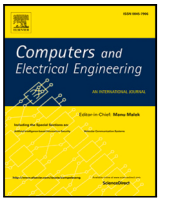
Computers and Electrical Engineering 127 (2025) 110530

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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



Edge-based freezing of gait recognition in Parkinson's disease

Luigi Borzì [a],*, Luis Sigcha [b], Farshad Firouzi [c], Gabriella Olmo [a], Florenc Demrozi [d]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b] Department of Physical Education and Sports Science, Health Research Institute, University of Limerick, V94 T9PX, Limerick, Ireland
[c] Department of Electrical and Computer Engineering, Duke University, NC, 27708, Durham, USA
[d] Department of Electrical Engineering and Computer Science, University of Stavanger, Rennebergstien 30, 4021 Stavanger, Norway

ARTICLE INFO

Keywords:
Parkinson's disease
Freezing of gait
Wearable sensors
Deep learning
Edge computing

ABSTRACT

Freezing of gait (FoG) stands as one of the most debilitating symptoms of Parkinson's disease (PD), occurring in more than half of patients with advanced PD. This condition manifests as a sudden blockage, significantly reducing the patients' quality of life. To improve gait and ameliorate FoG, cueing strategies involving audio, visual, or tactile stimulation have been evaluated. In particular, on-demand systems that can automatically detect FoG and administer cueing have emerged as promising solutions. In response, several wearable sensors and machine learning-based approaches have been proposed for accurate FoG recognition. However, existing techniques suffer from several critical challenges, notably suboptimal performance, and limitations for real-time operation and edge deployment. Addressing these issues, this study presents a groundbreaking advancement in real-time edge-based FoG recognition utilizing convolutional neural networks (CNN). We designed an optimized model, rigorously evaluating it across 62 PD patients using a cutting-edge reference dataset, achieving an F1-score of 92% and an area under the curve of 0.97. Further testing on an external dataset resulted in consistent detection performance, while a lower specificity was observed. The CNN implementation on a cost-effective processing device resulted in a 1 ms inference time and required only 6.3 KB of random access memory (RAM) and 37.8 Kb of Flash memory, meeting real-time demands and enhancing clinical applicability.

1. Introduction

Parkinson's disease (PD) is a complex and progressive neurodegenerative disorder, characterized by a wide spectrum of motor and non-motor symptoms, profoundly affecting the quality of life (QoL) of millions of people with PD (PwPD) worldwide [1]. Among motor symptoms, freezing of gait (FoG) is one of the most disabling symptoms, affecting more than half of PwPD, producing a higher risk of suffering more severe physical injuries [2]. FoG presents as an abrupt motor block, described as the feeling of having the feet glued to the ground [3]. It manifests in different forms, including trembling legs, shuffling steps, or total akinesia (i.e., loss of movement of the limbs or trunk) [4]. FoG episodes can vary in duration, ranging from very short episodes (1 s or less) to longer episodes exceeding 10 s [5]. The number and duration of FoG episodes increase as the effect of pharmacological therapy decreases [6]. In addition, certain activities and situations are known to trigger FoG manifestation, including turning, negotiating obstacles, experiencing stress, and engaging in both cognitive and motor dual tasks [7].

* Corresponding author.
E-mail addresses: luigi.borzi@polito.it (L. Borzì), luis.sigcha@ul.ie (L. Sigcha), farshad.firouzi@duke.edu (F. Firouzi), gabriella.olmo@polito.it (G. Olmo), florenc.demrozi@uis.no (F. Demrozi).

https://doi.org/10.1016/j.compeleceng.2025.110530
Received 26 January 2025; Received in revised form 28 April 2025; Accepted 7 June 2025

The current clinical evaluation of FoG is mainly based on questionnaires and controlled walking evaluations conducted in clinical and laboratory settings [7]. However, the subjectivity and qualitative assessment of FoG during brief and sporadic medical visits do not reflect the broad spectrum of FoG manifestations in daily life. FoG assessment represents a challenge mainly due to the heterogeneity of FoG manifestations, the fact that FoG events differ significantly among individuals, and the low rate of FoG events compared to activities of daily life under free-living conditions. Furthermore, variations in symptoms, disease progression, and medication responses among individuals in the heterogeneous PD population contribute to these differences, presenting challenges in establishing a standardized detection method [7]. Additionally, FoG episodes can occur under different pharmacological conditions, further complicating the identification and prediction of these events [6,8]. The combination of these multiple factors makes the accurate detection of FoG a formidable and intricate challenge.

On this wave, numerous technological solutions have emerged for the continuous and objective assessment of FoG, including optoelectronic systems and RGB cameras [9]. However, these systems are expensive, require considerable set-up time, and are limited to laboratory settings [10]. On the contrary, wearable devices that integrate inertial and/or physiological sensors have been widely used in the context of FoG recognition due to their small size, low cost, low power consumption, and stimulation capabilities [11,12]. The inherent advantages of wearables make them suitable for real-world scenarios, seamlessly integrating into daily activities [13]. Wearables can serve as walking assistance systems that improve FoG and gait in PD. External stimuli, including auditory, visual, and somatosensory stimuli, have been shown to be effective in improving gait and alleviating FoG [14]. However, the effectiveness of continuous cueing tends to decline over time, due to possible problems with habituation and compliance [15]. On the contrary, on-demand cueing approaches emerge as promising solutions, delivering stimulation only upon automatic detection of FoG. This strategy has been shown to be effective in reducing the duration of FoG episodes and may even prevent their onset if predicted before their occurrence [16]. Therefore, the critical need for accurate and timely recognition of FoG is emphasized to enable the implementation of targeted and responsive interventions.

In such direction, several wearable-based methods to recognize FoG have been proposed [12]. Early approaches used threshold-based methods on inertial data to identify increases in signal amplitude in the 3–8 Hz frequency band (freeze band) [17] or the computation of the freezing index, defined as the ratio between the power in the freezing band and that in the 0–3 Hz frequency band (locomotor band), that showed promising results in laboratory settings [18]. Subsequently, the advent of artificial intelligence (AI), particularly machine learning (ML), has improved FoG recognition accuracy using support vector machines (SVMs), k-nearest neighbors (k-NNs), and random forests (RFs) [19,20], achieving sensitivity and specificity in FoG detection up to 0.93 and 0.94, respectively [21]. Nevertheless, the performance of ML models is strongly influenced by the selection of relevant features from the recorded data [22,23].

On the other hand, deep learning (DL) methods have gained popularity due to their ability to automatically extract meaningful information from raw data, revolutionizing data processing pipelines [24,25]. For instance, convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and deep autoencoders have achieved remarkable performance in FoG recognition, with sensitivity and specificity reaching 0.92 and 0.98, respectively [12,26]. Although most of these solutions were evaluated in the laboratory, some incorporated simulated daily activities or unscripted tasks, such as random walking [27], vacuuming the floors and emptying a dishwasher [28], brushing teeth, painting/drawing/erasing, and cleaning windows [29]. However, very few studies have collected data in real scenarios and under unsupervised conditions, highlighting the complexity of accurate FoG detection in everyday life [30–32].

Although advances in ML/DL have played a key role in facilitating the development of FoG recognition systems, there is still a need to improve the computational efficiency of wearable systems, suboptimal performance, and overcome impracticality for real-time and edge applications. Furthermore, the paucity of available data on PwPD experiencing FoG has exacerbated the challenges in advancing the field of FoG detection methodologies [33,34].

2. Related work

Numerous studies have explored different methodologies and technologies to address the challenges associated with timely and accurate FoG detection. This section provides an overview of recent significant works that used DL-based approaches for FoG detection from wearable sensor data.

Yang et al. [35] equipped 12 PwPD with 5 inertial measurement units (IMUs), collecting a total of 530 FoG episodes. Raw data were input to a temporal CNN that provided an F-score of 0.47–0.72, depending on the tasks, medication state, and stops. The results evidenced that models trained on a specific task could not generalize to unseen tasks, while models trained on a specific medication state could generalize to unseen states. O'Day at al. [36] recorded inertial data from 7 PwPD using 6 IMUs on the feet, shanks, lower back and chest, collecting a total of 211 FoG events. Raw data were fed to a one-dimensional (1D) CNN. The best results (area under the receiver operating characteristic–AUROC of 0.83) were obtained using the combination of sensors on the ankles and lower back. Sun et al. [37] exploited data augmentation techniques (rotation, flip, addition of random noise) to balance the distribution of classes (FoG, non-FoG). A combination of manually selected features and deep features learned through a CNN was proposed, providing an F-score of 0.89 on a subject-dependent validation approach. Performance deteriorated when a general, subject-independent validation scheme was used. Huang et al. [38] equipped 12 PwPD with 5 IMUs on the ankles, legs, and lower back, registering 334 FoG events. Multiple consecutive 0.5s-windows were used as input to a DL model (FoG-Net), which exploited temporal information through attention mechanisms. The proposed model provided an F-score of 0.87 when processing the input data from six consecutive windows.

Shi et al. [39] equipped 63 PwPD with 2 ankle-mounted IMUs, recording 486 FoG episodes. Data were filtered, processed through the continuous wavelet transform (CWT), and input to a 2D-CNN, resulting in an F-score of 0.92. Klaver et al. [40] analyzed inertial data recorded using 7 IMUs on the lumbar region, upper legs, lower legs, and feet from a total of 80 PwPD enrolled across multiple studies, collecting a total of 1435 FoG episodes. Data were filtered and input to different DL architectures. The best results were obtained using a 1D-CNN, providing an AUROC of 0.72 on the hold-out set (31 subjects) and 0.90 on the external test set (10 subjects). The challenges associated with cross-dataset testing were further investigated by Sigcha et al. [41]. Different DL algorithms were trained and evaluated on three FoG datasets [18,36,42], all including accelerometer data recorded from the lower back. The results showed that although good results were obtained in each individual dataset (AUROC 0.79–0.94), performance declined when the models were trained and tested on different datasets (AUROC 0.65–0.82).

All of these studies contributed a novel approach to FoG detection and obtained promising results. In addition, they provided insights into the effect of sensor position, tasks, activities, and medication. However, results were evaluated on a small sample of 7–12 subjects [35–38], hindering comprehensive assessment of the generalizability of the findings. Some studies evaluated FoG recognition performance on larger datasets and performed cross-dataset testing [39–41]. However, none of them analyzed the computational complexity or evaluated the algorithm on a stand-alone wearable device. In some cases [37,39], the computational burden introduced by the data processing strategy (e.g., CWT computation, pre-processing, feature extraction, very deep neural network) hinders real-time edge applications.

Camps et al. [43] analyzed the Rempark dataset, which includes 1058 FoG episodes recorded at home on 21 PwPD using a waist-worn accelerometer. Data were processed using the fast Fourier transform (FFT), and two consecutive 2.56s-windows were input to a 1D-CNN, resulting in an accuracy of 0.89. Model complexity was evaluated in terms of the total number of parameters and memory requirement. Sigcha et al. [44] evaluated a 1D-CNN with an attention mechanism (CNN-Transformer) on the Rempark dataset. Temporal information was preserved by using a time-distributed layer comprising four consecutive 3.2s-windows. The proposed model achieved an AUROC of 0.96, and complexity was evaluated in terms of the input size, total number of parameters, and inference time. Borzì et al. [45] proposed a multi-head 1D-CNN for exploiting different spatial resolutions in the analysis of inertial data. The model, evaluated on the Rempark dataset, provided an AUROC of 0.95 and an F-score of 0.83. Model complexity was evaluated in terms of input size, total number of parameters, memory requirement, inference time, and number of floating-point operations (FLOPs).

All these studies represent initial attempts to assess the feasibility of real-time implementation of the algorithms. However, in some cases, the complexity of the model is high [43,44]. In addition, the experiments were performed on a personal computer, which has much larger computational resources than a small wearable device. Toward this objective, real-time implementation of the FoG recognition algorithm was carried out in [46–48].

Naghavi et al. [46] evaluated a CNN-based anomaly (one-class) detection approach (DGAD) on data recorded through 2 ankle-mounted IMUs from 7 PwPD. The model provided a sensitivity of 0.63 and a specificity of 0.98. The algorithm was implemented in a real-time FoG identification Android application running on a smartphone. Despite the high complexity produced by the combination of six-axis input signals and six-layer CNN, the testing time was reduced (16 ms). Koltermann et al. [47] proposed a closed-loop wearable system designed to treat FoG. The system consisted of 2 ankle-mounted IMUs streaming data to a smartphone for data processing. When FoG is detected, the smartphone activates vibro-tactile cues administered to subjects' feet. Data were filtered, normalized, and input to a 1D-CNN, providing an F-score of 0.76 when evaluated on data from 11 PwPD with a total of 716 FoG events. The model showed an inference time of 615 ms. A similar approach was proposed by Zoetewei et al. [48] to deliver auditory cues through earphones. The system was validated in the lab on 31 PwPD, showing an intra-class correlation coefficient (ICC) of 0.70 with the clinical rating of FoG. The system was further validated over four weeks of monitoring and on-demand cueing [16] on 63 PwPD. The results showed that the system significantly improved FoG by reducing the percent time spent with FoG (%TF).

These studies have taken a significant step toward real-time FoG detection and processing. However, computational complexity and memory usage have not been reported. In addition, the proposed systems are based on sensors that transmit data to a smartphone via Bluetooth, which may introduce transmission errors and communication delays. Finally, the developed model was evaluated on a single dataset, which does not allow to demonstrate generalization to different scenarios and environments.

The present study aims to develop an edge-AI algorithm specifically designed for real-time implementation on low-power, low-cost, and resource-limited wearable devices. These improvements are essential to enable real-time feedback as a therapeutic strategy to mitigate FoG events and to advance the deployment of this technology in real-life scenarios. The DL model is deployed directly on the sensor node, which functions both as data collector and as processing unit. This process speeds up computation and avoids the use of additional hardware, simplifying the sensor system and eliminating potential data loss and time delays due to communication errors between nodes. The main contributions of this study are based on the following strengths: (a) Comparison of classic feature-driven ML approaches with a more advanced data-driven DL model; (b) optimization of the data processing pipeline and design of a light and fast end-to-end DL algorithm that enables real-time implementation; (c) training and evaluation on a large and heterogeneous dataset, including more than sixty subjects with FoG and more than a thousand FoG episodes; (d) Exhaustive subject-independent validation and performance evaluation, calculated at sample-, episode-, and subject-level; (e) Test on an external dataset with more than a thousand FoG episodes to evaluate generalization to diverse subjects, devices, experiments, and environments; (f) Comprehensive estimation of on-device performance, including memory usage and inference time; (g) Comprehensive comparison against state-of-the-art models, in terms of dataset, data preprocessing, validation method, recognition performance, and model complexity.

The remainder of the paper is organized as follows. The proposed methodology, including datasets, pre-processing, DL model implementation and training, validation approach, and performance evaluation, is given in Section 3. Results are reported in Section 4 and discussed in Section 5. Conclusions are reported in Section 6, along with future work.

3. Material and methods

Fig. 1 describes the implemented FoG recognition pipeline. Starting from the original dataset (Section 3.1), data were pre-processed (Section 3.2) and input to ML (Section 3.3) and DL models (Section 3.4). FoG recognition performance was computed (Section 3.5) and compared among models. Finally, the on-device computational complexity of the best algorithm was computed (Section 3.6).

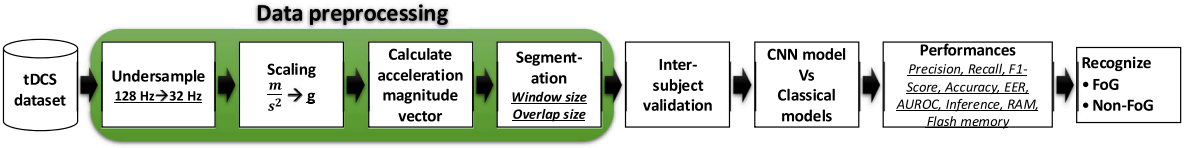


Fig. 1. Implemented FoG recognition pipeline. CNN: convolutional neural network; EER: equal error rate; AUROC: area under the receiver operating characteristic; RAM: random access memory.

3.1. Data

Main dataset: Among the available datasets that include FoG events, this study uses a dataset collected from a previous research work that aimed to investigate the impact of transcranial direct current brain stimulation (tDCS) in FoG [49,50]. This dataset comprises 71 PwPD with FoG, whose characteristics are provided in Table 1. This is by far the most comprehensive and extensive FoG dataset, including the largest number of PwPD with FoG.

Participants undertook a FoG-provoking test in a controlled lab setting, involving three difficulty levels (i.e., single task, dual-motor task, motor-cognitive task). Data were collected using IMUs attached to the participants' lower back, recording acceleration data at a sampling rate of 128 Hz. The study included assessments before and after tDCS administration or sham treatment in different sessions, both On (under dopaminergic therapy) and Off (not under dopaminergic therapy). The FoG-provoking test required participants to walk, turn in circles, enter a doorway, turn again, and return to a seated position. Each test was recorded and annotated offline to identify FoG episodes using specific criteria. Instances with short duration or annotation difficulties were reviewed by multiple evaluators. Our research is limited to analyzing a subset of the original data, consisting of 62 subjects, as only these data were made available to the community through an open repository.[1] A total of 15.32 h of acceleration data were recorded, including 1132 FoG episodes (4.75 h), accounting for 31% of total data. The mean FoG duration was 15.1s (median: 3.3s, inter-quartile range–IQR: 1.6–8.5s). A detailed overview of FoG events distribution in terms of the cumulative distribution function (CDF) and boxplot is presented in Fig. A.6.

External dataset: An independent dataset was employed in this study to test the generalization capability of the FoG detection algorithm. The Rempark dataset [42] was selected due to several factors, including the large number of FoG episodes registered, the environment (i.e., patients' home), and the wide range of activities performed. Subjects' demographic and clinical information are reported in Table 1. Participants were asked to complete a set of scripted tasks, as well as free-living activities. These activities included showing the researchers around their home, stand up and go test crossing through a doorway and turning back, walking outdoors, and cognitive dual tasks (e.g., walking while carrying an object). In addition, a false-positive protocol was designed, comprising activities in which the patient executed short and fast movements repeatedly whose inertial frequency content is similar to a FoG episode. These activities included brushing their teeth, painting/drawing/erasing on a sheet of paper, and cleaning windows. Data were collected using a single IMU attached to the left side of the waist, recording acceleration data at a sampling rate of 40 Hz. Experiments were performed both On and Off. A total of 9.1 h of acceleration data were recorded, including 1058 FoG episodes (1.55 h), accounting for 17% of total data. The mean FoG duration was 6.6s (median: 4.1s, IQR: 2.2–7.9s). A detailed overview of FoG events distribution in terms of CDF and boxplot is presented in Fig. A.7.

Table 1
Demographic and clinical information of the subjects involved in this study.

Dataset	Age (years)	Gender (M/F)	Disease duration (years)	FOG-Q	MMSE	UPDRS III ON (OFF)
tDCS	69.9 ± 7.8	57/14	9.2 ± 5.7	19.4 ± 4.3	28.0 ± 1.8	37.1 ± 14.5 (43.1 ± 16.9)
Rempark	69.3 ± 9.7	18/3	9.0 ± 4.8	15.8 ± 4.1	27.8 ± 1.9	16.2 ± 9.7 (36.3 ± 14.4)

FOG-Q: freezing of gait questionnaire; MMSE: mini-mental state examination; UPDRS: unified Parkinson's disease rating scale.

[1] Kaggle: Parkinson's Freezing of Gait Prediction

3.2. Pre-processing

The 3-axis accelerometer recordings (a_x, a_y, a_z), originally expressed in $\frac{m}{s^2}$, were converted to g by dividing by the gravitational acceleration. This allows the representation of most data in the [−1,+1] range. Most of the energy content of motion data during locomotion and FoG is in the 0–3 Hz and 3–8 Hz frequency band, respectively [17,51]. Therefore, the data were under-sampled to 32 Hz. The under-sampling process produced a mean absolute error (MAE) of 9 mg, 6 mg, and 11 mg on the x, y, and z-axis, respectively. This error, compared to an average acceleration range of 0.41 g, 0.41 g, and 0.35 g, corresponds to a percentage of 2.2%, 1.5%, and 3.1%. The small errors prove that under-sampling allows for maintaining a good signal representation while avoiding unnecessary calculations, thus reducing memory usage and battery consumption [45]. Moreover, this avoids using additional low-pass filters to remove high-frequency noise [40,47], which saves time and computations. The magnitude vector m was computed as in Eq. (1), where i represents the i^{th} sample. It allows for measuring the global movement intensity independently of the specific sensor orientation.

$$m_i = \sqrt{(a_{x_i}^2 + a_{y_i}^2 + a_{z_i}^2)} \tag{1}$$

To prepare the data for input to the classification model, they were organized into a structure of dimension (m, w_s, n_c), where m is the total number of windows, w_s is the window size, and $n_c = 4$ is the number of channels (i.e., 3-axis acceleration signals plus the magnitude vector). Data were segmented using fixed-length windows with overlap. In order to select the best configuration, a grid search was conducted on the window size (min: 1s, max: 4s, step: 1s) and overlap (min:0, max: 50%), and the performance was evaluated for each combination. Majority voting was used to assign the windows to the most represented class (i.e., FoG if at least 50% of samples were labeled as FoG, non-FoG otherwise). Finally, the mean values of each component were removed from each window, as done in previous studies [13,45]. The same pre-processing was applied to the external dataset. In addition, the order and direction of the axes were adjusted to match the sensor setting in the main dataset. A sample of raw data is shown in Fig. B.8, along with the effect of each pre-processing step.

3.3. Machine learning pipeline

A classic ML pipeline was implemented to provide baseline results for FoG detection, including feature extraction and classification. From the preprocessed data, a set of features was extracted from both the time and frequency domain. The extracted features consisted of simple descriptive statistics and more advanced temporal and spectral measures, able to discriminate human activities and to capture key characteristics of FoG (e.g., repetitive movements, amplitude, variations, regularity, complexity, energy content).

Four well-known ML algorithms, namely logistic regressor (LR), linear discriminant analysis (LDA), decision tree (DT), and RF, were implemented and validated. Internal model parameters were optimized using a grid-search approach. By systematically exploring multiple combinations of parameters within a defined range, the optimal configuration for each model can be selected. This avoids relying on arbitrary or guesswork-based parameter values but searches for the most effective hyperparameters.

The list of time- and frequency domain features extracted is provided in Table C.10, while the list of optimized parameters for each model is given in Table C.11, along with the best values.

3.4. Deep learning model

A 1D-CNN was developed to provide a lightweight FoG detection algorithm tailored for real-time implementation on edge devices. This decision prioritizes efficient performance for real-time deployment, in contrast to more complex models (e.g., transformers or LSTMs), that typically require higher computational resources, which can limit their suitability for edge deployment. Fig. 2 schematically reports the implemented DL model.

The first convolutional block comprises two consecutive convolutional layers. A max-pooling layer with a pool size of 2 was used to reduce the size of the feature maps while retaining essential features. A third convolutional layer is followed by a global average pooling layer, that computes the average value of each feature map along the spatial dimensions, generating a one-dimensional map. The latter is connected to a dense layer. Finally, the output layer comprises a single neuron with a sigmoid activation function.

The leaky rectified linear unit (LeakyReLU, Eq. (2)) activation function ($\alpha = 0.1$) was used in all convolutional and dense layers, introducing non-linearity in the network.

$$LeakyReLU(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \tag{2}$$

In addition, all weights were initialized using a random normal distribution. To reduce computation complexity, dilated convolutions (dilation rate = 2) were used in all convolutional layers. This allows the network to have a larger receptive field without increasing the kernel size and thus the number of parameters. To prevent overfitting, dropout layers were included after each convolutional block, with a dropout rate of 0.4, 0.3, and 0.2 respectively. In addition, L2 regularization ($l2 = 10^{-4}$) was used in all layers.

Internal model parameters were optimized, including the number of filters and kernel size for each convolutional layer, and the number of neurons in the fully connected layer. Specifically, the Hyperband method [52] was used to explore a wide set of combinations, reported in Table 2. The algorithm is a bandit-based optimization method that combines random search with

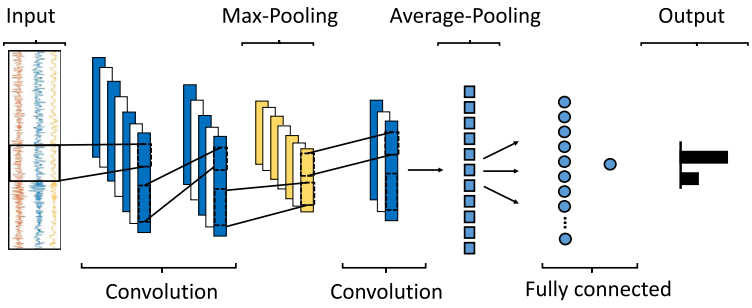


Fig. 2. One-dimensional convolutional neural network architecture.

Table 2
List of parameters and training settings tuned in the optimization process of the convolutional neural network.

Parameter	Search space
n. filters	$4 \cdot n, n \in [2, 8]$
kernel size	$2 \cdot n + 1, n \in [2, 6]$
n. neurons in dense layer	$2^n, n \in [3, 7]$
batch size	$2^n, n \in [6, 10]$
learning rate	$\alpha \cdot 10^{-n}, \alpha \in [1, 9], n \in [1, 6]$
weight decay	$\alpha \cdot 10^{-n}, \alpha \in [1, 9], n \in [1, 6]$

a successive halving strategy, aiming to efficiently allocate resources to promising sets of hyperparameter configurations. The combination that provided the best F-score in the validation set was finally selected. The model was trained using the adaptive moment estimation with weight decay (AdamW) algorithm, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a maximum number of epochs equal to 200. To avoid overfitting, an early stop condition was set, terminating training if the validation loss did not decrease by at least 10^{-3} for ten consecutive epochs. The optimization of the model weights was carried out using the binary cross-entropy loss function and the precision–recall metric. Learning rate, weight decay, and batch size were optimized using a grid-search approach (Table 2).

The final CNN configuration includes 20 and 16 filters in the first two convolutional layers, with kernel sizes of 5 and 7, respectively. The third convolutional block has 12 filters and kernel size of 9. The fully connected layer has 16 neurons. A batch size of 256, a learning rate of $4 \cdot 10^{-3}$, and a weight decay of $5 \cdot 10^{-4}$ were used to train the final model configuration. The resulting model has a total of 4641 trainable parameters, with 4416 coming from the convolutional layers and 225 from the fully connected layers. A detailed analysis of the effect of each parameter on model performance and complexity is provided in Figs. D.9–D.11 and discussed in Appendix D.

3.5. Evaluation methodology

The overall evaluation methodology is shown in Fig. 3. Twenty-one subjects who did not manifest any FoG episodes were initially filtered from the dataset. This resulted in a total of 41 subjects eligible for subsequent analysis. These 41 subjects were divided into a training (50% subjects), validation (25% subjects), and test (25% subjects) set. The division was carefully performed to ensure that each set contained subjects with similar total FoG duration, as described in previous research [49,50,53].

Specifically, subjects were ordered in descending order based on their total FoG duration. Subsequently, an alternating assignment approach was employed, wherein subjects were consecutively allocated to either the training or validation-test set. This process was then repeated for the validation-test set, further dividing it into the validation set and the test set. This resulted in 20, 10, and 11 subjects being included in the training, validation, and test subsets (see Fig. 3).

FoG recognition models were trained using the data from the training set and fine-tuned using the validation set. During this process, close attention was paid to the training and validation losses to prevent the model from over-fitting. Once optimized, the models were rigorously tested (Testing 1 in Fig. 3) on the independent test set, comprising new unseen subjects. It is worth noting that the size of the test set is comparable to the entire dataset used in previous research [27,36–38,46], thus contributing to a realistic performance estimation. An additional evaluation was conducted to further assess the model performance (Testing 2 in Fig. 3). The 21 subjects initially excluded due to their lack of FoG events were used as a separate, independent test group to evaluate the model ability to discard non-FoG instances. Finally, the Rempark dataset was used as a separate external dataset for evaluating the generalization capability of the algorithm to new sensing devices, activities, subject characteristics, and environments (Testing 3 in Fig. 3).

The number of true positives (tp) and false positives (fp), as well as true negatives (tn) and false negatives (fn), were calculated. Subsequently, classification metrics were computed, including precision (Eq. (3)), sensitivity (Eq. (4)), specificity (Eq. (5)), accuracy (Eq. (6)), and F1-score (Eq. (7)). The equal error rate (EER) and the AUROC were also determined. The EER represents the error rate observed at the point on the ROC curve where the sensitivity equals the specificity. The AUROC gauges the overall diagnostic

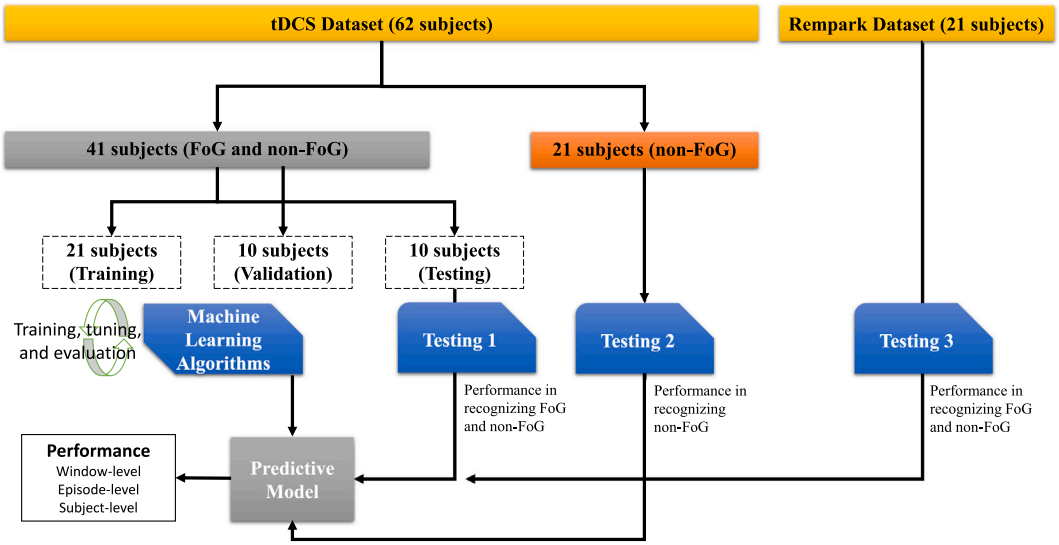


Fig. 3. Evaluation methodology over the main and external dataset.

capability of the model and corresponds to the area under the ROC curve. The value corresponding to the minimum EER was used as a classification threshold to compute the performance metrics. The classification performance of the present study was compared to those of related works in terms of sensitivity, specificity, and AUROC (when available). In addition, since sensitivity and specificity values depend on the selected classification threshold, the geometric mean of sensitivity and specificity was computed (Eq. (8)). This allows us to summarize the two evaluation metrics, penalizing situations in which one measure is much lower than the other.

$$precision = \frac{tp}{tp + fp} \tag{3}$$

$$sensitivity = \frac{tp}{tp + fn} \tag{4}$$

$$specificity = \frac{tn}{tn + fp} \tag{5}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{6}$$

$$F1 - score = \frac{2 \cdot sensitivity \cdot precision}{sensitivity + precision} \tag{7}$$

$$geometric_{mean} = \sqrt{sensitivity \cdot specificity} \tag{8}$$

To further evaluate the clinical validity of the classification model, the %TF was calculated as the percentage of the total time spent by each subject in FoG compared with the total recording time. The %TF was calculated from both the clinical assessment of FoG and the model predictions. The two measures were compared using Pearson's correlation coefficient and relative p-value. Finally, the absolute error between the true and the predicted measures was calculated. In addition to the window-level performance assessment, results at episode-level were evaluated, as detailed in [44,45]. Analyses were conducted by aggregating consecutive windows classified as FoG into episodes. The ability of the model to recognize FoG episodes was assessed by calculating the following metrics.

(a) Percentage of predicted FoG episodes and prediction horizon. The former corresponds to the percentage of episodes predicted before FoG onset. The second is the time interval between the prediction and the actual onset of FoG, expressed in seconds. These metrics describe the capability of predicting FoG before actual occurrence, which can help avoid FoG manifestation [54].
(b) Percentage of FoG episodes detected and detection delay. The former corresponds to the percentage of episodes detected after the onset of FoG. The second is the time interval between the onset of the actual FoG and the detected FoG, expressed in seconds. These metrics evaluate the delay in FoG recognition. The shorter the delay, the earlier cueing strategies can be activated, thus reducing the duration and severity of FoG [15].
(c) Percentage of missed episodes and their duration. The first corresponds to the percentage of real FoG episodes in which no FoG was recognized. The second is the duration of these episodes (expressed in seconds). These metrics describe the overall sensitivity of the model in FoG recognition. The number of missed episodes should be as low as possible. In addition, missing long FoG episodes is more serious than missing short episodes.

(d) Percentage of false episodes and their duration. The first corresponds to the percentage of episodes detected in which no real FoG was present. The second is the duration of these episodes (expressed in seconds). These metrics measure the amount and magnitude of false alarms. Both measures should be as low as possible, to avoid unnecessary and annoying activation of cueing strategies, which can compromise patient compliance with the wearable system.

These measures were calculated both at group-level (i.e., on the entire set of subjects included in the different test sets) and at subject-level (i.e., for each subject separately).

3.6. Edge testing

The 1D-CNN model was deployed on a wearable device to evaluate edge performance. Specifically, the Nordic Thingy:53[2] (Fig. 4) was used, consisting of an IoT prototyping solution tailored to streamline the creation of prototypes and proofs-of-concept, eliminating the need for customized hardware. Its operational core lies in the nRF5340 system on chip (SoC), a dual-core wireless module featuring two Arm Cortex-M33 processors. The Thingy:53 is powered by a rechargeable 1350 mAh Li-Po battery, integrates a range of environmental (temperature, humidity, gas, air quality) and low-power 6-axis IMU. The device enables data collection at a sampling rate spanning from 5 Hz to 200 Hz. Acceleration and angular velocity can be recorded with a settable full scale from ±2 g to ±16 g (accelerometer) and from ±125 dps to ±2000 dps (gyroscope) with 16-bit resolution. The nRF Edge Impulse app[3] provides the capability of deploying trained models on the Thingy:53 device, enabling real-time recognition capabilities. Data were input according to the defined window size and overlap, to closely mimic real-world operating conditions. All operations, including pre-processing and inference, were considered in the evaluation. Edge performance was computed in terms of accuracy, inference time, peak RAM usage, and flash memory utilization for both the float32 and int8 data types, highlighting the model performance before and after applying quantization techniques [55]. Quantization is a technique used to reduce the precision of numerical data, typically by converting floating-point numbers to integers. This process involves representing continuous numerical values with a limited set of discrete values. Moving from float32 (32-bit floating-point numbers) to int8 (8-bit integer numbers) is a form of quantization that reduces the memory footprint of the data, leading to more efficient storage and computation. In addition, the total number of parameters and the number of FLOPs required to predict a single window were computed.

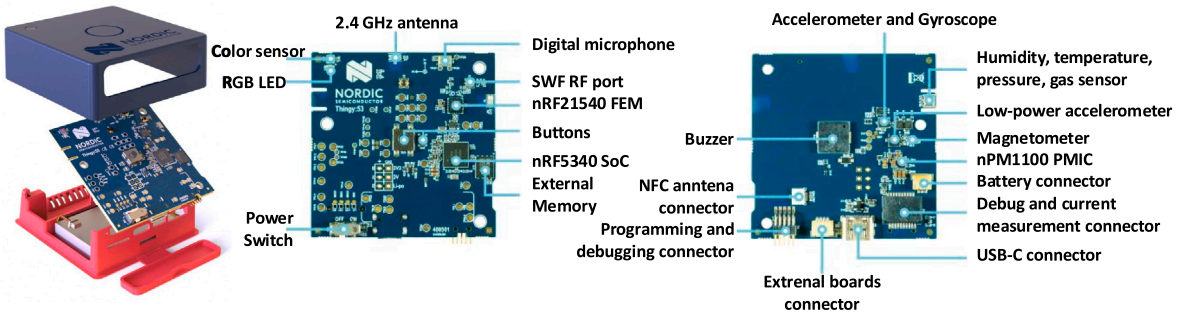


Fig. 4. Nordic Thingy:53 device featuring two Arm Cortex-M33 processors (dimension: 5 cm × 5 cm × 2 cm, price: 65 $).

4. Results

In this section, we present the findings of this study, which involves a comprehensive evaluation of the methodology outlined in Section 3 and the comparison of the CNN model with four traditional ML models.

The experiments were conducted on a computer equipped with a 2.3 GHz processor, 16 GB of RAM, and a 6 GB GPU. Pre-processing tasks were executed using Matlab (version R2023a), while the training and testing phases were carried out in Python (v3.11) utilizing the Keras and TensorFlow (v2.12) libraries. Furthermore, the CNN model was deployed on the Nordic Thingy:53 wearable device to assess its recognition accuracy and computational performance. Edge performance was computed using the Edge Impulse framework. This tool simulates execution on the target device (Nordic Thingy:53), providing accurate estimates of inference time, memory usage, and resource consumption.

4.1. Baseline results for FoG detection

Table 3 summarizes the performance metrics over the test set, including precision, sensitivity, F1-score, and accuracy for the baseline models. The performance does not increase monotonically as the length of the window increases. On the contrary, a maximum is observed around 2–3s. In addition, the overlap has a positive effect on the results. Overall, the configuration consisting

[2] Link to Nordic Thingy:53 producer
[3] Link to Android mobile App

Table 3
FoG recognition performance from baseline machine learning models.

		LR				LDA				DT				RF			
window	**overlap**	**pr**	**se**	**f1**	**acc**	**pr**	**se**	**f1**	**acc**	**pr**	**se**	**f1**	**acc**	**pr**	**se**	**f1**	**acc**
1 sec	0 %	67.8	64.5	64.8	69.1	69.9	67.2	67.7	71.2	69.1	65.0	65.3	69.9	72.1	67.7	68.2	72.2
1 sec	50 %	67.8	64.2	64.5	68.9	79.6	72.3	73.4	77.2	68.7	65.3	65.6	69.9	79.3	72.2	73.2	77.0
2 sec	0 %	67.5	64.0	64.2	68.7	70.4	66.3	66.8	71.2	67.2	63.3	63.4	68.6	70.5	67.0	67.5	71.2
2 sec	50%	68.0	63.7	63.8	68.8	78.8	71.5	72.5	76.7	67.2	63.3	63.5	68.7	79.1	72.1	73.1	76.8
3 sec	0 %	67.5	63.6	63.7	68.6	70.2	68.0	68.5	71.5	67.5	63.5	63.6	68.7	71.6	68.0	68.5	72.1
3 sec	50 %	68.1	63.0	62.9	68.6	78.6	72.1	73.1	76.8	67.9	63.2	63.2	68.7	78.8	71.9	72.9	76.6
4 sec	0 %	66.2	62.3	62.3	67.6	70.2	67.9	68.4	71.5	66.2	62.0	61.9	67.5	70.3	68.6	69.1	71.7
4 sec	50 %	67.5	61.5	61.0	67.9	76.5	70.3	71.1	75.1	67.3	61.3	60.7	67.6	77.0	70.5	71.3	75.4

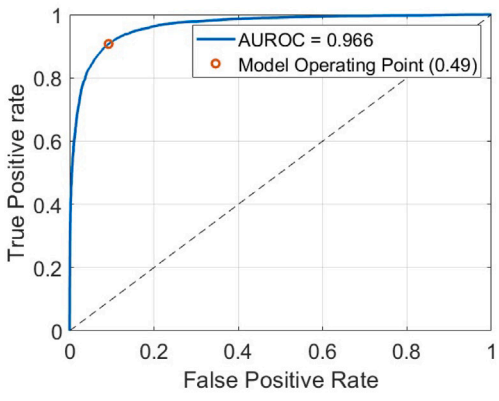
LR: logistic regressor; LDA: linear discriminant analysis; DT: decision tree; RF: random forest.
pr: precision; se: sensitivity; f1: f1-score; acc: accuracy.

of 2s-long windows with 50% overlap provided the best results in three out of four models. Moreover, the RF performed best, in line with previous findings [20,21].

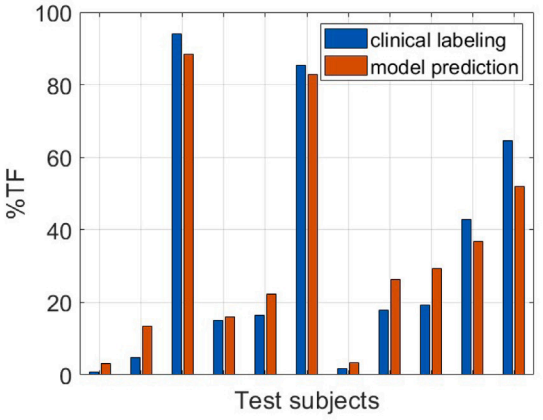
The original authors of the tDCS dataset [50] obtained 80.0% sensitivity, 82.5% specificity, and 86.6% accuracy. These are superior to the top-performing results achieved by the baseline models listed in Table 3, with comparable sensitivity but significantly better specificity (+10.4%) and accuracy (+9.8%). This may be due to several factors, including the different sensor configuration. In particular, two additional sensors on the legs and data from the gyroscopic sensors were analyzed in the original study, but these were not made publicly available. The additional data may significantly improve performance, considering that the sensors on the ankles showed the best results in detecting FoG [36,56,57]. In addition, differences in pre-processing procedures (segmentation, filtering, and feature selection) and classification model (SVM with radial basis function) may affect performance.

4.2. CNN results for FoG detection

Fig. 5 shows the CNN model performance in terms of the ROC curve (Fig. 5(a)) and %TF (Fig. 5(b)) over the test set. The model shows a very good discrimination capability, with an AUROC of 0.966. Good values of sensitivity (true positive rate) and specificity (1 - false positive rate) are observed for a wide range of classification thresholds. Specifically, increasing the threshold from 0.4 to 0.6 increases specificity from 88.2% to 93.7% and decreases sensitivity from 92.8% to 86.4%. The model operating point shown in Fig. 5(a) (0.49) corresponds to the minimum EER and generates the results reported in Table 4. The %TF calculated from the model prediction is very close to the true %TF calculated from the clinical labeling (Fig. 5(b)). More in detail, the absolute error in the %TF quantification is < 5% in 36% cases, between 5% and 10% in 45% cases, and > 10% in 18% cases (median: 5.9%, IQR: 2.4–8.6%). The two measures shows a strong linear relationship, quantified by a Pearson's correlation coefficient of 0.98 ($p < 0.001$). Considering that the results refer to the test set, which includes previously unseen subjects, the model demonstrates high clinical validity.



(a) Receiver operating characteristic curve, along with the model operating point.



(b) Percent time spent in freezing (%TF) calculated from the clinical labeling and the model output.

Fig. 5. FoG recognition performance of the proposed model. Results refer to the test set.

Table 4 presents the CNN performance on the test set, obtained by using 2s-windows with 50% overlap. Comparing the results with those reported in Table 3, it is evident that the CNN model outperformed all the classic ML algorithms, with improvements of 19.1% in F1-score and 13.9% in accuracy over the RF model. Moreover, the CNN model provided an improvement of 4.1% in accuracy, 10.7% in sensitivity, 8.2% in specificity, and 2.6% in AUROC over the original method.

Nonetheless, it is worth noting that the classical ML approach requires the calculation of a specific set of features, unlike the DL method, which relies solely on raw data. In this case, the only calculation required is that of the acceleration magnitude.

Table 4
Performance of the CNN model on the test set, along with the results from the original study. Down-arrows indicate better results corresponding to lower values of performance metrics.

Study	Accuracy	Sensitivity	Specificity	Precision	F1-score	EER (%, ↓)	AUROC
Original [50]	86.6	80.0	82.5	–	–	–	0.940
Present	90.7	90.7	90.7	93.8	92.2	9.26	0.966

EER: equal error rate; AUROC: area under the receiver operating characteristic.

Therefore, the latter approach achieves significantly better results and eliminates the need to calculate features, which would otherwise introduce additional complexity and lengthen the model inference time.

The results demonstrate that the model can successfully identify 90.7% of FoG instances while maintaining a precision rate of 93.8%. These outcomes, coupled with a low EER of 9.26% and an AUROC exceeding 96%, attest to the robust performance of the model and its ability to generalize effectively.

Finally, when testing the model on the 21 subjects initially excluded due to the lack of FoG events, the model achieved a specificity of 95.4%. This corresponds to a total of 255 false positives registered in 3 h. However, most false positives (65%) were of short duration (1–2s) and only 5% exceeded 5s. Discarding single-window FoG episodes (e.g., by detecting FoG only when at least two consecutive overlapped windows are classified as FoG) improves specificity up to 97.8% (+2.4%). This corresponds to a total number of 94 false alarms in 3 h, with a median duration of 2s. On the one hand, the high specificity demonstrates the capability of the model to reject false positives correctly. On the other hand, the number of false alarms should be strictly controlled to avoid additional burden for patients. An appropriate trade-off between sensitivity, precision and specificity should be carefully selected, specific to each subject and in accordance with the clinical staff.

From the total number of FoG episodes in the test set (266), 17.3% were detected at onset, 22.2% were predicted on average 2s before FoG onset (IQR: 1–3s), and 47% were detected with a median delay of 1s (IQR: 1–2s). The remaining episodes (13.5%) were not detected. However, 44% of these not detected episodes were short (1–2s) and only 5% had a duration longer than 5s. Finally, most false positives (62%) corresponded to isolated single windows (i.e., a window classified as FoG while the preceding and the following windows were classified as non-FoG). The remaining false positives had a mean duration of 1.7s (IQR: 1–2s).

In summary, more than 85% of FoG episodes were detected early or within 1s of FoG onset, while the model hardly recognized very short episodes. The number of false alarms is small, as evidenced by the high precision value. In addition, most of the false positives relate to short time periods (e.g., a single window of 1–2 s), which can be easily discarded.

4.3. On-device performance

Table 5 presents the assessment of the CNN model performance in terms of inference time, peak RAM usage, and flash memory utilization for both the float32 and int8 data types, highlighting the model performance before and after applying quantization techniques [55]. Quantization effectively reduces the model memory requirements and computation time, making the CNN model suitable for deployment on stand-alone wearable devices with limited resources.

The most precise configuration of the edge model achieved an inference time of 1 ms and utilized 6.3 Kb of RAM and 37.8 Kb of Flash memory, using a 2-second window with 50% overlap. The average recognition performance was consistent across all training and testing iterations, with accuracy, sensitivity, and specificity at 90.0%. The small variation observed between the on-the-edge and server results in Table 4 can be attributed to the inherent randomness in the initialization of network weights during each iteration. In accordance with the FoG recognition pipeline shown in Fig. 1, the only preprocessing steps performed before feeding data into the CNN model (Fig. 4) were the calculation of the acceleration magnitude vector and data segmentation into overlapping windows. These operations were implemented directly on the edge device using the Edge Impulse framework, ensuring lightweight processing suitable for real-time inference. Practical tests on the physical Nordic Thingy:53 device showed stable behavior during continuous operation. After deployment, the model ran in inference mode while maintaining a BLE connection with the Edge Impulse

Table 5
On the edge performance of the proposed model. Up/Down-arrows indicate better results corresponding to higher/lower values of performance metrics..

		int8				float32			
window (*sec*)	**overlap** (%)	**ram (↓)** (*Kb*)	**flash (↓)** (*Kb*)	**inference (↓)** (*ms*)	**acc (↑)** (%)	**ram (↓)** (*Kb*)	**flash (↓)** (*Kb*)	**inference (↓)** (*ms*)	**acc (↑)** (%)
1 sec	0 %	6.0	37.6	1	88.6	3.9	47.1	1	88.7
1 sec	50 %	6.0	37.6	1	89.4	3.9	47.1	1	89.8
2 sec	0 %	6.3	37.8	1	89.5	5.3	47.2	2	89.8
2 sec	50 %	6.3	37.8	1	90.0	5.3	47.2	2	90.0
3 sec	0 %	6.5	38.0	1	89.3	6.6	47.8	2	89.1
3 sec	50 %	6.5	38.0	1	89.5	6.6	47.8	2	89.6
4 sec	0 %	6.8	38.0	2	90.0	6.7	47.8	2	90.0
4 sec	50 %	6.8	38.0	2	90.0	7.9	47.8	2	90.0

ram: random access memory; acc: accuracy.

Android app for over 6 h without any time holes or data loss. The battery level dropped from 100% to 47%, indicating a runtime of approximately 12–13 h. Notably, this battery consumption includes continuous Bluetooth communication, which is known to account for a large portion of the total energy usage. Longer runtimes are expected in offline scenarios or reduced communication settings.

Overall, the evaluation of the edge model underlines the potential of on-device processing as a promising approach for FoG recognition. From a clinical perspective, real-time FoG recognition with an inference time of only 1 ms allows for swift detection and intervention on patients, improving their care and safety. From the perspective of daily living of the patient, this technology provides a sense of security by promptly identifying FoG episodes, triggering stimulations to help overcome FoG [14,58], and improving the overall QoL [12,13,19].

4.4. FoG detection performance at subject level

To evaluate the impact of subject-specific characteristics on FoG detection performance, performance were computed at subject-level. Table 6 reports the results for each subject, expressed in terms of timely detected, predicted, detected, and missed FoG episodes. The prediction horizon and detection delay are also provided, expressed in seconds. Finally, to contextualize the results, FoG characteristics are provided for each subject, in terms of the total number of FoG episodes manifested, total FoG duration, and %TF. A high heterogeneity is observed in FoG characteristics, with the number of episodes ranging from 5 to 54, and total FoG from 6 s (0.7% of the total recording time) to 2.1 h (94% of the total recording time). Detection performance shows generally good results, with an average of 19% of episodes timely detected, 25.4% predicted, 39% episodes detected, and only 15.2% missed. The average prediction horizon was 1.9 s (range: 1–3.5s), while the average detection delay was 1.4s (range:1–2 s). False positives, calculated excluding isolated single-window detections, were generally low (range: 1–6). The only exception is represented by subject 3, where 31 false alarms are registered. Overall, a significant inter-subject variability is observed in detection performance. This suggests that subject-specific thresholds should be used to optimize performance, maximizing prediction and detection rate and minimizing the detection delay.

Subject-level specificity, calculated on the set of 21 subjects initially excluded from the analysis, ranged from 71.6% to 99.3%, with an average value of 94.4% and a standard deviation of 6.5%. The average specificity at subject-level slightly differs from the specificity calculated at sample level, due to the different duration of recordings in each subject.

Table 6
Performance of the CNN model on the test set. Performance is calculated for each subject separately.

Subject	1	2	3	4	5	6	7	8	9	10	11
Total recording time (min)	13.8	26.1	23.1	137.1	9.8	20.4	16.9	10.4	2.7	13.0	14.0
# FoG episodes	5	40	18	54	20	43	7	27	5	27	20
Total FoG duration (s)	6	1338	66	7730	88	201	18	111	31	335	542
%TF	0.7	85.3	4.8	94.0	15.0	16.4	1.8	17.8	19.4	42.8	64.6
# Timely detected episodes (↑)	1	12	4	5	2	7	3	6	1	1	2
# Predicted episodes (↑)	1	7	4	7	8	18	0	12	3	0	4
Average prediction horizon (s, ↑)	2	1	1	2	3.5	2	–	1	2	–	3
# Detected episodes (↑)	1	17	8	30	6	11	4	8	1	18	7
Average detection delay (s, ↑)	1	1	1.5	2	1.5	1	2	1	1	1	2
# Missed episodes (↓)	2	3	2	6	4	6	0	1	0	8	6
# FP (excluding isolated FP, ↓)	6	1	31	1	4	4	2	5	5	0	3
Average FP duration (s, ↓)	2.3	4.0	3.8	4.0	2.5	4.8	2.5	2.8	3.2	–	2.3

%TF: percent time spent with FoG; FP: false positive.

4.5. The effect of therapy

Cross-medication tests were conducted to evaluate the effect of therapy on detection performance. Table 7 reports the results obtained when training the model with data from patients Off (On) and tested on data from patients On (Off) medication, as previously done in [8]. In addition, further tests were conducted training the model with all available data and testing on data from On and Off medications, respectively, as done in [35]. Table 7 shows that training the model on patients On and testing on patients Off produces better results than vice versa. Specifically, a reduction of 3% in accuracy, 11% in F-score and 2% in AUROC is observed in the second case. On the other hand, training the model with all available data and testing on data from patients On therapy provide better results, as evidenced by an increase of 3%, 10%, and 1.5% in accuracy, F-score, and AUROC, respectively.

4.6. Generalization to external datasets

Table 8 presents the results obtained when testing the trained model on the external Rempark dataset. Performance is reported using statistical measures (i.e., minimum, maximum, average, standard deviation) across all 21 subjects. The average percentage of timely detected, predicted, detected with delay, and missed FoG episodes is 14.5%, 32.8%, 34.2%, and 18.2%, respectively. These results are similar to that registered in the tDCS test set (19%, 25.4%, 39%, 15.2%). This proves the good generalization of the model to new data, recorded from different subjects in different environments. On the other hand, the number of false positives and their duration are significantly higher than that observed in tDCS test set. This can be partially explained by the longer duration of experiments in the Rempark dataset, accounting for an average of 54 min per subject, compared to the 26 min in the tDCS. In addition, a high variability in FoG characteristics (i.e., number and duration of episodes) and detection performance is observed, as evident from the high standard deviation and range found in all metrics.

Table 7
Results from the cross-medication tests. Performance refer to model trained on Off (On) data and tested on On (Off) data. Up/Down-arrows indicate better results corresponding to higher/lower values of performance metrics.

Train set	Test set	Accuracy (%, ↑)	Sensitivity (%, ↑)	Specificity (%, ↑)	Precision (%, ↑)	F1-score (%, ↑)	EER (%)	AUROC (↑)
On	Off	88.1	88.1	88.1	84.7	86.3	11.9	0.954
Off	On	85.0	85.0	85.0	67.6	75.3	15.1	0.933
All	Off	87.8	87.8	87.8	72.7	79.5	12.2	0.953
All	On	90.6	90.6	90.6	87.8	89.2	9.3	0.969

EER: equal error rate; AUROC: area under the receiver operating characteristic.

Table 8
FoG detection performance on the external dataset. Up/Down-arrows indicate better results corresponding to higher/lower values of performance metrics.

Statistics	Min	Max	Mean	Std
Total recording time (min)	15.2	90.1	53.9	18.3
# FoG episodes	1	119	44.9	29.9
Total FoG duration (s)	3	957	335.2	220.5
%TF	0.1	27.1	11.1	7.0
% Timely detected episodes (↑)	0	36	14.5	7.9
% Predicted episodes (↑)	0	80	32.8	21.7
Prediction horizon (s, ↑)	1	6	2.8	1.4
% Detected episodes (↑)	0	100	34.2	22.5
Detection delay (s, ↓)	1	3	1.9	0.7
% Missed episodes (↓)	0	54.6	18.2	17.0
# FP (↓)	8	173	55.4	39.7
Mean FP duration (s, ↓)	2.3	12	5.5	2.1
Total FP duration (% recording, ↓)	1.2	22.7	8.3	4.8

FP: false positive.

4.7. Comparison with state-of-the-art approaches

The results of the present work were comprehensively compared with those of related studies. Specifically, Table 9 reports the sample size, validation method, DL model, classification performance, computational complexity (i.e., total number of parameters and FLOPs), and on-device performance (i.e., memory requirement and inference time) of state-of-the-art approaches models. The information reported in Table 9 were sourced from the original publications.

Most studies evaluated the FoG detection algorithm on a sample of 7–22 PwPD, while a much larger population of more than 60 PwPD was only used in [39] and the present study. However, the model proposed in [39] presents suboptimal classification performance. In addition, the computational complexity given by both the two-dimensional CNN and the data pre-processing and transformation (i.e., CWT computation) methods hinders real-time edge implementations.

Although transformer-based models are gaining attraction, few studies have applied them to FoG detection [38,44], and their generalization on imbalanced or limited accelerometer data remains unclear. While the tDCS dataset includes a relatively large sample, transformers-based models such as vision transformers (ViT) typically require larger datasets to achieve state-of-the-art performance. Also, their lack of translation equivariance and locality (key for time-series tasks) may limit their effectiveness in this context [62]. This may partly explain the superior performance of our method, which was designed to address these domain-specific challenges. Moreover, the higher computational and memory demands of transformer architectures pose practical challenges for edge deployment.

As shown in Table 9, comparing results across studies is challenging due to differences in datasets and validation methods. LOSO validation was commonly used in studies with less than 21 subjects, with a geometric mean (GM) of sensitivity and specificity ranging from 0.79 to 0.89. Studies using a hold-out validation obtained a GM from 0.87 to 0.88, evaluated on a test set of 6–13 subjects. Overall, the approach proposed in this study shows promising classification performance, with the best AUROC (0.97) and GM (0.91) inferior only to [37] (0.92). However, it is worth noting that a subject-dependent approach was used in [37], which does not allow the development of general models that perform well on new, previously unseen subjects [12]. This underlines the importance of choosing appropriate validation strategies. Subject-dependent methods often lead to overoptimistic results on unseen data, whereas subject-independent approaches (such as LOSO or subject-independent hold-out) are better suited for evaluating models intended for real-world FoG detection, where the system must perform reliably across data from unseen subjects.

Unfortunately, most studies did not evaluate model complexity. Moreover, few studies reported inference times and memory requirements. Finally, only one study comprehensively evaluated the computational load and testing time for real-time applications. However, the experiments were limited to a personal computer [45]. Overall, the algorithm proposed in this study presents the smallest number of parameters and FLOPs, and the lowest memory requirement and inference time.

This is the result of a processing pipeline designed to reduce computational burden and promote real-time implementations. Specifically, the model requires the input of a single triaxial accelerometer, generating 3 channels. This represents a minimal

Table 9
Comparison of classification performance and computational complexity between the present study and related works. Up/Down-arrows indicate better results corresponding to higher/lower values of performance metrics. Best results are presented in bold type.

Study	Sample size	Validation method	Model	Performance (↑)	NP (↓)	FLOPs ($\cdot 10^6$, ↓)	IT (ms, ↓)	MU (Kb, ↓)
Camps et al. (2018) [43]	21	Hold out test:6	1D-CNN	**se: 91.9%** sp: 81.5% gm: 86.5% AUROC: 0.880	37 121	0.337	–	flash: 145
Sigcha et al. (2020) [26]	21	LOSO	CNN-LSTM	se: 87.1% sp: 87.1% gm: 87.1% AUROC: 0.939	–	–	–	-
Bikias et al. (2021) [27]	11	LOSO	1D-CNN	se: 83% sp: 88% gm: 85.5%	43 181	3.14	–	-
Shi et al. (2022) [39]	63	Hold out test:13	2D-CNN	se: 87.8% sp: 86.4% gm: 87.1%	–	–	–	-
Sigcha et al. (2022) [44]	21	LOSO	CNN Transformer	se: 89% sp: 89.1% gm: 89.0% AUROC: 0.957	87 825	8.93	45	-
Naghavi et al. (2022) [46]	7	LOSO	1D-CNN	se: 63.0 textbfsp: 98.6 gm: 78.8%	–	–	16	-
Borzìet al. (2023) [45]	21	Hold out test:5	Multi-head CNN	se: 87.7% sp: 88.3% gm: 88.0% AUROC: 0.946	10 834	0.399	43	flash: 55
O'Day et al. (2022) [36]	7	LOSO	1D-CNN	AUROC: 0.830	–	–	–	-
Sun et al. (2024) [37]	10	Hold out test:20% (SD)	1D-CNN	se: 86.2% sp: 98.8% **gm: 92.3%**	–	–	–	-
Huang et al. (2024) [38]	12	LOSO	CNN Transformer	se: 74.5% sp: 87.3% gm: 80.6%	-	-	-	–
Koltermann et al. (2024) [47]	11	LOSO	1D-CNN	acc: 86% f-score: 76%	–	–	615	-
Borzi et al. (2025) [59]	22	Hold out test:6	1D-CNN	sens: 82.6% spec: 82.5% f-score: 63.2% AUROC: 0.909	8300	–	60	-
Yang et al. (2025) [60]	18	LOSO	MS-TCN	f-score: 77.1%	–	–	–	-
Chen et al. (2025) [61]	24	LOSO	2D-CNN	sens: 89.2% spec: 79.2% f-score: 66.1%	–	–	–	-
Present study	62	Hold out test:10 external:21	1D-CNN	se: 90.7% sp: 90.7% gm: 90.7% AUROC: 0.966	**4641**	**0.132**	**1**	**flash: 37.8** ram: 6.3

LOSO: leave-one-subject-out; SD: subject-dependent; CNN: convolutional neural network; LSTM: long short-term memory.
MS-TCN: multi-stage temporal convolutional network; se: sensitivity; sp: specificity; gm: geometric-mean; AUROC: area under the receiver operating characteristic; acc: accuracy; NP: number of parameters; FLOPs: floating-point operations per second; IT: inference time; MU: memory usage.

sensor configuration compared to previous studies that employed 6–9 channels per device [27,38,43,46,47,59,60], which increased to 18–42 channels when multiple devices were placed on different body positions [36,39,61]. A smaller number of dimensions reduces the size of input data, controls model complexity, and speeds up computations. A reduced sampling rate (fs=32 Hz) was used, compared with other studies (fs = 100–128 Hz [27,38,46,47]), further reducing the size of the input data. The input data consists of a single time window of 2 s. This time frame is shorter than most previous approaches (3–4s windows [26,27,38,39]). In addition, some previous works exploited multiple (2–4) consecutive windows to account for temporal relationships between adjacent windows [26,43]. The use of a single short window contributes to light and fast data processing designed to promote real-time implementations. The model requires minimal preprocessing of the data, including only mean removal and magnitude calculation. Additional filtering, normalization, transformations (e.g., FFT [26,43], continuous wavelet transform [39], gramian summation angular field [61]) or feature extraction were avoided to keep computational resources low and speed up calculations. The designed 1D-CNN has 3 convolutional layers, dilated convolutions, global pooling and a single dense layer with a small number of neurons. This represents a lightweight DL architecture compared to previous studies (5–8 convolutional layers [39,46], 2 fully connected layers [26,43,46]). In addition, model parameters were controlled, with a reduced number of kernels (12–20) compared to previous works (64-128 [26,27,38,46], or even more than 500 [39]). Finally, the use of dilated convolutions and global pooling further controls the complexity of the model. The first method provides a larger spatial receptive field without increasing the number of parameters. The second approach reduces the spatial dimensions of feature maps to a single value per channel, significantly reducing the number of connections to the next dense layer.

5. Discussion

An edge-AI algorithm was developed in this study for real-time recognition of FoG in PwPD. The proposed approach consists of a fast and light 1D-CNN that process three-axis acceleration data recorded from a single wearable inertial sensor on the lower back. The algorithm was comprehensively trained, optimized and evaluated on a large dataset with more than 62 PwPD and 1132 FoG episodes, and further tested on an external dataset with 21 PwPD and 1058 FoG events. Window-level and episode-level classification metrics were calculated, and the effect of inter-subject variability, medication, and diverse environments and activities was evaluated. Finally, computational complexity and inference time were carefully calculated by processing data directly on a low-cost wearable device with limited resources.

Window-level performance was promising, with sensitivity, specificity, and precision over 90%, along with an AUROC of 0.97. When considering subject-independent approaches (i.e., generalized models), these results are superior to most previous studies [12,33,34]. Episode-level performance demonstrated good-to-excellent performance, with 22% episodes predicted on average 2s before FoG onset, 17% episodes detected at onset, and 47% detected with an average delay of 1s.

The evaluation of the model on patients without FoG (non-freezers) resulted in a specificity of 95.4%, increasing to 97.8% when discarding single-window FoG detections. This corresponds to a very high value, superior to most previous studies analyzing supervised gait tasks or simulated daily activities (68% in [40], 81.5% in [43], 86.4% in [39], 87.5% in [38], 88.3% in [45], 89.1% in [44]). A few studies [30–32] evaluated FoG detection algorithms in real-life scenarios for an extended period of time (5 to 7 consecutive days). These studies evaluated the number of episodes, their duration, and the %TF in freezers and non-freezers. They found that these measures were statistically different in the two groups. However, a significant number of false positives was evidenced by an average %TF of 15% in non-freezers and 20% in freezers [30], 0.07% in non freezers and 0.11% in freezers [31], and about 5% in non-freezers and 10% in freezers [32]. These results demonstrate that FoG detection algorithms applied to continuous recording in daily life produce a significant number of false positives. Finally, when calculating specificity at subject-level, this study highlighted heterogeneity among subjects, with specificity ranging from 72% to 99%. This suggests that subject-specific thresholds should be selected to maximize performance and minimize patient burden.

Performance at group level (i.e., evaluated on the entire test set) were complemented with results at subject level, calculated for each participant separately. When considering the average performance across subjects, the results of the two experiments were consistent. This is evidenced by the similar values of the percentage of timely detected, predicted, detected with delay, and missed FoG episodes, as well as the specificity calculated on non-freezers. However, subject-level analyses highlighted a wide heterogeneity across subjects in the recording time, number of FoG episodes and their duration, and model performance. Nevertheless, a small number of missed episodes and a low false alarm rate were observed in all test subjects. This shows that the model is robust to inter-subject variability in gait and FoG patterns.

Cross-medication experiments were conducted to evaluate the effect of therapy on performance. The results demonstrated a slightly better performance when training the model with On data and testing on Off data, in line with a previous work [8]. On the other hand, when training the model with all available data, tests in Off resulted in a higher F-score, confirming previous findings [35]. Overall, the results highlight the importance of including data from both On and Off medications to develop robust recognition algorithms, as suggested in [34].

The generalization capability of the developed FoG detection model was evaluated on an external, independent dataset. This included data collected in the patients' home during semi-supervised and unsupervised activities. The results were consistent with those obtained in the main dataset, with similar percentages of timely detected, predicted, detected with delay, and missed FoG episodes. However, a large number of false alarms was registered, and these false positives lasted longer than those presented in the main dataset. This can be partially explained by the intrinsic characteristics of the experiments. In particular, the duration of records in the external dataset was doubled compared to the main database, which contributes to a large number of false positives. In addition, the activities performed in the external dataset were weakly controlled, and also included activities specifically designed to produce signals patterns similar to FoG (e.g., brushing teeth). Finally, it is worth considering that the two datasets were collected using a different wearable device in a slightly different position (lower back and lateral waist, respectively), which can further contribute to errors.

In addition, it is worth noting that patients in the two datasets were evaluated in both On and Off conditions. In the case of the Rempark dataset, an additional "intermediate" class was included. Moreover, the clinical characteristics of patients differed in the two datasets, with a different distribution of FOG-Q and UPDRS-III. This, together with varying pharmacological conditions, adds complexity and heterogeneity to the data under evaluation. Despite these differences, the model proved robust, with comparable results across different conditions, clinical characteristics, and heterogeneous gait and FoG patterns.

The developed DL model has an extremely low computational complexity (i.e., less than 5k parameters), compared to previous studies (37k [43], 43k [27], 87k [44]). This is the result of designing a CNN with a few convolutional blocks and a single dense layer with a small number of neurons. This, coupled with the use of global pooling and dilated convolutions significantly reduce the total number of connections, improving computation time and memory usage. Performance and generalization capability were maximized by exploiting dropout, regularization, training settings (e.g., stop conditions), and a robust validation method.

The FoG recognition algorithm provides nearly real-time detection, with 22% of episodes predicted on average 1.9s before FoG onset, 17% detected at onset, and 47% detected with an average delay of 1s. This delay is in line with the findings from [48], where a closed-loop wearable system was described. The results suggest that some episodes may be avoided by activating specific cues before actual FoG manifestation. On the other hand, the detection is delayed in more than a half episodes. In these cases, cueing strategies can only reduce the duration of FoG episodes, and may be ineffective for very short (<2s) events. In addition,

majority voting strategies based on consecutive windows can enhance the robustness of the model and reduce false alarms. On the other hand, this will introduce an extra delay in FoG recognition. This can be controlled by using a small slide/step (i.e., larger overlap between consecutive windows), which allows for processing data more frequently (e.g., every 100–200 ms). A good trade-off between temporal resolution, performance, and computational complexity should be carefully evaluated. In addition to these technological considerations, subject-specific reaction time and effectiveness can significantly contribute to the real FoG treatment efficacy. Although these aspects go beyond the scope of this study, it should be carefully considered in future clinical studies.

Few studies have evaluated the efficacy of a closed-loop wearable system in improving FoG. In [16], on-demand auditory cues administered through a real-time system significantly reduced the %TF in patients with FoG, both On and Off medication. Patients with on-demand cues presented a lower number of FoG episodes and smaller FoG duration compared to subjects without cues. It is worth noting that the proposed system has an average detection delay of 1s [48], in line with the algorithm developed in the present work. On the one hand, the current results have not been evaluated in a clinical study. On the other hand, the developed edge-AI algorithm can minimize detection delay by avoiding data transmission to a smartphone (as done in [48]). In this case, the device functions both as a sensing node and processing unit. If the same device is used also as an actuator for administering cues (e.g., vibro-tactile stimulation [63]), technological delay can be further reduced.

The challenges of developing accurate FoG detection algorithms for real-world applications are magnified by the scarcity of high-quality datasets that capture daily-life scenarios. Most existing datasets are collected in controlled environments or during freezing-provoking tasks, which limits their generalization capability to unsupervised, real-life conditions. Furthermore, these datasets often rely on precise ground-truth annotations, which are labor-intensive and impractical to generate for long-duration recordings [59]. Weakly labeled datasets, where annotations are less granular or rely on higher-level event tagging rather than precise temporal labeling, can represent a promising alternative for addressing these limitations. Continuous data collection in remote, real-life settings can provide valuable insights into natural variations in movement patterns and contextual influences on FoG. These datasets can be leveraged by semi-supervised or self-supervised ML and DL algorithms to learn robust representations of FoG. These methods exploit large amounts of unlabeled data to improve performance. This allows for learning from more representative datasets while minimizing the reliance on exhaustive manual annotation. Additionally, the inherent variability and noise in real-world datasets can improve the robustness and generalization of algorithms, ultimately leading to more reliable FoG detection in daily-life conditions.

5.1. Limitations

This study has some limitations. First, comparing different studies is challenging, as each study is designed based on a specific dataset. The limited number of enrolled PwPD may undergo different activities, use different sensors, and have variations in sensor placement, resulting in a lack of uniformity among studies. This further hampers the ability to draw meaningful conclusions and make accurate comparisons among different recognition systems.

The analyzed dataset was collected in laboratory settings. Despite the well-structured experimental procedures, which included different activities and gait tasks, testing on data collected in the home environment can provide more realistic results. To address this issue, the developed model was further tested on an external dataset, which was collected in the home setting and included semi-supervised activities. However, experiments lasted between 15 min and 1.5 h, and this fails to provide continuous recordings over free-living activities in naturalistic environments.

Moreover, data from a single sensor on the lower back were analyzed. Such location is known to be comfortable and acceptable for long-term monitoring [64]. On the other hand, the combination of multiple sensors can improve FoG recognition performance [36]. It is worth noting that recording additional physiological signals, such as skin conductance and electrocardiogram, may not provide incremental performance over inertial data [60].

Although the system showed a relatively low false positive rate, its impact on patient experience, including potential cue fatigue and compliance over time, must be explored in future real-world studies. Evaluating the user experience with actual PwPD is essential to fully assess the system's practicality and to adapt cueing strategies accordingly.

Despite comprehensive evaluation of performance, computational complexity and inference time, the proposed system has not been yet validated for real-time FoG treatment in properly designed clinical studies. Therefore, real-world challenges such as subject-specific reaction time and effectiveness of different cues (e.g., vibro-tactile, auditory, visual) remain unexplored. Although this study did not include direct user testing with PwPD, previous research has demonstrated that patients generally accept wearable FoG detection systems, even in the presence of occasional false positives. This is particularly true when such systems are integrated with on-demand cueing mechanisms, which help mitigate or shorten FoG episodes, thereby improving perceived usefulness and overall user compliance [16,48,51].

Finally, although developing and evaluating a general subject-independent model provided promising performance, subject-specific approaches may improve accuracy and effectiveness by better adapting to specific patient conditions, activity patterns, and FoG manifestations. In this context, exploiting transfer learning to train a general model and fine-tune on the specific subject can improve performance while controlling for the amount of training data required [65]. Semi-supervised and self-supervised learning strategies further reduce the required data [66]. As the clinical conditions of patients evolve over time, adaptive models can be further leveraged to continuously fine-tune the model to new activities, habits, conditions, and manifestations.

6. Conclusion and future work

The present work contributes a novel and efficient data processing pipeline for accurate FoG detection. The results indicate that the proposed approach is accurate and robust to inter-subject variability and diverse sensor settings, environments, and activities. In addition, it provides low computational burden and very small inference time. Overall, the proposed method holds promise for enhancing the lives of individuals with PD by providing accurate, real-time FoG detection on everyday wearable devices. While our study has made significant strides in FoG detection, several avenues for future research and development remain open.

Future work will involve clinical validation with PwPD to assess usability, comfort, and real-world performance. Special attention will be given to evaluating the impact of false alarms on patient compliance and the potential development of cue fatigue during prolonged use.

The actuation strategy will be implemented on the wearable device itself to provide somatosensory (e.g., vibro-tactile) feedback upon FoG detection. Thus, the device will function as sensing node, processing unit, and actuator, avoiding external data transfer. This avoids communication errors and delays, preserve privacy, and enhance the system efficacy. Furthermore, the system will be validated in a properly designed clinical study, to face real-world challenges and investigate subject-specific responses. Subsequently, the system should be tested in real-life scenarios over continuous monitoring. As part of the current relevant challenges, future research efforts should prioritize the development and open sharing of weakly labeled FoG datasets collected continuously over extended periods in remote settings. These datasets should encompass diverse populations, including both freezers and non-freezers, and represent a wide range of daily activities and environmental conditions. Such initiatives would facilitate the training and evaluation of semi-supervised FoG detection algorithms, enabling a transition from laboratory-based to real-world applications.

CRediT authorship contribution statement

Luigi Borzì: Conceptualization, Methodology, Software, Investigation, Visualization, Validation, Writing- Original Draft. **Luis Sigcha:** Formal analysis, Software, Validation, Writing- Original Draft. **Farshad Firouzi:** Supervision, Writing – review & editing. **Gabriella Olmo:** Project administration, Resources, Supervision, Writing – review & editing. **Florenc Demrozi:** Conceptualization, Supervision, Visualization, Writing- Original Draft.

Data availability

The tDCS 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 proposed one-dimensional convolutional neural network is available at this link.

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.

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 “Telecommunications of the Future” (PE00000001 - program “RESTART”).

Appendix A. Distribution of FoG events

Fig. A.6 presents the distribution of FoG episodes in the tDCS dataset. The left panel presents the cumulative distribution function (CDF) of FoG episode durations up to approximately 650 s, showing that nearly all episodes are under 100 s, with a few extreme outliers extending beyond 10 min. The top-right panel depicts a box plot of FoG episode durations (clipped at 30 s for visualization), indicating a median duration of about 4 s, with the interquartile range extending roughly from 2 to 7 s. Numerous outliers above 17 s are marked as red crosses, emphasizing the skewed nature of the data. The bottom-right panel zooms in on the CDF for episodes lasting up to 30 s, revealing that around 90% of FoG episodes fall within this range. This comprehensive view illustrates the predominance of short FoG episodes while highlighting the presence of rare but considerably longer events (such longer events are also due to the fact that the dataset creators have merged FoG events divided by non FoG events shorter then a defined threshold), offering insights critical for designing responsive cueing and intervention technologies.

Fig. A.7 presents the distribution of FoG episodes in the Rempark dataset. The left panel shows the CDF of FoG episode durations (in seconds), illustrating that the majority of episodes are relatively short; approximately 90% of episodes are under 20 s, and nearly all are under 75 s. The top-right panel presents a box plot of FoG episode durations, highlighting the median (4 s), interquartile range, and presence of outliers (red crosses) beyond the whiskers. The bottom-right panel shows a zoomed-in version of the CDF for episodes duration up to 30 s, emphasizing the rapid accumulation of episodes in this shorter duration range. Together, these plots characterize the temporal profile of FoG events, providing insight into typical episode lengths and variability, which can inform the design of cueing systems and therapeutic interventions.

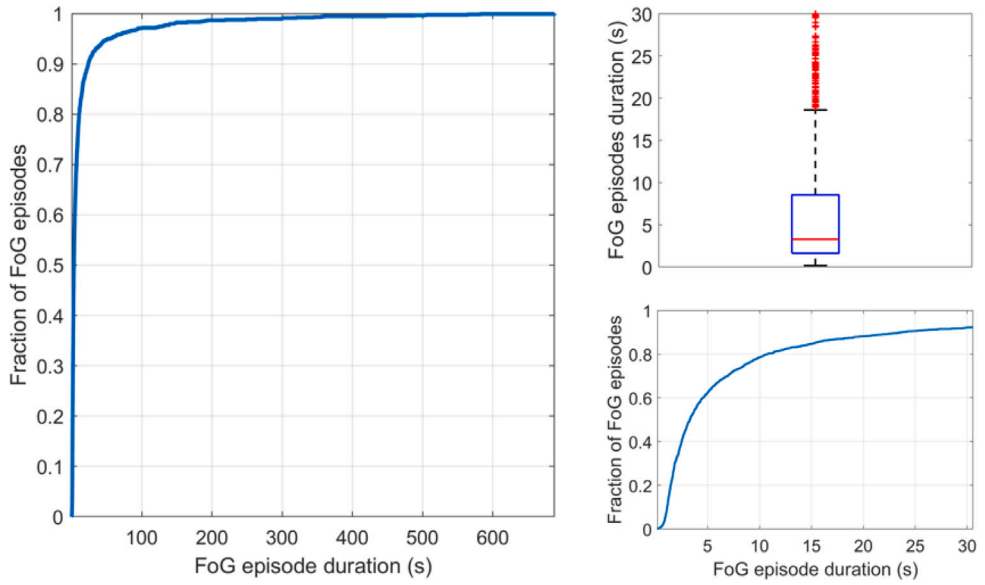


Fig. A.6. Distribution of FoG episodes durations in the tDCS dataset.

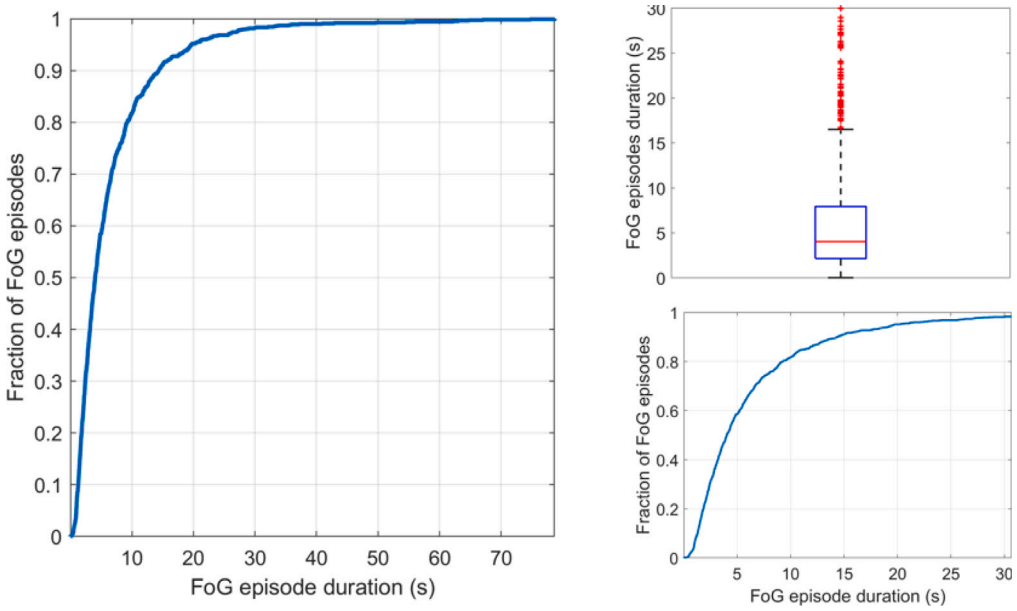


Fig. A.7. Distribution of FoG episodes duration in the Rempark dataset.

Appendix B. Signal pre-processing

Fig. B.8 shows a characteristic time frame of about 18 s, including two FoG events. The original acceleration signals (top-left image) are sampled at 128 Hz and measured in $\frac{m}{s^2}$. The effect of transformation to g-force unit is shown in the top-right image, where the x-axis has an average value of 1 g, produced by the constant gravity component. The effect of data undersampling is shown in the bottom-left image, where the original and resampled x-axis component are shown along a common timeline. As observed, the difference between the two signals is small. Finally, the bottom-right image shows the acceleration magnitude calculated from the 3-axis acceleration signals.

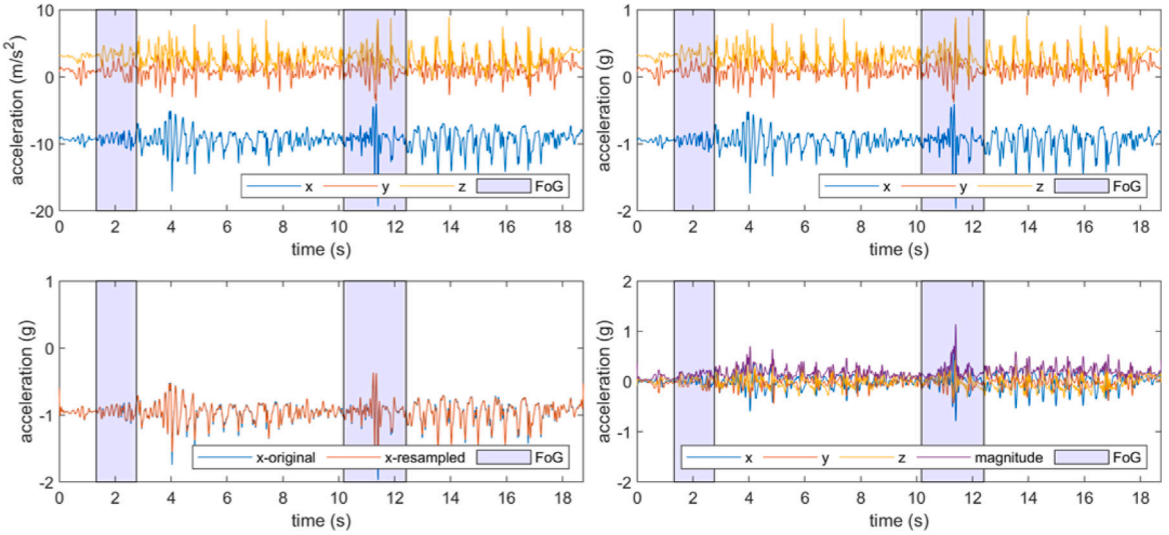


Fig. B.8. Preprocessing steps performed on raw acceleration data. Top-left: raw data; top-right: conversion to g-unit; bottom-left: down-sampling; bottom-right: magnitude computation.

Appendix C. Feature extraction and model optimization

Table C.10 reports the list of features extracted in the time and frequency domain. Table C.11 shows the list of parameters optimized for each ML model. The search space is provided, along with the best configurations (bold type).

Table C.10
List of features employed in this study.

Time domain features	Frequency domain features
1) maximum, 2) minimum, 3) mean, 4) standard deviation, 5) root mean square, 6) range, 7) median, 8) skewness, 9) kurtosis, 10) time-weighted variance, 11) interquartile range, 12) empirical cumulative density function, 13) percentiles (10, 25, 75, and 90), 14) sum of values above or below percentile (10, 25, 75, and 90), 15) square sum of values above or below percentile (10, 25, 75, and 90), 16) number of crossings above or below percentile (10, 25, 75, and 90), 17) mean amplitude deviation, 18) mean power deviation, 19) signal magnitude area, 20) signal vector magnitude, 21) covariance, 22) simple moving average of sum of range of a signal, 23) sum of range of a signal, 24) sum of standard deviation of a signal, 25) maximum slope of simple moving average of sum of variances of a signal, 26) autoregression.	1) Fast Fourier Transform (FFT) coefficients, 2) Discrete Fourier Transform (DFT), 3) Discrete Wavelet Transform (DWT), 4) first dominant frequency, 5) ratio between the power at the dominant frequency and the total power, 6) ratio between the power at frequencies higher than 3.5 Hz and the total power, 7) two signal fragmentation features, 8) DC component in FFT spectrum, 10) energy spectrum, 11) entropy spectrum, 12) sum of the wavelet coefficients, 13) squared sum of the wavelet coefficients and energy of the wavelet coefficients, 14) auto-correlation, 15) mean-crossing rate, 16) spectral entropy, 17) spectral energy, 18) wavelet entropy values, 19) mean frequency, 20) energy band.

Appendix D. The effect of parameters on model complexity and performance

Starting with the optimal model configuration, the effect of each parameter on model performance was evaluated, both in terms of F-score and model complexity (FLOP and parameters). Removal of dilated convolutions increases the number of FLOPs by 35%, while detection performance is not significantly affected (−0.2% in F-score). Using a classic flatten layer instead of global pooling increases the number of parameters by 22%, while detection performance is not significantly affected (−0.1% in F-score). Removing dropout slightly reduces performance (−0.8% in F-score). Using a classic ReLU instead of LeakyReLU has no significant effect (−0.2%

Table C.11
List of parameters and their respective search space for different models. The selected parameters are presented in **bold**.

Model	Parameters [Search space]
Logistic Regression	penalty: ['**l1**', 'l2'] C: [0.1, 1, **5**, 10] tol: [0.0001, **0.001**, 0.01] solver: ['newton-cg', '**lbfgs**', 'liblinear', 'sag', 'saga']
Linear Discriminant Analysis	solver: ['svd', '**lsqr**', 'eigen'] n_components: [2, 4, 6, 8, 10, 12, **14**, 16, 18, 20] tol: [0.0001, **0.001**, 0.01]
Decision Tree	criterion: ['**gini**', 'entropy'] max_depth: [3, 6, 9, 12, **15**, 18, 20] min_samples_split: [2, 4, 6, **8**, 10] min_samples_leaf: [1, 2, 3, **4**, 5] max_features: ['auto', '**sqrt**']
Random Forest	n_estimators: [50, 100, 150, 200, **250**, 300, 350, 400] criterion: ['**gini**', 'entropy'] max_depth: [3, 6, 9, 12, **15**, 18, 20] min_samples_split: [2, 4, 6, **8**, 10] min_samples_leaf: [1, 2, 3, **4**, 5] max_features: ['auto', '**sqrt**']

in F-score). Increasing the number of filters significantly increases the number of FLOPs (from 65 K to 12M) and parameters (from 1.4 K to 267 K), but not the classification performance. The maximum F-score corresponds to a number of filters of 16 (Fig. D.9). Increasing the kernel size slightly increases the number of FLOPs (from 191 K to 953 K) and parameters (from 2 K to 9.2 K) . Higher values of the F-score are registered with a kernel size between 5 and 9 (Fig. D.10). Increasing the number of neurons in the fully-connected layer does not particularly affect the number of FLOPs (from 263 K to 266 K) and parameters (from 4.5 K to 6.2 K). The best F-score corresponds to a dense layer with 16 neurons, and additional neurons worsen performance (Fig. D.11).

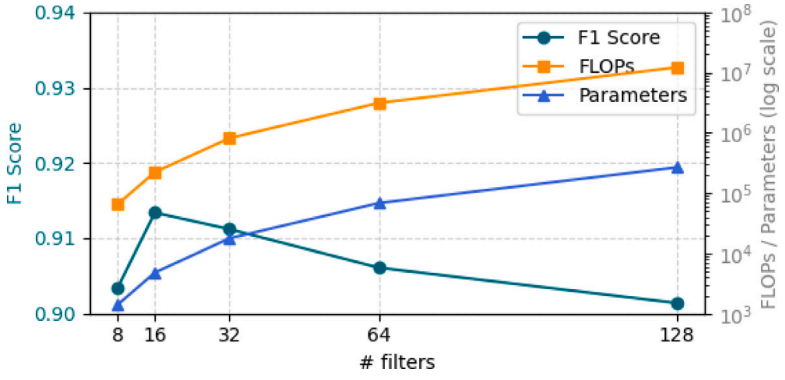


Fig. D.9. Effect of an increasing number of filters on model performance and complexity.

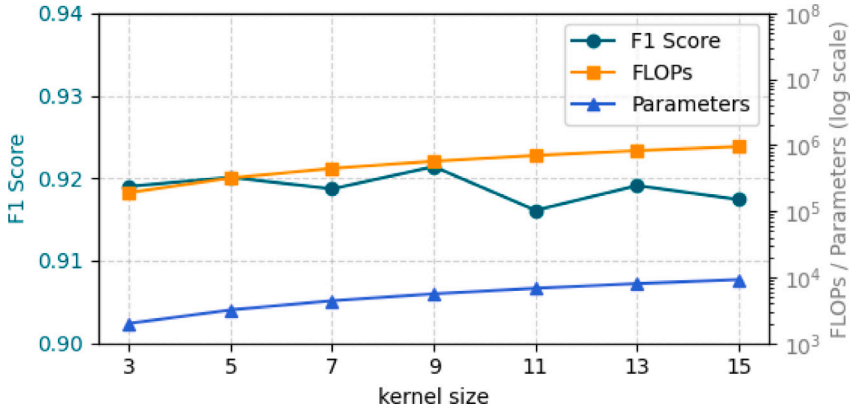


Fig. D.10. Effect of an increasing kernel size on model performance and complexity.

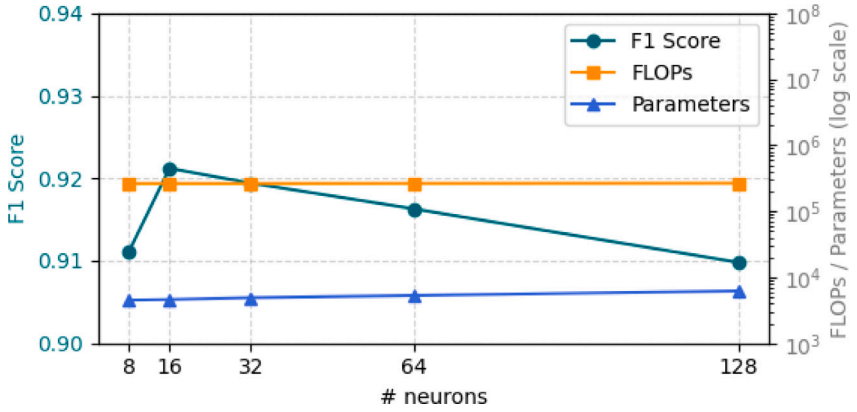


Fig. D.11. Effect of an increasing number of neurons on model performance and complexity.

References

[1] Magrinelli F, Picelli A, Tocco P, Federico A, et al. Pathophysiology of motor dysfunction in Parkinson's disease as the rationale for drug treatment and rehabilitation. Parkinsons Dis 2016;2016. http://dx.doi.org/10.1155/2016/9832839.

[2] Weiss A, Herman T, Giladi N, Hausdorff JM. New evidence for gait abnormalities among Parkinson's disease patients who suffer from freezing of gait: insights using a body-fixed sensor worn for 3 days. J Neural Transm 2015;122(3):403–10. http://dx.doi.org/10.1007/s00702-014-1279-y.

[3] Nutt JG, Bloem BR, Giladi N, Hallett M, et al. Freezing of gait: Moving forward on a mysterious clinical phenomenon. Lancet Neurol 2011;10(8):734–44. http://dx.doi.org/10.1016/S1474-4422(11)70143-0.

[4] Mazilu S, Blanke U, Roggen D, Tröster G, et al. Engineers meet clinicians: Augmenting Parkinson's disease patients to gather information for gait rehabilitation. In: Proceedings of the 4th augmented human international conference. New York, NY, USA: Association for Computing Machinery; 2013, p. 124–7. http://dx.doi.org/10.1145/2459236.2459257.

[5] Schaafsma JD, Balash Y, Gurevich T, Bartels AL, et al. Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease. Eur J Neurol 2003;10(4):391–8. http://dx.doi.org/10.1046/j.1468-1331.2003.00611.x.

[6] Suppa A, Kita A, Leodori G, Zampogna A, et al. L-DOPA and freezing of gait in Parkinson's disease: Objective assessment through a wearable wireless system. Front Neurol 2017;8(AUG). http://dx.doi.org/10.3389/fneur.2017.00406.

[7] Mancini M, Bloem BR, Horak FB, Lewis SJG, et al. Clinical and methodological challenges for assessing freezing of gait: Future perspectives. Mov Disord 2019;34(6):783–90. http://dx.doi.org/10.1002/mds.27709.

[8] Borzì L, Mazzetta I, Zampogna A, Suppa A, Olmo G, Irrera F. Prediction of freezing of gait in Parkinson's disease using wearables and machine learning. Sensors (Basel) 2021;21(2):1–19. http://dx.doi.org/10.3390/s21020614.

[9] Ferraris C, Amprimo G, Masi G, Vismara L, Cremascoli R, et al. Evaluation of arm swing features and asymmetry during gait in Parkinson's disease using the Azure kinect sensor. Sensors 2022;22:6282. http://dx.doi.org/10.3390/s22166282.

[10] Din SD, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: Vision, state of the art, and challenges ahead. J Parkinsons Dis 2021;11(s1):S35–47. http://dx.doi.org/10.3233/JPD-202471.

[11] Huang T, Li M, Huang J. Recent trends in wearable device used to detect freezing of gait and falls in people with Parkinson's disease: A systematic review. Front Aging Neurosci 2023;15:1119956. http://dx.doi.org/10.3389/fnagi.2023.1119956.

[12] Sigcha L, Borzi L, Amato F, Rechichi I, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. Expert Syst Appl 2023;229:120541. http://dx.doi.org/10.1016/j.eswa.2023.120541.

[13] Borzì L, Sigcha L, Olmo G. Context recognition algorithms for energy-efficient freezing-of-gait detection in Parkinson's disease. Sensors 2023;23(9). http://dx.doi.org/10.3390/s23094426.

[14] Demrozi F, Bacchin R, Tamburin S, Cristani M, Pravadelli G. Toward a wearable system for predicting freezing of gait in people affected by Parkinson's disease. IEEE J Biomed Heal Informatics 2019;24(9):2444–51. http://dx.doi.org/10.1109/JBHI.2019.2952618.

[15] Ginis P, Nackaerts E, Nieuwboer A, Heremans E. Cueing for people with Parkinson's disease with freezing of gait: A narrative review of the state-of-the-art and novel perspectives. Ann Phys Rehabil Med 2018;61(6):407–13. http://dx.doi.org/10.1016/j.rehab.2017.08.002.
[16] Zoetewei D, Herman T, Ginis P, Palmerini L, Brozgol M, Thumm PC, et al. On-demand cueing for freezing of gait in Parkinson's disease: A randomized controlled trial. Mov Disorders 2024;39(5):876–86. http://dx.doi.org/10.1002/mds.29762.
[17] Moore S, MacDougall H, W.G. O. Ambulatory monitoring of freezing of gait in Parkinson's disease. J Neurosci Methods 2008;167(2):340–8. http://dx.doi.org/10.1016/j.jneumeth.2007.08.023.
[18] Bachlin M, Plotnik M, Roggen D, Maidan I, Hausdorff JM, Giladi N, et al. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. IEEE Trans Inf Technol Biomed 2010;14(2):436–46. http://dx.doi.org/10.1109/TITB.2009.2036165.
[19] Borzì L, Olmo G, Lopiano CAAL. Detection of freezing of gait in people with Parkinson's disease using smartphones. In: 2020 IEEE 44th Annual Computers, Software, and Applications Conference. 2020, p. 625–35.. http://dx.doi.org/10.1109/COMPSAC48688.2020.0-186.
[20] San-Segundo R, Navarro-Hellín H, Torres-Sánchez R, Hodgins J, De la Torre F. Increasing robustness in the detection of freezing of gait in Parkinson's disease. Electronics 2019;8(2). http://dx.doi.org/10.3390/electronics8020119.
[21] Pardoel S, Kofman J, Nantel J, Lemaire ED. Wearable-sensor-based detection and prediction of freezing of gait in Parkinson's disease: A review. Sensors (Switzerland) 2019;19(23). http://dx.doi.org/10.3390/s19235141.
[22] Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci 2021;2(6):1–20. http://dx.doi.org/10.1007/s42979-021-00815-1.
[23] Guo Y, Huang D, Zhang W, Wang L, Li Y, Olmo G, et al. High-accuracy wearable detection of freezing of gait in Parkinson's disease based on pseudo-multimodal features. Comput Biol Med 2022;146:105629. http://dx.doi.org/10.1016/j.compbiomed.2022.105629.
[24] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8(1):53. http://dx.doi.org/10.1186/s40537-021-00444-8.
[25] Tăuţan A-M, Ionescu B, Santarnecchi E. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. Artif Intell Med 2021;117:102081. http://dx.doi.org/10.1016/j.artmed.2021.102081.
[26] Sigcha L, Costa N, Pavon I, Costa S, Arezes P, Lopez J, et al. Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. Sensors 2020;20(7):1895. http://dx.doi.org/10.3390/s20071895.
[27] Bikias T, Iakovakis D, Hadjidimitriou S, Charisis V, Hadjileontiadis LJ. DeepFoG: An IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. Front Robot AI 2021;8:537384. http://dx.doi.org/10.3389/frobt.2021.537384.
[28] May DS, Tueth LE, Earhart GM, Mazzoni P. Using wearable sensors to assess freezing of gait in the real world. Bioengineering 2023;10(3). http://dx.doi.org/10.3390/bioengineering10030289.
[29] Rodríguez-Martín D, Samà A, Pérez-López C, Català A, Moreno Arostegui J, et al. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PLoS One 2017;12(2). http://dx.doi.org/10.1371/journal.pone.0171764.
[30] Mancini M, Shah VV, Stuart S, Curtze C, Horak FB, Safarpour D, et al. Measuring freezing of gait during daily-life: an open-source, wearable sensors approach. J NeuroEng. Rehabil 2021;18:1–13. http://dx.doi.org/10.1186/s12984-020-00774-3.
[31] Zampogna A, Borzì L, Rinaldi D, Artusi CA, et al. Unveiling the unpredictable in Parkinson's disease: Sensor-based monitoring of dyskinesias and freezing of gait in daily life. Bioengineering 2024;11(5). http://dx.doi.org/10.3390/bioengineering11050440.
[32] Salomon A, Gazit E, Ginis P, et al. A machine learning contest enhances automated freezing of gait detection and reveals time-of-day effects. Nat Commun 2024;15. http://dx.doi.org/10.1038/s41467-024-49027-0.
[33] Elbatanouny H, Kleanthous N, Dahrouj H, Alusi S, et al. Insights into parkinson's disease-related freezing of gait detection and prediction approaches: A meta analysis. Sensors 2024;24(12). http://dx.doi.org/10.3390/s24123959.
[34] Zhang W, Sun H, Huang D, et al. Detection and prediction of freezing of gait with wearable sensors in Parkinson's disease. Neurol Sci 2024;45:431–53. http://dx.doi.org/10.1007/s10072-023-07017-y.
[35] Yang P, Filtjens B, Ginis P, et al. Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops. J NeuroEng. Rehabil 2024;21(24):2444–51. http://dx.doi.org/10.1186/s12984-024-01320-1.
[36] O'Day J, Lee M, Seagers K, Hoffman S, et al. Assessing inertial measurement unit locations for freezing of gait detection and patient preference. J NeuroEng. Rehabil 2022;19(20). http://dx.doi.org/10.1186/s12984-022-00992-x.
[37] Sun H, Ye Q, Xia Y. Predicting freezing of gait in patients with Parkinson's disease by combination of manually-selected and deep learning features. Biomed Signal Process Control 2024;88:105639. http://dx.doi.org/10.1016/j.bspc.2023.105639.
[38] Huang D, Wu C, Wang Y, Zhang Z, Chen C, Li L, et al. Episode-level prediction of freezing of gait based on wearable inertial signals using a deep neural network model. Biomed Signal Process Control 2024;88:105613. http://dx.doi.org/10.1016/j.bspc.2023.105613.
[39] Shi B, Tay A, Au WL, Tan DML, Chia NSY, Yen S-C. Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. IEEE Trans Biomed Eng 2022;69(7):2256–67. http://dx.doi.org/10.1109/TBME.2022.3140258.
[40] Klaver EC, Heijink IB, Silvestri G, van Vugt JPP, et al. Comparison of state-of-the-art deep learning architectures for detection of freezing of gait in Parkinson's disease. Front Neurol 2023;14. http://dx.doi.org/10.3389/fneur.2023.1306129.
[41] Sigcha L, Borzì L, Olmo G. Deep learning algorithms for detecting freezing of gait in Parkinson's disease: A cross-dataset study. Expert Syst Appl 2024;255:124522. http://dx.doi.org/10.1016/j.eswa.2024.124522.
[42] Rodríguez-Martín D, Samà A, Pérez-López C, Català A, et al. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PLoS One 2017;12(2):1–26. http://dx.doi.org/10.1371/journal.pone.0171764.
[43] Camps J, Samà A, Martín M, Rodríguez-Martín D, Pérez-López C, Moreno Arostegui JM, Cabestany J, et al. Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. Knowl-Based Syst 2018;139:119–31. http://dx.doi.org/10.1016/j.knosys.2017.10.017.
[44] Sigcha L, Borzì L, Pavón I, Costa N, Costa S, Arezes P, et al. Improvement of performance in freezing of gait detection in Parkinson's disease using transformer networks and a single waist-worn triaxial accelerometer. Eng Appl Artif Intell 2022;116:105482. http://dx.doi.org/10.1016/j.engappai.2022.105482.
[45] Borzì L, Sigcha L, Rodríguez-Martín D, Olmo G. Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. Artif Intell Med 2023;135:102459. http://dx.doi.org/10.1016/j.artmed.2022.102459.
[46] Naghavi N, Wade E. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: A novel deep one-class classifier. IEEE J Biomed Heal Inform 2022;26(4):1726–36. http://dx.doi.org/10.1109/JBHI.2021.3103071.
[47] Koltermann K, Jung W, Blackwell G, Pinney A, et al. Fog-finder: Real-time freezing of gait detection and treatment. In: Proceedings of the 8th ACM/IEEE international conference on connected health: applications, systems and engineering technologies. New York, NY, USA; 2024, p. 22–33. http://dx.doi.org/10.1145/3580252.3586980.
[48] Zoetewei D, Herman T, Brozgol M, Ginis P, Thumm PC, Ceulemans E, et al. Protocol for the DeFOG trial: A randomized controlled trial on the effects of smartphone-based, on-demand cueing for freezing of gait in Parkinson's disease. Contemp Clin Trials Commun 2021;24:100817. http://dx.doi.org/10.1016/j.conctc.2021.100817.
[49] Manor B, Dagan M, Herman T, Gouskova NA, Vanderhorst VG, Giladi N, et al. Multitarget transcranial electrical stimulation for freezing of gait: a randomized controlled trial. Mov Disorders 2021;36(11):2693–8. http://dx.doi.org/10.1002/mds.28759.

[50] Reches T, Dagan M, Herman T, Gazit E, Gouskova NA, Giladi N, et al. Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. Sensors 2020;20(16):4474. http://dx.doi.org/10.3390/s20164474.
[51] Bachlin M, Hausdorff JM, Roggen D, Giladi N, Plotnik M, Troster G. 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; 2009, http://dx.doi.org/10.4108/ICST.BODYNETS2009.5852.
[52] Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 2018;18(185):1–52, URL http://jmlr.org/papers/v18/16-558.html.
[53] Howard A, Salomon A, Gazit E, Jevster H, Hausdorff J, Kirsch L, et al. Parkinson's freezing of gait prediction. Kaggle; 2023, URL https://kaggle.com/competitions/tlvmc-parkinsons-freezing-gait-prediction.
[54] Sweeney D, Quinlan LR, Browne P, Richardson M, et al. A technological review of wearable cueing devices addressing freezing of gait in Parkinson's disease. Sensors 2019;19(6):1277. http://dx.doi.org/10.3390/s19061277.
[55] Lin D, Talathi S, Annapureddy S. Fixed point quantization of deep convolutional networks. In: International conference on machine learning. PMLR; 2016, p. 2849–58. http://dx.doi.org/10.48550/arXiv.1511.06393.
[56] Li B, Yao Z, Wang J, Wang S, et al. Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors. Electronics 2020;9(11). http://dx.doi.org/10.3390/electronics9111919.
[57] Ashfaque Mostafa T, Soltaninejad S, McIsaac TL, Cheng I. A comparative study of time frequency representation techniques for freeze of gait detection and prediction. Sensors 2021;21(19). http://dx.doi.org/10.3390/s21196446.
[58] Demrozi F, Bragoi V, Tramarin F, Pravadelli G. An indoor localization system to detect areas causing the freezing of gait in parkinsonians. In: 2019 design, automation & test in Europe conference & exhibition. IEEE; 2019, p. 952–5. http://dx.doi.org/10.23919/DATE.2019.8715093.
[59] Borzì L, Demrozi F, Bacchin RA, Turetta C, Sigcha L, Rinaldi D, et al. Freezing of gait detection: The effect of sensor type, position, activities, datasets, and machine learning model. J Parkinsons Dis 2025;15(1):163–81. http://dx.doi.org/10.1177/1877718X241302766.
[60] Yang P-K, Filtjens B, Ginis P, Goris M, Nieuwboer A, Gilat M, et al. Multimodal freezing of gait detection: Analyzing the benefits and limitations of physiological data. IEEE Trans Neural Syst Rehabil Eng 2025;33:956–65. http://dx.doi.org/10.1109/TNSRE.2025.3545110.
[61] Chen Q, Chen Z, Zhang F, Chen S, Ren K, Lu N. Dual-level freezing of gait recognition. IEEE Sensors J 2025. http://dx.doi.org/10.1109/JSEN.2025.3557407.
[62] Xu Z, Liu R, Yang S, Chai Z, Yuan C. Learning imbalanced data with vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 15793–803.
[63] Rossi S, Lisini Baldi T, Aggravi M, et al. Wearable haptic anklets for gait and freezing improvement in Parkinson's disease: a proof-of-concept study. Neurol Sci 2020;41:3643—-3651. http://dx.doi.org/10.1007/s10072-020-04485-4.
[64] Keogh A, Alcock L, Brown P, Buckley E, Brozgol M, et al. Acceptability of wearable devices for measuring mobility remotely: Observations from the mobilise-D technical validation study. Digital Health 2023;9:20552076221150745. http://dx.doi.org/10.1177/20552076221150745.
[65] Bursa SO, Incel OD, Isiklar Alptekin G. Personalized and motion-based human activity recognition with transfer learning and compressed deep learning models. Comput Electr Eng 2023;109:108777. http://dx.doi.org/10.1016/j.compeleceng.2023.108777.
[66] Petchhan J, Su S-F. Advances in inter-edge transfer learning with self-curriculum-labeling adaptive learning and lightweight attention. Comput Electr Eng 2024;116:109201. http://dx.doi.org/10.1016/j.compeleceng.2024.109201.