: Number of selected features by RFE for the occupational risk regressor.

Algorithm	Feature source	Selected features
Linear regression	Resting EMG	2
	Moving EMG	8
	Complete EMG	7
	Heart rate	0
	Noise	0
	Questionnaires	15
RF regression	Resting EMG	8
	Moving EMG	11
	Complete EMG	3
	Heart rate	1
	Noise	1
	Questionnaires	5
Support vector regression	Resting EMG	1
	Moving EMG	2
	Complete EMG	1
	Heart rate	0
	Noise	0
	Questionnaires	0

Table 6.22: Characterization of the regression models used for occupational risk prediction.

Algorithm	Hyperparameters
Linear regression	-
RF regression	criterion: absolute_error max_depth: 5 n_estimators: 500
Support vector regression	C: 1 gamma: 1 degree: - kernel: rbf

Table 6.23: Statistical analysis of the occupational risk regressor's MAEs.

Model	Minimum	Maximum	Mean	Standard deviation
Linear regression	1.02	1.08	1.05	0.02
RF regression	0.77	0.95	0.84	0.06
Support vector regression	0.81	0.91	0.85	0.04

Table 6.24: Results of the statistical tests performed to verify the fulfillment of the assumptions for linear regression.

Assumption	Test name	Subset	Test statistic	p-value
Autocorrelation	Durbin-Watson	Train	2.18	-
		Test	1.94	-
Heteroskedasticity	Goldfeld-Quandt	Train	0.70	0.99
		Test	1.55	0.05
Normality	Jarque-Bera	Train	187.04	2.42×10^{-41}
		Test	76.98	1.92×10^{-17}

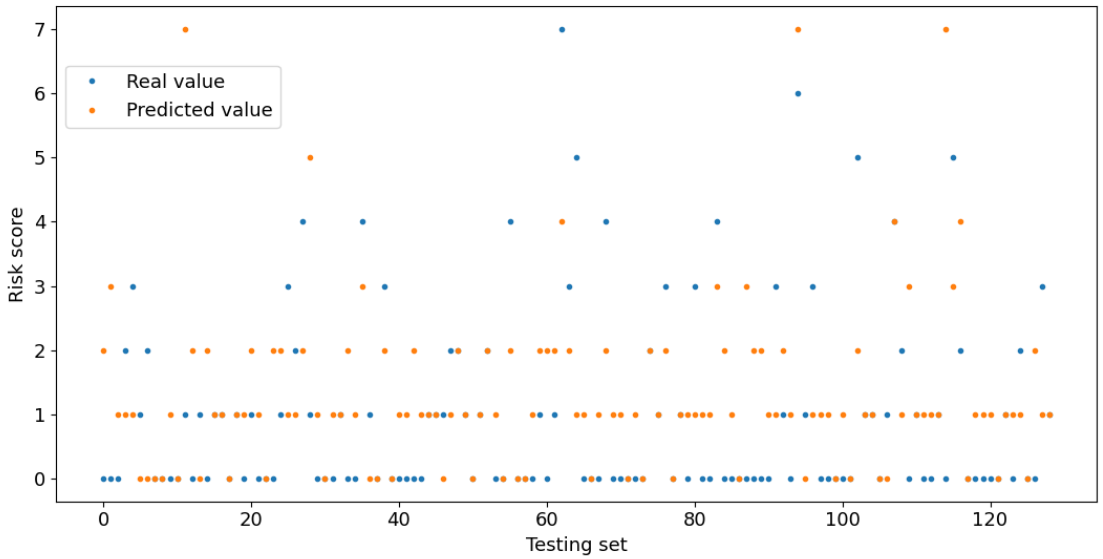

Figure 6.8: Predicted and real values for the linear regression model with lower MAE.

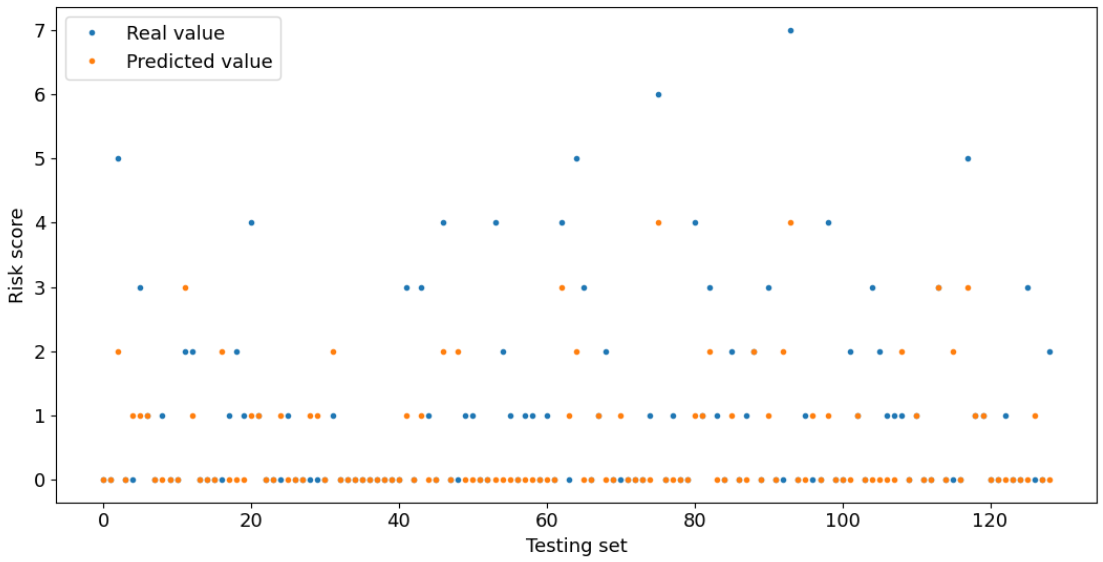

Figure 6.9: Predicted and real values for the RF regression model with lower MAE.

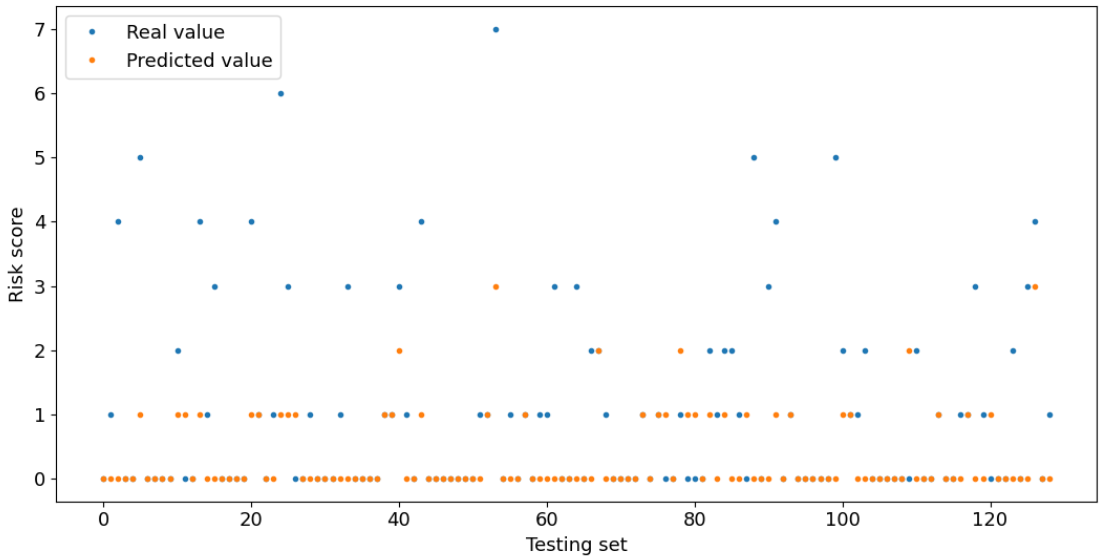

Figure 6.10: Predicted and real values for the support vector regression model with lower MAE.

6.3.4 Discussion

Regarding the classification models and Table 6.15, the logistic regression presents a very small number of selected features (2), as in the EMG model of pain prediction (discussed in Section 6.2.1). This reinforces the idea that the logistic regression algorithm behaves better when using fewer features, as long as they are the most informative, being sufficient for the definition of its sigmoid curve. The SVM model, this time, only considered one feature, which indicates that the addition of more features did not improve its results. The RF model selected a lot of features, as was the case in models discussed above. Therefore, when comparing to other types of algorithms, it can be assumed that the RF algorithm needs more features to define the nodes of its decision trees.

Taking a look at the classifiers' performance metrics (Tables 6.17 to 6.20), the logistic regression model clearly outperforms the other occupational risk classifiers, as evidenced by the mean values of its accuracy (65.7% vs 63.3% for RF vs 62.0% for SVM), precision (69.3% vs 63.4% for RF vs 68.3% for SVM), recall (45.1% vs 44.1% for RF vs 35.9% for SVM), and F_1 score (54.4% vs 51.8% for RF vs 45.3% for SVM).

Comparing the RF and SVM models, the RF presents better mean values of accuracy (63.3% vs 62.0%), recall (44.1% vs 35.9%), and F_1 score (51.8% vs 45.3%), but lower mean precision (63.4% vs 68.3%). This means that the RF model has more correct predictions and greater proportion of actual positives correctly identified, as indicated by accuracy and recall, respectively. On the other hand, the SVM model has greater proportion of correct positive identifications, as indicated by precision. In terms of balance between accuracy and recall, the RF model seems better, as indicated by its F_1 score.

Regarding the regression models, with respect to the number of selected features

(Table 6.21), the support vector regression is the model that uses less features (4), as was the case in the respective classification model (SVM). This makes sense, as the SVM classifier and support vector regressor work in a similar way, despite having distinct outputs.

Analyzing Table 6.24, it can be seen that not all of the assumptions needed to perform linear regression were fulfilled. For the Durbin-Watson test, a test statistic of 2 would indicate no autocorrelation in the residuals. As the values obtained for the test statistic are close to 2, both for the training (2.18) and testing (1.94) sets, it can be concluded that there is no significant autocorrelation. Additionally, regarding the results of the Goldfeld-Quandt test, which show p-values ≥ 0.05, specifically 0.99 and 0.05, we can assume that our linear regression model does not present significant heteroskedasticity. This means that the residuals have constant variance. However, the Jarque-Bera test allows us to conclude that the residuals of the linear regression model do not have a normal distribution, because its p-values are significantly below 0.05 (2.42×10^{-41} and 1.92×10^{-17}). Since one of the assumptions was violated, our linear regression model is not valid, which is why we also tested nonparametric algorithms such as RF regression and support vector regression.

If the regressors' MAEs are compared (Table 6.23), RF regression and support vector regression present better performance than the linear regression model, which did not fulfill all the assumptions, as aforementioned. Comparing the first two, they have very close mean values of the MAE (0.84 vs 0.85). However, RF regression presents a higher maximum MAE (0.95 vs 0.91), while the support vector regression model presents a higher minimum MAE (0.77 vs 0.81).

Overall, the MAE was never able to go below 0.77, which indicates that the regressors' predictions were not very close to their real values. It is relevant to note that a MAE of 0 would indicate that all the predicted values were equal to the real values, while a MAE of 1 would mean that the predicted values were, on average, 1 unit away (below or above) from the real values.

Regarding the regression plots, Figures 6.9 and 6.10 show that the best RF and support vector regressors were not able to regress risk scores above 4, at least for the testing set. This is probably due to the fact that there are not so many data samples for higher risks and thus the models rather focus on the risk scores where there are more values present. On the other hand, Figure 6.8 shows that, although the linear regression model is able to perform regression of higher risks, its predictions are not very accurate. This helps to explain why its residuals do not have a normal distribution, as discussed above.

As can be seen, none of the classification or regression models was able to make risk predictions with a high level of reliability. It is important to keep in mind that the occupational risk was modelled using postural variables, but we tried to predict it with signals and questionnaires not really related to posture. The results obtained for both classification and regression suggest that heart rate, noise, and different questionnaires comprising psychosocial, biomechanical, and environmental aspects do not have a strong relationship with the ergonomic risk modelled by posture. The features of the EMG

signals, although they did not provide high accuracies or low MAEs, seem to have a greater relationship with the risk modelled this way.

Additionally, the risk model used, which is based on the public administration sample available at this stage, may not be representative enough. It can be that the 40 subjects used to model the risk are not sufficient, or that they are not completely representative of all public administration workers and certainly all office workers. This is why a more diverse dataset (in terms of age and different working populations) would probably improve the risk model, and its respetive prediction models.

Furthermore, a deep learning approach (e.g., Recurrent Neural Networks (RNN)) could be useful to find patterns that traditional ML algorithms are not able to find, which could improve the obtained results.

7

Conclusion

This chapter highlights the main conclusions drawn from the work carried out. Additionally, it presents improvements that can be implemented in future work, in order to enhance the obtained results.

7.1 Main conclusions

In a world where people spend an important part of the day at work, occupational diseases have become a growing concern. Office workers, in particular, are prone to Musculoskeletal Disorders (MSD), which include a wide range of conditions. The PrevOccupAI project was born to try to answer this problem by educating office workers to reduce the risks of occupational diseases, through biosignals (including Electromyography (EMG), accelerometer, gyroscope, and magnetometer) and questionnaires (including pain, psychosocial, biomechanical, and environmental questionnaires).

The main objective of this master thesis was to model a relative occupational risk for a public administration population which consisted of 40 subjects. Then, Machine Learning (ML) classification and regression models were used to predict the defined risks.

To help achieve the main objective, a variety of additional steps were carried out. Firstly, data acquisition was performed, both within the main scope of the PrevOccupAI project and in a parallel way. The latter was performed due to the need to develop an algorithm to verify if someone is seated or not, which was a necessary step to the development of the final model. As the acquired data came from different devices and sensors, which include smartphone, smartwatch and muscleBANs, resampling and synchronization processes were performed.

Since it was intended to analyze data from office workers sitting at the desk, the acquisition periods in which the subjects were not seated were removed from the signals. For this purpose, ML and non-ML approaches were tested. The developed ML model, which uses the smartphone's accelerometer to distinguish between sitting and walking, has reached an outstanding accuracy of 100%.

Additionally, a model was developed to try to predict the pain reported by the participants twice per day of acquisition. Using different portions of the EMG signals and personal information gathered from the questionnaires, it was possible to reach a mean accuracy of 86.3% for the Random Forest (RF) model. This model may reduce the need for the participants to fill out the daily pain questionnaire.

Through several variables considered to characterize postural variability, to which Weibull distribution curves were associated, the relative risks were defined for each of the acquisitions of the 40-subject population. Then, through classification models, it was possible to reach an average precision of 65.7% for risk prediction. On the other hand, regarding the regression models, it was possible to achieve a Mean Absolute Error (MAE) with a mean value of 0.84.

The developed work and obtained results allowed important advances in the PrevOccupAI project, including the improvement of the risk reports provided to its participants. It is believed that this project can improve the perception of occupational risks by office workers and become a reference tool for the prevention of occupational diseases.

7.2 Future work

This work, despite having contributed with significant advances to the PrevOccupAI project, leaves some issues to be worked on, as well as some points to be improved.

The most important thing to be performed is the acquisition of more real-world data in large scale. This will lead to the improvement of the developed models, especially the pain prediction and occupational risk models. It is crucial to have a dataset as large as possible, but it is also important to take into account its diversity. The developed models would definitely benefit from a more diverse dataset, including different age groups and working populations.

The modelling of the occupational risk, specifically, would benefit a lot from the acquisition of more data, because this risk was defined relatively to the acquisition population. Additionally, the use of different populations, corresponding to subjects who do not work in the public administration, would potentially make the developed risk model applicable to all office workers. Moreover, risk modelling could also be based on more objective guidelines, which are not currently available.

In order to improve the developed models, deep learning techniques can also be used. The risk prediction model, in particular, could take advantage of using neural networks. Long Short-Term Memory (LSTM) networks, for example, which are a type of Recurrent Neural Networks (RNN) with the ability to learn long-term dependencies [87], appear to be indicated to potentially improve the classification accuracy of this model.

Bibliography

[1] J. M. Lourenço, *The NOVAthesis LaTeX Template User's Manual*, NOVA University Lisbon, 2021. [Online]. Available: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf.

[2] World Health Organization, *Protecting workers' health*, 2017. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/protecting-workers'-health.

[3] J. S. Boschman, T. Brand, M. H. Frings-Dresen, and H. F. van der Molen, "Improving the assessment of occupational diseases by occupational physicians", *Occupational Medicine*, vol. 67, pp. 13–19, 1 2017. DOI: 10.1093/OCCMED/KQW149.

[4] L. van der Zwaan, C. Oldenburg, M. van Emmerik, *et al.*, "Estimating the costs of work-related accidents and ill-health: An analysis of european data sources", 2017. DOI: 10.2802/566789.

[5] H. Perista, A. Cardoso, J. Nunes, P. Carrilho, and E. Quintal, "Inquérito às condições de trabalho em portugal continental: Trabalhadores/as", Autoridade para as Condições do Trabalho, 2016. [Online]. Available: https://www.act.gov.pt/(pt-PT)/Publicacoes/ProjetosApoiados/2017/Documents/Inqu%C3%A9rito%20%C3%A0s%20Condi%C3%A7%C3%B5es%20de%20Trabalho%20em%20Portugal%20Continental_Trabalhadores.pdf.

[6] E. Tompa, A. Mofidi, S. van den Heuvel, *et al.*, "Economic burden of work injuries and diseases: A framework and application in five european union countries", *BMC Public Health*, vol. 21, p. 49, 2021. DOI: 10.1186/S12889-020-10050-7/TABLES/4.

[7] Biosignals LIBPhys-UNL, *Biosignals - prevoccupai*. [Online]. Available: http://biosi.gnals.org/projects_pages/prevocupai/prevocupai.html.

[8] S. Silva, C. Cepeda, J. Rodrigues, P. Probst, and H. Gamboa, "Assessing occupational health with a cross-platform application based on self-reports and biosignals", Scitepress, 2022, pp. 549–556. DOI: 10.5220/0010846700003123.

[9] M. Farzandipour, E. Nabovati, S. Saeedi, and E. Fakharian, "Fuzzy decision support systems to diagnose musculoskeletal disorders: A systematic literature review", *Computer Methods and Programs in Biomedicine*, vol. 163, pp. 101–109, 2018. DOI: 10.1016/J.CMPB.2018.06.002.

[10] K. Storheim and J.-A. Zwart, "Musculoskeletal disorders and the global burden of disease study", *Annals of the Rheumatic Diseases*, vol. 73, pp. 949–950, 6 2014. DOI: 10.1136/ANNRHEUMDIS-2014-205327.

[11] World Health Organization, *Musculoskeletal conditions*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions.

[12] M. Moradi-Lakeh, M. H. Forouzanfar, S. E. Vollset, *et al.*, "Burden of musculoskeletal disorders in the eastern mediterranean region, 1990–2013: Findings from the global burden of disease study 2013", *Annals of the Rheumatic Diseases*, vol. 76, pp. 1365–1373, 8 2017. DOI: 10.1136/annrheumdis-2016-210146.

[13] D. Coggon, G. Ntani, S. Vargas-Prada, *et al.*, "International variation in absence from work attributed to musculoskeletal illness: Findings from the cupid study", *Occupational and Environmental Medicine*, vol. 70, pp. 575–584, 8 2013. DOI: 10.1136/OEMED-2012-101316.

[14] C. Crean, C. Mcgeough, and R. O'Kennedy, *Wearable biosensors for medical applications*, S. Higson, Ed., 2012. DOI: 10.1533/9780857097187.2.301.

[15] D. K. Shaeffer, "Mems inertial sensors: A tutorial overview", *IEEE Communications Magazine*, vol. 51, pp. 100–109, 4 2013. DOI: 10.1109/MCOM.2013.6495768.

[16] T. Tamura, *Wearable inertial sensors and their applications*, E. Sazonov and M. Neuman, Eds., 2014. DOI: 10.1016/B978-0-12-418662-0.00024-6.

[17] M. Kok, J. D. Hol, and T. B. Schön, "Using inertial sensors for position and orientation estimation", *Foundations and Trends in Signal Processing*, vol. 11, pp. 1–153, 1-2 2017. DOI: 10.1561/2000000094.

[18] E. M. Diaz, D. B. Ahmed, and S. Kaiser, *A review of indoor localization methods based on inertial sensors*, J. Conesa, A. Pérez-Navarro, J. Torres-Sospedra, and R. Montoliu, Eds., 2019. DOI: 10.1016/B978-0-12-813189-3.00016-2.

[19] M. A. C. Garcia and T. M. M. Vieira, "Surface electromyography: Why, when and how to use it", *Revista Andaluza de Medicina del Deporte*, vol. 4, pp. 17–28, 1 2011. [Online]. Available: https://www.elsevier.es/es-revista-revista-andaluza-medicina-del-deporte-284-articulo-surface-electromyography-why-when-how-X1888754611201253.

[20] E. Criswell and J. R. Cram, *Cram's introduction to surface electromyography*, 2nd edition. Jones and Bartlett, 2011.

[21] J. Celichowski and P. Krutki, *Motor units and muscle receptors*, 2019. DOI: 10.1016/B978-0-12-814593-7.00004-9.

[22] I. E. Naqa and M. J. Murphy, *What is machine learning?*, I. E. Naqa, R. Li, and M. J. Murphy, Eds., 2015. DOI: 10.1007/978-3-319-18305-3_1.

[23] T. M. Mitchell, *Machine Learning*, 1st edition. McGraw-Hill, 1997.

[24] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd edition, R. Roumeliotis and N. Tache, Eds. O'Reilly Media, 2019.

[25] N. Japkowicz and M. Shah, *Performance evaluation in machine learning*, I. E. Naqa, R. Li, and M. J. Murphy, Eds., 2015. DOI: 10.1007/978-3-319-18305-3_4.

[26] S. Suthaharan, *Decision tree learning*, 2016. DOI: 10.1007/978-1-4899-7641-3_10.

[27] D. Che, Q. Liu, K. Rasheed, and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics", *Advances in Experimental Medicine and Biology*, vol. 696, pp. 191–199, 2011. DOI: 10.1007/978-1-4419-7046-6_19/COVER.

[28] A. Cutler, D. R. Cutler, and J. R. Stevens, *Random forests*, 2012. DOI: 10.1007/978-1-4419-9326-7_5.

[29] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning", *Journal of Applied Science and Technology Trends*, vol. 1, pp. 140–147, 4 2020. DOI: 10.38094/jastt1457.

[30] C. R. Ahn, S. Lee, C. Sun, H. Jebelli, K. Yang, and B. Choi, "Wearable sensing technology applications in construction safety and health", *Journal of Construction Engineering and Management*, vol. 145, p. 03 119 007, 11 2019. DOI: 10.1061/(ASCE)CO.1943-7862.0001708.

[31] T. Vilarinho, B. Farshchian, D. G. Bajer, *et al.*, "A combined smartphone and smartwatch fall detection system", IEEE, 2015, pp. 1443–1448. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.216.

[32] S. Ayub, B. M. Heravi, A. Bahraminasab, and B. Honary, "Pedestrian direction of movement determination using smartphone", IEEE, 2012, pp. 64–69. DOI: 10.1109/NGMAST.2012.36.

[33] M.-S. Pan and H.-W. Lin, "A step counting algorithm for smartphone users: Design and implementation", *IEEE Sensors Journal*, vol. 15, pp. 2296–2305, 4 2015. DOI: 10.1109/JSEN.2014.2377193.

[34] T. R. Mauldin, M. E. Canby, V. Metsis, A. H. H. Ngu, and C. C. Rivera, "Smartfall: A smartwatch-based fall detection system using deep learning", *Sensors*, vol. 18, p. 3363, 10 2018. DOI: 10.3390/S18103363.

[35] E. Bulbul, A. Cetin, and I. A. Dogru, "Human activity recognition using smartphones", IEEE, 2018. DOI: 10.1109/ISMSIT.2018.8567275.

[36] Y. Kwon, K. Kang, and C. Bae, "Unsupervised learning for human activity recognition using smartphone sensors", *Expert Systems with Applications*, vol. 41, pp. 6067–6074, 14 2014. DOI: 10.1016/J.ESWA.2014.04.037.

[37] C. Cipriani, F. Zaccone, S. Micera, and M. C. Carrozza, "On the shared control of an emg-controlled prosthetic hand: Analysis of user-prosthesis interaction", *IEEE Transactions on Robotics*, vol. 24, pp. 170–184, 1 2008. DOI: 10.1109/TRO.2007.910708.

[38] P. Marshall and B. Murphy, "The validity and reliability of surface emg to assess the neuromuscular response of the abdominal muscles to rapid limb movement", *Journal of Electromyography and Kinesiology*, vol. 13, pp. 477–489, 5 2003. DOI: 10.1016/S1050-6411(03)00027-0.

[39] E. F. Shair, S. A. Ahmad, M. H. Marhaban, S. B. M. Tamrin, and A. R. Abdullah, "Emg processing based measures of fatigue assessment during manual lifting", *BioMed Research International*, vol. 2017, 2017. DOI: 10.1155/2017/3937254.

[40] I. M. Khairuddin, S. N. Sidek, A. P. P. A. Majeed, M. A. M. Razman, A. A. Puzi, and H. M. Yusof, "The classification of movement intention through machine learning models: The identification of significant time-domain emg features", *PeerJ Computer Science*, vol. 7, e379 2021. DOI: 10.7717/PEERJ-CS.379/SUPP-2.

[41] J. Coker, H. Chen, M. C. Schall, S. Gallagher, and M. Zabala, "Emg and joint angle-based machine learning to predict future joint angles at the knee", *Sensors*, vol. 21, p. 3622, 11 2021. DOI: 10.3390/S21113622.

[42] C. Miozzi, V. Errico, G. Saggio, E. Gruppioni, and G. Marrocco, "Uhf rfid-based emg for prosthetic control: Preliminary results", IEEE, 2019, pp. 310–313. DOI: 10.1109/RFID-TA.2019.8891964.

[43] G. Filios, S. Nikoletseas, and C. Pavlopoulou, "Efficient parameterized methods for physical activity detection using only smartphone sensors", Association for Computing Machinery, 2015, pp. 97–104. DOI: 10.1145/2810362.2810372.

[44] N. A. Capela, E. D. Lemaire, N. Baddour, M. Rudolf, N. Goljar, and H. Burger, "Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants", *Journal of NeuroEngineering and Rehabilitation*, vol. 13, 5 2016. DOI: 10.1186/S12984-016-0114-0/TABLES/5.

[45] R. San-Segundo, H. Blunck, J. Moreno-Pimentel, A. Stisen, and M. Gil-Martín, "Robust human activity recognition using smartwatches and smartphones", *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 190–202, 2018. DOI: 10.1016/J.ENGAPPAI.2018.04.002.

[46] S. Mekruksavanich, N. Hnoohom, and A. Jitpattanakul, "Smartwatch-based sitting detection with human activity recognition for office workers syndrome", IEEE, 2018, pp. 160–164. DOI: 10.1109/ECTI-NCON.2018.8378302.

[47] C.-W. Liao and Y.-H. Perng, "Data mining for occupational injuries in the taiwan construction industry", *Safety Science*, vol. 46, pp. 1091–1102, 7 2008. DOI: 10.1016/J.SSCI.2007.04.007.

[48] F. Frigerio and V. Virga, "Noise metering via mobile phones: Limitations, opportunities and findings in a workplace testing", IEEE, 2019. DOI: 10.1109/EEEIC.2019.8783379.

[49] N. D. Nath, R. Akhavian, and A. H. Behzadan, "Ergonomic analysis of construction worker's body postures using wearable mobile sensors", *Applied Ergonomics*, vol. 62, pp. 107–117, 2017. DOI: 10.1016/J.APERGO.2017.02.007.

[50] F. Caputo, E. D'Amato, A. Greco, I. Notaro, and S. Spada, "Human posture tracking system for industrial process design and assessment", vol. 722, Springer, 2018, pp. 450–455. DOI: 10.1007/978-3-319-73888-8_70.

[51] S. M. Cerqueira, A. F. Silva, and C. P. Santos, "Smart vest for real-time postural biofeedback and ergonomic risk assessment", *IEEE Access*, vol. 8, pp. 107 583–107 592, 2020. DOI: 10.1109/ACCESS.2020.3000673.

[52] F. Carnide, A. Veloso, H. Gamboa, S. Caldeira, and I. Fragoso, "Interaction of biomechanical and morphological factors on shoulder workload in industrial paint work", *Clinical Biomechanics*, vol. 21, S33–S38, Suppl. 1 2006. DOI: 10.1016/J.CLINBIOMECH.2005.09.010.

[53] S. E. Lenzi, C. E. Standoli, G. Andreoni, P. Perego, and N. F. Lopomo, "Comparison among standard method, dedicated toolbox and kinematic-based approach in assessing risk of developing upper limb musculoskeletal disorders", vol. 795, Springer, 2018, pp. 135–145. DOI: 10.1007/978-3-319-94619-1_13.

[54] S. Sarkar, V. Lodhi, and J. Maiti, "Text-clustering based deep neural network for prediction of occupational accident risk: A case study", IEEE, 2018. DOI: 10.1109/ISAI-NLP.2018.8692881.

[55] S. Dalal and D. Bassu, "Deep analytics for workplace risk and disaster management", *IBM Journal of Research and Development*, vol. 64, 14:1–14:9, 1/2 2020. DOI: 10.1147/JRD.2019.2945693.

[56] Y. Liu, S. Zhao, X. Yue, B. Muthu, and R. L. Kumar, "Ai-based framework for risk estimation in workplace", *Aggression and Violent Behavior*, p. 101 616, 2021. DOI: 10.1016/J.AVB.2021.101616.

[57] H. Liu, Z. Tang, Y. Yang, *et al.*, "Identification and classification of high risk groups for coal workers' pneumoconiosis using an artificial neural network based on occupational histories: A retrospective cohort study", *BMC Public Health*, vol. 9, 366 2009. DOI: 10.1186/1471-2458-9-366/TABLES/3.

[58] C. Yuan, G. Li, Z. Peihong, and C. Li, "Artificial neural network modeling of prevalence of pneumoconiosis among workers in metallurgical industry - a case study", vol. 1, IEEE, 2011, pp. 388–393. DOI: 10.1109/ISDEA.2010.111.

[59] Z. Huang, D. Yu, and J. Zhao, "Application of neural networks with linear and nonlinear weights in occupational disease incidence forecast", IEEE, 2000, pp. 383–386. DOI: 10.1109/APCCAS.2000.913515.

[60] A. di Noia, P. Montanari, and A. Rizzi, "Occupational diseases risk prediction by genetic optimization: Towards a non-exclusive classification approach", *Studies in Computational Intelligence*, vol. 620, pp. 63–77, 2016. DOI: 10.1007/978-3-319-26393-9_5.

[61] A. di Noia, A. Martino, P. Montanari, and A. Rizzi, "Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction", *Soft Computing*, vol. 24, pp. 4393–4406, 6 2020. DOI: 10.1007/S00500-019-04200-2.

[62] C.-W. Cheng, S.-S. Leu, Y.-M. Cheng, T.-C. Wu, and C.-C. Lin, "Applying data mining techniques to explore factors contributing to occupational injuries in taiwan's construction industry", *Accident Analysis and Prevention*, vol. 48, pp. 214–222, 2012. DOI: 10.1016/J.AAP.2011.04.014.

[63] C.-W. Cheng, H.-Q. Yao, and T.-C. Wu, "Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry", *Journal of Loss Prevention in the Process Industries*, vol. 26, pp. 1269–1278, 6 2013. DOI: 10.1016/J.JLP.2013.07.002.

[64] A. Martiniano, R. P. Ferreira, R. J. Sassi, and C. Affonso, "Application of a neuro fuzzy network in prediction of absenteeism at work", IEEE, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6263151.

[65] F. E. Ciarapica and G. Giacchetta, "Classification and prediction of occupational injury risk using soft computing techniques: An italian study", *Safety Science*, vol. 47, pp. 36–49, 1 2009. DOI: 10.1016/J.SSCI.2008.01.006.

[66] P. Bota, J. Silva, D. Folgado, and H. Gamboa, "A semi-automatic annotation approach for human activity recognition", *Sensors*, vol. 19, p. 501, 3 2019. DOI: 10.3390/S19030501.

[67] S. M. Sánchez and J. M. C. Rodriguez, "Smart protective protection equipment for an accessible work environment and occupational hazard prevention", IEEE, 2020, pp. 581–585. DOI: 10.1109/CONFLUENCE47617.2020.9058188.

[68] N. D. Nath, T. Chaspari, and A. H. Behzadan, "Automated ergonomic risk monitoring using body-mounted sensors and machine learning", *Advanced Engineering Informatics*, vol. 38, pp. 514–526, 2018. DOI: 10.1016/J.AEI.2018.08.020.

[69] I. Conforti, I. Mileti, Z. D. Prete, and E. Palermo, "Measuring biomechanical risk in lifting load tasks through wearable system and machine-learning approach", *Sensors*, vol. 20, p. 1557, 6 2020. DOI: 10.3390/S20061557.

[70] J. Manjarres, P. Narvaez, K. Gasser, W. Percybrooks, and M. Pardo, "Physical workload tracking using human activity recognition with wearable devices", *Sensors*, vol. 20, p. 39, 1 2020. DOI: 10.3390/S20010039.

[71] S. Koçer, "Classification of emg signals using neuro-fuzzy system and diagnosis of neuromuscular diseases", *Journal of Medical Systems*, vol. 34, pp. 321–329, 3 2010. DOI: 10.1007/S10916-008-9244-7/TABLES/7.

[72] A. Subasi, "Medical decision support system for diagnosis of neuromuscular disorders using dwt and fuzzy support vector machines", *Computers in Biology and Medicine*, vol. 42, pp. 806–815, 8 2012. DOI: 10.1016/J.COMPBIOMED.2012.06.004.

[73] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with numpy", *Nature*, vol. 585, pp. 357–362, 7825 2020. DOI: 10.1038/s41586-020-2649-2.

[74] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python", *Nature Methods*, vol. 17, pp. 261–272, 3 2020. DOI: 10.1038/s41592-019-0686-2.

[75] W. McKinney, "Data structures for statistical computing in python", SciPy, 2010, pp. 56–61. DOI: 10.25080/MAJORA-92BF1922-00A.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[77] M. Barandas, D. Folgado, L. Fernandes, *et al.*, "Tsfel: Time series feature extraction library", *SoftwareX*, vol. 11, p. 100 456, 2020. DOI: 10.1016/J.SOFTX.2020.100456.

[78] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning", 2021, pp. 142–151. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Fei_Z-Score_Normalization_Hubness_and_Few-Shot_Learning_ICCV_2021_paper.html.

[79] Fraunhofer AICOS, *List of available features - tsfel 0.1.4 documentation*. [Online]. Available: https://tsfel.readthedocs.io/en/latest/descriptions/feature_list.html.

[80] I. Syarif, A. Prugel-Bennett, and G. Wills, "Svm parameter optimization using grid search and genetic algorithm to improve classification performance", *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, pp. 1502–1509, 4 2016. DOI: 10.12928/TELKOMNIKA.V14I4.3956.

[81] P. M. Tomko, R. J. Colquhoun, M. A. Magrini, T. W. D. Muddle, and N. D. M. Jenkins, "Global electromyographic signal characteristics depend on maximal isometric contraction method in the knee extensors", *Journal of Electromyography and Kinesiology*, vol. 42, pp. 111–116, 2018. DOI: 10.1016/J.JELEKIN.2018.07.002.

[82] R. Goldman, "Understanding quaternions", *Graphical Models*, vol. 73, pp. 21–49, 2 2011. DOI: 10.1016/J.GMOD.2010.10.004.

[83] PLUX, *Introduction to android sensors*. [Online]. Available: http://notebooks.pluxbiosignals.com/notebooks/Categories/Other/intro_to_android_sensors_rev.html.

[84] M. V. Paterno, L. C. Schmitt, K. R. Ford, M. J. Rauh, and T. E. Hewett, "Altered postural sway persists after anterior cruciate ligament reconstruction and return to sport", *Gait and Posture*, vol. 38, pp. 136–140, 1 2013. DOI: 10.1016/J.GAITPOST.2012.11.001.

[85] R. W. Soames and J. Atha, "The spectral characteristics of postural sway behaviour", *European Journal of Applied Physiology and Occupational Physiology*, vol. 49, pp. 169–177, 1982. DOI: 10.1007/BF02334065.

[86] J. Miles, "Tolerance and variance inflation factor", *Wiley StatsRef: Statistics Reference Online*, 2014. DOI: 10.1002/9781118445112.STAT06593.

[87] S. Chandar, C. Sankar, E. Vorontsov, S. E. Kahou, and Y. Bengio, "Towards non-saturating recurrent units for modelling long-term dependencies", vol. 33, 2019, pp. 3280–3287. DOI: 10.1609/AAAI.V33I01.33013280.