POLITECNICO DI TORINO
Repository ISTITUZIONALE

Deep learning algorithms for detecting freezing of gait in Parkinson’s disease: A cross-dataset study

Original
Deep learning algorithms for detecting freezing of gait in Parkinson’s disease: A cross-dataset study / Sigcha, Luis; Borzì, Luigi; Olmo, Gabriella. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - ELETTRONICO. - 255:A(2024). [10.1016/j.eswa.2024.124522]

Availability:
This version is available at: 11583/2990028 since: 2024-07-01T07:01:54Z

Publisher:
Elsevier

Published
DOI:10.1016/j.eswa.2024.124522

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

01 August 2024

Expert Systems With Applications 255 (2024) 124522

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Deep learning algorithms for detecting freezing of gait in Parkinson's disease: A cross-dataset study

Luis Sigcha [a,b,c,*,1], Luigi Borzì [d,e,1], Gabriella Olmo [d]

[a] Department of Physical Education and Sports Science (PESS), University of Limerick, V94 T9PX, Limerick, Ireland
[b] Health Research Institute (HRI), University of Limerick, V94 T9PX, Limerick, Ireland
[c] Data-Driven Computer Engineering (D2iCE) Group, Department of Electronic and Computer Engineering, University of Limerick, V94 T9PX, Limerick, Ireland
[d] ANTHEA Lab–Data Analytics and Technologies for Health Lab, Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy
[e] PolitoBIOMed Lab–Biomedical Engineering Lab, Politecnico di Torino, 10129 Turin, Italy

ARTICLE INFO

Dataset link: this link, this link, this link

Keywords:
Wearable sensor
Accelerometer
Machine learning
Deep learning
Cross-dataset test
Motor symptoms

ABSTRACT

Freezing of gait is a complex and disabling symptom of Parkinson's disease, which has a significant impact on the patients' quality of life and increases the risk of falls and related injuries. This study aims to evaluate the generalization capability of deep learning algorithms in freezing of gait detection. To address this task, various machine learning and deep learning algorithms were implemented, fine-tuned, and evaluated using diverse data splitting and validation strategies. The experiments performed yielded mixed results. Although the implementations demonstrated competitive performance in single-dataset settings (area under the curve ranging from 0.77 to 0.94), all approaches showed limited robustness in cross-dataset tests and suboptimal generalization across different datasets (area under the curve ranging from 0.65 to 0.84). These results highlight the importance of standardized data collection procedures to ensure uniformity. The specification of sensor settings and predefined sensor locations can foster homogeneity in datasets, even when dealing with diverse subjects and environments. Such standardization efforts are crucial for advancing generalized methodologies in the detection of freezing of gait, applicable to both research and clinical applications.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that affects millions of people worldwide (Samii, Nutt, & Ransom, 2004). It is characterized by the progressive degeneration of dopaminergic neurons in the brain, leading to a wide range of motor and mental symptoms. The main motor symptoms include resting tremor, bradykinesia (slowed movements), muscle rigidity, and postural instability (Armstrong & Okun, 2020). These have a major impact on mobility and general motor function, leading to a gradual loss of autonomy and reducing quality of life (QoL) (Zhang et al., 2020).

Current treatment approaches for PD mainly involve the administration of dopaminergic drugs, such as levodopa, to relieve motor symptoms. However, the efficacy of drugs can vary from individual to individual, and long-term use can lead to complications and motor fluctuations (Reich & Savitt, 2019; Zhao et al., 2021).

Among the motor symptoms, freezing of gait (FoG) stands out as a complex and debilitating phenomenon that has a significant impact on the QoL of people with PD (PwPD) and increases the risk of falls and related injuries (Gao, Liu, Tan, & Chen, 2020). FoG is characterized by sudden, transient episodes of gait disruption, in which individuals experience the sensation of being "stuck" on the ground, unable to start or continue walking (Nutt et al., 2011). FoG typically occurs during complex motor activities, such as turning or traversing narrow spaces, and under motor or cognitive dual tasks. It affects about 50%–80% of PwPD, often in the advanced stages of the disease (Zhang, Gao, Tan, & Chen, 2021).

Accurate, objective, and continuous monitoring of FoG is critical for effective management and treatment of PD. Traditional evaluation methods involve clinical assessments and subjective patient reports (Kobylecki, 2020). However, these approaches have limitations, as symptoms can fluctuate throughout the day, making it difficult to capture the full extent of motor disturbances (Bhidayasiri & Martinez-Martin, 2017).

To address these challenges, wearable devices have emerged as a promising solution for continuous monitoring of PD symptoms, including FoG (Channa, Popescu, & Ciobanu, 2020; Del Din, Kirk, Yarnall, Rochester, & Hausdorff, 2021). Advances in wearable technology have recently paved the way for objective and non-intrusive monitoring of

* Corresponding author at: Department of Physical Education and Sports Science (PESS), University of Limerick, V94 T9PX, Limerick, Ireland.
E-mail addresses: luis.sigcha@ul.ie (L. Sigcha), luigi.borzi@polito.it (L. Borzì), gabriella.olmo@polito.it (G. Olmo).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2024.124522
Received 26 February 2024; Received in revised form 29 March 2024; Accepted 15 June 2024
Available online 20 June 2024
0957-4174/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

FoG in daily life (Pardoel, Kofman, Nantel, & Lemaire, 2019). Wearable devices typically incorporate inertial sensors, such as accelerometers and gyroscopes, that can capture movement and postural changes in real-time. These sensors can be strategically placed on various parts of the body, including the lower limbs, trunk, and wrists, to capture relevant data related to gait patterns and FoG events (Rovini, Maremmani, & Cavallo, 2017). The number and location of sensors may vary depending on the design of the device and the specific objectives of the monitoring system (Sigcha et al., 2023).

In order to facilitate the detection of FoG episodes from sensor data, machine learning (ML) techniques have been widely used. Traditional ML algorithms, such as support vector machine (SVM) and random forest (RF), have shown promise in detecting FoG episodes with reasonable accuracy (Giannakopoulou, Roussaki, & Demestichas, 2022). However, more recent advances in deep learning (DL), particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown even greater potential for FoG detection (Sigcha et al., 2023). DL models can automatically learn discriminative features from raw sensor data, enabling a more robust and accurate detection of FoG episodes (Alzubaidi et al., 2021). In this context, the integration of wearable inertial sensors and DL techniques offers a novel and promising approach for the detection and monitoring of FoG. By providing objective, real-time assessments, this technology has the potential to enhance personalized therapy and improve the overall management of PD, ultimately improving the QoL of PwPD.

The rest of this paper is organized as follows. Related works are discussed in Section 2, together with their limitations and the contribution of this study. Section 3 describes the material and methods, including the pre-processing procedures, the classic ML approaches and DL architectures implemented, the experimental procedures, the datasets used in this study, training and optimization processes, validation method, and performance evaluation strategy. Results are reported and discussed in Section 4, while the conclusions are drawn in Section 5, along with future directions.

2. Related work

In the realm of FoG detection, existing research exhibits a notable diversity in sensor settings and experimental protocols. One striking aspect of this diversity lies in the configuration of wearable sensors. Studies have employed varying numbers of sensors, ranging from single-sensor setups to complex multi-sensor arrays. This choice often reflects the trade-off between simplicity and comprehensive data capture. As stated by O'Day et al. (2022), the optimal technical configuration for FoG detection comprised three sensors placed on the lumbar region and both ankles. Moreover, these locations were considered highly wearable by a cohort of 16 PwPD.

In terms of experimental procedures, significant variability is apparent. Although some studies primarily analyze gait during standard walking tasks (Naghavi & Wade, 2022; Shi et al., 2022), others incorporate FoG-provoking activities designed to induce FoG episodes intentionally (Bikias, Iakovakis, Hadjidimitriou, Charisis, & Hadjileontiadis, 2021; Reches et al., 2020; Zhang et al., 2020). These provocations provide insights into the onset and characteristics of FoG but may differ in their ecological validity (e.g., FoG patterns collected in real-world settings). Furthermore, a subset of research explores gait patterns during simulated activities of daily living (ADL), with the aim of assessing gait in contexts that better mimic real-life scenarios (Rodríguez-Martín et al., 2017).

As far as concerns available data, the limited availability of publicly shared datasets is a notable challenge in FoG detection research. Most studies designed the experimental procedures and collected proprietary datasets (Sigcha et al., 2023). This scarcity hampers research progress, limits benchmarking opportunities, raises generalizability concerns, and hinders collaboration and innovation. Up to date, only a few datasets are available to the research community, and they were collected using a different sensor configuration. Specifically, three sensors on the ankle, thigh and lower back of 10 PwPD (Bächlin, Plotnik, Roggen, Maidan, J.M., Giladi, & Tröster, 2010); a single sensor on the wrist of 18 PwPD (Mazilu et al., 2016); and six sensors on the feet, shanks, and the lumbar and chest regions of 7 PwPD (O'Day et al., 2022).

Diverse sets of features and ML techniques have been used to detect FoG. ML algorithms include neural networks, decision trees, RF, naïve Bayes, k-nearest neighbor, and SVMs. Among these algorithms, SVM and RF implementations provided the best results, with 0.75–0.99 sensitivity and 0.79–1 specificity, and 0.66–0.98 sensitivity and 0.66–0.99 specificity, respectively (Pardoel et al., 2019). DL algorithms developed for FoG detection included CNNs, long short-term memory networks (LSTM), and deep autoencoders, providing 0.63–0.95 sensitivity and 0.67–0.99 specificity (Sigcha et al., 2023). Furthermore, contextualization of gait patterns has been investigated to exploit the inherent circumstances of FoG manifestations, aiming at reducing the computational burden of FoG detection algorithms (Borzì & Sigcha & Olmo, 2023; Pepa et al., 2020).

However, the comparison of performance between different approaches is not immediate, due to the diverse validation/test strategies. Specifically, high performance was obtained when data from all patients were merged and randomly shuffled to generate training and test sets (Ashfaque Mostafa, Soltaninejad, McIsaac, & Cheng, 2021; Kim et al., 2018). However, this approach does not ensure subject independence in the two sets and generates over-optimistic results. To overcome this issue, most studies used a leave-one-subject-out approach (Naghavi & Wade, 2022; San-Segundo, Navarro-Hellín, Torres-Sánchez, Hodgins, & De la Torre, 2019; Sigcha et al., 2022, 2020), consisting of iteratively using a single subject as test and all the remaining subjects for training. Finally, hold-out validation consists of using a portion of subjects for training and the remaining part for validation (Ashour, El-Attar, Dey, El-Kader, & Abd El-Naby, 2020; Noor, Nazir, Wahab, & Ling, 2021; Shi et al., 2022). Since the validation set is commonly used for model optimization, often an independent test set is used as the final test to provide a realistic estimate of performance in unseen data (Borzì, Sigcha, Rodríguez-Martín, & Olmo, 2023; Camps et al., 2018).

Furthermore, in Borzì et al. (2023), a dataset of 21 PwPD was used for training, validation, and test. To assess the model generalization capability, further testing was performed on two external datasets comprising 38 PwPD with FoG (Borzì & Olmo & Artusi & Lopiano, 2020) and 59 PwPD without FoG (Borzì, Olmo, Artusi, Fabbri, Rizzone, Romagnolo, Zibetti, & Lopiano, 2020). The results showed a net reduction in sensitivity and a slight increase in specificity. This represents the first evidence of how the performance of FoG detection algorithms can vary across datasets. However, the external datasets used as independent test sets included few (i.e., 50 FoG episodes) to any FoG. A recent work (Klaver et al., 2023) merged data from four separate studies and evaluated the performance of different predictive models. The study confirmed the potential of CNNs to detect FoG in a large (70 participants) and diverse dataset, achieving 0.85 sensitivity and 0.68 specificity in an external dataset. In addition, performance varied significantly from the training set to the test set to the external dataset. These results highlight the difficulty in developing a general FoG detection method. This is even more evident when considering that the study consistently used the same motion-capture equipment in all included datasets.

Although these studies have made significant progress in FoG detection, there are some limitations that still remain. Most studies have used data recorded in controlled laboratory environments. This raises questions about the applicability of these models in real-world scenarios, where contextual factors and varying environmental conditions may affect the accuracy of FoG detection. Moreover, the sample size is reduced in many studies, often consisting of fewer than 20 subjects. Even though

these studies have provided valuable insights, the generalizability of the developed DL models to larger populations remains uncertain. In addition, few studies have conducted tests on different datasets recorded from different subjects and settings. This scarcity of diverse test data raises questions about the robustness and generalizability of these models when employed in real-world scenarios.

Considering these limitations, the present study aims to address the need to evaluate the generalization capability of DL algorithms that have shown good performance on the original dataset, but whose performance in different datasets has not been thoroughly evaluated. Specifically, this study systematically evaluates ML and DL FoG detection algorithms using a merged and harmonized dataset comprised of heterogeneous data from three well-known reference datasets. By integrating diverse data sources, this study provides a novel perspective of the effectiveness of DL methods on heterogeneous data with main goal of providing insights to improve the algorithm development and identify potential challenges in the current FoG detection techniques.

By evaluating the performance of the models on different patients in different situations, this research aims to determine the reliability and robustness of these models in real-world settings, providing valuable insights for the development of more effective FoG detection systems in PD. The main contributions of this work are as follows. (a) Data from a single sensor (i.e., accelerometer) on the lower back were used; this aims to increase patient comfort and promotes continuous and long-term monitoring of FoG in daily life; (b) Different ML and DL models presented in the literature are reproduced, adapted, and appropriately trained and evaluated; this allows the performance of different approaches in FoG detection to be compared; (c) Different datasets are considered; this allows comparison of model performance with data collected from different samples during different experimental settings; (d) Appropriate validation and testing strategies are used to evaluate FoG detection performance; ensuring subject independence in training and test sets allows for robust and realistic performance estimates; (e) Cross-dataset evaluations are investigated; testing the model on datasets other than those included in the training phase provides a true estimate of the generalization ability of DL models in FoG recognition; (f) The contribution of features to model performance is compared across datasets.

3. Materials and methods

This section describes the materials and methods used in this study. More specifically, Section 3.1 provides an overview of the datasets used in this study; Section 3.2 reports the data preprocessing procedures; Section 3.3 describes the implemented ML and DL algorithms; the methods used for data splitting and validation strategies are reported in Sections 3.4 and 3.5; details on the optimization procedure and training settings are provided in Sections 3.6 and 3.7; finally, Section 3.8 reports the classification metrics used for performance evaluation.

3.1. Datasets

This study includes three different datasets, corresponding to different numbers of PwPD who participated in different sets of activities. Data from a single tri-axial accelerometer placed on the lower back were used for the analysis. This choice was made to ensure the compatibility of the data across various datasets. Furthermore, the lower back is considered the most suitable and commonly adopted location for placing a single sensor in the context of gait and FoG detection (Huang, Li, & Huang, 2023). Additionally, in the study conducted by Keogh et al. (2023), the placement of a wearable device on the waist for one week was considered comfortable and acceptable by a cohort of 106 PwPD. The combination of accuracy and comfort aims at providing a reliable monitoring system that can be implemented in real-world settings.

A brief description of the datasets is provided in Table 1, in terms of number of subjects, total amount of data, number of FoG episodes, and total amount of FoG.

Table 1
Description of the datasets used in this study.

Dataset	Rempark	Daphnet	Oday
Number of subjects	21	10	7
Total data (hours)	11.7	8.3	1.4
Number of FoG episodes	1075	272	211
Total FoG duration (minutes)	73.6	28.9	22
FoG percentage	10.5%	5.8%	26.2%

The Rempark dataset (Rodríguez-Martín et al., 2017) comprises data from twenty-one PwPD. The sample consisted of three women and eighteen men, with an average age of 69.3 ± 9.7. The participants had a disease duration of 9 ± 4.8 years, Hoehn and Yahr (H&Y) score of 3.1 ± 0.4, freezing of gait questionnaire (FoG-Q) score of 15.8 ± 4.1, mini-mental state examination score (MMSE) of 27.8 ± 1.9, and a total unified Parkinson's disease rating scale (UPDRS) part-III score of 16.2 ± 9.7 ON and 36.3 ± 14.4 OFF therapy. An inertial measurement unit (IMU) was attached to the left side of the waist using an elastic band to record three-axis acceleration data. The experiments were conducted in the participant's home, and data were collected both while the participants were ON and OFF dopaminergic therapy. The tasks performed included gait tasks such as walking outdoors, the stand-up-and-go test, and showing the participant's home. Additionally, false positive analysis tasks such as cleaning windows, brushing teeth, and painting/drawing/erasing on a sheet of paper were considered for the study.

The Daphnet dataset (Bächlin et al., 2010) comprises data from ten PwPD. The sample consisted of seven men and three women, with an average age of 66.4 ± 4.8 years, disease duration of 13.7 ± 9.7 years, and H&Y score of 2.6 ± 0.65 in ON condition. During the experiments, data were recorded from three accelerometers placed on the shank, thigh, and lower back. Participants were asked to complete three walking tasks that aimed to represent different aspects of daily walking. These tasks included walking back and forth in a straight line along the lab hallway and random walking in a reception hall space with initiated stops and 360-degree turns. In addition, walking while simulating ADL was considered in the protocol, including entry and exit of rooms and walking to the lab kitchen, getting a drink, and returning to the starting room with a cup of water. The experiments were carried out in the morning during the OFF stage of the medication cycle, which was more than 12 h after their last drug intake.

The dataset described in O'Day et al. (2022), hereafter referred to as Oday, comprises data from seven PwPD. The sample consisted of four men and three women, with an average age of 58.4 ± 5.1 years and disease duration of 10.1 ± 2.4 years. Six IMUs were strapped on the tops of both feet, the lateral side of both shanks, and the lumbar (L5) and chest regions. Each participant provided different walking sessions through the turning and barrier course specifically designed to elicit FoG. Each walking trial (walk) consisted of two ellipses and two figures of eight around tall barriers. Participants completed all trials OFF medication and OFF deep brain stimulation. A video of each walk was synchronized with the IMU system. The experiments were carried out over 2 to 6 clinic visits separated by up to 44 months.

3.2. Data preprocessing

Sensor settings were different in the datasets used in this study, as reported in Table 2. To standardize the different sensor configurations, data were scaled by converting the unit of measurement to g units, where g is the value of earth acceleration. In addition, the order and directions of the axes were adjusted so that the x, y and z axes pointed vertically (downward), posteriorly and mid-laterally (to the right). Finally, the data were resampled to 32 Hz. The latter value allows for a good representation of the frequency components of the acceleration signals during gait (0–3 Hz) and FoG (3–8 Hz) (Bächlin et al., 2010;

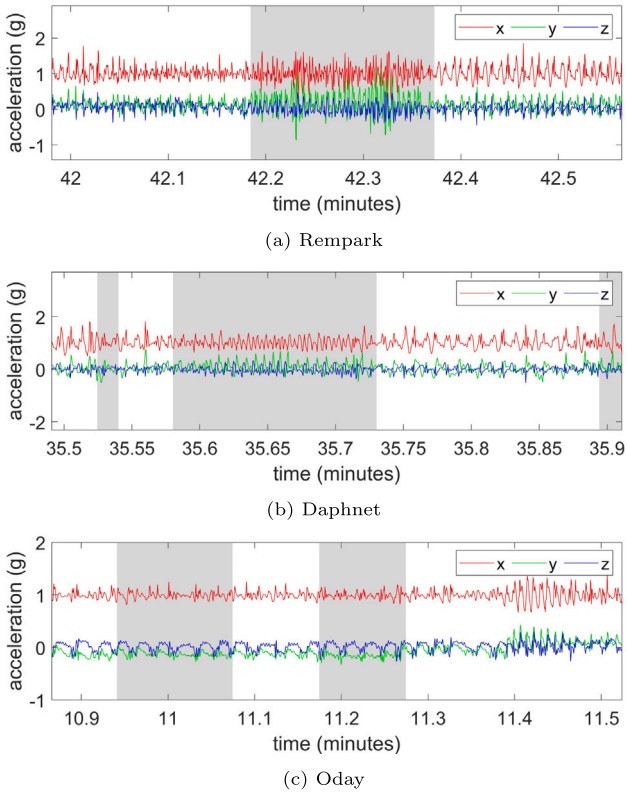


Fig. 1. Three-axis acceleration readings from different datasets. Gray zones identify FoG episodes.

Moore, MacDougall, & W.G., 2008) while removing high-frequency noise (Li et al., 2020).

A visual representation of the resulting acceleration signals is provided in Fig. 1, where gray areas identify FoG episodes. From Fig. 1, different gait and FoG patterns are clearly visible, differing between datasets in amplitude and frequency content.

The resampling strategy consisted of down-sampling the original signals using linear interpolation. A finite impulse response (FIR) anti-aliasing low-pass filter with a Kaiser window (β=5) was applied, compensating for the delay introduced by the filter. Information loss due to resampling was calculated using the mean absolute error (MAE) between the resampled and the original signal. The average MAE during FoG is reported in Table 2, expressed as an average over the three acceleration axes. For comparison, the average signal range during FoG is also reported. Overall, the resampling error was found to be two orders of magnitude less than the original signal, thus demonstrating a small loss of information.

Acceleration data were segmented into 2-second windows with 50% overlap. The data were then rescaled to be centered on zero, removing the mean value of each acceleration component in each window separately, as done in Borzì and Sigcha and Olmo (2023), Borzì et al. (2023).

On the one hand, raw accelerometer data were used in all experiments to evaluate the performance of DL approaches including a well-known and low-complexity CNN architecture (Bikias et al., 2021), used as a baseline DL approach. On the other hand, to obtain baseline results using a classical ML approach, the set of temporal and spectral features proposed by Mazilu et al. (2012) was extracted and used as input of a predictive the model. This set of features was selected due to its excellent trade-off between complexity and performance, as evidenced in previous studies (Camps et al., 2018; San-Segundo et al., 2019; Sigcha et al., 2023).

Table 2
Sensor orientation, unit of measurement, and sampling frequency in different datasets. MAE: mean absolute error.

Dataset	Rempark	Daphnet	Oday
Unit of measurement	$\frac{m}{s^2}$	mg	$\frac{m}{s^2}$
Sampling frequency (Hz)	40	64	128
X-axis	anterior	anterior	vertical (↓)
Y-axis	vertical (↑)	vertical (↓)	lateral (→)
Z-axis	lateral (←)	lateral (→)	posterior
Resampling MAE (mg)	2.7	6.5	3.3
Average signal range (g)	0.41	0.54	0.34

3.3. Machine and deep learning methods

Diverse classification algorithms were implemented and evaluated using different data-splitting strategies. A classic ML processing pipeline is described in Section 3.3.1, while the implementation of the DL-based algorithms is reported in Section 3.3.2.

3.3.1. Machine learning methods

Classic ML pipelines require the extraction of discriminative features from raw signals. In this study, the set of features proposed in Mazilu et al. (2012) was considered. The extracted features include time and frequency domains, namely: mean, standard deviation, variance, entropy, energy, freezing index (freezing band power 3–8 Hz divided by locomotor band power 0.5–3 Hz), and the sum of the freezing band and locomotor band power. The seven features were extracted from each component of the 3-axis acceleration signal, thus leading to a total number of 21 features. These were input of an RF algorithm (Breiman, 2001). RF represents a well-known and widely used classification algorithm that has provided top performance in FoG detection (Giannakopoulou et al., 2022; Pardoel et al., 2019; Zhang, Sun, Huang, et al., 2024) while maintaining low computational complexity. Moreover, it has been used as baseline evaluation model in previous research studies (Mazilu et al., 2012; San-Segundo et al., 2019; Sigcha et al., 2022, 2020).

3.3.2. Deep learning methods

Three different DL algorithms were implemented, fine-tuned and evaluated. All consisted of different types of CNNs. These DL algorithms excel at extracting intricate features and abstract patterns from extensive datasets through convolution operations applied to input data. CNNs capitalize on three key principles: sparse interactions, parameter sharing, and equivariant representations (LeCun, Bengio, & Hinton, 2015). They possess the ability to autonomously discern features from images and signals, leading to cutting-edge performance in various classification tasks. In domains such as time series classification, CNNs offer notable advantages over alternative models, particularly in terms of capturing local dependencies and maintaining scale invariance (Jindong, Yiqiang, Shuji, Xiaohui, & Lisha, 2019). Among DL classification algorithms, CNNs have been the most widely used in PD and also in FoG recognition problems. Moreover, they have demonstrated superior FoG recognition performance compared to RNNs and LSTMs (Sigcha et al., 2023).

To create a baseline based on DL, the CNN multilayer perceptron (CNN-MLP) architecture proposed in Bikias et al. (2021) was reproduced. Moreover, this study evaluated the potential of ConvMixer (Trockman & Kolter, 2022) adapted for inertial signals as a mechanism to reduce the computational burden. Finally, an architecture adapted from a previous work (Borzì et al., 2023) that employs three different convolutional heads was adapted and evaluated. In this work, we refer to this architecture as Wide-CNN, due to the adaptation performed in the number of input heads. The detailed implementation of these DL architectures is described in the following.

CNN-MLP: A network that uses CNN with max-pooling and MLP was adapted and evaluated using a data representation based on raw signals (Bikias et al., 2021). Although the principal settings (number of CNN layers and their parameters) of this architecture are similar to those reported in the corresponding article (Bikias et al., 2021), the present reproduction differs in the number of sensors and signals. Specifically, the output layer was adapted to perform a binary prediction using a dense layer with a sigmoid activation function, which outputs the probability of a FoG event. The implemented algorithm is shown in Fig. 2.

The input of the CNN-MLP model has a size of 64 × 3, where 64 corresponds to the window size (i.e., 2 s with a sampling rate of 32 Hz) and 3 refers to the three-dimensional acceleration signal. The two convolutional layers have 100 and 40 filters, respectively, and a kernel size of 10. A max-pooling layer with pool size of 2 was included between the two layers to reduce the feature map size. A dropout regularization technique was applied to the CNN layers (value of 0.3) and dense layer (value of 0.5) to avoid over-fitting. Next, a global average pooling (GAP) layer was used to reduce the CNN feature maps to a one-dimensional (1D) vector. Finally, this layer was connected to an output layer that with a single neuron with sigmoid activation.

ConvMixer: An adapted version of a ConvMixer (Trockman & Kolter, 2022) was implemented to determine its performance in FoG detection. The ConvMixer is a novel architecture proposed for image recognition. It consists of a patch embedding stage followed by isotropic convolutional blocks. This architecture aims to achieve competitive results with much lower complexity and computational cost. In more detail, the patch embedding stage divides the input signals into non-overlapping patches and projects the data from a multichannel sliding window into another dimension that can be exploited by classification algorithms. This patching strategy is achieved through a single convolution operation using kernel and stride parameters of the same dimension, generating non-overlapping patches in raster-scan order. Unlike traditional CNNs that rely on convolution and pooling layers, or Transformers networks that use self-attention layers, ConvMixer blocks employ processing blocks with deep separable convolutions, a common feature in modern CNN architectures. Furthermore, in the ConvMixer, the feature size remains consistent throughout the layers unlike traditional CNN architectures, where the feature size decreases. The implemented ConvMixer algorithm is shown in Fig. 3.

The input of the ConvMixer has size 64 × 3 as in the CNN-MLP model. The input is connected to the Patch embedding block. The extracted patches are connected to a gaussian error linear units (GELU) activation and batch normalization (BatchNorm) layer. After this, a ConvMixer section consisting of two ConvMixer blocks is used. Each ConvMixer block uses a depthwise separable convolution that consists of a depthwise convolution followed by a point-wise convolution with a residual connection; moreover, GELU and BatchNorm are applied after each convolution. Finally, the ConvMixer section is connected to a GAP and the output layer with a single neuron and sigmoid activation.

Wide-CNN: An adapted version of the multi-head CNN proposed in Borzì et al. (2023) was implemented and evaluated. This model uses different spatial resolutions to enhance the classification performance. The reproduction of this architecture differs from the original implementation in the use of a single input head instead of using a multi-head approach. In the Wide-CNN the input is copied to three different branches; however, the main idea of using three different spatial resolutions is held in order to capture useful features from the local to the global level (Borzì et al., 2023). Finally, the output layer was adapted to perform a binary prediction using a dense layer with sigmoid activation. The implemented Wide-CNN algorithm is shown in Fig. 4.

In specific, the input of the Wide-CNN is connected to three branches. Each of them comprises two convolutional layers with 16 filters and two max-pooling layers with a pool size of 3. The kernel size is different in each head and reduces in size in deeper layers. Specifically, kernel sizes of 6 and 3, 12 and 6, and 18 and 9 were used in each CNN branch, respectively. In addition, a dropout strategy was applied in each CNN layer and the dense layer, the values of each dropout were optimized for each experiment (see Section 3.6) with values ranging from 0.1 to 0.4. The resulting feature maps generated by each branch were flattened and concatenated with those of the other branches. Subsequently, a dense layer with 16 neurons and a rectified linear unit (ReLU) activation function was used. Finally, the output layer with a single neuron and sigmoid activation provides the class probability.

3.4. Data splitting and validation method

To comprehensively evaluate the performance of the models and accurately assess their robustness, several data-splitting strategies and validation methods were used. Fig. 5 summarizes the evaluation methods and the results obtained in this study.

In more detail, the datasets were independently evaluated using a single dataset validation to generate baseline results for both ML and DL approaches. After this, the datasets were consolidated into a unified dataset and evaluated using an all-in-one validation approach to develop a generalized model. Finally, cross-dataset validation was performed using only the data from a single dataset as a test subset.

A detailed overview of these methods is given in the following sections. It is worth noting that all validation strategies were performed by ensuring the independence of the subjects in the different sets, i.e., training, validation, and testing. In particular, data from the same subject belong only to a single subset, thus avoiding overfitting and providing more realistic estimates of model performance in unseen data.

3.4.1. Single dataset evaluation

Fig. 6 schematically reports the validation method at the single dataset level. This represents the most common validation approach, in which a single dataset is divided into training, validation, and test subsets. This validation approach was used to obtain baseline results, allowing to compare the performance of the different models implemented in this study. Additionally, these results allow for a comparison of the performance of our reproductions with state-of-the-art approaches proposed in similar studies.

Subjects were assigned to the training, validation, and test sets according either to the H&Y stage (Rempark and Daphnet datasets) or the percent time spent with FoG (Oday dataset), with a proportion of subjects of 0.5, 0.25 and 0.25 in the three sets. The distribution of subjects and the mean H&Y for each subset are provided in Table 3. This settings were used in subsequent experiments to create different subsets of training, validation and testing.

3.4.2. All-in-one evaluation

Fig. 7 describes the all-in-one dataset evaluation. For this task, the three datasets were combined to form a single one. This was done by combining each of the three subsets (e.g., training, validation, test). In this case, the different subsets include data from all datasets, but they are independent because they include data from different subjects. This validation approach aims to test whether increasing the size

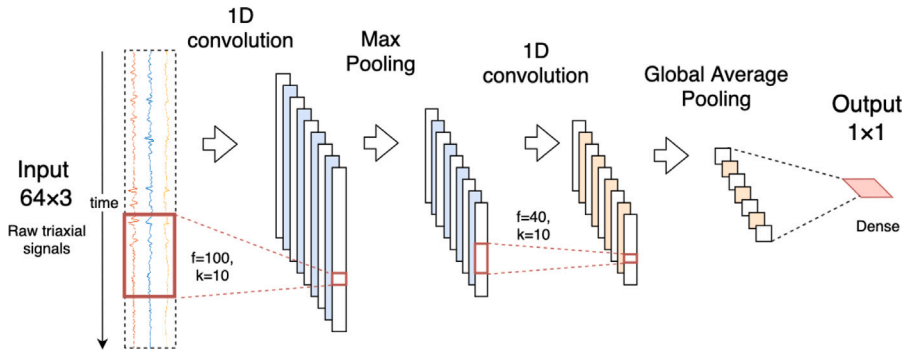


Fig. 2. CNN-MLP architecture. f: number of filters; k: kernel size.

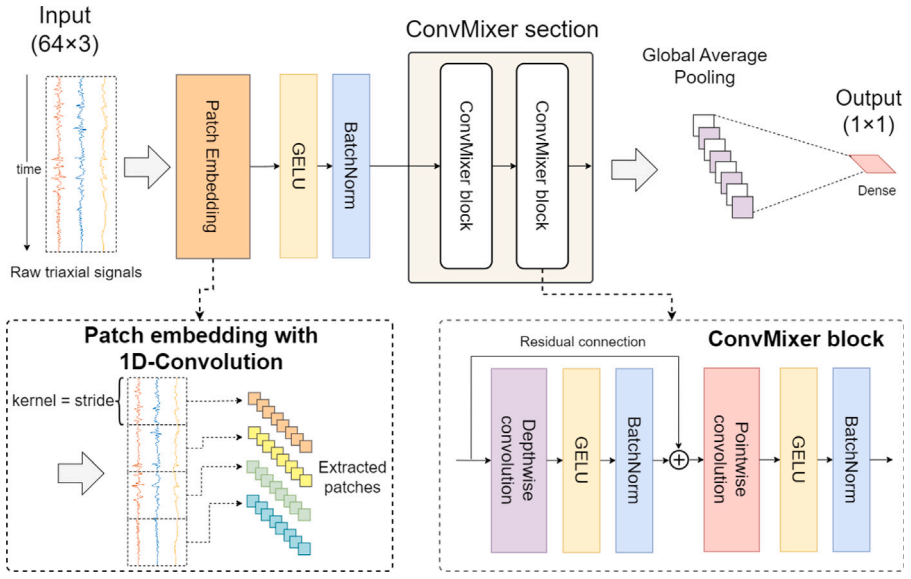


Fig. 3. ConvMixer architecture adapted for inertial signals. GELU: gaussian error linear unit; BatchNorm: batch normalization.

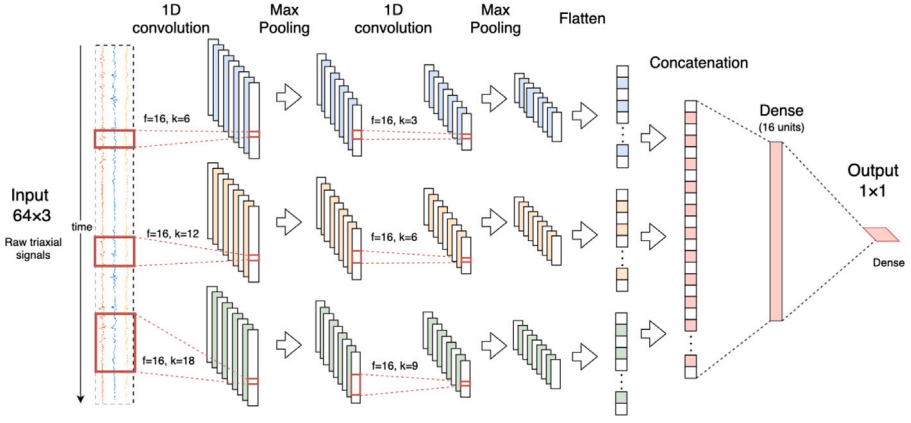


Fig. 4. Wide CNN. f: number of filters; k: kernel size.

and heterogeneity of the dataset is beneficial to the performance and robustness of the model.

3.5. Cross-dataset evaluation

A cross-dataset evaluation strategy was adopted to evaluate whether the combination of different datasets in the training and validation sets provides consistent performance in the test dataset.

Fig. 8 schematically shows the cross-dataset evaluation methodology. In this case, a single dataset is used as a test subset (e.g., combining train, validation, and test subsets), while the other two datasets were combined to create the training and validation subsets. In specific, the train subset was created by combining the 2 remaining train subsets plus the remaining 2 test subsets, while the validation subset was created with the 2 remaining validation subsets. This procedure was iterated for each dataset under test. This allows us to assess

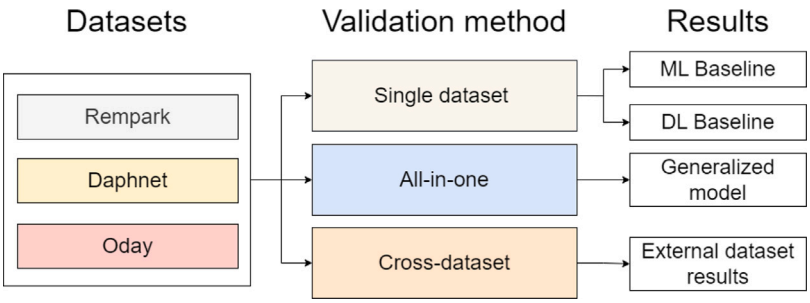


Fig. 5. Schematic representation of the datasets, validation methods, and results. ML: machine learning; DL: deep learning.

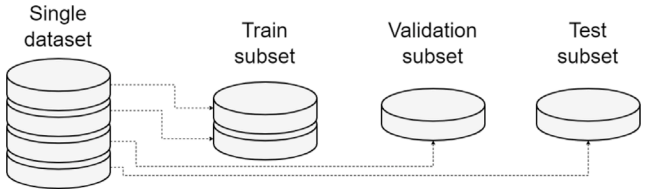


Fig. 6. Single dataset evaluation.

Table 3
Characteristics of the training, validation and test subsets in each dataset.

	Subset	Rempark	Daphnet	Oday
Number of subjects	Train	12	5	3
	Validation	4	3	2
	Test	5	2	2
Mean H&Y	Train	3.1	2.5	–
	Validation	2.8	2.5	–
	Test	3.2	2.6	–
Mean Age	Train	69.5	66.6	56.2
	Validation	66.5	63.7	63.5
	Test	72.2	70	57

how effectively models, trained on diverse datasets, can extend their generalization to a new and entirely unknown dataset.

3.6. Hyperparameter optimization

Training DNNs is an iterative process that aims to find a satisfactory solution to a given problem. During training, hyperparameters such as batch size, number of epochs, layer type and their parameters remain constant. Therefore, the precise optimization of these hyperparameters has a significant impact on the performance of the model and computational efficiency during training (Glorot & Bengio, 2010).

In this study, classification algorithms were subjected to hyperparameter optimization using the Hyperband method (Li, Jamieson, DeSalvo, Rostamizadeh, & Talwalkar, 2017). This method achieves an effective balance between performance and speed, particularly in problems characterized by high-dimensional space. It aims to efficiently allocate resources (i.e., computation time, number of iterations) to different hyperparameter configurations, with the goal of finding the best-performing configuration within a limited budget. A successive halving strategy is used to allocate resources. Specifically, the algorithm starts by training all configurations for a fixed number of iterations. Then, it discards the worst-performing half of the configurations and allocates additional resources (e.g., more iterations) to the remaining half. This process continues until only one configuration remains or the budget is exhausted. Given the multitude of experiments and configurations required for testing various data splitting strategies and evaluated DL models, this method allowed the implementation of an efficient optimization process.

Table 4
Range of values and steps used for the optimization of the machine and deep learning approaches. CNN: convolutional neural network; MLP: multi-layer perceptron; RF: random forest.

Architecture	Parameter	Range (Step)
CNN-MLP	learning rate	$1 \cdot 10^{-5}$ to $1 \cdot 10^{-1}$
	weight decay	$1 \cdot 10^{-6}$ to $1 \cdot 10^{-2}$
	batch size	[32, 64, 128, 256, 512, 1024]
ConvMixer	learning rate	$1 \cdot 10^{-5}$ to $1 \cdot 10^{-1}$
	weight decay	$1 \cdot 10^{-6}$ to $1 \cdot 10^{-2}$
	batch size	[32, 64, 128, 256, 512, 1024]
	patch size	[4, 5, 16, 32]
	ConvMixer blocks	1 to 5 (step=1)
	kernel size	[3, 5, 7, 9, 10]
	number of filters	[8, 16, 32, 64, 128]
Wide-CNN	learning rate	$1 \cdot 10^{-6}$ to $1 \cdot 10^{-1}$
	weight decay	$1 \cdot 10^{-6}$ to $1 \cdot 10^{-2}$
	batch size	[32, 64, 128, 256, 512, 1024]
RF	number of estimators	40 to 80 (step=10)
	max depth	[5, 6, 7, 10]
	max features	[5, 7, 9, 10]

As part of this optimization process, an emphasis was placed on fine-tuning the parameters of learning rate, the weight decay and batch size. The data used for the optimization processes correspond to the test and validation subsets that were generated for each validation experiment. In specific, for the CNN-MLP architecture, the learning rate, weight decay and batch size were fine-tuned, while the type of layers and layer parameters were kept as in the original study (Bikias et al., 2021). For the ConvMixer, the learning rate, weight decay, the number of ConvMixer blocks and their parameters, including the number of filters and kernel size, were fine-tuned. For the Wide-CNN, the learning rate, weight decay and batch size were tuned, while the parameters used within the CNN layers in the three branches were kept as in the original publication (Borzì et al., 2023). For the RF algorithm, the number of estimators, maximum depth, and maximum number of features were optimized. A minimum sample split of 2, minimum sample leaf of 1, and a split criterion based on the Gini impurity were set. The range of values used for parameter optimization are shown in Table 4. The parameter range were selected considering studies focusing on developing FOG detection algorithms (Borzì & Sigcha & Olmo, 2023; Camps et al., 2018; Sigcha et al., 2022).

3.7. Training settings

DL models were trained using the backpropagation algorithm and the adaptive moment estimation with weight decay (AdamW) optimizer (Loshchilov & Hutter, 2017). The training protocol used the binary cross-entropy loss function, a batch size ranging from 512 to 1024 according to the experiment, and maximum number of epochs of 300.

In addition, an early stopping strategy was used that terminates training when the validation loss shows no improvement over a span

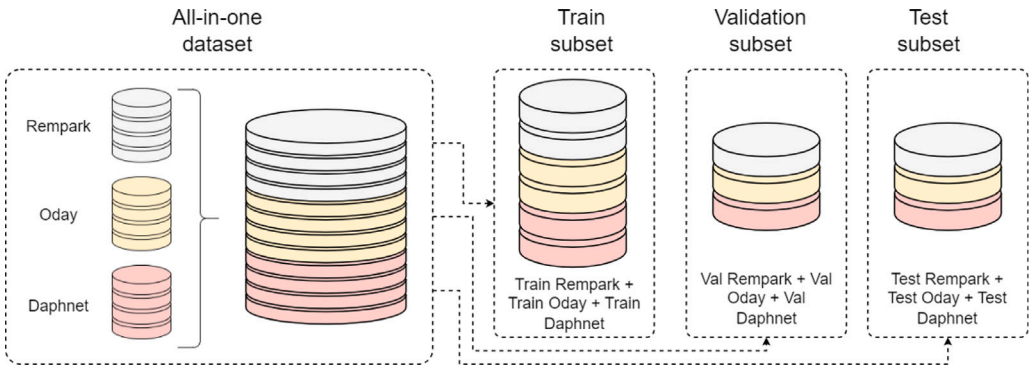


Fig. 7. All-in-one data splitting. Val: Validation subset.

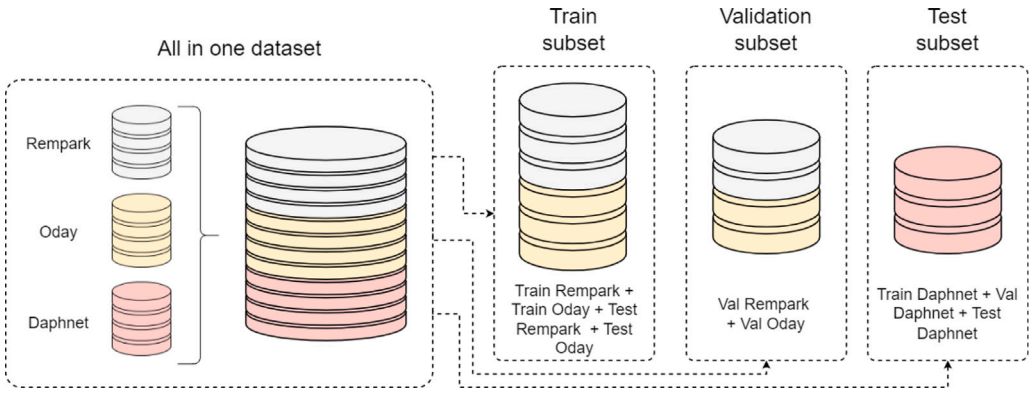


Fig. 8. Cross-dataset evaluation. Val: Validation subset.

of 10 consecutive epochs. This prevents over-fitting and minimizes unnecessary computational overhead (Shen, Gao, & Ma, 2022).

Tables 12–14 (Appendix) provide a summary of the hyperparameters utilized in training of the different ML and DL approaches and experiments.

3.8. Performance evaluation

To calculate and evaluate classification performance, the following metrics are defined. True positives (TP) are true samples (FoG) correctly identified by the model. False positives (FP) represent negative samples (non-FoG) incorrectly predicted as positive. False negatives (FN) correspond to positive samples that were not detected by the model. Finally, true negatives (TN) represent correctly classified negative instances. Sensitivity (Eq. (1)), also known as true positive rate or recall, measures the proportion of actual positive instances (FoG windows) that are correctly identified by the detection algorithm. High sensitivity indicates that the algorithm can effectively identify most of the actual FoG events. Specificity (Eq. (2)) measures the proportion of actual negative instances (non-FoG windows) that are correctly identified as negative by the detection algorithm. High specificity indicates that the algorithm can effectively distinguish between FoG and normal activities. Precision (Eq. (3)) measures the proportion of detected positive instances (identified as FoG) that are indeed positive events. High precision indicates a high confidence in detecting FoG. It is worth noting that the precision metric is often not reported in studies related to FoG detection. However, it is used in this study to assess the false alarms rate. Finally, the area under the curve (AUC) measures the ability of a classifier to distinguish between classes and is used as a summary of the receiver operating characteristic (ROC) curve. While the AUC is independent of the classification threshold (Hanley & McNeil, 1982), sensitivity, specificity, and precision highly depend on this. The threshold was selected according to the minimum equal error rate (EER) calculated on the training set. Then, it was applied to the validation and test set.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

4. Results and discussion

This section reports and discusses the experiments and results obtained with the different evaluation methods and algorithmic approaches. All experiments were repeated six times to mitigate variability due to stochastic processes in the training procedure. The reported results correspond to the mean values over multiple iterations.

4.1. Single dataset evaluation

Table 5 reports a summary of the FoG detection performance of the DL and ML models. The results refer to the test set. More detailed results obtained in the train, validation and test subsets are reported in Tables 8–11 (Appendix). All models provided the best performance on the Rempark dataset, followed by the Daphnet and Oday datasets. This reflects the dataset size, in terms of total amount of data, number of subjects, and number of FoG episodes. As expected, a large amount of data and FoG instances seems to promote performance and generalization capability.

From Table 5, it is evident that the DL models performed better than the ML approach in all datasets. Specifically, an improvement in AUC of 8.8–11.2%, 8.4–10.0%, and 3.1–5.6% was observed on the Rempark, Oday, and Daphnet datasets, respectively. This indicates that automatically extracted features from the CNN models perform

Table 5
Classification performance of different models in the different datasets. The results refer to the test set. CNN: convolutional neural network; MLP: multi-layer perceptron; RF: random forest; AUC: area under the curve of the receiver operating characteristic.

Model	Dataset	Sensitivity	Specificity	Precision	AUC
CNN-MLP	Rempark	0.836	0.883	0.566	0.934
	Oday	0.836	0.509	0.301	0.770
	Daphnet	0.632	0.949	0.602	**0.852**
ConvMixer	Rempark	0.798	0.865	0.519	0.914
	Oday	0.744	0.670	0.366	0.785
	Daphnet	0.599	0.933	0.521	0.827
Wide-CNN	Rempark	0.829	0.899	0.598	**0.938**
	Oday	0.848	0.557	0.330	**0.786**
	Daphnet	0.587	0.955	0.616	0.844
RF	Rempark	0.923	0.729	0.383	0.826
	Oday	0.711	0.662	0.346	0.686
	Daphnet	0.799	0.793	0.346	0.796

Table 6
Classification performance of different models when evaluated on the combination of all datasets. Results are reported separately for the train, validation and test sets. CNN: convolutional neural network; MLP: multi-layer perceptron; RF: random forest; AUC: area under the curve of the receiver operating characteristic.

Model	Subset	Sensitivity	Specificity	Precision	AUC
CNN-MLP	Train	0.827	0.827	0.385	0.911
	Validation	0.788	0.788	0.217	0.874
	Test	0.765	0.874	0.513	0.907
ConvMixer	Train	0.812	0.812	0.361	0.897
	Validation	0.797	0.797	0.226	0.880
	Test	0.743	0.805	0.401	0.872
Wide-CNN	Train	0.861	0.861	0.449	**0.935**
	Validation	0.810	0.810	0.241	**0.892**
	Test	0.759	**0.891**	**0.548**	**0.912**
RF	Train	0.827	0.827	0.386	0.914
	Validation	0.800	0.799	0.229	0.876
	Test	**0.870**	0.727	0.361	0.798

Table 7
Cross-dataset evaluation with an independent test subset. AUC: area under the receiver operating characteristic.

Test data	Subset	Sensitivity	Specificity	Precision	AUC
Rempark	Train	0.879	0.879	0.510	0.947
	Validation	0.828	0.828	0.300	0.908
	Test	0.532	0.871	0.315	0.829
ODAY	Train	0.890	0.890	0.516	0.955
	Validation	0.859	0.859	0.283	0.928
	Test	0.129	0.921	0.362	0.654
Daphnet	Train	0.877	0.877	0.503	0.947
	Validation	0.833	0.833	0.307	0.911
	Test	0.522	0.881	0.320	0.839

better than hand-crafted features used to feed the classic ML model. Although RF is one of the best choices among ML models for FoG detection (Pardoel et al., 2019), the DL networks outperformed the RF algorithm.

The results reported in Table 5 are consistent with those reported in related studies. As far as concerns the Rempark dataset, accuracy of 0.89 was obtained using a CNN (Camps et al., 2018). AUC of 0.94 was achieved using a combination of CNN and LSTM (Sigcha et al., 2020), 0.95 with a CNN (Borzì et al., 2023), and 0.96 with a Transformer-CNN (Sigcha et al., 2022). However, in the latter study, a LOSO validation was used.

As for the Oday dataset, an AUC of 0.75 was obtained when using a single sensor placed on the lower back, increasing up to 0.83 when combined with two additional sensors on the ankles (O'Day et al., 2022).

Regarding the Daphnet dataset, the performance varied based on the number of sensors included in the analysis. When using the three available sensors, an accuracy of 0.83 was obtained with an LSTM network (Ashour et al., 2020), 0.92 with a combination of CNN and LSTM (Li et al., 2020), 0.93 with a CNN (San-Segundo et al., 2019), and an AUC of 0.77 with a convolutional denoising autoencoder (CDA) (Mohammadian Rad, Van Laarhoven, Furlanello, & Marchiori, 2018). Furthermore, an accuracy of 0.79 was achieved using a single inertial sensor placed on the thigh by using a CDA (Noor et al., 2021).

Overall, the results suggest that the implemented DL models align with state-of-the-art approaches for FoG detection, with similar performance and generalization capability (see Appendix).

Finally, the Wide-CNN model provided the best results in two out of three datasets (Rempark and Oday), while the CNN-MLP model outperformed the other DL algorithms when evaluated on the Daphnet dataset.

4.2. All-in-one dataset evaluation

Table 6 reports the FoG detection performance of different DL and ML algorithms using the all-in-one evaluation strategy. The best results were obtained using the Wide-CNN model, which provided the best AUC in the train, validation, and test subsets. In addition, the Wide-CNN outperformed the other models in three out of four metrics (specificity, precision, and AUC), while the best sensitivity was obtained using the RF model.

The comparison of Tables 5 and 6 allows for assessing the effect of merging different datasets. As for the classic ML model, the performance recorded when merging all datasets (AUC of 0.798) is in line with that obtained in the Daphnet dataset (AUC of 0.796), superior to that recorded in the Oday dataset (AUC of 0.686), and lower than the results obtained in the Rempark dataset (AUC 0.826). As for the DL models, the AUC in the range 0.872–0.912 (test results in Table 6) is significantly better than the best results obtained in the Oday and Daphnet datasets (Table 5). However, performance is lower than that registered in the Rempark dataset. The latter represents the largest dataset in terms of the number of subjects and number of FoG episodes. In addition, it is the only one collected in the home environment. Based on the results presented in Table 6, subsequent experiments were performed only using the WideCNN model.

4.3. Cross-dataset evaluation

Table 7 reports the FoG detection performance of the Wide-CNN model when training iteratively on two datasets and testing on the remaining one. According to Table 7, it is evident that performance is impaired in all cases, with AUC in the range 0.654–0.839 for the test subsets. By comparing the results of the train, validation, and test subsets, a clear performance gap is observed moving from the validation to the test set.

Specifically, a significant decrease in AUC of 7.9%, 27.4%, and 7.2% is observed when comparing the test and validation subsets for the Rempark, Oday, and Daphnet dataset, respectively. Moreover, sensitivity represents the most affected metric, showing substantial impairment. The worst results were obtained when training the model on the Rempark and Daphnet datasets and testing on the Oday dataset. This is somehow surprising, as the former datasets comprise 31 PwPD manifesting a total of 1347 FoG episodes. Moreover, they include different activities and walking tasks, properly defined to represent common ADL and generate possible false-positive events. On the contrary, the Oday dataset includes only 7 subjects and pre-defined gait tasks. The results of the cross-dataset evaluation on the Rempark and Daphnet datasets are similar in terms of AUC (0.829 and 0.839). This suggests that the two datasets are somehow similar, in terms of the contribution of gait patterns, activities, and FoG manifestations.

From the comparison of Tables 5 and 7 it is evident that performance is impaired when testing DL models on datasets different

from those used in the training and validation stage. These results are very relevant, as it seems that DL models showing good performance on an adequate dataset, do not ensure similar performance on a different dataset. It is worth noting that the wearable devices used for data collection were different in the three datasets. However, technical requirements (e.g. full-scale, sensitivity, sampling frequency) were set to ensure proper data representation during ADL and FoG. In addition, the different datasets underwent a pre-processing stage aimed at uniforming signal scale and sampling frequency. On the other hand, the sensor location is slightly different in the datasets, but this should not significantly affect the performance of robust DL algorithms. Additionally, the sensor orientation was adjusted to ensure uniformity among datasets. In a previous work (Borzì et al., 2023), we trained a DL model similar to the Wide-CNN using the Rempark dataset, and we tested on a dataset (Borzì & Olmo & Artusi & Lopiano, 2020) including 38 PwPD that manifested a total of 52 FoG episodes. Despite the difference in the type of wearable device, its position and orientation, the algorithm exhibited similar performance in terms of AUC, while a reduction in sensitivity and an improvement in specificity were observed.

In recent decades, a lot of effort has been devoted to the development of accurate FoG detection algorithms. However, most of the studies collected proprietary datasets, using different prototype or commercial wearable inertial sensors and specific technical data acquisition settings. Some FoG datasets are available to the research community (Bächlin et al., 2010; Mazilu et al., 2016; O'Day et al., 2022; Rodríguez-Martín et al., 2017), however, they were collected using different sensor settings. The results of this study suggest that some kind of agreement is necessary to provide uniform data. The indication of sensor settings (i.e., range, sensitivity, sampling frequency), along with pre-defined sensor locations can help create homogeneous datasets in terms of data acquisition, but still including different subjects monitored in different environments.

In contrast to classical ML processing pipelines, data-driven DL models have demonstrated excellent performance on diverse datasets. However, the ability to generalize to external datasets is an important limitation to be considered with caution. Regardless of the training dataset, testing on a different dataset leads to an evident reduction in performance. Moreover, even with a small dataset, augmenting the training set with samples from different datasets does not lead to improved performance in the test set. This is surprising, as DL models are expected to exploit the size and heterogeneity of the dataset to provide more robust results. Finally, it is worth considering the precision metric, which represents the false alarm rate. Although it is desirable for most FoG episodes to be detected (high sensitivity), an excessive number of false positives (low precision) makes the algorithm inapplicable in real-world practice. Therefore, the trade-off between sensitivity and precision should be carefully selected when developing wearable applications oriented to real-life scenarios.

4.4. FoG manifestation in different datasets

To further investigate the possible differences in FoG recorded in the different datasets, analysis of variance (ANOVA) tests were performed on each extracted feature in the Mazilu feature set. Fig. 9 shows the histogram of some features in the three datasets. As can be seen, the distribution of feature values such as standard deviation, entropy, energy, and power is different among the datasets ($p < 0.001$).

Differences were observed in both time and frequency domains, suggesting different FoG characteristics in each dataset. This may be due to the great heterogeneity of FoG manifestation, which is influenced by intra- and inter-subject variability. It is worth noting that the difference between the Daphnet and Oday datasets is most obvious, while the Rempark dataset lies between them. This may be due to the larger number of subjects, greater heterogeneity in task performance, and higher number of FoG episodes recorded in the Rempark, compared to the other datasets. The large number of FoG windows collected from each dataset (more than a thousand) lends statistical significance to the results. Because subjects are independent in each dataset and the sample size is small (maximum 21 subjects), it appears that each dataset was not able to capture sufficient variability in FoG manifestation to ensure adequate generalization to a larger population.

To further assess inter-subject variability, the distribution of the freezing index (FI) values was compared between subjects included in the same dataset. The FI represents the most common and well-known FoG characteristic that allows to distinguish FoG from normal gait. Fig. 10 shows the histogram of FI values across multiple datasets. For each dataset, the distribution for each subject is represented in different colors.

As evident, there is inter-subject variability, with distributions for some subjects that are significantly different from the others. This further confirms that FoG manifestations are subject-dependent. It is worth noting that the distributions are similar in the Rempark dataset (Fig. 10(b)). This may be due to the large number of FoG episodes collected from each subject. On the other hand, some subjects from the Daphnet (Fig. 10(a)) and Oday dataset (Fig. 10(c)) differ significantly from the others. Moreover, the distribution of some subjects from the Oday dataset are not represented in the other datasets, and this may hinder adequate cross-dataset accuracy.

Shapley additive explanations (SHAP) analysis (Lundberg & Lee, 2017) was performed to better understand the reduction in performance observed in cross-dataset evaluation experiments. SHAP analysis allows for a consistent and objective explanation of the impact of each feature on model prediction. The RF model was considered and SHAP values were calculated on the test set.

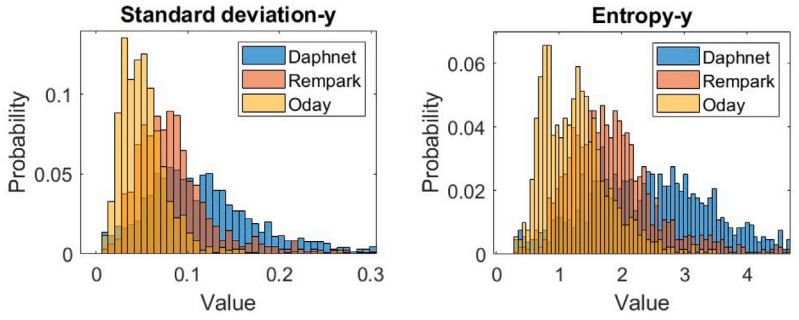
Detailed SHAP plots for each dataset can be found in Figs. 12,13,14 (Appendix). The results are summarized in Fig. 11, where the average absolute SHAP value of different features in each dataset are reported. Results refer to the test set, and features with low contribution in all datasets were not reported.

As expected, the freezing index proves to be the most significant feature in FoG detection. In general, spectral features have a larger contribution compared to temporal characteristics. In addition, features related to amplitude, intensity, and regularity of the acceleration signal were extracted from the x-axis (vertical acceleration). On the other hand, the freezing index computed from the y-axis (anterior–posterior acceleration) shows the strongest impact on the model output. However, significant heterogeneity is evident, with some features having a strong impact on some datasets but not on others.

In more detail, the sum of spectral power along the x-axis has a high contribution in the Daphnet and Rempark datasets, but not in Oday. The standard deviation along the x-axis has a significant impact in the Daphnet and Oday datasets, but not in Rempark. The freezing index along the z-axis has a significant impact on the Daphnet and Oday datasets, but not on Rempark. Overall, these differences in individual features help to explain why ML models trained on one dataset cannot generalize well to the other datasets.

4.5. Limitations

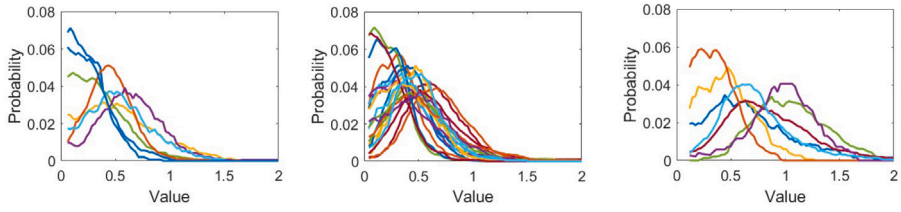
This study has some limitations that provide directions for future research. First, the performance and generalization capability of statistical and mathematical models that do not require ML were not considered. Such approaches showed lower performance than the ML and DL approaches (Giannakopoulou et al., 2022; Mancini et al., 2019; Pardoel et al., 2019). However, despite performing worse on a single dataset, they may show a more consistent performance in cross-dataset tests.



(a) Standard deviation along the y-axis. (b) Shannon entropy along the y-axis.

(c) Signal energy along the y-axis. (d) Signal power along the y-axis.

Fig. 9. Histogram of feature values in different datasets.



(a) Daphnet (b) Rempark (c) Oday

Fig. 10. Histogram of freezing index values for each subject in the different datasets.

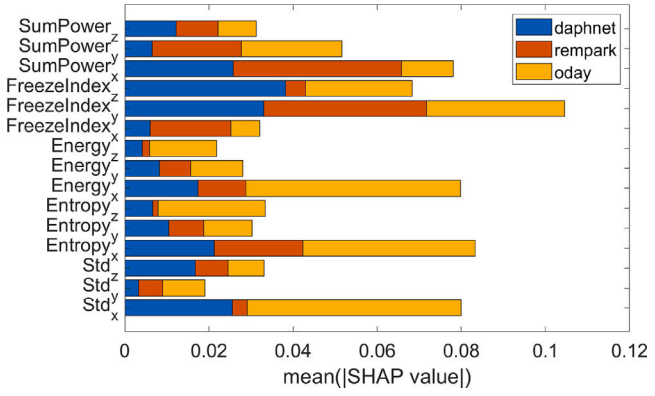


Fig. 11. Average shapley additive explanations (SHAP) values of different features in each dataset. Values represent the average contribution on the model output.

Despite the large number of FoG episodes (>200) recorded in each dataset, all datasets are relatively small and include 7 to 21 subjects. This might prevent the models from learning meaningful representations of the data that can be robust in tests on multiple datasets.

The datasets used were different in terms of recording device, environment, and experimental procedures. Cross-tests performed among datasets collected with the same device under different conditions could provide additional insights for the development of a generalizable method.

Finally, only one wearable device was considered, positioned near the center of mass of the body. It is not entirely clear whether similar results can be obtained using lower-limb sensors, which better capture leg dynamics during gait and FoG.

5. Conclusions

This study aims to evaluate the generalization capability of DL algorithms in FoG detection. For this purpose, different FoG datasets were used. In addition, the performance of DL models was compared with that of classical ML approaches, which involved the extraction of hand-crafted features. DL models showed better results than shallow ML algorithms, both in terms of performance on a single dataset and in terms of generalization capability. However, all approaches did not perform robustly in cross-dataset tests. Despite using the same type of sensor and similar location on the body, the models trained on one dataset failed to generalize well to other datasets. The insights provided by inter-dataset and inter-subject evaluation suggest that the peculiar characteristics of FoG differ significantly in different datasets. Moreover, within each dataset, the FoG pattern of some subjects is very different from that of others. From the model's perspective, the contribution of each feature to the final prediction varies significantly across

datasets, and this is reflected in a clear reduction in performance in cross-dataset tests. All together, these results suggest that personalized models (i.e., models specifically trained on each subject) or fine-tuning strategies are needed that adapt to specific FoG patterns and enhance the recognition capability. Future works should evaluate the generalization capability of other DL algorithms. In addition, more complex optimization methods (e.g., evolutionary and metaheuristics optimizers (Abdel-Basset, Abdel-Fatah, & Sangaiah, 2018)) can be exploited to further improve performance and robustness of DL models (Revin, Potemkin, Balabanov, & Nikitin, 2023). Finally, future studies should further investigate the effect of different wearable devices used for data collection. Other datasets should be collected using the same recording device under different conditions and environments. These can further demonstrate the true generalization capability of ML and DL algorithms and can contribute to the development of generalized methodologies for FoG detection strategies, useful for both research and clinical applications.

CRediT authorship contribution statement

Luis Sigcha: Resources, Conceptualization, Supervision, Methodology, Formal analysis, Investigation, Project administration, Writing – original draft. **Luigi Borzì:** Resources, Conceptualization, Supervision, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft. **Gabriella Olmo:** Investigation, Validation, Formal analysis, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The Daphnet freezing-of-gait dataset is available at this link. The imu-fog-detection (Oday) dataset is available at this link. The Rempark dataset belongs to the Technical Research Centre for Dependency Care and Autonomous Living (CETpD), Universitat Politecnica de Catalunya. The data were collected in the Rempark project and are available under reasonable request from the corresponding owners. The original multi-head convolutional neural network architecture is available at this link.

Acknowledgments

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, with particular reference to the partnership on the "Research and innovation on future telecommunications systems and networks, to make Italy more smart (RESTART)" program (PE00000001). The authors acknowledge to the Physical Education and Sports Science (PESS) department, the Health Research Institute (HRI), and the Data-Driven Computer Engineering (D2iCE) Group at University of Limerick.

Appendix

See Tables 8–14 and Figs. 12–14.

Table 8
Classification performance of the CNN-MLP network in single dataset experiments.

Dataset	Subset	Sensitivity	Specificity	Precision	AUC
Rempark	Train	0.876	0.876	0.442	0.949
	Validation	0.870	0.870	0.336	0.946
	Test	0.836	0.883	0.566	0.934
Oday	Train	0.776	0.777	0.663	0.839
	Validation	0.736	0.736	0.372	0.809
	Test	0.836	0.509	0.301	0.770
Daphnet	Train	0.857	0.857	0.476	0.933
	Validation	0.790	0.790	0.125	0.859
	Test	0.632	0.949	0.602	0.852

Table 9
Classification performance of the ConvMixer network in single dataset experiments.

Dataset	Subset	Sensitivity	Specificity	Precision	AUC
Rempark	Train	0.869	0.869	0.428	0.945
	Validation	0.871	0.871	0.340	0.943
	Test	0.798	0.865	0.519	0.914
Oday	Train	0.903	0.903	0.841	0.965
	Validation	0.710	0.711	0.343	0.780
	Test	0.744	0.670	0.366	0.785
Daphnet	Train	0.899	0.899	0.574	0.964
	Validation	0.820	0.820	0.148	0.894
	Test	0.599	0.933	0.521	0.827

Table 10
Classification performance of the Wide-CNN network in single dataset experiments.

Dataset	Subset	Sensitivity	Specificity	Precision	AUC
Rempark	Train	0.888	0.888	0.471	0.955
	Validation	0.869	0.869	0.335	0.944
	Test	0.829	0.899	0.598	0.938
Oday	Train	0.837	0.838	0.745	0.909
	Validation	0.719	0.719	0.352	0.788
	Test	0.848	0.557	0.330	0.786
Daphnet	Train	0.868	0.868	0.501	0.940
	Validation	0.804	0.805	0.136	0.883
	Test	0.587	0.955	0.616	0.844

Table 11
Classification performance of the RF algorithm in single dataset experiments.

Dataset	Subset	Sensitivity	Specificity	Precision	AUC
Rempark	Train	0.827	0.827	0.350	0.911
	Validation	0.812	0.812	0.247	0.893
	Test	0.923	0.729	0.383	0.826
Oday	Train	0.867	0.867	0.786	0.947
	Validation	0.611	0.609	0.249	0.641
	Test	0.711	0.662	0.346	0.686
Daphnet	Train	0.828	0.827	0.421	0.914
	Validation	0.749	0.749	0.101	0.815
	Test	0.799	0.793	0.346	0.796

Table 12
List of hyperparameters used to train machine and deep learning models in the single dataset experiments. CNN: convolutional neural network; MLP: multi-layer perceptron; RF: Random Forest.

Model	Dataset	Learning rate	Weight decay	Batch size	Other parameters
CNN-MLP	Rempark	$9 \cdot 10^{-3}$	$9 \cdot 10^{-4}$	512	
	Oday	$4 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	64	
	Daphnet	$7 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	512	
ConvMixer	Rempark	$9 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	128	patch size: 4
	Oday	$9 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	32	ConvMixer blocks: 2
	Daphnet	$8 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	32	filters (size): 128 (9)
Wide-CNN	Rempark	$7 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	512	
	Oday	$8 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	64	
	Daphnet	$7 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	1024	
RF	Rempark				n. estimators: 60
	Oday				maximum depth: 10
	Daphnet				maximum n. features: 7

Table 13
List of hyperparameters used to train machine and deep learning models for the all-in-one dataset experiments. CNN: convolutional neural network; MLP: multi-layer perceptron.

Model	Learning rate	Weight decay	Batch size	Other parameters
CNN-MLP	$1 \cdot 10^{-2}$	$3 \cdot 10^{-3}$	512	
Convmixer	$9 \cdot 10^{-3}$	$8 \cdot 10^{-4}$	512	patch size: 4 ConvMixer blocks: 2 filters (size): 128 (9)
Wide CNN	$1 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	256	
RF				n. estimators: 60 maximum depth: 10 maximum n. features: 7

Table 14
List of hyperparameters used to train the deep learning model for the cross-dataset experiments. CNN: convolutional neural network.

Model	Dataset	Learning rate	Weight decay	Batch size
Wide-CNN	Rempark	$7 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	256
	Oday	$6 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	256
	Daphnet	$7 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	1024

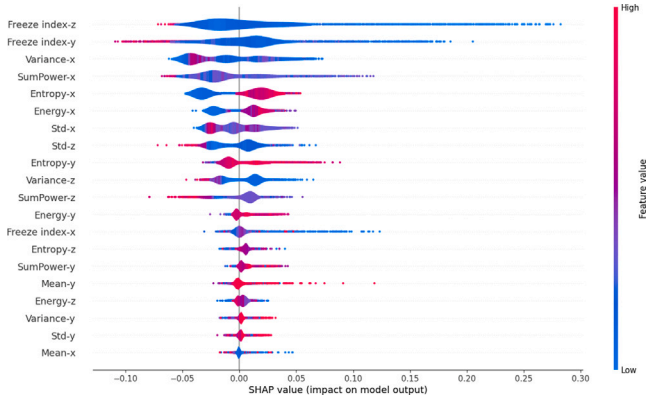


Fig. 12. Shapley additive explanations (SHAP) values of different features in the Daphnet dataset. Features are sorted in order of importance from top to bottom. Red and blue points indicate high and low feature values, respectively. Positive and negative SHAP values indicate positive (FoG) and negative (non-FoG) contribution on the model output.

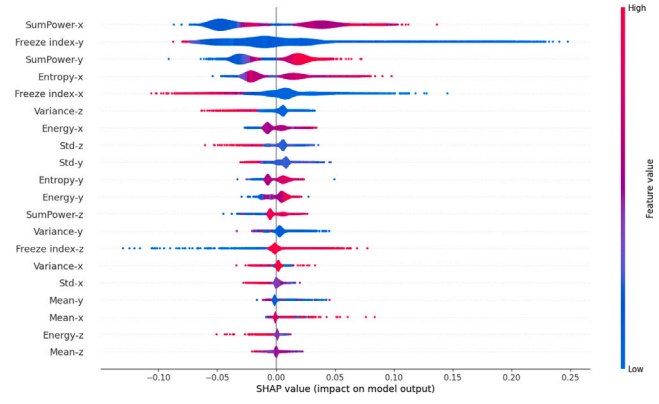


Fig. 13. Shapley additive explanations (SHAP) values of different features in the Rempark dataset. Features are sorted in order of importance from top to bottom. Red and blue points indicate high and low feature values, respectively. Positive and negative SHAP values indicate positive (FoG) and negative (non-FoG) contribution on the model output.

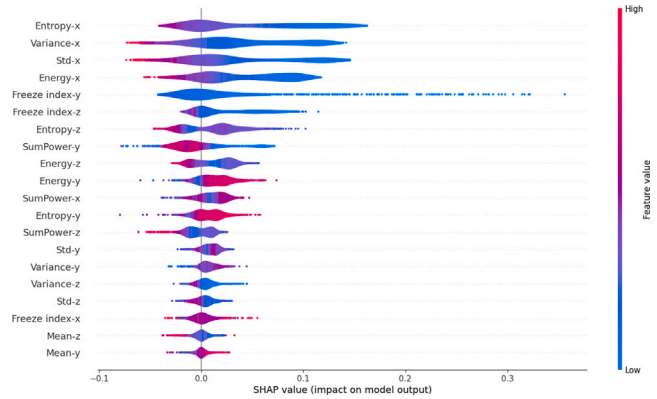


Fig. 14. Shapley additive explanations (SHAP) values of different features in the Oday dataset. Features are sorted in order of importance from top to bottom. Red and blue points indicate high and low feature values, respectively. Positive and negative SHAP values indicate positive (FoG) and negative (non-FoG) contribution on the model output.

References

Abdel-Basset, M., Abdel-Fatah, L., & Sangaiah, A. K. (2018). Chapter 10 - metaheuristic algorithms: A comprehensive review. In A. K. Sangaiah, M. Sheng, & Z. Zhang (Eds.), *Intelligent data-centric systems*, *Computational intelligence for multimedia big data on the cloud with engineering applications* (pp. 185–231). Academic Press, http://dx.doi.org/10.1016/B978-0-12-813314-9.00010-4.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. http://dx.doi.org/10.1186/s40537-021-00444-8.

Armstrong, M. J., & Okun, M. S. (2020). Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA*, *323*(6), 548–560. http://dx.doi.org/10.1001/jama.2019.22360.

Ashfaque Mostafa, T., Soltaninejad, S., McIsaac, T. L., & Cheng, I. (2021). A comparative study of time frequency representation techniques for freeze of gait detection and prediction. *Sensors*, *21*(19), 6446. http://dx.doi.org/10.3390/s21196446.

Ashour, A. S., El-Attar, A., Dey, N., El-Kader, H. A., & Abd El-Naby, M. M. (2020). Long short term memory based patient-dependent model for FOG detection in parkinson's disease. *Pattern Recognition Letters*, *131*, 23–29. http://dx.doi.org/10.1016/j.patrec.2019.11.036.

Bächlin, M., Plotnik, M., Roggen, D., Maidan, I., J.M., H., Giladi, N., et al. (2010). Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 436–446. http://dx.doi.org/10.1109/TITB.2009.2036165.

Bhidayasiri, R., & Martinez-Martin, P. (2017). Clinical assessments in Parkinson's disease: Scales and monitoring. *International Review of Neurobiology*, *132*, 129–182. http://dx.doi.org/10.1016/bs.irn.2017.01.001.

Bikias, T., Iakovakis, D., Hadjidimitriou, S., Charisis, V., & Hadjileontiadis, L. (2021). DeepFoG: An IMU-Based Detection of Freezing of Gait Episodes in Parkinson's Disease Patients via Deep Learning. *Frontiers in Robotics and AI*, *8*, Article 537384. http://dx.doi.org/10.3389/frobt.2021.537384.

Borzì, L., Olmo, G., Artusi, C., Fabbri, M., Rizzone, M., Romagnolo, A., et al. (2020). A new index to assess turning quality and postural stability in patients with Parkinson's disease. *Biomedical signal processing and control*, *62*, 1–7. http://dx.doi.org/10.1016/j.bspc.2020.102059.

Borzì, L., Olmo, G., Artusi, C., & Lopiano, L. (2020). Detection of freezing of gait in people with Parkinson's disease using smartphones. *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 625–635. http://dx.doi.org/10.1109/COMPSAC48688.2020.0-186.

Borzì, L., Sigcha, L., & Olmo, G. (2023). Context recognition algorithms for energy-efficient freezing-of-gait detection in Parkinson & rsquos disease. *Sensors*, *23*(9), http://dx.doi.org/10.3390/s23094426.

Borzì, L., Sigcha, L., Rodríguez-Martín, D., & Olmo, G. (2023). Real-time detection of freezing of gait in parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. *Artificial Intelligence in Medicine*, *135*, Article 102459. http://dx.doi.org/10.1016/j.artmed.2022.102459.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Camps, J., Sama, A., Martin, M., Rodriguez-Martin, D., Perez-Lopez, C., Arostegui, J., et al. (2018). Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge-Based Systems*, *139*, 119–131. http://dx.doi.org/10.5555/3163587.3163748.

Channa, A., Popescu, N., & Ciobanu, V. (2020). Wearable solutions for patients with parkinson's disease and neurocognitive disorder: A systematic review. *Sensors*, *20*(9), http://dx.doi.org/10.3390/s20092713.

Del Din, S., Kirk, C., Yarnall, A. J., Rochester, L., & Hausdorff, J. M. (2021). Body-Worn Sensors for Remote Monitoring of Parkinson's Disease Motor Symptoms: Vision, State of the Art, and Challenges Ahead. *Journal of Parkinsons Diseases*, *11*(s1), S35–S47.

Gao, C., Liu, J., Tan, Y., & Chen, S. (2020). Freezing of gait in Parkinson's disease: pathophysiology, risk factors and treatments. *Translational Neurodegeneration*, *9*, 12. http://dx.doi.org/10.1186/s40035-020-00191-5.

Giannakopoulou, K.-M., Roussaki, I., & Demestichas, K. (2022). Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. *Sensors*, *22*(5), http://dx.doi.org/10.3390/s22051799.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of machine learning research*: *9*, *International conference on artificial intelligence and statistics* (pp. 249–256). URL https://proceedings.mlr.press/v9/glorot10a.html.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.. *Radiology*, *143*(1), 29–36.

Huang, T., Li, M., & Huang, J. (2023). Recent trends in wearable device used to detect freezing of gait and falls in people with parkinson's disease: A systematic review. *Frontiers in aging neuroscience*, *15*, Article 1119956.

Jindong, W., Yiqiang, C., Shuji, H., Xiaohui, P., & Lisha, H. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11. http://dx.doi.org/10.1016/j.patrec.2018.02.010.

Keogh, A., Alcock, L., Brown, P., Buckley, E., Brozgol, M., Gazit, E., et al. (2023). Acceptability of wearable devices for measuring mobility remotely: Observations from the Mobilise-D technical validation study. *Digital Health*, *9*, http://dx.doi.org/10.1177/20552076221150745.

Kim, H. B., Lee, H. J., Lee, W. W., Kim, S. K., Jeon, H. S., Park, H. Y., et al. (2018). Validation of freezing-of-gait monitoring using smartphone. *Telemedicine and e-Health*, *24*(11), 899–907. http://dx.doi.org/10.1089/tmj.2017.0215.

Klaver, E. C., Heijink, I. B., Silvestri, G., van Vugt, J. P., Nonnekes, J., & Tjepkema-Cloostermans, M. C. (2023). Comparison of state-of-the-art deep learning architectures for detection of freezing of gait in parkinson's disease. *Frontiers in neurology*, *14*, Article 1306129.

Kobylecki, C. (2020). Update on the diagnosis and management of Parkinson's disease. *Clinical Medicine Journal*, *20*(4), 393–398. http://dx.doi.org/10.7861/clinmed.2020-0220.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. http://dx.doi.org/10.1038/nature14539.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *18*(1), 6765–6816.

Li, B., Yao, Z., Wang, J., Wang, S., Yang, X., & Sun, Y. (2020). Improved deep learning technique to detect freezing of gait in parkinson's disease based on wearable sensors.. *Electronics*, *9*(11), 1919. http://dx.doi.org/10.3390/electronics9111919.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, arXiv:1705.07874.

Mancini, M., Bloem, B. R., Horak, F. B., Lewis, S. J., Nieuwboer, A., & Nonnekes, J. (2019). Clinical and methodological challenges for assessing freezing of gait: Future perspectives. *Movement Disorders*, *34*(6), 783–790. http://dx.doi.org/10.1002/mds.27709.

Mazilu, S., Blanke, U., Calatroni, A., Gazit, E., Hausdorff, J. M., & Tröster, G. (2016). The role of wrist-mounted inertial sensors in detecting gait freeze episodes in Parkinson's disease. *Pervasive and Mobile Computing*, *33*, 1–16. http://dx.doi.org/10.1016/j.pmcj.2015.12.007.

Mazilu, S., Hardegger, M., Zhu, Z., Roggen, D., Tröster, G., Plotnik, M., et al. (2012). Online detection of freezing of gait with smartphones and machine learning techniques. In *IEEE international conference on pervasive computing technologies for healthcare (pervasiveHealth) and workshops* (pp. 123–130). http://dx.doi.org/10.4108/icst.pervasivehealth.2012.248680.

Mohammadian Rad, N., Van Laarhoven, T., Furlanello, C., & Marchiori, E. (2018). Novelty detection using deep normative modeling for IMU-based abnormal movement monitoring in parkinson's disease and autism spectrum disorders. *Sensors*, *18*(10), http://dx.doi.org/10.3390/s18103533.

Moore, S., MacDougall, H., & W.G., O. (2008). Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of Neuroscience Methods*, *167*(2), 340–348. http://dx.doi.org/10.1016/j.jneumeth.2007.08.023.

Naghavi, N., & Wade, E. (2022). Towards real-time prediction of freezing of gait in patients with parkinson's disease: A novel deep one-class classifier. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1726–1736. http://dx.doi.org/10.1109/JBHI.2021.3103071.

Noor, M., Nazir, A., Wahab, M., & Ling, J. (2021). Detection of freezing of gait using unsupervised convolutional denoising autoencoder. *IEEE Access*, *9*(11), 115700–115709. http://dx.doi.org/10.1109/ACCESS.2021.3104975.

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. http://dx.doi.org/10.1016/S1474-4422(11)70143-0.

O'Day, J., Lee, M., Seagers, K., Hoffman, S., Jih-Schiff, A., Kidziński, L., et al. (2022). Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of NeuroEngineering Rehabil*, *19*(20), http://dx.doi.org/10.1186/s12984-022-00992-x.

Pardoel, S., Kofman, J., Nantel, J., & Lemaire, E. D. (2019). Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: A review. *Sensors (Switzerland)*, *19*(23), http://dx.doi.org/10.3390/s19235141.

Pepa, L., Capecci, M., Andrenelli, E., Ciabattoni, L., Spalazzi, L., & Ceravolo, M. G. (2020). A fuzzy logic system for the home assessment of freezing of gait in subjects with parkinsons disease. *Expert Systems with Applications*, *147*, Article 113197.

Reches, T., Dagan, M., Herman, T., Gazit, E., Gouskova, N. A., Giladi, N., et al. (2020). Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. *Sensors*, *20*(16), http://dx.doi.org/10.3390/s20164474.

Reich, S. G., & Savitt, J. M. (2019). Parkinson's Disease. *Medical Clinics of North America*, *103*(2), 337–350. http://dx.doi.org/10.1016/j.mcna.2018.10.014.

Revin, I., Potemkin, V. A., Balabanov, N. R., & Nikitin, N. O. (2023). Automated machine learning approach for time series classification pipelines using evolutionary optimization. *Knowledge-Based Systems*, *268*, Article 110483. http://dx.doi.org/10.1016/j.knosys.2023.110483.

Rodríguez-Martín, D., Pérez-López, C., Samà, A., Català, A., Moreno Arostegui, J., Cabestany, J., et al. (2017). Waist-Worn Inertial Measurement Unit for Long-Term Monitoring of Parkinson's Disease Patients. *Sensors*, *17*, 827. http://dx.doi.org/10.3390/s17040827.

Rovini, E., Maremmani, C., & Cavallo, F. (2017). How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review. *Front Neurosci*, *11*, 555.

Samii, A., Nutt, J., & Ransom, B. (2004). Parkinson's disease. *Lancet*, *363*(9423), 1783–9173. http://dx.doi.org/10.1016/S0140-6736(04)16305-8.

San-Segundo, R., Navarro-Hellín, H., Torres-Sánchez, R., Hodgins, J., & De la Torre, F. (2019). Increasing robustness in the detection of freezing of gait in Parkinson's disease. *Electronics*, *8*(2), http://dx.doi.org/10.3390/electronics8020119.

Shen, R., Gao, L., & Ma, Y.-A. (2022). On optimal early stopping: Over-informative versus under-informative parametrization. arXiv preprint arXiv:2202.09885.

Shi, B., Tay, A., Au, W. L., Tan, D. M. L., Chia, N. S. Y., & Yen, S. (2022). Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. *IEEE Transactions on Biomedical Engineering*, *69*(7), 2256–2267. http://dx.doi.org/10.1109/TBME.2022.3140258.

Sigcha, L., Borzì, L., Amato, F., Rechichi, I., Ramos-Romero, C., Cárdenas, A., et al. (2023). Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review. *Expert Systems with Applications*, *229*, Article 120541. http://dx.doi.org/10.1016/j.eswa.2023.120541.

Sigcha, L., Borzì, L., Pavón, I., Costa, N., Costa, S., Arezes, P., et al. (2022). Improvement of performance in freezing of gait detection in parkinson's disease using transformer networks and a single waist-worn triaxial accelerometer. *Engineering Applications of Artificial Intelligence*, *116*, Article 105482. http://dx.doi.org/10.1016/j.engappai.2022.105482.

Sigcha, L., Costa, N., Pavón, I., Costa, S., Arezes, P., López, J., et al. (2020). Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors. *Sensors*, *20*(7), 1895. http://dx.doi.org/10.3390/s20071895.

Trockman, A., & Kolter, J. Z. (2022). Patches are all you need?. arXiv preprint arXiv:2201.09792.

Zhang, W., Gao, C., Tan, Y., & Chen, S. (2021). Prevalence of freezing of gait in Parkinson's disease: a systematic review and meta-analysis. *Journal of Neurology*, *268*(11), 4138–4150. http://dx.doi.org/10.1007/s00415-021-10685-5.

Zhang, W., Sun, H., Huang, D., et al. (2024). Detection and prediction of freezing of gait with wearable sensors in parkinson's disease. *Neurological Sciences*, *45*, 431—-453. http://dx.doi.org/10.1007/s10072-023-07017-y.

Zhang, Y., Yan, W., Yao, Y., Ahmed, J., Tan, Y., & Gu, D. (2020). Prediction of Freezing of Gait in Patients With Parkinson's Disease by Identifying Impaired Gait Patterns. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(3), 591–600. http://dx.doi.org/10.1109/TNSRE.2020.2969649.

Zhao, N., Yang, Y., Zhang, L., Zhang, Q., Balbuena, L., Ungvari, G., et al. (2021). Quality of life in Parkinson's disease: A systematic review and meta-analysis of comparative studies. *CNS Neuroscience & Therapeutics*, *27*(3), 270–279. http://dx.doi.org/10.1111/cns.13549.