U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Machine learning applied to fall prediction and detection using wearable sensors

Joana Raquel Cerqueira da Silva

Supervisor: Doutor Jaime dos Santos Cardoso

Co-Supervisor: Doutora Inês Nunes de Sousa Soares

Programa Doutoral em Engenharia Biomédica

November, 2020

Faculdade de Engenharia da Universidade do Porto

Machine learning applied to fall prediction and detection using wearable sensors

Joana Raquel Cerqueira da Silva

Dissertation submitted to Faculdade de Engenharia da Universidade do Porto
to obtain the degree of

Doctor Philosophiae in Biomedical Engineering

President: Doutor Fernando Jorge Mendes Monteiro
Referee: Doutor Lorenzo Chiari
Referee: Doutor Hugo Filipe Silveira Gamboa
Referee: Doutor Miguel Fernando Paiva Velhote Correia

November, 2020

Abstract

The worldwide population is aging rapidly. The process of aging affects the ability of a person to maintain balance, mobility, and muscle strength and to react properly to unexpected situations such as slipping or stumbling. There is currently no standardized protocol to assess the level of fall risk, neither in clinics and hospitals nor in daycare centers. Preventive strategies should be planned and implemented in order to decrease the prevalence of falls among the elderly. These strategies should be based on a multifactorial fall risk assessment that can be implemented with objective measures. The occurrence of falls is not always predictable and early assistance after a fall could decrease its negative effects. Currently, there is a shortage of automated procedures to rapidly deploy fall detection models adapted for different use-cases. Likewise, there is also the need for common datasets and methodologies to benchmark those models.

This thesis focused on the study of a multifactorial fall prediction system and the study of a wearable-based automatic fall detection system. The development of a multifactorial fall prediction system blended the study of feature extraction methods based on the instrumentation of fall risk assessment tests and the study of data fusion procedures to combine data sources, such as clinical, self-reported and sensor-retrieved data. For the advancement of automatic fall detection systems, we considered the impact of several variables in the system's performance: the type of dataset, composed of simulated or real-world data, the on-body positions to couple the wearable device, and restrictions related with the deployment hardware, such as sampling rate, algorithm's sensitivity level, and models complexity.

We contributed with different data fusion approaches for fall prediction based on the analysis of multimodal data collected according to a multifactorial screening protocol. The instrumentation of fall risk assessment tests with inertial sensors and pressure platform allowed to better discriminate the individuals with higher risk of falling. We also proposed a wearable solution for automatic fall detection, based on a low-power state machine algorithm, that can be adapted to different fall risk levels. We studied the impact of type of dataset, learning models, on-body positions and sampling rate in fall detection performance. A new machine learning pipeline was able to generalize to a new unseen position considering a user-independent validation and a lower sampling rate.

In the future, the work presented in the area of fall prediction could be used as a standard multifactorial fall prediction tool based on inertial and pressure devices, to provide a protocol to assess elderly fall risk in the community. The added value of features extracted from sensors could enhance the healthcare professional assessment of physical conditions such as balance, mobility and strength abilities, as well as personal and contextual information. The automation of fall detection systems will allow in the future to expedite the deployment of such systems and to accelerate the time to prototype after selecting the most suitable model's requirements.

Keywords: Aging. Fall Prediction. Fall Detection. Wearable Devices. Inertial Sensors. Pressure Platform. Signal Processing. Machine Learning. Statistics. Multifactorial Data Fusion.

Resumo

A população mundial está a envelhecer rapidamente. O processo de envelhecimento afeta a capacidade de uma pessoa manter o equilíbrio, a mobilidade e a força muscular e reagir adequadamente a situações inesperadas, como escorregões ou tropeções. Atualmente, não existe um protocolo padrão para avaliar o nível de risco de queda, nem em clínicas e hospitais nem em centros de dia. Estratégias preventivas devem ser planeadas e implementadas para diminuir a prevalência de quedas entre os idosos. Essas estratégias devem basear-se numa avaliação de risco de queda multifatorial que pode ser implementada com métricas objetivas.

A ocorrência de quedas nem sempre é previsível e a assistência rápida após uma queda pode diminuir os seus efeitos negativos. Atualmente, há uma escassez de procedimentos automatizados para implementar rapidamente modelos de deteção de queda adaptados a diferentes casos de uso. Da mesma forma, também há a necessidade de conjuntos de dados e metodologias comuns para comparar esses modelos.

Esta tese focou-se no estudo de um sistema multifatorial de previsão de quedas e no estudo de um sistema automático de deteção de quedas baseado em sensores. O desenvolvimento de um sistema de previsão de queda multifatorial combinou o estudo de métodos de extração de métricas com base na instrumentação de testes de avaliação de risco de queda e no estudo de procedimentos de fusão de dados para combinar fontes de dados, como dados clínicos, dados reportados e dados de sensores. Para o avanço dos sistemas automáticos de deteção de quedas, consideramos o impacto de várias variáveis no desempenho do sistema: o tipo de conjunto de dados, composto por dados simulados ou reais, as posições corporais para acoplar o sensor e restrições relacionadas com o hardware, como taxa de amostragem, o nível de sensibilidade do algoritmo e a complexidade dos modelos.

Contribuímos com diferentes abordagens de fusão de dados para previsão de quedas com base na análise de dados multimodais recolhidos de acordo com um protocolo de avaliação multifatorial. A instrumentação dos testes de avaliação de risco de queda com sensores inerciais e uma plataforma de pressão permitiu discriminar melhor os indivíduos com maior risco de queda. Também propusemos uma solução para deteção automática de quedas, com base num algoritmo de máquina de estados de baixo custo computacional, que pode ser adaptado a diferentes níveis de risco de queda. Estudamos o impacto do tipo de conjunto de dados, modelos de aprendizagem, posições no corpo e taxa de amostragem no desempenho da deteção de quedas. Um novo procedimento de aprendizagem computacional conseguiu generalizar para uma nova posição, considerando uma validação independente do utilizador e uma menor taxa de amostragem.

No futuro, o trabalho apresentado na área de previsão de quedas poderá ser usado como uma ferramenta padrão de previsão multifatorial de quedas baseada em dispositivos inerciais e de pressão, para fornecer um protocolo para avaliar o risco de queda em idosos na comunidade. O valor adicionado pelas métricas extraídas dos sensores pode melhorar a avaliação do profissional de saúde sobre condições físicas, como equilíbrio, mobilidade e força, além de informações pes-

soais e contextuais. A automação dos sistemas de deteção de queda permitirá, no futuro, acelerar a implementação de tais sistemas e acelerar o tempo de prototipagem após a seleção dos requisitos do modelo mais adequados.

Keywords: Envelhecimento. Previsão de queda. Deteção de Quedas. Dispositivos Vestíveis. Sensores Inerciais. Plataforma de Pressão. Processamento de Sinal. Aprendizagem Computacional. Estatística. Fusão de dados multifatoriais.

Acknowledgments

When I first joined Fraunhofer, seven years ago, I was passionate about this new filed, for me, of Machine Learning. At the time I didn't realize the importance that the field would have in my life. Allied to this field, there was also the deep study of inertial sensor applications in companion solutions for the elderly, which was the core of one of the main competence centers in this research center. All over my journey since then, I was able to deepen my knowledge in both fields, and for that, I am thankful for the lessons learned and knowledge acquired. This thesis is a proof of that accomplishment. It was a journey full of personal and professional growth, small big accomplishments, and among all, friendly and reliable colleagues, with whom I have followed my path until here.

I sincerely appreciate the vote of confidence and encouragement from my supervisors, Professor Jaime Cardoso and Dr. Inês Sousa. Professor Jaime was always passionate about learning new fields of research, and how to transfer knowledge among them, I have learned much with his experience, and I am very grateful and honored for all the encouragement along these years. Inês encouraged me to undertake the PhD since day one, and I can say we have both made a long journey until here. I consider you a friend with whom I have also learned a lot, and to whom I also thank for always letting me grow professionally and for supporting me in this journey.

I thank everyone with whom I have worked all these years, at Fraunhofer, and from whom I have also learned. I would like to thank my first colleagues from the Fall Competence Center: Bruno Aguiar, Tiago Rocha, Susana Carneiro, Vânia Guimarães, and Filipe Sousa, and my colleagues from the FallSensing team: João Madureira, Dinis Moreira, António Santos, Eduardo Pereira, Elsa Oliveira, Nuno Cardoso, and Diego Martins. For all the important discussions and accomplishments, I am thankful to João Machado, Carina Figueira, Duarte Folgado, Diana Gomes, José Alves, Francisco Nunes, Luís Rosado, David Ribeiro, João Gonçalves, Filipe Soares, Hugo Gamboa, Rui Castro, and Liliana Ferreira. During this journey, I also had the opportunity to meet new colleagues from the VCMI group at INESC-TEC, to whom I thank all the interesting discussions: Kelwin Fernandes, João Pinto, Diogo Pernes, and Ricardo Cruz.

I would like to express my acknowledgments to all the participants that enrolled in the data collection trials and also to the organizers of the publicly available datasets FARSEEING and UMAFall Dataset, that had an important role in my work. Moreover, I also would like to thank the colleagues that supervised the data collection within the FallSensing project: Professor Anabela Martins, Daniela Baltazar, Catarina Silva and Juliana Moreira from ESTEsC, Cláudia Tonelo, André Dias and Luís Ferreira from Sensing Future Technologies, Nuno Tavares from PhysioMondego, and Sílvia Rego from Colaborar Network.

I would like to thank Fraunhofer Portugal AICOS for supporting my Doctoral Program, and I also thank the financial support from National and European projects in the scope of which this thesis was conducted: *FallSensing* - Technological solution for fall risk screening and falls prevention, *Deus ex Machina* - Symbiotic technology for societal efficiency gains, and *IANVS* - Indoor Activity Notification for Vigilance Services.

Finally, I would like to thank all my family and friends for the support and encouragement they gave me. I honestly thank my grandmothers, Ondina and Helena, for helping me grow, for being strong and always care for me. I especially thank my mother Fátima and João for all the love, joy and guidance during these years. I also thank D. Darcília, Sr. Américo and Joana for always being kind and helping me when I needed it. My sincere words of gratitude to my best friend and love, Rui, for all the patience, help and guidance in the good and in the challenging moments.

Joana Raquel Silva

Publications

The work developed in this thesis was has been published in journal articles and conference papers, as listed below:

Conference papers:

P.1. J. Silva and I. Sousa, "Inertial Sensors-based Instrumented Timed Up and Go Tool for Fall Risk Assessment", in *Proceedings 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2016, pp. 5.

P.2. A. Martins, J. Silva, A. Santos, J. Madureira, C. Alcobia, L. Ferreira, P. Mendes, C. Tonelo, C. Silva, D. Baltazar, I. Sousa. "Case-Based Study of Metrics Derived from Instrumented Fall Risk Assessment Tests". *10th World Conference of Gerontechnology* 2016

P.3. J. Silva, J. Madureira, C. Tonelo, D. Baltazar, C. Silva, A. Martins, C. Alcobia and I. Sousa, "Comparing Machine Learning Approaches for Fall Risk Assessment", in *Proceedings BIOSIGNALS - 10th International Conference on Bio-inspired Systems and Signal Processing* 2017

P.4. J. Silva, I. Sousa, and J. Cardoso, "Transfer learning approach for fall detection with the FARSEEING real-world dataset and simulated falls," *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3509-3512, 2018.

Journal articles:

P.5. A. Martins, J. Moreira, C. Silva, J. Silva, C. Tonelo, D. Baltazar, C. Rocha, T. Pereira, and I. Sousa, "FallSensing, a multifactorial screening tool for fall-risk in community dwelling adults aged 50 years or over: Study protocol", *JMIR Research Protocols*, vol. 7, pp. 10304, 2018

P.6. J. Alves, J. Silva, E. Grifo, C. Resende, and I. Sousa, “Wearable Embedded Intelligence for Detection of Falls Independently of on-Body Location.” *Sensors MDPI* vol. 19, 11 2426, 2019.

P.7. J. Silva, I. Sousa, and J. Cardoso, "Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 50-56, 2020.

P.8. J. Silva, D. Gomes, I. Sousa, and J. Cardoso, "Automated development of custom fall detectors: position, model and rate impact in performance" *IEEE Sensors Journal*, vol. 20, no. 10, pp. 5465-5472, 2020.

White paper:

P.9. J. Silva, C. Figueira, M. Barandas, J. Gonçalves, J. Costa, L. Rosado, M. Vasconcelos, F. Soares, and H. Gamboa. "White Paper about Machine Learning @ Fraunhofer Portugal AICOS" 2017 [1]

The author of this thesis has conducted other research studies, that are not included in this thesis, but were published during the course of the thesis, in journal and conference papers:

Conference papers:

P.10. S. Carneiro, J. Silva, J. Madureira, D. Moreira, V. Guimarães, A. Santos, and I. Sousa. "Inertial sensors for assessment of joint angles" in *Proceedings of the 4th Workshop on ICTs for improving Patients Rehabilitation Research Techniques (REHAB '16)*, 2016

P.11 J. Silva, D. Moreira, J. Madureira, E. Pereira, A. Dias, I. Sousa. "A Technological Solution for Supporting Fall Prevention Exercises at the Physiotherapy Clinic" *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2018

P.12 J. Silva, E. Oliveira, D. Moreira, F. Nunes, M. Caic, J. Madureira, E. Pereira. "Design and Evaluation of a Fall Prevention Multiplayer Game for Senior Care Centres". *International Conference on Entertainment Computing*, 103-114, 2018

P.13 J. Silva, D. Gomes, F. Nunes, D. Moreira, J. Alves, A. Pereira, I. Sousa. "Position-independent Physical Activity Monitoring: Development and Comparison with Market Devices" *IEEE International Symposium on Medical Measurements and Applications (MeMeA)* 2019

Journal Articles

P.14 P. Bota, J. Silva, D. Folgado, H. Gamboa "A Semi-Automatic Annotation Approach for Human Activity Recognition" *Sensors MDPI* vol.3, 501, 2019

P.15 D. Gomes, J. Mendes-Moreira, I. Sousa, J. Silva. "Eating and Drinking Recognition in Free-Living Conditions for Triggering Smart Reminders". *Sensors MDPI* vol.12, 2803, 2019

[1] https://www.aicos.fraunhofer.pt/en/news_and_events_aicos/news_archive/older_archive/Machine_Learning_at_Fraunhofer_Portugal_AICOS.html

"Valeu a pena? Tudo vale a pena
Se a alma não é pequena.
Quem quer passar além do Bojador
Tem que passar além da dor.
Deus ao mar o perigo e o abismo deu,
Mas nele é que espelhou o céu."

Fernando Pessoa

Contents

List of Figures

List of Tables

List of Abbreviations

5-STS	5 Times Sit-To-Stand
AICOS	Fraunhofer Portugal AICOS
Acc	Accuracy
ADLs	Activities of Daily Living
AST	Alternate-Step Test
A-P	Antero-Posterior
AUC	Area Under the ROC Curve
BBS	Berg Balance Scale
BMI	Body Mass Index
CNN	Convolutional Neural Network
CRISP-DM	Cross Industry Standard Process for Data Mining
CoP	Center of Pressure
CV	Cross-Validation
DT	Decision Tree
F1	F1-score
FES	Tinetti Falls Efficacy Scale
FEUP	Faculdade de Engenharia da Universidade do Porto
FFT	Fast Fourier Transform
G	Geometric mean of sensitivity and specificity
HSDT	Tukey's Honest Significant Difference Test
IMU	Inertial Measurement Unit
iTUG	Instrumented version of TUG
k-NN	k-Nearest Neighbors
LOSO	Leave-One-Subject-Out
LogReg	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
M-L	Medio-lateral
MLP	Multilayer perceptron

NN	Neural network
OR	Odds Ratio
PCA	Principal Component Analysis
PERS	Personal Emergency Response Systems
POMA	Tinetti Performance Oriented Mobility Assessment
PPA	Physiological Profile Assessment
Prec	Precision
ROC	Receiver Operating Characteristic
RF	Random Forest
RMSE	Root Mean Squared Error
Se	Sensitivity
SMOTE	Synthetic Minority Oversampling Technique
Sp	Specificity
SVM	Support Vector Machine
STS	30-second Sit-To-Stand
TUG	Timed-Up and Go test
TOC	Total Operating Characteristic
YI	Youden Index

Part I

Introduction and Theoretical Background

Chapter 1

Introduction

1.1 Motivation

The global population is aging (WHO, 2007) and strategies to improve and increase life expectancy are emerging. The occurrence of falls is most prevalent among 55-year-old individuals, or older. Falls are one of the major causes of hospitalization and loss of independence in this population (Age UK. Stop Falling, 2013). Either at home, hospitals and at daycare centers, falls are considered an important issue, however, there is currently no standardized protocol to assess the fall risk level of an elderly, neither in clinics and hospitals nor in community settings.

Most of the functional tests used in hospitals and clinics to evaluate the fall risk are based on subjective and observational scales, most of them just evaluate a few fall risk factors and do not consider the multifactorial nature of the risk of falling. Even when several tests and questionnaires are applied, the elderly are only evaluated after the occurrence of a fall, when hospitalized. Preventive strategies should be planned and implemented in order to decrease the prevalence of falls among the elderly. These strategies should be based on a multifactorial analysis that should be implemented with objective scales in a daily and pervasive way. Since most of the fall risk factors vary with time, this evaluation could be improved when assessed more frequently in order to develop prevention strategies that are tailored for each person and to their abilities and daily habits (Avin et al., 2015).

The most commonly used standard tests for fall risk assessment are normally applied individually and most of them only evaluate one type of risk factor, such as gait speed or balance. The lack of a complete and multifactorial assessment tool could be overcome with the instrumentation of some standard tests that could help to evaluate multiple components during the application of traditional tests.

Fall risk assessment is essential for establishing adequate strategies for fall prevention that could help to revert or attenuate some of the fall risk factors among elderlies. Although fall prediction systems can contribute to preventing falls, the occurrence of falls is not only dependent on the physical stability of the individuals but also from external perturbations such as obstacles in the surrounding environment (Bruijn et al., 2013) or weather conditions. For this reason, the

occurrence of falls is not always predictable. Therefore, it urges to be able to detect the falls at the moment they occur.

Prompt assistance after a fall could decrease the negative effects of a fall event. Automatic fall detection systems have been developed in the past years and rely mostly on wearable or smartphones with integrated inertial sensors and location capabilities that facilitate the detection and trigger of a fall alert. However, there is still an extensive number of research studies built upon similar methodologies but addressing particular use-cases for fall detection. These requirements frequently motivate algorithm fine-tuning, making the modeling stage a time and effort consuming process. There is a lack of automated procedures to deploy fall detection models faster and adapted for different use-cases, and there is also the need for standard datasets and methodologies to benchmark those models.

1.2 Objectives

The work proposed in this thesis aims to instrument several fall risk assessment tests with wearable sensors in order to objectively assess an individual's risk of fall. Moreover, the thesis aims to implement fall detection approaches that pervasively analyze the activities of the elderly and quality of movements in order to automatically detect a fall, that could have an impact on the elder's the quality of life. Thereby, the main objectives for this thesis are two-fold: i) develop a multifactorial fall risk assessment approach based on multimodal sensor data and ii) develop a low-cost wearable-based system for automatic detection of falls.

The fall prediction and fall detection systems are illustrated in Figure 1.1 and will require two main components: sensing and processing. For sensing personal and contextual information of the user, the system makes use of sensing units, which could be: wearable devices, smartphones, pressure platforms, or dynamometers. The system also uses information from questionnaires about health and personal conditions, questionnaires about fall occurrence in a follow-up period, and self-reported data about social behavior, home hazards, and risks. In a pervasive way, the system sends the user's data to a processing unit that recognizes daily activities, detects falls, and evaluates the risk of falling. After processing the information, the system triggers a fall alarm in case of a fall and provides a prediction of fall based on the occurrence of a fall in a follow-up period of time.

In order to accomplish these objectives, several datasets were collected or requested and further analyzed:

- **FallSensing Dataset** - collected in the scope of the FallSensing project, during several data collection trials that began in May 2015. This dataset includes data from a multifactorial screening tool for fall-risk in community-dwelling adults aged 50 years or over (Martins et al., 2018). The data was collected by the physiotherapists of Coimbra Health School in community settings. The author of this thesis was involved in the definition of the protocol as well as in the data structuring and curation.

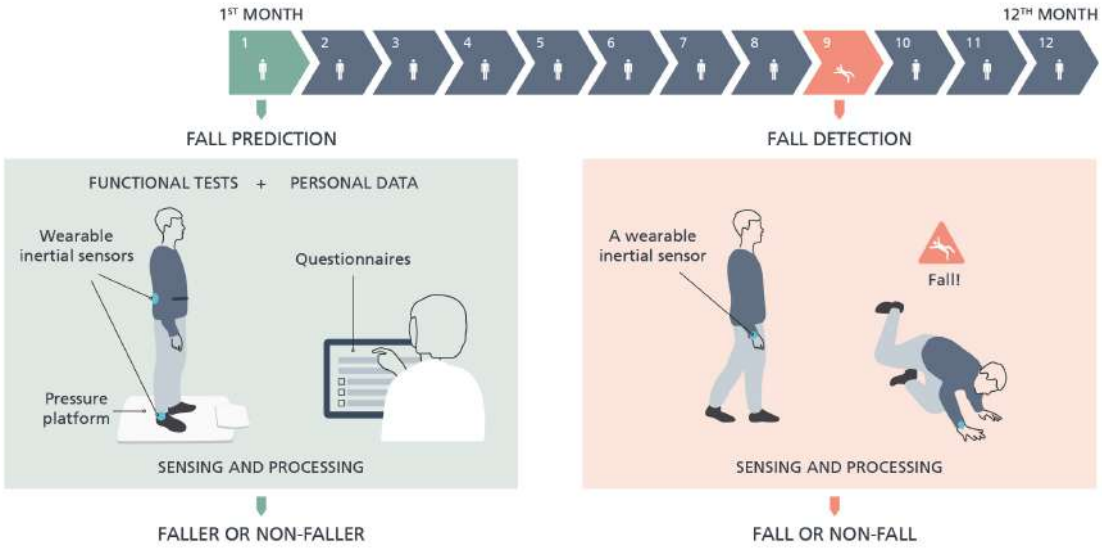

Figure 1.1: Fall prediction and fall detection systems overview.

- **AICOS Fall Dataset** - a large collection of simulated falls and non-falls acquired with young participants since 2009, in the living lab of Fraunhofer Portugal AICOS. The author was involved, since almost the beginning, in the definition of the data collection protocol, conducted most of the trials for data collection and was also involved in the work related to structuring and curation of the dataset.

- **FARSEEING Real-World Dataset** - from FARSEEING project with annotated real fall events for detection of falls and monitoring of daily activities (Klenk et al., 2016). The authors granted us access to a small dataset that includes 23 examples of real falls acquired from elderly patients in hospital settings.

- **UMAFall Dataset** - a publicly available dataset that contains simulated falls and non-falls acquired from young volunteers using wearable devices in several on-body positions (Casilari et al., 2018).

1.3 Contributions

The main contributions of this thesis in the areas of falls prediction and falls detection are detailed as follows:

- Validation that the instrumentation of fall risk assessment tests with inertial sensors and pressure platform could better discriminate the individuals at higher risk of falling. The added value of metrics derived from wearable devices has the potential to improve fall prediction systems (Silva and Sousa, 2016; Silva et al., 2017).

- Exploration of different data fusion approaches for fall prediction based on the analysis of multimodal data collected according to a multifactorial screening protocol. The richness of the collected data allowed to infer not only the functional capabilities of a person but also clinical and environmental information (Silva et al., 2020).

- Proposal of a transfer learning approach for combining a dataset of simulated falls and non-falls with the real-world FARSEEING dataset. The combination of simulated and real-world data allowed us to train a set of supervised classifiers for discriminating between falls and non-fall events (Silva et al., 2018).

- Development of a wearable solution for automatic fall detection, based on a low-power state machine algorithm. Study of different on-body positions and sensors' sampling rate using an optimization algorithm. The algorithm can also be adapted to different groups of people with different fall risk levels, by changing the algorithm's sensitivity (Alves et al., 2019).

- Study of the impact of learning models, on-body positions and sampling rate in fall detection performance, using a new machine learning pipeline that is able to deploy fall detection solutions adapted to the aforementioned system requirements.

1.4 List of Publications

The work developed in the area of *Multifactorial Fall Prediction* resulted in the following publications:

- Joana Silva and Inês Sousa, "Instrumented Timed Up and Go: Fall Risk Assessment based on Inertial Wearable Sensors". *11th IEEE International Symposium on Medical Measurements and Applications (MeMeA)* 2016

- Anabela Martins, Joana Silva, António Santos, João Madureira, Carlos Alcobia, Luís Ferreira, Pedro Mendes, Cláudia Tonelo, Catarina Silva, Daniela Baltazar, Inês Sousa. "Case-Based Study of Metrics Derived from Instrumented Fall Risk Assessment Tests". *10th World Conference of Gerontechnology* 2016

- Joana Silva, João Madureira, Cláudia Tonelo, Daniela Baltazar, Catarina Silva, Anabela Martins, Carlos Alcobia and Inês Sousa. "Comparing Machine Learning Approaches for Fall Risk Assessment". *10th International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)* 2017

- Anabela Martins, Juliana Moreira, Catarina Silva, Joana Silva, Claúdia Tonelo, Daniela Baltazar, Clara Rocha, Telmo Pereira and Inês Sousa, "FallSensing, a multifactorial screening tool for fall-risk in community dwelling adults aged 50 years or over: Study protocol", *JMIR Research Protocols*, vol. 7, pp. 10304, 2018

- Joana Silva, Inês Sousa and Jaime Cardoso, "Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls," *IEEE Journal of Biomedical and Health Informatics (J-BHI)*, vol. 24, no. 1, pp. 50-56, 2020.

The following publications were a result of the work conducted in the area of *Wearable-based Fall Detection*:

- Joana Silva, Inês Sousa, and Jaime Cardoso, "Transfer learning approach for fall detection with the FARSEEING real-world dataset and simulated falls," *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 3509-3512.
- José Alves, Joana Silva, Eduardo Grifo, Carlos Resende and Inês Sousa, "Wearable Embedded Intelligence for Detection of Falls Independently of on-Body Location." *MDPI Sensors* vol. 19, 11 2426. 2019.
- Joana Silva, Diana Gomes, Inês Sousa, and Jaime Cardoso, "Automated development of custom fall detectors: position, model and rate impact in performance" *IEEE Sensors Journal*, vol. 20, no. 10, pp. 5465-5472, 2020.

Only the publications for which the author of this thesis has contributed as the first author will be included in this thesis. With the exception of the publication by Alves et al. (2019), which will be adapted to this thesis, considering only the contributions to the work made by the author of this thesis.

1.5 Document Structure

This thesis is organized in four parts, with a total of 12 chapters, which are organized as follow:

- Part I, **Introduction and Theoretical Background**, includes Chapter 1, where the motivation and principal contributions and publications of the thesis are described, Chapter 2, where the fundamentals of inertial sensors are presented, since inertial sensors will be used as sensing device for developing methods for fall prediction and detection, and 3 explains the machine learning algorithms and data processing pipeline that will be used to process inertial sensor data for developing fall prediction and detection classification algorithms.
- Part II, **Multifactorial Fall Prediction**, is organized in four chapters: Chapter 4 includes an introduction to the topic, along with the related State-of-Art; Chapter 5 focus on the study of feature extraction using a wearable-instrumented functional test for fall prediction; Chapter 6 details a machine learning approach for fall prediction based on a set of functional tests instrumented with wearable devices; and Chapter 7 culminates the topic by combining the previous functional tests and feature extraction methods with data fusion approaches of clinical, self-reported and multisensor data for fall prediction.

- Part III, **Wearable-based Fall Detection**, is also organized in four chapters: Chapter 8 describes the main achievements of previous works, and explains some of the restrictions for developing automatic fall detection models; Chapter 9 describes an approach for combining simulated and real-world falls datasets to improve the performance of fall detection; Chapter 10 details a state-machine algorithm for fall detection, its parameters optimization and a study of the impact of the sampling rate in the performance; and Chapter 11 describes a framework for automating the development of fall detectors that take into account several constraints for model optimization.

- Part IV, **Conclusion**, comprises Chapter 12, that finally summarizes the main conclusions of the two parts of the thesis, along with the main contributions, and future work.

Chapter 2

Fundamentals of Inertial Sensors and Pressure Platform

This chapter describes the fundamentals of inertial sensors and pressure platform. The focus will be made on accelerometers, gyroscopes, and magnetometers, since these sensors are commonly used in applications of falls prediction and detection. An inertial measurement unit (IMU) combines three sensors of each to produce a three-dimensional measurement of the acceleration, angular velocity, and magnetic field, respectively. Most of the solutions commercially available are made with Micro-Electro-Mechanical Systems (MEMS) technology. The widely spread of MEMS sensors is due not only to its low dimension but also to its remarkable performance. There are a large number of microsensors for most of the sensing modalities: temperature, humidity, inertial forces, pressure, radiation, magnetic fields.

Since MEMs are produced by batch fabrication techniques, these small silicon chips can reach high levels of functionality and reliability at a relatively low cost.

> "MEMS technology is extremely diverse and fertile, both in its expected application areas, as well as in how the devices are designed and manufactured. Already, MEMS is revolutionizing many product categories by enabling complete systems-on-a-chip to be realized." (MEMS and Exchange)

2.1 Accelerometer

Accelerometers are currently the leaders of commercial solutions with MEMs technology. There are high, medium and low-grade IMUs with different accuracies and costs. Low-grade accelerometers are mostly used in the automotive industry, in crash airbags and can also be used for pedestrian navigation and attitude and heading reference systems. High-grade IMUs are used for marine applications in submarines and aviation navigation systems for example, because of its high performance, reduced drift, but can have costs of million dollars, while low-grade sensors cost around a dollar (Groves, 2008).

The physical mechanisms of MEMS accelerometers include capacitive, piezoresistive, electromagnetic, piezoelectric, ferroelectric, optical, and tunneling (Bouchaud, 2009).

The most successful type of accelerometer is based on capacitive transduction because of its low power consumption and good stability over temperature. Its function principle is illustrated in Figure 2.1.

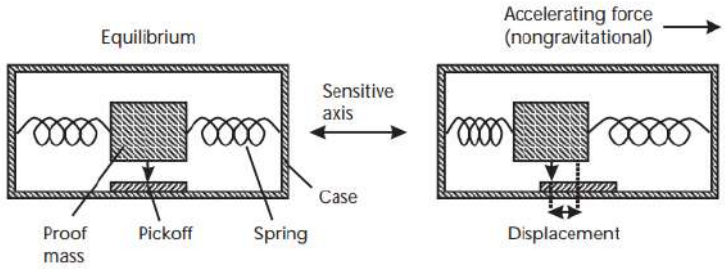

Figure 2.1: Capacitive accelerometer principle. Adapted from Groves (2008).

Accelerometers are inertial sensors because their functionality is based on the principle of inertia, which states that "a body with no net force acting on it will either remain at rest or continue to move with uniform speed in a straight line, according to its initial condition of motion" (Newton's first law). In Figure 2.1 the proof mass is free to move in the accelerometer case and the pickoff measures its position with respect to the case. When a force is applied in the sensitive axis, the case will move with respect to the mass, compressing one spring and stretching the other. The forces from the springs alter the force transmitted to the mass. When the acceleration of the mass reaches the external force, the relative position of the mass (measured by the pickoff) relative to the case is proportional to the acceleration applied to the case. In opposite, the gravitational force acts directly on the mass and there is any relative movement of the mass in relation to the case. Accelerometers measure both static (gravity) and dynamic (movement) acceleration, therefore an accelerometer in equilibrium will measure the Earth's gravitational acceleration, which is approximately 9.81 m/s^2 (Groves, 2008).

In 2006, the top 5 accelerometer suppliers, Freescale, Analog Devices, Bosch, VTI and Denso, almost only provided sensors for the automotive market. However, nowadays, these inertial sensors can be found on regular smartphones, digital audio players, personal digital assistants, game controllers, mobile PCs, and camcorders.

The incorporation of IMUs in smartphones has been used for orientation view adjustment, pedestrian navigation, movement sensing, gaming controllers and pedometer applications. For most sensors, the coordinate system is defined relative to the device's screen when the device is held in a default orientation, as depicted in Figure 2.2.

This sensor is almost 10 times less consuming than other inertial sensors, however, the acceleration should be filtered to remove noise and also the gravitational component, for movement analysis. The output unit is in m/s^2.

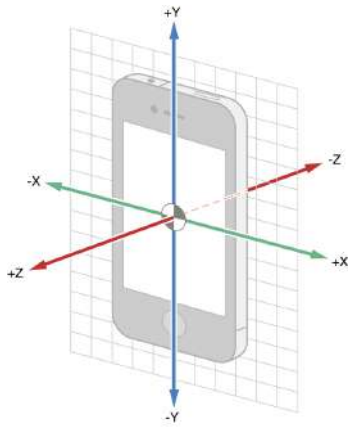

Figure 2.2: Coordinate system relative to the device.

2.2 Gyroscope

There are three main types of a gyroscope: mechanical, optical, and vibratory. Vibratory gyroscopes take advantage of the Coriolis effect to measure the angular velocity. When a mass m rotates with an angular velocity Ω, a force F is applied to the mass, that moves with velocity v, according to Figure 2.3. The sensor's principle is to detect the Coriolis acceleration of the vibrating element when the gyroscope is rotated. The Coriolis effect causes the object to exert a force on its support and the rate of the rotation is determined by this force, as in Figure 2.3. The Coriolis acceleration is perpendicular both to the direction of the velocity of the moving mass and to the frame's rotation axis. The physical displacement is read with a capacitive sensing interface and the output unit is radian/second.

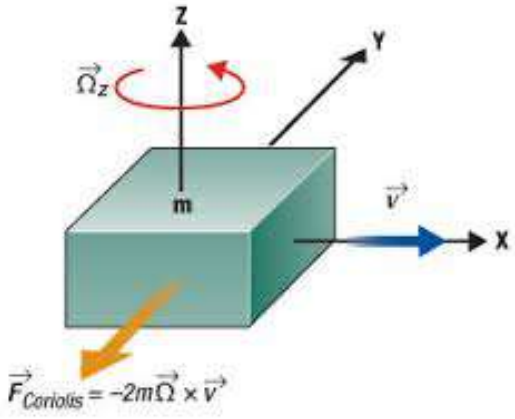

Figure 2.3: Coriolis effect illustration. Adapted from Vigna et al. (2010)

There are several conventions to represent the rotation angles around the three axes. The most commonly used are the Euler angles and the Tait–Bryan angles. The latter are also called

pitch, roll, and yaw and are defined as the rotation around X, Y and Z axis, respectively, and are represented in Figure 2.4.

MEMS gyroscopes have known limitations, such as the output drift over time, output offset when stationary and limited sensitivity. According to the application, the gyroscope can be combined with an accelerometer to compensate for the output drift.

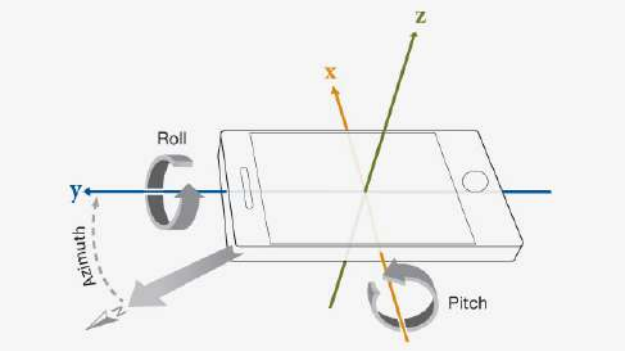

Figure 2.4: Magnetometer azimuth calculation. It is also illustrated the pitch and roll angles, around axis x and y, respectively.

2.3 Magnetometer

Magnetometers are sensors that measure the strength and/or direction of the magnetic field in a point at the space, based on the Hall effect. These sensors are used for measuring the Earth's magnetic field, to detect magnetic anomalies of various types, searching for mineral deposits or locating lost objects and also in the military to detect submarines.

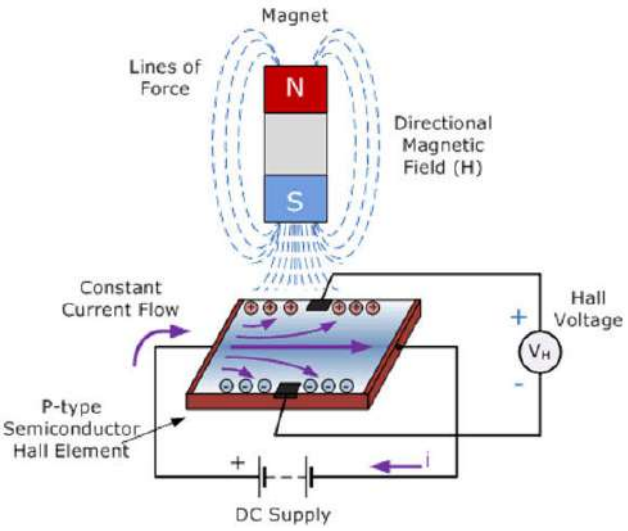

Figure 2.5: Forces that are applied to the electrons in the presence of Hall effect on a conductor. Adapted from hal (2013).

When a magnetic field with a perpendicular component is applied to a conductor, the charges that constitute the current of the conductor (electrons, holes, and ions) experience the Lorentz force, that accumulates charges on one face of the material. The force creates a difference potential between the sides of the conductor, as in Figure 2.5.

Most of the actual smartphones have built-in magnetometers, that allow measuring the Earth's magnetic field strength. This sensor provides raw field strength data (in micro Tesla) for each of the three coordinate axes and they are commonly used as an electronic compass. The sensor determines the azimuth component of the device orientation, according to Figure 2.4.

The output of the sensor can be influenced by the location on the planet, the weather or season of the year and also nearby electromagnetic devices such as magnets, electric coils or objects with ferrite elements.

2.4 Pressure Platform

The plantar pressure distribution data can be measured with PhysioSensing platform (Sensing Future Technologies, Lda)[1]. PhysioSensing is a portable balance and pressure platform with visual biofeedback technology. PhysioSensing allows to evaluate clinical practice and make it objective and quantified in a clinical report. PhysioSensing is indicated for balance, biofeedback, rehabilitation, physical, vestibular. It is a CE Medical Device Class I.

It contains 1600 pressure sensors of size 10mm by 10mm with a maximum value of 100N/sensor. Voltage data is converted with an 8-bit A/D converter and is transmitted via USB (Universal Serial Bus). In this way, it is possible to receive raw data of each pressure sensor as well as the raw center of pressure coordinates (CoP), in cm. The main specifications of the pressure platform are detailed below:

- Size: 61 x 58 cm
- Thickness: 1 cm
- Weight: 4 kg
- Active surface: 40 x 40 cm
- Number of sensors: 1600
- Sensor size: 1 x 1 cm
- Sensor type: Resistive
- Sensor thickness: 4 mm
- Sensor life time: more than 1 000 000 actuations

[1] https://www.physiosensing.net/

- Maximum pressure (each sensor): 100 N/cm2
- Temperature range: from 0ºC to 60ºC
- Frequency: 100 Hz, 100 acquisitions/second
- Data transmission: via USB (Universal Serial Bus)
- 8-bit A / D conversion
- Power: via USB cable
- Output: Raw data of each pressure sensor (8 bits) and coordinates of pressure center (x, y)

The pressure values are obtained through an 8-bit Analog Digital converter (A / D converter), that is, values from 0 to 255. Pressure values can be visualized through a conversion to color, where red represents 255 and white represents 0, as in Figure 2.6. These values can be converted into kg knowing the user's weight.

In order to obtain more precision in CoP displacements, an algorithm was employed to obtain CoP positions in mm, using the matrix of pressure sensors (Hsi, 2016).

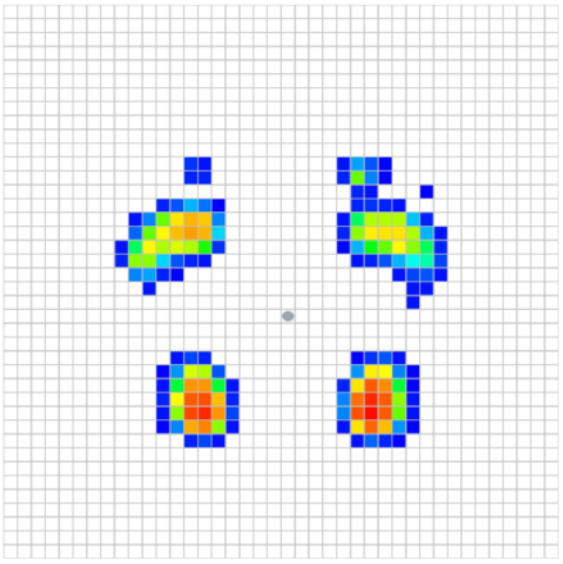

Figure 2.6: Data display of the pressure platform matrix, and the pressure center in gray.

2.5 Summary

The IMUs integrate a suite of three sensors: accelerometer, gyroscope, and magnetometer. The low cost, low dimension and remarkable performance of these sensors enabled their utilization in most of the wearable devices we carry every day. We can find these sensors integrated into nowadays smartphones, smartwatches, personal assistants, mobile PCs, and smart home devices. New

research areas, in which these sensing devices are applied, are emerging, such as fall detection, fall risk assessment and physical activity monitoring. The use of these sensors in personal companion solutions for the elderly allows to pervasively monitor the eventual occurrence of falls and to provide insights into the elderly movement patterns during the execution of specific activities. The most used sensor for movement analysis is the accelerometer, which provides an estimate in three axes of the acceleration of the user when the sensor is carried closer to the body. However, different body segments have different movement signatures, and most of the solutions based on inertial sensors have restrictions on the place to use the wearable device. The accelerometer signal requires further processing and filtering to be used for movement analysis, as it will be explained in the next sections. Combing movement data with plantar pressure data can allow us to obtain more information about the balance of a person, and at the same time to give visual feedback in order to correct unbalanced foot positions during the exercises. This way, this thesis will study the viability of using these inertial sensors and pressure platform, combined with signal processing and machine learning techniques, for developing fall detection and fall prediction solutions for the elderly population.

Chapter 3

Fundamentals of Machine Learning

Adapted from J. Silva, C. Figueira, M. Barandas, J. Gonçalves, J. Costa, L. Rosado, M. Vasconcelos, F. Soares, and H. Gamboa. "White Paper about Machine Learning @ Fraunhofer Portugal AICOS" (2017)[1]

3.1 Overview

In Artificial Intelligence, Machine Learning (ML) can be defined as a technology to learn autonomously from training data. It is a branch of computer science concerned with induction problems where an underlying model for predictive or descriptive purposes has to be discovered, based on known properties learned from a training set. Machine Learning algorithms can be divided into different categories, depending on the nature of the learning process:

- **Supervised learning** uses labeled data (data inputs and their desired outputs) to train an algorithm, which becomes able to map new inputs (Wilde, 2010). The supervised learning problems can be divided into Classification or Regression problems. In classification problems, the output to predict is discrete, whereas in regression problems the output is continuous;

- **Unsupervised learning** uses unlabelled data to build recognition models. Generally, the main objective is to identify and organize a dataset into different clusters through their similarity, providing significant information from the original dataset (Wilde, 2010; Ghahramani, 2004);

- **Semi-supervised learning** uses unlabelled data together with labeled data, i.e., in the learning process the training dataset is composed by, typically, a small amount of labeled data and a large amount of unlabelled data, falling between unsupervised learning and supervised learning (Witten and Frank, 2005).

[1] https://www.aicos.fraunhofer.pt/en/news_and_events_aicos/news_archive/older_archive/Machine_Learning_at_Fraunhofer_Portugal_AICOS.html

- **Deep Learning** has also been a big trend in Machine Learning. It is composed of several Artificial Neural Networks (ANN) processing layers. ANN are inspired by the structure and functions of the biological neurons. An artificial neuron has a finite number of inputs with the respectively associated weights and an activation function. Through the application of the activation function to the weighted sum of inputs, the output is obtained. An ANN is the result of the connection of many artificial neurons (Dayhoff and DeLeo, 2001; Moujahid, 2017).

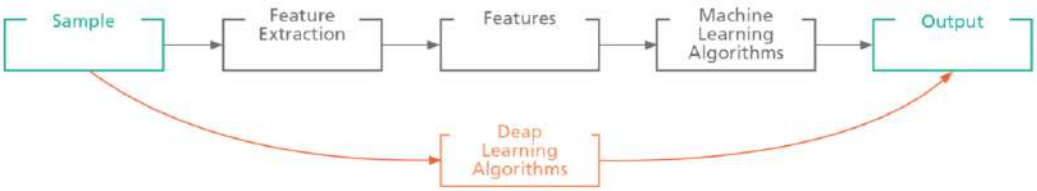

Figure 3.1: Traditional Machine Learning (above) and Deep Learning (below) flow.

Figure 3.1 illustrates that the main difference between traditional Machine Learning and Deep Learning algorithms lies in feature engineering. Traditional Machine Learning algorithms involves a feature extraction process, in order to provide relevant information that will have an essential role in the classification process (Figure 3.1 – above). On the other hand, Deep Learning algorithms perform feature engineering in an automatic way (Figure 3.1 – below) (Moujahid, 2017).

As represented in the figure above, a Machine Learning pipeline starts with a collection of data (the *Sample*). Afterward, a *Feature Extraction* process is performed, which consists of transforming a large quantity of data into a set of values, providing relevant information that will have an essential role in the classification process. A perfect feature type has a wide variation between different classes and a small one between the same class data (Øivind Due Trier et al., 1996). After *features* extraction, the next step is to apply *Machine Learning techniques* in order to construct a classification algorithm.

After the development of a classification algorithm, there are several different ways to evaluate its performance. First of all, it is necessary to have in mind the possible problem of overfitting. If we train and test the classifier with the same data, this situation may occur. The model constructed would just repeat the labels of the samples that it has just seen and exaggerated in minor fluctuations of data, leading to a perfect score. But testing the classifier with unseen data, the results would show poor predictive performance. A simple way to avoid it is using, for instance, a k-fold Cross-Validation (CV) method (Pedregosa et al., 2011). K-fold Cross-Validation divides randomly the original sample into k equal sized subsamples. From the k subsample, it is considered a single subsample as the validation set for testing the model, and the remaining k subsamples are used as the training set. This process is then repeated k times, where each of the k subsamples are used exactly once as the validation set. The k results obtained for each fold can then be averaged to produce a single estimation. When k is equal to the number of subjects, the K-fold Cross-Validation

becomes the Leave-One-Subject-Out (LOSO) Cross Validation, where each learning set is created by taking all the samples except the ones from one subject, and the test set is composed by the samples of the subject left out. In this case, for n subjects, there are n different training sets and n different test sets. Typically, this kind of validation is used when there are multiple samples per subject and we want to avoid subject bias, i.e., the model can learn features specific of each subject and it can fail to generalize to new subjects (Pedregosa et al., 2011; Kohavi, 1995).

Nested cross-validation applies a k-fold split between train and test and repeat it several times. It was first proposed by Dietterich (1998) as a way to obtain not only an estimate of the generalization error but also an estimate of the variance of that error, in order to perform statistical tests. Nested CV is not relevant if the dataset is large and without outliers, but if data have outliers than CV performance may be different depending on what folds the outliers are.

After the validation process, there are several metrics to evaluate classifier performance. One of the most used is the accuracy. *Accuracy* measures how close a value is to its true value (Eq. 3.1). Usually, we also report sensitivity (or recall), precision and specificity of the model. *Sensitivity* represents the ability of a model to correctly identify the positive class (true positive rate) (Eq. 3.2), whereas *specificity* is the ability of the model to correctly identify the negative class (true negative rate) (Eq. 3.3). *Precision* (positive predictive value) is the fraction of instances classified as positive class that were actually positive instances (Eq. 3.4). For binary classification, it is common to report *F1-score* (F-score or F-measure) and Youden's J statistic (J index). F1-score is the harmonic mean of precision and recall and it is used as a single measure of the test's performance for the positive class (Eq. 3.5). The *J index* combines sensitivity and specificity in a single statistic that captures the performance of a binary test (Eq. 3.6), as well as the *geometric mean*, which is the squared root of the product of the sensitivity and specificity (G) (Eq. 3.7).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Se = \frac{TP}{TP + FN} \tag{3.2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3.3}$$

$$Prec = \frac{TP}{TP + FP} \tag{3.4}$$

$$F1 = \frac{2 \times Prec + Se}{Prec + Se} \tag{3.5}$$

$$YI = Sp + Se - 1 \tag{3.6}$$

$$G = \sqrt{Sp \times Se} \tag{3.7}$$

Where *TP* are the True Positives, *TN* the True Negatives, *FP* the False Positives and *FN* the False Negatives.

3.2 Performance evaluation of binary predictors

The Receiver Operating Characteristic curve (ROC) and Total Operating Characteristic curve (TOC) are used to visualize a binary classifier performance for all possible threshold choices in a single graph. However, these are based on two different coordinate systems. A ROC curve is plotted into an FPR vs. TPR coordinate system, for all threshold values between 0% and 100%. A TOC curve, is plotted into a (TP+FP) vs. TP coordinate system, for each threshold value.

Each point on a ROC curve determines the corresponding confusion matrix if the total number of positive and negative samples are known. However, the ROC graph does not contain the total test set composition as visual information. The ROC curve has long been used, but this shortcoming has recently been addressed by the introduction of the TOC curve by Pontius and Si (2014). A TOC curve displays the full ROC information and additionally allows to visualize the total information, i.e., the test set's composition and all the four entries of the confusion matrix, for each point on the curve.

- **Receiver Operating Characteristic curve (ROC)** is a graphical plot of the diagnostic ability of a binary classifier system at all classification thresholds. The ROC curve is composed by plotting in the y-axis the true positive rate (TPR) against the false positive rate (FPR), in the x-axis, at various threshold settings, as can be seen in Figure 3.2 (left graph). The true-positive rate is also known as sensitivity, recall or probability of detection. The false-positive rate is also known as probability of false alarm and can be calculated as (1 - specificity). The points in the upper left of the ROC curve are the good ones, since these present higher sensitivity and higher specificity. Any point on the diagonal line represents a classifier that is guessing randomly. The Area under ROC curve (AUC) can be interpreted as an estimate of the probability that the classifier will give a random positive instance a higher score than a random negative instance.

- **Total Operating Characteristic Curve (TOC)** shows the total information in the confusion matrix for each threshold. TOC maintains desirable properties of ROC, while revealing more information than ROC. To analyze the results obtained for each threshold, these can be plotted in a TOC graph. In order to generate this curve, the hits, true positives (y-axis), are plotted against the hits plus the false positives, this is, against the total of positive predictions (x-axis) (Pontius and Si, 2014). Therefore, this curve allows better visualization of the balance between false positives and the number of true positives, as illustrated in Figure 3.2 (right graph).

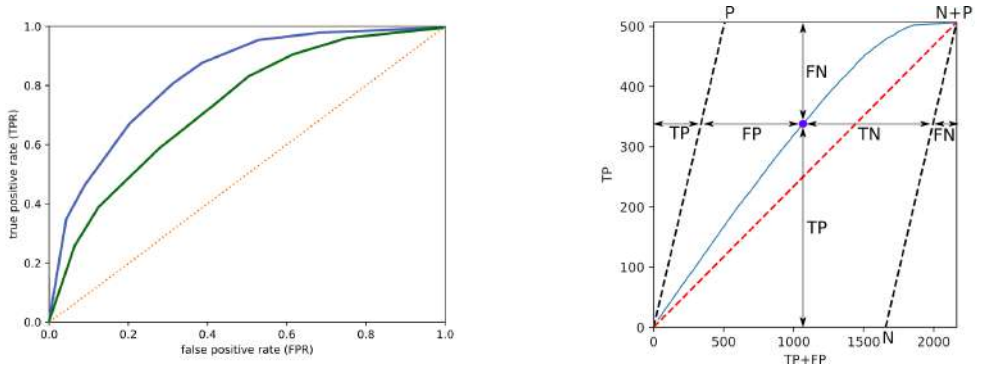

Figure 3.2: Receiver Operating Characteristic curve (ROC) (Left) and Total Operating Characteristic Curve (TOC) (Right).

3.3 Learning Methods

The choice of the algorithm depends on the type of dataset available and the objectives of the research. In Machine Learning, several algorithms are used in different contexts and applications, which will be now described.

- **K-Nearest Neighbours (*k*-NN)** find the k closest instances of the training set according to a metric measure, where the resulting class is the most frequent class label of the k nearest instances Kotsiantis (2007);

- **Decision Tree (DT)** creates a model that classifies instances by sorting them based on data features values. The major goal is to determine the best decisions (Kotsiantis, 2007);

- **Naïve Bayes (NB)** predicts class membership probabilities, based on the Bayes Theorem. Independence between the features and prior probabilities are assumed. Based on previous experience, these probabilities are then used to predict outcomes before they actually happen (nai, 2017);

- **Support Vector Machine (SVM)** is a binary classifier that builds a model that assigns new data into one category or other from a set of training examples. The main objective is to find the maximum hyperplane which separates the data classes (Kotsiantis, 2007);

- SVM classifiers require the solution of high quadratic programming (QP) optimization problem. Based on it, a new algorithm to train SVM was created, the **Sequential Minimal Optimization (SMO)**. With SMO, the QP problem is solved by being divided into the smallest QP problems, which are then solved analytically (Platt, 1998);

- **Markov Models** are based in probabilities and are used in cases where certain conditions may happen repeatedly over time or for modeling predictable events that take place over time. Hidden Markov Models (HMM) is a special case of Markov Models. In this case, the

Markov process shows unobserved (i.e. hidden) states and there is no knowledge regarding the existing states and transition probabilities between them (Fosler-Lussier, 1998);

- **Adaboost** is a boosting algorithm, i.e., an algorithm based on the combination of weak and inaccurate rules in order to construct a prediction rule (Schapire, 2013);

- **Spectral Clustering** uses the main eigenvectors of the similarity matrix resulted from points distance. With this, data dimension is reduced into fewer dimensions (Ng et al., 2002);

- **Convolutional Neural Networks (CNN)** are a type of neural network where the input is an image and the fully connected layers are replaced by layers of convolutional filters. The use of convolutional filters instead of fully connected layers, with neurons connected to every neuron in the previous layer, allows to significantly decrease the number of parameters of the network, thus enabling the network to efficiently learn features in relatively large images. The goal of the training phase is to learn the weights in those convolutional filters, in order to minimize a given loss function;

- **Recurrent Neural Networks (RNN)** are popular models which, unlike traditional ANN, consider that inputs and outputs are dependent on each other. Long Short-Term Memory (LSTM) networks are a type of RNN, which are capable of learning long-term dependencies (Sak et al., 2014).

3.4 Imbalance Learning

- **Synthetic Minority Over-sampling Technique:** (SMOTE) (Bowyer et al., 2011) was used to oversample real-world samples in the train set. Using this approach, the minority class is oversampled by creating "synthetic" examples rather than by oversampling with replacement. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. A variation of SMOTE called SMOTE-NC (Nominal and Continuous) allows to use both continuous and nominal data.

- **Balance Cascade:** creates an ensemble of balanced sets by iteratively undersample the imbalanced dataset using an estimator (Liu et al., 2009). This method iteratively select subset and make an ensemble of the different sets. The selection is performed using a specific classifier. SMOTE and Balance Cascade are implemented in Python's *imbalanced-learn* (v.0.2.1) package (Lemaître et al., 2017).

- **Ranking Models:** are used for tackling class imbalance with ranking, using an application of learning pairwise rankers. Several models can be used with this approach, such as Adaboost, Balanced linear SVC, Linear SVC, Rankboost and Rank SVM (Cruz et al., 2016).

3.5 Statistical Analysis

- **Fisher Exact Test** is statistical test used to determine if there are nonrandom associations between two categorical variables, when the sample sizes are small. Fisher's exact test returns a test decision for the null hypothesis that there are no nonrandom associations between the two categorical variables, against the alternative that there is a nonrandom association. The result is 1 if the test rejects the null hypothesis at the 5% significance level, or 0 otherwise.

- **Odds Ratio** (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure Szumilas (2010). If the OR is greater than 1, then the exposure and the outcome are associated (correlated). Conversely, if the OR is less than 1, then the exposure and the outcome are negatively correlated, and the presence of one event reduces the odds of the other event.

- **Analysis of variance (ANOVA)** is a statistical comparison analysis test that analyses the means between groups and determines if any of those means are statistically significantly different from each other. ANOVA provides a statistical test of whether two or more groups' means are equal, and therefore generalizes the t-test beyond two means. ANOVA allows to know if the differences are significant or not, but does not allow to know between which pairs the differences are significant.

- **Tukey's Honest Significant Difference Test (HSDT)** is a *post-hoc* test that allows to interpret the statistical significance of ANOVA test and find out which specific groups' means (compared with each other) are different. After performing each round of ANOVA, one should use a Tukey Test, with 95% confidence level, to all possible pairs to find out where the statistical significance is occurring in the data.

- **p-value** is a measure of the probability that an observed difference could have occurred just by random chance. The lower the p-value, the greater the statistical significance of the observed difference. For a confidence level of 5%, a p-value < 0.05 indicates a statistically significant difference between groups, and a p-value > 0.05 indicates there is not a statistically significant difference between groups.

3.6 Machine learning pipeline

The Cross Industry Standard Process for Data Mining (CRISP-DM) described by Wirth and Hipp (2000) proposed a standard process model for carrying out data mining projects. The CRISP-DM process model is useful for planning, communication within and outside the project team, and documentation. The main steps of the data mining life cycle are presented in Figure 3.3.

Based on CRISP-DM, one can define a machine learning pipeline which considers not only the modeling phase, with the tuning of a learning algorithm, but all the steps for data preprocessing

and feature engineering that precedes the model inference step. Most of the time, the prediction algorithm is the easier part to deploy, but the other steps of the pipeline could be more difficult to integrate into other platforms (e.g. sensor fusion algorithms, feature extraction algorithms). When we use deep learning models, the feature extraction and selection steps are usually not needed, and we focus only on the preprocessing steps and the neural network implementation. We will explain following the main tasks underlying each step of the ML pipeline:

- **Domain understanding** Discuss with domain experts their needs for using machine learning, formulate the problem task, the variables that need to be collected, how the data will be collected and structured. Discuss the specifications of the data and the problem.
- **Data collection and annotation** Define data requirements, structure, loggers and annotation procedures, as well as sample size needed. Consider ethical, fairness, bias and privacy issues.
- **Data verification**
  - Data cleaning: imputation of missing values, removal of duplicate samples, data verification procedures for detection of acquisition errors
  - Unbalanced data: evaluate the need for data over/undersampling or augmentation.
  - Oversampling: should only be applied in the training set, the validation set should have the same unbalance nature as the test set.
  - Sampling techniques: uniformize each window/segment – for example, ensure that each sample has the same sampling rate, i.e. the same number of points
- **Data partitioning** Divide the entire dataset into 3 partitions to be used for train, validation, and test, respectively. The train and validation sets constitute normally 2/3 of the dataset and the test set comprises the other 1/3 of data.
- **Data segmentation** Split each data sample into small windows/segments for analysis, if applicable.
- **Feature extraction** (if not deep learning) Use libraries or define algorithms for feature extraction that are applied for each window/segment. The output will be a feature vector. The rows correspond to each instance, i.e. window/segment, and the columns correspond to each feature value.
- **Feature selection and normalization** (if not deep learning) Remove correlated features or useless features with specific algorithms, such as forward feature selection. Normalize features in the train set and apply the same normalization parameters to the test set. The output will be a feature vector with only the selected features and a set of parameters for feature normalization of the test set.
- **Models' hyperparameters tuning**

  - Define a cross-validation procedure, either by K-fold Cross-Validation or Leave-One-Subject-Out Cross-Validation
  - Define routines or use toolboxes for models' hyperparameters optimization.
  - Define the models to be tested and each parameter's range to be considered.
  - Define a performance metric for retrieving the best models.
  - Define requirements for model deployment if needed at this stage.
  - The output will be the best model's hyperparameters and the model implementation (if it is deep learning, the architecture of the network).

- **Model performance evaluation** Using the test set, report the performance metrics for this dataset partition. Consider informative metrics for the problem task at hand. Be careful with unbalanced data when reporting the model's performance.

- **Model deployment** Implement the same data verification, segmentation, feature extraction and selection, feature normalization and models in the deployment platform. Use this model to infer for new samples.

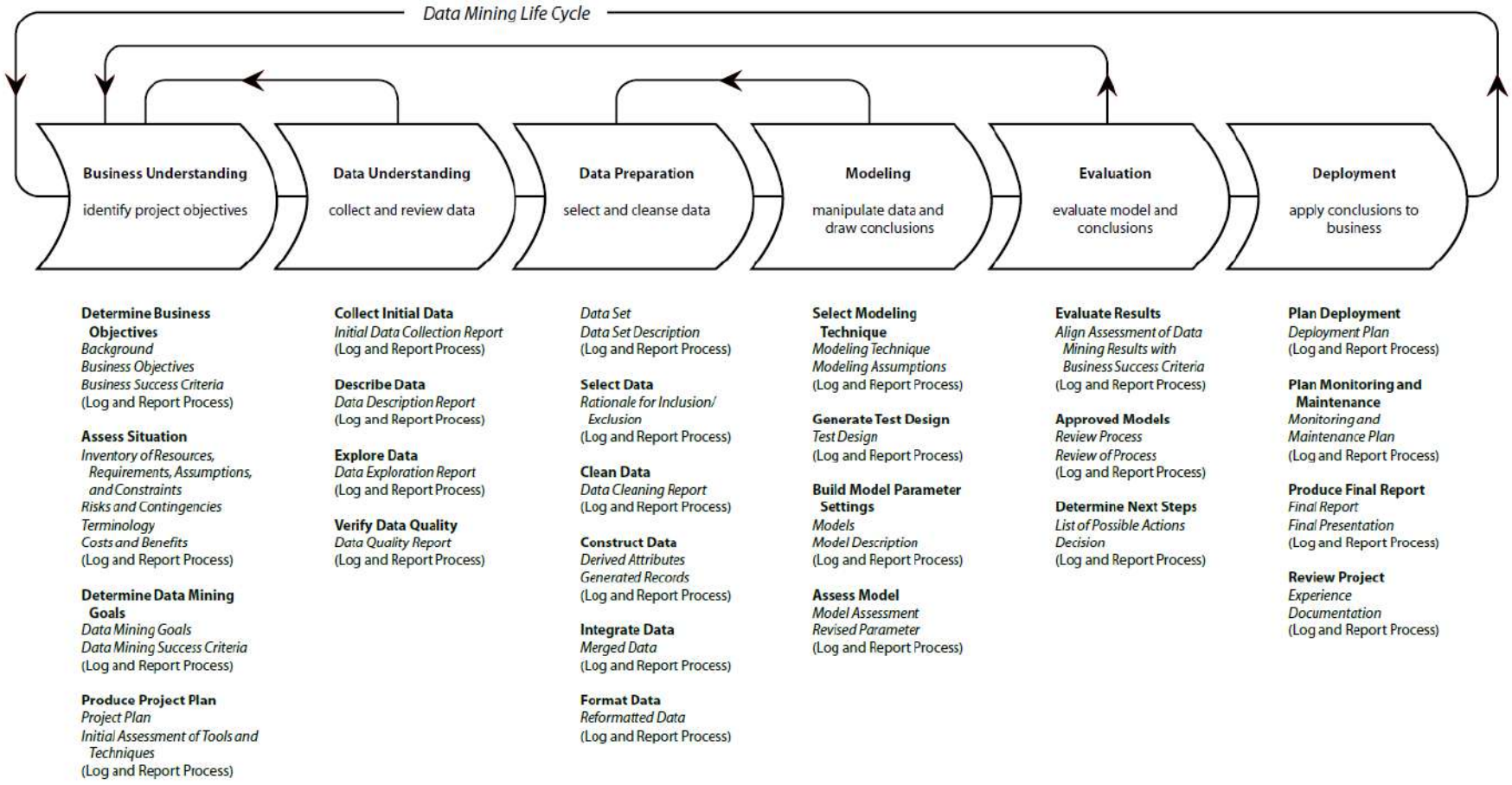

Figure 3.3: Data mining life cycle, adapted from Leaper (2009).

3.7 Fall prediction and detection pipeline

The research problems related to Signal Processing, such as fall prediction or detection, use time series collected from inertial and environmental sensors, which are embedded in smartphones or wearable sensors. Currently, the methodology for collecting samples from inertial sensors uses a recording tool to acquire samples of movements or activities that we want to analyze. The data can be collected from smartphones and wearable platforms, which contain an inertial measurement unit with an accelerometer, gyroscope, magnetometer and barometer sensors. After data acquisition, the processing pipeline follows the data verification, feature extraction, modeling and deployment of the best performing model. In the next sections, it will be explained in detail the ML pipelines used for each classification problem. For fall prediction, we want to discriminate between fallers and non-fallers based on multifactorial data and the prospective 1-year number of fall occurrences. For fall detection, we want to discriminate between events of fall and other non-fall movements based on annotated samples of movements collected for each type of event.

Part II

Multifactorial wearable-based fall prediction: study of feature extraction and data fusion

Chapter 4

Introduction

4.1 The problematic of falls

The worldwide population aged over 65 is growing rapidly. The consequences of this phenomenon are not only social and health-related but also economic. The process of aging affects the ability of a person to maintain balance, mobility, and muscle strength and to react properly to unexpected situations such as slipping or stumbling. There are also cross-related factors resulting from health conditions, including loss of auditory and visual capabilities, side effects of medications, dizziness, body pain, depression, and slow walking speed. Aside from these intrinsic risk factors, falls among older people are also associated with extrinsic factors, such as environmental hazards, footwear malfunctioning, improper use of assistive devices, and recent hospitalizations (Ambrose et al., 2013).

Falls are described as a complex phenomenon caused by the interaction of multiple risk factors. To assess the risk of falling, it is necessary to identify the factors that increase an older person's risk of falling. Intensive research has been conducted in order to identify specific risk factors (Ambrose et al., 2013; Rubenstein, 2006; Oliver et al., 2004), which can increase the likelihood of a fall occurrence. The idea behind these studies is to develop preventive strategies based on the identified risk factors.

Given a wide range of factors contributing to falls in the context of an aging population, it becomes extremely important to frame strategies that properly evaluate the risk factors of falls in older people. Several scales, questionnaires, functional tests, and protocols have been proposed in the past years to overcome the lack of standardized clinical and medical procedures for assessing the risk of falls (Howcroft et al., 2013). However, in the majority of public sectors, risk factors of falls among the elderly are only assessed after the occurrence of a fall leading to hospitalization or the need for other forms of medical care. When a fall risk assessment is conducted after an occurrence of a fall, the collected parameters are altered as a consequence. On the other hand, the majority of the proposed assessment scales and questionnaires are subjective and self-reported and do not consider all major fall risk factors. Proper methods for the objective assessment of individual gait, strength, and balance are confined to laboratory settings requiring specialized personnel

and equipment, thus leading to higher costs. All these solutions rely on on-time assessments that do not reflect the variation of risk factors over time.

Fall risk assessment (also referred to as fall prediction) methods have been studied aiming to estimate the risk of falling in order to identify those at higher risk and timely apply the appropriate actions to prevent falls. This kind of assessment can take the form of questionnaires, simple screenings or more comprehensive multidimensional fall risk assessments.

4.2 Fall risk factors

The majority of falls can be attributed to a physiologic cause, 78% of falls are labeled *anticipated* (i.e., physiological falls that can be predicted in patients exhibiting clinical signs that contribute to increased falls risk), and 8% labeled *unanticipated* (i.e., physiological falls that cannot be predicted before their first occurrence). The remaining 14% of falls are labeled *accidental* (i.e., often attributed to environmental causes) (Feil and Gardner, 2012).

Most of the anticipated falls could be predicted and the associated fall risk factors are reversible. Commonly referred to as main causes and risk factors for falls include muscle weakness, gait and balance problems, visual impairment, cognitive impairment, depression, functional decline, and particular medications (especially in the presence of environmental hazards) as described by Rubenstein and Josephson (2002) in Table 4.1. The authors surveyed 16 studies and compared common risk factors among fallers and non-fallers. Muscle weakness was identified as a significant risk factor, that increases the odds of falling over 4.4-fold in average. This factor is followed by the history of falls, gait and balance deficit that increase the odds of falling by 3.0-fold on average.

Table 4.1: Most common risk factors for falls, adapted from (Rubenstein and Josephson, 2002).

Risk factor	Mean Relative Risk Ratio	Range
Muscle weakness	4.4	1.5–10.3
History of falls	3	1.7–7.0
Gait deficit	2.9	1.3–5.6
Balance deficit	2.9	1.6–5.4
Use assistive device	2.6	1.2–4.6
Visual deficit	2.5	1.6–3.5
Arthritis	2.4	1.9–2.9
Impaired activities of daily living	2.3	1.5–3.1
Depression	2.2	1.7–2.5
Cognitive impairment	1.8	1.0–2.3
Age >80 years	1.7	1.1–2.5

4.3 Assessment methodologies

Revised American Geriatrics Society (AGS) guidelines specify for individuals above 65 years of age an annual fall risk screening and assessment that should evaluate fall history, balance, and gait assessment, with a complete multifactorial fall risk assessment performed by a specialized clinician when the subjects have positive balance or gait disabilities (DePasquale, 2014).

Most of the fall risk assessment tests evaluated whether the instrument or test can accurately differentiate fallers from non-fallers (Hempel et al., 2012).

Fall risk assessment is a vast research area with widely disparate approaches being used. There are various scoring systems intended for use in hospitals, nursing homes or outpatient settings. The available indices are designed to be used by different professionals (e.g., geriatric doctors, nurses or physical therapists) and are based on questionnaires, observations, physical examinations or their combination (Shany et al., 2012).

The fall risk assessment aims to determine an older person's risk of falling in order to plan coordinated fall prevention strategies and long-term follow up. The assessment is sometimes performed in specialized settings like a fall clinic and includes methods that are specifically designed and tested for the risk of falling (e.g. gait speed, static balance, strength, dual-task measures, cardio-vascular diagnosis, etc) (Becker and Lamb, 2007).

A systematic review of multifactorial and functional mobility assessment tools for fall risk (Scott et al., 2007) compares several studies for community settings.

4.3.1 Questionnaires

Some of the widely used questionnaires to assess a variety of fall risk factors are listed below. The first questionnaire is one of the most used questionnaires to assess fear of falling in daily activities, and the other two questionnaires are mostly used in hospital settings, and the last one is a recent web-based questionnaire to assess fall risk:

- **Falls Efficacy Scale** (FES) – Initially proposed by Tinetti, and later adapted to several languages, it is a 10 items questionnaire that is scored on a scale of 0 to 10. The higher the score the greater the fear of falling (Tinetti et al., 1990).

- **St. Thomas' Risk Assessment Tool** (STRATIFY) – This tool can be used to identify risk factors for falls in hospitalized patients. The total score may be used to predict future falls, but it is more important to identify risk factors using the scale and then plan care to address those risk factors. It evaluates the history of falls, vision, transfer and mobility capabilities. 0 = Low, 1 = Moderate, >2 = High Risk (Agency for Healthcare Research and Quality, 2013).

- **Morse Fall Scale** – The Morse Fall Scale (MFS) is a rapid and simple method of assessing a patient's likelihood of falling. Evaluate history of falling, secondary diagnosis, ambulatory

aid, intravenous therapy/heparin lock, gait, and mental status. Sample risk level is categorized into three levels: no risk (0-24), low (24-44) and high risk (>45). Maximum score is 125 points (Schwendimann et al., 2006).

- **Fall Risk Assessment Tool for Community-Dwelling Older People** (FRAT-up) – This is a web-based fall risk assessment tool for elderly people living in the community. This fall prediction tool is based on a meta-analysis of fall risk factors. Based on the fall risk factor profile, this tool calculates the individual risk of falling over the next year (Cattelani et al., 2015).

Other questionnaires commonly used are Fall Risk assessment and Screening Tool (FRAST); Modified Falls Efficacy Scale, (mFES), Modified Gait Efficacy Scale (mGES), Activities-specific Balance Confidence short version (ABC-6), Fear of Falling Avoidance Behavior Questionnaire (FFABQ), Survey of Activities and Fear of Falling in the Elderly (SAFE), Physical Functioning Scale of the Short-Form (SF).

4.3.2 Functional Tests

There are several functional tests used for fall risk assessment, most of them use observational scales, manual counting, timing metrics, and few of them use objective and automatic scaling processes. The functional more used in previous studies are described below:

- **Physiological Profile Approach** (PPA) – Considers gait, balance, vision, proprioception and vibration sense and strength, but omits assessments of medication, medical condition or home hazards. PPA tests produces a *gold standard* fall risk score that categorizes subjects from very low risk (-5), up to marked risk (5) (Redmond et al., 2010).

- **Tinetti Performance Oriented Mobility Assessment Tool** (POMA) – The Tinetti Assessment Tool is a simple, easily administered test that measures a patient's gait and balance. The balance component has 13 maneuvers such as sitting balance, sit to stand, immediate standing balance (first 3–5 seconds), standing balance, balance with eyes closed, turning 360^{o}, nudging the sternum, turning the neck, unilateral stance, extending the back, bending down and picking up an object, and sitting down. The gait analysis consists of the 9 components of initiation of gait, step height and length, step symmetry and continuity, path deviation, trunk stability, walking stance, and turning while walking (Faber et al., 2006). The maximum score for the gait component is 12 points. The maximum score for the balance component is 16 points. The maximum total score is 28 points. In general, patients who score below 19 are at high risk for falls. Patients who score in the range of 19-24 indicate that the patient has a risk for falls. Scores between 25 and 28 are associated with low fall risk (Faber et al., 2006).

- **Berg Balance Scale** (BBS) – evaluates individual's balance capabilities in sitting, standing, transfers, stand with eyes open and eyes closed, reaching forward with an outstretched arm,

retrieving object from floor, turning to look behind, turning 360^{o}, placing alternate foot on a stool, standing with one foot in front and standing on one foot. Each task is rated on a four-level scale (0 = lowest level of function 4 = highest level of function) and the final score is cumulative of all task scores. The maximum score is 56 and a score of less than 45 is indicative of balance impairment (Newton, 1997).

- **Timed Up and Go Test** (TUG) at normal pace – the subject is asked to start seated on a chair and when test starts, the person should stand up, walk straight for 3 meters, turn around and walk back to the chair and sit down. The final score of this test corresponds to the time needed to perform the TUG test. Scores higher than 14 seconds are considered associated with high fall risk. Some studies reported a variation of TUG at a fast pace, the subjects are asked to perform it as fast as they can. No normative values were found, however, a mean score of 9.85 $\pm$ 1.44 seconds was obtained by 120 healthy older adults between the ages 60-87 (Hofheinz and Schusterschitz, 2010).

- **10-meters walking speed test** – the subject is instructed to walk at his/her fastest walking speed. It requires a 20m straight path, with 5m for acceleration, 10m for steady-state walking, and 5m for deceleration. Markers are placed at the 5m and 15m positions along the path and the time to traverse is registered. The range for normal walking speed is between 1.2 and 1.4 m/s. Normative values are: $<$ 0.4m/s indicate a probability of needing marching help at home; 0.4 to 0.8 m/s is correlated to limited mobility; 0.8 to 1.25 m/s indicate subjects wander in community with some risks (Fritz and Lusardi, 2009).

- **Four-Stage Balance Test** – the subject is instructed to stand in four balance positions: stand with feet together, semi-tandem stand with the instep of one foot so it is touching the big toe of the other foot, tandem stand with one foot in front of the other, heel touching toe and one leg stand. Each position should be maintained for 10 seconds without moving the feet or needing support. Scoring is binary, as able or not able (Murphy et al., 2003).

- **Romberg (R) & Sharpened Romberg** (SR) - the subject is instructed to perform two balance positions with eyes open and with eyes closed: feet together, firm surface and compliant surface (i.e. foam) and feet heel-to-toe (dominant foot behind non-dominant foot) (Agrawa et al., 2011).

- **30s Sit-to-Stand** – maximum sit and stand transfers for 30 seconds. The cut-off for risk level is divided by age and gender. For example, for a female with 65 years old, the threshold is 15 repetitions (Rikli and Jones, 2013).

- **Step Test** - was designed to assess dynamic standing balance and reproduce lower extremity motor control and coordination. To perform the test, the person is asked to step on and off a block (7.5 cm height, 55 cm width, and 35 cm depth) placed against a wall as many times as possible for 15 seconds. The total number of completed steps in 15 seconds is recorded.

This test is performed only for the dominant side, as indicated by the person being tested. A performance of <10 steps indicates a higher risk of falling (Hill, 1996).

- **Alternate-Step Test** (AST) is a modified version of the stool stepping task in a BBS evaluation. It evaluates the participant's weight-shifting ability in the forward and upward directions. The participant is instructed to place the entire left and right foot alternately on the step as fast as possible, 8 times for each foot. Each successful step involved placing the entire foot on the step and returning it to the floor. The time required for completing a total of 8 steps is measured using a stopwatch (Chung et al., 2014).

- **Grip Strength** – evaluates the maximum voluntary force of manual grip. The grip strength relates to the lower limb strength and with the individual's functional capacity. The grip strength is measured with a dynamometer and measured in kilograms (Pizzigalli et al., 2016).

4.4 Wearable approaches for fall prediction

Recently, solutions for the fall risk assessment based on low-cost technologies have been proposed (Howcroft et al., 2013), including solutions based on inertial sensors embedded into wearable devices or smartphones. There are also solutions based on force and pressure platforms aiming at assessing multiple factors of balance and correlated fall risks.

Wearable devices containing inertial sensors have been used for collecting movement data during the execution of standard mobility tests such as the TUG (Salarian et al., 2010; Greene et al., 2010), STS (Doheny et al., 2011), 5 times sit-to-stand (5-STS) (Narayanan et al., 2010; Liu et al., 2011; Doheny et al., 2013) or stance balance (Doheny et al., 2012). The advantage of using inertial sensors during the execution of a standard test is the additional quantitative information that can be derived. This information may be helpful to better assess and characterize the mobility and balance conditions of a person. Moreover, some characteristics that are not perceived with subjective assessments, such as POMA, may become relevant when computing objective metrics from the inertial sensors signal. By eliminating the need for observation of movements and subjective assessment the output extracted is potentially more reliable and reproducible.

The inertial sensors built-in smartphones or wearables have been used in our previous studies for fall detection (Aguiar et al., 2014a) and activity monitoring (Aguiar et al., 2014b) and similar metrics and algorithms can be applied to the signal collected during standard fall risk tests. The introduction of this accessible and easy to use technology can contribute to making fall risk assessments more widespread and reach a larger potentially affected population.

The TUG test evaluates the time that a person takes to stand up from a chair, walk forward, turn around, walk back to the chair and sit down. This test was reported to be predictive of falls with a sensitivity of 81% and a specificity of 39% (Whitney et al., 2005) and is currently widely used as a standard fall risk assessment tool. The instrumented version of TUG (iTUG) has recently been used to extract quantitative information during the test, besides the total time of execution,

which was the only score of the original TUG test. The analysis of the sensors used to instrument the iTUG could give more insights during the walking phase, the sit and stand transition phases, and the turning phases of the TUG test. Salarian et al. (2010) used the iTUG test to assess mobility impairments in Parkinson's patients versus age-matched controls. Seven inertial measurement units containing accelerometers and gyroscopes were attached to the forearms, shanks, thighs, and sternum of the participants while performing the iTUG test with 7m walkway at their normal speed. Automatic segmentation of the four major components of iTUG, sit-to-stand, steady-state gait, turning, and turn-to-sit, was performed. Despite the similarity of the total time to perform iTUG, several measures related to some components of iTUG, namely gait, turns and turn-to-sit, showed significant differences between groups. Test-retest reliability was also generally good. Greene (2014) used the iTUG test and logistic regression models to classify individuals with and without falls history, using temporal and angular velocity parameters derived from two sensors attached to the shanks. The models demonstrated good overall performance, with a mean sensitivity of 77.3% and mean specificity of 75.9%. This performance is higher than the one obtained using logistic regression models based on the standard TUG timing and the BBS.

In addition to the iTUG, waist-mounted accelerometers have been used for data collection during the Alternate-Step Test (AST) and 5-STS, for which time-domain (Narayanan et al., 2010) and frequency-domain (Liu et al., 2011) features have been extracted. After mapping the extracted features toward the target value of the PPA score using a linear least-squares model, the leave-one-out cross-validation was employed. A good correlation with PPA score and low root mean squared error (RMSE) were found especially when including frequency-domain features.

The related work described before comprises studies that used any type of sensors to retrieve metrics during the execution of fall risk functional tests. The studies focused only on clinical, self-reported, or measurable variables (e.g., (Nelson et al., 2001; Vassallo et al., 2008)), are not discussed in this section. We only note that the sensitivity achieved in these studies varied from 43% to 100% (median = 80%), whereas the specificity ranged from 38% to 96% (median = 75%).

The work of Palumbo et al. (2014) estimated that a theoretical maximum accuracy of a fall prediction model, attempting to identify people with at least one fall incident over a year from the time of the testing, would not exceed 0.81 (maximum AUC of 0.89), which has a moderate effect size. They expect the statistical effect of fall prediction models to be small, which is the reason why many research studies in the area report negative results, especially if the sample size is small.

Howcroft et al. (2013) reviewed previous studies focusing on fall risk assessment with inertial sensors. The authors concluded that future research should: i) consider investigating the relationship between the models' predictive variables and specific fall risk factors and ii) focus on groups with an increased fall risk due to some diseases. As weak points of the previous studies, the authors reported that 50% of them had not used separate datasets for model training and validation, which could have impacted the models' applicability beyond the training set population. The same pitfalls were identified by Shany et al. (2015), who bemoaned that most of the previous literature is reporting over-optimistic results due to small sample sizes used, questionable feature selection processes, and biased validation methodologies that lack external validation sets. Moreover, ap-

plying the most commonly used cut-off values in clinical assessment tests could have biased the decisions made since the thresholds typically used to split classes had produced false positives and false negatives, introducing inaccuracies when evaluating sensor-based models. Another aspect to be considered is that clinical assessment thresholds were not used consistently across the research studies included in the review. The prospective fall occurrence rate is considered to be the most reliable criterion for dividing subjects into non-fallers and fallers (Howcroft et al., 2013); however this criterion was only used in 15% of the studies. Regarding the retrospective fall assessment, the most relevant limitations are the inaccurate recording of fall histories most commonly assessed by self-reported questionnaires and the fact that balance, strength, and gait parameters can change due to past falls.

4.4.1 Retrospective studies

Bigelow and Berme (2011) studied posturography for clinical fall risk screening of older adults. They recruited 150 adults aged 65 and above from local senior centers and independent living facilities. The subjects were categorized as recurrent fallers and non-recurrent fallers based on their fall status in the previous year. The participants performed four standing tasks on a force platform. The authors extracted "traditional and fractal measures from the center of pressure data" (Bigelow and Berme, 2011). Their logistic regression model exhibited a sensitivity (recall) of 75% and a specificity of 94%. The authors highlighted the importance of combining multiple variables rather than using only a single measure to compute the fall risk.

Qiu et al. (2018) reported a study conducted with multiple wearable inertial sensors for multifactorial fall risk assessment on 196 community-dwelling older women. The sequence included the TUG, 5-STS, and Limits of Stability tests. A model built using inertial sensor data and support vector machine was able to classify between fallers (N = 82) and non-fallers (N = 114) based on fall histories. The model achieved an overall accuracy of 89.4% (92.7% sensitivity and 84.9% specificity). The results of the study support the idea that inertial sensors allow the identification of individuals with a high risk of falls, who should be followed with fall prevention strategies.

Greene et al. (2012) performed a quantitative estimation of the fall risk using multiple sensors during the standing balance exercise. The authors acquired data from 120 community-dwelling older adults aged over 60 by using a pressure-sensitive platform sensor and attaching a body-worn inertial sensor to the lower back of the participants. The estimation of the fall risk was compared with the BBS. The results were analyzed by gender using a support vector machine model, which returned a mean classification accuracy of 73.07% for the participants with a self-reported history of falling in the past 5 years. These results compared favorably with those obtained using solely the BBS (with a mean classification accuracy of 59.42%).

4.4.2 Prospective studies

Liu et al. (2014) reported an accelerometer-based fall prediction model that was trained using wearable inertial sensor data obtained in a routine assessment, including the TUG test, AST, and

5-STS. The study sample included 68 subjects aged from 72 to 91 from a previous study and a second group of 30 subjects aged from 68 to 92 who were newly recruited. The authors have assessed the prospective falls that occurred in the following 12 months based on fall diaries. The best classification performance allowing to distinguish fallers from non-fallers with a sensitivity of 68% and a specificity of 73% was achieved by a logistic regression model that was trained using only AST data.

van Schooten et al. (2015) performed several studies focusing on the assessment of the ambulatory fall risk. The study participants aged over 65 wore an inertial sensor for one week. The authors extracted metrics related to the amount of physical activity and gait characteristics and reported several approaches ranging from logistic regression to deep learning methods to discriminate between fallers and non-fallers. A logistic regression model trained on accelerometry-derived parameters of gait obtained from 139 participants allowed to substantially improve the area under the curve (AUC) up to a value of 0.82, compared with using questionnaires and functional test scores alone. Deep learning models built using a dataset of 296 older adults achieved an accuracy similar to that of the logistic regression model. Aicha et al. (2018) highlighted the fact that deep learning models have the advantage of not requiring the implementation of feature extraction methods. On the other hand, deep learning models lack interpretability, which limits their application in medical contexts. The same authors (van Schooten et al., 2016) also demonstrated that the gait quality in daily life is "predictive for both time-to-first and time-to-second falls in both univariate and multivariate models" with adequate to good accuracy.

4.5 Overview

The fall risk factors that are underling the occurrence of a fall are multiple and range from personal intrinsic factors, such as gait and balance disorders, to external uncontrolled factors, such as weather conditions and home hazards. The research area of fall risk assessment, or fall prediction, focuses on developing strategies to assess the risk of falling of a person in order to trigger specific fall prevention strategies that aim to revert some of the fall risk factors. However, there is still a lack in a standard protocol, application guidelines, and frequency to apply fall risk assessment in the elderly population, either in clinical or community settings. Some efforts have been made in order to provide scales, questionnaires and functional tests to that end, however, a multifactorial fall risk assessment should combine multiple fall risk factors and integrate objective and measurable outcomes.

One of the most commonly reported limitations of fall risk assessments relates to the fall risk parameters being evaluated and the data sources used for feature extraction. The majority of the existing studies focusing on fall prediction are based on a single source of data, either clinical and self-reported or extracted from inertial sensors, instead of a combination of multiple data sources. Another challenge presented in previous studies is the small sample size of datasets. Considering prospective studies, with the exception of the work by van Schooten et al. (2016), the majority of existing studies are based on data collected from less than 150 participants. The collection of

larger datasets is time and resource consuming, whereas small-size datasets can impact the quality of the analysis and generalizations retrieved from that data. Moreover, the low incidence of falls (less than 30% in the older population) leads to unbalanced datasets which can negatively impact results. While there is also a lack of consensus regarding the output metric that should be used to divide population groups into fallers and non-fallers, the 1-year follow-up occurrence of falls has been pointed out by Howcroft et al. (2013) as the most reasonable metric.

In the next sections, we will introduce a multifactorial fall risk screening protocol and corresponding data collection and analysis, that we have applied to 403 participants. Only a part of the population aged over 65 was taken into consideration for analysis, resulting in a total of 281 participants. Based on the dataset collected we aim to:

- *Study signal processing methods applied to functional tests instrumentation:* we developed a signal processing methodology of the sensor data collected during the execution of functional tests, for segmentation of relevant time periods, and computation of metrics. This study will provide the basis for the application of machine learning algorithms for fall prediction based on features extracted from sensor data.

- *Analyse the added value of sensor data compared to tests scores:* we applied signal processing and feature extraction methods for several instrumented functional tests using inertial sensors and a pressure platform. We have used a fall level to divide population groups into fallers and non-fallers, based on previous falls history and the use of walking aid. We aim to study the predictive power of sensor-based features compared to functional test scores and personal self-reported data.

- *Study multimodal data fusion approaches for fall prediction:* we analyzed the added value of several data sources, including not only clinical and self-reported data but also information about functional capabilities, such as mobility, balance, and strength. These capabilities were obtained from inertial sensors and a pressure platform during the execution of a multifactorial screening protocol. This protocol combined the most relevant tests for assessing grip strength, balance, mobility and muscle strength. The multifactorial nature of the collected data provided an opportunity to compare models based on a single source with models based on data fusion. Three alternative approaches were explored for data fusion: an *early fusion approach*, that combines all data in a unified feature vector; a *late fusion approach* that combines the predictions of three classification pipelines trained with each of the individual data sources; and a *slow fusion approach* that fuses information from each data source individually in the first layers of a neural network. The record of the occurrence of falls over a 1-year period based on monthly follow-up phone calls was used to differentiate groups.

In both studies, we based our analysis on reported fall occurrences instead of automatic detection of falls. While there are tools for automatic detection of falls, such as personal emergency response systems (PERS), these have never been reported to be used during the follow-up period. Automatic fall detection systems based on wearable sensors will be studied in detail in Part III.

Chapter 5

Instrumented Timed Up and Go: Fall Risk Assessment based on Inertial Wearable Sensors

J. Silva and I. Sousa,
Published in Proceedings of IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2016, pp. 5.

Strategies for fall risk assessment are currently not multifactorial neither implemented as a regular assessment of health status in clinics or hospitals. The reason could be related to the lack of an easy to implement, complete and objective test to assess the elderly's fall risk level. More recently, inertial wearable sensors have been used in combination with standard tests to evaluate the performance of the person during each phase of the test in an objective way. This work proposes a methodology for collecting and analyzing the Timed-Up and Go (TUG) test instrumented with wearable inertial sensors. An automatic algorithm to segment the TUG test into three components was implemented prior to feature extraction. Overall, features from the walking and the first turning phase of the test could provide meaningful information to differentiate groups of high and low fall risk.

5.1 Methods

5.1.1 Participants

A group of 18 community-dwelling older adults has been invited and gave their informed consent to participate in this study. The group has an average age of 73± 5 years old, body mass index (BMI) of 26.7± 4 kg/m^2 and is composed of 5 males.

5.1.2 Protocol

The participants were asked about their previous falls and the number of falls reported in the previous 12 months was registered. A medical history questionnaire was also given to them to evaluate chronic conditions, audition and vision problems, other relevant medical conditions and exercise habits.

They were also evaluated using an adapted iconographic version of the Tinetti Falls Efficacy Scale (FES). A set of ten questions about the confidence level in performing daily life activities was presented using an illustration and a small phrase in Portuguese in a smartphone (Guimarães et al., 2013). The questions should be rated on a three-point Likert scale from 1 (very confident) to 3 (not confident). This scale is converted to a final score ranging between 0 and 100, where each question has a value of 10 final points and each point in the Likert scale has a value of 3.3 final points. Final scores higher than 70 are associated with a higher fear of falling (Tinetti et al., 1990).

POMA was also applied. This is a task-oriented test to assess gait and balance abilities, scoring each task by an ordinal scale from 0 (highest level of impairment) to 2 (independent) by means of observational judgment. The balance assessment was based on sit to stand transitions, turns and standing balance. The gait assessment was based on gait analysis in a straight 3-meter walk. The final score is a sum of the result for each task of the assessment and is ranged between 0 and 28. Scores lower than 19 are associated with high fall risk and scores higher than 25 are associated with low fall risk (Faber et al., 2006).

The standard TUG was also performed by the participants. They were asked to start sitting down on a chair and when the test starts, the person should stand up, walk straight for 3 meters at their normal pace, turn around, walk back to the chair and sit down. The final score of this test corresponds to the time needed to perform the TUG test. Scores higher than 14 seconds are considered associated with high fall risk (Shumway-cook et al., 2000).

An instrumented version of the TUG test was performed, using the inertial measurement unit of a smartphone, placed on the pocket or fixed at the waist or at the leg, and video records to annotate data during the test. In contrast to the traditional timed-up and go test, instead of keeping the walking distance of 3 meters fixed, the iTUG test had a fixed duration of 30 seconds. Auditory cues were used to instruct the person to stand up and walk forward in the first 15 seconds and then turn, walk back and sit down in the last 15 seconds. Accelerometry data were collected using a smartphone built-in 3-axial accelerometer, sampled at 200Hz. The gyroscope signal was

also recorded with a sampling rate of 200 Hz. Both signals were synchronized and were used to segment the TUG test into its several components.

5.1.3 Signal Segmentation

For the data collected during the iTUG test, manual and automatic segmentation of the accelerometer signal was performed to obtain three segments corresponding to three components of the TUG test: stand up, walk forward and turn around. Since it could happen that for some of the higher risk persons the given time was not enough to return to the chair, the second segment of walking and sitting down were not considered.

Manual segmentation was based on the visual inspection of the video recorded during the performance of the tests. The signal and video collected were synchronized and the timestamps of the transition between phases were registered.

Automatic segmentation is based on the integral of the gyroscope signal to identify the turning points. A turning was considered when an angle of 150^{o} was detected in the integral. To identify the duration of the turn, the previous minimum and the next maximum were used as start and end of a turn, Figure 5.1. In order to segment the sit-to-stand and stand-to-sit transitions, the angle with the gravity vector was calculated based on the accelerometer readings. Consecutive differences of 3 degrees in the angle signal were associated with transition phases. After segmented these phases, the in-between phases were considered as walking components.

5.1.4 iTUG Feature Extraction

For each segment, the duration and the number of steps were also manually retrieved from the accelerometer signal and confirmed by video records. The magnitude of the accelerometer signal was computed and several statistical features were retrieved, namely, number of times the magnitude signal crosses the mean value (MeanCrossCount), interquartile range (IQR), energy, entropy, standard deviation (Stdev), mean value, median deviation (MedianDev), root mean square (RMS), skewness and kurtosis. Maximum and minimum values were also considered as signal features and the average value of the minimum and the average value of the maximum (MinAvg and MaxAvg). The feature AvgPeak Height is defined as the difference between MaxAvg and MinAvg. Several features were extracted from the FFT of the magnitude of the accelerometer signal. The maximum amplitude (FFT Max Amp) and the second maximum amplitude (FFT 2^{nd} Max) in the spectrum were considered as features. The ratio (FFT Amp scale) and difference (FFT Amp dif) between these two features were also computed.

5.2 Results

In this section, the results of the standard fall risk assessment tests are detailed and also the results of the Timed-Up and Go test instrumented with inertial sensors. According to the results of the standard tests, the segmentation in two different groups of risk of falling is not straightforward and

so the metrics extracted from the instrumented TUG could help to better differentiate the groups of fall risk.

5.2.1 Standard Tests Results

Table 5.1: Standard tests results.

Participant	Previous Falls	FES	POMA	TUG time (sec)
1	1	13	26	14.5
2	0	27	25	15.7
3	5	100	9	40.5
4	6	67	23	13.6
5	1	87	7	41.2
6	1	0	24	9.4
7	0	20	20	19.9
8	3	23	25	9.2
9	0	0	26	12.3
10	2	20	24	10.6
11	0	30	24	19.0
12	0	17	26	13.0
13	0	20	26	18.0
14	0	60	20	14.0
15	0	0	26	14.0
16	0	10	26	10.4
17	1	0	26	10.0
18	0	0	27	7.0

State-of-the-art fall risk assessment tests and questionnaires were performed and the results are compiled in Table 5.1. The results of these standard fall risk assessment tools were used to segment the population in higher and lower fall risk subgroups. A higher fall risk group was defined with the persons that have a Tinetti POMA test result lower than 25 and TUG time equal or higher than 14 seconds. The persons identified with a higher risk of falling in this dataset are highlighted in grey in Table 5.1. This group is composed of 5 persons. The number of previous falls was not considered to divide the two groups because the number reported by the participants not always correspond to the true value, due to missing reports or confusion in the timeline of events.

Since participants 1, 2, 4, 6, 10 and 13 only satisfied one of the criteria, they were not grouped in the higher risk group.

Some statistics concerning other relevant fall risk factors were also computed. Only participants 2, 15, 16, 17 and 18 were male. For the higher risk group, only participants 3, 5 and 7 used assistive walking devices. Participants 4 and 5 had a knee and a hip prosthesis, respectively.

The incidence of some diseases in this population was: rheumatic diseases were presented in 11 of the 18 participants, 3 of them belong to the higher risk group. Chronic pain was reported by 14 of the 18 participants, 4 of them belong to the higher risk group. Most of the participants,

reported hypertension and dizziness in the previous year, 9 of the 18 participants reported both conditions, only participants 3, 14 and 15 reported dizziness but not hypertension and only participants 5, 6, 10 and 18 reported hypertension but not dizziness. For each condition, only 3 participants belong to the higher risk group. Polypharmacy was reported by 10 of the 18 participants, all the participants of the higher risk group reported polypharmacy.

When asked about physical activities during the week, 10 out of 18 participants practice any physical activity more than twice a week, only one of them belongs to the higher risk group.

5.2.2 Automatic Segmentation of iTUG phases

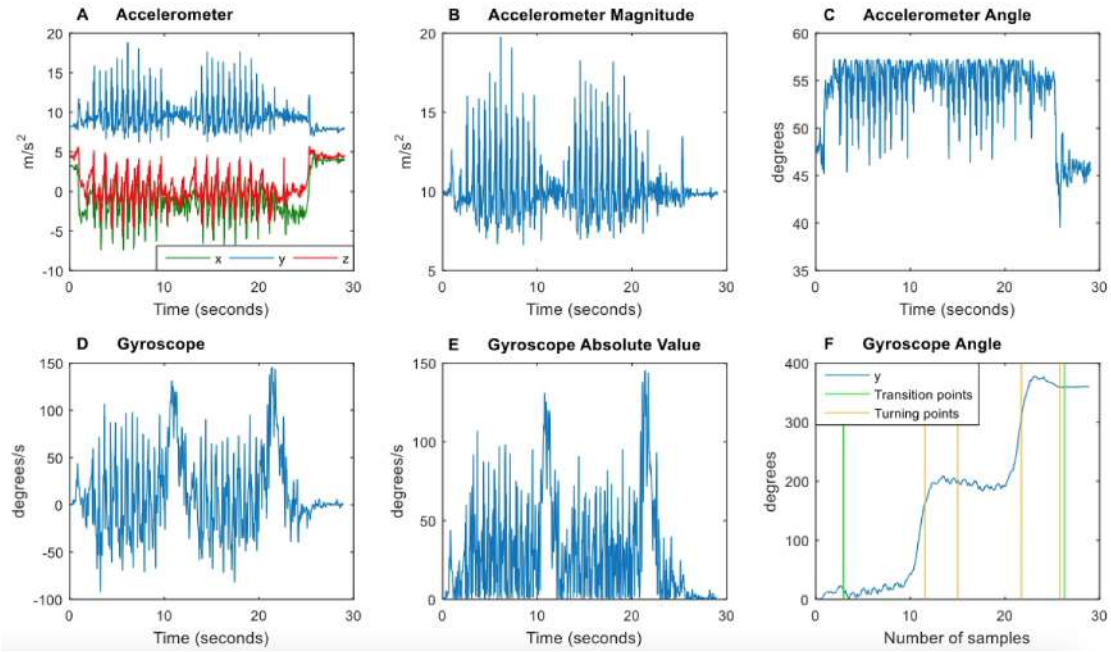

Figure 5.1: Example of automatic segmentation of TUG components for participant 1. Accelerometer signals were recorded for the three axes (x,y,z) (Figure 5.1-A). The magnitude of the accelerometer (Figure 5.1-B) was used to calculate the accelerometer angle (Figure 5.1-C) with the gravity vector. The signal of the gyroscope was only analyzed for the y-axis (Figure 5.1-D). The variations of the gyroscope absolute value were not distinguishable (Figure 5.1-E). However, the gyroscope angle (Figure 5.1-F) allowed a better identification of segments: green lines represent the transition points and red lines represent the turning points (start and end of turning).

The segmentation was done based on the smartphone built-in inertial sensors, accelerometer and gyroscope, as illustrated in Figure 5.1, for participant 1.

The transition phases at the beginning and at the end of the test are present in the accelerometer signal of the first row of images of Figure 5.1. The flat lines in Figure 5.1-A represent variations in the orientation of the accelerometer that occur when the person changes between sitting and standing positions. These variations are also observed in the accelerometer angle of Figure 5.1-C and marked with green lines in Figure 5.1-F. The accelerometer angle disputed in Figure 5.1-C is calculated with the magnitude signal of Figure 5.1-B and the gravity vector (0, 1, 0).

For the identification of the turning segments, the integration of the gyroscope y-axis signal was used (Salarian et al., 2010). As shown in Figure 5.1-D the two higher peaks of the signals represent the points when the person is turning. These variations correspond to absolute values of the gyroscope signal higher than 120 degrees per second. The detection was performed with the signal of Figure 5.1-F, the angle around the y-axis. After identifying the points with an angle equal to 150 degrees, the signal was filtered using a moving average with a window size of 100 samples. The start and end points of each turning phase were identified based on the previous minimum and next maximum around the point with an angle of 150 degrees. These points are marked in Figure 5.1-F with red lines.

5.2.3 iTUG Features

Table 5.2: TUG features for walking and turning segments.

Feature	Low Risk[a]	High Risk[a]	p-Value[b]
Walking Segment			
RMS	10.55± 0.35	10.20 ± 0.16	0.05
Stdev	2.32 ± 0.67	1.47 ± 0.40	0.02
MedianDev	1.21± 0.31	0.54 ± 0.30	0.00
IQR	2.52 ± 0.62	1.13 ± 0.66	0.00
Skewness	0.89 ± 0.53	2.42 ± 1.24	0.00
Kurtosis	1.97 ± 2.55	13.72 ± 9.61	0.00
MeanCrossCount	86.23 ± 23.62	135.80 ± 67.90	0.03
FFT Max Freq	1.67 ± 0.23	3.47 ± 2.86	0.03
FFT 2nd Max Freq	3.48 ± 1.50	5.53 ± 2.31	0.04
FFT Max Amp	1.61 ± 0.53	0.65 ± 0.36	0.00
FFT 2nd Max	0.76 ± 0.22	0.50 ± 0.24	0.04
FFT Amp scale	2.12 ± 0.51	1.26 ± 0.10	0.00
FFT Amp dif	0.84 ± 0.42	0.15 ± 0.12	0.00
Turning Segment			
Stdev	1.91 ± 0.93	0.80 ± 0.27	0.02
Median Dev	1.06 ± 0.76	0.31± 0.22	0.05
IQR	2.15 ± 1.50	0.66 ± 0.47	0.05
MinAvg	6.72 ± 0.94	8.22 ± 0.24	0.00
MaxAvg	13.97 ± 2.10	11.91± 0.53	0.05
AvgPeak Height	7.25 ± 2.83	3.70 ± 0.74	0.02
Energy	59803.05 ± 23109.31	98759.52 ± 33372.40	0.01
Entropy	8.81 ± 0.92	9.83 ± 0.51	0.03
Kurtosis	2.50 ± 3.10	10.54 ± 8.98	0.01
MeanCrossCount	27.69 ± 13.01	73.20 ± 41.25	0.00
FFT 2nd Max	0.70 ± 0.30	0.27 ± 0.11	0.01

[a]Mean ± standard deviation; [b]significance level of 5%

After extracting the features from the first three components of the test, only the ones belonging to the component of the first walking phase and the first turning phase showed significant

differences between the mean of high risk group and the mean of low risk group. Applying the t-test with a significance level of 5% the most relevant features are presented in Table 5.2 divided per test segment. As expected, a person with a higher risk of falling will walk slowly and the foot impacts will be lower than a person with a lower risk of falling. This behavior is explained with the features RMS and Stdev that will be lower for a higher risk group than for a lower risk group. Also, the MedianDev and the IQR were expected to be lower for a higher risk group. Conversely, during the turning phase, a person with mobility disabilities will have more difficulties to turn and will for consequence take more steps to turn than a lower risk person. Features energy and entropy are higher for the higher risk group as a consequence of the higher movement during turns.

5.3 Discussion

Despite the fact that the standard tests as FES, POMA and TUG could differentiate the two groups with a significance level of 5% (p-values are 0.0005 (FES); 0.0002 (POMA) and 0.0008 (TUG)), sometimes is difficult to evaluate a person based only on these three tests. As for example, there are participants that might be categorized in the higher risk group but the values of the standard tests are very similar to the lower risk participants. The number of previous falls did not have a significant difference between higher and lower risk groups (p-value of 0.576) and so this factor was not used to differentiate the groups. In accordance with the findings of Salarian et al. (2010) the standing up component of iTUG did not reveal significant differences between the two fall risk groups. However, the walking and turning components were the most significant ones in terms of features that could be useful to differentiate between higher risk and lower risk person.

In a previous study (Guimarães et al., 2014), walking features were extracted from inertial sensors signals and have shown to be well correlated with those obtained from a kinematic evaluation using high-speed cameras. The features extracted were different from the ones computed in the present study, since they were matched to the outputs of the usual kinematic evaluations. Some examples of such features are gait cadence and speed, pelvic sway, asymmetry of step duration and length. While those features may better match human observation, be more physiologically meaningful and easier to interpret, a number of approximations and assumptions are required to compute them from inertial sensors signal. These assumptions often introduce inaccuracies and errors that may degrade the quality of the fall risk assessment outcome. The features computed in the present study show significant differences between the higher and the lower fall risk groups. The feature extraction process should be implemented for a larger dataset with different fall risk profiles in order to identify the best set of features that could differentiate lower and higher risk persons with very different profiles and different combinations of fall risk factors. The final set of features would then be used to create a model in order to distinguish higher and lower fall risk groups.

This study indicates that the iTUG is a viable tool for fall risk assessment, with the potential to be implemented in clinical or hospital environments. The test is quick and the instrumentation is easy and does not require any specialized technician to perform it.

Chapter 6

Comparing Machine Learning Approaches for Fall Risk Assessment

J. Silva, J. Madureira, C. Tonelo, D. Baltazar, C. Silva, A. Martins, C. Alcobia and I. Sousa, Published in Proceedings of BIOSIGNALS 2017 - 10th International Conference on Bio-inspired Systems and Signal Processing.

Traditional fall risk assessment tests are based on timing certain physical tasks, such as the timed up and go test, counting the number of repetitions in a certain time-frame, as the 30-second sit-to-stand or observation such as the 4-stage balance test. A systematic comparison of multi-factorial assessment tools and their instrumentation for fall risk classification based on machine learning approaches were studied for a population of 296 community-dwelling older persons aged above 50 years old. Using features from inertial sensors and a pressure platform by opposition to using solely the tests scores and personal metrics increased the F-Score of Naïve Bayes classifier from 72.85% to 92.61%. Functional abilities revealed a higher association with fall level than personal conditions such as gender, age and health conditions.

6.1 Methods

6.1.1 Subjects

A total of 296 subjects voluntarily participated in the study. Informed consents were obtained from all participants who responded to personal information, health, previous falls inquiries and completed the three instrumented assessment tests: TUG, STS and 4-stage. The data collection took place in different environments, mostly at community (76.0%), at day-care centers (15.9%), and at nursing homes (8.1%).

Demographic and anthropometric information was annotated for all the subjects along with health-related information from two questionnaires: health conditions and medication intake. Fall-related information was inquired using a history of falls questionnaire.

The mean age of the sample was 70.2 years (93 persons with age below 65 years), the majority of the subjects were women (68.2%), 25.0% lived alone, 51.0% only have primary education and 11.5% use an assistive device. Diabetes was the most prevalent health condition (15.5%), followed by osteoarthritis (14.2%) and osteoporosis (10.8%).

Urinary incontinence was reported by 22.3% (by answering the question: do you leak urine when you cough, laugh, sneeze or lift an object?); fear of falling was reported by 47.0% (by answering the question: are you afraid of falling?); 57.4% of the persons referred to intake 4 or more different medicines per day (mean was 4.52 medicines).

During the previous year, 30.7% of the persons have fallen (18.9% outdoors) and 8.1% underwent to the emergency service (hospital). The wrist/hand fracture was the most common injury (2.4%) among these fallers.

6.1.2 Screening Protocol

This section describes the fall risk assessment tests applied in this study:

Timed Up and Go Test (TUG) fast pace: the person is asked to start seated on a chair and when the test starts, the person should stand up, walk straight for 3 meters, as fast as the person can, turn around, walk back to the chair and sit down (Beauchet et al., 2011). Test score corresponds to the time needed to perform TUG test (*TUG duration*). A threshold of 10s has been found to be associated with falls occurrence in 12 months follow up period for community-dwelling older adults (Rose et al., 2002).

30 Seconds Sit-to-stand Test (STS): the person is instructed to sit on a chair and repeatedly stand up and sit down as many times as possible over 30 seconds (Jones et al., 1999). The person must be seated in the middle of the chair, feet should approximately width apart and placed on the floor, and arms crossed by the wrists placed against the chest. The final score of this test is the number of times the person completes a cycle of sit-to-stand and stand-to-sit (*number of STS cycles*). While normative levels are dependent on age and sex (Rikli and Jones, 2010), a score of fewer than 15 transitions in the 30 seconds test duration has been used to identify "fallers" in a group of elderlies (Cho et al., 2012).

4 Stage Balance Test "*modified*"*:* the person is instructed to progressively maintain four-foot positions for 10 seconds each, without moving his/her feet or needing support. The foot positions are side by side stance, semi-tandem stance, tandem stance, and unipedal stance (Rossiter-Fornoff et al., 1995; Thomas et al., 2014). For each position, the subjects were instructed to stand quietly without shoes on the pressure platform, with their arms along the body. In this study, except for the one-leg stand position (unipedal stance), all positions must be performed with eyes open and then closed. The final score of this test is the number of positions a person can hold for 10 seconds without losing balance (*number of 4-stage exercises*). The inability to complete the tandem stance position has been associated with a higher risk of falling (Murphy et al., 2003).

The tests were applied by trained health professionals. Prior to the execution of tests, the test procedure was explained to each person and it was demonstrated how the test should be performed. Auditory cues were also used to instruct the person during the execution of the tests. Only persons who performed the three functional tests (TUG, STS, and 4-stage) were included in this study.

6.1.3 Instrumentation

The participants were instrumented with one wearable inertial sensor during the execution of TUG and 30-seconds sit-to-stand tests. The 4-stage balance test was performed on a pressure platform, as can be seen in Figure 6.1.

The wearable sensor was developed and assembled at Fraunhofer AICOS and was placed at the lower back. Inertial data was collected using the built-in 3-axial accelerometer and 3-axis gyroscope, both sampled at 50 Hz. Raw data from the accelerometer sensor was acquired for all the tests in m/s^2, and in rad/second for the gyroscope sensor.

The pressure distribution data was measured with PhysioSensing platform (Sensing Future Technologies, Lda) running at a frequency of 50Hz. It contains 1600 pressure sensors of size 10mm by 10mm with a maximum value of 100N/sensor. Voltage data is converted with an 8-bit A/D converter and is transmitted via USB (Universal Serial Bus). In this way, it is possible to receive raw data of each pressure sensor as well as the raw center of pressure coordinates (CoP), in cm. In order to obtain more precision in CoP displacements, an algorithm was employed to obtain CoP positions in mm, using the matrix of pressure sensors (Hsi, 2016).

6.1.4 Inertial Sensors Data Analysis

The accelerometer and gyroscope signals were synchronized and used to segment the TUG test into its several components (stand up, walk forward, turn around, walk back to the chair and sit down) as previously described in (Silva and Sousa, 2016) and to identify the stand and sit phases of the STS test. The identification of the STS transition points was made analyzing the y-axis of the gyroscope signal. After filtering the signal with a moving average filter of 20 samples window size, zero crossings were identified (Guimarães et al., 2014). In order to remove outliers, a minimum of 20 samples was used as the difference between consecutive transition points. Since the score is given by the total number of complete cycles, it was considered one cycle between

Figure 6.1: Example of a test set-up, with the pressure platform in the floor and an illustration of the inertial sensor placement of at the lower back, since it is covered by the clothes.

two transitions points, one sit-to-stand and one stand-to-sit. The number of cycles is, therefore, half the number of transitions points identified, as illustrated in Figure 6.2.

For each one of the TUG segments and for the whole STS test, statistical and frequency domain features were extracted from the magnitude of the accelerometer signal. The list of features has been reported in (Silva and Sousa, 2016) and corresponds to *mean, median, maximum, minimum, signal height, standard deviation, median deviation, root mean square, interquartile range, number of times the magnitude signal crosses the mean value, energy, entropy, skewness, kurtosis, average of minima, average of maxima, average signal height, fundamental harmonic of Fast Fourier Transform (FFT) spectrum and fundamental amplitude.*

Additional metrics for each test were calculated from the inertial data: for the TUG test, the *duration of the stand* segment (duration of the first segment) and the *number of steps* (calculated with a step counter algorithm reported by Aguiar et al. (2014b)) taken during the test; for the STS test, the *number of STS cycles* and the *STS power* (Zhang et al., 2014).

6.1.5 Pressure Platform Data Analysis

For each posture of the 4-stage balance test executed, the pressure values on each sensor of the pressure platform were recorded. The CoP coordinates were then obtained and several parameters, which are typically used in postural sway and fall risk assessment (Bigelow and Berme, 2011; Guimarães et al., 2014; Raymakers et al., 2005) were calculated.

For all the medio-lateral (ML) and antero-posterior (AP) CoP position coordinates obtained during each posture execution, the mean (*mean AP CoP positions, mean ML CoP positions*), standard deviation (*std AP CoP positions, std ML CoP positions*), root mean square (*rms AP CoP positions, rms ML CoP positions*), maximum (max *AP CoP positions, max ML CoP positions*) and minimum (min *AP CoP positions, min ML CoP positions*) were calculated.

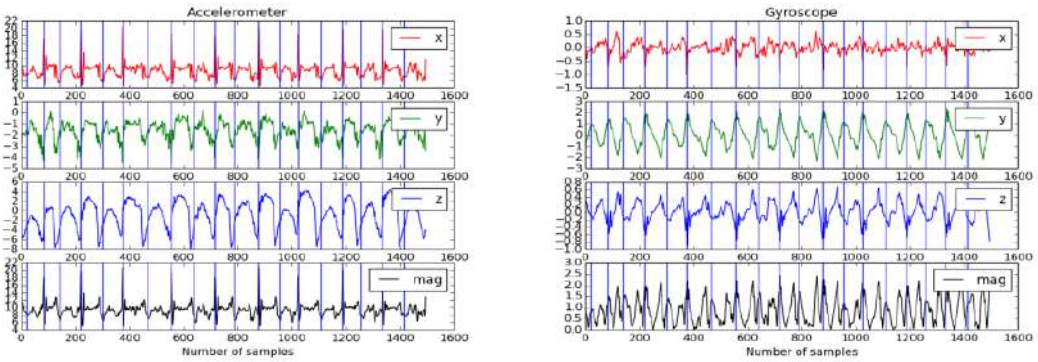

Figure 6.2: Axis x (red), y (green), z (blue) and magnitude signals (black) of the accelerometer and gyroscope signals for STS test with identification of transition points with blue vertical lines. The interval between two consecutive lines is considered as one STS cycle. Figures are from a low risk person.

The displacement of CoP in each direction per time unit gave rise to the mean velocity of CoP displacement (*vm CoP position AP, vm CoP position ML*) metrics.

Another metric extracted was the area of a confidence ellipse containing 95% of the CoP coordinates projected in a 2D plan (*Ellipse area*). Figure 6.3 shows a comparison of CoP displacements in ML and AP directions for two persons with different fall risk levels during the semi-tandem stance with eyes closed. For a low fall risk person (left figure) the displacement is concentrated around the center, however, for a high fall risk person, more outliers in ML and especially in AP direction are identified, reflecting unbalance situations.

Sway can be defined, in this scope, as the amplitude or absolute distance of CoP oscillations. The sum of all the distances accumulated during the execution of each posture is computed resulting in the CoP path length (*total Sway distance*). The standard deviation of sway distances (*std Sway*) and the maximum and minimum amplitude of CoP oscillations (*maxSway* and *minSway*) were also included as pressure platform metrics.

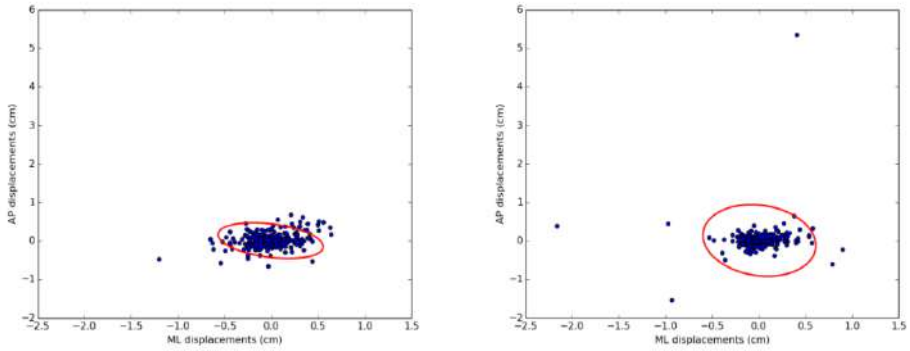

Figure 6.3: CoP displacements in ML and AP directions and 95% confidence ellipse area (red line) during semi-tandem stance with eyes closed of 4-stage test. Left figure is from a low risk person and right figure is from a high risk person, showing more outliers in ML and AP directions.

6.1.6 Machine Learning Methods

Classification and regression methods were tested to differentiate between high and low fall risk groups using metrics extracted from inertial sensors and pressure platform. Rapid Miner Toolkit was used for the train and test processes. Ten-fold cross-validation with a random split was used for all the processes. In order to define a metric to divide the groups, a *fall level* was determined based on the history of falls questionnaire and usage of walking aid, as presented in Figure 6.4, since these two factors have evidence to be more related with the risk of falling. The *fall level* is merely an indication if the person shows more or less probability of falling since the occurrence of the fall in 12-months follow-up period was not possible to measure. The dimension of the population is 296 subjects. The low risk group represents 83% of the dataset and is composed of 245 subjects (35% within 50-65y.o. and 65% above 65y.o.). The high risk group represents 17% of the dataset and contains the remaining 51 subjects (16% within 50-65y.o. and 84% above 65y.o.). This distribution is in agreement with the fall incidence in the elderly population, which is less than 30% (Bergen et al., 2016).

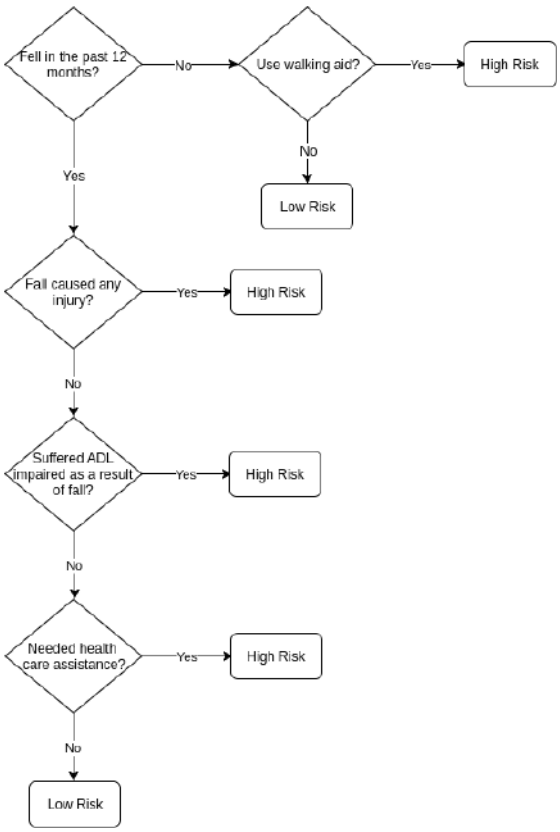

Figure 6.4: Fall level definition based on history of falls and usage of walking aid.

Two approaches were compared: first, only personal metrics and test scores were used to construct the feature vector, and then this vector was replaced with features extracted from inertial sensors and pressure platform. The objective was to study the added value of the sensors' features to differentiate between fall risk groups.

The performance of several classification and regression methods was compared based on accuracy, precision, recall and F-Score. It was considered low risk as the positive class and high risk as the negative class. TP states for true positive, FP for false positive, TN for true negative and FN for false negative. The performance metrics are calculated as follows:

$$\text{Precision (Prec)} = TP / (TP+FP) \quad (1)$$
$$\text{Recall (R)} = TP / (TP+FN) \quad (2)$$
$$\text{Accuracy (Acc)} = (TP+TN) / (TP+TN+FP+FN) \quad (3)$$
$$\text{F1-Score (F1)} = (2P \times R) / (P+R) \quad (4)$$

6.2 Results

6.2.1 Statistical Analysis

A statistical analysis has been conducted for the variables: gender, age, body mass index (BMI), number of medicines, number of health conditions, fear of falling, TUG score, STS score, and 4-stage score. Cut-off values that have been used in previous studies referred to in section 6.1.2 to distinguish high and low fall risk levels were applied to each one of these variables. The Fisher's exact test was applied with the null hypothesis that there are no non-random associations between the two categorical variables: fall level and each one of the variables considered. The Fisher's exact test *p*-value and odds ratio (OR) are reported in Table 6.1 and were calculated with Matlab function *fishertest*.

Table 6.1: Odds Ratio and Fisher's exact test p-value for personal metrics and tests scores with the fall level.

Variable	**Odds Ratio**	**p-value**
Feminine Gender	1.04	1.00
Age >65	2.86	0.01
BMI <13.7 or BMI >29.7	1.58	0.18
More than 4 Medicines	1.96	0.05
More than 2 Health Conditions	1.56	0.38
Has Fear of Fall	3.35	0.00
TUG Duration >10 s	6.51	0.00
STS Cycles <15	11.25	0.00
Not completed 10s Tandem Stance (eyes open)	3.59	0.00

Presence of fear of falling, TUG duration above 10 seconds, number of STS cycles below 15 and not completed the tandem stance with eyes open were the metrics with higher odds ratio with the fall level and *p*-value below 0.05. Thus, the hypothesis of a random association between fall level and the variables in shaded lines of Table 6.1 can be rejected. Age above 65 years old and take more than 4 medicines per day also showed a *p*-value below 0.05 but the OR was lower than for the previously mentioned variables. For the remaining variables, the conclusion is that female individuals, or individuals that have BMI lower than 13.7 or higher than 29.7 or that have more

than two health conditions do not have greater odds of having a high fall level than individuals that are male, have a normal BMI and have less than two health conditions. In general, test scores showed a higher association with fall level than personal metrics, reflecting that functional abilities have a higher impact on fall level than personal conditions of a person.

6.2.2 Machine Learning Approaches

Classification and regression methods were studied for the differentiation between low and high fall risk groups using the fall level as the label. All algorithms applied were retrieved from the Rapid Miner predictive models.

6.2.2.1 Functional tests scores

As a first analysis, personal metrics (age, gender, BMI, fear of fall, number of health conditions and number of medicines) and test scores (TUG duration, number of STS cycles and number of 4-stage exercises) were used to define the feature vector and fall level was used as the label. The results are summarized in Table 6.2.

Table 6.2: Classification and regression results for personal metrics and functional tests scores. Accuracy, precision, recall and F-Score are in percentage (%).

Algorithm	Accuracy	Precision	Recall	F-Score
k-NN, k=4	81.41	69.33	63.00	66.01
Naïve Bayes	84.82	74.58	71.19	72.85
Random Forest	83.13	59.37	53.05	56.03
Decision Tree	81.44	68.28	60.33	64.06
Neural Net	82.45	69.22	64.84	66.96
SVM	82.45	49.08	51.21	50.12
Linear Regression	83.11	69.01	56.05	61.86
Logistic Regression	82.13	67.48	64.88	66.15

Naïve Bayes classifier obtained higher accuracy, 84.82%. Precision was 74.58% and recall was 71.19%. Random Forest and Linear Regression also obtained acceptable results. In general, all algorithms showed higher precision than recall.

6.2.2.2 Sensors features

In order to compare the previous results based on tests scores with the features extracted from inertial sensors and pressure platform, a feature vector containing 224 sensors features was used. For each TUG segment (stand, walk, turn and walk back) 19 statistical and frequency domain features were extracted, yielding 76 features plus 2 metrics, time to stand and the number of steps. For the STS test, the same 19 features were extracted plus 2 metrics, the number of STS cycles and the STS power. For the 4-stage test, 17 CoP metrics were extracted for each one of the 7 exercises (when available), yielding 119 features. Additionally, 6 personal metrics were added: age, gender,

BMI, fear of fall, number of health conditions and number of medicines. The fall level was used as the label. Since the number of features was considerably high, forward feature selection was applied prior to cross-validation. Results are presented in Table 6.3.

Table 6.3: Classification and regression results for personal metrics and features extracted from sensors. Number of features selected by forward feature selection follows the name of the algorithm. Accuracy, precision, recall and F-Score are in percentage (%)

Algor.	**Accuracy**	**Precision**	**Recall**	**F-Score**
k-NN, k=4 [5F.]	85.78	87.79	95.88	91.66
Naïve Bayes [4F.]	87.16	88.18	97.50	92.61
Neural Net [5.]	87.20	88.05	97.94	92.73
SVM [3F.]	84.82	84.95	99.23	91.54
Random Forest [3F.]	87.48	87.92	98.43	92.88
Decision Tree [5F.]	88.17	89.47	97.10	93.13
Linear Regression [3F.]	85.89	85.66	99.55	92.08
Logistic Regression [4F.]	86.54	86.74	98.78	92.37

The decision tree classifier obtained higher accuracy, 88.17%. The precision was 89.47% and the recall was 97.10%. Comparing the results of Naïve Bayes with the previous analysis, the features obtained from sensors yield higher accuracy than only test scores. Moreover, features from TUG and 4-stage tests were frequently selected with a forward feature selection method. For all algorithms tested, using features from sensors provide higher precision and recall values. The F-Score obtained with features from sensors was the same across all algorithms tested and considerably higher than the F-Score obtained only with tests scores and personal metrics (91-93% against 50-72%).

6.3 Discussion and Conclusion

Previous studies from Scott et al. (2007) have compared the accuracy of several functional tests and fall risk tools to differentiate groups with different levels of fall risk. Despite the differences in protocol and population analyzed (only for community settings and validated in a prospective study), similar accuracy and sensitivity were reported. Murphy et al. (2003) concluded that 'floor transfer' and '50 ft walk' tests combined can discriminate fallers from non-fallers with an overall accuracy of 96% (82% sensitivity and 100% specificity).

A similar study from Liu et al. (2011) has used metrics from instrumented TUG, alternate step test and 5 times STS to classify between fallers and non-fallers and the best models have achieved 70% accuracy (68% sensitivity and 73% specificity).

The objective of this study was to compare the performance of functional test scores and features obtained from inertial sensors and pressure platforms to discriminate between groups of low and high risk of fall. A fall level was defined based on the history of falls and usage of walking aid and was used as label in classification and regression algorithms. Only subjects who performed the three functional tests (TUG, STS, and 4-stage) were included in this study.

The association between functional test scores and fear of falling with fall level are not random (Fisher's exact test p-value < 0.05), concluding that individuals with functional disabilities and fear of falling have greater odds of having a higher fall level than individuals without physical disabilities and without fear of falling. Moreover, when comparing personal metrics with fall level, it was concluded for some personal metrics that random association with fall level cannot be excluded.

The differentiation power of personal metrics and test scores was considerably different when tested with classification and regression methods. Accuracies above 80% were obtained for all algorithms. Naïve Bayes outperforms with an accuracy of 84.82% (74.58% of precision and 71.19% of recall).

However, features from inertial sensors and pressure platform obtained better results for the same algorithms than only test scores. Naïve Bayes classifier obtained an accuracy of 87.16% (88.18% of precision and 97.50% of recall).

These results support the conclusion that instrumentation of fall risk assessment tests with inertial sensors and pressure platform could better discriminate the individuals at a higher risk of falling.

Chapter 7

Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls

J. Silva, I. Sousa and J. Cardoso,
Published in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, pp. 50-56, Jan. 2020

Falls are among the frequent causes of the loss of mobility and independence in the elderly population. Given the global population aging, new strategies for predicting falls are required to reduce the number of their occurrences. In this study, a multifactorial screening protocol was applied to 281 community-dwelling adults aged over 65, and their 12-month prospective falls were annotated. Clinical and self-reported data, along with data from instrumented functional tests, involving inertial sensors and a pressure platform, were fused using early, late, and slow fusion approaches. For the early and late fusion, a classification pipeline was designed employing stratified sampling for the generation of the training and test sets. Grid search with cross-validation was used to optimize a set of feature selectors and classifiers. According to the slow fusion approach, each data source was mixed in the middle layers of a multilayer perceptron. The three studied fusion approaches yielded similar results for the majority of the metrics. However, if recall is considered to be more important than specificity, then the result of the late fusion approach providing a recall of 78.6% is better compared with the results achieved by the other two approaches.

The main contributions of this study are the following: i) the use of multimodal data collected according to a multifactorial screening protocol for predicting falls; ii) the richness of the collected data allowing to infer not only functional capabilities of a person but also clinical and environmental information; and iii) the exploration of different fusion approaches (Figure 7.1).

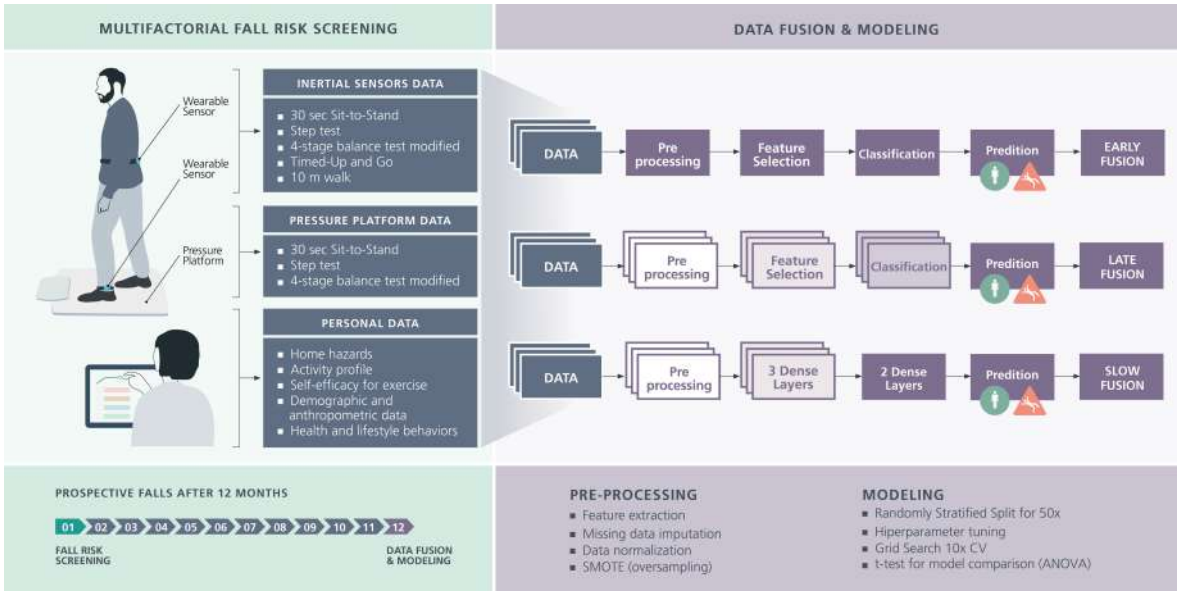

Figure 7.1: Graphical representation of the main contributions of the study: multifactorial fall risk screening, data fusion and modeling.

7.1 Methodology

7.1.1 Data collection

7.1.1.1 Subjects

Four hundred and three Portuguese community-dwelling adults aged over 50 (mean age of 69.69 $\pm$ 10.31; 70% women) were recruited from parish councils, physical therapy clinics, senior's universities, and other community facilities. The inclusion criterion was the ability to independently stand and walk with or without walking aids. The excluding criterion was the presence of severe sensory (deafness or blindness) or cognitive impairments (Martins et al., 2018). Only adults aged over 65 were considered for the analysis given that many previous studies used this age as a threshold for patient recruitment. The sample used in this study consisted of 281 subjects. The research was approved by the Ethics Committee at the Polytechnic Institute of Coimbra (N^{o}6/2017). All participants gave their written informed consent before the data collection in accordance with the principles of the Declaration of Helsinki (Martins et al., 2018).

7.1.1.2 Protocol

A multifactorial screening protocol for assessing the risk of falls in community-dwelling adults was defined based on relevant literature. The protocol included demographic and anthropometric data; lifestyle and health behavior data; six functional tests (*handgrip strength test*, *TUG test, 30s STS*, Step test (*Step*), "modified" 4-Stage Balance test (*4Stage*), and 10-m walking speed test (*10 Meter Walk*) instrumented with inertial sensors and a pressure platform); and questionnaires about environmental home hazards, activity and participation profile related to mobility, and self-efficacy to exercise (Martins et al., 2018).

7.1.1.3 Data sources

Several types of data were collected:

- *Clinical data,* including demographic, anthropometric and data such as place of residence, age, sex, medical conditions, and medications taken, as well as functional tests outcomes, such as test timing, number of repetitions, and grip strength.

- *Self-reported data* from questionnaires, such as home hazards, previous number of falls, and fear of falling;

- *Three-dimensional (3D) time series* extracted from the 3D accelerometer and 3D gyroscope used in the functional tests, including the time to stand and average acceleration along x, y, and z axes.

- *Two-dimensional (2D) time series* extracted from the pressure platform used in the functional tests, including the center of pressure oscillation in the mediolateral and anteroposterior directions.

Clinical and self-reported data were combined to form one data source named the *personal data.*

7.1.1.4 Prospective falls after 12 months

The participants were followed for 12 months via monthly phone calls. "The rate of falls was recorded from the day of inclusion until voluntary dropout, loss of phone contact or the end of the follow-up period" (Martins et al., 2018). The participants who reported at least one fall in the 12-month follow-up period were categorized as fallers, whereas those who did not report any falls during this period were categorized as non-fallers. The incidence of fallers in the study sample was 26.3%, which is in accordance with the literature reporting that approximately one-third of people over 65 will fall each year (Howcroft et al., 2013).

Table 7.1: Features extracted from clinical reports, self-reported, inertial, and pressure platform data.

Source	Features extracted
Clinical data Martins et al. (2018)	**sex**, age, **height**, weight, dwelling place, **benzodiazepines**, **antidepressants**, anti-psychotic, anti-inflammatory analgesics, anti-hypertensive, **total medication**,**+ 4 medicines daily**, STS score, TUG score, 4Stage score, Step score, 10m-Walk score
Self-reported data Martins et al. (2018)	**retrospective falls**, prospective falls, fear of falling, **live alone**, sedentary lifestyle, assistive device, **upper extremities assistance to stand**, **home risks**, not applicable home risks, items answered, risk home entrance, risk stairs out, risk stairs in, risk living areas, risk kitchen, **risk bathroom**, **risk bedroom**, risk outdoor, **index of home risk**, **index of home risk percentage**, self-efficacy score
Inertial sensors Silva et al. (2017)	mean, **median**, max, min, rms, std dev, median dev, **iqr**, min avg, max avg, **peak height**, avg peak height, mean cross count, fft max freq, **fft max amp**, **energy**, entropy, skewness, kurtosis, walking steps, walking variability, walking speed, STS power, time to stand
Pressure platform Silva et al. (2017)	sway velocity, **sway range**, sum oscillation, std oscillation, **area ellipse**, transfer time, **left foot force**, **right foot force**, left foot higher pressure zone, right foot higher pressure zone, **rising index**, weight symmetry

max: maximum, min: minimum, rms: root mean square, std dev: standard deviation, iqr: interquartile range, avg: average, fft: fast fourier transform, freq: frequency, amp: amplitude.

7.1.2 Feature extraction

During the walking tests (i.e., TUG and 10 Meter Walk), two wearable inertial sensors (AICOS, 2016), were placed on the lower back and ankle of the support leg. The sensors were sampled at 50 Hz. For the static tests (i.e., STS Step, and 4Stage) the PhysioSensing pressure platform (Sensing Future Technologies, 2018) sampled at 50 Hz was used in addition to the two inertial sensors. The hand-grip strength was assessed using a Jamar hydraulic hand dynamometer (Martins et al., 2018). Each functional test was divided into phases, e.g., 4Stage was divided into seven balance positions. Several features, as detailed in Table 7.1, were extracted from the four sources of data, i.e., clinical, self-reported, inertial sensor, and pressure platform data. Overall, 230 features were extracted.

7.1.2.1 Inertial sensors

An analysis and segmentation of the TUG test involving inertial sensors were previously reported by Silva and Sousa (2016). Later on, Silva et al. (2017) presented an analysis of the TUG, STS, and 4Stage tests performed with inertial sensors and a pressure platform, reporting a feature extraction process for both types of sensors. We adopted the analysis procedures reported in these studies. Our analysis of the 10 Meter Walk test was based on a previous work by Aguiar et al. (2014b).

The inertial features presented in Table 7.1 were extracted from the magnitude of the accelerometer signal.

7.1.2.2 Pressure platform

For the Step test, the number of steps was segmented based on the information provided by the pressure platform when a subject raised the leg. If a variation in the number of active cells was detected compared with the initial bipodal position, a step was identified. As the leg was lowered toward the pressure platform, the number of active cells increased, and the end of the segmentation phase was reached. The pressure platform features extracted for the Step and STS tests were the same as previously described for the STS test (Silva et al., 2017).

7.1.3 Classification pipeline

7.1.3.1 Data profiling

First, nominal data, such as therapist id, patient id, local id, were removed from the feature vector, and the remaining variables were converted to numerical values. The clinical and self-reported data were converted to numerical values using categorical/dichotomous variables when appropriate. Then, data profiling was performed, and the features with a correlation coefficient above 0.90 were removed. A statistical description of the database was achieved by depicting grouping variables such as the last year falls, follow-up falls, need of walking aid, and need of assistance to stand up in scatter and box plots. We performed an independent samples t-test on each grouping variable, with 95% confidence level. A segmentation of the database for each type of the data source, i.e., personal data (comprising the clinical and self-reported data), inertial sensor data, and pressure platform data, was also considered for testing different fusion approaches.

7.1.3.2 Feature preprocessing

Several feature preprocessing methods were employed, mainly for dealing with missing values. As 28 participants were unable to perform at least one of the functional tests due to physical limitations, the data from these tests had missing values. Moreover, missing values were present when participants were unable to reach the last positions among the seven balance positions of the 4Stage test. The last position of the 4Stage test had 80% missing values. The missing values for the remaining features accounted on average for $5.8 \pm 11.8\%$ of all values. Due to time constraints during the data collection, the database also contained some missing answers for participants who filled in the questionnaire. Since the inability to accomplish a functional test could yield valuable information related to functional capabilities, the missing values in such cases were replaced by zero. Removing the participants with missing values would have resulted in a significant reduction of the sample dimension, preventing from accurately representing the target population. All features were normalized by removing the mean and scaling to the unit variance.

7.1.3.3 Data fusion approaches

Three approaches to data fusion could be considered using the collected dataset: 1) *data-level fusion*, i.e., combining the data obtained from the inertial sensors and pressure platform to extract features resulting from the joint analysis of both signals; 2) *feature-level fusion*, i.e., extracting features from the three data sources separately and combining all features in the same feature vector; and 3) *decision-level fusion*, i.e., training a model for each data source and combining the predictions of all models. In our study, we experimented with three different data fusion approaches. The first approach, called *early fusion*, involves fusing data after the feature extraction stage and before the classification stage (i.e., feature-level fusion). The second approach, called *late fusion*, involves fusing data after the classification stage (i.e., decision-level fusion). Finally, the third approach, called *slow fusion* is based on the combination of the first two approaches. In particular, it gradually fuses multisource information in a multilayer perceptron (MLP), in such a way that higher layers of the network are provided with progressively more information (Karpathya et al., 2014).

The *late fusion approach* uses the majority voting mechanism, where the predicted class label for a specific instance is assigned based on the class label predicted by the majority of individual classifiers. The *slow fusion approach* combines the information of each data source in the middle layers of a neural network (Fig. 7.2). For the implementation of the *slow fusion approach*, we employed the Keras library to train a multi-input sequential model, receiving three data sources in a single network. For each data source, we combined three feedforward fully connected (dense) layers with the *ReLu* activation function, intercalated with dropout layers, and with a sequential decrease in the number of layers' nodes. The last layers of each model were concatenated in a stack of two dense layers with *sigmoid* activation. This model was optimized using *binary cross entropy* loss and *Adam* optimization.

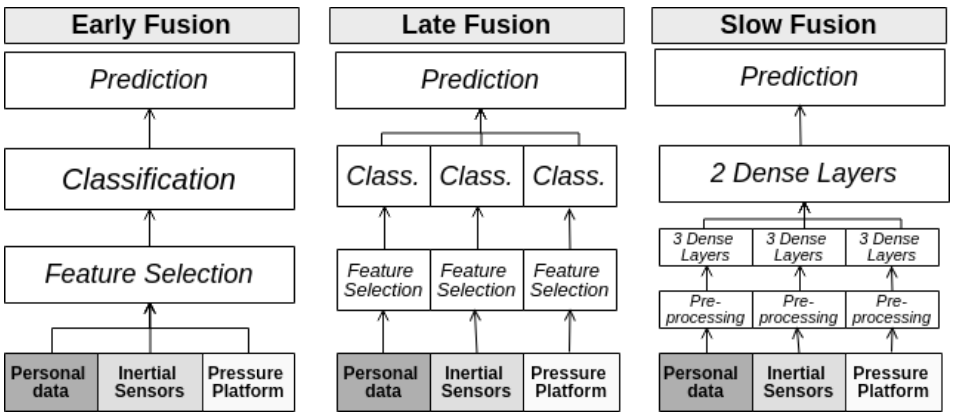

Figure 7.2: Early, late, and slow fusion approaches for combining personal, inertial sensor, and pressure platform data, for fall prediction.

7.1.3.4 Classification pipeline

A randomly stratified train-test split was performed 50 times to ensure the variability between train and test splits, with 33% of the data being selected for the test set. Each split yielded a training set of 188 samples (138 non-fallers, 50 fallers) and a test set of 93 samples (69 non-fallers, 24 fallers). Using a grid search with cross-validation (CV) over the training set, a classification pipeline was defined to optimize a range of parameters for the three stages: feature selection, classification, and grid search scoring (Fig. 7.3).

For the feature selection, we optimized the number of components for principal component analysis (PCA) and the threshold for the variance threshold method. For the classification stage, we optimized the following hyperparameters of each of the considered classifiers: the variable k and search algorithm of the k-Nearest Neighbors (k-NN) classifier; the maximum depth, number of estimators, and minimum samples to split for the Decision Tree and Random Forest classifiers; and the solver and maximum number of iterations for the Logistic Regression (LogReg) classifier. For the grid search scoring, we considered precision, recall, AUC, F1-score, and accuracy.

Since the incidence of fallers in the database was only 26.33%, we applied an oversampling procedure, namely, the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), to the training set employed in the grid search. In particular, the SMOTE was used to oversample the minority class in the feature space. In this way, the minority class was oversampled by creating synthetic examples rather than oversampling with replacement.

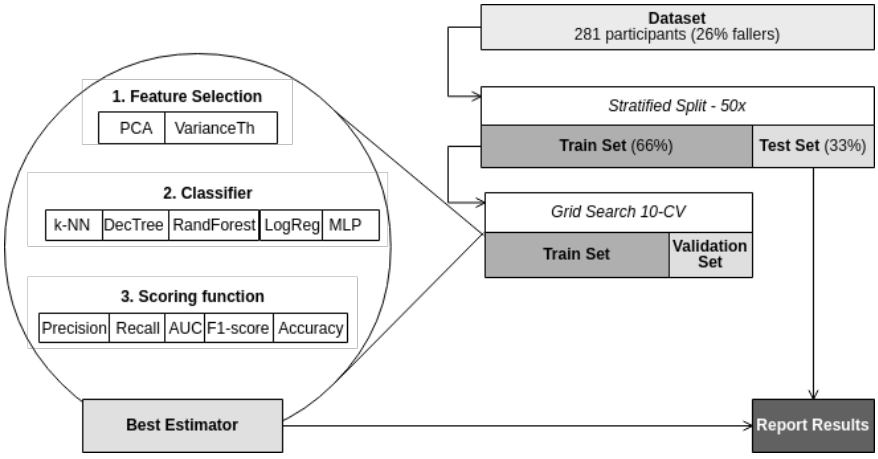

Figure 7.3: Classification pipeline for optimizing the feature selector, classifiers, and scoring function; grid search with CV is applied to the training set, whereas results are reported for the test set.

7.1.3.5 Validation

Since the grid search was performed over 50 partitions of the dataset and for the three stages of the pipeline, we obtained several combinations of parameters evaluated with different partitions of the initial dataset. We decided to present the mean and standard deviation across the 50 iterations

for each tested classifier combined with different feature selection methods, estimator's hyperparameters, and grid search optimization scores. We report the obtained accuracy, AUC, F1-score, precision, recall, and specificity. To compare the performance metrics across the different fusion approaches, we used ANOVA multiple comparison analysis testing. As a *post-hoc* test, we applied Tukey's Honest Significant Difference Test (HSDT) with 95% confidence level to all possible pairs among the three data fusion methods.

7.2 Results

7.2.1 Descriptive characteristics

7.2.1.1 Demographic and anthropometric information

A total of 281 older people aged over 65 were included in this study. Out of them, 65% were female, 74% were community-dwelling, and 17% used a walking aid. Participants were 75.1 $\pm$ 6.9 years old, 160 $\pm$ 7.9 cm tall and weighed 72.1 $\pm$ 11.1 kg.

7.2.1.2 Retrospective and prospective falls

Out of the 281 participants, 94 (33.5%) reported at least one fall in the previous year and 74 subjects (26.3%) experienced at least one fall during the 1-year follow-up. Among the 94 subjects that reported previous falls, 35 fell during the follow-up period.

7.2.1.3 Self-reported questionnaires

Self-reported questionnaires revealed that 38.8% of the participants required an upper extremity assistance to stand up from a chair. Among all participants, 35.6% reported living alone, 69% reported taking more than four medicines daily, and 50.5% reported having a sedentary lifestyle. When asked if they were afraid of falling, 52.7% answered affirmatively.

7.2.1.4 Functional tests scores

The majority of the subjects (253 out of 281) were able to complete all functional tests. Out of the 281 subjects, 13 subjects did not perform the TUG test, 14 subjects were unable to complete the Step test, and 17 subjects were unable to do the 30s STS test. Only eight subjects were unable to perform any standing position of the 4Stage test, whereas all participants completed the 10 Meter Walk test. Data from the subjects who were only capable of performing one or two tests were still considered for analysis.

7.2.1.5 Individual predictive value

We performed a statistical analysis of the individual predictive value of each feature for the prediction of 12 months prospective falls. The differences in the functional test scores between fallers

and non-fallers were not statistically significant (p-value>0.05). The difference between the two groups was statistically significant for the features highlighted in bold in Table 7.1.

7.2.2 No data fusion - individual data sources

The classification performance metrics using each data source individually were retrieved from the inner loop of the late fusion approach to access the predictive value of each data source. The results were grouped by the data source, feature selector, classifier, and grid search score. The average results for the 50 test sets were computed, and the highest recall values were retrieved as listed in Table 7.2.

Table 7.2: Average results for each data source (mean and standard deviation of the 50 test sets, in %).

Source	**Personal**	**Inertial**	**Platform**
Selector	PCA	PCA	PCA
Model	Decision Tree	Decision Tree	Log Reg
Score	Recall	Recall	Recall
Accuracy	40.5 ± 10.2	40.0 ± 7.8	39.8 ± 6.9
AUC	47.8 ± 4.8^{I}	50.5 ± 3.7^{S}	49.8 ± 4.4
F1-score	$34.0 \pm 7.9^{I,P}$	38.0 ± 3.9^{S}	37.1 ± 5.6^{S}
Precision	$23.8 \pm 4.9^{I,P}$	26.1 ± 2.1^{S}	25.5 ± 3.1^{S}
Recall	62.9 ± 22.9^{I}	72.2 ± 15.1^{S}	70.6 ± 16.6
Specificity	32.7 ± 20.6	28.8 ± 14.8	29.1 ± 13.7

S: sig. different from personal; I: sig. different from inertial; P: sig. different from platform

According to Tukey's HSDT performed for the single-step multiple comparison between the data sources, the averages of the accuracy and specificity were not significantly different across all data sources. For the AUC and recall, only the average of the personal data was significantly different from that of the inertial data. For the F1-score and precision, only the averages of the personal data were significantly different from that of the inertial and platform data.

7.2.3 Early, late, and slow fusion approaches

The same classification pipeline was employed for the early and late fusion approaches. For the early fusion approach, we used a combined feature vector with information from the three sources of data and ran it through the classification pipeline illustrated in Fig. 7.3. For the late fusion approach, the data were split into inertial, pressure platform, and a combination of clinical and self-reported data. The pipeline shown in Fig. 7.3 was optimized using each source of data individually and then the best estimator for each source of data was combined using voting classification. Finally, the evaluation using the test set was performed.

7.2.3.1 Early fusion approach

According to the early data fusion approach, clinical, self-reported, and multisensor data were fused using feature fusion prior to the classification pipeline. In addition to the clinical and self-reported data retrieved mainly from questionnaires (categorical data) and measured variables (e.g., timed tests or anthropometric characteristics), we employed features engineered from the raw signals of the inertial sensors and pressure platform. The initial analysis was performed individually for each type of sensor data. After retrieving features from the inertial sensor and pressure platform signals, they were combined in a unified feature vector together with clinical and self-reported data. This feature vector was then used for the optimization of the grid search pipeline and for retrieval of the best estimator. The resulting feature vector included 229 features extracted from the three data sources for 281 participants (aged over 65). The results were grouped by the data source, feature selector, classifier, and grid search score. The average results for the 50 test sets were computed and the highest recall values were retrieved as listed in Table 7.3. The best combination was PCA, Decision Tree, and recall as the grid search score function.

7.2.3.2 Late fusion approach

In the case of the late fusion approach, the same procedure as described for the early fusion was applied; however individual data sources were used in this case. We combined the predictions of three different estimators (based on individual inertial, pressure platform, and clinical/self-reported data) using a voting classifier. The feature vector constructed based on the inertial data comprised 125 features. In addition, we extracted 59 features from the pressure platform data and 44 features from the clinical/self-reported data. The model selection method was the same as described for the early fusion. The best combination was PCA, Decision Tree, and recall as the grid search score (Table 7.3).

7.2.3.3 Slow fusion approach

The slow fusion approach slows the process of fusing estimations by using a MLP to combine multiple data sources. In this case, we mixed information from each data source in the middle layers of the MLP, where the output from each individual stack of layers for each data source was concatenated in the last layers of the MLP. The three branches operated independently from each other until they were concatenated. In this way, we designed a network with three inputs and one output. The average results for the 50 partitions of the dataset are reported in Table 7.3.

According to the Tukey's HSDT performed for the single-step multiple comparisons between the fusion methods, the averages of the AUC and precision were not significantly different across the fusion methods. For the remaining performance metrics (accuracy, F1-score, recall and specificity), only the difference between the averages of the early and late fusion approaches was not significantly different.

Table 7.3: Average results for early, late, and slow fusion (mean and standard deviation of the 50 test sets, in %).

Fusion	**Early fusion**	**Late Fusion**	**Slow Fusion**
Selector	PCA	PCA	n.a.
Model	Decision Tree	Decision Tree	MLP
Score	Recall	Recall	Cross Entropy
Accuracy	35.8 ± 10.3^{S}	37.0 ± 9.8^{S}	59.2 ± 4.8
AUC	49.5 ± 3.9	50.5 ± 4.3	50.3 ± 4.4
F1-score	37.1 ± 7.5^{S}	38.2 ± 6.0^{S}	28.5 ± 6.4
Precision	25.2 ± 5.1	26.2 ± 3.2	26.3 ± 5.9
Recall	77.8 ± 24.1^{S}	78.6 ± 22.2^{S}	31.8 ± 8.5
Specificity	21.2 ± 21.6^{S}	22.5 ± 19.9^{S}	68.7 ± 6.9

S: statistically significant different from slow

7.3 Discussion and Conclusion

We tested three approaches for multisource data fusion, namely, early, late, and slow fusion, using the procedure illustrated in Fig. 7.2. We investigated the impact of fusing data at different stages of the pipeline on the obtained results. In this study looking at predicting falls in elderly, similar results were found for the majority of the considered performance metrics. Nevertheless, it should be noted that the late and slow fusion approaches can provide a set of advantages regarding the deployment of a prediction system. For example, a system capable of dealing with fewer sources of information can be designed and trained when a certain data source is not available. Moreover, we found that recall was more important than specificity, for the predictive system considered in this study, since the fall risk screening was used to select elderly with a higher risk of fall that should be considered for fall prevention. It would be preferable to minimize the error of losing a potential faller (i.e., maximizing recall) instead of losing a potential non-faller (i.e., maximizing specificity). This rationale was used to select the best models among all possible combinations in the optimization pipeline.

We compared each data source, regarding their predictive value individually. The inertial and platform data alone revealed a higher F1-score and precision compared with those of the personal data. Furthermore, the inertial data alone allowed to achieve a higher AUC and recall compared with those obtained when considering only the personal data. These results reinforce the added value of sensor instrumentation in fall risk screening protocols.

The average results obtained for the early and late fusion approaches were not statistically different from each other, which may indicate that the different data sources were highly correlated. The early fusion approach can be preferred given its lower computational requirements. The slow fusion approach obtained a higher accuracy score but a lower F1-score. The standard deviation of all scores achieved by this fusion approach was lower compared with those of the other approaches because the pipeline for slow fusion was only optimized for one loss function. By optimizing for cross-entropy, the model with slow fusion retrieved a higher specificity and lower recall compared with the early and late fusion approaches. The slow fusion approach can also be useful in scenarios

where specificity is more important than recall.

To the best of our knowledge, no previously published work has attempted to study different approaches to data fusion using multiple sources of data for prospective fall prediction. However, we found one previous work that reported a late fusion approach with clinical and inertial data for retrospective fall prediction validated using nested CV (Greene et al., 2018). The authors reported a significant added value of data fusion compared with analyzing individual data sources. The majority of previous studies report the use of a combination of personal (clinical and self-reported) data and one source of sensor data (either inertial sensors, pressure platform, or other types of sensor-based data) in an early fusion approach.

Furthermore, the classification and validation pipeline used in this study covers different stages of optimization. We reported the results for a test set that was not used during the training of the proposed grid search pipeline. The lack of an external test set, which is considered to be essential for the evaluation of trained models to avoid overfitting, has been considered as one of the main disadvantages of previous studies (Howcroft et al., 2013). Moreover, few studies have used neural networks for the prediction of falls or employed slow fusion approaches, which are more common for video classification (Karpathya et al., 2014).

Providing the results of this study, as our future work we consider studying different methods for feature processing and training different types of classifiers that are more suitable for each data source. Furthermore, it is possible that the nature of falls is not completely covered by the screening protocol used in this study. For example, once an elderly person with poor functional capabilities and clinical history of associated fall risk factors is institutionalized, the falling probability is reduced due to the resulting movement restriction. Adding strategies for data preprocessing and variables that better describe the unexpected nature of fall occurrences should be considered.

Part III

Wearable-based fall detection: impact of dataset, hardware restrictions, and model's requirements

Chapter 8

Introduction

8.1 Automatic detection of falls

Among the elderly population, falls are one of the most common causes of death and injury. More than 30% of people over 65 years old fall each year and the prevalence increases for people above 80 years old (WHO, 2007). Even a minor fall can severely affect the physical and mental health of an elder due to the fear of falling again. Thus, the elderly quality of life and of their carers can be affected (Age UK. Stop Falling, 2013). Besides the social and personal effects, falls also play an important role in healthcare costs. For instance, in 2015 the direct costs for fatal and nonfatal fall injuries were 637.5 million and 31.3 billion dollars respectively (Burns et al., 2016). Some studies have made some relevant developments on fall prediction through gait stability assessment. van Schooten et al. (2016) performed a study using wearable sensors to analyze the relation between common gait characteristics and the time to the next fall. Their findings reveal that with the daily measurement of these gait characteristics it is possible to assess the elderly risk of falling. Our previous study (Silva et al., 2020) provides a comparison between different data fusion approaches for fall prediction based on prospective falls. Although these systems can contribute to preventing falls, the occurrence of falls is not only dependent on the physical stability of the individuals but also on external perturbations such as, for instance, home hazards in the involving environment (Bruijn et al., 2013) or weather conditions. For this reason, the occurrence of falls is not always predictable. Therefore, it urges to be able to detect the falls at the moment they occur. Earlier detection of a fall will allow a faster intervention, decreasing the severity of injuries and, in some cases, avoiding deaths (Noury et al., 2007). Therefore, in the past years, the scientific community is making an effort in order to develop systems for automatic fall detection (Pannurat et al., 2014).

8.2 Related Studies

According to a recent review (Ren and Peng, 2019) of fall detection systems, the taxonomy of these systems can be divided into context-based and wearable-based systems. Context-based systems can sense and process data from the environment where the sensing device is integrated, instead of

using a device attached to the person. Examples of context-based systems are pressure platforms, cameras, acoustic and infrared sensors (Chaccour et al., 2017). Most wearable-based systems resort to the analysis of inertial sensors to detect falls. However, there are also systems that rely on sensor fusion, ranging from the fusion of inertial sensors to a combination of these sensors with barometer, microphone, heart rate sensors or cameras. Wang et al. (2016), for example, combined accelerometer and barometric information in a fall detector with low-power consumption.

The main problem with ambient devices and vision-based systems is the restriction of its use to the room where the sensors/cameras are placed. Additionally, they usually require complicated installation and setup when compared with wearable sensor-based systems (Mubashir et al., 2013). Smartphones' embedded sensors have also been used for fall detection (Aguiar et al., 2014a; Shahza and Kim, 2019), however, in some situations, they are not carried by the user or placed in the user's clothes, which will disable this specific function. Therefore, approaches based on wearable sensors, even though being more intrusive and less accurate (Mubashir et al., 2013), emerge as a potential solution to overcome these problems. As reported in (Schwickert et al., 2013) many proposals involving wearable sensors have been studied with the aim of solving the problem of automatic fall detection. The performance reported in such studies depends on several variables, for instance, the type of algorithm used, type of sensor, number of sensors, sensor position, and type of data used for training.

8.2.1 Falls datasets

Prior reported methods used accelerometer signal processing for the development of a supervised algorithm for fall detection with a dataset of simulated falls and activities of daily living (ADLs) considered as non-falls, acquired from young volunteers (Casilari et al., 2017a). The occurrence of a fall event is very rare, compared to the number of daily living activities, and the annotation process involved in creating a real-world falls dataset is very time and resource consuming, which makes the availability of such datasets extremely scarce. Even though, there are some research groups that have managed to collect data from real-world falls (Klenk et al., 2016).

Fall detection has been tackled with devices such as cameras, floor pressure sensors, infrared sensors, inertial sensors, heart rate sensors, and microphones. Most of the approaches based on wearable devices rely on inertial sensors to discriminate between falls and ADLs. Moreover, a high percentage of these studies have used datasets of simulated falls to develop and validate the fall detector. Previous studies on this topic have reported the comparison between accidental falls in elderly and simulated falls of younger volunteers, such as Klenk et al. (2011), Kangas et al. (2012) and Bourke et al. (2015). These studies concluded that the limitations of simulated falls should be considered and the protocol should be adapted to better match real-world falls.

The fall detectors training should consider the imbalance in real and simulated datasets, by employing imbalance learning methods on the train set and evaluate the trained models on imbalance conditions that mimic real-world scenarios. Machine learning approaches have successfully been applied to discriminate between falls and ADLs based on inertial sensors data. High sensitivities and specificities have been achieved for simulated falls datasets, however the validation of

these approaches has been questioned for real-world conditions (Lindemann et al., 2005)(Kangas et al., 2012). More recently, Bourke et al. (2016) has reported that a decision tree classifier trained with real-world falls is capable to discriminate between falls from ADLs with accuracy comparable with previous studies that used simulated falls datasets. However, a hybrid approach that considers both simulated and real-world falls for the development and validation of a fall detector could be of utmost interest, given that real-world falls datasets are rare and difficult to achieve because the limited number of available samples could impact the development of machine learning algorithms. Given that young volunteers have potentially more false positives than more sedentary older volunteers, non-fall events acquired from young volunteers should also be incorporated in the validation datasets. The FARSEEING real-world dataset was acquired from hospitalized patients, which may not be representative of a broad population. This way, the combination of this real-world dataset with simulated non-falls events could enrich the validation set, since more diverse data is used to validate the models.

The naive use of real data in models learned from simulated data may face difficulties due to the differences between both settings. However, the use of domain adaptation or transfer learning techniques has the potential to leverage the benefits of both.

Techniques of data augmentation have been used with these real falls in some studies. However, these methodologies can introduce some bias in the results (Khan and Hoey, 2017). Therefore, most of the studies presented in the literature have been using simulated fall data in order to train the algorithms. Some studies have reported significant differences in the simulated falls when compared with real falls data (Klenk et al., 2011), however, some have found that several characteristics are common between both types of falls (Kangas et al., 2012).

8.2.2 Wearable-embedded solutions

Most of the fall detection algorithms developed for wearable devices are based on simulated data acquired from a single (Kangas et al., 2012, 2009; Pannurat et al., 2017; Sucerquia et al., 2018) or multiple accelerometer sensors (Özdemir and Barshan, 2014). Often multiple inertial sensors are used, for instance, the combination of gyroscope and accelerometer (Li et al., 2009; Huynh et al., 2015) or accelerometer and barometer (Wang et al., 2016). Regarding the type of algorithm, works vary from the most simple threshold-based algorithms (Kangas et al., 2012, 2009; Pannurat et al., 2017; Bourke et al., 2010) to machine learning algorithms (Özdemir and Barshan, 2014; Ozdemir, 2016). As the number of wearable's sensors and algorithm complexity increases, more processing power will be required in order to execute the algorithm. This can be achieved with more powerful and expensive devices or by streaming the data to a more capable device, like a smartphone or computer, to process the data (Pannurat et al., 2017). However, in this case the need to always carry the smartphone is not eliminated (Aguiar et al., 2014a; Shahza and Kim, 2019). Also, the use of various sensors can be uncomfortable for the subject. There is a need of a single body-worn unit with inertial sensors that can be carried everywhere and used in most daily situations (Mubashir et al., 2013).

The review by Schwickert et al. (2013) concluded that there is no consensual position where the sensor should be placed in order to achieve the better fall detection rate. The most used positions are usually the waist (Kangas et al., 2009; Sucerquia et al., 2018; Bourke et al., 2010), hip and trunk (Pannurat et al., 2017). Some studies compared the performance of their algorithms for different positions. Kangas et al. (2008) compared the performance of the algorithm in three body positions: wrist, waist and head. They concluded that the head and waist would be the most suitable positions. Furthermore, Gjoreski et al. (2011) studied the best accelerometer placement for posture recognition and fall detection and concluded that both waist and chest positions had the best performance among waist, chest, thigh, and ankle. However, similarly to the smartphone, sometimes the sensors cannot be placed in a specific position due to the user's health condition or acceptability. Pannurat et al. (2017) have tackled this problem by developing an algorithm that works in several positions of the body such as the head, upper arm, wrist, ankle, chest, waist and thigh. However, this algorithm requires a calibration step for the position where the sensor will be used, and an external computer to perform the algorithm computation. If the person wants to change the position of the sensor, a new calibration step is always required. Since the user may not be an expert, the calibration step might be performed wrongly and this may hinder the proper functioning of the system.

8.2.3 Impact of models' requirements

Fall detection systems have been a trend research topic over the past years, motivated by the damaging impact of fall events in the quality of life, especially of the elder, and the importance of prompt assistance to minimize their consequences. Among the variety of available solutions, wearable-based systems, relying on ubiquitous equipment (e.g. smartphone, smartwatch, fitness trackers) to enable pervasive monitoring of users' motion parameters, are some of the most common. As such, there is a tendency to generate multiple fall detection solutions adapted to each different use case and shaped by each system's hardware limitations. This leads to an overflow of custom-made systems built upon similar methodologies but fine-tuned to particular objectives, constraints or even target populations.

Common examples of specific requirements and constraints are related to the wearable design, such as the place of usage, the way it can be attached to the body; the device's processing capability, memory and battery; or limitations in the accelerometer sampling rate. Fall detection systems' fine-tuning implies the collection of a significant amount of data examples, in conditions as similar as possible to those of the intended use, to train and test a new fall detection model. Hardware specifications may also influence the choice of the modeling approach and adaptations in the implementation of the model may be required. In summary, adjusting multiple fall detection solutions is a time and effort consuming process.

Regardless of the data source, the most common data analysis algorithms in the state-of-the-art can be divided into three main groups: threshold-based algorithms, binary or multiclass machine learning supervised algorithms and one class classification or novelty detection algorithms. The threshold-based approaches are simple algorithms that trigger a fall alarm when the sensor values

exceed certain predefined thresholds or a set of rules. Contrarily, machine learning approaches based on pattern recognition are more complex and sophisticated compared to threshold-based approaches. In novelty detection algorithms only data from daily life movements are used for training, and falls are detected as outliers.

Fall detection systems are usually trained and evaluated in simulated scenarios, given the difficulty of acquiring data from real-world falls. There are also research studies that attempted to collect data from real falls in uncontrolled settings; however, fall events are very rare and its respective number of samples is frequently insufficient to train robust supervised models, as in the two datasets used in the work of Aziz et al. (2017). Most previous works have acquired data from simulated falls and ADLs. Besides acquiring scripted samples in laboratory conditions, other studies have focused on acquiring and evaluating the trained models in free-living scenarios, from continuous usage of the wearable devices. Nevertheless, fall detection is usually an unbalanced problem, with a higher percentage of non-falls compared to falls in most prior works.

The system setup depends on several variables that could influence fall detection performance, and there has been some effort in previous studies towards understanding the impact of the wearable device usage position and the sensors' sampling rate. Intuitively, it is possible to acknowledge that the wearable position can influence the type of movements that could be misinterpreted as a fall, e.g. trunk, waist and pocket positions are expected to trigger fewer false alarms than the wrist, given its higher number of degrees of freedom as in Ozdemir (2016) work. Santoyo-Ramón et al. (2018) investigated the impact of number and positions of wearable sensors in fall detection. Their findings suggest that the best usage positions for the wearable devices are the chest and/or waist. On the other hand, the sampling rate has an impact on computational efficiency and battery life of the system. Liu et al. (2018) studied this topic and tested several models with lower sampling rates, and obtained 98% and 97% accuracy, with sampling rates of 11.6 and 5.8 Hz. In this sense, position and rate are both important to consider at the design stage.

Some studies benchmarked their method with the publicly available UMAFall dataset (Casilari et al., 2017b). Tsinganos and Skodras (2018) have extracted features from the accelerometer magnitude, considering only the belt position. These features were used to train a *k*-NN classifier. Their validation method was not user-independent, because they did not use leave-one-subject-out (LOSO) validation, and achieved a F1-score of 96.7%. The work of Wisesa and Mahardika (2019) revealed a F1-score of 97.4% using a Long Short-Term Memory (LSTM) model solely trained with data from the X-axis of the accelerometer from the belt position. For validation of results, the authors have divided the dataset into two static parts at random, disregarding user-independence; thus, the obtained results are not only orientation-dependent, but may also be optimistic if aiming real-world utilization with unseen users. The work of Barri Khojasteh et al. (2018) compared a decision tree (DT) model with a feed-forward neural network (NN). The authors have validated their models considering only the wrist position and applying a 5x2 cross-validation, which can also be considered user-dependent. The DT slightly outperformed the NN model regarding the geometric mean of sensitivity and specificity (DT - 92.4%; NN - 91.8%). The work of Wang et al. (2018) was the only study found with evaluation with the UMAFall dataset using LOSO cross-

validation. Their best approach combines data from accelerometer and gyroscope in a threshold-based algorithm. The highest obtained results were 95.3% sensitivity and 81.5% specificity (i.e, 88% geometric mean), although the authors did not refer if all UMAFall dataset positions were considered to evaluate their results.

8.3 Overview

Fall detection systems have been a trend research topic over the past years, motivated by the damaging impact of fall events in the quality of life, especially of the elder, and the importance of prompt assistance to minimize their consequences. Among the variety of available solutions, the most common are wearable-based systems, which rely on ubiquitous equipment (e.g. smartphone, smartwatch, fitness trackers) to enable pervasive monitoring of users' motion parameters. As such, there is a tendency to generate multiple fall detection solutions adapted to each different use case and shaped by each system's hardware limitations. This leads to an overflow of custom-made systems built upon similar methodologies but fine-tuned to particular objectives, constraints or even target populations.

Common examples of specific requirements and constraints are related to the wearable design, such as the place of usage, the way it can be attached to the body; the device's processing capability, memory, and battery; or limitations in the accelerometer sampling rate. Fall detection systems' fine-tuning implies the collection of a significant amount of data examples, in conditions as similar as possible to those of the intended use, to train and test a new fall detection model. Hardware specifications may also influence the choice of the modeling approach and adaptations in the implementation of the model may be required. In summary, adjusting multiple fall detection solutions is a time and effort consuming process.

In the next sections, we introduce several machine learning pipelines, trained with data from a comprehensive proprietary AICOS dataset, to model and deploy custom-made fall detection algorithms, based on which we will:

- *Type of dataset:* study the combination of simulated falls and real-world falls to improve the performance of the models.
- *Model complexity:* take into account hardware constraints that require low computational power features and algorithms for developing the fall detection algorithms. We have studied threshold-based algorithms and ML algorithms. We have also compared traditional feature-based models and deep learning models.
- *Single vs. multiple on-body positions:* compare models solely trained with data from sensors placed on a certain body position and models trained with data acquired at multiple positions. We also evaluated the generalization of those models for a new unseen position.
- *Sampling rate:* decrease the accelerometer sampling rate and evaluate the impact in fall detection performance.

Chapter 9

Transfer learning approach for fall detection with the FARSEEING real-world dataset and simulated falls

J. Silva, I. Sousa, and J. Cardoso,
Published in Proceedings of 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 3509-3512.

Falls are very rare and extremely difficult to acquire in free-living conditions. Due to this, most of the prior work on fall detection has focused on simulated datasets acquired in scenarios that mimic the real-world context, however, the validation of systems trained with simulated falls remains unclear. This work presents a transfer learning approach for combining a dataset of simulated falls and non-falls, obtained from young volunteers, with the real-world FARSEEING dataset, in order to train a set of supervised classifiers for discriminating between falls and non-falls events. The objective is to analyze if a combination of simulated and real falls could enrich the model. In the real-world, falls are a sporadic event, which results in imbalanced datasets. In this work, several methods for imbalanced learning were employed: SMOTE, Balance Cascade and Ranking models. The Balance Cascade obtained fewer misclassifications in the validation set. There was an improvement when mixing the real falls and simulated non-falls compared to the case when only simulated falls were used for training. When testing with a mixed set with real falls and simulated non-falls, it is even more important to train with a mixed set. Moreover, it was possible to conclude that a model trained with simulated falls generalize better when tested with real falls than the opposite. The overall accuracy obtained for the combination of different datasets were above 95%.

9.1 Data acquisition and processing

9.1.1 Simulated falls dataset

Simulated falls and non-falls were collected in the laboratory facilities of Fraunhofer Portugal AICOS (AICOS), using a smartphone inside the trousers' pocket, following the protocol described by Noury et al. (2007), adapted with additional non-fall activities. The dataset was collected from 7 volunteers that performed 8 types of falls (backward, forward, lateral falls) and 8 types of ADLs (sit-to-stand and stand-to-lay transitions, walk, run, bend, drop phone on a table, walk and sit, sit with rotations), repeated three times (Aguiar et al., 2014a). This dataset comprises 650 falls and 410 non-falls.

9.1.2 FARSEEING real-world fall database

In the scope of the FARSEEING project, a dataset with real falls were recorded from three different wearable devices: MiniMod and Hybrid were located in the lower back, sampled at 100Hz, and ActivPAL3 was placed on the thigh, sampled at 20Hz. The authors made available 22 accelerometer files from a set of 100 files and 1908 sequences of ADLs (Klenk et al., 2016). Each person, of a group of 15 persons, used a wearable device to collect falls and sequences of ADLs, at the Geriatric Rehabilitation Unit in Robert Bosch Hospital, Stuttgart. The available dataset comprises 22 files with 20 minutes of wearable sensor data. From these files, 21 have only one fall event and one file has two fall events (with a short duration of approximately 10 seconds). The sensor data prior to the fall event were considered as non-fall events.

9.1.2.1 Resampling

From the 22 samples, 7 were collected at 20Hz and the remaining were collected at 100Hz. Original files were resampled in order to uniformize the sampling frequency to 100Hz. First, the timestamp was converted to seconds: an artificial timestamp was created for all files in order to have unique timestamps. Second, the 20Hz files were upsampled to 100Hz to ensure concordance with the remaining of the dataset. The method used to upsample was based on replicating samples in order to have 100 samples in each second elapsed.

9.1.2.2 Segmentation

Centered in the timestamp of the fall (previously annotated by the FARSEEING group), a window with 7.5 seconds was defined as the fall period. The non-fall period was considered as the period until 10 seconds before the fall timestamp, yielding a total of 9 minutes and 50 seconds for the non-fall period. Then, this period was sequentially divided into 7.5 seconds windows. Since the number of samples of the non-fall period is considerably higher than the number of samples in the fall period, several imbalanced learning approaches were used to overcome this difference.

9.1.3 Comparison between datasets

Two datasets were used for the development and validation of fall detection approaches based on machine learning techniques: a simulated dataset acquired by AICOS's young volunteers in simulated conditions and a set of FARSEEING real-world falls acquired with elderly patients in a hospital. The major differences between the two datasets are different inertial sensors placement (simulated falls were acquired with the smartphone on the trousers' pocket and real falls were acquired with wearable sensors on the lower back and thigh), different sampling rates (simulated falls were acquired at 100Hz and some real falls were acquired at 20Hz), different inertial sensors specifications concerning the accelerometer amplitude range (certain sensors involved in the real-world data collection were limited to 2G range). In addition, the context of the fall is different: the elderly do not stand in front of the mattress to fall as in simulated conditions. The acceleration of the fall impact is also different since the impact is on the ground and not on the mattress.

9.2 Machine Learning Pipeline

9.2.1 Pipeline overview

Using Python's *scikit-learn* package (v.0.19.1), a classification pipeline was designed:

- *Input signal:* the accelerometer magnitude was computed, at 100Hz. The signal was divided into windows with 7.5 seconds (750 samples), without overlap. Windows with a low signal standard deviation were removed, in order to discard samples where the signal was mainly stationary, and were considered useless for the train and test sets.

- *Feature extraction:* a set of time-domain features were extracted for each time-window and include *mean, standard deviation, median, median deviation, maximum (max), minimum, energy, root mean square (rms), inter quartile range (iqr), histogram (10 bins), skewness and kurtosis* (Aguiar et al., 2014a).

- *Feature selection:* features with a correlation higher than 0.90 were removed. The removed features were *energy, iqr, max, median, rms*, yielding a feature vector with 16 features.

- *Feature standardization:* features were standardized by removing the mean and scaling to unit variance, in order to ensure that all features are standard normally distributed.

- *Train and test split:* for the real-world dataset, there were 23 falls and 877 non-falls (after removing stationary windows). The stratified split train and test, with 30% test size, was used, which yield a train set with 630 samples (14 falls and 616 non-falls) and a test set with 270 samples (9 Falls and 261 non-falls).

- *Supervised classifiers:* Nearest Neighbors, Decision Tree, Random Forest, Multi-layer Perceptron and AdaBoost.

- *Hyperparameters:* of each classifier were optimized, within a given interval, using Grid Search with 10 folds cross-validation, for the train set (630 samples). The stratified split train and test, with 30% test size, was once more applied in order to divide the train set into the train (9 falls and 413 non-falls) and the validation (5 falls and 203 non-falls) sets, for hyperparameters tuning. This process was repeated 40 times to ensure statistical significance.

9.2.2 Imbalance learning

A dataset is considered imbalanced when the classification classes are not approximately equally represented. Often real-world datasets are mainly composed of normal events with only a small percentage of abnormal or of interest events (Bowyer et al., 2011). For the real-world fall dataset, the amount of non-fall events is considerably higher than the number of fall events (due to its rare and sudden nature). To overcome the class imbalance, several approaches were considered, in order to ensure that the imbalance dataset does not have an impact on performance of the classifiers:

- *Synthetic Minority Over-sampling Technique:* (SMOTE) (Bowyer et al., 2011) was used to oversample real-world samples in the train set. Using this approach, the minority class is oversampled by creating "synthetic" examples rather than by oversampling with replacement.

- *Balance Cascade:* creates an ensemble of balanced sets by iteratively undersample the imbalanced dataset using an estimator (Liu et al., 2009). SMOTE and Balance Cascade are implemented in Python's *imbalanced-learn* (v.0.2.1) package (Lemaître et al., 2017).

- *Ranking Models:* were used for tackling class imbalance with ranking. Models tested with features extracted from real-world falls were: Adaboost, Balanced linear SVC, Linear SVC, Rankboost and Rank SVM (Cruz et al., 2016).

9.2.3 Transfer learning

We propose the use of domain adaptation/transfer learning techniques to cope with differences between simulated and real falls datasets. For means of comparison with a setup that only uses the real-world dataset, an initial test was made for training with real samples and test with real samples.

Given that the real-world falls dataset is highly imbalanced, 23 falls and 877 non-falls, imbalance methods were also applied in order to overcome this disproportion, namely the SMOTE algorithm. Moreover, given that the real-world dataset was collected with hospitalized patients, it could be expected that the activities the users undertook do not include high accelerometer variations, as expected when running or jumping or other activities with higher impacts. Due to this, a dataset of simulated non-falls, that includes samples with high accelerometer variations, were mixed with the real-world dataset (mixed set) in order to challenge the train and test sets (cases 1, 2 and 3). This approach was evaluated for three different cases: 1) train with a mixed set and

test with real samples and 2) train with real samples and test with mixed set and 3) train with mixed set and test with the remaining mixed set. These tests were meant to analyze the number of false positives achieved on a mixed set when the model is trained only with real samples (case 2) or when the model is trained with a mixed set (case 3). The opposite was also tested in order to assess the added value of training a model with a mixed set (case 1) comparatively to train only with the real-world dataset (case 6). The latter two cases were also compared with the case of training a model exclusively with simulated falls and non-falls and test it with real-world samples (case 4) and the opposite (case 5).

9.3 Results

9.3.1 Imbalance learning

9.3.1.1 SMOTE

SMOTE was applied for each classifier, using a stratified train and test split for 40 splits. Overall, the SMOTE algorithm obtained high accuracy for the set of five classifiers tested using a balanced train set with 413 falls and 413 non-falls. When evaluated in an imbalanced validation set, the MLP classifier had a higher area under the curve and achieved only one false positive and one false negative.

9.3.1.2 Balance Cascade

Balance Cascade divided the dataset into 53 sets and the classifiers were trained for each set. The model with the higher area under the curve was the MLP Classifier. Using the Balance Cascade, the accuracy with the training set was also very high. Despite the Random Forest, all remaining classifiers obtained accuracy, precision and recall above 90%. For the imbalanced test set, the best trained model obtained only one false positive.

9.3.1.3 Ranking Models

Ranking models were trained for 10 sets using the stratified split. The trained model with the highest area under the curve was evaluated with the test set. The results obtained for the ranking models were even higher than for the latter cases, however, when evaluated with the test set, the best model obtained a higher number of false positives (six non-falls were misclassified as falls).

9.3.2 Transfer learning

For the cases 1 and 6 (Table 9.1), when testing only with real data, if we include simulated non-falls in the training (case 1), the number of false negatives (FN, actual falls predicted as non-falls) will not change, as expected, however the number of false positives (FP, actual non-falls predicted as falls) will increase because the train set has more variability and also more samples. Even though, this model is expected to be more robust against potential non-falls events, since it was

trained with non-falls simulated in conditions that usually trigger more false positives. Comparing these two cases with case 4, when training only with simulated falls and testing with real falls, there is an improvement when mixing the real and simulated falls (case 1) compared to the case when only simulated falls were used for training (case 4).

For the cases 2 and 3, when testing with a mixed set with real falls and non-falls and simulated non-falls, it is indeed more important to train with a mixed set (case 3) than train only with real data (case 2), because less false positives were found. However, this model is also more prone to false negatives. In order to conclude which combination is better, it should be considered the cost of misclassifying a fall and the cost of detecting a false fall.

Comparing cases 4 and 5, it was possible to conclude that a model trained with simulated falls (Simul.) generalizes better when tested with real falls (case 4), than the opposite (case 5), a model trained with real falls and tested with simulated falls. Moreover, these two cases, that do not include any mixed samples, were the ones with lower accuracy (Acc) (expecting for case 6), highlighting the fact the mixing real and simulated non-falls improves the results.

Table 9.1: Transfer learning results for the combination of simulated and real falls (in %).

#	Train	Test	Accuracy	AUC	Precision	Recall	FP	FN
1	Mixed	Real	96	97	99	96	21	0
2	Real	Mixed	97	98	99	97	20	0
3	Mixed	Mixed	99	67	99	99	1	4
4	Simul.	Real	72	61	98	92	144	3
5	Real	Simul.	71	66	71	71	73	20
6	Real	Real	99	99	100	99	3	0

9.4 Conclusions

Detecting a fall in non-restricted nor simulated scenarios has been accomplished in most of the past works using wearable inertial sensors. Comparatively to camera-based approaches, wearable sensors avoid the need for environmental adaptations and fixed placement, allowing the monitored device to follow the user continuously. Since fall events are very rare and extremely difficult to acquire in free-living conditions, most of the prior work has focused on simulated datasets acquired in scenarios that mimic the real context. Even though, the validation of the systems that were trained with simulated falls remains unclear.

The approach presented combined two datasets: one with real falls and non-falls, from the FARSEEING real-world dataset, and another with simulated falls and non-fall events, acquired with younger and more active volunteers. Both datasets were used with different combinations for training and validation, in order to obtain a fitted supervised classifier that better generalizes to new fall events. In the real-world, falls are a sporadic and rare event, which results in imbalanced datasets. In this work, several methods for imbalanced learning were employed, to deal with this dataset, namely SMOTE algorithm, Balance Cascade and Ranking Models. Among the three

approaches, the accuracies obtained for a set of different classifiers were very high, but the Balance Cascade obtained fewer misclassifications in the test set.

Combined sets of simulated and real falls presented advantages compared to using only simulated falls. There is an improvement when mixing real falls and simulated non-falls compared to the case when only simulated falls were used for training. When testing with a mixed set containing real falls and simulated non-falls, it is indeed more important to train with a mixed set. Moreover, it was possible to conclude that a model trained with simulated falls generalize better when tested with real falls than the opposite. Compared to previous works that have used the FARSEEING real-world dataset, the sensitivity obtained with this approach overcome the one obtained by Bourke et al. (2016) using a decision tree classifier by 10%. Moreover, few samples of falls were used in this work for train and test, highlighting the need to employ imbalance learning and transfer learning approaches.

Chapter 10

Wearable Embedded Intelligence for Detection of Falls Independently of on-Body Location

Adapted from *J. Alves, J. Silva, E. Grifo, C. Resende, and I. Sousa,*
Published in MDPI Sensors (Basel, Switzerland) vol. 19, 11 2426. May 2019.

Falls are one of the most common problems in the elderly population. Therefore, each year more solutions for automatic fall detection are emerging. This study proposes a single accelerometer algorithm for wearable devices that works for three different body locations: chest, waist, and pocket, without a calibration step being required. This algorithm is able to be fully executed on a wearable device and no external devices are necessary for data processing. Additionally, a study of the accelerometer sampling rate, that allows the algorithm to achieve a better performance, was performed. The algorithm was validated with a continuous dataset with daily living activities and 272 simulated falls. Considering the trade-off between sensitivity and the number of false alarms the most suitable sampling rate found was 50 Hz. In conclusion, this study presents a reliable solution for automatic fall detection that can be adapted to different usages and conditions, since it can be used in different body locations and its sensitivity can be adapted to different subjects according to their physical activity level.

10.1 Materials and Methods

In this chapter the collection of datasets for training and validating the algorithm is described in Section 10.1.1, followed by the definition of the fall detection algorithm, Section 10.1.2. Furthermore, a method to study the accelerometer sampling rate that best fits our algorithm is presented in Section 10.1.3 and the algorithm optimization process is presented in Section 10.1.4.

10.1.1 Datasets

In order to train the proposed fall detection algorithm, 3 axial accelerometer data from simulated falls and non-fall movements were collected. This dataset will be referred to in the remaining document as *DS-1*. Besides ambulatory movements, some data from movements that, due to their hard impacts, could more likely trigger false alarms (FAs) were also collected. One of the most common movements performed during daily usage is to put the sensor on a table when the user will not use it, or, for instance, to charge the sensor. During a preliminary study, it was verified that this type of movement was one of the most likely to trigger false positives (FPs). Therefore a high amount of movements of this type were acquired to train the algorithm. Another type of non-fall movement acquired was, for instance, getting up, bend and pick up an object from the floor, based on the protocol presented by Noury et al. (2007); Ozdemir (2016). The acquisitions were made using the wearable sensor inside the user's pants frontal pocket, on the waist (fixed on the belt) or on the chest. Data were collected from 19 subjects, 5 women and 14 men, with an average age of 25 $\pm$ 2 years old who gave their informed consent and participated voluntarily in the data acquisition. The simulated falls were performed in ambulatory conditions and the users fell to a 10 cm high gym mattress while wearing a helmet for their safety. This dataset includes 1399 non-fall movements (6.5 h of data) and 1009 simulated falls (4.5 h) in a total of 2408 movements (11 h of data). Each sample acquired is considered as an activity, i.e., each sample acquired of walking a few meters is considered as a single activity that can be well classified as non-fall, or misclassified as fall.

To validate the algorithm in similar conditions to daily life, 22 young subjects, 5 women and 17 men, average age of 26 $\pm$ 3 years old, have performed a continuous data collection including non-fall movements intercalated with falls, referred to as dataset *DS-2*. From these 22 young subjects, only 9 subjects have participated in the data collection of the dataset *DS-1*. However, from those 9 subjects, 7 have only performed non-fall movements. The remaining two have contributed with data from simulated falls and non-fall movements. In this acquisition, each subject performed 6 min of each of the following ambulatory activities: standing still, sitting, walking, running, standing with freedom of movements and laying. Between these activities, each subject has performed 4 different simulated falls. The complete protocol had a duration of approximately 40 min per subject. Regarding the type of simulated falls, the subjects were divided into two different groups. A group of eleven subjects (1 woman, 10 men) has performed a forward fall, a backward fall, a sit-stand transfer fall, and a stand-sit transfer fall. The remaining eleven subjects (4 women, 7 men) have performed movements of stumble and fall forward, lateral fall, vertical fall

(faint simulation) and a forward fall preceded by getting up and walking for a few meters. These movements were chosen based on the protocol presented by Noury et al. (2007); Ozdemir (2016). During the acquisitions, each subject was wearing 3 wearable devices, in the three considered on-body positions: chest, waist, and pocket (14 subjects used it in the right pocket and 8 used it in both pockets). Eight of these subjects were wearing an additional sensor, making a total of 4 devices, placed in the chest, waist and both left and right pants frontal pockets. A total of 44.67 h of accelerometer data containing 272 falls were acquired. The algorithm was tested individually for each subject and the overall performance was analyzed.

10.1.2 Fall Detection Algorithm

The proposed algorithm is a modified version of the state machine algorithm developed in our previous work, adapted to be implemented in a wearable device, as described in Figure 2 from Aguiar et al. (2014a). Although the structure of the state machine is the same, since the device can be placed more freely on the body, the conditions of transition between states are different, as well as the thresholds of each state. The features used in the smartphone algorithm were dependent on the orientation of the device. This new version of the algorithm should work independently for three different on-body positions. In these positions, the device will have different orientations. Therefore, as we do not want to require a calibration step, the calculation of features was changed in order to make the features between each state independent of the wearable device orientation. The objective is for the user to wear the wearable in the most convenient place without having additional concerns. So, the thresholds between each state are defined by a feature that characterizes each new state. These features were defined in our previous study of Aguiar et al. (2014a) using machine learning techniques in order to obtain the best features that characterize each phase of the fall. Then the best features obtained using these machine learning techniques were used in order to build the state machine. To transit from state to state, a single or multiple features are calculated from the accelerometer data. This data is processed in real-time and sample by sample. If the value of the calculated feature crosses the threshold previously defined, the transition between consecutive states is performed. In this algorithm, in order to trigger a fall event, all the fall states should be successively detected through the analysis of the data collected by the accelerometer placed on the user's body. The state machine has five different states: Stable, Unstable, Falling, Impact and Unconscious Watcher, as presented in Fig.2 from Aguiar et al. (2014a). The transitions between states are described as follow:

1. **Stable to Unstable State:** When the fall detector algorithm is enabled and the subject is not moving, the system will start in the Stable state. Then, if some relevant acceleration changes are detected, the algorithm transits to the Unstable state. This change in acceleration is evaluated calculating the magnitude of the acceleration.

2. **Unstable to Falling State:** When in the unstable state, a significant decrease in acceleration can indicate that the user is experiencing a free fall, Falling State. To check if the decrease in acceleration values occur according to what is expected during a fall, the ratio between

the magnitude of the linear acceleration and the magnitude of acceleration at each moment is evaluated.

3. **Falling to Impact State:** If a fall is truly occurring, when the subject hits the ground, a sudden and significant increase in the acceleration and a large difference in body orientation occurs, corresponding to the change from standing/sitting positions to the lying position occurring after the fall. Thus, when the state machine is in the falling state, two different features are evaluated: the magnitude of acceleration and the angle between two different vectors: the average acceleration vector in this state and the average acceleration vector obtained before entering in the Falling state.

4. **Impact to Broadcast of the fall:** After the impact, the system starts an Unconscious watcher that will check if the user has recovered from the fall or not. If the user does not move during five seconds after the fall, a fall alert will be broadcasted. If some movement is detected, the system will restart the process in the Unstable state. This detection of movement is accomplished by evaluating the values of the acceleration magnitude during the Unconscious watcher.

10.1.3 Accelerometer Sampling Rate Analysis

There is no consensus in the literature regarding the ideal accelerometer sampling frequency for fall detection. In the review of Schwickert et al. (2013), 61 studies focused on the development of algorithms using accelerometer data for fall detection were analyzed and the range of sampling rates used varies from 6 to 3200 Hz. Furthermore, Kangas et al. (2012) has used dynamic sampling frequencies depending on the fall phases. Therefore, it is important to study the sampling rate that would allow our algorithm to achieve the best performance. Still, the accelerometer used in our wearable device only allows sampling rates of 8, 50, 100, 250, 333 and 500 Hz. As the samples of the datasets *DS-1* and *DS-2* were collected at 100 Hz, the performance of the algorithm can only be tested with the data undersampled to 8, 50 or 100 Hz. In Kangas et al. (2012) the authors discussed that during the pre-impact phase of a fall, the data sampled at 6.25 Hz were not enough to identify the movements in detail. Consequently, 8 Hz should also not be enough to discriminate the fall features and, for this reason, the algorithm optimization and validation processes were repeated using the data at 100 Hz and undersampled to 50 Hz.

10.1.4 State Machine Thresholds Optimization

Since the main objective of this work is the creation of a robust fall detection algorithm independent of the on-body sensor location, samples from the three on-body positions in dataset *DS-1* were mixed to train the algorithm.

The optimization process iterates over a set of thresholds of the state machine. For each threshold, a higher and lower bound of its possible value is defined. Then in each iteration, a random

combination between each threshold values will be tested and the results will be summarized for all combinations. This optimization process was implemented following these steps:

1. Perform a random stratified sampling of the dataset *DS-1* 10 times with a train/test ratio of 0.7, meaning, 70% for the train set and 30% for the test set.
2. For each split, randomly undersample the majority class on the train set, in this case the non-fall movements, 10 times.
3. For each undersampled set, randomly generate 100 thresholds sets according to the allowed lower and higher bounds set for each parameter.
4. Train each set of thresholds with the corresponding train dataset. Save the 50 best sets and their respective test set.
5. Test the 50 selected sets of thresholds with the corresponding test set.
6. From the results obtained for all iterations of dataset splitting and undersampling, the 50 sets of thresholds that presented the best result during the evaluation with the test set of the dataset were chosen.

The train/test ratio chosen was equal to 0.7 in order to take advantage of as much data as possible. Also, as it was also collected a validation set *DS-2*, the test part is only used for a preliminary choice and for this reason the majority of data can be used for training the algorithm.

10.1.4.1 Sensitivity Level Sets Selection

After the iterative process explained in the last section, since it is not possible to obtain a perfect score in the detection of falls and non-falls, it was decided to choose three different sets of thresholds according to their levels of sensitivity: high, medium and low sensitivity sets of thresholds. With this objective, from the 50 results obtained in step 6, the 10 sets of thresholds with the highest J-Index score were chosen in order to generate a Receiver Operating Characteristic (ROC) curve. To obtain the ROC, the value of sensitivity, true positive rate, obtained was plotted against the false positive rate (1-specificity). Then, the sets corresponding to the optimal point of the ROC curve, the point with the highest specificity and the point with highest sensitivity were selected as the medium, low and high sensitivity sets of thresholds, respectively. This process was repeated using the results of the iterative process performed with the data at both 100 and 50 Hz of sampling rate.

10.1.5 Algorithm Validation

In order to validate the set of thresholds chosen after optimization at 100 and 50 Hz, these were tested using the dataset *DS-2*, described in Section 10.1.1. To analyze the results obtained on this test, these were firstly plotted in order to create a Total Operating Characteristic (TOC) curve (Pontius and Si, 2014). In order to generate this curve, the hits, true positives (y-axis), are plotted

against the hits plus the false positives, this is, against the total of positive predictions (x-axis) (Pontius and Si, 2014). Therefore, this curve allows better visualization of the balance between FAs and the number of correctly detected falls for the six sets of thresholds (high, medium and low sensitivity levels for each frequency) chosen in the previous process of Section 10.1.4.1. Since the fall detection algorithm evaluates each of the data samples (44.67 h of data at 100 Hz corresponds to more than 16 million samples), the number of non-fall cases that can possibly generate FAs is very large and difficult to represent. Thus, the maximum value of the x-axis of the TOC chart, which represents the worst case in which all the non-fall samples are wrongly classified as falls, is not represented. On the other hand, the maximum of the y-axis, where the well-classified falls are represented, is the number of falls that the dataset *DS-2* contains. To analyze the TOC chart, it should be taken into account that when there were more points below the maximum of the y-axis, the number of falls that were not detected increased. Also, if the algorithm has no FAs the number of well-detected falls is the same as the number of well-detected falls plus the false alarms, the value in x is equal to the value in y. Therefore, the more points there are to the right of the diagonal line, which represents the existence of no FAs, the higher the number of the FAs.

The performance of the fall detection algorithm when tested with the dataset *DS-2* is also analyzed by calculating the sensitivity, precision, and F-score. In order to better depict the prevalence of FAs and their possible impact on the daily life, besides the number of FAs and the precision, we also take into account the number of FAs per day, considering 16.5 h of daily usage (Bourke et al., 2010).

10.2 Results

10.2.1 Threshold Optimization - 100 Hz

The ROC curve obtained after the threshold optimization process using the data at 100 Hz is presented in Figure 10.1. From this ROC curve the sets corresponding to the three levels of sensitivity, black points, were selected. The medium sensitivity point (black dot in Figure 10.1) presents 95.0% of sensitivity and 94.4% of specificity. The low sensitivity set which is also the one with higher specificity, black square, has 97.3% specificity and 89.4% of sensitivity, and the high sensitivity set (black triangle in Figure 10.1) has 96.4% of sensitivity and 92.8% of specificity.

10.2.2 Thresholds Optimization - 50 Hz

The ROC curve with the 10 best results obtained in the test using the data undersampled at 50 Hz is presented in Figure 10.2. The black square in Figure 10.2 represents the results of the low sensitivity level, 97.4% of specificity and 93% of sensitivity. The medium sensitivity level is represented with the black dot, 96.4% of specificity and 96.7% of sensitivity. Lastly, the black triangle in Figure 10.2, represents the high sensitivity level, 98.3%, that had also 94.5% of specificity.

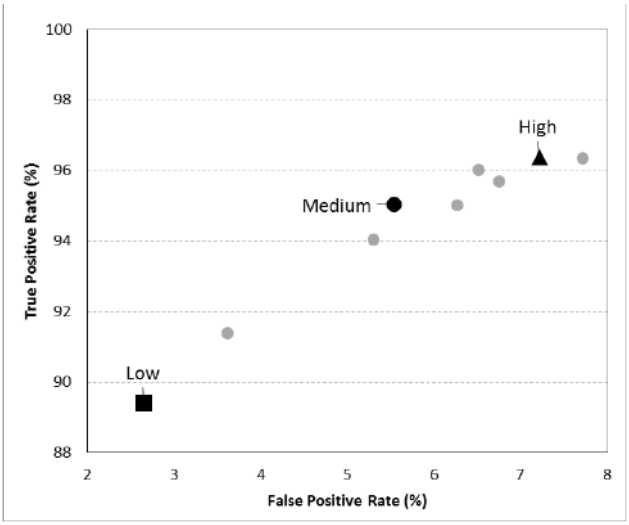

Figure 10.1: Receiver Operating Characteristic (ROC) curve with the 10 sets of thresholds that presented a better J-index when the algorithm was tested with the test set sampled at 100 Hz—Black square: Low sensitivity level; Black doth: Medium sensitivity level; Black triangle: High sensitivity level.

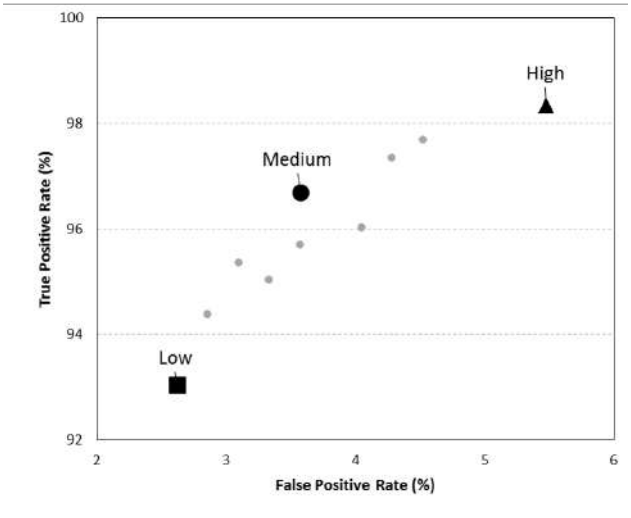

Figure 10.2: ROC curve with the 10 sets of thresholds that presented the best J-index when the algorithm was tested with the test set sampled at 50 Hz—Black square: Low sensitivity level; Black doth: Medium sensitivity level; Black triangle: High sensitivity level sensitivity.

10.2.3 Comparison between 50 Hz vs. 100 Hz Sets

Table 10.1 summarizes the results for the three levels of sensitivity chosen in both 100 and 50 Hz optimization, already mentioned in Sections 10.2.1 and 10.2.2. The results regard the test with 30% of the *DS-1*. As our iterative process randomly splits the *DS-1* in each iteration, the test results are obtained for different test sets. Therefore each set of thresholds was tested with a different test set. Even so, as can be observed in Table 10.1, the levels of sensitivity optimized for 50 Hz perform better than 100 Hz levels. The specificity increase was already expected when decreasing the

Table 10.1: Comparison of 100 Hz and 50 Hz sets of thresholds when tested with each respective test set, 30% of the dataset *DS-1*.

Sensitivity	**High**		**Medium**		**Low**	
Frequency (Hz)	100	50	100	50	100	50
Sensitivity (%)	96.4	98.3	95.0	96.7	89.4	93.0
Specificity (%)	92.8	94.5	94.4	96.4	97.3	97.4
J index (%)	89.2	92.8	89.4	93.1	86.7	90.4

sampling frequency since with a lower number of samples the possibility for outliers is reduced. This comparison between 50 and 100 Hz levels indicates that the algorithm performs better with data sampled at 50 Hz. However, since the test sets used are different, further validation with the dataset *DS-2* was conducted as described in Section 10.1.5.

10.2.4 Algorithm Validation in Continuous Usage

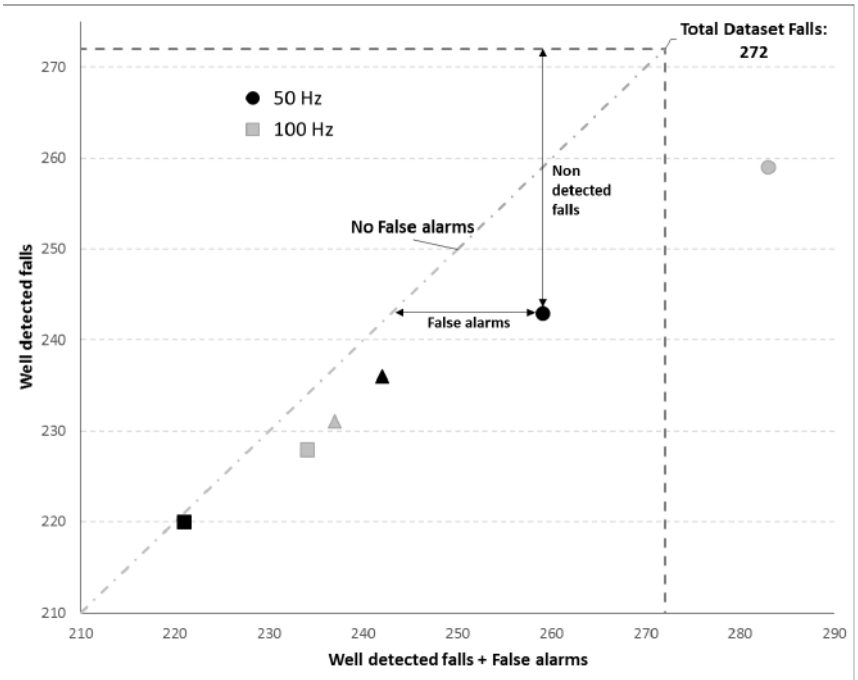

Figure 10.3: Total Operating Characteristic (TOC) curve with the results of the algorithm on *DS-2* using each set of thresholds chosen for each frequency. Legend: Squares—Low sensitivity levels; Triangles—Medium sensitivity levels; Circles—High sensitivity levels; Black marks—50Hz; Grey marks—100 Hz.

The TOC curve obtained with the results from the validation test of the algorithm using each set of thresholds is presented in Figure 10.3, as explained in Section 10.1.5 of the methods. Analyzing this Figure 10.3 it can be observed, as expected, that results from the high sensitivity sets of thresholds, circles, are the ones closer to the "Total Dataset Falls" line which means that the number of falls correctly detected is, as expected, the highest for these sets. Simultaneously, these

points are the ones that are further away to the right from the diagonal "No False alarms" line, which means that these sets are also the ones with a higher number of FAs. The 50 Hz sets are, with exception of the medium sensitivity level, black triangle, closer to this line comparing with the 100 Hz sets. On the other hand, the 100 Hz sets points, are always above the corresponding 50 Hz sets points, meaning that the number of well-detected falls is higher. The exception to this behavior is the medium sensitivity sets since the 50 Hz set is able to detect a higher number of falls, as observed by the black triangle being above the grey triangle in Figure 10.3, while having the same number of FAs. In Section 10.2.3, the 50 Hz threshold sets showed the best performance regarding both sensitivity (fall detection rate) and specificity when tested in the test set of *DS-1*. It does not occur during this validation since the sensitivity is lower for the high and low sensitivity levels at 50 Hz when compared with the corresponding 100 Hz sets.

Table 10.2: Results of the test using the *DS-2* with both 50 and 100 Hz sets of thresholds.

Sensitivity Levels	**High**		**Medium**		**Low**	
Amount of data	44.67 h					
Number of falls	272					
S_{rate} (Hz)	100	50	100	50	100	50
True Positives	259	243	228	236	231	220
False Alarms	24	16	6	5	6	1
Sensitivity %	95.2	89.3	83.8	85.6	84.9	80.9
Precision %	91.5	93.8	97.5	97.5	97.5	99.5
F-score %	93.3	91.5	90.1	91.8	90.8	89.3
FA per day	8.9	5.9	2.1	1.9	2.1	0.4

The remaining metrics obtained during this process of validation of the algorithm were summarized in Table 10.2. It shows that for all sets of thresholds, independently of the sampling rate, the sensitivity is always higher than 80%. As expected the set that has the lowest sensitivity, 80.9%, also has the lowest number of FAs. One FA was triggered using almost 45 h of data for the low sensitivity level thresholds set optimized at 50 Hz. Regarding the low sensitivity set at 100 Hz, the fall detection rate is higher, 84.9%. However, the number of FAs also increases from 1 to 6. These results are even more relevant when analyzing the average FAs per day, since the 100 Hz set would broadcast, in average, more than 2 FA per day while the 50 Hz set would only broadcast, in average, less than 0.5 FA per day.

On the other hand, the high sensitivity thresholds set optimized at 100 Hz is the one that presents the best fall detection rate, 95.2%. However, this set has also triggered a large number of FAs (24) with an average of almost 9 FAs per day. Comparing the high sensitivity set optimized for 100 Hz, which presents 95.2% of sensitivity, with the one optimized for 50 Hz data, the sensitivity decreases to 89.3%. These values reveal that the amount of non detected falls increases from 13 to 29 between the 100 Hz and the 50 Hz sets. However, the precision increases from 91.5 to 93.8%, from the 100 to the 50 Hz set, revealing that using the data at 50 Hz can reduce the number of FAs

in 3 FAs per day.

Regarding the medium sensitivity level, both 50 Hz and 100 Hz sets have performed similarly regarding the number of FAs, 6 at 100 and 5 at 50 Hz. However, the 50 Hz set has a higher value of fall detection rate, 86.8% vs. 83.8%. The number of FAs with these sets are lower than with the sets with high sensitivity level, however, they are still quite high with an average of 2.1 and 1.9 FAs per day, for the 100 Hz and 50 Hz sets, respectively.

Summing up, the use of a sampling rate of 50 Hz will benefit the algorithm's performance by reducing the number of FAs, which is particularly relevant for the low sensitivity level mode.

10.3 Discussion

In this study, we propose a fall detection algorithm that works independently for three different body locations, waist, chest, and pocket. This algorithm is a low complexity accelerometer state machine that was implemented in a wearable device. The device is responsible for executing all the data processing and communicates the occurrence of a fall to, for example, a smartphone that would be responsible to transmit this information to a caregiver. In case the wearable contains a GSM feature, the fall alert could be directly broadcasted to the caregiver. Additionally, three different sensitivity/specificity optimization modes were developed, and a study about which accelerometer sampling rate would allow a better algorithm performance was carried out.

Different people have a different fall risk, depending on several intrinsic factors, such as, the physical activity level and specific health conditions. Factors like movements with huge impacts have similar patterns to falls which can influence the performance and reliability of the algorithm due to the probability of triggering false alarms. Therefore, it is particularly interesting to have an algorithm that allows an adjustment to the trade-off between specificity and sensitivity. This means that, if a person has a high risk of falling, a high sensitivity level of the algorithm should be used, while for a healthy person a low/medium sensitivity level should suffice. Thus, some sensitivity levels tested in this work should be considered for daily use. For instance, for a low risk patient, the best sensitivity level set at 50 Hz can be used, having 80.9% of fall detection rate and, on average, less than one FA per day, supporting three on-body positions. If the patient suffers from a condition that increases his fall risk, a high sensitivity level set should be selected, for instance, the 50 Hz set that, in our validation, had almost 90% of fall detection and 6 FA per day. The FA values will also depend on how active the user is since active people will be more likely to have sudden movements that can be confused with falls and cause FAs. Since the data used in validation was acquired from young subjects, that are usually more active than elderly users, the results regarding the number of FAs can be biased representing a worst case scenario. For the subjects with a high risk of falling, it is more important to have a better fall detection rate, due to the highest probability of occurring a fall. At the same time, these subjects are usually people with low mobility, which consequently have fewer movements that can generate FAs. Therefore one of the high sensitivity sets can be suitable for this group of high fall risk since this sensitivity level has a high fall detection rate and the number of FAs would probably be reduced given the users'

lower mobility. The sets of thresholds deployed should also depend on the physical activity of the user. A high physically active person will have more movements that can have similar patterns to the falls and therefore sets with higher specificity should be deployed. Further validation with elders with different levels of physical activity and people with a high risk of falling is, for this reason, still required.

When varying the sampling rate of the accelerometer, the performance of the algorithm improves when the sampling rate decreases from 100 to 50 Hz, mainly regarding the FA rate. Gao et al. (2014) have studied the performance of some single accelerometer activity classifiers for sampling rates between 10 and 200 Hz. They verified that the accuracy of the algorithm increases from 10 to 50 Hz and stabilizes above this frequency. These results are in accordance with those obtained in this study. Between the supported frequencies, 50 Hz was the frequency that showed the best results in our study considering the analysis of the trade-off between fall detection rate and FAs for the three sensitivity levels. The 100 Hz sets have shown, however, better performance regarding the fall detection rate, mainly for the high sensitivity set. Hence, the use of different sampling rates depending on the intended sensitivity level can be considered. For instance, when a high sensitivity level is chosen, the algorithm would use the data at 100 Hz and the respective high sensitivity sets of thresholds. For the low and medium levels the data would be undersampled to 50 Hz and the respective sets of thresholds, optimized to this sampling rate, would be used.

A low-power fall detector using accelerometer and barometer data was proposed by Wang et al. (2016). It was considered the usage of the sensor on the user's chest. Their objective, like ours, is to develop an algorithm that runs on a wearable device and allows its battery to last as long as possible. They obtained a sensitivity of 93% and a FA rate of 0.023 alarms per hour, 0.3795 per day considering 16.5 h of daily usage. In our work, for instance, using the medium sensitivity level at 50 Hz wearing the sensor on the user's chest, no FA were obtained. Additionally, the value of the fall detection rate is 8.8% lower than the obtained in Wang et al. (2016). In general, the algorithm using the remaining set of thresholds outperforms their in terms of FA, but is outperformed regarding the obtained sensitivity. The worst performance of our algorithm using most of the sensitivity levels compared with their work could be explained, mainly, by the use of the barometer sensor in their work.

Pannurat et al. (2017) developed a hybrid framework that combines activity classification using machine learning and rule-based knowledge representation for the detection of different phases of falls. Similarly to our algorithm, it also works for several positions, namely, head, arm, wrist, ankle, chest, side waist, front waist, and thigh. For all the positions, the specificity of their algorithm is lower than the one obtained with our algorithm using any sensitivity level set. When testing their best algorithm with a dataset containing activities of daily living they found false alarm rates below 0.05% for both waist and chest positions. The dataset they used has smaller periods of data of each activity (15 s) when compared to our continuous dataset of Section 10.1.1. As they use 0.5 s windows, for the chest position they have 61,200 data samples. Therefore, the value of false alarm rate reported, 0.05%, representing a total of 6 FAs in less than half an hour of data. With some specific levels of thresholds for both considered positions, chest and waist, our algorithm

did not trigger any false positive in almost 45 h of data. The fall detection rate is similar to both algorithms. Although our algorithm presents better results, a further comparison using the same dataset for validation is still required. Even though, the algorithm presented has some advantages when compared with the one presented by Pannurat et al. (2017), such as the independence of the on-body sensor position for the three positions considered (waist, chest, and pocket), as well as the fact that it does not require a calibration step before changing the wearable position.

A study of Bagalá et al. (2012) demonstrated that most common algorithms decrease their accuracy when tested with real falls instead of simulated ones. For this reason, the algorithm presented in this work still requires validation with real falls. Even though, this data is really difficult to obtain since falls are rare and unpredictable events (Khan and Hoey, 2017).

10.4 Conclusions

In this work, we present a fall detection algorithm implemented in a wearable device that can be used in three different body positions, chest, waist, and pocket. It uses single accelerometer data and classifies the movements using a state machine. Even while having to respect hardware constraints that require a very simple algorithm and impose implementation approximations, and is optimized and tested for three different on-body positions, it presents a performance level similar to more complex or single position algorithms currently presented in the literature. Additionally, it shows some versatility since it can be adjusted to three different levels of sensitivity that can be used to better suit the subjects' needs depending on different risks of falling and mobility patterns. In this study it has also been showed that decreasing the accelerometer sampling rate does not largely affect the accuracy of the detection, being even beneficial for avoiding false alarms. The decrease in the accelerometer sampling rate has a positive side effect in decreasing the demand for the processing capabilities, resulting in a more suitable use of the algorithm in wearable devices.

To summarize, the fall detection algorithm described in this work presents a reliable, simple, and wearable solution for automatic fall detection. This is a versatile solution that can be adapted to different groups of people with different fall risk levels. Also, it can be used in different on-body positions, without requiring any calibration step, which makes this system less intrusive and easier to use.

Chapter 11

Automated development of customised fall detectors: position, model and rate impact in performance

J. Silva, D. Gomes, I. Sousa, and J. Cardoso,
Published in IEEE Sensors Journal, vol. 20, no. 10, 2020

The past years have witnessed a boost in fall detection-related research works, disclosing an extensive number of methodologies built upon similar principles but addressing particular use-cases. These use-cases frequently motivate algorithm fine-tuning, making the modeling stage a time and effort consuming process. This work contributes towards understanding the impact of several of the most frequent requirements for wearable-based fall detection solutions in their performance (usage positions, learning model, rate). We introduce a new machine learning pipeline, trained with a proprietary dataset, with a customizable modeling stage which enabled the assessment of performance over each combination of custom parameters. Finally, we benchmark a model deployed by our framework using the UMAFall dataset, achieving state-of-the-art results with an F1-score of 84.6% for the classification of the entire dataset, which included an unseen usage position (ankle), considering a sampling rate of 10 Hz and a Random Forest classifier.

In this study, we introduce a new machine learning pipeline, trained with data from a comprehensive proprietary dataset, to model and deploy custom-made fall detection algorithms, based on which we shall:

1. Study the cases in which customization is indeed necessary
    - *Model complexity:* Do models of higher complexity outperform models of more modest complexity at detecting falls?
    - *Sensor position generalization:* Do models that were not trained with data from sensors placed on a certain body position maintain their performance when evaluated with these data?
    - *Single vs. multiple training positions:* Do models solely trained with data from sensors placed on a certain body position A perform better than models trained with data acquired at multiple positions when evaluated with data from A?
    - *Sampling rate:* Does the accelerometer sampling rate have an impact in fall detection performance?
2. Evaluate the performance of our framework against the state-of-the-art
    - *External data generalization:* Do models deployed by our framework perform adequately at detecting falls using datasets acquired under different conditions?
    - *Positioning within state-of-the-art:* Is the performance of a model deployed by our framework competitive within the state-of-the-art?

All in all, this study makes significant contributions towards i) understanding if customization is indeed necessary for a specific use case, namely regarding usage position, accelerometer sampling rate, and processing/performance trade-off requirements; ii) the automated creation of mature ready-to-go fall detection solutions adapted to several of the most frequent customization requirements for wearable-based systems.

11.1 Methods

Figure 11.1 depicts an overview of the proposed approach for the automated development of custom fall detectors, enabling a clearer understanding of the relation between each stage within the flow of the method. The following subsections detail the steps at each of these stages.

11.1.1 Data acquisition

11.1.1.1 Protocol

Fraunhofer AICOS has been acquiring simulated falls and non-falls since 2009. The protocol for data collection was first described by Aguiar et al. (2014a); Alves et al. (2019), which followed

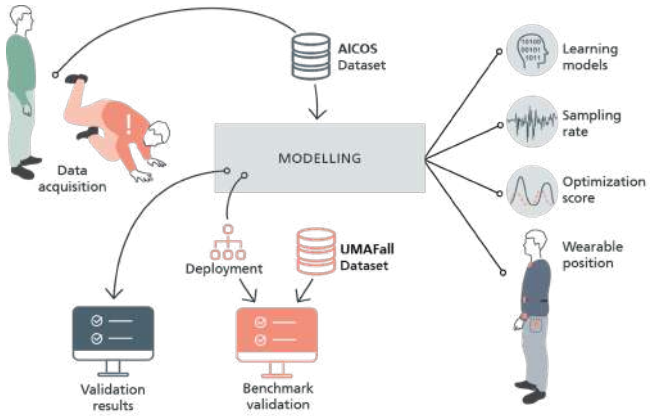

Figure 11.1: Study design overview.

the protocol defined by Noury et al. (2007), and considers data collection for the on-body sensor positions of chest, belt, and pocket. Recently, that protocol was extended to include the wrist position and non-fall movements specific to the wrist. The dataset was collected in laboratory conditions, at AICOS' living lab, where two mattresses were placed in the ground. The living lab also included a sofa, a table with chairs, a bed and an open space for acquiring running and walking samples. The activities of daily living recorded as non-falls included drop the sensor on the table, sit on a lower chair, catch an object from the floor while walking, run a few meters, laying on a bed, among others. The type of falls recorded included forward, backward and lateral falls (without recovery) ending lying on the floor. The protocol was previously described by Aguiar et al. (2014a) and Alves et al. (2019). Overall, the dataset comprises 36 different types of falls and 43 types of non-falls. Data was collected using a data logger Android application that provides access to the inertial sensors either directly built-in the smartphone or in wearable devices paired with the smartphone. The wearable devices used are proprietary of Fraunhofer AICOS and include a 3-axis Inertial Measurement Unit (IMU) (AICOS, 2016). Several smartphone models were used for data collection, namely: Samsung S3, S3 Mini, S4, Nexus S, Galaxy Nexus, Nexus 5, Moto G XT1032, and Vodafone 985N.

11.1.1.2 Data distribution

Data was collected in several occasions, from different participating subjects who wore a set of devices in different on-body locations. For this reason, none of the subjects has collected data for the complete set of usage positions considered in this study. For each subject, the positions for which only one class is available (fall or non-fall) were removed prior to the analysis. The cleaned dataset is composed of 42 subjects (34 males) with average age of 25.0 $\pm$ 2.9 years, an average weight of 72.4 $\pm$ 12.6 kg, and an average height of 176.0 $\pm$ 7.9 cm. The percentage of samples that were captured by the built-in sensors of the smartphones was 54.17% and the percentage acquired with the wearable devices was 45.81%. The average sampling rate for the smartphone samples was 102.26 $\pm$ 24.11 Hz and for the wearable samples the average sampling rate was 97.68 $\pm$ 8.50

Hz. The accelerometer range was $\pm$ 2G for all used smartphone models and wearable devices. The distribution of samples across the two classes is presented in Table 11.1. The belt and pocket sensor positions have a higher percentage of samples than chest and wrist positions, because belt and pocket positions include samples from the smartphone and from wearable devices, whereas the chest and wrist include only samples from wearable devices. Overall, the distribution of falls and non-falls per position for the entire dataset may be considered nearly balanced. On average the fall events have a duration of 15.20 $\pm$ 4.99 seconds and the ADLs activities have a duration of 14.94 $\pm$ 5.30 seconds.

Table 11.1: Distribution of dataset across different positions in terms of number of subjects, fall and non-fall samples.

Position	**Subjects**	**Fall**	**Non-fall**	*Total*
Belt	24	1731	1407	3138
Pocket	28	1305	1146	2451
Wrist	7	887	1112	1999
Chest	12	455	401	856
Total	42 (unique)	4378	4066	8444

11.1.2 Modeling

Figure 11.2 illustrates the pipeline for automated modeling, using the AICOS dataset. This pipeline is prepared to receive several input parameters that enable the customized modeling (see Figure 11.1): 1) train and test positions; 2) learning models; 3) target sampling rate; 4) grid-search optimization score. In the scope of this work, all experiments were performed using the F1-score as the optimization score.

11.1.2.1 Data preprocessing

A resampling strategy was firstly implemented with the aim of correcting the time distribution of all arriving samples and compensating for eventual sensor reading gaps. The accelerometer signal magnitude was evenly sampled, according to the target sampling rate. To that end, we computed the expected time of arrival of each sample (t_e). Samples arriving before t_e were stacked and their average was computed and set to correspond to t_e; if there were no samples arriving before t_e, the value of the last sample which arrived in the stream was considered. This combination of up and down-sampling techniques resulted in the computation of the accelerometer signal magnitude, evenly distributed in time, according to the required sampling rate. The data stream was segmented into windows of 7.5 seconds, without overlap, centered in the signal magnitude maximum. If there were not enough samples in the beginning or in the end of the window, after centering it in the maximum, the first and/or the last samples, respectively, were replicated until the pre-defined window size is reached. Windows with a standard deviation of low accelerometer magnitude were removed in order to discard samples that were useless for training the fall detection algorithm.

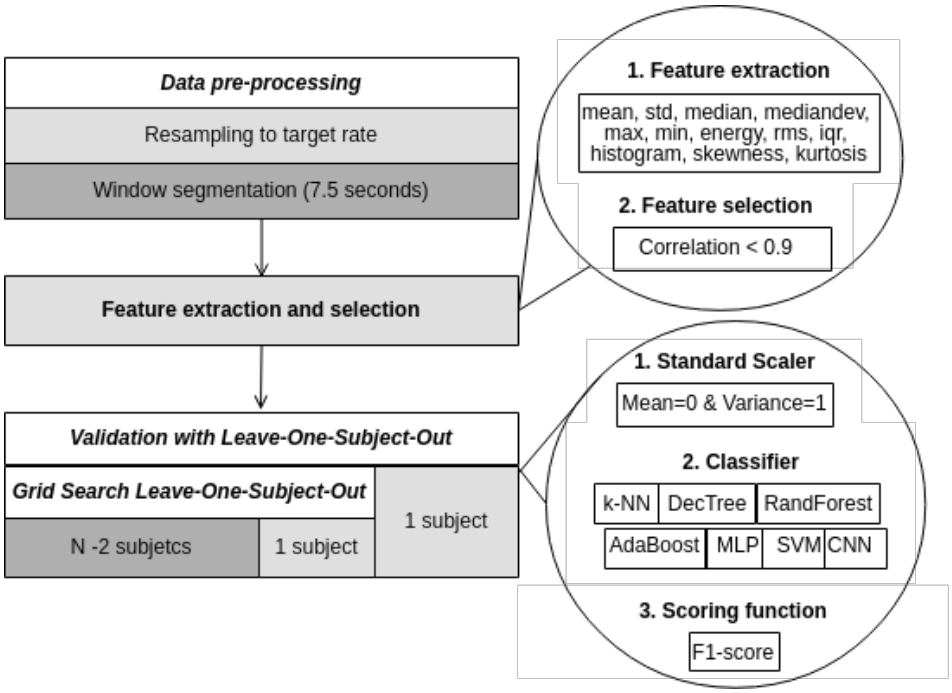

Figure 11.2: modeling stage: data preprocessing, feature extraction and selection, and nested leave-one-subject-out validation with grid search.

11.1.2.2 Features

Several time-domain features were extracted for each time-window signal magnitude: mean, standard deviation, median, median deviation, maximum, minimum, energy, root mean square, interquartile range, histogram (10 bins), skewness and kurtosis, using our open-source Time Series Feature Extraction Library (AICOS, 2019). These features require low computation power and are the most commonly used features for fall detection according to Pannurat et al. (2014). Features with a correlation higher than 0.90 were removed. All features of the training set were standardized by removing the mean and scaling to unit variance. The same parameters were used to standardize the test set. These features constituted the input for all classifiers, with the exception of CNN. CNN received a feature vector of raw signal magnitude (for each time-window), re-scaled to $[0, 1]$ range by subtracting the minimum and dividing by the difference between the maximum and the minimum signal magnitude.

11.1.2.3 Leave-one-subject-out validation

Two nested LOSO loops were used for training and validation assessment. The inner LOSO was used to optimize the hyperparameters of the learning models via grid search (except for the CNN-1D model) using N-2 participants for training and 1 subject for validation.

- *Grid search for hyperparameters optimization:* The hyperparameters of the learning models were optimized for F1-score metric. The following hyperparameters were optimized for each classifier, *k*-Nearest Neighbours (*k*-NN): parameter *k* and search algorithm; Decision

Tree (DT) & Random Forest (RF): maximum depth, number of features and estimators, and minimum samples to split; AdaBoost: number of estimators; Multi-layer Perceptron (MLP): variable alpha, activation function and learning rate; Support Vector Machine (SVM): variable C, degree, gamma and type of kernel.

- *CNN-1D architecture:* The architecture of the network encompasses two stacked 1-Dimensional Convolutional Neural Network (CNN-1D) with a kernel size of 5, with 4 filters, and a tangent activation function. CNN-1D layers were interleaved with max-pooling and 0.25 dropout layers. The sigmoid function was used in the last activation layer. The loss function was set to the binary cross-entropy and optimized with the Adam algorithm.

The outer LOSO was used to assess the performance of the best set of parameters, retrieved from the grid search (inner LOSO), in the remaining subject of the dataset. The final output metrics, presented in section 11.2, were computed by mapping correct and misclassifications by user, position and learning model. This process enabled the computation of single (cumulative) confusion matrices with respect to each of these parameters, from which all performance metrics were extracted: accuracy (Acc), sensitivity (Se), specificity (Sp), precision (Prec), F1-score (F1), Youden index (YI), and geometric mean of sensitivity and specificity (G). As such, this outer LOSO was paramount to enable the fair comparison of algorithms defined by different input parameters, maintaining complete user-independence in the validation process.

11.1.3 Multiple comparisons

Table 11.2: Different combinations of input parameters tested using the modeling pipeline.

Type of test	**Positions**		**Target rate (Hz)**	**Learning model**
	Train	**Test**		
Baseline	P, B, C, W	P, B, C, W	100	All
Unseen test position	C, B, W	P	100	Random Forest
	P, B, W	C	100	
	P, C, W	B	100	
	P, C, B	W	100	
Single position	P	P	100	Random Forest
	C	C	100	
	B	B	100	
	W	W	100	
Rate variation	P, B, C, W	P, B, C, W	50	Random Forest
	P, B, C, W	P, B, C, W	20	
	P, B, C, W	P, B, C, W	10	
	P, B, C, W	P, B, C, W	5	
	P, B, C, W	P, B, C, W	3	
	P, B, C, W	P, B, C, W	1	

Legend: P–Pocket; B–Belt; C–Chest; W–Wrist

ANOVA multiple comparison analysis was used for comparison of performance metrics between different tests, using vectors of metrics by user obtained from the outer LOSO validation loop as input. As *post-hoc* test, we used the Tukey's Honest Significant Difference test (95% confidence level) between pairs of different learning models, usage positions or sampling rates. These tests aimed the identification of statistically significant differences between different combinations of input parameters (Table 11.2), in order to address the research questions of this work.

All learning models were considered and compared pair-wise for training and testing with all positions at 100 Hz (*Baseline*). For simplicity of analysis, we selected a single model - Random Forest - for conducting all remaining tests, based on the results of the aforementioned comparison and the fact that it is a decision-based classifier. Algorithms based on decision trees are very interpretable, do not require much computation, and are ease to implement in any platform. A more detailed explanation of this selection process is provided in subsection 11.2.1.

11.1.4 Deployment

The output metrics of the LOSO validation in the modeling stage shall assist the process of selecting the most adequate learning model for deployment, considering the requirements of each specific use case, i.e. the selection process should consider performance, complexity and/or other requirements initially set up for the algorithm.

After the selection of the classification algorithm, all data of the AICOS dataset corresponding to the required positions (and resampled to the desired target rate) are used to refit the classifier, with the respective best set of hyperparameters derived from the process of LOSO grid search. This step completes the deployment of a final fall detection algorithm.

To evaluate the effectiveness of our method, we have deployed a fall detector algorithm using a Random Forest classifier, expecting a sampling rate of 10 Hz, and trained with all positions available in the AICOS dataset. This algorithm was then tested using all data from the UMAFall dataset for performance comparison with other fall detection works using the same data.

11.1.5 Benchmark validation using the UMAFall dataset

We benchmarked our framework with the publicly available UMAFall dataset (Casilari et al., 2018) described in Casilari et al. (2017b); Santoyo-Ramón et al. (2018). Several ADLs and simulated falls were collected from 17 volunteers with an average age of 26.7 $\pm$ 10.5 years old. Each subject wore four different wearable devices – chest, belt, wrist, ankle – and carried one smartphone in the pocket. Overall, 11 types of ADLs and 3 types of falls were simulated, yielding a total of 970 falls and 2444 non-falls, with an average of 683 samples for each usage position. Accelerometer, gyroscope and magnetometer data was collected at a sampling rate of 20 Hz from the wearables and 200 Hz from the smartphone.

The UMAFall dataset was selected for its representation of all sensor positions included in AICOS dataset. Interestingly, it also contains data from wearables positioned in a new position –

the ankle –, which our framework is not expecting, and shall thus allow us to assess the generalization of the deployed fall detector for this new usage position.

11.2 Results

11.2.1 Multiple comparisons

Even though we analysed multiple comparisons for several performance metrics, we opted for solely presenting the results for the F1-score for simplicity of analysis, since it was selected as the scoring metric in the optimization process. Moreover, the F1-score will allow us to assess the performance of the algorithm taking into account an harmonic mean of precision and recall.

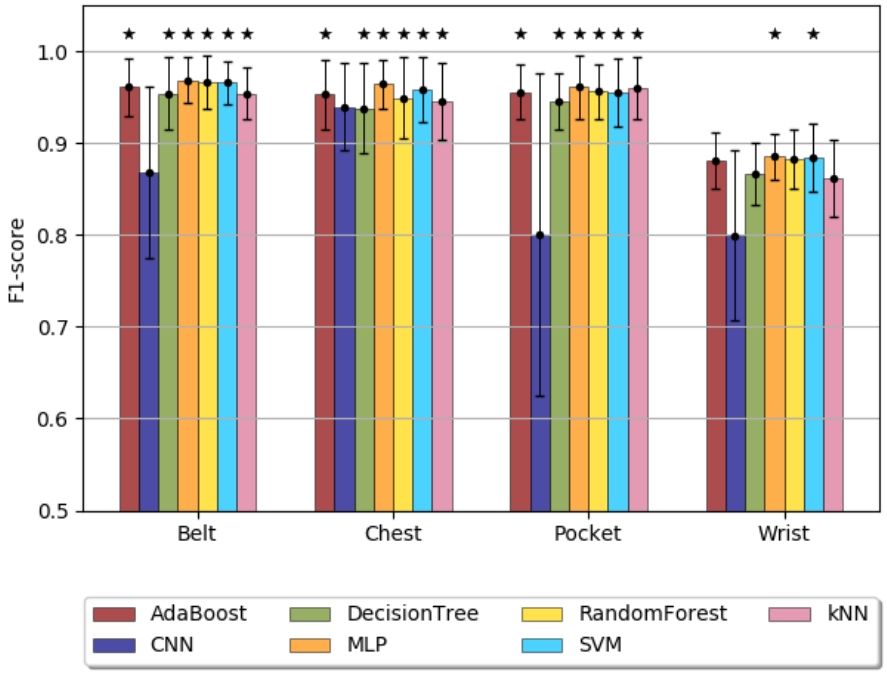

Figure 11.3: F1-score for all tested classifiers, considering the baseline input parameters. Classifiers with SSD from CNN for each sensor position are marked with stars.

The first set of comparisons corresponded to the performance of different learning models for the same set of input parameters (defined as *Baseline* in Table 11.2). Results were arranged by position and classifier and exhibited in Figure 11.3. We looked for statistically significant differences (SSD) between all pairwise combinations of classifiers. No SSD was found among the conventional supervised binary models tested within each position; however, CNN's performance was frequently significantly inferior to that of the remaining models. Given the equivalence of all the conventional models tested, all subsequent experiments were performed using a single classification model. We prioritized decision-based models (Decision Tree and Random Forest) in this selection, due to their low prediction expensiveness which is valuable for wearable implementations. Decision trees are easy to deploy in firmware and are also fast at giving predictions. Random

Forest was finally selected since it consistently led to higher average F1-scores than Decision Tree classifiers.

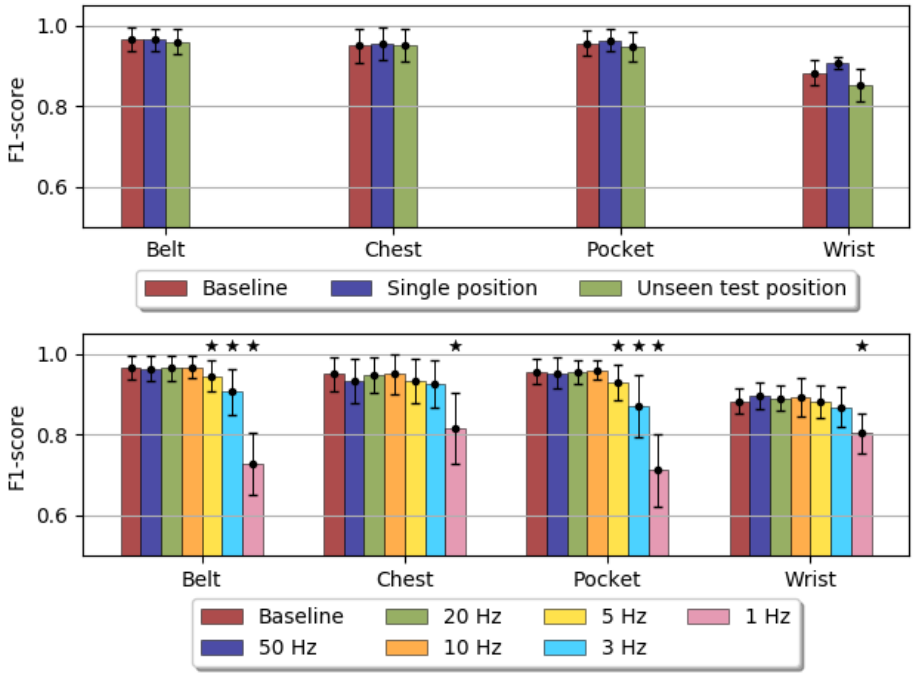

Figure 11.4: F1-score for Random Forest classification, considering the described combinations of input parameters. Pipelines with SSD from *Baseline* for each sensor position are marked with stars.

Figure 11.4 presents the results for different combinations of train/test sensor positions, organised by sensor position of test data. Multiple comparison analysis was performed between results for respective positions derived from setting as input parameters: 1) *Baseline* vs. *Unseen test position*; 2) *Baseline* vs. *Single position*. No SSD were found between either of them. This means that, for example, for 1), the performance of detecting falls in data from sensors in the pocket remains unchanged irrespective of whether data from sensors in this position are included in the training set or not; and, for 2), solely using data from sensors in the pocket for training does not improve the performance falls detection in data from sensors in the pocket, relatively to including data from all the different sensor positions in the training set.

Finally, fall detection performance results using data sampled at different rates are depicted in Figure 11.4. A Random Forest classifier was trained and tested using data from sensors in all the positions available in AICOS dataset and varying the accelerometer sampling rate (*Baseline* and *Rate variation* entries of Table 11.2). Considering rates of 100 Hz, 50 Hz, 20 Hz or 10 Hz did not lead to SSD between the fall detection performance for respective test positions. However, statistically significant decays of performance were verified for belt and pocket positions for rates of 5 Hz and 3 Hz, and, more evidently, for all positions with data sampled at 1 Hz.

11.2.2 Benchmark validation

Table 11.3: Evaluation results with the UMAFall dataset. All performance metrics are in %.

Position	F1	Acc.	Recall	Specificity	Precision	Youden I.	G. Mean	N
B	95.0	96.9	95.2	97.7	94.7	92.9	96.4	617
P	90.2	94.2	92.8	94.7	87.7	87.5	93.7	723
C	90.8	94.9	87.5	97.9	94.3	85.4	92.5	725
W	81.5	87.3	97.1	83.4	70.1	80.5	90.0	726
A	65.3	78.9	78.3	79.1	55.9	57.5	78.7	623
PBCW	88.8	93.2	93.1	93.2	84.9	86.3	93.2	2786
PBCWA	84.6	90.6	90.7	90.5	79.2	81.2	90.6	3414

Legend: P–Pocket; B–Belt; C–Chest; W–Wrist; A–Ankle; N–No. Samples

Table 11.3 combines the results obtained using the framework's model specifically deployed for benchmark validation, as previously described, organised by testing data sensor positions, for all sensor positions included in AICOS dataset (i.e. except ankle), and for the entire dataset. All computed metrics were presented for analysis to instigate further comparisons with previous and future works in the field.

The belt sensor position presented, overall, the best results, immediately followed by pocket and chest positions - the first associated with more false positive occurrences (lower specificity) and the latter associated with more false negative occurrences (lower sensitivity). For data from sensors placed on the wrist a decrease of performance was verified, as compared with the previous positions, which is coherent with the results obtained using the AICOS dataset (Figure 11.3). Finally, considering the testing data from sensors on the ankle yielded the poorest performance for all compared metrics. Combining the samples of all positions, we achieved an F1-score of 84.6%, which increased until 88.8% by not considering the unexpected ankle position.

11.3 Discussion

This section will provide an overall discussion of results, considering general results, CNNs compared with standard techniques, and impact of positions and sampling rate in the performance of the models. Moreover, a state-of-the-art performance comparison will be presented along with potential limitations of this study.

11.3.1 Need for customization

Figures 11.3 and 11.4 provide important information towards understanding the cases worthy of investment in customization.

Starting with the problem of selecting the most adequate learning model, considering trade-offs of performance and available resources in wearable implementations, one can take the results depicted in Figure 11.3, which unveiled that there is no SSD between the performance of all standard binary classification models in our tests for all considered sensor positions. If we describe

model complexity as a function of its consumed resources and prediction expensiveness, one can observe that *there is no evidence that higher model complexity leads to improved results* in the conditions under which these tests took place. This means that selecting the least complex model for implementation may be beneficial for the final system, because it shall lead to lower resource consumption while achieving statistically similar results. If this conclusion is taken under consideration at the moment of system design, there may not be a need to develop several fall detectors with custom learning models to improve performance under different restrictions on the availability of resources.

Figure 11.4 enables a discussion of the role of considering (or not) sole data from the intended place of usage of the sensing device in the training stage. Our tests verified that the *fall detection performance on data from each of the 4 sensor positions is similar, regardless of its inclusion in the training stage*, using AICOS dataset. While this conclusion is not particularly surprising for belt and pocket (both at the waist), or even chest (all in the trunk region), to achieve similar performance for the wrist regardless of its consideration in the training stage is not intuitive. This conclusion can reiterate a claim for position generalization of our method, even though further tests should be conducted to thoroughly understand if there is a more significant impact for other performance metrics. Moreover, to *solely consider data from the intended sensor position to train the models leads to statistically similar results than considering all positions as training data*. As such, it may be beneficial to consider all positions at the modeling stage, regardless of the effective place of usage of the final system, so that its portability is facilitated under different conditions, if needed.

From the rate impact study, one can conclude that the *lowest sampling rate considered that did not present SSD from the baseline 100 Hz pipeline was 10 Hz*. This conclusion appears to be coherent with findings of previous works (Liu et al., 2018), setting a valuable landmark in the field of fall detection towards the efficiency of wearable systems.

11.3.2 State-of-the-art performance

The quality of the AICOS dataset, regarding its variety of usage positions, the representative amount of samples for each position, and expression of relevant different types of falls and non-falls, allowed us to deploy a robust Random Forest classifier trained with all usage positions of this dataset for a target rate of 10 Hz, since no SSD were found between these models and models trained considering higher sampling rates. This process based the conclusion that *our framework is able to deploy models that perform adequately when tested with data acquired under different conditions* (not controlled by the authors), as Table 11.3 corroborates.

The authors of the UMA dataset have achieved their best results for chest and belt (Santoyo-Ramón et al., 2018), comparing with other usage positions, consistently with our findings, to which we can add the pocket position in our case (performance similar to chest and belt). Moreover, the geometric mean achieved in that work was always inferior to 75% for any combination that included the sensor in the ankle, which means that even though our dataset did not feature any sample acquired from the ankle position, our method still outperforms the method of the authors

of the dataset at detecting falls in this position (79% geometric mean for the sole classification of ankle samples).

Directly comparing previous studies with our user-independent approach is, however, a difficult task, since the validation methods previously reported are mostly user-dependent; thus, it is unclear if these methods would lead to the same results under user-independent conditions, typically more challenging. The work of Wang et al. (2018) was the only study found employing LOSO cross validation. Comparing that work with ours, one can verify that our method achieved better results (geometric mean of 91% *vs.* 88%) for all positions considered in the UMA dataset. However, the authors did not explicitly refer if all UMA dataset positions were considered to evaluate their results.

It is also worth mentioning that the model that we have deployed was trained with data downsampled to 10 Hz, instead of using the most frequent sampling rate of the UMA dataset, 20 Hz. In spite of that, *the results obtained with our models are in line with those of other studies using the same data*.

11.3.3 Limitations

These conclusions may not be true for all datasets, but only for datasets similar to AICOS; they are maybe only true due to the quality of our dataset, and the higher amount of samples for each usage position, that allowed us to generalize better to new unseen positions. Moreover, these results were obtained using the F1-score as the optimization score. One can also analyse all of the pipelines' comparisons for other scoring metrics, and the conclusions found with the F1-score may not stand. The model deployed by our framework retrieved from the pipeline described in this study should also be validated with more datasets, and ideally with data from real fall events.

11.4 Conclusion

In this work, we studied the impact of learning models, on-body positioning and sampling rate in fall detection performance, using a new machine learning pipeline which is able to deploy fall detection solutions adapted to the aforementioned system requirements. Our experiments did not verify any relation between model complexity and performance. Moreover, using our dataset and method, considering 3 positions in the training set was enough for achieving model generalization for the 4th (unseen) position, and considering solely data from a certain position *vs.* all positions in the training stage led to statistically similar results when detecting falls at that position. We were also able to decrease the sampling rate expected by our pipeline until 10 Hz without any statistically significant impact in performance.

Finally, we used the UMAFall dataset to benchmark a solution deployed by our framework. This solution is expected to receive data sampled at 10 Hz and uses a Random Forest classifier previously trained with data from AICOS dataset. This experiment unveiled that our solution led to state-of-the-art results for the UMAFall dataset, even under our demanding test conditions

(considering an unseen test position, the ankle; lower sampling rate; test data acquired under conditions not controlled by the authors).

As future work, we can optimize our pipeline for different performance metrics (other than F1-score), to deploy models that require a specific trade-off between sensitivity and specificity. For example, in a specific case or disease it can be more important to detect falls than to have a higher rate of false alarms. This framework will ease the fast deployment of fall detection models that are adjusted to different use cases. After selecting the most suitable model and the target performance metric, we expect to implement our pipeline in a wearable solution to assess the model's performance in free-living conditions.

Part IV

Conclusion

Chapter 12

Conclusions and Future Work

12.1 Conclusions

Fall risk assessment is essential for establishing adequate strategies for fall prevention that could help to revert or attenuate some of the fall risk factors among the elderly population. Recently, researchers in this field have been proposing solutions for fall risk assessment based on low-cost hardware, such as inertial sensors embedded into wearable devices or smartphones. Additionally, there are other solutions based on force and pressure platforms, which aim to assess multiple factors of balance and correlated fall risks. The instrumentation of fall risk tests with these devices allows extracting objective and relevant metrics to help provide more insights about the elderly physical capabilities. Although fall prediction systems can contribute to preventing falls, the occurrence of falls is not only dependent on the physical stability of the users but also on external perturbations. For this reason, the occurrence of falls is not always predictable. Therefore, it is important to detect falls at the moment they occur, because immediate assistance after a fall could decrease its negative effects. Automatic fall detection systems have been developed in the past years and rely mostly on devices with integrated inertial sensors and location capabilities that facilitate the detection and triggering of a fall alert.

This thesis focused on the study of a **multifactorial fall prediction system** and the study of a **wearable-based automatic fall detection system**.

For the **multifactorial fall prediction system**, we focused on the study of feature extraction methods based on the instrumentation of fall risk assessment tests and the study of data fusion procedures to combine data sources, such as clinical, self-reported and sensor-retrieved data. The main conclusions for the fall prediction study are detailed as follows:

- We employed signal processing methods for segmentation and feature extraction from sensor data collected during the execution of instrumented functional tests. We presented the data processing methods for extracting features from inertial sensors during the execution of the iTUG test. For a small group of persons, we divided higher and lower fall risk persons based on the POMA and TUG test scores. The number of previous falls did not show a significant difference between high and low risk groups, whereas the sensors' features showed

significant differences between the high and the low fall risk groups. Although the sample size was small, this study contributed towards the identification of a set of sensor-based features that have higher predictive power than standard test scores or self-reported data (Silva and Sousa, 2016).

- The previous study was extended by proposing signal processing methods for other functional tests and we also included features extracted from a pressure platform. We compared the performance of functional test scores with features obtained from inertial and pressure platform sensors to discriminate between persons with low and high fall levels. To differentiate the two groups we proposed a fall level, defined by us, that combines previous falls questionnaire and the need for using a walking aid because these two variables have a higher associated fall risk. Using the features extracted from the inertial sensors and pressure platform we obtained better results for the same machine learning algorithms than using only test scores. The Naïve Bayes classifier obtained an accuracy of 87.16% (88.18% of precision and 97.50% of recall). We concluded that the added value of metrics derived from wearable devices and the pressure platform has the potential to improve fall prediction systems (Silva et al., 2017).

- We proposed a new multifactorial screening protocol for individuals aged above 50 years old living in the community. The protocol combines clinical data, self-reported data, and data from wearable inertial sensors and a pressure platform, which were used to instrument six fall risk assessment tests: Timed-Up and Go Test, 30 seconds Sit-To-Stand test, 10-meter Walking Speed test, "Modified" 4-Stage Balance test, Step test, and Handgrip Strength test (Martins et al., 2018). This screening protocol was applied to 403 participants living in Portugal, however, only a part of the population aged over 65 years old was taken into consideration for analysis, resulting in a total of 281 participants. We were able to follow them during 12-months to register their monthly occurrence of falls.

- The collected information allowed us to compare three data fusion approaches for prospective fall prediction based on the analysis of multimodal data collected according to the predefined protocol. The richness of the collected data allowed to infer not only the functional capabilities of a person but also clinical and environmental information. In this study, we employed an oversampling technique to deal with the unbalanced nature of the collected dataset, a procedure that is rarely reported in the literature for fall prediction. Furthermore, we divided the dataset into a training set and a hold-out test set for model validation, something that has been lacking in previous research. We investigated the impact of fusing data at different stages of the machine learning pipeline on the obtained results. We considered the recall of the system more important than specificity since we foresee that correctly identifying a higher risk person is more relevant than the opposite. The early, late and slow data fusion approaches revealed similar results in terms of fall prediction performance. To the best of our knowledge, no previously published work has attempted to study different approaches to data fusion using multiple sources of data for prospective fall prediction. The

study also focused on several optimization stages in the ML pipeline, and the final results are presented for a test set that was not considered during the evaluation of the trained models. The result of the late fusion approach providing a recall of 78.6% is better compared with the results achieved by the other two approaches (Silva et al., 2020).

For the development of an **automatic fall detection system**, we considered the impact of several variables in the performance of the system: the type of dataset, composed of simulated or real-world data, the on-body positions to place the wearable device, and restrictions related with the deployment hardware, such as sampling rate, algorithm's sensitivity level, and models complexity. The main conclusions for the several analyses made are described below:

- A transfer learning approach was presented for combining a dataset of simulated falls and non-falls with the real-world FARSEEING dataset. Since most of the previous studies have used simulated falls to develop the models, we studied the combination of simulated data, acquired with younger and more active volunteers, with a small dataset of real-world falls, acquired from hospitalized older persons. The combination of simulated and real-world data allowed to train a set of supervised classifiers for discriminating between falls and non-falls. In the real-world, falls are a sporadic event, which results in imbalanced datasets. To overcome this issue, three methods of imbalanced learning were employed. The accuracy obtained was very similar, but the Balance Cascade obtained fewer misclassifications in the test set. Combined sets of simulated and real falls presented advantages compared to using only simulated falls. There is an improvement when mixing datasets compared to the case when only simulated falls were used for training. When models are tested with a mixed set it is indeed more important to train with a mixed set. We have also concluded that a model trained with simulated falls generalizes better when tested with real falls than the opposite. The sensitivity obtained outperformed the one reported in the state-of-art with the FARSEEING dataset by 10% (Silva et al., 2018).

- It was proposed a reliable, simple, and wearable solution for automatic fall detection, that can be adapted to different groups of people with different fall risk levels. Also, it can be used in different on-body positions, chest, pocket, and waist, without requiring any calibration step, which makes this system less intrusive and easier to use. The proposed solution is based on a state machine algorithm and an optimization routine that allowed the model to be adapted to different sensitivity levels, different sampling rates, and different on-body positions. We have taken into account hardware constraints that require a very simple algorithm and impose implementation approximations for developing the algorithm. We concluded that when decreasing the sampling rate of the accelerometer from 100 to 50 Hz, the performance of the algorithm improves by 3 to 4% in the J index metric, for the three sensitivity levels. We collected a continuous dataset of falls and ADLs to validate the algorithm in a more realist approach. The algorithm trained with a sampling rate of 50 Hz obtained 6 false alarms per day for the higher sensitivity level and 0.4 false alarms per day for the lower sensitivity level, which compares favorably with previous works (Alves et al., 2019).

- Finally, it was studied the impact of learning models, on-body positions and sampling rate in fall detection performance, using a new machine learning pipeline that is able to deploy fall detection algorithms adapted to the aforementioned system requirements. We propose a new machine learning pipeline, trained with our proprietary AICOS dataset, with a customizable modeling stage which enabled the assessment of performance over each combination of custom parameters. By using our AICOS dataset, we did not find any evidence that higher model complexity leads to higher performance. Moreover, using our dataset and pipeline, considering three positions in the training set was enough for achieving model generalization for the fourth unseen position. We also concluded that considering solely data from a certain position vs. all positions in the training stage led to statistically similar results when detecting falls at that position. We were also able to decrease the sampling rate expected by our pipeline until 10 Hz without any statistically significant impact in performance. The validation of this pipeline in the publicly available UMAFall dataset allowed the model to generalize for a new unseen position that was not part of the training set, and also outperformed the previous state-of-art in terms of fall detection performance (geometric mean of 91% vs. 88%), considering a user-independent validation.

12.2 Future work

In the future, the work presented in the area of fall prediction could be used as a standard multifactorial fall prediction tool based on wearable devices, to provide a standard protocol to assess elderly fall risk in the community. The added value of features extracted from wearable sensors could enhance the healthcare professional assessment of physical conditions such as balance, mobility and strength abilities, as well as personal and contextual information. The next step will be to undertake the validation of such a system in a long-term trial, together with an industry partner and a healthcare professional institution, to validate if the fall prediction is correlated with the prospective falls occurrence. In terms of research work, we still have open issues to discuss, concerning the study of other preprocessing methods for different data sources, and the addition of variables that better explain the unexpected nature of a fall event, that could also be used as the output of such fall prediction systems. Other possible outcomes are the predicted time until the first fall (in months), or the probability of suffering a fall in a given period during the year after the assessment. The problem can also be formulated as multiclass classification, allowing to distinguish groups of first-time fallers after the assessment, recurrent fallers, and non-fallers, for example. Moreover, we also consider combining both the fall prediction and the fall detection systems, in order to provide automatic detection of falls that can be used to validate the fall prediction estimation.

In the area of fall detection, the work conducted towards the development of a low-power wearable-based solution for automatic fall detection and the work developed for the automation of the development of such systems taking into account several constraints will certainly be used in future projects. We have made considerable efforts for transferring such technology to indus-

try partners, we prospect the use of such systems for different use cases, as the use in different hardware devices, or to be adapted to different on-body positions or use cases. Moreover, the automation of such fall detection systems will allow us in the future to accelerate the deployment of a fall detection system and to accelerate the time to prototype in further projects. We have also been in contact with other partners to conduct a long-term trial for assessing this technology in real-world conditions. This is one of the main goals for future work, to provide validation for this technology in long-term trials with elderly, to assess in one hand the performance of the automatic fall detection system and to validate the use cases previously defined, and in the other hand, to collect data from real-world falls.

Bibliography

"Hall Effect Sensor and How Magnets Make It Works," Aug. 2013. [Online]. Available: http://www.electronics-tutorials.ws/electromagnetism/hall-effect.html

"Naive bayes classifier," 2017. [Online]. Available: http://www.statsoft.com/textbook/naive-bayes-classifier

Age UK. Stop Falling, "Start saving lives and money," Age UK, Available online, Tech. Rep., 2013. [Online]. Available: http://www.ageuk.org.uk/documents/en-gb/campaigns/stop_falling_report_web.pdf?dtrk=true

Agency for Healthcare Research and Quality, "Tool 3G: STRATIFY Scale for Identifying Fall Risk Factors: Preventing Falls in Hospitals: A Toolkit for Improving Quality of Care." Jan. 2013. [Online]. Available: https://www.ahrq.gov/professionals/systems/hospital/fallpxtoolkit/fallpxtk-tool3g.html

Y. Agrawa, J. P. Carey, H. J. Hoffman, D. A. Sklare, and M. C. Schubert, "The modified romberg balance test: Normative data in U.s. adults," *Otology & Neurotology*, vol. 32, no. 8, pp. 1309–11, 2011.

B. Aguiar, T. Rocha, J. Silva, and I. Sousa, "Accelerometer-based fall detection for smartphones," in *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2014, pp. 1–6.

B. Aguiar, J. Silva, T. Rocha, S. Carneiro, and I. Sousa, "Monitoring physical activity and energy expenditure with smartphones," in *IEEE-EMBS BHI*, June 2014, pp. 664–667.

A. N. Aicha, G. Englebienne, K. S. van Schooten, M. Pijnappels, and B. J. A. Kröse, "Deep learning to predict falls in older adults based on daily-life trunk accelerometry," in *Sensors*, 2018.

AICOS, "A day with pandlets," Fraunhofer Portugal, White Paper, 2016. [Online]. Available: https://www.aicos.fraunhofer.pt/en/news_and_events_aicos/news_archive/older_archive/-a-day-with-pandlets---the-most-recent-white-paper-from-fraunhof.html

F. AICOS, "TSFEL: time series feature extraction library," https://github.com/fraunhoferportugal/tsfel, 2019.

J. Alves, J. Silva, E. Grifo, C. Resende, and I. Sousa, "Wearable embedded intelligence for detection of falls independently of on-body location," *Sensors*, vol. 19, no. 11, 2019.

A. F. Ambrose, G. Paul, and J. M. Hausdorff, "Risk factors for falls among older adults: a review of the literature." *Maturitas*, vol. 75 1, pp. 51–61, 2013.

K. G. Avin, T. A. Hanke, N. Kirk-Sanchez, C. M. McDonough, T. E. Shubert, J. Hardage, and G. Hartley, "Management of falls in community-dwelling older adults: clinical guidance statement from the Academy of Geriatric Physical Therapy of the American Physical Therapy Association," *Physical therapy*, vol. 95, no. 6, p. 815—834, June 2015.

O. Aziz, J. Klenk, L. Schwickert, L. Chiari, C. Becker, E. J. Park, G. Mori, and S. N. Robinovitch, "Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets," *PLOS ONE*, vol. 12, no. 7, pp. 1–11, 07 2017.

F. Bagalá, C. Becker, A. Cappello, L. Chiari, K. Aminian, J. M. Hausdorff, W. Zijlstra, and J. Klenk, "Evaluation of accelerometer-based fall detection algorithms on real-world falls," *PLoS ONE*, 2012.

S. Barri Khojasteh, J. Villar, E. Marín, V. González, and C. Chira, "Comparing model performances applied to fall detection," *International Conference On Mathematical Applications*, 2018.

O. Beauchet, B. Fantino, G. Allali, S. W. Muir, M. Montero-Odasso, and C. Annweiler, "Timed up and go test and risk of falls in older adults: A systematic review," *The Journal of Nutrition, Health & Aging*, vol. 15, no. 10, pp. 933–38, 2011.

C. Becker, L. Schwickert, S. Mellone, F. Bagalà, L. Chiari, J. Helbostad, W. Zijlstra, K. Aminian, A. Bourke, C. Todd, S. Bandinelli, N. Kerse, and J. Klenk, "Proposal for a multiphase fall model based on real-world fall recordings with body-fixed sensors," *Zeitschrift für Gerontologie und Geriatrie*, vol. 45, no. 8, pp. 707–715, 2012.

C. Becker and S. Lamb, "Profane manual for the fall prevention classification system," Apr. 2007.

K. Berg, S. L Wood-Dauphinee, J. I Williams, and B. Maki, "Measuring balance in the elderly: Validation of an instrument," *Canadian journal of public health. Revue canadienne de santé publique*, vol. 83 Suppl 2, pp. S7–11, 11 1991.

G. Bergen, M. R. Stevens, and E. R. Burns, "Falls and fall injuries among adults aged $\geq$ 65 years - united states, 2014," *MMWR. Morbidity and Mortality Weekly Report*, vol. 65, no. 37, pp. 993–98, 2016.

K. E. Bigelow and N. Berme, "Development of a protocol for improving the clinical utility of posturography as a fall-risk screening tool." *The journals of gerontology. Series A, Biological sciences and medical sciences*, vol. 66 2, pp. 228–33, 2011.

J. Bouchaud, "Accelerometers Set to Become Leading MEMS Device in 2013 - IHS Technology," 2009. [Online]. Available: https://technology.ihs.com/389196/accelerometers-set-to-become-leading-mems-device-in-2013

A. Bourke, P. van de Ven, M. Gamble, R. O'Connor, K. Murphy, E. Bogan, E. McQuade, P. Finucane, G. Olaighin, and J. Nelson, "Evaluation of waist-mounted tri-axial accelerometer based fall-detection algorithms during scripted and continuous unscripted activities," *J. Biomech*, pp. 3051–3057, 2010.

A. K. Bourke, J. Klenk, L. Schwickert, K. Aminian, E. A. F. Ihlen, J. L. Helbostad, L. Chiari, and C. Becker, "Temporal and kinematic variables for real-world falls harvested from lumbar sensors in the elderly population," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5183–5186.

A. K. Bourke, J. Klenk, L. Schwickert, K. Aminian, E. A. F. Ihlen, S. Mellone, J. L. Helbostad, L. Chiari, and C. Becker, "Fall detection algorithms for real-world falls harvested from lumbar sensors in the elderly population: A machine learning approach," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 3712–3715.

K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011.

S. M. Bruijn, O. G. Meijer, P. J. Beek, and J. H. van Dieën, "Assessing the stability of human locomotion: A review of current measures," *J. R. Soc. Interface*, 2013.

E. R. Burns, J. A. Stevens, and R. Lee, "The direct costs of fatal and non-fatal falls among older adults—United States," *J. Safety Res*, pp. 99–103, 2016.

E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "UMAFall: A multisensor dataset for the research on automatic fall detection," *Procedia Computer Science*, vol. 110, pp. 32 – 39, 2017, 14th MobiSPC 2017 / 12th FNC 2017 / Affiliated Workshops.

E. Casilari, J.-A. Santoyo-Ramón, and J.-M. Cano-García, "Analysis of public datasets for wearable fall detection systems," *Sensors*, vol. 17, no. 1513, 2017.

E. Casilari, J. Santoyo-Ramón, and J. Cano-García, "UMAFall dataset," https://doi.org/10.6084/m9.figshare.4214283.v7, 2018.

L. Cattelani, P. Palumbo, L. Palmerini, S. Bandinelli, C. Becker, F. Chesani, and L. Chiari, "Fratup, a web-based fall-risk assessment tool for elderly people living in the community," *Journal of medical Internet research*, vol. 17, p. e41, 02 2015.

K. Chaccour, R. Darazi, A. H. El Hassani, and E. Andrès, "From fall detection to fall prevention: A generic classification of fall-related systems," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 812–822, Feb 2017.

N. Chawla, K. Bowyer, L. O. Hall, and W. Philip Kegelmeyer, "Smote: Synthetic minority oversampling technique," *JAIR*, vol. 16, pp. 321–357, 01 2002.

K. H. Cho, S. K. Bok, Y.-J. Kim, and S. L. Hwang, "Effect of lower limb strength on falls and balance of the elderly," *Annals of Rehabilitation Medicine*, vol. 36, no. 3, pp. 386–93, 2012.

M. Chung, R. Chan, Y.-K. Fung, S. Fong, S. Lam, C. Lai, and S. Ng, "Reliability and validity of alternate step test times in subjects with chronic stroke." *Journal of rehabilitation medicine*, vol. 46, 08 2014.

CMSIS-DSP, *CMSIS-DSP Software Library*. Available online: ARM, 2015. [Online]. Available: http://www.keil.com/pack/doc/CMSIS/DSP/html/index.html

R. Cruz, K. Fernandes, J. S. Cardoso, and J. F. P. Costa, "Tackling class imbalance with ranking," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 2182–2187.

J. E. Dayhoff and J. M. DeLeo, "Artificial neural networks," *Cancer*, vol. 91, no. S8, pp. 1615–1635, 2001.

L. DePasquale, "Fall Prevention: Current Perspectives, Tools with Evidence," Apr. 2014. [Online]. Available: https://www.physicaltherapist.com/articles/fall-prevention-current-perspectives-tools-with-evidence/

T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

E. Doheny, C. W. Fan, T. Foran, B. Greene, C. Cunningham, and R. Kenny, "An instrumented sit-to-stand used to examine differences between older fallers and non-fallers," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2011, pp. 3063–6, 08 2011.

E. P. Doheny, D. McGrath, B. R. Greene, L. Walsh, D. J. McKeown, C. Cunningham, L. Crosby, R.-A. Kenny, and B. Caulfield, "Displacement of centre of mass during quiet standing assessed using accelerometry in older fallers and non-fallers," *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3300–3303, 2012.

E. P. Doheny, C. Walsh, T. Foran, B. R. Greene, C. W. Fan, C. Cunningham, and R. A. M. Kenny, "Falls classification using tri-axial accelerometers during the five-times-sit-to-stand test." *Gait & posture*, vol. 38 4, pp. 1021–5, 2013.

M. J. Faber, R. J. Bosscher, and P. C. v. Wieringen, "Clinimetric Properties of the Performance-Oriented Mobility Assessment," *Physical Therapy*, vol. 86, no. 7, pp. 944–954, Jul. 2006.

M. Feil and L. A. Gardner, "Falls Risk Assessment: A Foundational Element of Falls Prevention Programs," *Pennsylvania Patient Safety Advisory*, pp. 73–81, 2012.

E. Fosler-Lussier, "Markov models and hidden markov models: A brief tutorial," *International Computer Science Institute*, 1998.

S. Fritz and M. Lusardi, "White Paper: "Walking Speed: the Sixth Vital Sign"," *Journal of Geriatric Physical Therapy*, vol. 32, no. 2, 2009.

L. Gao, A. K. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," in *Med. Eng. Phys*, 2014, pp. 779–785.

Z. Ghahramani, *Unsupervised Learning*. Berlin, Heidelberg: Advanced Lectures on Machine Learning: ML Summer Schools, 2004, pp. 72–112.

H. Gjoreski, M. Lustrek, and M. Gams, "Accelerometer placement for posture recognition and fall detection," in *Proceedings of the Seventh International Conference on Intelligent Environments*. UK, pp. 47–54: Nottingham, July 2011, pp. 25–28.

B. R. Greene, "Wireless sensor based quantitative falls risk assessment," Worldwide Patent US8 805 641 B2, 2014.

B. Greene, A. O'Donovan, R. Romero-Ortuno, L. Cogan, C. Ni Scanaill, and R. Kenny, "Quantitative falls risk assessment using the timed up and go test," *IEEE transactions on bio-medical engineering*, vol. 57, pp. 2918–26, 10 2010.

B. Greene, K. Mcmanus, and B. Caulfield, "Automatic fusion of inertial sensors and clinical risk factors for accurate fall risk assessment during balance assessment," in *IEEE Conference on Biomedical and Health Informatics (BHI) 2018At: Las Vegas, Nevada*, 03 2018, pp. 1–4.

B. R. Greene, D. McGrath, L. Walsh, E. P. Doheny, D. J. McKeown, C. Garattini, C. Cunningham, L. Crosby, B. Caulfield, and R. A. M. Kenny, "Quantitative falls risk estimation through multi-sensor assessment of standing balance." *Physiological measurement*, vol. 33 12, pp. 2049–63, 2012.

P. D. Groves, "Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems - Chapter 4: Inertial Sensors | Engineering360," 2008. [Online]. Available: http://www.globalspec.com/reference/74016/203279/chapter-4-inertial-sensors

V. Guimarães, D. Ribeiro, L. Rosado, and I. Sousa, "A smartphone-based fall risk assessment tool: Testing ankle flexibility, gait and voluntary stepping," *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, 2014.

V. Guimarães, D. Ribeiro, and L. Rosado, "A smartphone-based fall risk assessment tool: Measuring one leg standing, sit to stand and falls efficacy scale," in *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, Oct 2013, pp. 529–533.

S. Hempel, S. Newberry, Z. Wang, P. G. Shekelle, R. M. Shanman, B. Johnsen, T. Perry, D. Saliba, and D. A. Ganz, "Review of the Evidence on Falls Prevention in Hospitals," 2012. [Online]. Available: http://www.rand.org/pubs/working_papers/WR907.html

K. Hill, "A new test of dynamic standing balance for stroke patients : Reliability, validity and comparison with healthy elderly," *Physiotherapy Canada*, vol. 48, pp. 257–262, 1996.

M. Hofheinz and C. Schusterschitz, "Dual task interference in estimating the risk of falls and measuring change: a comparative, psychometric study of four measurements," *Clinical Rehabilitation*, vol. 24, no. 9, pp. 831–842, Sep. 2010.

J. Howcroft, J. Kofman, and E. D. Lemaire, "Review of fall risk assessment in geriatric populations using inertial sensors," *Journal of NeuroEngineering and Rehabilitation*, vol. 10, no. 1, p. 91, Aug 2013.

W.-L. Hsi, "Analysis of medial deviation of center of pressure after initial heel contact in forefoot varus," *Journal of the Formosan Medical Association - Taiwan Yi Zhi*, vol. 115, no. 3, pp. 203–9, 2016.

Q. T. Huynh, U. D. Nguyen, L. B. Irazabal, N. Ghassemian, and B. Q. Tran, "Optimization of an accelerometer and gyroscope-based fall detection algorithm," *J. Sens*, 2015.

R. Igual, C. Medrano, and I. Plaza, "Challenges, issues and trends in fall detection systems," *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 66, 2013.

C. J. Jones, R. E. Rikli, and W. C. Beam, "A 30-s chair-stand test as a measure of lower body strength in community-residing older adults," *Research Quarterly for Exercise and Sport*, vol. 70, no. 2, pp. 113–19, 1999.

M. Kangas, I. Vikman, J. Wiklander, P. Lindgren, L. Nyberg, and T. Jamsa, "Sensitivity and and specificity of fall detection in people aged 40 years and over," *Gait Posture*, pp. 571–574, 2009.

M. Kangas, I. Vikman, L. Nyberg, R. Korpelainen, J. Lindblom, and T. Jämsä, "Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects," *Gait & Posture*, vol. 35, no. 3, pp. 500 – 505, 2012.

M. Kangas, A. Konttila, P. Lindgren, I. Winblad, and T. Jämsä, "Comparison of low-complexity fall detection algorithms for body attached accelerometers," *Gait & Posture*, vol. 28, no. 2, pp. 285 – 291, 2008.

A. Karpathya, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE CVPR'14*, 2014, pp. 1725–1732.

S. S. Khan and J. Hoey, "Review of fall detection techniques: A data availability perspective," in *Med. Eng. Phys*, 2017, vol. 39, pp. 12–22.

J. Klenk, C. Becker, F. Lieken, S. Nicolai, W. Maetzler, W. Alt, W. Zijlstra, J. Hausdorff, R. van Lummel, L. Chiari, and U. Lindemann, "Comparison of acceleration signals of simulated and real-world backward falls," *Medical Engineering & Physics*, vol. 33, no. 3, pp. 368 – 373, 2011.

J. Klenk, L. Schwickert, L. Palmerini, S. Mellone, A. Bourke, E. Ihlen, N. Kerse, K. Hauer, M. Pijnappels, M. Synofzik, K. Srulijes, W. Maetzler, J. Helbostad, W. Zijlstra, K. Aminian, C. Todd, L. Chiari, and C. Becker, "The FARSEEING real-world fall repository: a large-scale collaborative database to collect and share sensor signals from real-world falls," *European Review of Aging and Physical Activity*, vol. 13, no. 1, p. 8, 2016.

R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 1137–1143.

S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering*. IOS Press, 2007, p. 3–24.

C. Krupitzer, T. Sztyler, J. Edinger, M. Breitbach, H. Stuckenschmidt, and C. Becker, "Beyond position-awareness: Extending a self-adaptive fall detection system," *Pervasive and Mobile Computing*, vol. 58, p. 101026, 2019.

N. Leaper, "A visual guide to CRISP-DM methodology," 2009. [Online]. Available: https://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf

G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

Q. Li, J. Stankovic, M. Hanson, A. T. Barth, J. Lach, and G. A. Zhou, "Fast fall detection using gyroscopes and accelerometer-derived posture information," *In Proceedings of the IEEE International Workshop on Wearable and Implantable Body Sensor Networks, Berkerey, CA, USA, 3–5*, vol. 6, pp. 138–143, June 2009.

U. Lindemann, A. Hock, M. Stuber, W. Keck, and C. Becker, "Evaluation of a fall detector based on accelerometers: A pilot study," *Medical and Biological Engineering and Computing*, vol. 43, no. 5, pp. 548–551, 2005.

K.-C. Liu, C.-Y. Hsieh, S. J.-P. Hsu, and C.-T. Chan, "Impact of sampling rate on wearable-based fall detection systems based on machine learning models," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9882–9890, 2018.

X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

Y. Liu, S. J. Redmond, T. Shany, J. Woolgar, M. R. Narayanan, S. R. Lord, and N. H. Lovell, "Validation of an accelerometer-based fall prediction model," in *2014 36th IEE-EMBS*, Aug 2014, pp. 4531–4534.

Y. Liu, S. J. Redmond, N. Wang, F. Blumenkron, M. R. Narayanan, and N. H. Lovell, "Spectral analysis of accelerometry signals from a directed-routine for falls-risk estimation," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 2308–2315, 2011.

S. Lord, H. Menz, and A. Tiedemann, "A physiological profile approach to falls risk assessment and prevention," *Physical therapy*, vol. 83, pp. 237–52, 04 2003.

A. C. Martins, J. Moreira, C. Silva, J. Silva, C. Tonelo, D. Baltazar, C. Rocha, T. Pereira, and I. Sousa, "Multifactorial screening tool for determining fall risk in community-dwelling adults aged 50 years or over (fallsensing): Protocol for a prospective study," *JMIR Res Protoc*, vol. 7, no. 8, p. e10304, Aug 2018.

S. Mathias, U. S. L. Nayak, and B. W. Isaacs, "Balance in elderly patients: the "get-up and go" test." *Archives of physical medicine and rehabilitation*, vol. 67 6, pp. 387–9, 1986.

C. Medrano, R. Igual, I. Plaza, and M. Castro, "Detecting falls as novelties in acceleration patterns acquired with smartphones," *PLOS ONE*, vol. 9, no. 4, pp. 1–9, 04 2014.

M. MEMS and N. Exchange, "What is MEMS Technology?" [Online]. Available: https://www.mems-exchange.org/MEMS/what-is.html

A. Moujahid, "A practical introduction to deep learning with caffe and python," Data Analytics and more, 2017. [Online]. Available: http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/

M. Mubashir, L. Shao, and L. A. Seed, "Survey on fall detection: Principles and approaches," *Neurocomputing*, pp. 144–152, 2013.

M. A. Murphy, S. L. Olson, E. J. Protas, and A. R. Overby, "Screening for falls in community-dwelling elderly," *Journal of Aging and Physical Activity*, vol. 11, no. 1, pp. 66–81, 2003.

M. R. Narayanan, S. J. Redmond, M. E. Scalzi, S. R. Lord, B. G. Celler, and N. H. Lovell, "Longitudinal falls-risk estimation using triaxial accelerometry," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 534–541, 2010.

A. Nelson, K. L. Perell, L. Z. Rubenstein, N. Prieto-Lewis, R. L. Goldman, and S. L. Luther, "Fall Risk Assessment Measures: An Analytic Review," *The Journals of Gerontology: Series A*, vol. 56, no. 12, pp. M761–M766, 12 2001.

R. A. Newton, "Balance screening of an inner city older adult population," *Archives of Physical Medicine and Rehabilitation*, vol. 78, no. 6, pp. 587–591, Jun. 1997.

A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural Inf. Process. Syst*, vol. 14, 04 2002.

N. Noury, A. Fleury, P. Rumeau, A. Bourke, G. Laighin, V. Rialle, and J. E. Lundy, "Fall detection–principles and methods," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 29, pp. 1663–1666, August 2007.

D. Oliver, F. Daly, F. C. Martin, and M. E. T. McMurdo, "Risk factors and risk assessment tools for falls in hospital in-patients: A systematic review," *Age and Ageing*, vol. 33, no. 2, pp. 122–30, 2004.

A. T. Ozdemir, "An analysis on sensor locations of the human body for wearable fall detection devices: Principles and practice," *Sensors (Basel, Switzerland)*, 2016.

A. T. Özdemir and B. Barshan, "Detecting falls with wearable sensors using machine learning techniques," *Sensors (Basel, Switzerland)*, pp. 10 691–10 708, 2014.

P. Palumbo, L. Palmerini, and L. Chiari, "A probabilistic model to investigate the properties of prognostic tools for falls." *Methods of information in medicine*, vol. 54 2, pp. 189–97, 2014.

N. Pannurat, S. Thiemjarus, and E. A. Nantajeewarawat, "Hybrid temporal reasoning framework for fall monitoring," *IEEE Sensors J*, pp. 1749–1759, 2017.

N. Pannurat, S. Thiemjarus, and E. Nantajeewarawat, "Automatic fall monitoring: A review," *Sensors*, vol. 14, no. 7, pp. 12 900–12 936, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

L. Pizzigalli, M. Micheletti Cremasco, A. Mulasso, and A. Rainoldi, "The contribution of postural balance analysis in older adult fallers: A narrative review," *Journal of Bodywork and Movement Therapies*, vol. 20, no. 2, pp. 409–417, Apr. 2016.

J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines." Microsoft Research MSR-TR-98-14, 1998.

R. Pontius and R. Si, "The total operating characteristic to measure diagnostic ability for multiple thresholds," *Int. J. Geogr. Inform. Sci*, pp. 570–583, 2014.

H. Qiu, R. Z. U. Rehman, X. J. Yu, and S. Xiong, "Application of wearable inertial sensors and a new test battery for distinguishing retrospective fallers from non-fallers among community-dwelling older people," in *Scientific Reports*, 2018.

J. A. Raymakers, M. M. Samson, and H. J. J. Verhaar, "The assessment of body sway and the choice of the stability parameter(s)," *Gait & Posture*, vol. 21, no. 1, pp. 48–58, 2005.

S. Redmond, M. Scalzi, M. Narayanan, S. Lord, S. Cerutti, and N. Lovell, "Automatic segmentation of triaxial accelerometry signals for falls risk estimation," in *Conf Proc IEEE Eng Med Biol Soc.*, 2010, pp. 2234–2237.

L. Ren and Y. Peng, "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, vol. 7, pp. 77 702–77 722, 2019.

R. E. Rikli and J. C. Jones, "Functional fitness normative scores for community-residing older adults, ages 60–94," *Human Kinetics Journals*, vol. 21, April 2010.

R. E. Rikli and C. J. Jones, "Development and validation of criterion-referenced clinically relevant fitness standards for maintaining physical independence in later years," *The Gerontologist*, vol. 53, no. 2, pp. 255–267, Apr. 2013.

D. J. Rose, J. C. Jones, and N. Lucchese, "Predicting the probability of falls in community-residing older adults using the 8-foot up-and-go: A new measure of functional mobility," *Journal of Ageing and Physical Activity*, vol. 10, no. 4, pp. 466–75, 2002.

J. E. Rossiter-Fornoff, S. L. Wolf, L. I. Wolfson, and D. M. Buchner, "A cross-sectional validation study of the ficsit common data base static balance measures. frailty and injuries: Cooperative studies of intervention techniques," *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, vol. 50, no. 6, pp. M291–297, 1995.

L. Z. Rubenstein, "Falls in older people: Epidemiology, risk factors and strategies for prevention," *Age and Ageing*, vol. 35, 2006.

L. Z. Rubenstein and K. R. Josephson, "The epidemiology of falls and syncope," *Clinics in Geriatric Medicine*, vol. 18, no. 2, pp. 141–58, 2002.

H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014.

A. Salarian, F. B. Horak, C. Zampieri, P. Carlson-Kuhta, J. G. Nutt, and K. Aminian, "iTUG, a sensitive and reliable measure of mobility," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 3, pp. 303–310, June 2010.

J. A. Santoyo-Ramón, E. Casilari, and J. M. Cano-García, "Analysis of a smartphone-based architecture with multiple mobility sensors for fall detection with supervised learning," *Sensors*, vol. 18, no. 4, 2018.

R. Schapire, *Explaining AdaBoost*, 10 2013, pp. 37–52.

R. Schwendimann, S. D. Geest, and K. Milisen, "Evaluation of the Morse Fall Scale in hospitalised patients," *Age and Ageing*, vol. 35, no. 3, pp. 311–313, May 2006.

L. Schwickert, C. Becker, U. Lindemann, C. Maréchal, A. Bourke, L. Chiari, J. L. Helbostad, W. A. Zijlstra, K. Todd, and C., "Fall detection with body-worn sensors: A systematic review," *Z. Gerontol. Geriatr*, pp. 706–719, 2013.

V. Scott, K. Votova, A. Scanlan, and J. Close, "Multifactorial and functional mobility assessment tools for fall risk among older adults in community, home-support, long-term and acute care settings," *Age and Ageing*, vol. 36, no. 2, pp. 130–39, 2007.

Sensing Future Technologies, "PhysioSensing balance and pressure plate," Sensing Future Technologies, Tech. Rep., 2018. [Online]. Available: https://www.physiosensing.net/

A. Shahza and K. F. Kim, "An automated smart phone based fall detection system using multiple kernel learning," *IEEE Trans. Ind. Inform*, pp. 35–44, 2019.

T. Shany, S. Redmond, M. Narayanan, and N. Lovell, "Sensors-Based Wearable Systems for Monitoring of Human Movement and Falls," *Sensors Journal, IEEE*, vol. 12, no. 3, pp. 658–670, Mar. 2012.

T. Shany, K. Wang, Y. Liu, and N. Lovell, "Review: Are we stumbling in our quest to find the best predictor? over-optimism in sensor-based models for predicting falls in older adults," *IET Healthcare Technology Letters (HTL)*, vol. 2, pp. 79–88, 08 2015.

A. Shumway-cook, S. G. Brauer, and M. Woollacott, "Predicting the probability for falls in community-dwelling older adults using the timed up & go test." *Physical therapy*, vol. 80 9, pp. 896–903, 2000.

J. Silva, I. Sousa, and J. Cardoso, "Transfer learning approach for fall detection with the FARSEEING real-world dataset and simulated falls," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 40, pp. 3509–3512, July 2018.

J. Silva, I. Sousa, and J. S. Cardoso, "Fusion of clinical, self-reported, and multisensor data for predicting falls," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 50–56, Jan 2020.

J. Silva and I. Sousa, "Instrumented timed up and go: Fall risk assessment based on inertial wearable sensors," in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2016.

J. Silva, J. Madureira, C. Tonelo, D. Baltazar, C. Silva, A. Martins, C. Alcobia, and I. Sousa, "Comparing machine learning approaches for fall risk assessment," in *BIOSIGNALS*, 2017.

A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "Real-life/real-time elderly fall detection with a triaxial accelerometer," *Sensors (Basel, Switzerland)*, 2018.

M. Szumilas, "Explaining odds ratios." *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, vol. 19 3, pp. 227–9, 2010.

J. C. Thomas, C. Odonkor, L. Griffith, N. Holt, S. Percac-Lima, S. Leveille, P. Ni, N. K. Latham, A. M. Jette, and J. F. Bean, "Reconceptualizing balance: Attributes associated with balance performance," *Experimental Gerontology*, vol. 57, pp. 218–23, September 2014.

M. Tinetti, D. Richman, and L. W. Powell, "Falls efficacy as a measure of fear of falling." *Journal of Gerontology*, vol. 45 6, pp. 239–43, 1990.

M. E. Tinetti, "Performance-oriented assessment of mobility problems in elderly patients," *Journal of the American Geriatrics Society*, vol. 34, no. 2, pp. 119–126, 1986.

P. Tsinganos and A. Skodras, "On the comparison of wearable sensor data fusion to a single sensor machine learning technique in fall detection," *Sensors*, vol. 18, p. 592, 02 2018.

K. S. van Schooten, M. Pijnappels, S. M. Rispens, P. J. M. Elders, P. J. A. Lips, A. Daffertshofer, P. J. Beek, and J. H. van Dieën, "Daily-life gait quality as predictor of falls in older people: A 1-year prospective cohort study," *PLoS ONE*, 2016.

K. S. van Schooten, M. Pijnappels, S. M. Rispens, P. J. M. Elders, P. T. A. M. Lips, and J. H. van Dieën, "Ambulatory fall-risk assessment: amount and quality of daily-life gait predict falls in older adults." *The journals of gerontology. Series A, Biological sciences and medical sciences*, vol. 70 5, pp. 608–15, 2015.

M. Vassallo, L. Poynter, J. C. Sharma, J. Kwan, and S. C. Allen, "Fall risk-assessment tools compared with clinical judgment: an evaluation in a rehabilitation ward," *Age and Ageing*, vol. 37, no. 3, pp. 277–281, 01 2008.

B. Vigna, F. Pasolini, R. de Nuccio, M. Capovilla, L. Prandi, and F. Biganzoli, "Low cost silicon coriolis' gyroscope paves the way to consumer IMU," in *Advanced Materials and Technologies for Micro/Nano-Devices, Sensors and Actuators*. Springer Netherlands, 2010, pp. 67–74.

C. Wang, W. Lu, M. R. Narayanan, D. C. W. Chang, S. R. Lord, S. J. Redmond, and N. H. Lovell, "Low-power fall detector using triaxial accelerometry and barometric pressure sensing," *IEEE Trans. Ind. Inform*, pp. 2302–2311, 2016.

F.-T. Wang, H. lung Chan, M.-H. Hsu, C.-K. Lin, P.-K. Chao, and Y.-J. Chang, "Threshold-based fall detection using a hybrid of tri-axial accelerometer and gyroscope." *Physiological measurement*, vol. 39 10, p. 105002, 2018.

J. Whitney, S. R. Lord, and J. C. T. Close, "Streamlining assessment and intervention in a falls clinic using the timed up and go test and physiological profile assessments." *Age and ageing*, vol. 34 6, pp. 567–71, 2005.

WHO, "WHO global report on falls prevention in older age," WHO, Tech. Rep, 2007.

A. G. Wilde, "An overview of human activity detection technologies for pervasive systems," vol. 212. Department of Informatics University of Fribourg, Switzerland, 2010, p. 72–112.

R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining," 2000.

I. Wisesa and G. Mahardika, "Fall detection algorithm based on accelerometer and gyroscope sensor data using recurrent neural networks," *IOP Conference Series: Earth and Environmental Science*, vol. 258, p. 012035, 05 2019.

I. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques (second edition): Morgan kaufmann," 01 2005.

W. Zhang, G. R. H. Regterschot, H. Schaabova, H. Baldus, and W. Zijlstra, "Test-retest reliability of a pendant-worn sensor device in measuring chair rise performance in older persons," *Sensors (Basel, Switzerland)*, vol. 14, no. 5, pp. 8705–17, 2014.

Øivind Due Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition - A survey," *Pattern Recognition*, vol. 29, no. 4, pp. 641 – 662, 1996.