Detection of Freezing of Gait in Patients with Parkinson's Disease Using Deep Recurrent Neural Networks

Spyroula Masiala

Master Thesis DSBG

THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE OF DATA SCIENCE
AT THE FACULTY OF HUMANITIES
OF TILBURG UNIVERSITY

Thesis committee:

Dr. Martin Atzmüller

Dr. Willem Huijbers

Tilburg University
School of Humanities
Tilburg, the Netherlands

Preface

This thesis has been written as the partial fulfillment of the requirements for the Master track Data Science: Business and Governance and completes my education at Tilburg University. During my time at Tilburg University, I had the opportunity to meet many interesting people, who inspired me and contributed to the accomplishment of this thesis. It would be impossible to complete this work without the support and the motivation of those people.

First of all, I would like to thank my supervisor Dr. Martin Atzmüller for his excellent guidance, the encouragement during this process and all the fruitful discussions we had. I am sincerely grateful to him for supervising my thesis with great care and for having pushed me to go always a step farther.

Moreover, I would like to thank Dr. Willem Huijbers for evaluating my thesis in its final stage. Thank you for the time spent in assessing my work and making it possible for me to graduate.

Without the friendly atmosphere, created by my classmates and the inspirational meetings with the data science consultants, my time at Tilburg University would not be so enjoyable. I would like to especially thank them for all the fun we had together during the coffee breaks at the University's library and the interesting discussions we had in Montesque building, the last year.

Finally, this thesis would not be possible without the gens and moral support of my family, and the continued motivation and love of my amazing and beloved sisters, Dora and Penelope.

Spyroula Masiala

Tilburg, December 2017

Summary

"I have a form of Parkinson's disease, which I don't like. My legs don't move when my brain tells them to. It's very frustrating!" -George H.W.Bush.

Freezing of gait (FoG) is a common gait disorder among patients with Parkinson's disease. The disorder appears in the advanced stages of the disease and negatively affects the patients' quality of life. FoG is resistant to the existing medication, hence there is an urgent need for non-pharmacologic treatment. In the past two decades, a number of researchers have sought to address this problem and determine a non-pharmacological, yet effective treatment of Freezing of gait. Recent evidence suggests that the gait performance and gait stability of patients with Parkinson's disease can be improved with the help of continuous external rhythmic auditory cues. The following studies have focused on the design of a Wearable Assistant Hardware that relies on 3D-accelerometer sensors and assists walking of Parkinson´s Disease patients. It was deployed to detect FoG events from sensors information and help patients suppress them by applying acoustic cueing. Therefore, the main goal of this thesis is to determine to what extent FoG episodes can be automatically detected in Parkinson's disease using 3D-accelerometer sensors. This goal leads to the following problem statement: *To what extent can FoG episodes be detected from sensor information?*

The existing studies developed systems for FoG detection based on different machine learning algorithms, however, most of them were designed in subject-dependent settings. In the subject-dependent method, the system is optimized for a specific user, i.e. training and testing on the same subject. The performance is higher compared to the user-independent method, however, the system does not generalize as well to other subjects. In this thesis, we aim to answer the problem statement, by developing a ''deep'' FoG detector, for both subject-dependent and subject independent settings, and achieve an objective FoG detection as well.

After extracting highly informative features based on the existing scientific literature, we apply a Recurrent Neural Network with Long Short-Term Memory cells, to detect FoG episodes. We perform thirteen experiments in total and in each one of them, the input for our deep model is a different feature group, determined by the sensor placement, i.e. ankle, thigh and trunk, and the axis of the signal, i.e. x, y, and z-axes. In the subject-independent method, the experiment where the selected feature group is the frequency domain features extracted from the trunk sensor is identified as the best performing

experiment. Our model detects successfully FoG episodes with an average AUC score of 93%, Specificity of 90% and Sensitivity of 81%.

Moreover, in the subject dependent method, the experiment where the selected feature group is the frequency and statistical features of all the sensors is identified as the best performing experiment. Our system detects FoG events with an average AUC score of 97%, Specificity of 96% and Sensitivity of 87%.

The results of the current study are promising, especially for the development of a system more robust to intraclass variability and the walking characteristics of each user. This study could be used in future research to build upon as a basis for a real-time FoG detector, optimized for working with hundreds of different patients. A generalized model, independent of the subject could be an effective and low-cost approach to the Freezing of Gait problem and could improve the life of the patients vitally.

Keywords: Detection, Freezing of Gait, subject-independent, Recurrent Neural Network, LSTM.

Contents

List of Figures

List of Tables

Nomenclature

AD: Alzheimer's Disease

ANN: Artificial Neural Networks

DTF: directed transfer function

FFT: Fast Fourier Transform

FI: Freeze Index

$FI_{MC:}$ multi-channel Freeze Index

FoG: Freezing of Gait

H&Y: Hoehn and Yahr

HCS: Healthy control subjects

LD: Levodopa

LSTM: Long Short-Term Memory

MLP-NN: Multilayer Perceptron Neural Network

$P_{L:}$ Power in the locomotor band

$P_{H:}$ Power in the "freeze" band

PCC: Pearson's correlation coefficient

PD: Parkinson's Disease

PDC: partial directed coherence

PWF: Patient with PD with FOG

PWoF: Patient with PD without FOG

PWP: Patient with PD

RAS: Rhythmic Auditory Stimulation

RNN: Recurrent Neural Network

SEN: Sensitivity

SMOTE: Synthetic Minority Oversampling Technique

SP: Specificity

Chapter 1: Introduction

The first part of this chapter contains a general introduction into the Freezing of Gait in Parkinson's disease. The second part describes the main research problem. Section 1.3 formulates the problem statement and provides the research questions. In Section 1.4 we discuss the contribution of this research.

1.1 Freezing of Gait in Parkinson's disease

Next to Alzheimer's disease (AD), Parkinson's disease (PD) is known as the second most common, age-related neurodegenerative disorder. It is caused by the progressive loss of dopaminergic and other sub-cortical neurons (Steven T. Moore, MacDougall, & Ondo, 2008). Dopamine is a chemical released by dopaminergic neurons and it plays an important role in the efficiency of human motion. Considering that, the loss of the dopaminergic neurons manifests motor symptoms such as tremor at rest, akinesia or bradykinesia, rigidity, postural instability, impaired balance, forward-flexed posture and freezing (Jankovic, 2008; Mhyre, Boyd, Hamill, & Maguire-Zeiss, 2012). In addition to the main motor symptoms, a non-motor symptomology (personality change, anxiety, dementia, depression, sleep disorder, and hallucinations) may be present. After several years of the Parkinson's disease, the above-mentioned symptoms may become not only troublesome but even deadly (Factor & Weiner, 2007).

The increasing recognition of PD as a serious, worldwide health threat impelled the scientific research in the disease. Firstly, in the past decades, researchers have sought to determine the cause of PD, however, it remains unknown. Nevertheless, several studies claim that the cause of the disease involves environmental and genetic factors and age is recognized as the most important risk factor, while more than 10 million people over age 50 globally suffer from Parkinson's disease. Secondly, treatment options for PD have been explored in several studies. These studies reported medications and surgery as treatment options to improve the symptoms of PD, however, a complete cure remains absent (Bloem, Hausdorff, Visser, & Giladi, 2004). The continuous rise of elderly people in the society indicates the increase of patient with PD (PWP) in the following years. Taking the absence of a cure into consideration, we predict that, by 2040, neurological diseases such PD will be the second most common cause of death worldwide (WHO Program on Neurological Diseases and Neuroscience, World Health Organization Department of Mental Health and Substance Abuse, & World Federation of Neurology, 2004).

Balance and gait disorders are the most considerable therapeutic challenges in PD, considering that they respond poorly to the available treatment options, except for the early stage of the disease. Freezing of gait (FoG) is a major gait disorder in the later stage of the PD (Nutt et al., 2011). Approximately 50% of

patients with Parkinson's disease (PWP) suffer from the freezing of gait episodes (FoG), i.e. a sudden and brief inability to walk (Fahn, 1995). As reported by patients, during a FoG episode, they are halted with the feeling "their feet are glued to the earth and they are temporarily unable to walk", despite their intention to start walking or moving forward (Schaafsma et al., 2003). Per a research of 6620 patients with Parkinson's disease (PWP), nearly 47% of them reported regular freezing episodes and 28% of the subjects experience FoG episodes daily (Macht et al., 2007). FoG affects social and clinical outcomes in the PWP. Falls, interferences on daily activities and substantial harm to the quality of life are the most common of the above-mentioned outcomes (Bloem et al., 2004; de Boer, Wijker, Speelman, & de Haes, 1996). Finally, FoG appears resilient to the most common pharmacological treatment in Parkinson's disease (Levodopa) (Bloem et al., 2004).

1.2 Problem Formulation

The absence of a complete cure in Parkinson's disease and the inefficacy of pharmacological treatment in FoG incited clinicians, and patients to develop numerous behavioral "tricks", to overcome freezing episodes. To mention a few: stepping over cracks in the floor, marching to command, shifting body weight, and walking to a beat (Rahman, Griffin, Quinn, & Jahanshahi, 2008). Rhythmic Auditory Stimulation (RAS) forms one of the most efficient instrument in gait improvement among patients suffering from FoG (PWF) (Hashimoto, 2006). RAS can provide a regular metronome ticking sound upon the detection of a FoG episode. The sound alerts the patient about the upcoming FoG event and they enhance their speed and improve their gait stability (Hausdorff et al., 2007). Recent clinical studies confirmed that the rhythmical ticking sound synchronizes with the gait and helps the PWP suppress the freezing episode and continue walking (Donovan et al., 2011; Suteerawattananon, Morris, Etnyre, Jankovic, & Protas, 2004).

Nevertheless, despite its safety and efficacy, RAS suffers from several major drawbacks: to work effectively, FoG episodes need to be properly detected. While it is proven that with the assistance of RAS, the duration of FoG episodes is shorter and the patients return to the normal walking patterns (Bachlin et al., 2010), its effectiveness weakens through time. Consequently, long lasting cueing is not recommended (Cubo, Leurgans, & Goetz, 2004; Nieuwboer, 2008; Rubinstein, Giladi, & Hausdorff, 2002) and context-aware cueing systems, that are capable to provide the rhythmical ticking on the onset of a FoG event, are required. The main challenge in the context-aware cueing systems is the reliable detection of FoG episodes (Bachlin et al., 2010), hence, extensive research has been carried out on the off-line and on-line detection of FoG episodes (Bachlin et al., 2010; Cole, Roy, & Nawab, 2011; Delval et al., 2010; Djurić-Jovicić et al., 2014; J. H. Han, Lee, Ahn, Jeon, & Park, 2003; Handojoseno et al., 2012; S. Mazilu et al.,

2012; Sinziana Mazilu et al., 2013; S. T. Moore et al., 2017; Steven T. Moore et al., 2008; Niazmand et al., 2011; Popovic, Djuric-Jovicic, Radovanovic, Petrovic, & Kostic, 2010).

Considering that PD is related to gait disorders and tremor, the movement patterns of such shivers and FoG episodes can be captured by wearable sensors. Several attempts have been made to detect freezing episodes from sensors information. Bächlin et al. (2010) introduced a wearable Assistant Research Hardware, based on 3D-accelerometer sensors that record the movement time-series of PWP. Their wearable assistant designed to detect FoG events and provide rhythmic auditory cueing on their onset (Bachlin et al., 2010). The online FoG detector achieved Sensitivity of 73.1% and Specificity of 81.6% respectively, in subject-independent settings. The user-specific performance varies from one patient to another. Moore et al. analyzed offline accelerometer data collected from 11 PWP and they created a FoG detector with accuracy up to 89% in subject dependent settings (Steven T. Moore et al., 2008).

In numerous studies, machine learning classifiers have worked well for patient-dependent or group dependent settings (Cole et al., 2011; S. Mazilu et al., 2012), however, few researchers have been able to draw on any systematic research into the detection of a FoG episode in subject-independent settings. Moreover, the few of them reported poor user independent performance in contrast to the user-specific (Bachlin et al., 2010; Cole et al., 2011; S. Mazilu et al., 2012). Bachlin et al. (2010) reported that the walking style differentiates among PWP, while we can separate them into two groups; the smooth walkers and the intensified stepping walkers. Furthermore, training machine learning algorithms with general data from a large number of users result in poor performances when we test them in specific users (S. Mazilu et al., 2012). In addition, the same classifier works well for some subjects (patient), while for some other performs significantly worse.

In the light of the above general lack of research on FoG detection in subject-independent settings and the poor user independent performance that has been reported in the few existing studies, this thesis seeks to introduce a novel system that detects successfully FoG episodes from signal information, not only in subject-dependent settings but in subject-independent as well. A FoG detector can be used to enable preemptive RAS on time and repel the episode. Moreover, a FoG detector could improve the quality of life for PWP and reduce significantly their medical care costs.

1.3 Problem statement and Research Questions

As mention above, existing FoG algorithms, have been designed in user-dependent settings or reported poor performance in subject-independent method. Bachlin et al. (2010) argued that the poor user-independent performance is due to the differences in the walking style of patients.

Given that we are provided with a dataset with signals acquired from body sensors placed on PWP, the major objective of this study is to build a detection model that utilize the sensor information and detect

accurately a FoG episode not only in subject-dependent settings but in subject-independent as well. We should highlight the importance of detection a FOG event on its onsets so that a RAS system could support the patients to return in their normal gait and successfully suppress the FoG episode.
Consequently, the problem statement of the thesis is formulated as follows:

Problem statement: *To what extent can FoG episodes be detected from sensor information?*

In the field of activity recognition, there are numerous successful applications of machine learning techniques using acceleration data (Precce et al., 2009; Mannini et al., 2010). The machine learning algorithms use features extracted from motion data and they successfully detect activities such as walking, running, opening a door, or turning around. Freezing of Gait can be observed as a specific activity in the field of activity recognition, even though it is not a deliberate motion (Mazilu et al., 2012; Zach et al., 2009).

In addition, numerous studies in Human Activity Recognition (HAR) explored deep learning models and showed that deep learning algorithms outperform shallow ones (Alsheikh et al., 2016). However, there is still a lack of methodological research on deep learning capabilities, especially on the detection of a FoG event. Hence, our first research question aims to investigate whether a recurrent deep learning algorithm, specifically a recurrent neural network with Long Short-Term Memory (LSTM) cells, could successfully find hidden trends in the sensor data. Therefore, our first and second research questions state as follows:

Research Question 1 (RQ1): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-independent method?*

Research Question 2 (RQ2): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-dependent method?*

The key to solve successful a Human Activity classification problem, as for any pattern recognition task, is the proper selection of the features extracted from body sensor data and the design of the appropriate classifier (Plotz et al., 2011). FoG episodes hold a unique frequency range and project a typical motion pattern, which can be visually distinguished from normal gait (Moore et al., 2008). The motion patterns can be analyzed, and the unique pattern of freezing episodes can be detected, by using features extracted

from body sensor data. Hence, the utilization of relevant features is crucial, as the different motion patterns are represented diversely in the feature space and render the distinction between the different motion possible (Mazilu et al., 2012). Apart from Plotz et al. (2011), there is a general lack of systematic research in feature learning, which leads to a major weakness in the current HAR system (Lukowicz et al., 2010). An additional goal of this thesis is to utilize different feature groups, namely frequency domain features and statistical and time-domain features and investigate the most relevant features in the FoG detection task. Therefore, the third research question of the thesis addresses this matter:

Research Question 3 (RQ3): *Which features contribute most to detect FoG in Parkinson's disease patient?*

1.4 Outline of the thesis

The outline of the thesis is structured as follows. Chapter 2 contains a literature review on FoG detection. Chapter 3 describes the experimental setup, where the dataset and experimental procedure will be described in detail. In Chapter 4, we present the results of our experiments. Finally, in chapter 5, we provide a general discussion of the problem statement and research questions, present the conclusion of our research, and suggest recommendations and directions for further research.

Chapter 2: Related work

This chapter describes relevant studies in the field. Reviewing previous scientific approaches formulates our approach and explains how this thesis can be placed in the existing literature. Moreover, it guides us to identify important features in advance and select the proper algorithm. This chapter is divided into five sections. First, we discuss recent studies in the field of Freezing of Gait detection (Section 2.1). Second, we discuss deep learning algorithms which have been used in the field (Section 2.2). After that, we explore feature learning techniques which have been frequently used for the FoG detection in Parkinson's disease (Section 2.3). Finally, we conclude the chapter with a precise explanation of how this thesis will contribute to the existing studies (Section 2.4).

2.1 Current approaches in FoG detection

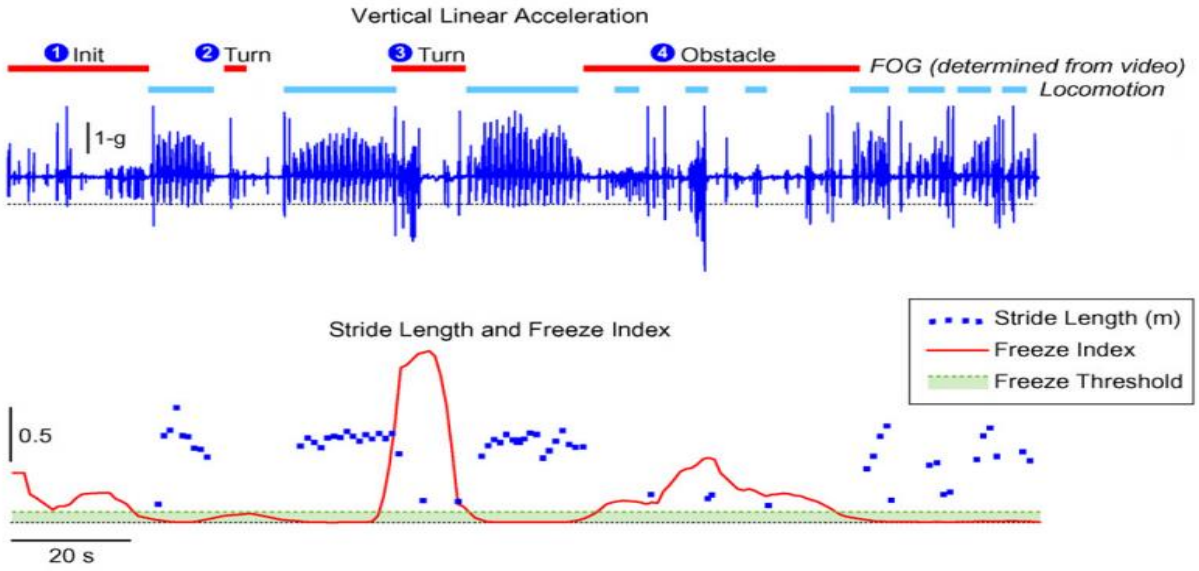
As already pointed out in the introduction, approximately 50% of advanced PD patients suffer from Freezing of Gait episodes (Fahn, 1995). During a "freeze" episode, a PWP exhibits gait characteristics that are noticeably different than the gait characteristics manifested during normal walking. The first and most common characteristic is the tremble of the legs in their effort to overcome a FoG event. The disability to move forward, or *akinesia*, is the second and best- known characteristic. The third characteristic is moving forward with small steps, but without fully lifting the legs from the ground (Bloem et al., 2004). Studies on the three characteristics determined the quantifiable changes in the gait between healthy persons and PWP. To mention a few: reduced walking speed, footstep length and increased rhythm at the beginning of the motion (Nieuwboer, 2008). Those findings form the fundamental aspect for many FoG detection approaches.

In 2003, Han et al. published a research in the FoG detection, based on the need for a "un-freeze" treatment and the clinical observation of the gait disorder. In their research, Han et al. (2003) measured the leg "swaying" using data collected from 2 biaxial accelerometer sensors placed on 2 PWF and 5 healthy control subjects. Their research outlines that the frequency acceleration of the "freezing" gait (6-8 Hz) per PWF is significantly higher in comparison to the normal gait (2 Hz). Next, they formulated a three class classification problem -normal, FoG and resting state- and solved it by evaluating the differences between the wavelet power levels on different frequency bands. Following their preliminary study, in 2006, the group developed a wearable activity monitoring system (J. Han, Sun Jeon, Suk Jeon, & Park, 2006), consisting of a 3D accelerometer with a foot pressure system and a camcorder. Finally, their gait detection algorithm was capable of detecting any irregularity of the gait, such as "freezing" of

gait, bradykinesia, and toe-walking. According to the evaluation of the gait parameters, the performance of their algorithm is 94% accuracy for normal gait and 93% accuracy for PD gait.

Numerous studies on FoG detection used data collected from body sensors, while the 3D-accelerometer sensor appears to be the most popular. For instance, Moore et al. (2008) conducted a study in the ambulatory monitoring of FoG. They used data collected from a 3D-accelerometer sensor placed on the left shank of 11 PWP and they introduced the freezing index (FI) as the power of the body acceleration signal in the freeze band (3 to 8 Hz) divided by the power in the "locomotor" band (0.5 to 3 Hz). The FI has been described extensively (see Section 2.3). Moreover, the group showed that the width of the optimal window was two times the duration of the shortest detected "freeze" episode. Finally, the group showed that the accuracy and sensitivity of the FoG detection increases with an individual FI threshold. Their study is recognized as a milestone in the field since FI remains the most common extracted feature for FoG detection (Mazilu et al. 2012).

Figure 2.1: The Freeze Index (FI) was calculated from the power in the "freeze" band (3 to 8 Hz) divided by the power in the locomotor band (0.5 to 3Hz). During the FoG event the FI (red line) occurs large peak values. (Moore et al. 2008)



The idea expressed by Moore et al. (2008) on ambulatory monitoring of FoG, led Popovic et al. (2010) to design a FoG detector, using a force sensitive resistor and Pearson's correlation coefficient (PCC). Accelerometer data from 9 PWF were collected for the purpose of this study and they were recorded and processed by Bluetooth technology. Their detection model managed to classify accurately all the FoG episodes in patient-dependent settings, while only 24 episodes were successfully detected by observing the video. Moreover, they determined the "small steps" during the FoG episodes as the ones with PCC values of less than ±1. As a final remark of their work, Popovic et al. (2010) claimed that this method is equally sensitive to different freezing patterns of gait, however, the recognition of non -freezing pattern of gait which might still be different from normal gait remains unclear.

In an interesting analysis of the nearly undetectable FoG, Delval et al. (2010) moved out of the box and they utilized data collected by a goniometer sensor placed on the subjects's knee. Gait parameters were evaluated in 10 PWF, 10 PWoF and 10 healthy control subjects. During the experiments, the subjects were walking on a treadmill, trying to avoid randomly dropped obstacles. In the time-frequency analysis of the data, Delval et al. (2010) extracted the FI (Moore et al., 2008) from sliding windows with size of 4.1 seconds and they reported a significant decrease in the step duration, FI and footstep length of each subject during a FoG episode. They proposed the new features (footstep length and step duration) as important parameters related to FoG and they reported Specificity of 95% and Sensitivity of 75-88%.

While the interest in the field increases, more researchers attempt to explore the FoG detection problem. For instance, Niazmand et al. (2011) used data collected from five 3D accelerometer sensors placed on the shank and belt of 6 PWF and they applied an algorithm with frequency dominant thresholds to detect "freeze" events. Their system, or *MiMed-Pants*, achieved Specificity of 85.3% and Sensitivity of 88.3% in the user-specific method. Moreover, Niazmand et al. (2011) acknowledged that their model could not detect the akinesia during a FoG episode or classify the lack of movement. Another study conducted from Mancini et al. (2012), highlighted the significant increase in "freeze" band power (3-8 Hz) during a FoG episode and pointed out the FI as the most useful indicator of the gait disorder during the event. A more recent study by Coste et al. (2014) introduced a new method for the observation of gait anomalies and the FoG detection. They argued that the FoG criterion (FOGC) -the uninterrupted evaluation of frequency and stride length-, provides a better indicator of freezing compared to the FI.

In the same vein, Bachlin et al. (2010) conducted a study on the online FoG detection at the Movement Disorders Unit, Department of Neurology at the Tel Aviv Sourasky Medical Center (TAMSC). They developed a real-time wearable device for automatic FOG detection. The data were collected from 3D-accelerometer sensors, placed on the ankle, thigh, and trunk of 10 PWF. A wireless Bluetooth transmitted all the data to a wearable computing system and a rhythmic auditory stimulation (RAS) system. Inspired

by Moore et al. (2008), they extracted the Freeze Index (FI) and Power Index (PI) from each signal window. The chosen detection tolerance was a maximal delay of 2 seconds. In their study, Bachlin et al. (2010) reported an average Sensitivity of 73.1% and Specificity of 81.6% in the patient-independent method. However, due to the different gait patterns of each subject, the group noticed a large variation in the user-specific performance. To mention a few: patient 1 achieved Sensitivity of 99.1% (with Specificity of 39.7%) and the patient 8 achieved Specificity of 88.9% (with Sensitivity of 34.1%). Finally, Bachlin et al. (2010) pointed out that the detection performance increased up to 88.6% Sensitivity and 92.4% Specificity, by optimizing the FI and PI thresholds for each patient.

The significance of cueing in the FOG treatment, inspired Johanov et al. (2009) to propose a wearable automatic system designed to detect FoG events in real-time and deliver acoustic cues to the patient only upon a "freeze" event detection. Their FoG detector, or *deFog system*, was evaluated on a real PWP for a limited test run and managed to detect FoG episodes with a maximum latency of 580 milliseconds and average latency of 332 milliseconds.

Similarly, Djurić-Jovicić et al., (2014) designed a custom - build wireless sensor system, consisting of 6 inertial measurement units (IMUs) on each leg segment and a non-wearable computer device. They collected data from 4 PWP, at the stage from 1 to 4 of the disease. All the patients had a clinical history of FoG episodes. They developed an algorithm for gait classification and they used signal data collected from the gyroscope sensors placed on the foot and the thigh and from the accelerometer sensors placed on the foot segment. Their model reported an average accuracy of 88% per patient.

Year by year, the development of a wearable on-line FoG detector gains more ground and scientists became fascinated by achieving the challenge. In 2012, Mazilu et al. developed a wearable FoG detector based on a smartphone, and they utilize data collected from 3D-accelerometer sensors placed on PWF. They evaluated numerous supervised algorithms, namely Random Forest, C4.5 Decision Trees, Naive Bayes, multilayer perceptron, as well as boasting and bagging methods. Their research showed that machine learning techniques can adapt successfully the high in dimensionality features of the FoG episodes, instead of the commonly-used manual thresholds. The proposed FoG detector was tested offline on data collected from three 3D-accelerometer sensors placed on the ankle, the thigh, and the trunk of 10 PWF. The detection latency and the general performance of the system were optimized by the sensor location, the window size, and a combination of machine learning algorithms. Interestingly, their research demonstrates that the combination of machine learning algorithms outperforms the single classifier. According to their study, in the user-dependent method, Decision Tree C4.5 and Random Forest classifier provide the shortest latency, 0.24 s, and 0.35 s respectively, and the AdaBoost reported the highest Sensitivity and Specificity, of 98.35%, and 99.72% respectively. Moreover, the Random Forest classifier

reported Sensitivity of 62.05% and Specificity of 95.15% for the 1s window size. As final remarks of their study, the group pointed out, firstly that the large deviation between the walking styles of each patient was the main reason for the low user-independent performance, and secondly that a single sensor could collect sufficient data for the FoG detection system. In their most recent work, Mazilu et al. (2014) developed a second daily-life assistant for the PWP, or *the GaitAssistant.* Their second system provides FoG detection and daily-life support to the PWF by using a C4.5 Decision Trees as classifier and FI as its feature. The *GaitAssistant* showed the positive short-term impact on the participants' movement.

In the same vein, Tripoliti et al. (2013) developed a FOG detection system and they tested 4 machine learning algorithms; Random Forest, Random Tree, Decision Trees, and Naive Bayes. They used signal data collected from 6 accelerometers and 2 gyroscopes placed on 5 PWF, 6 PWoF, and 5 healthy control subjects. Tripoli et al. (2013) used a sliding window of 1 s and overlap of 0.5 s and they extract entropy from each window. The Random Forest classifier was reported as the one with the best user-specific performance, with Sensitivity of 81.94%, Specificity of 98.74% and Accuracy of 96.11%.

In another major study, Handojoseno et al. (2012) presented a novel FoG detection model and instead of accelerometer data, they utilized Electroencephalogram (EEG) signals. In their study, they used wavelet decomposition to analyze the dynamics of EEG signals during the onset of the FoG event and the freezing periods. The aim of their study was to achieve an early detection of a FoG episode and help the patients to avoid the upcoming "freeze" episode. Their system achieved accuracy of 80.2%, Sensitivity of 86% and Specificity of 74.4% in the user-dependent settings. Their results demonstrate the promising use of EEG signals for the future research on FoG treatment.

In their most recent research, Moore et al. (2017) introduced the novel multi-channel Freezing Index (FI_{MC}) - the ratio of the powers (freeze and locomotor bands), that are summations of single powers over the N axis. In their study they selected the 7 most relevant features, out of 244, by using voting process with clusterability and mutual information criterion. Next, they designed an anomaly score FoG detector with adapting thresholding and they achieved a Sensitivity of 96% and Specificity of 79% in subject-independent settings.

In view of all the above mentioned studies, we may suppose that several research groups have proposed promising wearable FoG detection systems, while some of them provide feedback to the users as well. Table 2.1 contains an overview of the characteristics and evaluation results of the above-mentioned systems.

Table 2.1: Current strategies on FoG detection

Authors	*Dataset name (if applicable)*	*Dataset characteristics*	*Sensors*	*Results*
Bachlin et al., 2010	Daphnet FoG Dataset	10 PWF	3D-accelerometer (on ankle, thigh and lower back) and one wearable device	SEN=78.1%, SP=86.9%
Moore et al., 2008 (1) & 2017 (2)	Not applicable (1); Daphnet FoG Dataset (2)	11 PWF (1) 10 PWF (2)	Accelerometer (on the left shank) and wearable device (1); 3D-accelerometer (on ankle, thigh and lower back) and one wearable device (2)	Significantly higher 'freeze' band power (3-8 Hz) during FoG episode (1); SEN=96%, SP=79% (2)
Han et. al., 2003 (1) & 2006 (2)	Not applicable (1) & (2)	5 HCS and 2 PWP (1); 5 HCS and PWP (2)	biaxial accelerometer sensors (1); a camcorder and a foot pressure system (2)	Significantly higher frequency component during FoG episode (1); Accuracy = 93% for PD gait and Accuracy = 94 for normal gait (2)
Johanov et al., 2009	Not applicable	1 PWF and simulation data	3D-accelerometer and a gyroscope	Detection Latency = 332ms (on average)
Popovic et al., 2010	Not applicable	9 PWF	Force Sensitive Resistors and video	Significantly lower correlation between normal locomotion and the "freezing" gait
Delval et al., 2010	Not applicable	10 PWF, 10 PWoF, 10 HCS	3D motion analysis system, video, and goniometers	SEN=75-83%, SP= 95%
Djurić-Jovičić et al. 2014	Not applicable	4 PWF	6 Inertial Measurement Units (on each leg segment) and a non-wearable computer device	Accuracy = 88%
Cole et al., 2011	Not applicable	10 PWF and 2 HCS	3D accelerometer (on ankle, thigh and lower back) and an Electromyography (EMG) device	SEN=82.9%, SP=97.3%
Niazmand et al., 2011	Not applicable	6 PWF	5 accelerometers (shanks and belt) with a non-wearable computer device	SEN=88.3%, SP=85.3%
Handojoseno et al, 2012	Not applicable	26 PWF	Wireless Electroencephalography (EEG) system	Accuracy= 80.2% SEN=86% and SP= 74.4%
Mancini et al., 2012	Not applicable	21 PWF, 27 PWoF, 21 HCS	3D accelerometer and 3D gyroscopes	Significantly higher 'freeze' band power (3-8 Hz) during FoG episode
Zhao et al., 2016	Not applicable	10 PWF	5 accelerometers (shank and belt) with a non-wearable computer device	SEN=81,7%
Mazilu et al., 2012 (1), 2014 (2)	Daphnet FoG Dataset (1); not applicable (2)	10 PWF (1); 5 PWF (2)	3D-accelerometer (on ankle, thigh and trunk) and smartphone (1); motion sensors (ankle) and a wearable device (smartphone) (2)	SEN=99.69%, SP=99.96% Latency=340ms (1); Accuracy =97% and Latency=0.25s (2)

** Not applicable refers to anonymous/ nonpublic available datasets. The data was collected from patients into lab sessions and used for the current study

2.2 Deep neural networks in FoG recognition

The first and second research question of the thesis is the following: *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-independent method?* and *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-dependent method?* In order to answer these research questions, a review of deep learning algorithms in the FoG studies is necessary.

Freezing of Gait detection can be formulated into a two-class classification problem, namely no-FoG (walking) versus FoG. In previous studies, some researchers used a simple linear classifier, while some other adapted more sophisticated machine learning classification strategies. Moore et al. (2003 & 2006), Bachlin et al. (2010), Niazmand et al. (2011), Johanov et al. (2009), Delval et al. (2010) applied manually thresholds to detect FoG episodes. In their work, Mazilu et al. (2012 & 2014) implemented the Random Forest, C4.5 Decision Trees, Naive Bayes, multilayer perceptron, boasting and bagging methods to solve the FoG detection problem. Zhao et al. (2016) developed a time-frequency combined algorithm and Tripoliti et al. (2013) applied a Naive Bayes classifier. Together, the different strategies reached detection Sensitivity above 80%, and the latency of the detection was mostly of a few hundred milliseconds.

A broader perspective has been adopted by Handojoseno et al. (2012), who utilized first a deep learning algorithm for their FoG detection model. In their study, they extracted features from Electroencephalogram (EEG) signals with Discrete Wavelet Transform calculations and they utilized pattern recognition techniques. The Wavelet Energy and Total Wavelet Entropy were the selected features due to their significant advantages in detecting changes in EEG signal patterns. Afterwards, they applied a Multilayer Perceptron Neural Network and they proved the significant change in the EEG signal values prior a FoG episode. Their model reported average user-specific performance of Accuracy, Sensitivity, and Specificity around 75 %.

In order to recognize anomalies in the walking patterns of PWP, Djurić-Jovičić et al. (2010) recorded walking motions via a wireless sensor system, consisting of 6 measurement units placed on the segment of each leg. Afterwards, they developed an algorithm for automatic FoG detection and walking patterns classification. The main structure of the algorithm was the combination of a perceptron neural network and rule-based signal classification. Simultaneously, a video camera was recording the walking routine of the patients, so as medical experts would be able to identify FOG episodes from the recorded video and use their ground truth annotations for the evaluation of the FoG detection model. The results demonstrated that the user-specific performance was up to 84%.

One year after their first attempt, Handojoseno et al. (2013) used a different approach for their deep FoG detection model. In their first attempt (2013) they extracted spectral and spatial features from the EEG signals by employing statistical analysis and wavelet decomposition of its coefficients. Next, they used those electrophysiological signatures as inputs for the Multilayer Perceptron Neural Network and k-Nearest Neighbor classifier. Their method could detect the change from normal locomotion to "freezing" gait with an accuracy up to 73 % and Sensitivity up to 87 %. The preliminary results of the study asserted the functional collapse between areas in the brain during FOG events and suggested that EEG signals offer potentials as a therapeutic strategy in advanced PD. One year later, the group continued their study in FoG prediction and they used a directed transfer function (DTF) and partial directed coherence (PDC) to analyze the patterns of PWP brain signals during FoG episodes. The values of DTF and PDC were the selected features used as input for the MLP-NN. In the user-dependent method, their model achieved average values of Sensitivity of 82%, Specificity of 77%, and accuracy of 78%. The results of their studies point out a significant improvement compared to conventional methods for FOG detection, indicating a promising future for the PD patients.

Despite the remote findings in the literature, neural networks seem to be promising and powerful tools for FoG detection.

2.3 Feature Learning Techniques in FoG Detection

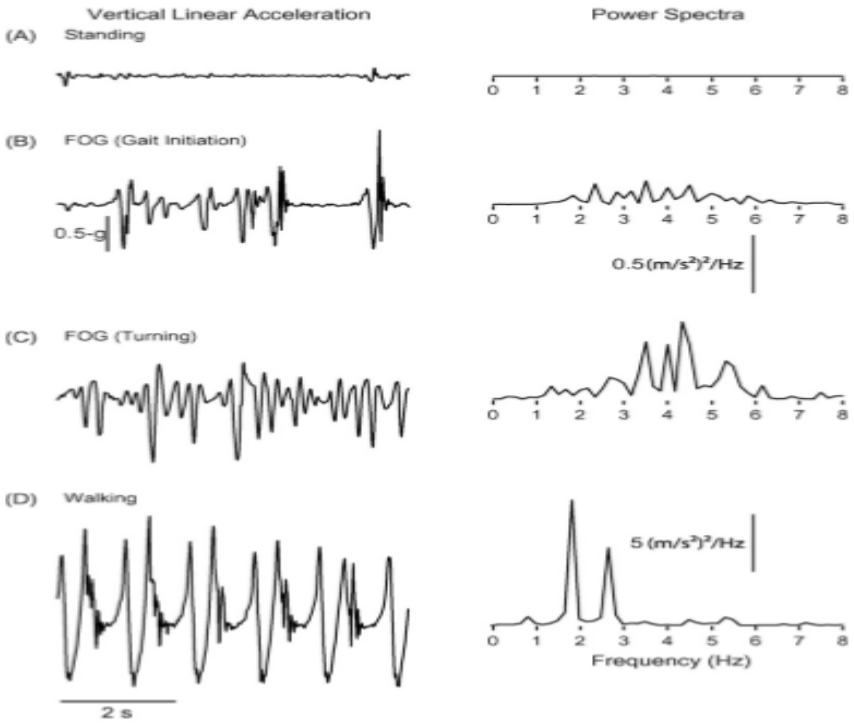
In order to provide an answer to the third research question: *Which features contribute most to detect FoG in Parkinson's disease patients?* we review existing feature learning techniques in the FoG detection literature. As described in section 2.1, researchers inclined to develop wearable systems which detect FoG events in real time. Most studies on FoG detection used signal data collected from wearable sensors, which were attached to the body of the subjects. The sensors could record subjects' physiological states such as speed, moving, change of location etc. Such sensors include accelerometers and/or barometers (Bachlin et al., 2009 & 2010; Moore et al., 2003 & 2006; Han et. al., 2003 & 2006; Johanov et al., 2009; Niazmand et al., 2011; Mancini et al., 2012; Zhao et al., 2012; Mazilu et al., 2012 & 2014), force sensitive resistors (Popovich et al., 2010), electroencephalography (EEG) (Handojoseno et al, 2012) and electromyography (EMG) (Cole et al., 2011). After the signal collection, feature extraction techniques must be applied to transform the signals into features that are discriminative for a specific activity. The features extracted from the signals model the feature space. In general, the recognition performance of a system is high when each activity can be clearly separated into the feature space. Hence, a major challenge in activity recognition task is the extraction of discriminative features, in a way that features equivalent to a specific activity should be grouped together into the feature space, whilst features

equivalent to a completely different activity should be clustered far apart in the feature space (Bulling, 2014). Signal-based features are the type of features that are mostly extracted in human activity recognition field. Signals can be transformed into time and frequency domain with mathematical operators, hence signal-based features can be divided into two categories; statistical and time domain features and frequency domain features.

Statistical and time domain features, such as mean, kurtosis, standard deviation, range etc., are popular for their simplicity, their high performance, and their ability to show the changes of the signal over time. Mazilu et al. (2012) and Moore et al. (2017) investigated the influence of the statistical and time-domain features on the FoG detection. Their research confirmed that the top-ranked features for FoG detection include statistical and time-domain such as mean, variance, range, standard deviation, minimum, and maximum.

Moreover, frequency-domain features are widely extracted for activity recognition task, due to their ability to demonstrate how the energy of a signal is distributed over a range of frequency components. In other words, frequency domain features are representations that model the body motion related to each activity. In the field of FoG detection, the current standard frequency-domain feature extracted from raw signals is the Freezing Index (FI). Freeze Index (FI) was introduced by Moore et al. (2008), in their study on ambulatory monitoring of Freezing of Gait. In their research, 11 PWP were asked to perform different walking tasks, while the movement of the left leg was measured by an ambulatory device designed by the authors. After the signal data collection, FFT was applied to transform the physical signals to frequency domain representation, or *spectrum of frequency components*. FoG episodes were identified based on the spectrum of frequency components. Moreover, all the walking tasks were recorded on a video camera and movement disorder specialist analyzed them and identified the FoG episode. Then, the ground truth annotations were used to evaluate the ambulatory monitoring system. Based on the frequency characteristics of the patients, they concluded that the leg movement during a FoG episode showed higher frequency component in the 3-8 Hz band and lower frequency component in the 0.5-3 Hz. Hence, the locomotor band was defined as 0.5-3 Hz and the “freeze” band as 3-8 Hz. Moreover, they computed the power spectra, which describes how the power of a signal is distributing into frequency components, over the locomotor and freeze band. They reported significantly higher power spectra in the “freeze” band than in the locomotor band (Figure 2.3). Next, a Freeze Index (FI) was calculated as the power in the “freeze” band (3-8 Hz) divided by the power in the locomotor band (0.5-3 Hz). Finally, a threshold was determined such that FI values above this limit were identified as FOG. The global ‘freeze’ threshold detected 78% of FoG events and the subject calibrated ‘freeze’ threshold detected 89%.

Figure 2.3: The vertical linear acceleration of the left leg and corresponding power spectra in the patient 6 (Moore et al., 2008)



Bachlin et al. (2010) developed an online FoG detector, based on the principle introduced by Moore et al. (2008), however, they included the power threshold as well. They collected accelerometer data from 10 PWP at 64 Hz sample frequency and then a sliding window was applied at window length of 4 s and overlapping of 0.5 s. On each window, a 256-point FFT was computed following by the power spectrum. Then the power in the "freeze" band (3-8 Hz), or P_H, and the power in the locomotor band (0.5-3 Hz), or P_L, were computed. The FI was calculated from the P_H divided by P_L, as suggested by Moore et al. (2008). Moreover, a new gait feature, the "energy", or *power index*, was also calculated from the summation of the P_H and P_L. Bachlin et al. (2010) determined a new power threshold, called $Power_{TH}$, such that values above this limit were identified as FoG. The user-specific threshold parameters, $Power_{TH}$ and $Freeze_{TH}$, optimization improved the detection performance up to 92.4% Specificity and 88.6% Sensitivity, while the user- independent threshold parameters optimization improved the detection performance up to Sensitivity of 73.1% and Specificity of 81.6%.

Johanov et al. (2009), and Delval et al. (2010) used the FI as the gait feature in their FoG detection study. Similarly, Tripoliti et al. (2013) calculated the entropy of raw signals collected from 6 accelerometers and

2 gyroscopes. The entropy of a signal measures the distribution of the frequency components. Mazilu et al. (2012), besides the statistical and time-domain features, they extracted numerous frequency-domain features. They calculated the FI as suggested by Moore et al. (2008) and the sum of the power in the 'freeze' (3-8 Hz) and locomotor (0.5-3 Hz) band as suggested by Bachlin et al. (2010).

Han et al., (2003 & 2006) compared the wavelet power on different frequency bands, to detect the FoG events. Wavelet analysis reveal the time-frequency distribution, meaning how the power of the signal changes over time. The group distinguished the freezing from normal locomotion by comparing the different wavelet power values, which we can observe at the signals during the two activities.

Mancini et al. (2012) followed the feature extraction technique suggested by Moore et al. (2008) and they used the FI, power in the "freeze" and the locomotor band as gait features. They verified an increase in power in the "freeze" band and the importance of FI as a gait feature, as reported by Moore et al. (2008),

In a different approach, Niazmand et al. (2011) used the standard frequency ration (FI), however, they added two extra features in their FoG detector, the number of pulses that the patient's heart beats during a FoG episode and the shaking foot duration. Moving beyond standard procedures, Handojoseno et al (2012) performed wavelet analysis, more specifically, they computed the wavelet energy and total wavelet entropy of EEG signals. The wavelet entropy calculates the degree of disorder of a signal, so it can provide valuable information about the changes within the EEG signals during the freezing events. The aim of their research was to develop an early FoG detection model using brain activity that may help patients to overcome an upcoming FoG episode. Finally, Djurić-Jovičić et al. (2010) used the energy of the sensor signals and the stride length from their signals as selected features. The energy of a signal is calculated by the summation of the squared Fast Fourier transformed signals. Popovic et al. (2010) analyzed the time series with Pearson correlation coefficient (PCC). First, they selected one step from the "normal" locomotion. The "normal" step was used to calculate the PCC with the entire dataset. Steps with PCC values of less than ±1 were considered as "irregular" and they hypothesized that the irregularity of the PCC values was the result of complex freezing events.

Finally, in a more recent study, Moore et al. (2017) introduced four new feature extraction approaches. The first two are: the number of peaks and the maximum in the spectral coherence, i.e. C_{XYNpks} and C_{XYmax}). The other two are the multichannel FI (FI_{MC}) and FI_k. The FI_{MC} is defined as the deviation between the powers P_H and P_L (i.e., the "freeze" and locomotor band) that are the sum of single powers over N channels (axis). The FI_k results from a Koopman spectral analysis.

Table 2.3: Current feature learning techniques for FoG detection

Authors	*Feature Learning technique*	*Description*
Bachlin et al., 2010	FI=$\frac{PS(3-8HZ)}{\text{PS}(0.5-3\text{HZ})}$, PI= $PS(3-8HZ)+PS(0.5-3HZ)$	Freeze Index (FI) is the ratio between the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz). Power Index (PI) is the sum of the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz).
Moore et al., 2008 & 2017	FI=$\frac{PS(3-8HZ)}{\text{PS}(0.5-3\text{HZ})}$; Average, standard deviation, variance, median, entropy, energy, power, FI, C_{XYNpks},$C_{XY\ max,}$ FI_K, FI_{MC}	Freeze Index as described above, the number of peaks, the maximum in the spectral coherence, the multichannel Index and various statistical domain features
Han et. al., 2003 & 2006	$\frac{WE(0-2.5HZ)}{WE(10-20HZ)}$;	Wavelet power on different frequency bands, i.e. how the power of the signal changes over different frequency bands
Johanov et al., 2009	$FI=\frac{PS(3-8HZ)}{PS(0.5-3HZ)}$	Freeze Index (FI) is the ratio between the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz)
Popovic et al., 2010	Pearson's correlation coefficient	Pearson's correlation coefficient measure the similarity between two signals
Delval et al., 2010	$FI=\frac{PS(3-8HZ)}{PS(0.5-3HZ)}$	Freeze Index (FI) is the ratio between the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz)
Djurić-Jovičić et al. 2010	Energy= $E_s(f)= \|X(f)\|^2$, stride length	The energy of a signal measures the signal strength
Niazmand et al., 2011	FI, the time of shaking foot, the number of pulses	Freeze Index (FI) is the ratio between the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz), the time of shaking foot, the number of pulses
Tripoliti et al. (2013)	Entropy = $-\sum_{i=1}^{n} p_i l_n p_i$	The entropy of a signal measures the distribution of the frequency components
Handojoseno et al, 2012	Wavelet Energy and Total Wavelet Entropy	The wavelet entropy calculates the degree of disorder of a signal and the wavelet power how the power of the signal changes over time
Mancini et al., 2012	$FI=\frac{PS(3-8HZ)}{PS(0.5-3HZ)}$	Freeze Index (FI) is the ratio between the "freeze" band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz)
Mazilu et al., 2012 & 2014	FI, power, energy, mean, standard deviation, entropy, variance, etc.	Freeze Index as described above and multiple statistical domain features

2.5 Contribution of the thesis

The contributions of this thesis are presented as follows:

- Firstly, we propose a new deep learning detection system in both subject independent settings and subject dependent settings witch uses a recurrent neural network with Long Short-Term Memory (LSTM) cells. To the best of our knowledge, our model achieved higher detection performance at least in terms of Sensitivity compared to previous studies.
- Secondly, we perform a combination of feature extraction techniques, included the new features suggested by Moore et al. (2017) and we evaluate different feature groups in each experiment.

Chapter 3: Experimental Setup

This section describes the dataset and the experimental method we conducted in order to answer the research questions of this thesis. In Subsection 3.1, we provide a detailed description of the dataset and the data acquisition method. Secondly, in Subsection 3.2, we define the pre-processing steps. The procedure used to extract features is presented in Subsection 3.3. We define the structure of our neural network model in Subsection 3.4. After that, in Subsection 3.5, we describe the construction of the training and the test set. We discuss the evaluation criteria for the experiments in Subsection 3.6. Finally, Subsection 3.7 presents the software used for the experiments.

3.1 Dataset Description

In this project, we utilize the publicly available Daphnet Freezing of Gait dataset, which was developed to benchmark automatic methods to recognize the FoG events from wearable 3D accelerometer sensors attached on the leg, the thigh, and the trunk of PWP (Bachlin et al., 2010). The dataset is the result of a research conducted by the Laboratory for Gait and Neurodynamics at Tel Aviv Sourasky Medical Center (TASMC) in Israel and the Wearable Computing Laboratory at ETH Zurich in Switzerland. The local Human Subjects Review Committee approved the research and it was executed under the honorable standards of the Declaration of Helsinki. Prior to describing the feature extraction approaches and the methodology of our study, we first direct attention to the data collection process.

3.1.1 Participants

The data was collected from 10 idiopathic PWP, who suffer from FoG episodes and could walk without assistance in the OFF period, i.e. when the Parkinsonian medication is no longer effective. Parkinson's disease patients with signs of other orthopedic or neurological diseases, dementia, severe hearing, and vision loss were excluded from the research. The subjects of the study were 7 males, and 3 females at the age of 59 to 75 years. They were diagnosed with different stages of PD and different HY score (table 3.1). At the end of the study, the results showed a high variability in the movement performance of the patients, meaning that during a non-freezing episode, the locomotion of some patients could not be distinguished from the locomotion of healthy elderly people, while others were walking in a more unbalanced and slow way.

3.1.2 Protocol

Three wireless accelerometer sensors placed on the ankle, thigh, and trunk of each patient provide the data of the study. The sensors recorded 3D accelerations at 64Hz and transferred the acceleration data to a wearable computer, which was attached to the trunk of the subjects (along with the 3rd sensor). The sensors were wirelessly connected to the wearable computer device, meaning that every acceleration data was transmitted via a Bluetooth link.

The ten patients were tested in the morning when eight of them were on the OFF stage of the medication cycle, i.e. they have not consumed any anti-Parkinsonian medication for the last 12 hours. Two patients were asked not to avoid their anti-Parkinsonian medication intake since they suffered from frequent freezing episodes during the ON state.

First, the patients received the instruction for use for the wearable assistance and they were informed about how to use the auditory cues in case of a freezing event, by synchronizing their locomotion with it. The research protocol is based on two sessions, both designed to replicate a normal daily walking routine. During the first session, the wearable computer collected all the data and conducted online FoG detection, without RAS feedback. In the second session the same procedure was followed, however, the RAS feedback was activated. The subjects performed three basic walking tasks in both 10-15 minutes sessions, specifically:

1. Walking in a straight line in the lab hall (back and forth with several 180˚ turns)
2. Walking randomly in the reception hall, with several 360˚ turns and initiated stops. The patients should turn and stop in different directions.
3. Walking according to daily life activities (ADL), such as moving from one room to another, going to the lab kitchen, picking up a cup of coffee or a glass of water and return to the examination room.

At the end of the study, the subjects returned to the lab room, received their medication, and were debriefed by the physiotherapist. All the participants followed the protocol, without any side effect or special accommodation.

Table 3.1: Detailed patient characteristics for the study of Bächlin et al. (2010)

Patient ID	*Gender*	*Age*	*Disease Duration*	*H&Y in ON*	*Tested in*
01	Male	66	*16*	*2*	*OFF*
02	Male	67	7	2	ON
03	Male	59	30	2.5	OFF
04	Male	62	3	3	OFF
05	Male	75	6	2	OFF
06	Female	63	22	2	OFF
07	Male	66	2	2.5	OFF
08	Female	68	18	4	ON
09	Male	73	9	2	OFF
10	Female	65	24	3	OFF
Mean ± STD		66.4± 4.8	13.7± 9.67	2.6±0.65	

Figure 3.1: FOG detection and feedback device developed by Bächlin et al. (2010)

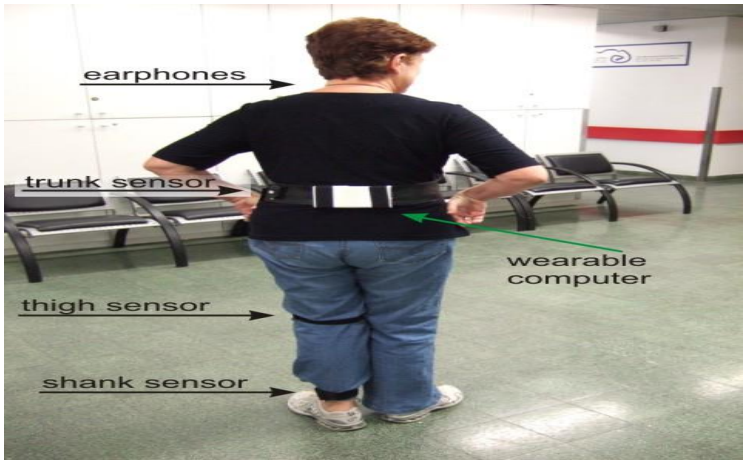
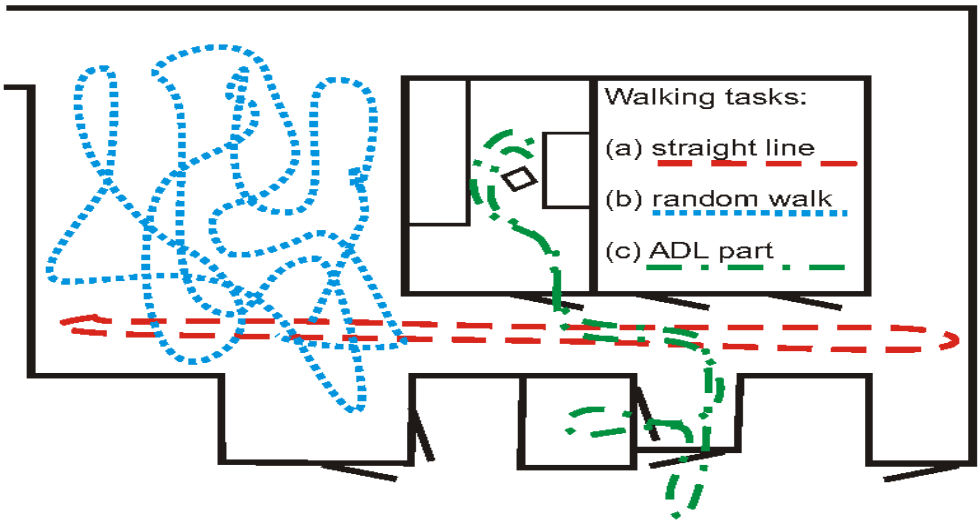


Figure 3.2: Sketch of the three basic walking tasks performed by the patients



3.1.3 Annotation of ground truth

During the sessions, a therapist was close to the patients for safety reasons. In addition, a physiotherapist recorded the sessions on a digital camera to capture all the FoG episodes. An assistant therapist was taking notes and another assistant labeled the real-time activity of each subject; standing, turning, walking, and freezing. Afterward, the physiotherapists analyzed the recorded video along with the manual labels to identify the ground truth labels, the duration, the onset, and the end of the FOG episodes as well. The onset of a FoG episode was detected when the locomotion pattern, more specifically the alternation from left to right step, was decelerating, and the end of the FoG episode was considered as the moment that the pattern was accelerating. The three 3D wearable accelerometer sensors, placed on the ankle, thigh and trunk of each patient, recorded acceleration at 64 Hz. The orientation and the acceleration of each patient were assessed by the accelerometer in three dimensions, X, Y, and Z respectively. The Y-axis measured the vertical acceleration of the subject, while the X-axis calculated the horizontal forward acceleration and the Z-axis measured the horizontal lateral acceleration. The physiotherapists labeled manually four activities, specifically standing, walking, turning, and freezing. However, after the video analysis, they categorized the activities standing, walking, and turning as “no freezing”. Altogether, 237 FoG events were detected (23.7±20.7 per subject), with duration of between 0.5s and 40.5s (7.3±6.7s). 93.2% of the FoG episodes lasted less than 20 seconds, where 50 percent were shorter than 5.4s.

Consequently, for each 3D accelerometer sensor reading of the Daphnet Dataset, the final ground truth class labels are 1 for “no freeze” and 2 for “freeze”. It is worth pointing out that there is an extra class

label (annotation) in the Daphnet dataset with the value 0, which is not part of the experiment, for instance the subject war performing activities unrelated to the protocol.

3.2 Preprocessing

One of the essential steps in the data mining process is data preprocessing. The data preprocessing includes data filtering, detection of outliers, replacement of missing values and feature selection and extraction (Attal et al., 2015; Rinnan et al., 2009). The set of preprocessing steps is standard in human activity recognition problems as well (Casale, Pujol, & Radeva, 2011). The first step involves sampling the sensor signals and compute the magnitude of the acceleration signals. The next and fundamental step is applying windowing techniques, where the sensor signals are sliced into partially overlapping windows. After the segmentation, features are extracted from each window and the resulting feature vectors are used as training data (Attal et al., 2015).

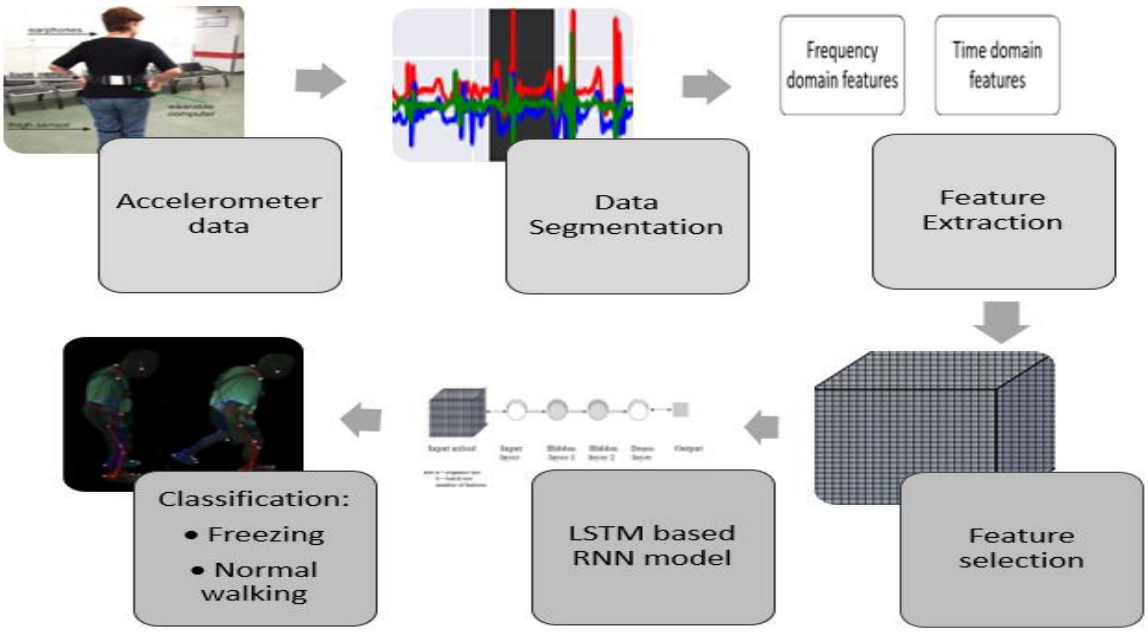
In our project, the first step of data preprocessing was to delete the annotation 0, since it is not part of the experiment and save the new data set into a new directory, while we keep the original file name. Moreover, an additional time series, A_{mag}, was obtained by computing the magnitude of the three accelerations: $A_{mag} = \sqrt{A^2x + A^2y + A^2z}$.

Our following step was the data windowing. Generally, there are three types of windowing techniques:

1. Sliding window - the signals are divided into a prefixed window length
2. Event defined window - specific events are identified at the preprocessing stage and used to define data segmentation
3. Activity defined window - the signals or data are divided based on the detection of an activity change.

In our research, we used a window function with a window length of 4 secs (256 samples) and an overlap of 0.5s (32 samples). The general preprocessing method and classification that we followed are represented in Figure 3.3.

Figure 3.3: FoG detection system



3.3 Feature Extraction Approaches

Previous studies in the human activity recognition suggest two main feature groups, which are usually extracted from the signals within a window (segment) (Attal et al., 2015; Casale et al., 2011; Preece, Goulermas, Kenney, & Howard, 2009; M. Zhang & Sawchuk, 2011). The two groups include a range of time- and frequency- domain features, which are used as input to train the recognition algorithms, at the training phase (Wilde, 2011). However, a combination of the two mentioned groups is usually desired.

The feature extraction approaches for this research are chosen based on the theoretical framework (see Subsection 2.3). We study two feature extraction schemes, as suggested by Mazilu et al. (2012) and Moore et al. (2017) and we compute statistical and time-domain and frequency-domain features over the signal data within each window. Moreover, in most of studies in FoG detection, a single input is used to extract features; for instance, a single axis from a single sensor (the vertical acceleration (X) of the sensor placed on the ankle) or the calculated magnitude of a single sensor. In our project we choose both inputs to extract features, however, we use an extra method suggested by Moore et al. (2017). We use multiple inputs as well, where the feature values are computed from a matrix of inputs, i.e. multiple axis of multiple body sensors. Overall, we extracted 145 features for our study. The interested reader can find a detailed table of the extracted feature in Appendix A (Table A).

3.3.1 Time-domain and Statistical Feature extraction

In our first feature extraction scheme, we extract a group of statistical and time-domain features. Time-domain features are simple statistical and mathematical metrics, easily computed, and applied to extract basic and significant signal information within the sliced window.

Studies in human activity recognition have reported a huge range of frequently extracted statistical and time-domain features. In our study, we extract the statistical features based on expert knowledge. Mazilu et al., (2013) and Moore et al., (2017) extracted a wide range of statistical and time domain features and investigated their relevance on the FoG detection. We decide to extract the following statistical features, based on their reported top ranked features:

- the average,
- standard deviation,
- variance,
- median,
- range,
- maximum
- and minimum.

We extracted the above-mentioned statistical features for each of the three accelerometer axis -x,y,z-, for each body sensor -ankle, thigh and trunk- and for the magnitude of each sensor. We obtain 84 statistical and time domain features in total. A detailed description of the computed statistical features and the feature extraction function is given in the table 3.2.

Table 3.2: Detailed description of the statistical and time domain features

Feature	*Description*	*Formula* *where T= window length*
Mean	The average signal value over the window	Mean ($\mu_{y)} = \sum_{i=1}^{T} \frac{y_i}{T}$
Standard deviation	The mean signal deviation compared to the average signal value over the window	$(\sum_{i=1}^{T} \frac{y_i - \mu_y}{T})^{\frac{1}{2}}$
Variance	The square of the standard deviation	$\sum_{i=1}^{T} \frac{y_i - \mu_y}{T}$
Median	The median signal value over the window	$median_{yi}(y_i)$
Range	The difference between the maximum and the minimum signal values over the window	$\lvert maxy_i(y_i) - miny_i(y_i) \rvert$
Maximum	The maximum signal value over the window	$max_{yi}(y_i)$
Minimum	The minimum signal value over the window	$min_{yi}(y_i)$

3.3.1 Frequency-Domain Features extraction

Frequency domain features capture the individual nature of a sensor signal (Figo, Diniz, Ferreira, & Cardoso, 2010). A signal can be transformed into the frequency domain by applying Fast Fourier transformation and show the distribution of the signal's energy over a range of frequencies (Wilde, 2011). In our second feature extraction scheme, we extract a range of frequency-domain features, based on the previous studies in FoG detection (Section 2.3).

As described in Section 2.3, in the field of FoG detection, the current standard frequency-domain feature extracted from raw signals is the Freezing Index (FI). Freeze Index (FI) was introduced by Moore et al. (2008), in their study on ambulatory monitoring of Freezing of Gait. They collected signal data from PWP and they applied FFT to transform the physical signals to frequency domain representation, or *spectrum of frequency components*. They analyzed the frequency characteristics of the patients and they proved that the leg movement during a FoG episode shows higher frequency component in the 3-8 Hz band and lower frequency component in the 0.5-3 Hz. Hence, the locomotor band was defined as 0.5-3 Hz and the "freeze" band as 3-8 Hz. Moreover, they computed the power spectra, which describes how the power of a signal is distributing into frequency components, in the locomotor and freeze band. They

reported significantly higher power spectra in the "freeze" band than in the locomotor band. Next, a Freeze Index (FI) was calculated as the power in the "freeze" band (3-8 Hz) divided by the power in the locomotor band (0.5-3 Hz). The power in "freeze" band or P_H, the power in locomotor band or P_L and the freeze ratio (FI) as well are considered as the standard feature group that required for the FoG detection. Afterwards, Bachlin et al. (2010) introduced another gait feature, the *power index,* which is calculated by summing up the P_H and P_L Besides the power in "freeze" band, the power in locomotor band, and the FI, the power index is considered as a standard gait feature, used in FoG detection studies (Capecci, Pepa, Verdini, & Ceravolo, 2016).

The energy of a signal is another frequency-domain feature, usually extracted in human activity recognition studies. The energy of a signal measures the "size" of a signal, or more precisely the signal strength. The energy of a signal is calculated by summing up the squared magnitudes of each FFT component of the signal. Then the sum is divided by the window length for normalization (S. Mazilu et al., 2012).

Recently, Moore et al. (2017) extracted 244 features and they employed new feature selection techniques based on voting methods with clusterability and correlation metrics, to identify the most discriminative ones. Moreover, they introduced a novel frequency domain feature, the multi-channel FI (FI_{MC}), which was ranked as the most informative feature for their anomaly score detector. Similar to the single input FI, the multi-channel FI is calculated from the power of the freeze band (P_H) divided by the power of locomotor band (P_L), which are summations of single powers over the N axis (N = 3 axis x 3 sensors).

Taking into consideration the studies of Mazilu et al. (2013) and Moore et al. (2017) and their reported top-ranked features we extract the following frequency domain features:

- the Freeze Index,
- the Power Index,
- the energy,
- the power in the freezing frequency bands
- the power in the locomotor frequency bands
- Multi-channel Freeze Index (FI_{MC})

of each of the three accelerometer axis -x,y,z-, for each body sensor -ankle, thigh, and trunk- and for the magnitudes of each sensor.

We obtain 61 frequency-domain features in total. A detailed description of the computed frequency domain features and the feature extraction formula is given in table 3.3.

Table 3.3: Detailed description of the frequency domain features

Feature	*Description*	*Formula*
Energy	The summation of the squared magnitudes of each FFT component of the signal, divided by the window length for normalization	$E_x = \sum_{w=0}^{L} \text{MagFx(w)}^2$, where L = window length
Freeze Index	The power in the "freeze" band (3-8Hz) divided by the power in the locomotor band (0.5-3Hz)	$FI = \frac{P_H}{P_L}$
Power	The sum of the power in the "freeze" band (3-8Hz) divided by the power in the locomotor band (0.5-3Hz)	$Power = P_H + P_L$
Power in the freeze band	The sum of the power spectrum in the "freeze" band of frequencies (3-8Hz) divided by the sampling frequency	$P_H = \frac{1}{2f_s}[\sum_{i=H_1+1}^{H_2}[Pxx_n(i)] + \sum_{i=H_1}^{H_2-1}[Pxx_n(i)]]$
Power in the locomotor bands	The sum of the power spectrum in the locomotor band of frequencies (0.5-3Hz) divided by the sampling frequency.	$P_L = \frac{1}{2f_s}[\sum_{i=L+1}^{H_1}[Pxx_n(i)] + \sum_{i=L}^{H_1-1}[Pxx_n(i)]]$
Multi-channel FI (FI_{MC})	The power of the freeze band (P_H) divided by the power of locomotor band (P_L), that are summations of single powers over the N axis (N = 3 axis x 3 sensors)	$FI_{MC} = \frac{P_H}{P_L}$, where $P_H = \frac{1}{2f_s}\sum_{n=1}^{N}[\sum_{i=H_1+1}^{H_2}[Pxx_n(i)] + \sum_{i=H_1}^{H_2-1}[Pxx_n(i)]]$ $P_L = \frac{1}{2f_s}\sum_{n=1}^{N}[\sum_{i=L+1}^{H_1}[Pxx_n(i)] + \sum_{i=L}^{H_1-1}[Pxx_n(i)]]$

* Given x is the signal, w the frequency, N the number of inputs, f_s the sampling frequency

** Given the FFT transform of the signal is conducted as: $F_x(w) = \int_{-\infty}^{+\infty} x(t)e^{-jwt}\,dt$

*** Given $\overline{F_x(w)}$ the conjugate of the FFT transform of the signal

**** Given power spectrum of a signal is $Pxx_{(w)} = F_x(w)\,\overline{F_x(w)}$

***** $H_1 = \frac{3N_{FFT}}{f_s}$, $H2 = \frac{8N_{FFT}}{f_s}$, $L = \frac{0.5N_{FFT}}{f_s}$

3.4 Freezing of Gait detection model

3.4.1 Recurrent Neural Networks

Artificial Neural Networks (ANNs) are computing systems inspired by the biological neural networks that compose the human and animal brain. The ANNs are designed to mathematically imitate the function of the human brain and learn how to perform a specific task. ANNs refer to as connectionist systems since their structure is based on the collection of connected artificial neurons (units or nods). A signal is transmitted from one neuron to another through their connection, or *synapse*. Then, the postsynaptic (receiving) neuron can process the signal and afterwards transmit it to another connected neuron (presynaptic). Neurons and their connections may have weights that differ as the learning phase proceeds and the strength of the signal is highly depended on the weights. As for the structure of an artificial neuron, it consists of layers, whithin signals travel through. The first layer receives the inputs/ feature vectors and the hidden layers of the neurons "learn" and understand the high in dimensionality feature representation. Finally, the last (output) layer receives the outputs of the hidden layers and estimates the probability of a specific class.

The original goal of the artificial neural networks method was to solve issues in the same manner the human brain would. However, a neural network does not have any memory of the inputs that has previously received, in other words, an ANN does not "remember" what has already seen. An ANN assumes that the main source of the data is an independent distributor. Thus, every time a new input is given to a neural network, it is just the same to the neural network as the input it received before. To fill this gap, recurrent neural networks were designed to use their internal memory to process the sequences of an input. RNNs are a powerful type of neural network and they are called "recurrent", due to their "memory" which allows them to transfer information along the network, for an infinitive time. Specifically, for every element of a sequence, they perform the same task and the output depends on the previous computations. In an RNN, due to its own loop (Appendix B, figure B1), the classification (output) of an input (feature vector) received by a hidden neuron at time t will be guided by all the previous inputs the same hidden neuron received at the time prior to t. Their unique architecture qualifies them for tasks such as language modeling, speech, and handwriting and image recognition.

3.4.2 Long Short-Term Memory Neural Network

As discussed in the previous subsection, an important advantage of RNNs is their ability to use information when mapping between sequences of inputs and outputs. However, for the standard RNN models, the range of contextual information that can be approached is limited. The influence of a given

input declines exponentially as it cycles around the recurrent connections of the network. In the scientific literature, the problem is referred to as the vanishing gradient problem (Bengio, Simard, & Frasconi, 1994; Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001; Hochreiter & Schmidhuber, 1997) and a detailed illustration is provided in Figure C.1 (Appendix C). In the 1990s, a large volume of studies was conducted to address the problem of vanishing gradients for RNNs. The studies introduced the non-gradient based algorithms, hierarchical sequence computation and time constants and delays. A favored approach is the Long Short-Term Memory (LSTM) architecture published by Hochreiter & Schmidhuber in 1997.

The Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN), consisting of a set of recurrently connections and memory blocks (Figure C.2, Appendix C). Generally, an LSTM network resembles the traditional recurrent neural network, with the main difference; the hidden layers contain memory blocks. The output layers are the same as the traditional RNNs. The memory blocks can be considered as different computer memory chips. Each memory block of an LSTM network includes four main parts that control the information flow:

- an input gate
- an output gate
- "forget" gate(s)
- and the self-connected memory cell(s)

A detailed illustration of an LSTM memory block with one memory cell is provided in Figure C.3 (Appendix C). The input, output and forget gait of the memory block use activation functions and compute values between 0 and 1. Through the small black circles, multiplication is applied to those values to control which information will flow into the memory block. The forget, input and output gates multiply the cell's previous state, input, and output respectively. The logistic sigmoid is usually the gate activation function ("f") so that the gate is closed when it has an activation near 0 and open when the activation is near 1. Logistic sigmoid or tahn are usually the activation functions of the cell input and output ('g' and 'h'). The weighted connection from the gates to the cell is presented with dashed lines. All the other recurrent connections have a pre-fixed weight of 1.0, so as the constancy of the memory cell's state between the steps is secured, as any interference is excluded.

In summary, the input gate permits or prevents an incoming signal to "awake" the state of the memory block. As long as, the activation of the input gate is 0, i.e closed, the new input will not overwrite the activation of the cell. Similarly, by opening the output gate, the memory state will be available to the network, since they are responsible to control whether the value in memory will be used to compute the

output of the block. Finally, the forget gate controls whether a value will remain in memory and based on the needs it allows the cell to forget or remember its previous state. The original form of LSTM networks contained only input and output gates. However, later in 2000, forget gates and peephole weight, were added to give a more extended and sophisticated LSTM. The forget gate provides the ability to the memory cell to reset itself. This ability was important in tasks requiring the erase of any information from previous inputs. The peephole weighted connections improved the LSTM's ability to learn tasks requiring precise timing. Moreover, the vanishing gradient problem decreases as the gates allow LSTM memory cells to save and access information over an infinite period.

During the past 30 years, LSTM networks have proved successful at a range of tasks requiring long memory, including music generation (Eck & Schmidhuber, 2002), handwriting recognition (Alex Graves, 2012; Liwicki, Graves, Bunke, & Schmidhuber, 2007), and speech recognition (A. Graves & Schmidhuber, 2005; Alex Graves, Fernández, Gomez, & Schmidhuber, 2006). The main reason for the success is that LSTM networks use a long range of contextual information. For our study, we will utilize an LSTM based RNN since a human activity recognition task is a classical sequence analysis problem, suitable for an LSTM network. In the next section, we will introduce our LSTM-based FoG predictor.

3.4.3 FoG deep detector model

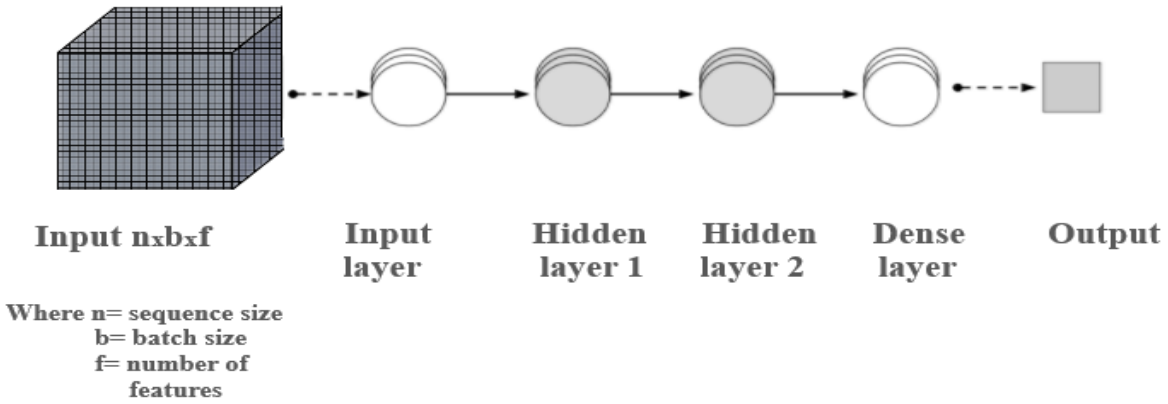
LSTM-based RNNs with deeper architecture are built by stacking multiple LSTM layers. Similar to the neural networks with deep architectures, multilayer LSTM models have been successfully used in speech recognition (A. Graves, Mohamed, & Hinton, 2013; Alex Graves, Jaitly, & Mohamed, 2013; Y. Zhang et al., 2015), as the experimental results suggest that this type of recurrent neural network outperforms the normal LSTM networks. Moreover, in the human activity recognition task, the deep LSTM based RNN model outperforms the normal LSTM models (Inoue, Inoue, & Nishida, 2016). According to (Y. Zhang et al., 2015), in a typical LSTM model, the features from a given instance are processed only by a single nonlinear layer before calculating the output of that time instant. Hence, the depth of the multilayer LSTM recurrent neural network has an additional meaning. The input, or more precisely, the feature vectors at a given time step are processed from multiple LSTM layers, i.e. more non-linear operations per time step. Hermans and Schrauwen (2013), suggested that deep multilayer LSTM recurrent networks allow the network to learn over the input at different time scales. Moreover, deep multilayer LSTM RNNs present another advantage over standard LSTM RNNs; they can successfully distribute the parameters over the space through the multiple layers.

In order to form a model capable of learning richer data representations, we build a neural network, with two LSTM layers (Figure 3.4). Given the 3D accelerometer data, collected from three sensors placed on

user's body, we use a sliding window with a length of 4s, we extract features from each window and we utilize them as input sequences for the LSTM model.

In RNNs model, the input involves three dimensions instead of two like the most machine learning algorithms. The three dimensions are the sequence size, the batch size, and the number of features. In each experiment of this thesis, the input into our LSTM-based RNN model will depend on the group of the extracted features we choose to feed the network. Thus, the input layer receives the 3D volumetric input. After receiving the input values, the input layer feeds into the first LSTM recurrent layer with several memory blocks ("smart" neurons), that afterward feeds into the second LSTM recurrent layer. Finally, since we treat FoG prediction as a binary classification problem, the second LSTM layer, feeds into a fully connected dense layer with a sigmoid activation function, which will return the predictions for the two classes (non-freeze and freeze) in the problem.

Figure 3.4: The structure of our LSTM base RNN model consisting of an input layer which receives the 3D input (feature pool), two hidden layers containing LSTM cells and a final dense layer with a sigmoid activation function which return the output (freeze or non-freeze)



3.5 Constructing a training and a test dataset

The construction of the training and the test dataset is the first step to start modeling. The method of constructing is different in user-dependent settings and user-independent settings. Before we explain the construction of a training and test dataset in the two methods, a more detailed description of the dataset is

required. Daphnet Dataset consists of 3d-accelerometer signals of ten Parkinson's disease patient, who manifested 237 FoG episodes in total during the experiment. The number of FoG instances (label '1') is very small compared to the number of" non-freeze" instances (label '0'). In 35531-time series, only the 3459 are labeled as freeze. In other words, our dataset appears to be heavily skewed in favor of the" non-freeze" class (label '0'). If we train our model on this skewed dataset, the representation of the data would not be perfect. To overcome the skewed imbalance, we use the Synthetic Minority Oversampling Technique (SMOTE), a technique proposed by Chawla et al. (2011). SMOTE, as its name would suggest, is an oversampling method that creates synthetic samples from the minority, or *abnormal*, class rather than copies. The SMOTE algorithm uses a distance measure and selects two or more similar instances. Afterward, it discomposes an instance one attribute at a time by a random number within the range of the neighboring instances. Finally, the method of over-sampling creates synthetic minority class examples and achieves higher classification performance in ROC space.

First, in the user-dependent method, we divide the data from each patient in a balanced way. A 70/30 division is chosen, such that 70% of the instances create the training dataset and 30% of the instances the test dataset. We preprocess the training dataset using SMOTE. After preprocessing of our skewed dataset, we get a new over-sampled training dataset where both "freeze" and "non-freeze" classes were equally represented. In the user-dependent method, we train our RNN model on the training set of each patient and evaluate the performance of our neural network model on the test dataset (unseen data) of the same patient.

In the user-independent method, the procedure is slightly different. Our training set consists of data from five participants, specifically patients 5, 6,7,8,9 and our test set consists of data from three remaining participants, specifically patients 1, 2 and 3. Similar to the user-dependent technique, we preprocess the training dataset using SMOTE, to obtain a more balanced training set. Afterwards, we train our RNN detector on general data obtained from the five patients, namely the training set, and we evaluate the performance of our RNN model on the unseen test data obtained from the rest 3 patients.

3.6 Evaluation Method

The most popular evaluation criteria in FoG detection are the Sensitivity and the Specificity (see Table 2.1). The prediction performance is based on window evaluation. The classifiers' output for each window is compared to the ground truth label. The windows that are correctly labeled as "freeze" episodes are counted as True Positive (TP), while the wrongly labeled as FoG episode are counted as False Positives (FP). The windows that the system failed to correctly label as FoG episode are counted as False Negatives

(FN) and the windows correctly labeled as no FoG are counted as True Negatives (TN). The Sensitivity (Sens =$\frac{TP}{TP+FN}$) measures the ratio of the correctly labeled FoG windows to the number of the referenced FoG windows, while the Specificity (Spec =$\frac{TN}{TN+FP}$) calculates the ratio of correctly predicted no-FoG windows to the number of the referenced no-FoG windows. The thesis will follow the above-mentioned recommendations.

Additionally, the area under the curve (AUC) in the ROC space is reported as performance metrics to evaluate our predictive model. The ROC curve plots the True Positive Rate (on the y-axis) versus the False Positive Rate (on the x-axis) for every possible classification threshold. Specifically, the True Positive Rate represents the ratio of the correctly labeled as "freeze" episodes to the number of actual "freeze" episodes, while the False Positive Rate represents the ration of the wrong predicted as "freeze" episodes to the number of the actual "non-freeze" episodes. The main purpose of Area under the Curve (AUC), is to quantify the performance of a classifier by using the ROC curve. Area under the Curve (AUC), is literally just the percentage of the area under the ROC curve. A classifier with an AUC score around 0.5 is considered as a poor classifier, while an AUC score above 0.8 indicates a successful classifier.

3.7 Software

The experiments are conducted using the programming language Python in Jupyter Notebook (version 5.0.0). We used the following Python libraries for data preprocessing, data analysis and visualizing the data:

- pandas
- sklearn
- scipy
- keras
- sys
- os
- imbalanced-learn
- numpy
- time
- matplotlib
- random
- math
- randomizedsearchcv

Chapter 4: Experiments and Results

This section presents the results of the experiments we performed to answer the research questions of the thesis (Section 1.3). In Section 4.1, we present the performance of our baseline model. The next four sections report the results of the experiments we conducted to evaluate the performance of our recurrent neural network. In each experiment, we use as input different feature groups from different sensor places. The performance of the model and the influence of each feature group will be discussed in further depth. More specifically, Section 4.2, presents the results of the first experiment, where we investigate the influence of the statistical domain features on the FoG detection. In Subsection 4.3, we investigate the influence of the frequency domain features on the FoG detection. Section 4.4, represents the results of the third experiment, where the combination of statistical and frequency domain features is used as input for the RNN model. Finally, in section 4.5, we represent the results of the fourth and last experiment, where we use features obtained from the evaluation of the feature importance using forests of trees. We will conclude this section with a summary of the experimental results and some preliminary conclusions.

4.1 Random Forest Classifier: A baseline for the study

In order to develop a baseline for the thesis and use it as a benchmark to compare the performance of our RRN model, we use a Random Forest algorithm both statistical domain and frequency domain features as input. For the development of our baseline model, i.e. ranking features and optimizing parameters, we utilize a random sample of five participants (70%) who experienced FoG events, namely patient 3, 5, 6, 7, 8, and 9 (with age for 66 ± 5.9 years old, with disease duration for 16.2 ± 10.15 years and H&Y score: 2.3 ± 0.44). For the out-of-sample tests, we utilize the remaining three participants (30%) who experienced FoG events, namely 1, 2 and 3 (with age for 66.8±4.1 years old, with disease duration for 11.2±9.6 years and H&Y score: 2.9±0.74). The optimal values of the parameters ware determined by use of Randomized Search, with a 5-fold cross validation splitting. After identifying the optimized parameters, namely n_estimators = 10 and max_features = ‘log2’, we determine the ranking of the features by recursive feature elimination. Table 4.3 presents the top 25 most informative features.

Afterwards, we trained the Random Forest classifier with only of the most important features and we tested the classifier on previously unseen data, i.e. the test set. In order to include statistical significance tests in our analysis and demonstrate whether the differences between the performances of the baseline and our RNN model are statistically significant, we perform multiple repeats experimental protocol, where we repeat the experiment 30 times. Table 4.1 presents the average AUC score, Specificity, and Sensitivity of the baseline model obtained after the multiple repeats. Afterwards, we evaluate the baseline

model in subject dependent settings, where we train our algorithm in a random sample of data (70%) of each patient separately and we tested it within 30 times in the remaining (30%) of unseen data of the same patient. Table 4.2 presents the grand mean of the AUC, Specificity and Sensitivity scores obtained from all the patients and the standard deviation as well. Table E.1 (Appendix E) presents the average AUC score, Specificity, and Sensitivity results of each patient separately.

Table 4.1: Performance of the baseline model (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
Random Forest	72%	92%	52 %

Table 4.2: Performance of the baseline model (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
Random Forest	90% ± 4%	97% ± 2%	83% ± 8%

From the table 4.1 and 4.2, we can derive that the baseline model performs slightly better in the subject dependent settings, where the overall differences between the AUC, Specificity, and Sensitivity scores of the two different methods are 18%, 5 % and 31% respectively.

Table 4.3: The Top 25 most informative features

1) FI_ankle_mag	**2) median_thigh_mag**	**3) max_thigh_mag**
4) mean_ankle_X	**5) fr_band_thigh_mag**	**6) median_ankle_X**
7) FI_ankle_X	**8) mean_ankle_Z**	**9) FI_ankle_Z**
10) range_thigh_*X*	**11) mean_thigh_Y**	**12) loc_band_thigh_Y**
13) FI_thigh_Y	**14) mean_thigh_Z**	**15) FI_thigh_Z**
16) mean_trunk_X	**17) max_trunk_X**	**18) fr_band_trunk_X**
19) mean_trunk_Y	**20) fr_band_trunk_Y**	**21) mean_trunk_Z**
22) loc_band_trunk_z	**23) FI_trunk_Z**	**24) FImc**
25) fr_band_thigh_mag		

4.2 Experiment 1: The influence of the statistical and time domain features

In the first experiment, the focus lays on the statistical and time domain features. We investigate how well our LSTM based RNN model can detect a freezing episode when we use as input the statistical features from each axis and each sensor. We perform 4 different rounds in the first experiment. In the first three rounds we use as input the statistical features from each sensor, ankle, thigh, and trunk respectively and in the fourth round a summation of them. The Random forest model was used as the baseline.

4.2.1 Round one: Statistical features of the ankle sensors

Following the same procedure as with the baseline, we performed a subject independent and a subject dependent method. In the subject independent method, we utilize a sample of five PWF for the RNN model development, i.e. hyperparameters optimization. For the out-of-sample tests, we utilize the data obtained from the remaining three PWF. The optimal value of hyperparameters are determined by the Randomized Search, with a 5-fold cross validation splitting. They are determined as optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After training our optimized model on the training set, we tested its performance on the unseen (test) data within 30 repeats. Afterwards, we performed the Student's t-test statistical test to determine whether the sets of scores (AUC, Sensitivity, and Specificity) in both model (RNN model and baseline model) are significantly different from one another. The table 4.4 presents the

average AUC, Sensitivity and Specificity scores of our optimized model obtained from the multiple repeats. The performance of the baseline model is provided as well.

Table 4.4: Performance of the RNN model for statistical features of the ankle sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	80%	78%	58%
Baseline model	72%	92%	52%

The results in table 4.4 show that the RNN model outperforms the baseline model in terms of AUC and Sensitivity scores. The overall difference in the AUC score between the two classifiers is 8% and significant at the $p<0.00001$ level, while the difference in the Sensitivity between the two models is 6% and significant at the $p= 0.0005$ level. However, the Specificity is worse in the RNN model, with a significant difference ($p<0.00001$) of 25%. Onwards, we perform the subject dependent evaluation, where the RNN model is trained with a training set consisted of a random sample (70%) and tested on the remaining (30%) unseen data of the same patient. Table 4.5 presents the grand mean and the standard deviation of AUC, Specificity and Sensitivity scores obtained from all the patients. Table E.2 (Appendix E) presents the average AUC score, Specificity, and Sensitivity results of each patient separately.

Table 4.5: Performance of the RNN model for statistical features of the ankle sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	94% ± 3%	92% ± 4%	79% ± 8%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

As shown in Table 4.5, in the subject dependent method, the RNN model outperforms the baseline in terms of AUC score, where the difference is 4%. However, the differences in Specificity and Sensitivity scores between the two models are 5% and 4% respectively in the advantage of the baseline model. Finally, comparing the performance of our model on both methods, the difference is in the maximal advantage of the subject dependent method. The overall differences between the average AUC, Specificity, and Sensitivity scores of the subject dependent and the subject independent method are 14%, 14%, and 22% respectively.

4.2.2 Round two: Statistical and time domain features of the thigh sensors

In the second round, we use statistical and time domain features, extracted from the signals obtained from the thigh sensor. In the subject independent method, we apply Randomized Search, with a 5-fold cross validation splitting, to tune the hyperparameters of the RNN model. The optimized hyperparameters are optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After the hyperparameters optimization, we trained our optimized model on a sample of five participants who experienced FoG events. As for out-of-sample tests, we utilize the unseen data obtained from the remaining three participants. The table 4.6 presents the average evaluation metric scores of our optimized model and the baseline model.

Table 4.6: Performance of the RNN model for statistical features of the thigh sensor (subject independent method)

Algorithm	*AUC score (average)*	*Specificity (average)*	*Sensitivity (average)*
RNN model	77%	78%	56%
Baseline model	72%	92%	52%

The RNN model is the best performing classifier in terms of AUC score and Sensitivity, while the differences with respect to the baseline model performance are 5% and 4% respectively. The above-mentioned differences are statistically significant at $p<0.00001$ level for the AUC score and at $p= 0.01$ level for the Sensitivity score. Nevertheless, the baseline model achieves higher Specificity score by 14%.

Following the subject independent method, we perform a subject dependent method, as well. In user dependent method both training and test set are obtained from the same patient. We evaluated our model in each one patient of the DAPHNET dataset. The table 4.7 reports the grand mean and standard deviation of the average performance measures of each participant. A full representation of the average AUC, Specificity, and Sensitivity scores from each patient separately is donated in Table E.3 in Appendix E.

Table 4.7: Performance of the RNN model for statistical features of the thigh sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model (Mean ±st.dev.)	95% ± 2%	94% ± 2%	78% ± 8%
Baseline model (Mean ±st.dev.)	90% ± 4%	97% ± 2%	83% ± 8%

It is apparent from the Table 4.7, that our RNN model outperforms the baseline in terms of the AUC score, with a difference of 5%. However, the Specificity and Sensitivity scores of the baseline are higher by 3% and 5% respectively.

Finally, comparing the results of Table 4.6 and 4.7, we conclude that our model performs better in the patient dependent method when we use as input statistical and time domain features of the thigh sensor.

4.2.3 Round three: Statistical and time domain features of the trunk sensors

In the third round of the first experiment, we used the statistical and time domain features of the trunk sensors as input for the RNN model. The procedure is the same with the previous two rounds. In the subject independent method, we determine the optimal hyperparameters as optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After training our optimized model on the training set (5 patient), we tested it on the unseen (test) data (the 3 remaining patients). The table 4.8 presents the results of the experiment.

Table 4.8: Performance of the RNN model for statistical features of the trunk sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	80%	88%	38%
Baseline model	72%	92%	52%

From the results in Table 4.8, we can conclude that the baseline performs better in terms of Specificity and Sensitivity with significant difference ($p<0.00001$) of 4% and 14% respectively. The average AUC score of the RNN model is higher by 8% and the statistical significant by $p<0.00001$value.

Afterwards, we conduct the subject dependent method. We report the grand mean and the standard deviation of the performance measures on the whole dataset (all the participants) for the RNN model and the baseline in the Table 4.9. The results on average performance measures on each patient separately is presented in Table E.4 in Appendix E.

Table 4.9: Performance of the RNN model for statistical features of the trunk sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	95% ± 3%	99% ± 4%	83% ± 10%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

Surprisingly, in the user dependent method, the RNN model outperforms the baseline model in terms of AUC score and Specificity, while the Sensitivity seems to be the same in both models.

Finally, the reported performance metric scores of the RNN model for the subject dependent method are higher than the subject independent method.

4.2.4 Round four: Statistical features of all the sensors

In the fourth and last round of the first experiment, we use as input for the RNN model all the statistical and time domain features from all the sensors. Following the same procedure with the previous three rounds, we used both a subject independent and a subject dependent method. In the subject independent method, we utilize a random sample of five PWF, for the RNN model development, i.e. hyperparameters optimization. For out-of-sample tests, we utilize the data obtained from the remaining three PWF. The Randomized Search helped us determine the optimized hyperparameters as optimizer = Nadam, number of neurons for the first LSTM cell = 50, number of neurons for the second LSTM cell= 20, epochs=20 and batch size=100. After training our optimized model on the training set, we test it on the unseen (test) data. The table 4.10 presents the results of the subject independent method of the fourth round.

Table 4.10: Performance of the RNN model for statistical features of all the sensors (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	79%	89%	34%
Baseline model	72%	92%	52%

As the table 4.10 shows, in the subject independent method, the RNN model performs significantly worse than the baseline in terms of Sensitivity and Specificity. The differences in Specificity and Sensitivity between the two models are 3% and 18% respectively. The above-mentioned differences are significant at the $p<0.00001$ value. However, the RNN model outperforms the baseline model in terms of the AUC score with a significant difference ($p<0.00001$) of 7%.

Afterwards, we performed the same steps, for the subject dependent method. The Table 4.11 presents the grand mean and standard deviation of the performance metrics scores obtained from all the patients of the Daphnet dataset. Table E.5 reports the results on average performance metrics scores on each patient separately.

Table 4.11: Performance of the RNN model for statistical features of all the sensors (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model (**Mean ±st.dev.**)	97% ± 1%	96% ± 2%	83% ± 8%
Baseline model (**Mean ±st.dev.**)	90% ± 4%	97% ± 2%	83% ± 8%

From the AUC point of view, the results in the table show that our model leads to better results than the baseline. The difference between the best performing models is 7%, while there is no difference between the Sensitivity of the two models. Comparing the Specificity of both model, the difference is in the minimal advantage of the baseline model.

Finally, the overall performance of our model is better in the subject dependent method than in the subject independent method, as the AUC score is higher by 18%, the Specificity by 7% and the Sensitivity by 49%.

4.3 Experiment 2: The influence of the frequency domain features

In the second experiment of the thesis, the focus lied on the frequency domain features. We analyze how well our model performs, when we use as input the frequency features from each axis and each sensor. Similarly, to the first experiment, we perform 4 different rounds. In each one of the first three round we use as input for our RNN model the statistical features from each one of the three sensor, ankle, thigh, and trunk respectively. In the fourth round we use a summation of them. The Random forest model was used as the baseline.

4.3.1 Round one: Frequency domain features of the ankle sensor

Following the same procedure as in the previous rounds, we used both a subject independent and a subject dependent method. In the subject independent method, we utilize a random sample of five PWF for the RNN model development, i.e. hyperparameters optimization, and we utilize the data obtained from the

remaining three PWF, for out-of-sample tests. The optimal value of hyperparameters were determined as optimizer = Adagrad, number of neurons for the first LSTM cell = 50, number of neurons for the second LSTM cell= 1, epochs=25 and batch size=75. After training our optimized model on the training set, we tested it on the unseen (test) data within 30 repeats. The table 4.12 presents the results on the average performance scores for our optimized model.

Table 4.12: Performance of the RNN model for frequency domain features of the ankle sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	93%	90%	81%
Baseline model	72%	92%	52%

The results in table 4.12 show that the RNN model outperforms the baseline model in terms of AUC and Sensitivity scores. The overall difference in the AUC score between the RNN model and the baseline is 21% and significant at the $p<0.00001$ level, while the difference in the Sensitivity between the two models is 29% and significant at the $p<0.00001$ level. However, the Specificity is worse in the RNN model, with a significant difference ($p<0.00001$) of 2%. Onwards, we perform the subject dependent evaluation, where the RNN model is trained with a training set consisted of a random sample (70%) and tested on the remaining (30%) unseen data of the same patient. Table 4.13 presents the grand mean and the standard deviation of AUC, Specificity and Sensitivity scores obtained from all the patients. Table E.6 (Appendix E) presents the average AUC score, Specificity, and Sensitivity results of each patient separately.

Table 4.13: Performance of the RNN model for frequency domain features of the ankle sensor (subject dependent method)

Algorithm	*AUC score* (average)	*Specificity* (average)	*Sensitivity* (average)
RNN model (**Mean ±st.dev.**)	94% ± 2%	91% ± 2%	81% ± 6%
Baseline model (**Mean ±st.dev.**)	90% ± 4%	97% ± 2%	83% ± 8%

From these findings, we can conclude that in the subject dependent method, the RNN model outperforms the baseline in terms of AUC score, where the difference is 4%. However, the differences in Specificity and Sensitivity scores between the two models are 2% and 4% respectively in the advantage of the baseline model. Finally, comparing the performance of our model on both methods, the difference is in the maximal advantage of the subject dependent method. The overall differences between the average AUC and Specificity scores of the subject dependent and the subject independent method are 1%. The Sensitivity is the same for both method.

4.3.2 Round two: Frequency domain features of the thigh sensor

In the second round, we utilize frequency domain features extracted from the signals of the thigh sensor. In the subject independent method, we trained our RNN model on data from 5 PWF. The optimized hyperparameters of our model were determined as optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After the training, we test the performance of our model on the unseen data (3 PWF) within 30 repeats. The table 4.14 presents the performance of our optimized model.

Table 4.14: Performance of the RNN model for frequency domain features of the thigh sensor (subject independent method)

Algorithm	*AUC score* (average)	*Specificity* (average)	*Sensitivity* (average)
RNN model	91%	75%	89%
Baseline model	72%	92%	52%

The RNN model is the best performing classifier in terms of AUC score and Sensitivity, while the differences with respect to the baseline model performance are 19% and 37% respectively. The above-mentioned differences are statistically significant at $p<0.00001$ level. Nevertheless, the baseline model achieves higher Specificity score by 17%. The difference is significant at the $p<0.00001$ value.

Following the subject independent method, we perform a subject dependent method, as well. In user dependent method both training and test set are obtained from the same patient. The table 4.15 reports the grand mean and standard deviation of the average performance measures of each participant. A full representation of the average AUC, Specificity, and Sensitivity scores from each patient separately is donated in Table E.7 in Appendix E.

Table 4.15: Performance of the RNN model for frequency domain features of the thigh sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model (**Mean ±st.dev.**)	93% ± 3%	91% ± 4%	78% ± 8%
Baseline model (**Mean ±st.dev.**)	90% ± 4%	97% ± 2%	83% ± 8%

From the AUC point of view, the results in the table show that our model leads to better results in the user dependent method. The difference between the two models is 3%. The baseline model outperforms our model in terms of Specificity and Sensitivity, with differences of 6% and 5% respectively.

Looking at the best-performing method, the subject dependent method achieves highest AUC and Specificity scores. Although small, the difference of the AUC score between the two methods is 3%, while the Specificity is 16%. Finally, only the Sensitivity score of our model is higher in the user independent method by 10%.

4.3.3 Round three: Frequency domain features of the trunk sensor

In the third round of the second experiment, we use the frequency domain features of the trunk sensor as input for the RNN model. The procedure is the same with the previous two rounds. In the subject independent method, we determined the optimal hyperparameters as optimizer = Adagrad, number of

neurons for the first LSTM cell = 50, number of neurons for the second LSTM cell= 1, epochs=25 and batch size=75. After training our optimized model on the training set (5 patient), we tested it on the unseen (test) data (the 3 remaining patients) within 30 repeats. The table 4.16 presents the performance of our optimized model.

Table 4.16: Performance of the RNN model for frequency domain features of the trunk sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	94%	86%	89%
Baseline model	72%	92%	52%

As the table 4.16 shows, our model performs worse than the baseline in terms of Specificity in the subject independent method. The significant difference ($p<0.00001$) is 6%. However, the AUC and Sensitivity scores of our model are higher than the baseline, with significant differences ($p<0.00001$) of 22% and 37% respectively.

Following the subject independent method, we performed a user dependent method, as well. The table 4.17 reports the grand mean and standard deviation of the performance metric scores on the whole dataset. A full representation of the average AUC, Specificity, and Sensitivity scores of each patient separately is presented in Table E.8 in Appendix E.

Table 4.17: Performance of the RNN model for frequency domain features of the trunk sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	94% ± 2%	92% ± 3%	81% ± 10%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

In the subject dependent settings, our model performs best in terms of AUC score by 4%. However, the baseline model performs best in terms of Specificity and Sensitivity. The differences are 5% and 2% respectively. Finally, the performance of our model is higher in terms of Sensitivity in the subject dependent method. Specifically, the Sensitivity is higher by 8%. The Specificity is lower by 6% and the AUC score remains the same in both methods.

4.3.4 Round four: Frequency domain features of all the sensors

In the fourth and last round of the second experiment, we use as input for our model the frequency domain features of all the sensors. Following the same procedure with the previous three rounds, in the subject independent method, we trained our model with data obtained from 5 PWF. We determined the optimized hyperparameters as optimizer = Adagrad, number of neurons for the first LSTM cell = 50, number of neurons for the second LSTM cell= 1, epochs=25 and batch size=75. After the training, we test the performance of our model on the unseen data (3 PWF) within 30 repeats. The table 4.18 presents the performance of our optimized model and the baseline.

Table 4.18: Performance of the RNN model for frequency domain features of all the sensors (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	86%	80%	77%
Baseline model	72%	92%	52 %

From the AUC and Sensitivity point of view, the results in Table 4.18 show that, our model performs better than the baseline, with significant differences ($p<0.00001$) of 14% and 25% respectively. The baseline model achieves better Specificity score with a difference of 12%.

Afterwards, we performed the same tasks, for the subject dependent method. The Table 4.19 presents the mean of the means of the performance metrics scores obtained by the patients of the Daphnet dataset. Table E.9 reports the results on average performance measures on each patient separately.

Table 4.19: Performance of the RNN model for frequency domain features of all the sensors (subject dependent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model **(Mean ±st.dev.)**	95% ± 1%	95% ± 2%	79% ± 9%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

It can be seen from the data in the table 4.19 that the difference in the AUC score between the two models is 5% in the favor of the RNN model, in the user specific method. However, the differences in the Specificity and Sensitivity are 2% and 4% respectively in the favor of the baseline model.

Finally, the overall average performance of our model is higher in the subject dependent method. Specifically, the average AUC score is higher by 10%, the Specificity by 15% and the Sensitivity by 4%.

4.4 Experiment 3: The influence of the statistical and frequency domain features

In the third experiment of the thesis, we focus on to what extend does our RNN model performs well, when we use as input the combination of statistical and frequency domain features from each axis and each sensor. Similarly, to the previous experiments, we perform 4 different rounds. In each one of the first three round we use as input for our model the statistical and frequency domain features from each sensor, i.e. ankle, thigh, and trunk and in the fourth one a summation of them. The Random forest model is used as the baseline.

4.4.1 Round one: Statistical and Frequency domain features of the ankle sensor

Following the standard procedure of our experiments, we perform both a subject independent method and a subject dependent. First, we train and evaluate our model in user-independent settings. We train our model with the optimized hyperparameters, more specifically optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After the training of our model with data obtained form 5 PWF, we test the performance of

our model on the unseen data (the rest 3 patients) within 30 repeats. The table 4.20 presents the performance of our optimized model.

Table 4.20: Performance of the RNN model for statistical and frequency domain features of the ankle sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	92%	86%	84%
Baseline model	72%	92%	52 %

The results in table 4.20 how that the RNN model outperforms the baseline model in terms of AUC and Sensitivity scores. The overall differences between the RNN model and the baseline model are 20% and 32% respectively and they are significant at the $p<0.00001$ level. However, the Specificity is worse in the RNN model, with a significant difference ($p<0.00001$) of 6%. Onwards, we perform the subject dependent evaluation, where the RNN model is trained with a training set consisted of a random sample (70%) of a patient and tested on the remaining (30%) unseen data of the same patient. Table 4.21 presents the grand mean and the standard deviation of the performance scores obtained from all the patients. Table E.10 (Appendix E) presents the average AUC score, Specificity, and Sensitivity results of each patient separately.

Table 4.21: Performance of the RNN model for statistical and frequency domain features of the ankle sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	96% ± 1%	94% ± 2%	82% ± 6%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

As shown in Table 4.21, in the subject dependent method, the RNN model outperforms the baseline in terms of AUC score, where the difference is 6%. However, the differences in Specificity and Sensitivity scores between the two models are 3% and 1% respectively in the advantage of the baseline model. Finally, comparing the performance of our model of both methods, the difference in AUC and Specificity scores are in the advantage of the subject dependent settings. The AUC score and Specificity are higher by 4% and 8% respectively. However, the Sensitivity of our model is slightly better (2%) in the subject independent method.

4.4.2 Round two: Statistical and Frequency domain features of the thigh sensor

In the second round, we use statistical and frequency domain features, however extracted from the signals of the thigh sensor. Following the same procedure as in the previous experiments, in the subject independent method, we trained our RNN model with optimized hyperparameters of optimizer = Adam, number of neurons for the first LSTM cell = 10, number of neurons for the second LSTM cell= 10, epochs=10 and batch size=25. After the training of our model with data collected from 5 PWF, we test the performance of our model on the unseen data (3 PWF) within 30 repeats. The table 4.22 presents the performance of our optimized model.

Table 4.22: Performance of the RNN model for statistical and frequency domain features of the thigh sensor (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	89%	82%	81%
Baseline model	72%	92%	52 %

The RNN model is the best performing classifier in terms of AUC score and Sensitivity, while the differences with respect to the baseline model performance are 17% and 29% respectively. The above-mentioned differences are statistically significant at $p<0.00001$ level. Nevertheless, the baseline model achieves higher Specificity score by 10%.

Following the subject independent method, we perform a subject dependent method, as well. In user dependent method both training and test set are obtained from the same patient. We evaluated our model on each one patient of the DAPHNET dataset. The table 4.23 reports the grand mean and standard deviation of the average performance measures of each participant. A full representation of the average AUC, Specificity, and Sensitivity scores from each patient separately is donated in Table E.11 in Appendix E.

Table 4.23: Performance of the RNN model for statistical and frequency domain features of the thigh sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	96% ± 1%	94% ± 2%	82% ± 6%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

It is apparent from the Table 4.23, that our RNN model outperforms the baseline in terms of the AUC score, with a difference of 6%. However, the Specificity and Sensitivity scores of the baseline are higher by 3% and 1% respectively.

Finally, comparing the results of Table 4.22 and 4.23, we conclude that the overall performance of our model is better in the patient dependent method. Specifically, the AUC score is higher by 7%, the Specificity by 12% and the Sensitivity by 1%.

4.4.3 Round three: Statistical and Frequency domain features of the trunk sensor

In the third round of the third experiment, we used the statistical and frequency domain features of the trunk sensor as input for the RNN model. The procedure is the same with the previous two rounds. In the subject independent method, we determined the optimal hyperparameters as optimizer = Adagrad, number of neurons for the first LSTM cell = 100, number of neurons for the second LSTM cell= 10, epochs=25 and batch size=50. After training our optimized model on the training set (5 patient), we tested

it on the unseen (test) data (the 3 remaining patients). The table 4.24 presents the performance of our model for the third round of this experiment.

Table 4.24: Performance of the RNN model for statistical and frequency domain features of the trunk sensor (subject independent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model	87%	89%	56%
Baseline model	72%	92%	52 %

As the table 4.24 shows, our model performs worse than the baseline in terms of Specificity in the subject independent method. The significant difference ($p<0.00001$) is 3%. However, the AUC and Sensitivity scores of our model are higher than the baseline, with significant differences ($p<0.00001$) of 15% and 4% respectively.

Following the subject independent method, we performed a user dependent method, as well. The table 4.25 reports the grand mean and standard deviation of the performance metric scores on the whole dataset. A full representation of the average AUC, Specificity, and Sensitivity scores of each patient separately is presented in Table E.12 in Appendix E.

Table 4.25: Performance of the RNN model for statistical and frequency domain features of the trunk sensor (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	96% ± 2%	96% ± 2%	81% ± 5%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

In the subject dependent settings, our model performs better in terms of AUC score by 6%. However, the baseline model performs better in terms of Specificity and Sensitivity. The differences are 1% and 2% respectively. Finally, the overall average performance of our model is higher in the subject dependent method, as the average AUC score is higher by 11% and the Specificity by 7% and the Sensitivity by 25%.

4.4.4 Round four: Statistical and Frequency domain features of all the sensors

In the fourth and last round of the third experiment, we use as input for the RNN model all the statistical and frequency domain features from all the sensors. Specifically, we used the whole feature pool of the thesis. Following the same procedure as in the previous three rounds, we used both a subject independent and a subject dependent method. In the subject independent method, we determined the optimal hyperparameters as optimizer = Nadam, number of neurons for the first LSTM cell = 25, number of neurons for the second LSTM cell= 1, epochs=10 and batch size=75. After training our optimized model on the training set (5 PWF), we tested it on the unseen (test) data (3 PWF). Afterwards, we performed the same tasks, for the subject dependent method. The Table 4.26 presents the performance of the RNN model for the subject independent method.

Table 4.26: Performance of the RNN model for statistical and frequency domain features of all the sensors (subject independent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model	87%	80%	79%
Baseline model	72%	92%	52 %

From the AUC and Sensitivity point of view, the results in Table 4.26 show that, our model performs better than the baseline, with significant differences ($p<0.00001$) of 15% and 27% respectively. The baseline model achieves better Specificity score with a difference of 12%.

Afterwards, we performed the same tasks, for the subject dependent method. The Table 4.27 presents the mean of the means of the performance metrics scores obtained by the patients of the Daphnet dataset. Table E.13 reports the results on average performance measures on each patient separately.

Table 4.27: Performance of the RNN model for statistical and frequency domain features of all the sensors (subject dependent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model **(Mean ±st.dev.)**	97% ± 1%	96% ± 2%	87% ± 3%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

From the AUC and Sensitivity point of view, the results in the table 4.27 show that our model leads to better results than the baseline. The difference between the best performing models is 7% and 4% respectively. Comparing the Specificity of both model, the difference is in the minimal advantage of the baseline model.

Finally, the overall performance of our model is better the model in the subject dependent method than in the subject independent method, as the average AUC score is higher by 10%, the Specificity by 16% and the Sensitivity by 9%.

4.5 Experiment 4: The influence of the 25 most informative features

In the fourth and last experiment of the thesis, we focus on the 25 most informative features determined by the recursive feature elimination. Unlikely the previous experiments, we perform only one round in the fourth experiment. The Random forest model was used as the baseline.

Following the same procedure as in the previous rounds, we use both a subject independent method and a subject dependent. In the subject independent method, we trained our RNN model with optimized hyperparameters of optimizer = Adadelta, number of neurons for the first LSTM cell = 75, number of neurons for the second LSTM cell= 25, epochs=25 and batch size=75. After training our optimized model on the training set (5 PWF), we tested it on the unseen (test) data (3 PWF) within 30 repeats. The table 4.28 presents the performance of our model for the 25 most informative features in subject independent settings.

Table 4.28: Performance of the RNN model for 25 most informative features (subject independent method)

Algorithm	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
RNN model	85%	89%	60%
Baseline	72%	92%	52%

The results in table 4.28 show that the RNN model outperforms the baseline model in terms of AUC and Sensitivity scores. The overall difference in the AUC score between the models is 13% and significant at the $p<0.00001$ level, while the difference in the Sensitivity is 8% and significant at the $p<0.00001$ level. However, the Specificity is worse in the RNN model, with a significant difference ($p<0.00001$) of 3%.

Onwards, we perform the subject dependent evaluation, where the RNN model is trained with a training set consisted of a random sample (70%) of a patient and tested on the remaining (30%) unseen data of the same patient. Table 4.29 presents the grand mean and the standard deviation of AUC, Specificity and Sensitivity scores obtained from all the patients. Table E.14 (Appendix E) presents the average AUC score, Specificity, and Sensitivity results of each patient separately.

Table 4.29: Performance of the RNN model for 25 most informative features (subject dependent method)

Algorithm	*AUC score*	*Specificity*	*Sensitivity*
RNN model **(Mean ±st.dev.)**	97% ± 1%	95% ± 2%	86% ± 5%
Baseline model **(Mean ±st.dev.)**	90% ± 4%	97% ± 2%	83% ± 8%

As shown in Table 4.29, in the subject dependent method, the RNN model outperforms the baseline in terms of AUC score and Sensitivity. The differences are 7% and 3% respectively. However, the difference in Specificity scores between the two models is 2% in the advantage of the baseline model.

Finally, comparing the performance of our model of both methods, the difference is in the maximal advantage of the subject dependent settings. The overall differences between the average AUC, Specificity, and Sensitivity scores of the subject dependent and the subject independent method are 12%, 6%, and 26% respectively.

4.6. A summary of the outstanding results

In the following Table 4.30, we present a summary of our outstanding result through the different experiments and rounds of the thesis.

	User independent			*User dependent*			
Experiment	***AUC score***	***Specificity***	***Sensitivity***	***AUC score***	***Specificity***	***Sensitivity***	*Experiment*
Baseline	72%	92%	52%	90%	97%	83%	Baseline
RNN model, experiment 2, round 1 *(frequency features, ankle sensor)*	**93%**	**90%**	**81%**	**95%**	**97%**	**83%**	*RNN model, experiment 1, round 3 (statistical features, trunk sensor)*
RNN model, experiment 2, round 3 (frequency features, trunk sensor)	**93%**	**85%**	**89%**	**97%**	**96%**	**87%**	*RNN model, experiment 3, round 4 (frequency and statistical features from all the sensor)*
RNN model, experiment 3 round 1 (frequency and statistical features from the ankle sensor)	92%	86%	84%	95%	94%	83%	*RNN model, experiment 1, round 2 (statistical features, thigh sensor)*
RNN model, experiment 3, round 2 *(frequency and statistical features from thigh sensor)*	89%	82%	80%	97%	96%	83%	RNN model, experiment 1, round 4 (statistical features, all the sensors)
RNN model, experiment 3, round 4 (frequency and statistical features from all the sensor)	87%	80%	78%	97%	95%	86%	RNN model, experiment 4 (top 25 informative features)

Table 4.30: Summary of the top performing results

Chapter 5: General Discussion and Conclusions

The following sections provide a general discussion on deep FoG detection, present the conclusions of our research, and give recommendations and suggest a direction for future research. First, we revisit the results and conclusions of the different experiments and answer the research questions that were formulated in Subsection 1.2. Afterwards, we provide answers to the problem statement. Subsequently, in Subsection 5.3 we conclude with several directions and suggestions for future research.

5.1. Answers to the research questions

In this thesis, we formulate the problem statement as:

Problem statement: *To what extent can FoG episodes be detected from sensor information?*

In order to find the answer to this problem statement, we focus on three research questions:

Research Question 1 (RQ1): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-independent method?*

Research Question 2 (RQ2): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-dependent method?*

Research question 3 (RQ3): *Which features contribute most to detect FoG episodes in Parkinson's disease patient?*

In the remainder of this section, we will discuss and conclude upon each of the research questions.

Research question 1 (RQ1): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject-independent method?*

In Subsection 3.4, we present our deep FoG detector, a recurrent neural network model with LSTM cells. To investigate to what extent our LSTM-based RNN model perform well on the FoG detection task, we perform an offline detection in subject-independent and subject dependent settings. We train our LSTM based recurrent model with features extracted from the accelerometer data of the Daphnet Dataset. Afterwards, we validate the performance of our deep learning detector on unseen test sets. We perform several experiments and in each experiment, we used different feature groups. More specifically, we performed 3 experiments, each one of them consist of 4 rounds. The number of the experiment indicate a different feature group, i.e. statistical features for the first one, frequency features for the second one and a summation of statistical and frequency features for the third one. The number of each round indicates the different body sensor, i.e. ankle for the first one, thigh for the second, trunk for the third and a summation of them for the fourth one. Moreover, we perform a fourth experiment, where we use as input for our model the 25 most informative features which we determine by recursive feature elimination.

One of the main challenges in the human activity recognition tasks is the subject sensitivity. The subject sensitivity occurs due to the fact that different people have different walking patterns. Especially in the FoG detection task, where the subjects are mostly elderly people with gait patterns affected by the PD, the development of a system that is robust to intrapersonal variability is highly challenging. In our first research question, we tried to address this issue and design a model that captures as much variability as possible from accelerometer data of specific subjects and works reliably when it is tested on data from different subjects.

The results of the experiments revealed in Table 4.30 show that the best performing model in subject independent settings is in experiment 2, round 3 where we use as input for our model frequency domain data extracted from the signals of the trunk sensor. Specifically, our recurrent model outperforms the baseline; achieving AUC score of 93%, Specificity of 85% and Sensitivity of 89%. In other words, our model is able to detect 85 non-FoG episodes out of 100 non-FoG episodes and 89 FoG episodes out of 100 FoG episodes. The performance of our model is satisfying since it can not only detect the “freeze” events, but also the non-freeze events so as not to disturb the patients with unnecessary rhythmic sounds. The findings of the current study are consistent with those of Moore et al. (2017) who showed that the frequency domain features extracted from the trunk sensor, especially the FI, are the most informative.

Regarding the model design, to the best of our knowledge, achieves better published performance in subject-independent settings in offline FoG detection, at least in terms of Sensitivity. Specifically, results obtained by (Bachlin et al., 2010; Hammerla, Kirkham, Andras, & Ploetz, 2013; S. Mazilu et al., 2012; S. T. Moore et al., 2017), were lower than the results obtained from the presented approach. As reported in

Table 5.1, Mazilu et al. (2012) and Bachlin et al. (2010) achieved a higher Specificity but with a lower Sensitivity.

Table 5.1: Comparison of our proposed model against previous methods in terms of FoG detection

Reference	*Window size (Tolerance)*	*Specificity*	*Sensitivity*
(Bachlin et al., 2010)	4s (1s)	81.6% (online) 86.9% (offline)	73.1% (online) **87.1% (offline)**
(Hammerla et al., 2013)	-	82%	82%
(S. Mazilu et al., 2012)	4s (1s)	95.38%	66.25%
(S. T. Moore et al., 2017)	8s (0.5s)	84.5%	**87.5%**
Our model	4s (0.5s)	85%	89%

Our answer to the first research question therefore reads: a recurrent neural network with LSTM cells successfully applied to the FoG detection task in subject independent settings. Moreover, it performs well in terms of AUC, Specificity, and Sensitivity when the inputs consist of frequency domain features extracted from the trunk sensor.

Research question 2 (RQ2): *To what extent can a recurrent deep learning model detect successfully a FoG episode in the subject dependent method?*

To investigate to what extent the LSTM-based RNN model perform well on the FoG detection task, we conducted a subject dependent method as well and we determined the efficiency of our model in each user separately. In the user specific method, we train our model with features extracted from the accelerometer data of each patient of the Daphnet Dataset. Afterwards, we validate the performance of our deep learning detector on unseen test sets from the same patient. Similarly, with the user independent method, we perform exactly the same number of experiments and rounds.

The results of the experiments revealed in Table 4.30, show that the best performing model is in experiment 1, round 3, where statistical and time domain data extracted from the signals of the trunk sensor are used as input for our model. Specifically, in the subject dependent method, our model, outperforms the baseline in terms of average AUC score by 95%, while both model performs equally in terms of average Specificity by 87% and average Sensitivity by 83%. Moreover, our model outperforms the baseline based on AUC score, in every experiment and every round and on Sensitivity in five out of the thirteen rounds. However, our model failed to outperform the baseline based on Specificity, yet the differences between the two models are not extent.

The performance of our model in the subject dependent method is satisfying as well. Firstly, the average Sensitivity score was always higher than 55% and the average Specificity score was always higher than 80% for each patient. At this point, we would like to draw the attention to the fact that Bachlin et al. (2010) reported significant worst result in terms of Specificity for the patient 01 (38.7% Specificity and 97.1% sensitivity) and in terms of Sensitivity for the patient 08 (28.7% sensitivity and 87.7% specificity). In our study, the high variability in the movement performance of the patients which was reported in previous studies, is absent.

Our answer to the second research question therefore reads: a recurrent neural network with LSTM cells successfully applied to the FoG detection task in subject dependent settings and it performs well in terms of AUC, Specificity, and Sensitivity.

Research question 3 (RQ3): *Which features contribute most to detect FoG episodes in Parkinson's disease patient?*

In Subsection 3.3, we state that in the field of HAR the most popular extracted features are extracted the statistical and/or time-domain features and frequency domain features. In our research, we extracted features from both domains and from all the axis from all the body sensors. Moreover, we extracted a novel feature, the multichannel Freeze Index (FI_{MC}), introduced by Moore et al. (2017). In order to provide an answer to this research question, we use different feature groups as input for our RNN model. We performed 13 experiments where we used each feature group as input to our deep detector model. Moreover, we determined the 25 most informative features which we determine by recursive feature elimination.

The findings on the contribution of each feature group are quite revealing in several ways. Firstly, in the subject independent method, the performance of our model trained with frequency domain features is significantly better compared to the performance of our model trained with statistical and time domain

features. The results of all the rounds of the second experiments, i.e. frequency domain features as input for our model, demonstrate better scores that the results of the first experiment, where the statistical features are used. Moreover, at least in the subject independent method, our model with frequency domain features outperforms the baseline in terms of AUC score and Sensitivity. The same outperformance is observed in the third experiment, where we use the combination of statistical and frequency domain features as input, however with slightly lower scores than the scores achieved in the second experiment (frequency features as input).

Moreover, comparing the performance of our model in the user-specific method of both, we observe that there is a minimal advantage of the statistical domain features. However, the differences between the performance of our model using statistical domain and frequency domain features in the user-specific method are really small. A final remark of our study is that a single sensor, namely the trunk, could collect sufficient signal data for the FoG detection system in both methods.

Our answer to the third research question therefore reads: in the user independent method, the frequency domain features are the most promising candidates for the FoG deep detector model, especially the frequency domain features extracted from the signals of the trunk sensor. In the user depended method the statistical and time domain features extracted from the trunk sensor are the most promising candidate.

5.2. Answer to the problem statement

In this thesis, we addressed the following **problem statement:** ***To what extent can FoG episodes be detected from sensor information?***

Our answer reads as follows:

A recurrent neural network with LSTM cells successfully detect FoG episodes from sensor information. Moreover, a LSTM based RNN performs well in terms of AUC score, Sensitivity, and Specificity. By answering the research questions, we demonstrated that with the right combination of features extracted from sensor signal and the optimization of the hyperparameters of the RNN model we can achieve significant high performance on objective and subjective FoG detection. The proposed model outperforms the baseline and achieved better scores in comparison with existing FoG detection studies at least in terms of Sensitivity.

To conclude this subsection, we discuss the relevance of the thesis for the research community. The problem statement of the thesis arises due to the lack of a FoG detection model in the Parkinson's disease. In our research, the proposed model achieves high-performance scores in the user independent method

and it can later be implemented on wearable devices such as smartphone to enable objective real-time detection of the FoG episodes and provide rhythmic sounds to the patients. Moreover, it has the advantage of being user dependent as well, meaning that the same deep learning model can be used by any patient successfully.

5.3 Limitations and future research

Pointing out the major limitation of this research, will help us suggest directions for future research. The presented research is limited by its applicability to online detection in order to provide rhythmic cues. This limitation represents a major challenge that the FoG detection task faces, since in real-time detection, a short delay between the onset of a FoG episode and its detection is desired. Further research could focus on testing the presented RNN model online and report whether the model is capable of process the accelerometer data and detect a FoG episode successfully and on its onset so as the rhythmic cue will be provided to the patients and suppress the freeze episode.

A second limitation relates to the window size optimization, with respect to the work of Moore et al. (2017). Hence, a second possible area of future research would be to investigate which is the optimal window size and tolerance of the presented model. An interesting approach to this limitation could be running different experiments, with several window sizes and tolerance. Enlarging or shortening the window size and/or the tolerance, could yield other results.

Finally, in this thesis, we identified the optimized number of batches, epochs, optimizer, and number of neurons in the first and second LSTM cell. However, in the deep learning models, there are numerous hyperparameters that need to be set and tuned. Therefore, future research should concentrate on the investigation of different optimized hyperparameters of our RNN model. A great example could be a research in the optimization of learning rate, momentum, network weight initialization, activation functions and dropout regularization. The findings could be of great value for the presented deep FoG detection model.

Bibliography

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical Human Activity Recognition Using Wearable Sensors. *Sensors*, *15*(12), 31314–31338. https://doi.org/10.3390/s151229858

Bachlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J. M., Giladi, N., & Troster, G. (2010). Wearable Assistant for Parkinsons Disease Patients With the Freezing of Gait Symptom. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 436–446. https://doi.org/10.1109/TITB.2009.2036165

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bloem, B. R., Hausdorff, J. M., Visser, J. E., & Giladi, N. (2004). Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena. *Movement Disorders: Official Journal of the Movement Disorder Society*, *19*(8), 871–884. https://doi.org/10.1002/mds.20115

Bulling, A. (2014). A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Computing Surveys*, *46*, 1–33. https://doi.org/10.1145/2499621

Capecci, M., Pepa, L., Verdini, F., & Ceravolo, M. G. (2016). A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. *Gait & Posture*, *50*, 28–33. https://doi.org/10.1016/j.gaitpost.2016.08.018

Casale, P., Pujol, O., & Radeva, P. (2011). Human Activity Recognition from Accelerometer Data Using a Wearable Device. In *Pattern Recognition and Image Analysis* (pp. 289–296). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21257-4_36

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. *ArXiv:1106.1813 [Cs]*. https://doi.org/10.1613/jair.953

Cole, B. T., Roy, S. H., & Nawab, S. H. (2011). Detecting freezing-of-gait during unscripted and unconstrained activity. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2011*, 5649–5652. https://doi.org/10.1109/IEMBS.2011.6091367

Cubo, E., Leurgans, S., & Goetz, C. G. (2004). Short-term and practice effects of metronome pacing in Parkinson's disease patients with gait freezing while in the "on" state: randomized single blind evaluation. *Parkinsonism & Related Disorders*, *10*(8), 507–510. https://doi.org/10.1016/j.parkreldis.2004.05.001

de Boer, A. G., Wijker, W., Speelman, J. D., & de Haes, J. C. (1996). Quality of life in patients with Parkinson's disease: development of a questionnaire. *Journal of Neurology, Neurosurgery, and Psychiatry*, *61*(1), 70–74.

Delval, A., Snijders, A. H., Weerdesteyn, V., Duysens, J. E., Defebvre, L., Giladi, N., & Bloem, B. R. (2010). Objective detection of subtle freezing of gait episodes in Parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, *25*(11), 1684–1693. https://doi.org/10.1002/mds.23159

Djurić-Jovicić, M. D., Jovicić, N. S., Radovanović, S. M., Stanković, I. D., Popović, M. B., & Kostić, V. S. (2014). Automatic identification and classification of freezing of gait episodes in Parkinson's disease patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, *22*(3), 685–694. https://doi.org/10.1109/TNSRE.2013.2287241

Donovan, S., Lim, C., Diaz, N., Browner, N., Rose, P., Sudarsky, L. R., … Simon, D. K. (2011). Laserlight cues for gait freezing in Parkinson's disease: an open-label study. *Parkinsonism & Related Disorders*, *17*(4), 240–245. https://doi.org/10.1016/j.parkreldis.2010.08.010

Eck, D., & Schmidhuber, J. (2002). *A First Look at Music Composition Using LSTM Recurrent Neural Networks*. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale.

Factor, S. A., & Weiner, W. (2007). *Parkinson's Disease: Diagnosis & Clinical Management, Second Edition*. Demos Medical Publishing.

Fahn, S. (1995). The freezing phenomenon in parkinsonism. *Advances in Neurology*, *67*, 53–63.

Figo, D., Diniz, P. C., Ferreira, D. R., & Cardoso, J. M. (2010). Preprocessing Techniques for Context Recognition from Accelerometer Data. *Personal Ubiquitous Comput.*, *14*(7), 645–662. https://doi.org/10.1007/s00779-010-0293-9

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks* (Vol. 385). https://doi.org/10.1007/978-3-642-24797-2

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369–376). New York, NY, USA: ACM. https://doi.org/10.1145/1143844.1143891

Graves, A., Jaitly, N., & Mohamed, A. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU*.

Graves, A., Mohamed, A. r, & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). https://doi.org/10.1109/ICASSP.2013.6638947

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (Vol. 4, pp. 2047–2052 vol. 4). https://doi.org/10.1109/IJCNN.2005.1556215

Hammerla, N. Y., Kirkham, R., Andras, P., & Ploetz, T. (2013). On Preserving Statistical Characteristics of Accelerometry Data Using Their Empirical Cumulative Distribution. In *Proceedings of the 2013 International Symposium on Wearable Computers* (pp. 65–68). New York, NY, USA: ACM. https://doi.org/10.1145/2493988.2494353

Han, J. H., Lee, W. J., Ahn, T. B., Jeon, B. S., & Park, K. S. (2003). Gait analysis for freezing detection in patients with movement disorder using three dimensional acceleration system. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)* (Vol. 2, p. 1863–1865 Vol.2). https://doi.org/10.1109/IEMBS.2003.1279781

Han, J., Sun Jeon, H., Suk Jeon, B., & Park, K. (2006). Gait detection from three dimensional acceleration signals of ankles for the patients with Parkinson's disease.

Handojoseno, A. M. A., Shine, J. M., Nguyen, T. N., Tran, Y., Lewis, S. J. G., & Nguyen, H. T. (2012). The detection of Freezing of Gait in Parkinson's disease patients using EEG signals based on Wavelet decomposition. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2012*, 69–72. https://doi.org/10.1109/EMBC.2012.6345873

Handojoseno, A. M. A., Shine, J. M., Nguyen, T. N., Tran, Y., Lewis, S. J. G., & Nguyen, H. T. (2013). Using EEG spatial correlation, cross frequency energy, and wavelet coefficients for the prediction of Freezing of Gait in Parkinson's Disease patients. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in*

Medicine and Biology Society. Annual Conference, *2013*, 4263–4266. https://doi.org/10.1109/EMBC.2013.6610487

Hashimoto, T. (2006). Speculation on the responsible sites and pathophysiology of freezing of gait. *Parkinsonism & Related Disorders*, *12*(Supplement 2), S55–S62. https://doi.org/10.1016/j.parkreldis.2006.05.017

Hausdorff, J. M., Lowenthal, J., Herman, T., Gruendlinger, L., Peretz, C., & Giladi, N. (2007). Rhythmic auditory stimulation modulates gait variability in Parkinson's disease. *The European Journal of Neuroscience*, *26*(8), 2369–2375. https://doi.org/10.1111/j.1460-9568.2007.05810.x

Hermans, M., & Schrauwen, B. (2013). Training and Analysing Deep Recurrent Neural Networks. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 190–198). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5166-training-and-analysing-deep-recurrent-neural-networks.pdf

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. A field guide to dynamical recurrent neural networks. IEEE Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Inoue, M., Inoue, S., & Nishida, T. (2016). Deep Recurrent Neural Network for Mobile Human Activity Recognition with High Throughput. *ArXiv:1611.03607 [Cs]*. Retrieved from http://arxiv.org/abs/1611.03607

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *79*(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

Liwicki, M., Graves, A., Bunke, H., & Schmidhuber, J. (2007). A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*.

Macht, M., Kaussner, Y., Möller, J. C., Stiasny-Kolster, K., Eggert, K. M., Krüger, H.-P., & Ellgring, H. (2007). Predictors of freezing in Parkinson's disease: a survey of 6,620 patients. *Movement Disorders: Official Journal of the Movement Disorder Society*, *22*(7), 953–956. https://doi.org/10.1002/mds.21458

Mancini, M., Priest, K. C., Nutt, J. G., & Horak, F. B. (2012). Quantifying Freezing of Gait in Parkinson's disease during the Instrumented Timed Up and Go test. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2012*, 1198–1201. https://doi.org/10.1109/EMBC.2012.6346151

Mazilu, S., Blanke, U., Hardegger, M., Tröster, G., Gazit, E., Dorfman, M., & Hausdorff, J. M. (2014). GaitAssist: A wearable assistant for gait training and rehabilitation in Parkinson's disease. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)* (pp. 135–137). https://doi.org/10.1109/PerComW.2014.6815179

Mazilu, S., Calatroni, A., Gazit, E., Roggen, D., Hausdorff, J. M., & Tröster, G. (2013). Feature Learning for Detection and Prediction of Freezing of Gait in Parkinson's Disease (pp. 144–158). Presented at the International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39712-7_11

Mazilu, S., Hardegger, M., Zhu, Z., Roggen, D., Tröster, G., Plotnik, M., & Hausdorff, J. M. (2012). Online detection of freezing of gait with smartphones and machine learning techniques. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops* (pp. 123–130). https://doi.org/10.4108/icst.pervasivehealth.2012.248680

Mhyre, T. R., Boyd, J. T., Hamill, R. W., & Maguire-Zeiss, K. A. (2012). Parkinson's Disease. *Sub-Cellular Biochemistry*, *65*, 389–455. https://doi.org/10.1007/978-94-007-5416-4_16

Moore, S. T., MacDougall, H. G., & Ondo, W. G. (2008). Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of Neuroscience Methods*, *167*(2), 340–348. https://doi.org/10.1016/j.jneumeth.2007.08.023

Moore, S. T., Pham, T. T., Lewis, S. J. G., Nguyen, D. N., Dutkiewicz, E., Fuglevand, A. J., … Leong, P. H. W. (2017). Freezing of Gait Detection in Parkinson's Disease: A Subject-Independent Detector Using Anomaly Scores. Retrieved from https://opus.lib.uts.edu.au/handle/10453/113362

Niazmand, K., Tonn, K., Zhao, Y., Fietzek, U. M., Schroeteler, F., Ziegler, K., … Lueth, T. C. (2011). Freezing of Gait detection in Parkinson's disease using accelerometer based smart clothes. In *2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 201–204). https://doi.org/10.1109/BioCAS.2011.6107762

Nieuwboer, A. (2008). Cueing for freezing of gait in patients with Parkinson's disease: a rehabilitation perspective. *Movement Disorders: Official Journal of the Movement Disorder Society*, *23 Suppl 2*, S475-481. https://doi.org/10.1002/mds.21978

Nutt, J. G., Bloem, B. R., Giladi, N., Hallett, M., Horak, F. B., & Nieuwboer, A. (2011). Freezing of gait: moving forward on a mysterious clinical phenomenon. *The Lancet. Neurology*, *10*(8), 734–744. https://doi.org/10.1016/S1474-4422(11)70143-0

Popovic, M. B., Djuric-Jovicic, M., Radovanovic, S., Petrovic, I., & Kostic, V. (2010). A simple method to assess freezing of gait in Parkinson's disease patients. *Brazilian Journal of Medical and Biological Research*, *43*(9), 883–889. https://doi.org/10.1590/S0100-879X2010007500077

Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., & Howard, D. (2009). A Comparison of Feature Extraction Methods for the Classification of Dynamic Activities From Accelerometer Data. *IEEE Transactions on Biomedical Engineering*, *56*(3), 871–879. https://doi.org/10.1109/TBME.2008.2006190

Rahman, S., Griffin, H. J., Quinn, N. P., & Jahanshahi, M. (2008). The factors that induce or overcome freezing of gait in Parkinson's disease. *Behavioural Neurology*, *19*(3), 127–136.

Rinnan, \AAsmund, Nørgaard, L., van den Berg, F., Thygesen, J., Bro, R., & Engelsen, S. B. (2009). *Data pre-processing*. Academic Press, San Diego, CA. Retrieved from http://www.academia.edu/download/33499848/2009_Data_pre-processing.pdf

Rubinstein, T. C., Giladi, N., & Hausdorff, J. M. (2002). The power of cueing to circumvent dopamine deficits: a review of physical therapy treatment of gait disturbances in Parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, *17*(6), 1148–1160. https://doi.org/10.1002/mds.10259

Schaafsma, J. D., Balash, Y., Gurevich, T., Bartels, A. L., Hausdorff, J. M., & Giladi, N. (2003). Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease. *European Journal of Neurology*, *10*(4), 391–398.

Suteerawattananon, M., Morris, G. S., Etnyre, B. R., Jankovic, J., & Protas, E. J. (2004). Effects of visual and auditory cues on gait in individuals with Parkinson's disease. *Journal of the Neurological Sciences*, *219*(1–2), 63–69. https://doi.org/10.1016/j.jns.2003.12.007

WHO Program on Neurological Diseases and Neuroscience, World Health Organization Department of Mental Health and Substance Abuse, & World Federation of Neurology. (2004). Atlas : country resources for neurological disorders 2004 : results of a collaborative study of the World Health Organization and the World Federation of Neurology. Retrieved from http://www.who.int/iris/handle/10665/43075

Wilde, A. G. (2011, April). *Activity recognition for motion-aware pervasive systems* (masters). University of Fribourg (Switzerland). Retrieved from https://eprints.soton.ac.uk/272433/

Zhang, M., & Sawchuk, A. A. (2011). A Feature Selection-based Framework for Human Activity Recognition Using Wearable Multimodal Sensors. In *Proceedings of the 6th International*

Conference on Body Area Networks (pp. 92–98). ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). Retrieved from http://dl.acm.org/citation.cfm?id=2318776.2318798

Zhang, Y., Chen, G., Yu, D., Yao, K., Khudanpur, S., & Glass, J. (2015). Highway Long Short-Term Memory RNNs for Distant Speech Recognition. *ArXiv:1510.08983 [Cs]*. Retrieved from http://arxiv.org/abs/1510.08983

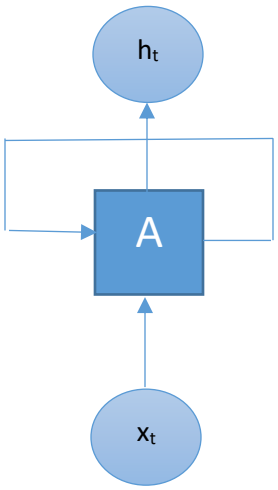
Appendix A

Table A1: The extracted features examined in this work

Sensor	*Axis*	*Extraction*	*Feature ID*
Ankle	X	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	*1-12*
Ankle	Y	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	13-24
Ankle	Z	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	25-36
Ankle	A_{mag}	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	37-48
Thigh	X	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	49-60
Thigh	Y	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	61-72
Thigh	Z	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	73-84
Thigh	A_{mag}	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	85-96
Trunk	X	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	97-108
Trunk	Y	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	109-120
Trunk	Z	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	121-132
Trunk	A_{mag}	Average, standard deviation, variance, median, range, maximum, minimum, FI, the power, the energy, the power of freezing and locomotor bands	133-144
Ankle, Thigh, Trunk	X, Y, Z	Multi-channel FI ($FI_{MC)}$	145

Appendix B

Figure B1: Recurrent Neural Network



In the above diagram, a chunk of a recurrent neural network, A, receives an input x at time t and passes an outputs h. Though the loop of the network, the information is passed from one step to the next one.

Appendix C

Figure C1: RNNs vanishing gradient problem (Alex Graves, 2012)

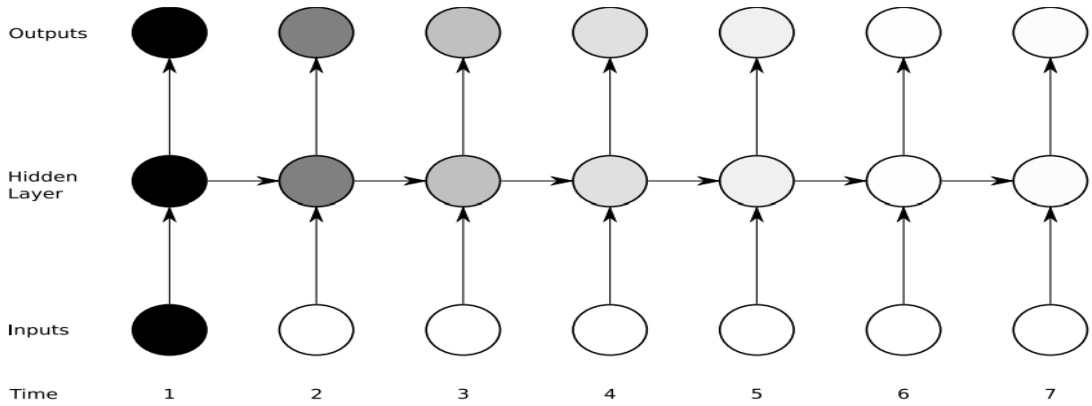
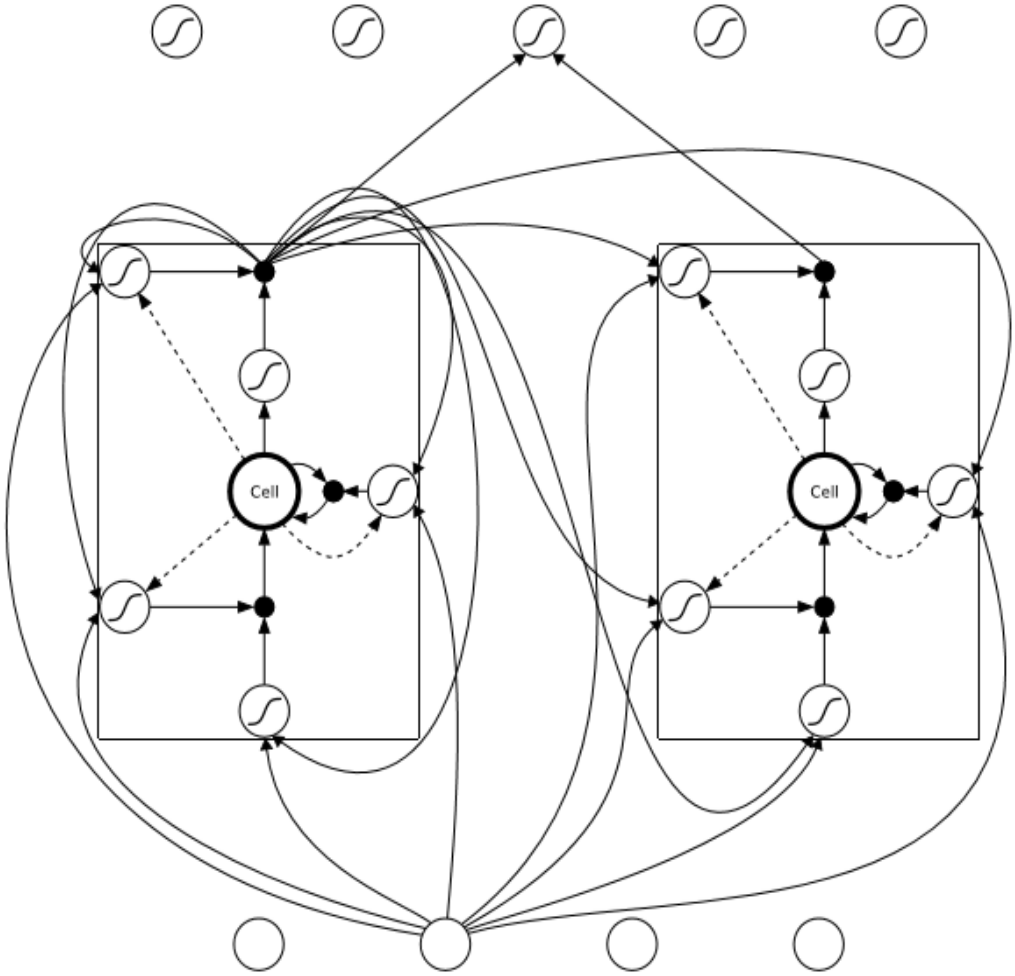
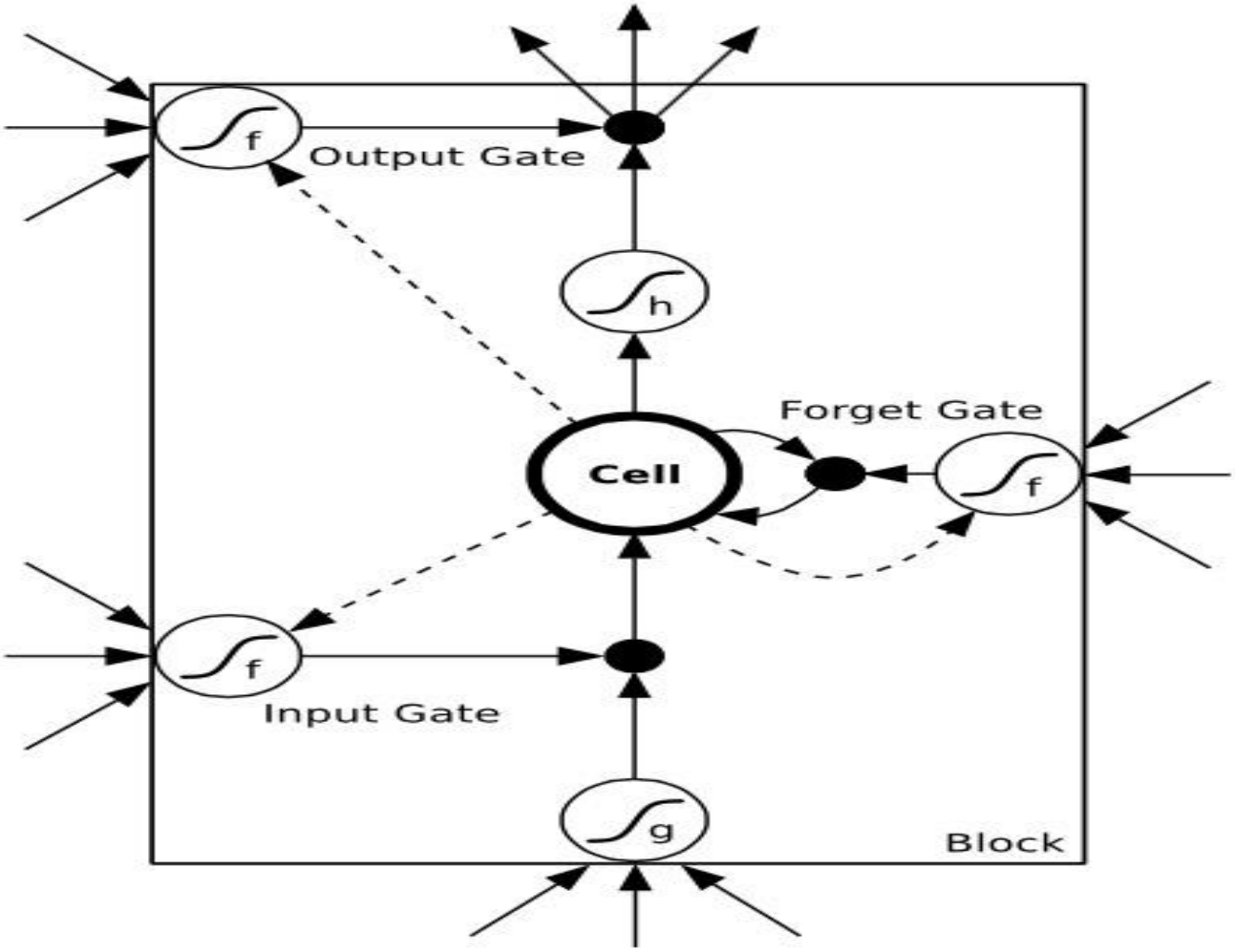


Figure C2: An LSTM network (Alex Graves, 2012)



The above network consists of four input units, a hidden layer of two LSTM memory blocks (single-cell) and five output units. Each block has four inputs but only one output and not all the connections are shown in the figure.

Figure C3: LSTM memory block with one cell (Alex Graves, 2012)



Appendix E

In order to compare our model with the baseline we used the population of 30 test scores of the two models in each experiment. Afterwards we used the Student's t-test statistical test to determine whether the two sets of data are significantly different from one another. In the following tables, we present the performance metric scores of each. The highlighted scores are the highest between the two models, with statistical significant difference (the pvalue of the t-test is provided). Due to the lack of space, we report only the statistical significant scores.

Table E.1: Performance of the baseline model (subject dependent method)

Patient ID	*AUC score* *(average)*	*Specificity* *(average)*	*Sensitivity* *(average)*
1	84%	98%	69%
2	93%	98%	87%
3	87%	93%	81%
5	88%	93%	83%
6	92%	99%	84%
7	86%	98%	74%
8	94%	97%	92%
9	92%	97%	93%

Table E.2: Performance of the RNN model for statistical features of ankle sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.94)** (p<0.0001) **Sensitivity (0.82)** (p<0.0001) Specificity (0.92)	AUC score (0.84) Sensitivity (0.69) **Specificity (0.98)** (p<0.0001)
	Patient 2	**AUC score (0.97)** (p<0.0001) Specificity (0.95) Sensitivity (0.81)	AUC score (0.93) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.87)** (p<0.0001)
	Patient 3	**AUC score (0.92)** (p<0.0001) Specificity (0.86) **Sensitivity (0.86)** (p<0.0001)	AUC score (0.87) **Specificity (0.93)** (p<0.0001) Sensitivity (0.81)
	Patient 5	**AUC score (0.92)** (p<0.0001) Specificity (0.87) Sensitivity (0.80)	AUC score (0.88) **Specificity (0.93)** (p<0.0001) **Sensitivity (0.83)** (p=0.001)
	Patient 6	**AUC score (0.96)** (p<0.0001) Specificity (0.94) Sensitivity (0.83)	AUC score (0.92) **Specificity (0.99)** (p<0.0001) Sensitivity (0.84)
	Patient 7	**AUC score (0.89)** (p<0.0001) Specificity (0.94) Sensitivity (0.60)	AUC score (0.86) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.74)** (p<0.0001)
	Patient 8	AUC score (0.943) Specificity (0.95) Sensitivity (0.76)	AUC score (0.947) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.92)** (p<0.0001)
	Patient 9	**AUC score (0.97)** (p<0.0001) Specificity (0.96) Sensitivity (0.84)	AUC score (0.92) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.93)** (p<0.0001)

Table E.3: Performance of the RNN model for statistical features of thigh sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.96)** (p<0.00001) Specificity (0.97) **Sensitivity (0.75)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.95)** (p<0.00001) Specificity (0.94) Sensitivity (0.76)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) **Sensitivity (0.87)** (p<0.00001)
	Patient 3	**AUC score (0.96)** (p<0.00001) Specificity (0.92) **Sensitivity (0.85)** (p<0.00001)	AUC score (0.87) **Specificity (0.93)** (p<0.00001) Sensitivity (0.81)
	Patient 5	**AUC score (0.93)** (p<0.00001) Specificity (0.90) Sensitivity (0.74)	AUC score (0.88) **Specificity (0.93)** (p<0.00001) **Sensitivity (0.83)** (p<0.00001)
	Patient 6	**AUC score (0.98)** (p<0.00001) Specificity (0.96) **Sensitivity (0.89)** (p<0.00001)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) Sensitivity (0.84)
	Patient 7	**AUC score (0.92)** (p<0.00001) Specificity (0.95) Sensitivity (0.66)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) **Sensitivity (0.74)** (p<0.00001)
	Patient 8	AUC score (0.90) Specificity (0.95) Sensitivity (0.68)	**AUC score (0.94)** (p<0.00001) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.97)** (p<0.00001) Specificity (0.93) Sensitivity (0.87)	AUC score (0.95) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.93)** (p<0.00001)

Table E.4: Performance of the RNN model for statistical features of trunk sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.96)** (p<0.0001) Specificity (0.94) **Sensitivity (0.97)** (p<0.0001)	AUC score (0.84) **Specificity (0.98)** (p<0.0001) Sensitivity (0.69)
	Patient 2	**AUC score (0.94)** (p<0.0001) Specificity (0.94) Sensitivity (0.72)	AUC score (0.93) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.87)** (p<0.0001)
	Patient 3	**AUC score (0.95)** (p<0.0001) Specificity (0.90) **Sensitivity (0.87)** (p<0.0001)	AUC score (0.87) **Specificity (0.93)** (p<0.0001) Sensitivity (0.81)
	Patient 5	**AUC score** (0.91) (p<0.0001) Specificity (0.85) Sensitivity (0.81)	AUC score (0.88) **Specificity (0.93)** (p<0.0001) **Sensitivity (0.83)** (p<0.0001)
	Patient 6	**AUC score (0.98)** (p<0.0001) Specificity (0.98) **Sensitivity (0.89)** (p<0.0001)	AUC score (0.92) **Specificity (0.99)** (p<0.0001) Sensitivity (0.84)
	Patient 7	**AUC score (0.96)** (p<0.0001) Specificity (0.95) **Sensitivity (0.81)** (p<0.0001)	AUC score (0.86) **Specificity (0.98)** (p<0.0001) Sensitivity (0.74)
	Patient 8	AUC score (0.88) Specificity (0.91) Sensitivity (0.66)	**AUC score (0.94)** (p<0.0001) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.92)** (p<0.0001)
	Patient 9	**AUC score (0.98)** (p<0.0001) Specificity (0.96) Sensitivity (0.89)	AUC score (0.95) **Specificity (0.97)** (p=0.0054) **Sensitivity (0.93)** (p<0.0001)

Table E.5: Performance of the RNN model for statistical features from all the sensors (subject dependent method)

		RNN model	*Baseline*
Patient Dependent Method	*Patient 1*	**AUC score (0.94)** (p<0.00001) Specificity (0.97) Sensitivity (0.66)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) **Sensitivity (0.69)** (p<0.00001)
	Patient 2	**AUC score (0.98)** (p<0.00001) Specificity (0.97) **Sensitivity (0.89)** (p<0.00001)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	**AUC score (0.96)** (p<0.00001) **Specificity (0.95)** (p<0.00001) Sensitivity (0.79)	AUC score (0.87) Specificity (0.93) Sensitivity (0.81)
	Patient 5	**AUC score (0.96)** (p<0.00001) **Specificity (0.95)** (p<0.00001) Sensitivity (0.82)	AUC score (0.88) Specificity (0.93) **Sensitivity (0.83)** (p<0.03)
	Patient 6	**AUC score (0.99)** (p<0.00001) Specificity (0.97) **Sensitivity (0.92)** (p<0.00001)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) Sensitivity (0.84)
	Patient 7	**AUC score (0.95)** (p<0.00001) Specificity (0.97) **Sensitivity (0.78)** (p<0.00001)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) Sensitivity (0.74)
	Patient 8	**AUC score (0.95)** (p<0.00001) Specificity (0.90) Sensitivity (0.85)	AUC score (0.94) **Specificity (0.97)** (p=0.001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.99)** (p<0.00001) Specificity (0.97) Sensitivity (0.90)	AUC score (0.95) Specificity (0.97) **Sensitivity (0.93)** (p<0.005)

Table E.6: Performance of the RNN model for frequency features of the ankle sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.95)** (p<0.0001) Specificity (0.91) Sensitivity (0.87)	AUC score (0.84) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.96)** (p<0.0001)
	Patient 2	**AUC score (0.95)** (p<0.0001) Specificity (0.92) Sensitivity (0.80)	AUC score (0.93) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.87)** (p<0.0001)
	Patient 3	**AUC score (0.92)** (p<0.0001) Specificity (0.87) Sensitivity (0.80)	AUC score (0.87) **Specificity (0.93)** (p<0.0001) Sensitivity (0.81)
	Patient 5	**AUC score (0.92)** (p<0.0001) Specificity (0.88) Sensitivity (0.78)	AUC score (0.88) **Specificity (0.93)** (p<0.0001) **Sensitivity (0.83)** (p<0.0001)
	Patient 6	**AUC score (0.97)** (p<0.0001) Specificity (0.95) Sensitivity (0.85)	AUC score (0.92) **Specificity (0.99)** (p<0.0001) Sensitivity (0.84)
	Patient 7	**AUC score (0.92)** (p<0.0001) Specificity (0.94) Sensitivity (0.746)	AUC score (0.86) **Specificity (0.98)** (p<0.0001) Sensitivity (0.744)
	Patient 8	AUC score (0.91) Specificity (0.90) Sensitivity (0.72)	**AUC score (0.94)** (p<0.0001) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.92)** (p<0.0001)
	Patient 9	**AUC score (0.97)** (p<0.0001) Specificity (0.94) Sensitivity (0.89)	AUC score (0.95) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.93)** (p<0.0001)

Table E.7: Performance of the RNN model for frequency features from the thigh sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.92)** (p<0.00001) Specificity (0.93) **Sensitivity (0.76)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	AUC score (0.96) Specificity (0.93) Sensitivity (0.83)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) **Sensitivity (0.87)** (p<0.00001)
	Patient 3	**AUC score (0.91)** (p<0.00001) Specificity (0.83) **Sensitivity (0.85)** (p<0.00001)	AUC score (0.87) **Specificity (0.93)** (p<0.00001) Sensitivity (0.81)
	Patient 5	**AUC score (0.92)** (p<0.00001) Specificity (0.86) Sensitivity (0.81)	AUC score (0.88) **Specificity (0.93)** (p<0.00001) **Sensitivity (0.83)** (p=0.02)
	Patient 6	**AUC score (0.94)** (p<0.00001) Specificity (0.96) Sensitivity (0.67)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) **Sensitivity (0.84)** (p<0.00001)
	Patient 7	**AUC score (0.95)** (p<0.00001) Specificity (0.90) **Sensitivity (0.87)** (p<0.00001)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) Sensitivity (0.74)
	Patient 8	AUC score (0.87) Specificity (0.89) Sensitivity (0.63)	**AUC score (0.94)** (p<0.00001) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score** (0.97) (p<0.00001) Specificity (0.95) Sensitivity (0.85)	AUC score (0.95) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.93)** (p<0.00001)

Table E.8: Performance of the RNN model for frequency features of the trunk sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.95)** (p<0.00001) Specificity (0.96) Sensitivity (0.70)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.97)** (p<0.00001) Specificity (0.93) **Sensitivity (0.89)** (p<0.0001)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	AUC score (0.90) (p<0.00001) Specificity (0.86) Sensitivity (0.75)	AUC score (0.87) **Specificity (0.93)** (p<0.00001) **Sensitivity (0.81)** (p<0.00001)
	Patient 5	**AUC score (0.93)** (p<0.00001) Specificity (0.87) Sensitivity (0.82)	AUC score (0.88) **Specificity (0.93)** (p<0.00001) Sensitivity (0.83)
	Patient 6	**AUC score (0.96)** (p<0.00001) Specificity (0.96) Sensitivity (0.80)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) **Sensitivity (0.84)** (p<0.00001)
	Patient 7	**AUC score (0.96)** (p<0.00001) Specificity (0.89) **Sensitivity (0.98)** (p<0.00001)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) Sensitivity (0.74)
	Patient 8	AUC score (0.92) Specificity (0.94) Sensitivity (0.67)	**AUC score (0.94)** (p<0.00001) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.96)** (p<0.00001) Specificity (0.93) Sensitivity (0.84)	AUC score (0.95) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.93)** (p<0.00001)

Table E.9: Performance of the RNN model for frequency features from all the sensors (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.95)** (p<0.00001) Specificity (0.94) **Sensitivity (0.83)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.97)** (p<0.00001) Specificity (0.96) Sensitivity (0.82)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) **Sensitivity (0.87)** (p<0.00001)
	Patient 3	**AUC score (0.94)** (p<0.00001) Specificity (0.90) Sensitivity (0.82)	AUC score (0.87) **Specificity (0.93)** (p<0.00001) Sensitivity (0.81)
	Patient 5	**AUC score (0.94)** (p<0.00001) Specificity (0.92) Sensitivity (0.76)	AUC score (0.88) Specificity (0.93) **Sensitivity (0.83)** (p<0.00001)
	Patient 6	**AUC score (0.96)** (p<0.00001) Specificity (0.98) Sensitivity (0.81)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) **Sensitivity (0.84)** (p<0.00001)
	Patient 7	**AUC score (0.92)** (p<0.00001) Specificity (0.96) Sensitivity (0.56)	AUC score (0.86) **Specificity (0.98)** (p=0.0003) **Sensitivity (0.74)** (p=0.0001)
	Patient 8	AUC score (0.95) Specificity (0.95) Sensitivity (0.81)	AUC score (0.94) **Specificity (0.97)** (p=0.02) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.98)** (p<0.00001) Specificity (0.96) Sensitivity (0.88)	AUC score (0.95) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.93)** (p<0.00001)

Table E.10: Performance of the RNN model for statistical and frequency features from the ankle sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.94)** (p<0.0001) Specificity (0.91) **Sensitivity (0.80)** (p<0.0001)	AUC score (0.84) **Specificity (0.98)** (p<0.0001) Sensitivity (0.69)
	Patient 2	**AUC score (0.97)** (p<0.0001) Specificity (0.95) Sensitivity (0.84)	AUC score (0.93) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.87)** (p<0.0001)
	Patient 3	**AUC score (0.95)** (p<0.0001) Specificity (0.93) Sensitivity (0.82)	AUC score (0.87) Specificity (0.93) Sensitivity (0.81)
	Patient 5	**AUC score (0.94)** (p<0.0001) Specificity (0.91) Sensitivity (0.80)	AUC score (0.88) **Specificity (0.93)** (p<0.0001) **Sensitivity (0.83)** (p<0.0001)
	Patient 6	**AUC score (0.98)** (p<0.0001) Specificity (0.98) Sensitivity (0.83)	AUC score (0.92) **Specificity (0.99)** (p<0.0001) Sensitivity (0.84)
	Patient 7	**AUC score (0.94)** (p<0.0001) Specificity (0.94) Sensitivity (0.69)	AUC score (0.86) **Specificity (0.98)** (p<0.0001) **Sensitivity (0.74)** (p<0.0001)
	Patient 8	**AUC score (0.96)** (p<0.0001) Specificity (0.92) Sensitivity (0.86)	AUC score (0.94) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.92)** (p<0.0001)
	Patient 9	**AUC score (0.98)** (p<0.0001) Specificity (0.95) Sensitivity (0.90)	AUC score (0.95) **Specificity (0.97)** (p<0.0001) **Sensitivity (0.93)** (p<0.0001)

Table E.11: Performance of the RNN model for statistical and frequency features of the thigh sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.97)** (p<0.00001) Specificity (0.97) **Sensitivity (0.77)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.97)** (p<0.00001) Specificity (0.94) **Sensitivity (0.89)** (p<0.001)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	**AUC score (0.94)** (p<0.00001) **Specificity (0.94)** (p=0.0001) Sensitivity (0.73)	AUC score (0.87) Specificity (0.93) **Sensitivity (0.81)** (p<0.00001)
	Patient 5	**AUC score (0.94)** (p<0.00001) Specificity (0.91) Sensitivity (0.79)	AUC score (0.88) **Specificity (0.93)** (p=0.0001) **Sensitivity (0.83)** (p<0.00001)
	Patient 6	**AUC score (0.98)** (p<0.00001) Specificity (0.98) **Sensitivity (0.87)** (p<0.00001)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) Sensitivity (0.84)
	Patient 7	**AUC score (0.96)** (p<0.00001) Specificity (0.93) **Sensitivity (0.82)** (p<0.00001)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) Sensitivity (0.74)
	Patient 8	AUC score (0.93) Specificity (0.93) Sensitivity (0.78)	**AUC score (0.94)** (p<0.0006) **Specificity (0.97)** (p=0.0001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.97)** (p<0.00001) Specificity (0.94) Sensitivity (0.89)	AUC score (0.95) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.93)** (p<0.00001)

Table E.12: Performance of the RNN model for statistical and frequency features of the trunk sensor (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.96)** (p<0.00001) Specificity (0.98) **Sensitivity (0.81)** (p<0.00001)	AUC score (0.84) Specificity (0.98) Sensitivity (0.69)
	Patient 2	**AUC score (0.98)** (p<0.00001) Specificity (0.96) Sensitivity (0.87)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	**AUC score (0.95)** (p<0.00001) **Specificity (0.94)** (p=0.0008) Sensitivity (0.75)	AUC score (0.87) Specificity (0.93) Sensitivity (0.81) (p<0.00001)
	Patient 5	**AUC score (0.96)** (p<0.00001) Specificity (0.93) Sensitivity (0.82)	AUC score (0.88) Specificity (0.93) Sensitivity (0.83)
	Patient 6	**AUC score (0.97)** (p<0.00001) Specificity (0.98) Sensitivity (0.84)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) Sensitivity (0.84)
	Patient 7	**AUC score (0.95)** (p<0.00001) Specificity (0.98) Sensitivity (0.74)	AUC score (0.86) Specificity (0.98) Sensitivity (0.74)
	Patient 8	AUC score (0.92) Specificity (0.93) Sensitivity (0.76)	**AUC score (0.94)** (p<0.00001) **Specificity (0.97)** (p<0.00001) **Sensitivity (0.92)** (p<0.00001)
	Patient 9	**AUC score (0.99)** (p<0.00001) Specificity (0.97) Sensitivity (0.90)	AUC score (0.95) Specificity (0.97) **Sensitivity (0.93)** (p<0.00001)

Table E.13: Performance of the RNN model for statistical and frequency features from all the sensors (subject dependent method)

		RNN model	***Baseline***
Patient Dependent Method	*Patient 1*	**AUC score (0.95)** (p<0.00001) Specificity (0.95) **Sensitivity (0.88)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.99)** (p<0.00001) Specificity (0.97) **Sensitivity (0.92)** (p<0.00001)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	**AUC score (0.95)** (p<0.00001) Specificity (0.91) Sensitivity (0.83)	AUC score (0.87) **Specificity (0.93)** (p=0.001) Sensitivity (0.81)
	Patient 5	**AUC score (0.96)** (p<0.00001) **Specificity (0.95)** (p<0.00001) Sensitivity (0.85)	AUC score (0.88) Specificity (0.93) Sensitivity (0.83)
	Patient 6	**AUC score (0.98)** (p<0.00001) Specificity (0.97) Sensitivity (0.84)	AUC score (0.92) **Specificity (0.99)** (p<0.00001) Sensitivity (0.84)
	Patient 7	**AUC score (0.99)** (p<0.00001) Specificity (0.97) **Sensitivity (0.92)** (p<0.00001)	AUC score (0.86) Specificity (0.98) Sensitivity (0.74)
	Patient 8	**AUC score (0.96)** (p<0.00001) Specificity (0.96) Sensitivity (0.86)	AUC score (0.94) **Specificity (0.97)** (p=0.0002) Sensitivity (0.92) (p<0.00001)
	Patient 9	**AUC score (0.98)** (p<0.00001) Specificity (0.98) (Sensitivity (0.89)	AUC score (0.95) Specificity (0.97) **Sensitivity (0.938)** (p<0.00001)

Table E.14: Performance of the RNN model for 25 most informative features (subject dependent method)

		RNN model	*Baseline*
Patient Dependent Method	*Patient 1*	**AUC score (0.97)** (p<0.00001) Specificity (0.97) **Sensitivity (0.81)** (p<0.00001)	AUC score (0.84) **Specificity (0.98)** (p<0.00001) Sensitivity (0.69)
	Patient 2	**AUC score (0.98)** (p<0.00001) Specificity (0.97) Sensitivity (0.85)	AUC score (0.93) **Specificity (0.98)** (p<0.00001) Sensitivity (0.87)
	Patient 3	**AUC score (0.95)** (p<0.00001) Specificity (0.90) **Sensitivity (0.86)** (p<0.00001)	AUC score (0.87) **Specificity (0.93)** (p<0.00001) Sensitivity (0.81)
	Patient 5	**AUC score (0.95)** (p<0.00001) Specificity (0.92) Sensitivity (0.81)	AUC score (0.88) **Specificity (0.93)** (p=0.0008) **Sensitivity (0.83)** (p=0.0003)
	Patient 6	**AUC score (0.99)** (p<0.00001) Specificity (0.98) **Sensitivity (0.92)** (p<0.00001)	AUC score (0.92) **Specificity (0.99)** (p=0.002) Sensitivity (0.84)
	Patient 7	**AUC score (0.97)** (p<0.00001) Specificity (0.96) **Sensitivity (0.77)** (p=0.0004)	AUC score (0.86) **Specificity (0.98)** (p<0.00001) Sensitivity (0.74)
	Patient 8	**AUC score (0.98)** (p<0.00001) Specificity (0.94) Sensitivity (0.90)	AUC score (0.94) **Specificity (0.97)** (p<0.00001) Sensitivity (0.92)
	Patient 9	**AUC score (0.99)** (p<0.00001) Specificity (0.97) Sensitivity (0.928)	AUC score (0.95) Specificity (0.97) **Sensitivity (0.938)** (p=0.02)