Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit

Julià Camps[a], Albert Samà[a,i], Mario Martín[b], Daniel Rodríguez-Martín[a], Carlos Pérez-López[a,i], Joan M. Moreno Arostegui[a,i], Joan Cabestany[a,i], Andreu Català[a,i], Sheila Alcaine[g], Berta Mestre[g], Anna Prats[g], Maria C. Crespo-Maraver[g], Timothy J. Counihan[c], Patrick Browne[c], Leo R. Quinlan[d], Gearóid Ó Laighin[d], Dean Sweeney[d], Hadas Lewy[e], Gabriel Vainstein[e], Alberto Costa[f], Roberta Annicchiarico[f], Àngels Bayés[g], Alejandro Rodríguez-Molinero[h,i]

[a] *Technical Research Centre for Dependency Care and Autonomous Living, CETPD, Universitat Politècnica de Catalunya, Barcelona Tech., Rambla de l'Exposició 59-69, Vilanova i la Geltrú 08800, Spain*
[b] *Knowledge Engineering and Machine Learning Group, Universitat Politècnica de Catalunya, Barcelona Tech., C/ Jordi Girona 1-3, Barcelona, 08034, Spain*
[c] *School of Medicine, National University of Ireland Galway (NUIG), Galway, Ireland*
[d] *Electrical and Electronic Engineering Department, National University of Ireland Galway (NUIG), Galway, Ireland*
[e] *Maccabi Healthcare Services, Tel Aviv, Israel*
[f] *IRCCS Fondazione Santa Lucia, Rome, Italy*
[g] *Unidad de Parkinson y trastornos del movimiento (UParkinson), Barcelona, Spain*
[h] *Clinical Research Unit, Consorci Sanitari del Garraf, Vilanova i la Geltr, Spain*
[i] *Sense4Care, Barcelona, Spain*

Abstract

Among Parkinson's disease (PD) motor symptoms, freezing of gait (FOG) may be the most incapacitating. FOG episodes may result in falls and reduce patients' quality of life. Accurate assessment of FOG would provide objective information to neurologists about the patient's condition and the symptom's characteristics, while it could enable non-pharmacologic support based on rhythmic cues.

This paper is, to the best of our knowledge, the first study to propose a deep learning method for detecting FOG episodes in PD patients. This model is trained using a novel spectral data representation strategy which considers

Email address: julcamps@gmail.com (Julià Camps)

information from both the previous and current signal windows. Our approach was evaluated using data collected by a waist-placed inertial measurement unit from 21 PD patients who manifested FOG episodes. These data were also employed to reproduce the state-of-the-art methodologies, which served to perform a comparative study to our FOG monitoring system.

The results of this study demonstrate that our approach successfully outperforms the state-of-the-art methods for automatic FOG detection. Precisely, the deep learning model achieved 90% for the geometric mean between sensitivity and specificity, whereas the state-of-the-art methods were unable to surpass the 83% for the same metric.

Keywords: Deep learning, Signal processing, Freezing of gait, Parkinson's disease, Wearable device

1. Introduction

Parkinson's disease (PD) is a progressive neurological condition, resulting from the degeneration of dopamine-producing neurones. Furthermore, with a prevalence of approximately of 1% among people of age above 65. This condition is the second most common neurodegenerative disorder after Alzheimer's disease [1, 2, 3, 4, 5, 6, 7]. Although there are indicators that relate PD with genetic factors, the cause of it is still unknown [8, 9]. Patients with this disease manifest several motor symptoms such as bradykinesia (slowness of movements), tremor, muscle stiffness, posture alteration and freezing of gait (FOG) [6, 10, 11, 12]. Even though the current drug therapies can successfully mitigate most of these symptoms, FOG episodes may prevail regardless of the medication's effect [13, 14]. On the other hand, duration and frequency of FOG episodes can be effectively reduced using cueing techniques, which imply inducing external cues, such as visual, somatosensory, or rhythmic acoustic ones [15, 16], to patients in order to improve their gait.

FOG is a poorly understood symptom associated with PD [17, 18]. PD patients describe this symptom as if they had their feet glued to the floor.

According to Nieuwboer and Giladi [19], FOG is defined as the inability to deal with concurrent cognitive, limbic and motor inputs, causing an interruption of locomotion. FOG is manifested in episodes usually shorter than 10 seconds (s) [14]. Moreover, manifesting FOG episodes significantly increases the risk of falling during activities related to walking [20]. This symptom has, furthermore, ambient triggers which are commonly found at home, such as walking through a door, and sidestepping obstacles that patients may encounter in their way, such as avoiding a chair in the living room [20, 10]. Consequently, this symptom can affect activities of daily living (ADL); thus, lowering the patient's autonomy and quality of life (QOL) [6, 10]. However, the characteristics of FOG complicate its clinical assessment since its frequency and severity are closely bound to the environment and the activity performed by patients.

Currently, the gold standard for FOG assessment is based on in-lab movement tests and specific symptom questionnaires, namely FOG-questionnaire (FOG-Q) [21, 22, 23, 24]. These strategies can provide biased information on the patient's daily experiences: on the one hand, compared to the patient's home, in-lab tests are performed in artificial environments; thus, patients may show different FOG patterns than those shown by doing ADL. Moreover, patients may be conditioned by other factors, which are well-known side effects from the clinical environment [22, 25, 14]. For instance, being under continuous observation in a controlled environment can have a similar impact than the Hawthorne effect [26]. On the other hand, although the FOG-Q has been proven to provide relevant indicators for the identification and characterisation of FOG [22, 27, 28], questionnaires rely on the subjective perception of the patient, which could be inaccurate [27]. In addition, these evaluations are performed few times per year throughout most of the disease's span. Most patients are assessed by their neurologist once every 3 to 6 months [29]. Therefore, techniques based on continuous observation and evaluation mechanisms are the most suitable to ensure optimal therapy control for PD patients [29]. In fact, accurate and automatic detection of FOG in PD patients has the potential of providing neurologists with relevant indicators about the condition status and

its evolution [30], while enabling to design useful cueing wearable devices since these systems become more effective when applied only during the symptom's episodes [15, 16].

Recent technological advancements have produced light and comfortable devices, which can easily handle data collection and processing [31, 32]. Consequently, many undergoing projects are trying to implementing artificial intelligence algorithms for PD symptom-monitoring tasks inside non-intrusive wearable technologies. These systems aim to be suitable for daily addressing this monitoring task [33, 34, 35, 36, 37, 38]. Within the same trend, state-of-the-art for automatic FOG detection are shallow machine learning (ML) algorithms applied to signals acquired from inertial measurement units (IMU) [39, 40, 41, 35]. These systems are able to reach performances about 85% for the geometric mean between sensitivity and specificity (GM) [38]. However, the complexity of designing handcrafted features and the scarcity of data from PD patients collected under real-life-like conditions for developing reliable solutions are the major impediments preventing the research community from mastering this task.

On the other hand, feature learning is a set of methods that learns a transformation of raw data input to a representation that can be exploited by ML methods. Deep learning (DL) methods are feature learning methods with multiple levels of representation [42]. DL models can easily learn feature extractions from any data, namely multimodal data, missing information and high dimensional feature spaces [42, 43]. Thus, these techniques, as opposed to shallow ML algorithms, are not constrained by the engineering ability for handcrafting features or the complexity of the data representation. Furthermore, DL approaches can outperform shallow ML ones when enough data are available, which may vary depending on the expressiveness of the representation strategies adopted. In fact, DL models have already exhibited a breakthrough in several complex problems, such as image classification [44] and playing games [45]. Therefore, as stated by Eskofier et al. (2016) [46], these techniques seem a promising alternative to the traditional ML ones for IMU-based movement disorder assessment in PD patients.

This paper presents a novel approach for automatic FOG detection based on DL and wearable sensors data, which is able to outperform the state-of-the-art methods reaching performances of 90% for the GM. The data employed were acquired from 21 PD patients who manifested FOG episodes at their homes while performing ADL. This approach consists of both a DL architecture and a data representation strategy. More specifically, the architecture proposed is an eight-layered light one-dimensional (1D) convolutional neural network (ConvNet), and the data representation presented was the spectral window stacking (SWS), which combines information from the current window and its preceding one in the spectral domain. Moreover, the methodologies composing the state-of-the-art for FOG detection were reproduced and fairly evaluated on the same data to be compared to our DL method.

The reproduced feature extractions from other authors were the following: Bächlin et al. (2009) [39], which is hereafter referred as the Moore-Bächlin FOG Algorithm (MBFA); Mazilu et al. (2012) [40], which is an extension of the MBFA; Tripoliti et al. (2013) [41]; and Samà et al. (2017) [38]. However, instead of employing the same classifiers that these authors considered in their work, it was decided to use their features to feed better suited binary classification ML algorithms [47] due to the fact that our data are more complex by being recorded in home environments while performing ADL. The ML algorithms trained were tree bagging [48], adaptive boosting (AdaBoost) [49], adaptive logistic regression boosting (LogitBoost) [50], random undersampling boosting (RUSBoost), robust adaptive boosting (RobustBoost) [51] and support vector machine (SVM) [52]. The results from these models served to objectively compare our proposed DL approach to the state-of-the-art, and, thus, confirm that our method is able to outperform the state-of-the-art by 7% for the GM.

This paper is organised as follows: Section 2 illustrates other studies about automatic FOG detection, which comprise the state-of-the-art for this task, while reviewing the latest applications of DL to the biomedical field. Section 3 introduces the DL theory applied in our approach. Section 4 shows the DL architecture and the data representation techniques composing our approach.

Section 5 describes the experiments and assessment methodologies conducted. Section 6 reports and discusses the experiments' results in detail. Section 7 comments the conclusions and contributions of this work while outlining the next steps to take within our research line.

2. Related work

Automatic FOG detection is still an open research issue despite having been widely addressed by several combinations of devices and algorithms. This section reviews some of these approaches.

Moore et al. in 2008 [53] made the first attempt to automatically detect FOG. They proposed a novel method based on a frequency analysis of the accelerometer signals, from which they defined specific bands related to FOG. More specifically, they implemented a freeze index (FI) threshold, where FI was defined as the ratio between the power spectral density in the gait freezing band (FB) (i.e. 3–8 hertz (Hz)), and in the locomotion band (LB) (i.e. 0.5–3 Hz). This threshold was applied on tri-axial accelerometer signals, which were windowed into splits of 6 s. The data for performing their study were composed of inertial signals recorded by microelectromechanical systems (MEMS) placed at the left shank of 11 PD patients who manifested FOG episodes. Considering the simplicity of their method, it achieved highly accurate results; precisely, it detected 78% of FOG events. However, they performed a poor evaluation of the method due to the fact that it obviated the rate of wrong FOG predictions and the data employed contained only 46 FOG events.

Later in 2009, Bächlin et al. [39] presented the MBFA, which was an extension of the method designed by Moore et al. (2008) [53]. The main changes introduced were the following: adding a power index (PI) (i.e. the power spectral density in the 0.5–8Hz band) threshold to discard standing periods as FOG candidates, and changing the window duration to 4 s in order to reduce latency time of the algorithm. They reported 73.1% and 81.6% for sensitivity and specificity, respectively. These results were, however, computed allowing an

offset margin of 2 s of error for the predictions. They evaluated the MBFA on the Daphnet dataset [54], which was composed by data collected in the Dynamic Analysis of Physiological Networks (DAPHNet) project. Due to its simplicity and acceptable outcomes, the MBFA method has been adopted by the research community as the basic performance reference for automatic FOG detection in PD patients inertial sensors data.

The Daphnet dataset [54] comprises inertial signals data from 10 PD patients, from whom 8 manifested FOG episodes. Specifically, this dataset comprises one hour of tri-axial accelerometer measurements recorded with three accelerometer MEMS, placed on the shank (above the ankle), the thigh (above the knee) and the lower back. The collection was conducted in clinical facilities and under controlled conditions. The protocol required patients to complete three walking tasks: walking back and forth in a straight line and making several turns of 180 degrees; walking in the lab room while performing stops and turns; and, finally, walking to another room and coming back with a cup of water. Even though this dataset was created to design systems for automatic FOG detection, the conditions defining the data collection protocol, such as the limited set of activities and the clinical settings, may overestimate the results of the approaches tested on it compared to performances obtained using data recorded at the patients' homes.

Mazilu et al. in 2012 [40] presented a novel FOG monitoring system based on the work from Bächlin et al. (2009) [39]. Their approach combined the use of smartphones and wearable accelerometers to collect data. Moreover, they employed for the first time ML algorithms to address this detection task. Some of the ML algorithms they tested were: random forests, decision trees, naive Bayes and k-nearest neighbours (k-NN). They reported top results of 66.25% and 95.38% for sensitivity and specificity, respectively, using random forests as classifier algorithm on the Daphnet dataset.

Within the same year, Zhao et al. [55] presented an online threshold-based FOG detection system using accelerometers integrated pants. Their approach was able to achieve 81.7% of sensitivity when testing the system on 8 PD patients

(from which only 6 manifested FOG events); however, they did not report any other performance metric. On the other hand, Handojoseno et al. (2012) [56] implemented a method based on electroencephalography (EEG) data. Specifically, they employed EEG subbands Wavelet Energy and Total Wavelet Entropy features to detect early FOG episodes. In addition, they proposed to complement their system with a sensory cueing to mitigate FOG episodes before they affect the patients movements, preventing them this way from falling. When testing their method on data collected from 26 PD patients who manifested FOG, they were able to report results about 75% for accuracy, sensitivity and specificity.

In 2013, Moore et al. [57] published a comparative study of different configurations of sensors and placements and signal processing parameters for the MBFA. The data for their study were composed of inertial signals from 25 PD patients, which were recorded from 7 sensors in different positions of the body. They demonstrated that best performances were reached using window times from 2.5 s to 5 s, and that, not surprisingly, the best sensors configuration was to employ all 7 sensors simultaneously. Furthermore, they stated that the most convenient configurations for placing only one sensor are the shank and back.

Within the same year, Tripoliti et al. [41] proposed a new system for automatic FOG detection. Their methodology consisted of four stages: data cleaning, filtering, feature extraction and classification. As classifiers, they tested the following ML algorithms: naive Bayes, random forests, decision trees and random trees. The data they employed were collected from 16 people: 5 healthy subjects, 6 PD patients who did not suffer from the FOG symptom, and 5 PD patients who manifested FOG episodes. These data were collected using 6 accelerometers and 2 gyroscopes attached to different body parts of the participants. They achieved 89.3% and 79.15% for sensitivity and specificity, respectively, when considering only the results associated with PD patients who suffered the FOG symptom.

In 2014, Coste et al. [58] presented a pre-FOG detection approach based on a single wireless sensor placed at the lower limbs of the patient. From their

experiments, which were conducted on data from 4 PD patients, they concluded that properly measuring stride length and cadence could enhance FOG detection in PD patients.

Within the same year, Rodríguez et al. [59] proposed to complement the FOG detection approaches with postural assessment to increase the specificity of the methods. In fact, they were able to increase this metric by about 5% due to preventing the system from detecting FOG events when the patient was not in a standing position. These experiments were performed using a single waist-worn tri-axial accelerometer on 20 PD patients who manifested FOG episodes.

In 2015, Zach et al. [60] published a study in which they investigate the sensitivity and specificity of the FI to detect FOG. Specifically, they evaluated this technique using a single accelerometer placed at the lower back, such as in the Daphnet dataset, of 23 PD patients who manifested FOG episodes. However, their evaluation only targeted episodes taking place during full rapid turns and while walking rapidly with short steps. Finally, their results suggested that, as Moore et al. (2013) [57] stated, FOG could be detected from a single lumbar accelerometer; moreover, they reported results of sensitivity and specificity of 75% and 76%, respectively.

During the same period, Mazilu et al. [61] present a study in which they employ electrocardiogram and skin conductance signals, from the CuPiD dataset, in order to predict FOG episodes before they take place. They were able to predict 71.3% of FOG episodes about 4 seconds before they took place (in average). However, they tuned the models for each patient to evaluate them.

Recently, in 2017, Rodríguez et al. [35] presented an approach for FOG detection in PD patients during their ADL. They proposed a set of 55 features in combination with an SVM as the binary classifier to detect FOG episodes. The data was composed of inertial signal recordings at 40 Hz from 21 PD patients who manifested FOG episodes. These data were collected using a single IMU placed at the left side of the waist while the patients performed ADL at their homes. They achieved performance results of 76.8% for the GM using an episode-based evaluation strategy. However, this strategy reduces the specificity

achieved compared to sample based evaluations due to long sitting and standing episodes; consequently, their approach could achieve higher performances using the same evaluation strategy than other authors.

Within the same year, Samà et al. [38] presented a simplification of the approach proposed by Rodríguez et al. (2017) [35], which reduces the number of features to be considered from 55 to 28. Furthermore, they performed a comparative study of the validation performances of their approach and the other ones composing the state-of-the-art for automatic FOG detection. More precisely, they reproduced the feature extraction approaches presented in the following papers: Bächlin et al. (2009) [39], Mazilu et al. (2012) [40], Tripoliti et al. (2013) [41] and Rodríguez et al. (2017) [35]. Later, they trained ML models using these features and compared the 10-fold cross-validation error performance of all of them. From this work, they demonstrated that the 28-feature simplification could reach equivalent performances to the 55-feature former one. In addition, they reported 85.15% for the GM using leave-one-patient-out on data from 15 PD patients.

3. Convolutional neural networks

ConvNets [62] are a type of feed-forward deep neural network (DNN), which typically combine convolutional layers with traditional dense layers to reduce the number of weights composing the model. Convolutional layers enforce local connectivity between neurones of adjacent layers to exploit spatially local correlation. Concretely, convolutional layers are formed by kernels that share weights and, thus, permit to learn position invariant features from the input data. While traditional DL models are composed of stacked dense layers, which lead to an overwhelming number of weights, ConvNets implement a powerful and efficient alternative if the target data present underlying spatial patterns. In fact, the convolutional layers can extract features from data that have local underlying spatial or temporal patterns; moreover, stacking these layers leads to extracting progressively more abstract patterns. Although ConvNets are con-

sidered the most powerful and efficient location invariant feature extractors, the key strategies when training these models are to employ sample normalisation and augmentation techniques [63]. However, when dealing with sequence data, its temporal information may restrict the scope of coherent data augmentation strategies. For example, the meaning of signal data might change if these are rotated or inverted, whereas patterns in images are usually invariant to these operations.

The next subsections review some of the properties of DL models exploited in our approach and some work of ConvNet models for biomedical data problems.

3.1. Characteristics of DL models

3.1.1. Activation functions

DL strategies have a computational bottleneck due to their huge amount of weights to train; thus, the traditional activation functions employed in artificial neural networks, namely the hyperbolic tangent function $f(x) = \tanh(x)$ and the logistic sigmoid function $f(x) = (1 + e^{-x})^{-1}$ have been replaced by the rectified linear unit (ReLU) $f(x) = max(x, 0)$, which is cheaper to compute and allow non upper-bounded output values. Indeed, the ReLU is the most popular activation function for ConvNets [64, 65].

3.1.2. Regularisation for DL models

Regularisation techniques are those that modify a learning algorithm to reduce its generalisation error while preserving training accuracy [43]. There are several regularisation techniques such as the L2-norm regularisation, dropout and early stopping, which are hereafter reviewed.

Parameter norm penalties. This type of techniques aim to limit the representation power of a learning model by adding a penalisation cost $\Omega(\boldsymbol{\theta})$ to the objective function $J(\boldsymbol{Y}_t, \boldsymbol{Y}_p)$ such that the function to be optimised is redefined as

$$\tilde{J}(\boldsymbol{\theta}; \boldsymbol{Y}_t, \boldsymbol{Y}_p) = J(\boldsymbol{Y}_t, \boldsymbol{Y}_p) + \lambda\Omega(\boldsymbol{\theta}) \quad , \tag{1}$$

where $\boldsymbol{\theta}$ represents the model's parameters, which would correspond to DL's weights and/or activations; $\boldsymbol{Y}_t$ are the ground truth labels of the data samples; while $\boldsymbol{Y}_p$ represent the model's predictions; and λ is an hyperparameter that balances the model's learning capacity and its training error.

The most extended parameter norm penalty is the L2 parameter regularisation. This technique penalises the parameters' norm, emphasising on abnormally high values. This technique is defined as

$$\Omega(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 = \frac{1}{2}\sum_i \theta_i^2 \quad . \tag{2}$$

Thus, it will positively reward the model's objective function when the representation responsibility is distributed among all patterns, rather than being concentrated on a subset of them.

Early stopping. DL models are trained iteratively to reach suboptimal, but acceptable, solutions to a problem. These iterations are denoted as epochs. Correctly establishing the number of epochs is important to prevent the model from overfitting to the training data, while avoiding useless computation. Ideally, one would determine the number of epochs that a model will require to converge and set it as a fixed parameter. However, this number is correlated to several factors, some of which are semi-random such as the weights initialisation. Therefore, it is usually convenient to establish convergence criteria to stop the training process [66, 67].

Data augmentation. The best strategy for enhancing ML models' generalisation is to increase the training data available [63]. Data augmentation approaches are those that artificially increment the training data available to maximise the information gain from exploiting it. For example, in the image data context, a picture of a dog will still coherently represent a dog whether it is rotated, inverted, changed of colour or intensity, shifted some pixels to any direction, or cropped; thus, any of these transformations would provide a 'new' sample of a dog to the training set.

Dropout. Dropout is an efficient and effective regularisation method [68, 65]. This method has a similar effect to implementing the bagging ensemble strategy over numerous DL models while preserving inexpensive computational costs. The main intuition behind dropout is to employ only a random subset of the network each time a new instance is fed to the model; thus, the only parameter to be tuned is the probability of dropping a neurone.

3.1.3. Training DL models

Training DL models implies tuning an overwhelming number of hyperparameters [43]. Moreover, DL training procedures are non-deterministic due to stochastic operations that take place during the process such as weights initialisation and dropout. Consequently, the configuration for these models is rarely determined by exhaustive exploration or cross-validation strategies due to the unaffordable overhead of computation resources required by these techniques. DL models are trained using different datasets for the training and testing procedures instead of cross-validation or leave-one-out strategies, while part of the hyperparameters is defined according to previous knowledge or similar studies instead of performing blind explorations.

Optimisation algorithms and the learning rate. DL models implement stochastic gradient-based algorithms to optimise the error loss after each batch is processed. These stochastic training methods perform the optimisation in minibatches; thus, models compute the gradients and apply the weight corrections per minibatch, rather than performing a single update per epoch. The hyperparameter associated to the magnitude of these corrections is referred as learning rate. The optimal value for the learning rate is hard to determine; however, it has been stated that if these corrections are small enough, the model will eventually converge.

Recent optimisation algorithms, such as adaptive gradient algorithm (AdaGrad) [69], implement adaptive learning rate techniques. Algorithms with adaptive learning rates enable using large learning rates to accelerate the training

procedure. Other popular stochastic gradient-based algorithms adopted in DL are the following: stochastic gradient descent [70]; root mean square propagation [71], which extends AdaGrad; AdaDelta [72], which is another extension of AdaGrad; and adaptive momentum (Adam) [73].

3.2. Deep learning for the medical field

A trend of adopting DL techniques to address various tasks has recently arisen in the biomedical research field [74, 75]. In general terms, the biomedical tasks being addressed so far belong to two main categories: problems defined by images, and those involving sequences.

3.2.1. Image data

ConvNets provide very powerful representations on image-data problems [76]. Two recent applications of these models for biomedical research were to diagnose mild cognitive impairment from resting state functional MRI (fMRI) data [77], and to segment 3D biomedical image, such as MRI, fMRI and computed tomography [78]. In fact, ConvNets received great acceptance within this field due to their extraordinary capacity for exploiting image data.

3.2.2. Time series data

The latest advances in DL strategies specially designed to tackle time series data problems, such as Memory Networks [79] and Differential Neural Computers [80], have recently become the state-of-the-art on several complex problems, such as to answer questions about a text or dialogue, to find the shortest path and to infer missing links in graphs [80]. However, while these architectures intend to address reasoning-like tasks, most of the biomedical time-series problems can be solved by simpler classifying or regression models. Moreover, these classification problems can usually be addressed by 1D ConvNets. For example, Shashikumar et al. (2017) [81] present a ConvNet based system for atrial fibrillation detection, which uses pulsatile photoplethysmographic and accelerometer signals recorded from a wrist-worn device.

However, while ConvNets can easily handle patterns in biomedical images, this is not the case for biomedical signals. Biomedical signals data are more complex than image data due to the temporal dependencies between samples, and, thus, the impossibility to freely rotate and distort the information without losing coherence. Moreover, patterns in biomedical signals are rarely invariant to direction, orientation, scale and position. From these differences, it can be noted that while images are extremely friendly for applying techniques such as normalisation and augmentations, signals should be modified carefully to maintain their meaning.

A common technique to overcome the time series disadvantages for training DL models is to represent the time series data as 2D images to train ConvNets. In fact, Pereira et al. published two papers in 2016, [82] and [83], in which they represented inertial data from handwriting dynamics as 2D images to train a PD diagnosis DL model. In [82], the data employed resulted into two datasets of 308 images of size 128 $\times$ 128 pixels, and, in [83], they also had two datasets, one of meanders and another one of spirals, composed by 264 (256 $\times$ 256 pixels) images each. However, in several cases related to PD data the samples to be classified may be too small to build large images, which take the most advantage of the ConvNets. For example, Frid et al. (2016) [84] present a speech based ConvNet approach for PD diagnosis, which is feed with 1D raw speech signal.

Although most studies focus on diagnosis and symptoms monitoring, deep learning has also shown to be able to efficiently reduce the duration of clinical tests as shown in Stamate et al. (2017) [85]. In their work, they present a deep learning based approach which enables to reduce the clinical movement protocol for PD patients from approximately 25 minutes to below 4 minutes. Their approach requires the patients to perform a predefined sequence of iterated movements which are recorded by sensors on a smartphone.

4. DL for FOG detection

4.1. Architecture

The architecture proposed is an eight-layered 1D-ConvNet, which is illustrated in Figure 1. The input layer of this model was fed using 9-channel signals recorded from a waist-worn IMU, which were represented adopting the SWS (see Section 4.3.1) resulting into samples of 64 measures and 18 channels. Next, there is the first convolutional layer, which differs from the successive ones since it fuses the information from the signal's channels using a kernel of shape 3×18. Moreover, this layer, as the remaining convolutional ones, implements 16 kernels with the stride equal to 1 and without padding. The subsequent three convolutional layers had the same settings as their predecessor, except for the kernels' shapes, which were set to 3. After the fourth convolutional layer, there are two fully-connected dense layers, each of which has 32 neurones connected to all the input and output cells from itself. Finally, there is the output layer, which is implemented by a single fully-connected neurone. However, whereas all previous activation functions of the model were set to ReLUs, the one associated with this classification neurone was set to the linear activation function as activation.

We consider that 1D-ConvNets are the most suitable approach to detect FOG, concerning other approaches that mainly are 2D-ConvNets due to the specific characteristics of the symptom. On the one hand, FOG episodes may last less than 1 s and typically have a duration below 2 s. On the other hand, 2D-ConvNet require constructing images by concatenating the frequency content of several windows. However, given that FOG lasts few seconds, these windows would have a very short duration and images would have an extremely reduced size, which is not suitable for ConvNets. On the other hand, 1D-ConvNet allows to automatically learn the frequency characteristics along window sizes above 2 s, which we consider that contains enough significant patterns to solve the classification task.

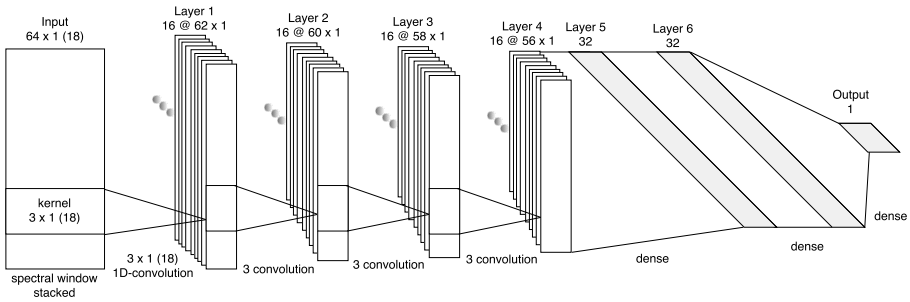

Figure 1: Diagram of the 1D-ConvNet's architecture.

4.2. Error loss

Since FOG detection is an imbalanced binary classification task, the weighted hinge loss (5) algorithm was chosen as the error loss function.

$$\boldsymbol{l}_b^+ = max(0, 1 - \boldsymbol{y}_{true}^+ * \boldsymbol{y}_{pred}) * (1 - \rho) \tag{3}$$

$$\boldsymbol{l}_b^- = max(0, 1 - \boldsymbol{y}_{true}^- * \boldsymbol{y}_{pred}) * \rho \tag{4}$$

$$l_w = mean(\boldsymbol{l}_b^+ * w^{-1} + \boldsymbol{l}_b^- * w) \quad , \tag{5}$$

where $\boldsymbol{y}_{true}^+$ and $\boldsymbol{y}_{true}^-$ are the true positive (i.e. 1) and negative (i.e. -1) labels of the data, respectively; $\boldsymbol{y}_{pred}$ are the model's predictions; ρ is the prior of the FOG class in the Training-train data samples; thus, $1 - \rho$ is the prior of the non-FOG one; and w ($1 < w < \frac{(1-\rho)}{\rho}$) is a weighting coefficient that allows to adjust the balancing factor.

4.3. Data representation

4.3.1. Spectral window stacking

The data representation strategy proposed for training the DL models is the SWS. This strategy consists in taking the information of two consecutive windows from the 9-channel signal data recorded by the three tri-axial sensors (i.e. accelerometer, gyroscope and magnetometer) and joining them in the spectral domain. More precisely, the already preprocessed data was windowed using a window size of 2.56 s since sensor measurements employed were acquired at 50 Hz, according to Moore et al. (2013) [57], window sizes for FOG detection

should be at least of 2.5 s to achieve accurate results, and GPUs are more efficient for input sizes equals to powers of two. Thus, a window was composed of 128 nine-channel instances. Then, the SWS method was applied to these windows.

The SWS function takes two arguments W_t and W_{t-1}, which refer to the window to be analysed at time t and its previous one, respectively. The function's process, a diagram of which is illustrated in Figure 2, is composed of the following steps: first, the fast Fourier transform (FFT) is computed for both windows; next, only the first symmetric half of each window is kept; and, finally, both windows are stacked alongside each other forming a single sample. The shape of the resulting samples corresponds to half the original window's length and twice its width (i.e. 64×18).

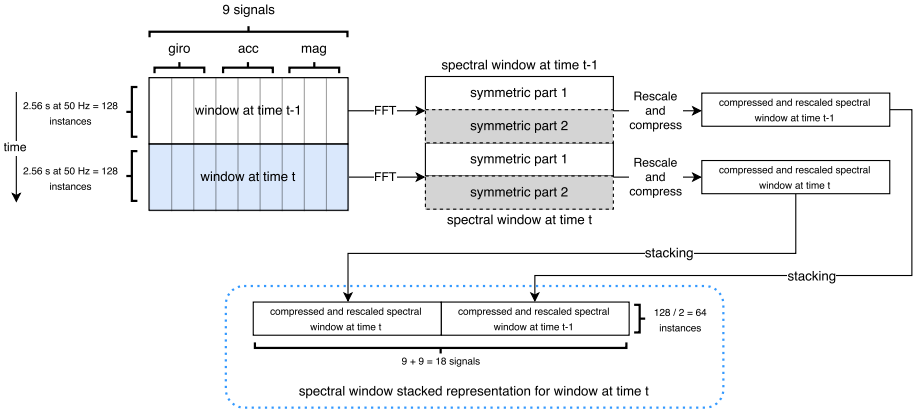

Figure 2: Spectral window stacking process diagram.

4.3.2. Data augmentation

The proposed data augmentation strategy stochastically quadrupled the Training-train dataset differently for each epoch. More precisely, this technique implied that a subset of the samples in the dataset was transformed into reasonably coherent versions of themselves with a certain probability. In addition, this probability was learned during the hyperparameters tuning process. The presence of these modified instances introduced stochastic noise in the training

process, and, thus, prevented the model from overfitting.

This method was composed of two consecutive operations: shifting and rotating. More precisely, the starting measure of each file in the data was randomly shifted within the range of one window's size to have the same data in different parts of the samples. Next, the windows of all files were generated accordingly. Later, for each window before performing the SWS, it was decided whether to use it unchanged or transformed using a rotation matrix with probability 0.5. Moreover, these matrices implemented rotation angles which simulated variations of the IMU placement due to inter-patients waist form differences or device misplacement. The implementation of these matrices was

$$\boldsymbol{R} = R_z(\alpha)R_y(\beta)R_x(\gamma) \quad , \tag{6}$$

where $\alpha \in [-30, 30]$ was generated through a normal distribution $N \sim (0, 10)$, $\beta \in [-10, 10]$ with $N \sim (0, 15)$ and $\gamma \in [-10, 10]$ with $N \sim (0, 2.5)$ (see Figure 3 to view these axes associated with a patient's IMU placement).

5. Experiments

5.1. Data collection

The data employed were composed of inertial signals from 21 PD patients. The clinical characteristics of the sample are shown in Table 1. The inertial data were recorded using a single IMU with three tri-axial sensors: accelerometer, gyroscope and magnetometer. The IMU was of size 99 × 53 × 19 mm^3 and weighed 78 g. It was placed on the left side of the patient's waist, as shown in Figure 3 [86], to collect the data. Admittedly, the state-of-the-art placement for sensing FOG episodes includes the shank, whereas the dataset considered only contained data collected from the waist. However, according to Moore et al. [57] the lower back achieves similar performances to the shank in this task. Furthermore, other symptoms, such as full body dyskinesias, are easier to monitor from waist than the shank. Thus, successfully detecting FOG from the waist enables the possibility of creating PD multi-symptom detection systems,

which was one of the goals of the Remote and Autonomous Management of Parkinson's Disease (REMPARK) [87] project, where these data were collected.

Table 1: Clinical characteristics of the participants.

	Mean	Std. dev.
Age	69.29	9.72
Years since diagnosis	11.3	3.96

	Median	Interquartile range
UPDRS III - OFF	37	26.5
UPDRS III - ON	14.5	12
Mini Mental	28.5	3
H&Y	3	0
FOG-Q	16	6.5

Figure 3: The data collector IMU and its location on patients' body (i.e. left side of their waist).

The REMPARK's experimental protocol was designed by medical experts, complied with the ethical approval and was video-recorded using an HD camera from a Google Nexus S smartphone. This protocol included data from 92 PD patients. The protocol's inclusion criteria were being diagnosed with PD according to the UK Brain Bank [88]; having Hoehn & Yahr stage above 2 in OFF-state; not having dementia according to DSM-IV criterion; and giving written informed consent. From these dataset 21 PD patients' (i.e. 3 women and 18 men) data were selected for the Improving Quality of Life with an Au-

tomatic Control System (MASPARK) project, which aimed to analyse FOG in PD patients. The selection criteria for these patients was having at least one minute of FOG labelled data and having reached a score of at least 6 in the FOG-Q.

The REMPARK's collection protocol was executed by four medical institutions: Teknon Barcelona (Barcelona, Spain), Fondazione Santa Lucia (Rome, Italy), National University of Ireland Galway (Galway, Ireland), and Maccabi Healthcare Services (Tel Aviv, Israel). This protocol was completed at the home of each patient, while these participants performed several ADL wearing the waist-placed IMU. The collection trials were structured in two parts: the first part was performed when the patients were in OFF-state; the same tests were repeated in ON-state. The researchers visited patients at home at the period they usually were in the OFF-state, and occasionally it was facilitated by reducing or skipping the previous dopaminergic medication dose. Once the OFF-state was confirmed by the patient and the researchers, the inertial sensor was placed on the patients waist, and the protocol was followed by the patient. Once finished, the patient took his medication and researchers waited until the patient entered the On phase. Then, the protocol was followed again with the waist inertial sensor. The protocol included the following activities: walking in the apartment while showing every room of it to the researchers as if they were interested in selling it, walking ten meters outdoors, and a specific test to elicit FOG episodes. This test included at least 3 repetitions of the following sequence of actions: standing up from a chair or sofa, walking 6 meters, turning 180 degrees, walking the 6 meters back to the chair or sofa, and sitting down on it. In addition to this, other special activities were recorded to increase the complexity of detecting FOG: cleaning a cup, carrying a glass of water, typing in a computer, brushing one's teeth, and drawing and erasing on a sheet of paper. During the data recording, many non-scripted situations arose which introduced variability on the collection trials. For example, some patients had to answer a phone call or avoid colliding with their pets; furthermore, some of these events triggered FOG episodes. All the data were labelled by clinicians participating in

the REMPARK project, using the strategy presented by Samà et al. (2013) [87] to synchronise the video with the IMU measurements. Finally, all the data were relabelled relying only on the video recordings using a specific labelling software to ensure that all episodes were correctly labelled in the videos. This task was performed by clinicians from UParkinson within the MASPARK project.

5.2. Data preprocessing

The data were composed by 18.64 hours of 9-channel signal data sampled at 200 Hz recorded from 21 PD patients. These raw data were processed as follows. Firstly, missing values due to sensor errors and unlabelled values (i.e. measures taken while the video was not video recording the patient) were discarded. Secondly, the data was downsampled to 50 Hz since gait patterns commonly have associated frequencies lower than 20 Hz. Thirdly, signals were filtered using an eighth-order low-pass filter with cut-off frequency set to 20 Hz. Fourthly, the Training-train, Training-validation and Testing datasets were generated by randomly splitting total data into three sets while fulfilling these restrictions: every dataset should contain at least one patient from each medical institution participating in the data collection protocol; every dataset should contain at least one patient of each gender; the relative difference between FOG percentages among any pair of datasets should be less than 50% (e.g. if a set had 15% of FOG instances, others should at least 7.5% and at most 22.5% of it); and the number of patients included in the Training-train dataset should be maximised. This splitting process concluded with 13, 4 and 4 patients for the Training-train, Training-validation and Testing sets, respectively. The properties of these datasets are shown in Table 2. Next, each data channel was normalised using its sample standard deviation (ST) computed from the entire Training data (i.e. Training-train plus Training-validation). Lastly, these data were windowed into 128-instance windows (see Subsection 4.3), and these windows were labelled as FOG only if at least 50% of their measures were labelled as FOG, whereas non-FOG instances had to be purely composed by non-FOG measurements. Note that all samples containing some but less than 50% of FOG measurements were

discarded. Consequently, the final number of available samples were 15568, 5040 and 5312 for Training-train, Training-validating and Testing, respectively. Which, reinforced the need for data augmentation strategies since DL strategies usually require huge datasets.

Table 2: Data properties per dataset and patient in it.

Dataset	# Patients	# Instances	FOG %
Training-train	13	2230800	16.27
Training-validation	4	818400	18.04
Testing	4	844800	13.04

5.3. DL training settings

The hyperparameters were chosen conducting several suboptimal explorations during which only subsets of hyperparameters were tuned at a time. During this process, the following hyper-parameters were tuned: learning rate, number of convolutional and fully connected layers, weight decay values, loss function, and early stopping. The resulting configuration of this process is hereafter described.

The models were trained via backpropagation using the following configuration: the weight initialisation method was set to 'Xavier initialisation' [89]. The Optimisation method was set to Adam. The learning rate was set to $5 \cdot 10^{-5}$. Gradient clipping was implemented with clip value set to 1. The w parameter of Equation (5) was set to 1.5. The batch size of the minibatch training strategy was set to 16. The maximum number of epochs was set to 3000, and early stopping was implemented using the minimum between the Training-train and the Training-validation GMs as the observed metric; in other words, the training was ended when this metric did not increase for 500 contiguous epochs. The numbers of random shifts and rotations per epoch and file were set to 2, each; moreover, the probabilities of randomly shifting a window and multiplying it by a rotation matrix were set to 1 and 0.5, respectively. The L2-norm weight regularisation with the penalty parameter set to 10^{-5} was implemented in all

layers except the output one, which had the penalty parameter set to 10^{-2}. The dropout dropping rate was set to 0.5 in the convolutional layers, and to 0.25 in the fully-connected ones.

Although our proposed representation was the SWS, three simpler strategies were implemented and compared to this one before proceeding to test our approach. More precisely, the representation methodologies compared were the following: single window in the temporal domain (TW); single window in the spectral domain (SW); two windows concatenated in the temporal domain (TWC); and the SWS. The results of these experiments are shown in Table 3 which contains the best training models out of 50 runs for each configuration. The procedure for choosing the models to be compared followed the same strategy that the one presented in Subsection 5.5, but considering only 50 runs from which at least 20 models should achieve selection metrics above 50%. From Table 3 it can be observed that the SWS undoubtedly outperformed all alternative representations considered in this task; the SWS was, therefore, implemented in the final DL model.

Table 3: Validation performances for each representation.

	Training-train			Training-validation		
Data	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
TW	78.4	87.0	76.1	75.1	71.3	78.2
SW	90.8	91.5	90.3	85.0	90.5	86.0
TWC	81.5	90.2	79.5	76.8	82.3	78.4
SWS	**93.9**	**94.5**	**93.5**	**87.9**	**92.6**	**88.7**

5.4. Reproduction of the state-of-the-art

5.4.1. Data representation and preprocessing

According to the comparative study presented by Samà et al. (2017) [38], all the approaches selected for being reproduced had high performances when

using window sizes of 3.2 s with overlapping of 50%, sensors sampling frequency equals to 40 Hz and being filtered using a second-order low-pass filter with cut-off frequency set to 15 Hz. Although this strategy slightly differs regarding window size and sampling frequency from our DL approach, since FOG frequencies are still entirely represented by using 40 Hz and the window size is almost the same, both representations are considered as equivalent. Therefore, the recommendations of this previous work were adopted, and the data were processed accordingly to ensure having a fair reproduction of other authors' methods. The window labelling was implemented as in our DL approach.

5.4.2. Reproduced feature extractions

The following features were implemented to reproduce the existing state-of-the-art methods and compare them to our DL approach:

MBFA presented by Bächlin et al. (2009) [39]. This feature extraction was applied to each window using accelerometer data, from which computed the FI and the PI.

Online FOG detection presented by Mazilu et al. (2012) [40]. This feature extraction extends the previous approach by adding the following characteristics computed from the acceleration data in the current window: the mean, ST and variance (VAR) of each axis; the entropy of the amplitudes of the Y-axis; and the energy of each axis.

Four-stage FOG detection presented by Tripoliti et al. (2013) [41]. This feature extraction was applied for each window, consisting in the entropy of each axis for the accelerometer and gyroscope signals.

FOG detection for home environments presented by Samà et al. (2017) [38]. This feature extraction was applied for each window in the accelerometer data using characteristics computed from itself and its preceding one. The features composing it are the following: the mean and ST for each axis from the current window; the difference between each pair of axes' means; the difference between

the means of the current window and the ones of the previous one; the skewness of each axis from the current window; the skewness of the current acceleration's magnitude in the spectral domain; the skewness of the LB amplitudes, the same for the FB ones, and for the combination of both amplitude ranges; the ST of the amplitudes in the posture transition band (i.e. 0.1–0.68 Hz), the FB, the LB; and from those above the LB, namely from 8 Hz to half the sampling frequency; the frequency of the centre of mass; the correlation coefficient between each pair of axes; the frequencies corresponding to the first and second amplitude peaks in the Y-axis; and the first three components from the principal components analysis of the spectral representation of the Y-axis.

5.4.3. ML algorithms implemented

According to Caruana et al. (2006) [47], tree bagging [48], AdaBoost [49], LogitBoost [50], RUSBoost [90], RobustBoost [51] and SVM [52] are powerful algorithms for solving two-class classification problems. These were therefore implemented in order to reproduce the state-of-the-art for automatic FOG detection.

Tree bagging [48]. This algorithm trains an ensemble of decision trees using subsets from the training data, which are generated by sampling with replacement as many instances as the number of samples in the training dataset. Then, the predictions are performed by majority voting from all the trees in the ensemble.

AdaBoost [49]. This algorithm trains an ensemble of decision trees sequentially, such that the new trees aggregated to the ensemble are focused on the previously misclassified samples. The prediction is performed as a weighted average on over all the predictions of the trees in the ensemble.

LogitBoost [50]. This algorithm extends AdaBoost by reducing the weight assigned to badly misclassified samples to achieve higher performances in poorly separable data.

RUSBoost [90]. This algorithm extends AdaBoost by training the learners with class balanced subsets of the training data to achieve higher performances in class-imbalanced problems.

RobustBoost [51]. The traditional AdaBoost focuses each iteration on classifying previously misclassified samples. However, this strategy may lower the average accuracy of the classifier if there are incorrect labels in the data. The RobustBoost algorithm extends AdaBoost by maximising the number of undoubtedly well-classified samples, namely above a certain threshold, rather than minimising the models' train error.

SVM [52]. SVMs are a powerful shallow ML algorithm, specially in binary classification tasks. The implemented version of SVM was the one with the radial basis function (RBF) kernel [91], namely SVM-RBF.

5.4.4. Shallow ML training settings

The shallow ML algorithms implemented were tuned by performing 15 repetitions of the hyperparameters explorations using leave-one-patient-out cross-validation over the patients in the Training dataset. The configurations considered during the training of these algorithms are hereafter outlined. The ensemble ML algorithms were all trained using the following types of decision trees as weak classifiers: decision stumps, trees with minimum leaf size set to 3, trees with minimum leaf size set to 5% of the training data, and trees with minimum leaf size set to 10% of the training data. The numbers of weak classifiers considered were 128, 256, 512 and 1024. From these algorithms, some of the boosting ones allowed to set a learning rate parameter; precisely, AdaBoost, LogitBoost and RUSBoost considered learning rates of 0.1, 0.5 and 1. On the other hand, RobustBoost permits to tune an error goal parameter, which determines an error tolerance to avoid overfitting on partially mislabelled data. This parameter was tested using 0.01, 0.05 and 0.1, where these values refer to the rates of misclassification tolerance. The SVM-RBF considered all combinations of the following values for both cost factor λ and the parameter of the RBF

kernel γ: 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2} and 10^{3}. However, while other methods were trained 15 times, the SVMs were trained only once per configuration since their training process is deterministic.

From these hyperparameters explorations, the best validation configurations were chosen to be later compared to our DL approach on the same 4 test PD patients' data, which are presented in Table 4. The number of weak classifiers and the learning rate were fixed to 1024 and 0.1, respectively, in all models that had these parameters. The other hyperparameters optimal configuration differed depending on the feature extraction and the algorithm. The table presents the number of weak learners used by the ensemble ML algorithms, which is denoted as 'Tree-type'. The minimum and the maximum number of leafs is presented as an absolute value by 'min_x' and 'max_x', respectively, or as a relative value by 'min_x%' and 'max_x%', where X is the number of leafs.

Table 4: Best validation shallow ML configurations.

Algorithms	Parameters	Bächlin	Mazilu	Tripoliti	Samà
Tree bagging	Tree-type	min_3	min_5%	min_5%	min_5%
AdaBoost	Tree-type	min_3	min_3	min_3	min_10%
LogitBoost	Tree-type	min_3	min_5%	min_5%	min_5%
RUSBoost	Tree-type	max_1	max_1	max_1	max_1
RobustBoost	Tree-type	min_3	min_5%	min_3	min_10%
	Error goal	0.05	0.1	0.05	0.05
SVM-RBF	λ	10^3	10^2	10^3	10
	γ	10^3	10^{-3}	1	10^{-1}

5.5. Evaluation

The evaluation of the DL and the shallow ML methods was performed following the same rules: first, each metric was computed independently per patient, and later all patients' metrics were averaged together. This strategy implied

that the GM metric, which is our main performance reference, should be high in all patients to reach high values. In other words, if a model classified all data from a patient as FOG or non-FOG, rather than discerning between both classes, this model would receive 0% in the GM for that patient, while the 100% in sensitivity or specificity would be ignored. Consequently, this assessment approach forced models to learn patterns for both FOG and non-FOG data that were effective in all patients.

The optimal DL model used for testing a selection metric which was the minimum between Training-train and Training-validation GMs of sensitivity and specificity and the following steps: Firstly, 100 runs were performed for each configuration. Secondly, these configurations were only further considered if at least 40 models reached values of the selection metric above 50%. After that, all models that trained for less than 300 epochs were discarded. Later, only configurations which had less than 10% difference between their top-5 models regarding selection metric were maintained. Next, the best model (in terms of selection metric) of each of the remaining configurations was chosen to represent them. Finally, these candidate models were sorted by selection metric and the best one was chosen for testing, while the remaining ones forming the top-5 were employed to ensure that our study was reproducible. In fact, all top-5 models were able to outperform the state-of-the-art approaches reproduced in this work, while achieving testing performances higher than 88.5%.

Regarding the shallow ML models, the reported configurations were chosen by running 15 times the hyperparameters search first, using the same 17 Training patients that in the DL experiments, and leave-one-patient-out cross-validation. Next, the best validation configuration for each pair of shallow algorithm and feature extraction was retrained on the entire training data and tested on the same remaining patients than our DL approach.

5.6. Tools and technologies

The experiments in this work were executed using three PCs, each of which had the following features: the Intel Core i7-7700 (8 cores at 3.60 GHz) as CPU;

and 16 GB DDR3 of memory; the GTX 1060 (6GB DDR5) as GPU. The code for training the DL models implemented was written in Python (version 3.4), using Keras library (version 1.2.2) [92] running on top of TensorFlow (version tensorflow-gpu 1.0.1) [93]. Furthermore, the code implemented for the training and processing algorithms is publicly available at [94]. The reproduction of the state-of-the-art was coded and run using Matlab (version R2017a), precisely, its Statistics and Machine Learning Toolbox.

6. Results and discussion

6.1. Shallow ML results

Table 5 presents the testing results obtained from reproducing the state-of-the-art approaches and training shallow ML algorithms using the same Testing dataset than for the DL models' evaluation. Testing results presented were obtained from the optimal model found through the validation performances. From it, it can be observed that the best results from the approaches reproduced were achieved by the approach proposed by Samà et al. (2017) [38]. Specifically, this approach achieved several GM performances above 80% and a top performance of 83% when combined with the SVM-RBF. On the other hand, the algorithms trained on the features proposed by Tripoliti et al. (2013) [41] achieved the poorest results among all.

Table 6 presents the comparison between our reproduction of the state-of-the-art and itself but on the literature. The performances reported by the original authors in their studies were compared to our reproduction of these methods to assess the correctness of the replications. More precisely, each feature extraction was paired to its best performing classifier to form the reproduction of the other authors' approach. This was done to verify the correctness of the reproduction in our data.

From Table 6 it can be observed that the best performances of the methods reproduced differ in some cases from the ones reported in the literature. However, these differences can be justified by the following observations. Bächlin et

Table 5: Testing results from the state-of-the-art reproduction.

Features	Algorithm	Accuracy	Sensitivity	Specificity	GM
Bächlin et al. (2009)	Tree bagging	82.90	30.00	91.89	52.50
	AdaBoost	80.56	36.25	88.10	56.51
	LogitBoost	80.46	34.64	88.25	55.29
	RUSBoost	66.05	96.96	60.80	76.78
	RobutBoost	79.60	36.96	86.85	56.66
	SVM-RBF	62.57	94.46	57.15	73.48
Mazilu et al. (2012)	Tree bagging	83.49	64.29	86.76	74.68
	AdaBoost	81.88	59.82	85.64	71.57
	LogitBoost	80.28	68.75	82.24	75.19
	RUSBoost	66.52	98.04	61.16	77.43
	RobutBoost	77.32	62.86	79.78	70.81
	SVM-RBF	64.81	97.50	59.25	76.00
Tripoliti et al. (2013)	Tree bagging	83.57	12.68	95.63	34.82
	AdaBoost	82.01	18.75	92.77	41.71
	LogitBoost	82.92	16.79	94.17	39.76
	RUSBoost	59.62	96.07	53.42	71.64
	RobutBoost	78.61	37.32	85.64	56.53
	SVM-RBF	59.77	97.50	53.36	72.13
Samà et al. (2017)	Tree bagging	88.99	51.89	95.35	70.34
	AdaBoost	85.36	30.16	94.83	53.48
	LogitBoost	85.18	77.20	86.55	81.74
	RUSBoost	72.04	96.95	67.76	81.05
	RobutBoost	85.52	59.25	90.02	73.03
	SVM-RBF	75.19	96.23	71.58	**83.00**

al. (2009) [39] evaluated their approach using error tolerance between windows misclassified, considering that events were correctly classified if the correct label was predicted from 2 s before to 2 s after the event. Besides, their approach was designed using the Daphnet dataset, which as mentioned in Section 2 contains simpler activities than ours. Mazilu et al. (2012) [40] also employed the Daphnet dataset. Tripoliti et al. (2013) [41] employed a collection protocol similar to for the Daphnet dataset using 5 relevant patients, which also justifies the lower performance obtained in our data. Samà et al. (2017) [38] performed leave-one-patient-out cross-validation using 15 PD patients' data, rather than testing over 4 unseen patients as in this study, which has lead to slight differences in terms of performance. Therefore, it can be stated that the reproduction of the state-of-the-art was successful, and, thus, the results reported in Table 5 fairly represent the top existing solutions for automatic FOG detection.

Table 6: Comparison of the reproduction and the literature of the state-of-the-art.

	Data representation	Model	GM
Literature	Bächlin et al. (2009) [39]	Thresholds	77.23
	Mazilu et al. (2012) [40]	Random forests	79.49
	Tripoliti et al. (2013) [41]	Random forests	84.07
	Samà et al. (2017) [38]	SVM-RBF	85.15
Reproduction	Bächlin et al. (2009)	RUSBoost	76.78
	Mazilu et al. (2012)	RUSBoost	77.43
	Tripoliti et al. (2013)	SVM-RBF	72.13
	Samà et al. (2017)	SVM-RBF	**83.00**

6.2. DL results

The 1D-ConvNet selected, which was composed by 37121 parameters, trained for 456 epochs achieving a Training-validation GM of 90.2%. Table 7 illustrates the results of this model in all three datasets, which were consistent with the remaining candidate models from the top-5 (these results are not included in this table). Testing results presented were obtained from the optimal model found through the selection metric performances, as described in Subsection 5.5. Moreover, from these results, it can be observed that the DL approach outperformed the state-of-the-art methodologies for automatic FOG detection, which had testing GM performances of 83% (see Table 5).

Table 7: Performance of our 1D-ConvNet.

Dataset	Accuracy	Sensitivity	Specificity	GM
Training	93.9	94.5	93.5	93.9
Validation	87.9	92.6	88.7	90.2
Testing	89.0	91.9	89.5	**90.6**

DL models are usually hard to train; they require large amounts of resources concerning both time and computation power. Moreover, the training process

may be even non-deterministic. Therefore, generating adjustable models is an interesting feature to consider since it would allow the final user to modify the performance ratio of the system in order to match his preferences, without, of course, the overhead of retraining a slightly different model. From Figure 4, which shows the receiver operating characteristic curve (ROC Curve) of our 1D-ConvNet approach, it can be observed that this model can be regulated to numerous well-performing configurations. Furthermore, the GM of several of the thresholded models remained above 90%, and the area under the curve (AUC) was equal to 0.88.

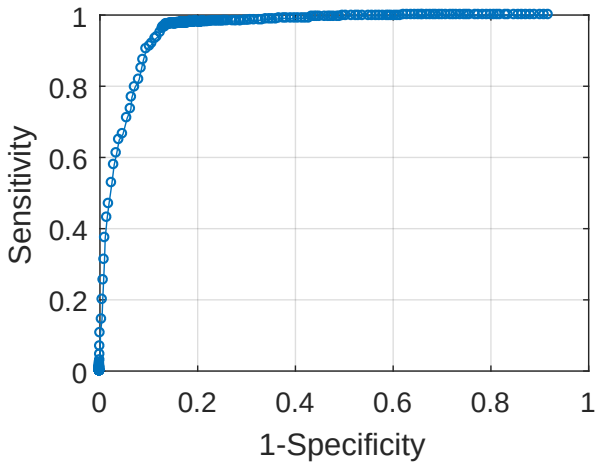

Figure 4: ROC Curve of the 1D-ConvNet model (AUC = 0.88).

Table 8 summarises the results reported from our DL experiments together with the best results from Table 5 to compare the results from our approach to the best possible methods in the state-of-the-art. This table confirms that our approach significantly outperforms all existing strategies for automatic FOG detection.

6.3. Limitations

Results achieved by DL models are clearly above the state-of-the-art; however, the real-time implementation of our approach would require more compu-

Table 8: Comparison to the state-of-the-art.

Data representation	Model	Accuracy	Sensitivity	Specificity	GM
SWS	1D-ConvNet	89.0	91.9	89.5	**90.6**
Bächlin et al. (2009)	RUSBoost	66.05	96.96	60.80	76.78
Mazilu et al. (2012)	RUSBoost	66.52	98.04	61.16	77.43
Tripoliti et al. (2013)	SVM-RBF	59.77	97.50	53.36	72.13
Samà et al. (2017)	SVM-RBF	75.19	96.23	71.58	83.00

tational resources than other existing alternatives such as SVM based systems [38]. More precisely, classifying one sample in our ConvNet involves 37121 parameters, whereas the SVM-RBF model presented by Samà et al. [38] considers 27776 parameters, which is about 25% less than the DL strategy.

The DL models were trained using 3 tri-axial signals data; while most of the previous approaches only used accelerometers or, in some cases, accelerometers and gyroscopes. Thus, this method requires being implemented into a complete 9-axis IMU instead of a single accelerometer. However, 9-axis sensors, such as the LSM9DS1 [95], are currently quite extended.

The dataset comprised 21 PD patients' data. This volume of participants may be insufficient to capture and accurately represent Parkinson's disease inter-patients variability since this condition affects each patient differently. Consequently, the proposed approach should be further validated in larger datasets.

7. Conclusions

This paper is, to the best of our knowledge, the first study to present a method for FOG detection on home environments based on DL techniques. More precisely, the DL model presented is feed-forward 1D-ConvNet, which achieved performances about 90% for the GM. Moreover, by achieving these results, our approach was able to outperform the state-of-the-art methodologies for this task, which was double checked by accurately reproducing these

methodologies and testing them on the same data that the DL models.

Our approach, which significantly improves the current performance of existing automatic FOG detection methods, may serve to improve the medical monitoring of FOG's evolution in PD patients; thus, allowing neurologists to better understand this symptom. Moreover, it may allow clinicians to objectively evaluate the effect of drugs (e.g. during clinical trials) over the symptom's characteristics from automatically gathered indicators.

Future work. Interesting extensions of our work could be to assess the proposed methods on more extensive datasets and to explore alternative DL architectures such as the long-short-term-memory [96] and the gated-recurrent-unit [97, 98], which are specially thought for working on time-series data. Additionally, it would be interesting to implement and evaluate our approach on real-time. This implementation seems feasible since it requires 145 KB to store its 37121 parameters, and, thus, could fit into a microcontroller's memory (e.g. the STM32F415RG [99] used in Rodríguez et al. (2017) [36] has 1 megabyte of flash memory).

Acknowledgements

Part of this project has been performed within the framework of the Freezing in Parkinson's Disease: MASPARK [100] project which is funded by La Fundació La Marató de TV3 20140431. This work also forms part of the framework of the FP7 REMPARK [101] project ICT-287677, which is funded by the European Community. The authors, thus, would like to acknowledge the contributions of the members from MASPARK and REMPARK consortium.

References

[1] C. M. Tanner, S. M. Goldman, Epidemiology of parkinson's disease, Neurologic clinics 14 (2) (1996) 317–335.

[2] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, et al., Mutation in the α-synuclein gene identified in families with parkinson's disease, science 276 (5321) (1997) 2045–2047.

[3] R. L. Nussbaum, C. E. Ellis, Alzheimer's disease and parkinson's disease, New England journal of medicine 348 (14) (2003) 1356–1364.

[4] L. M. De Lau, M. M. Breteler, Epidemiology of parkinson's disease, Lancet Neurology 5 (6) (2006) 525–535.

[5] W. H. Organization, Neurological disorders: public health challenges, World Health Organization, 2006.

[6] B. Post, M. P. Merkus, R. J. De Haan, J. D. Speelman, Prognostic factors for the progression of parkinson's disease: a systematic review, Movement disorders 22 (13) (2007) 1839–1851.

[7] T. Pringsheim, N. Jette, A. Frolkis, T. D. Steeves, The prevalence of parkinson's disease: A systematic review and meta-analysis, Movement disorders 29 (13) (2014) 1583–1590.

[8] G. W. Ross, H. Petrovitch, R. D. Abbott, J. Nelson, W. Markesbery, D. Davis, J. Hardman, L. Launer, K. Masaki, C. M. Tanner, et al., Parkinsonian signs and substantia nigra neuron density in decendents elders without pd, Annals of neurology 56 (4) (2004) 532–539.

[9] L. V. Kalia, A. E. Lang, Parkinson disease in 2015: Evolving basic, pathological and clinical concepts in pd, Nature reviews Neurology.

[10] O. Moore, C. Peretz, N. Giladi, Freezing of gait affects quality of life of peoples with parkinson's disease beyond its relationships with mobility and gait, Movement disorders 22 (15) (2007) 2192–2195.

[11] J. Jankovic, Parkinsons disease: clinical features and diagnosis, Journal of Neurology, Neurosurgery & Psychiatry 79 (4) (2008) 368–376.

[12] J. Marusiak, K. Kisiel-Sajewicz, A. Jaskólska, A. Jaskólski, Higher muscle passive stiffness in parkinson's disease patients than in controls measured by myotonometry, Archives of physical medicine and rehabilitation 91 (5) (2010) 800–802.

[13] N. Giladi, M. McDermott, S. Fahn, S. Przedborski, J. Jankovic, M. Stern, C. Tanner, P. S. Group, et al., Freezing of gait in pd prospective assessment in the datatop cohort, Neurology 56 (12) (2001) 1712–1721.

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

[15] P. Arias, J. Cudeiro, Effect of rhythmic auditory stimulation on gait in parkinsonian patients with and without freezing of gait, PloS one 5 (3) (2010) e9675.

[16] W. R. Young, L. Shreve, E. J. Quinn, C. Craig, H. Bronte-Stewart, Auditory cueing in parkinson's patients with freezing of gait. what matters most: Action-relevance or cue-continuity?, Neuropsychologia 87 (2016) 54–62.

[17] S. Gandhi, H. Plun-Favreau, Mutations and mechanism: how pink1 may contribute to risk of sporadic parkinsons disease, Brain 140 (1) (2017) 2–5.

[18] A. Giri, K. Y. Mok, I. Jansen, M. Sharma, C. Tesson, G. Mangone, S. Lesage, J. M. Bras, J. M. Shulman, U.-M. Sheerin, et al., Lack of evidence for a role of genetic variation in tmem230 in the risk for parkinson's disease in the caucasian population, Neurobiology of aging 50 (2017) 167–e11.

[19] A. Nieuwboer, N. Giladi, Characterizing freezing of gait in parkinson's disease: models of an episodic phenomenon, Movement Disorders 28 (11) (2013) 1509–1519.

[20] B. R. Bloem, J. M. Hausdorff, J. E. Visser, N. Giladi, Falls and freezing of gait in parkinson's disease: a review of two interconnected, episodic phenomena, Movement Disorders 19 (8) (2004) 871–884.

[21] C. Jenkinson, V. Peto, R. Fitzpatrick, R. Greenhall, N. Hyman, Self-reported functioning and well-being in patients with parkinson's disease: comparison of the short-form health survey (sf-36) and the parkinson's disease questionnaire (pdq-39), Age and ageing 24 (6) (1995) 505–509.

[22] N. Giladi, H. Shabtai, E. Simon, S. Biran, J. Tal, A. Korczyn, Construction of freezing of gait questionnaire for patients with parkinsonism, Parkinsonism & related disorders 6 (3) (2000) 165–170.

[23] C. Jenkinson, C. Clarke, R. Gray, P. Hewitson, N. Ives, D. Morley, C. Rick, K. Wheatley, A. Williams, Comparing results from long and short form versions of the parkinson's disease questionnaire in a longitudinal study, Parkinsonism & Related Disorders 21 (11) (2015) 1312–1316.

[24] A. Rizos, P. Martinez-Martin, S. Pal, C. Carroll, D. Martino, R. Sophia, C. Falup-Pecurariu, B. Kessel, A. Sauerbier, A. Martin, et al., The first parkinson's disease pain questionnaire (king's pd pain quest)–an interim analysis of a multicentre study of the patient's perspective, Parkinsonism & Related Disorders 22 (2016) e41.

[25] N. Giladi, T. Treves, E. Simon, H. Shabtai, Y. Orlov, B. Kandinov, D. Paleacu, A. Korczyn, Freezing of gait in patients with advanced parkinson's disease, Journal of neural transmission 108 (1) (2001) 53–61.

[26] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, P. Fisher, The hawthorne effect: a randomised, controlled trial, BMC medical research methodology 7 (1) (2007) 30.

[27] A. Nieuwboer, L. Rochester, T. Herman, W. Vandenberghe, G. E. Emil, T. Thomaes, N. Giladi, Reliability of the new freezing of gait question-

naire: agreement between patients with parkinson's disease and their carers, Gait & posture 30 (4) (2009) 459–463.

[28] N. Giladi, J. Tal, T. Azulay, O. Rascol, D. J. Brooks, E. Melamed, W. Oertel, W. H. Poewe, F. Stocchi, E. Tolosa, Validation of the freezing of gait questionnaire in patients with parkinson's disease, Movement Disorders 24 (5) (2009) 655–661.

[29] PD patient self care blog, http://www.riggare.se/1-vs-8765/, accessed: 2017-04-06.

[30] S. Del Din, A. Godfrey, C. Mazzà, S. Lord, L. Rochester, Free-living monitoring of parkinson's disease: Lessons from the field, Movement Disorders 31 (9) (2016) 1293–1313.

[31] W. Maetzler, J. Domingos, K. Srulijes, J. J. Ferreira, B. R. Bloem, Quantitative wearable sensors for objective assessment of parkinson's disease, Movement Disorders 28 (12) (2013) 1628–1637.

[32] C. F. Pasluosta, H. Gassner, J. Winkler, J. Klucken, B. M. Eskofier, An emerging era in the management of parkinson's disease: wearable technologies and the internet of things, IEEE journal of biomedical and health informatics 19 (6) (2015) 1873–1881.

[33] C. Pérez-López, A. Samà, D. Rodríguez-Martín, A. Català, J. Cabestany, J. M. Moreno-Arostegui, E. de Mingo, A. Rodríguez-Molinero, Assessing motor fluctuations in parkinsons disease patients based on a single inertial sensor, Sensors 16 (12) (2016) 2132.

[34] S. Mazilu, U. Blanke, A. Calatroni, E. Gazit, J. M. Hausdorff, G. Trster, The role of wrist-mounted inertial sensors in detecting gait freeze episodes in parkinsons disease, Pervasive and Mobile Computing 33 (2016) 1 – 16. doi:http://dx.doi.org/10.1016/j.pmcj.2015.12.007.
URL http://www.sciencedirect.com/science/article/pii/S157411921600002X

[35] D. Rodríguez-Martín, A. Samà, C. Pérez-López, A. Català, J. M. M. Arostegui, J. Cabestany, À. Bayés, S. Alcaine, B. Mestre, A. Prats, et al., Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer, PloS one 12 (2) (2017) e0171764.

[36] D. M. Rodríguez Martín, C. Pérez López, A. Samà Monsonís, A. Català Mallofré, J. M. Moreno Aróstegui, J. Cabestany Moncusí, B. Mestre, S. Alcaine, A. Prats, M. Cruz Crespo, et al., A waist-worn inertial measurement unit for parkinsons disease long-term monitoring, Sensors 17 (4) (2017) 1–28.

[37] A. Samà, C. Pérez-López, D. Rodríguez-Martín, A. Català, J. M. Moreno-Aróstegui, J. Cabestany, E. de Mingo, A. Rodríguez-Molinero, Estimating bradykinesia severity in parkinson's disease by analysing gait through a waist-worn sensor, Computers in biology and medicine 84 (2017) 114–123.

[38] A. Samà, D. Rodríguez-Martín, C. Pérez-López, A. Català, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo, À. Bayés, Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments, Pattern Recognition Letters.

[39] M. Bächlin, J. M. Hausdorff, D. Roggen, N. Giladi, M. Plotnik, G. Tröster, Online detection of freezing of gait in parkinson's disease patients: A performance characterization, in: Proceedings of the Fourth International Conference on Body Area Networks, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009, p. 11.

[40] S. Mazilu, M. Hardegger, Z. Zhu, D. Roggen, G. Troster, M. Plotnik, J. M. Hausdorff, Online detection of freezing of gait with smartphones and machine learning techniques, in: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on, IEEE, 2012, pp. 123–130.

[41] E. E. Tripoliti, A. T. Tzallas, M. G. Tsipouras, G. Rigas, P. Bougia, M. Leontiou, S. Konitsiotis, M. Chondrogiorgi, S. Tsouli, D. I. Fotiadis, Automatic detection of freezing of gait events in patients with parkinson's disease, Computer methods and programs in biomedicine 110 (1) (2013) 12–26.

[42] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[43] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, `http://www.deeplearningbook.org`.

[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385.

[45] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

[46] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz, et al., Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment, in: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, 2016, pp. 655–658.

[47] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 161–168.

[48] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140.

[49] Y. Freund, R. E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: European conference on computational learning theory, Springer, 1995, pp. 23–37.

[50] J. Friedman, T. Hastie, R. Tibshirani, et al., Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), The annals of statistics 28 (2) (2000) 337–407.

[51] Y. Freund, A more robust boosting algorithm, arXiv preprint arXiv:0905.2138.

[52] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[53] S. T. Moore, H. G. MacDougall, W. G. Ondo, Ambulatory monitoring of freezing of gait in parkinson's disease, Journal of neuroscience methods 167 (2) (2008) 340–348.

[54] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, G. Troster, Wearable assistant for parkinsons disease patients with the freezing of gait symptom, IEEE Transactions on Information Technology in Biomedicine 14 (2) (2010) 436–446.

[55] Y. Zhao, K. Tonn, K. Niazmand, U. M. Fietzek, L. T. D'Angelo, A. Ceballos-Baumann, T. C. Lueth, Online fog identification in parkinson's disease with a time-frequency combined algorithm, in: Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, IEEE, 2012, pp. 192–195.

[56] A. A. Handojoseno, J. M. Shine, T. N. Nguyen, Y. Tran, S. J. Lewis, H. T. Nguyen, The detection of freezing of gait in parkinson's disease patients using eeg signals based on wavelet decomposition, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 69–72.

[57] S. T. Moore, D. A. Yungher, T. R. Morris, V. Dilda, H. G. MacDougall, J. M. Shine, S. L. Naismith, S. J. Lewis, Autonomous identification of freezing of gait in parkinson's disease from lower-body segmental ac-

celerometry, Journal of neuroengineering and rehabilitation 10 (1) (2013) 19.

[58] C. Azevedo Coste, B. Sijobert, R. Pissard-Gibollet, M. Pasquier, B. Espiau, C. Geny, Detection of freezing of gait in parkinson disease: preliminary results, Sensors 14 (4) (2014) 6819–6827.

[59] D. Rodríguez-Martína, A. Samàa, C. Pérez-Lópeza, A. Catalàa, J. Cabestanya, P. Browneb, A. Rodríguez-Molinerod, Posture detection based on a waist-worn accelerometer: an application to improve freezing of gait detection in parkinsons disease patients, Recent Advances in Ambient Assisted Living-Bridging Assistive Technologies, E-Health and Personalized Health Care 20 (2015) 3.

[60] H. Zach, A. M. Janssen, A. H. Snijders, A. Delval, M. U. Ferraye, E. Auff, V. Weerdesteyn, B. R. Bloem, J. Nonnekes, Identifying freezing of gait in parkinson's disease during freezing provoking tasks using waist-mounted accelerometry, Parkinsonism & related disorders 21 (11) (2015) 1362–1366.

[61] S. Mazilu, A. Calatroni, E. Gazit, A. Mirelman, J. M. Hausdorff, G. Tröster, Prediction of freezing of gait in parkinson's from physiological wearables: an exploratory study, IEEE journal of biomedical and health informatics 19 (6) (2015) 1843–1854.

[62] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[63] N. Jaitly, G. E. Hinton, Vocal tract length perturbation (vtlp) improves speech recognition, in: Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, 2013, pp. 625–660.

[64] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[66] L. Prechelt, Automatic early stopping using cross validation: quantifying the criteria, Neural Networks 11 (4) (1998) 761–767.

[67] Z. Cataltepe, Y. S. Abu-Mostafa, M. Magdon-Ismail, No free lunch for early stopping, Neural computation 11 (4) (1999) 995–1009.

[68] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting., Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[69] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, Journal of Machine Learning Research 12 (Jul) (2011) 2121–2159.

[70] L. Bottou, Online learning and stochastic approximations, On-line learning in neural networks 17 (9) (1998) 142.

[71] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2).

[72] M. D. Zeiler, ADADELTA: an adaptive learning rate method, CoRR abs/1212.5701.
URL http://arxiv.org/abs/1212.5701

[73] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.
URL http://arxiv.org/abs/1412.6980

[74] J-BHI Special Issue on 'Deep Learning for Biomedical and Health Informatics', Journal: Journal of Biomedical and Health Informatics; Editor-in-Chief: Guang-Zhong, Yang, http://jbhi.embs.org/special-issues/

deep-learning-for-biomedical-and-health-informatics/, accessed: 2017-04-11.

[75] D. Rav, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, G. Z. Yang, Deep learning for health informatics, IEEE Journal of Biomedical and Health Informatics 21 (1) (2017) 4–21. doi:10.1109/JBHI.2016.2636665.

[76] H. Greenspan, B. van Ginneken, R. M. Summers, Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique, IEEE Transactions on Medical Imaging 35 (5) (2016) 1153–1159.

[77] H.-I. Suk, C.-Y. Wee, S.-W. Lee, D. Shen, State-space model with deep learning for functional dynamics estimation in resting-state fmri, NeuroImage 129 (2016) 292–307.

[78] J. Chen, L. Yang, Y. Zhang, M. Alber, D. Z. Chen, Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation, in: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 3036–3044.
URL http://papers.nips.cc/paper/6448-combining-fully-convolutional-and-recurrent-neura
pdf

[79] J. E. Weston, Dialog-based language learning, in: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 829–837.
URL http://papers.nips.cc/paper/6264-dialog-based-language-learning.pdf

[80] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, et al., Hybrid computing using a neural network with dynamic external memory, Nature 538 (7626) (2016) 471–476.

[81] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, S. Nemati, A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology, in: Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on, IEEE, 2017, pp. 141–144.

[82] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, J. P. Papa, Deep learning-aided parkinson's disease diagnosis from handwritten dynamics, in: Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on, IEEE, 2016, pp. 340–346.

[83] C. R. Pereira, D. R. Pereira, J. P. Papa, G. H. Rosa, X.-S. Yang, Convolutional neural networks applied for parkinsons disease identification, in: Machine Learning for Health Informatics, Springer, 2016, pp. 377–390.

[84] A. Frid, A. Kantor, D. Svechin, L. M. Manevitz, Diagnosis of parkinson's disease from continuous speech using deep convolutional networks without manual selection of features, in: Science of Electrical Engineering (ICSEE), IEEE International Conference on the, IEEE, 2016, pp. 1–4.

[85] C. Stamate, G. D. Magoulas, S. Küppers, E. Nomikou, I. Daskalopoulos, M. U. Luchini, T. Moussouri, G. Roussos, Deep learning parkinson's from smartphone data, in: Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on, IEEE, 2017, pp. 31–40.

[86] D. Rodríguez-Martín, C. Pérez-López, A. Samà, J. Cabestany, A. Català, A wearable inertial measurement unit for long-term monitoring in the dependency care area, Sensors 13 (10) (2013) 14079–14104.

[87] A. Samà Monsonís, C. Pérez López, D. M. Rodríguez Martín, J. Cabestany Moncusí, J. M. Moreno Aróstegui, A. Rodríguez Molinero, A heterogeneous database for movement knowledge extraction in parkinson's disease, in: ESANN 2013 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 24-26 April 2013, 2013, pp. 413–418.

[88] A. J. Hughes, S. E. Daniel, L. Kilford, A. J. Lees, Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases., Journal of Neurology, Neurosurgery & Psychiatry 55 (3) (1992) 181–184.

[89] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks., in: Aistats, Vol. 9, 2010, pp. 249–256.

[90] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Rusboost: Improving classification performance when training data is skewed, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE, 2008, pp. 1–4.

[91] J.-P. Vert, K. Tsuda, B. Schölkopf, A primer on kernel methods, Kernel Methods in Computational Biology (2004) 35–70.

[92] F. Chollet, Keras, `https://github.com/fchollet/keras` (2015).

[93] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).
URL `http://tensorflow.org/`

[94] source code github, `https://github.com/juliacamps/FOG`, accessed: 2017-04-22.

[95] iNEMO inertial module lsm9ds1, `http://www.st.com/en/mems-and-sensors/lsm9ds1.html`, accessed: 2017-07-04.

[96] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[97] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259.

[98] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555.

[99] micro controller stm32f415rg, `http://www.st.com/en/microcontrollers/stm32f415rg.html`, accessed: 2017-07-04.

[100] Improving Quality of Life with an Automatic Control System (MAS-PARK), `http://futur.upc.edu/15557508`, accessed: 2017-04-11.

[101] Remote and Autonomous Management of Parkinson's Disease (REMPARK), `http://www.rempark.eu/`, accessed: 2017-04-11.