# FoG-Track: Semi-supervised Real-Time Response to Freezing of Gait

LUYAO YANG, CEMSE Division, King Abdullah University of Science and Technology, KSA
OSAMA AMIN, CEMSE Division, King Abdullah University of Science and Technology, KSA
BASEM SHIHADA, CEMSE Division, King Abdullah University of Science and Technology, KSA

Freezing of Gait (FoG) is a prevalent motor dysfunction experienced by patients with Parkinson's disease, and while extensive research has focused on detecting FoG episodes in both clinical and home environments over the past decade, predictive approaches that enable preemptive prompting are still scarce. Most existing studies rely on fully supervised learning conditions, which pose significant challenges. This study introduces a semi-supervised learning framework utilizing a prototype network designed to leverage a FoG prediction model based on impaired gait patterns indicative of pre-FoG, simultaneously harnessing information from unlabeled sensor data for real-time FoG prediction. To validate our framework, we establish pre-FoG state labeling across four datasets: Daphnet, CPGDD, PDTURN, and BXHC. Our developed real-time detection model demonstrates strong performance, achieving up to 96% sensitivity, 99% specificity, and 91% average accuracy on the cross-validation dataset. Furthermore, we compare the accuracy between semi-supervised and fully supervised modes, revealing that the semi-supervised approach yields improvements in average accuracy ranging from 0.01 to 0.02 across three datasets. Thus, our proposed method represents an effective strategy for the accurate prediction of FoG in real-time algorithmic settings.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Freezing of Gait, accelerometer, Gyroscope, Parkinson's disease, gait impairment, semi-supervised, wearable sensors

**ACM Reference Format:**
Luyao Yang, Osama Amin, and Basem Shihada. 2025. FoG-Track: Semi-supervised Real-Time Response to Freezing of Gait. 1, 1 (March 2025), 20 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Freezing of Gait (FoG), a profound and distressing symptom of Parkinson's disease (PD), presents a substantial challenge to those affected. Characterized by a sudden, episodic inability to move the feet forward despite the intention to walk, FoG is often analogized to the sensation of having one's feet "glued" to the ground [1]. This debilitating phenomenon disrupts the normal rhythmic and fluid movement of walking, severely limiting mobility and significantly heightening the risk of falls. Such falls are not only common among PD patients but also represent a major source of injury, further diminishing quality of life. As a pronounced form of gait disorder, FoG typically manifests during gait initiation, while turning, or when navigating through tight spaces, compounding its impact on daily activities and overall well-being [2].

Recent clinical observations have revealed that prior to the onset of FoG, patients often exhibit a stage of impaired gait patterns, including accelerated rhythm, deterioration of rhythmic gait, gait asymmetry, and poor

Authors' Contact Information: Luyao Yang, luyao.yang@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA; Osama Amin, osama.amin@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA; Basem Shihada, Basem.shihada@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA.


ACM XXXX-XXXX/2025/3-ART
https://doi.org/XXXXXXX.XXXXXXX

balance control [3]. This stage is characterized by serialized features in gait patterns that precede FoG [3, 4]. Recognizing these preliminary symptoms, some researchers have proposed the concept of a pre-FoG stage, where gait alterations before the actual freezing event occurs [5–8]. However, a significant challenge in studying these premonitory signs is the nature of existing public datasets, which predominantly classify gait into binary categories: FoG and non-FoG, without specific annotations for the pre-FoG stage. Typically, these labels are assigned by clinical experts who review and time-stamp video recordings of patients, a process that is not only labor-intensive but also lacks precision in marking the onset of changes in serialized signals.

To address this gap, some researchers have started to include annotations for the pre-FoG stage within these datasets, which usually spans a few time windows immediately before FoG occurs [5–8]. Accurate prediction of this pre-FoG stage could enable systems to anticipate and react to the imminent onset of actual FoG episodes in a timely and effective manner. Building on this approach, this paper incorporates and analyzes four public datasets [9–12], applying a unified annotation method for the pre-FoG phase (2s-3s) to these datasets. This enhanced annotation aims to improve the predictive capabilities of FoG detection systems, ensuring more proactive management of Parkinson's disease symptoms. These preprocessed datasets have been organized and will be available after the paper is published.

In addition, in the real-world setting, the wearable sensor data collected from patients often consists of a large number of parameters, and the labeling of FoG episodes can be a complex and time-consuming process, typically requiring at least two experts' observation and annotation. FoG data annotation is labor-intensive and time-consuming, particularly given the rapid expansion of sensor data collection and concerns about user privacy [9–11, 13]. The sheer volume of data and the intricacies of manual annotation hinder the scalability of supervised learning methods, which typically rely on well-labeled datasets that are limited in size and scope. Furthermore, these traditional supervised approaches often struggle with generalization when applied to broader, more diverse populations. These challenges highlight the importance of semi-supervised learning for FoG detection, where the approach becomes particularly valuable. By effectively leveraging abundant unlabeled sensor data alongside limited expert-annotated FoG episodes, semi-supervised methods can substantially alleviate the labeling burden while maintaining detection reliability. More importantly, such frameworks enhance model adaptability to individual gait variations and improve robustness against real-world signal variability.

In response to these challenges, we introduce a novel semi-supervised real-time detection method for FoG. In summary, we break the limitations of fully supervised models while ensuring the lowest latency in model prediction of FoG, and achieve efficient real-time monitoring of FoG through an efficient and low-latency prediction pre-FoG stages. Our approach leverages both labeled and unlabeled sensor data, enhancing the model's ability to generalize across different scenarios and datasets. The key contributions of our method are:

- We propose a novel semi-supervised prototype network that dynamically refines class prototypes using both labeled and unlabeled data through an attention-based weighting mechanism, while integrating a metric learning loss to explicitly separate class representations.
- We incorporate a metric learning loss that explicitly optimizes the embedding space. This loss ensures high inter-class separation and low intra-class variance, leading to more distinct and reliable class representations.
- Our method utilize a larger database, comprising a total of 73 participants across four public datasets. And the results demonstrate the robustness and generalizability of our approach.
- We obtained significant results across four distinct datasets, achieving a notable highest sensitivity of 96% and a specificity of 99% and a latency of 0.5 seconds of predicting the onset of FoG 2 to 3 seconds in advance. The overall average accuracy was 91%.

## 2 Related Works

### 2.1 Sensing Modalities for FoG Detection

Currently, a substantial body of research focuses on sensor-based approaches for detecting Freezing of Gait (FoG). These approaches encompass both physiological and kinematic signals. For instance, electroencephalogram (EEG) records cortical activity through scalp electrodes and provides direct insight into neural mechanisms of FoG, although it is costly and intrusive for daily use [8, 14]. Electrocardiogram (ECG) and skin conductance (SC) signals are obtained via chest electrodes or wrist sensors and can capture autonomic responses related to stress and anxiety, but they are indirect markers with limited specificity to FoG [8, 15]. Geophone-based systems offer a low-cost and data-rich approach to gait monitoring by capturing fine-grained vibration signals that directly reflect the interaction between the foot and the ground. However, their deployment still relies on fixed sensor locations and wired connections, which limits their scalability and reduces their practicality in real-world environments [16–18]. Camera-based systems provide a non-contact means of capturing whole-body movement patterns and have demonstrated promise for FoG detection. However, they are highly sensitive to environmental conditions and raise significant privacy concerns. Furthermore, multi-camera systems introduce additional drawbacks such as complex calibration, higher costs, and the need for view synchronization—further limiting their feasibility for scalable, long-term home monitoring [10]. Pressure sensors embedded in insoles or walkways capture gait dynamics with high accuracy, yet they may be uncomfortable for long-term use and are less suitable for large-scale deployment [19, 20]. Inertial measurement units (IMUs), integrating accelerometers (ACC) and gyroscopes (GYRO), offer a low-cost and unobtrusive means of continuously monitoring gait kinematics in daily life, despite challenges such as sensor drift and placement variability [5, 6, 9, 11, 15, 20–25]. Importantly, IMUs are already embedded in widely available consumer devices such as smartphones and smartwatches, and several studies have leveraged these devices as practical data acquisition platforms for FoG detection, further highlighting their feasibility for home-based monitoring [26]. In addition, multimodal approaches that combine complementary signals have been investigated for their potential to enhance detection accuracy [8, 20, 27].

Table 1. Comparative analysis of sensing modalities for FoG detection.

| Modality | Cost | Unobtrusiveness | Data Richness | Deployment Feasibility |
|---|---|---|---|---|
| **IMU** | Low | Wearable | Kinematic time-series | Smartphones; smartwatches |
| **EEG** | High | Cumbersome headset | Complex brain signals | Uncomfortable for long-term |
| **ECG** | Medium | Chest strap | Physiological signals | Chest strap; patches |
| **SC** | Low | Wrist band | Event-driven arousal signals | Wristband |
| **Camera** | Medium | Ambient | Rich visual context | Privacy concerns, fixed view |
| **Geophone** | low | Ambient | Event-based | Suitable for in-home setting |

A systematic comparison of these modalities is presented in Table. 1, highlighting differences in deployment feasibility, data richness, and user comfort. Although multiple sensing modalities have been explored for FoG detection, each entails trade-offs in terms of cost, unobtrusiveness, and deployment feasibility. IMUs, by contrast, provide the most balanced solution, offering unobtrusive, low-cost, and versatile gait monitoring with proven compatibility for integration into widely available consumer devices such as smartphones and smartwatches. On this basis, our study employs IMU-based data to investigate and advance FoG detection.

## 2.2 IMU-based Methods

IMU-based methodologies for detecting FoG have progressively focused on the critical aspect of real-time detection. Notable among these is the approach by Naghavi et al., which utilizes a moving window technique capable of identifying FoG episodes with a latency of approximately two seconds [15]. Similarly, Borzi et al. have employed a convolutional neural network (CNN) strategy, achieving the detection of 32.5% of FoG events 1.3 seconds before FoG onset and 50.0% of events 1.1 seconds prior, as validated on an independent dataset [23]. FoG-Finder system, developed by Koltermann et al., demonstrated an average detection delay of 615 milliseconds during a leave-one-out test involving seven participants [28]. Additionally, Koltermann et al. also developed GaitGuard, a Transformer-based FoG detection system that integrates two ankle-worn IMUs, an Android smartphone, and a vibrotactile feedback device. In a leave-one-subject-out (LOSO) evaluation, the system achieved a FoG detection accuracy of 96.5% [21]. Recently, Chen et al. proposed a two-level FoG recognition algorithm that combines out-of-distribution (OOD) detection with an anomaly detection (AD) module and introduces a Gaussian kernel-based label fusion strategy. However, its limitation lies in the limited ability to recognize diverse daily activities in non-laboratory environments. [25]. Despite this, all these approaches rely on binary classification to predict the imminent occurrence of FoG, enabling the systems to forecast the onset of FoG with a specific delay time, thereby offering potential improvements in patient mobility management and safety.

In the field of FoG detection, numerous studies have leveraged IMU, ACC, and GYRO sensors to develop methods for predicting the FoG stage [9, 20, 22, 23]. However, the majority of these approaches have focused on binary classification, distinguishing between the FoG and non-FoG states.

Despite these advances in IMU-based FoG detection systems, a fundamental limitation persists: the reliance on detecting FoG as it begins or milliseconds prior still leaves little time for preventive intervention. The sequential degradation of gait characteristics preceding a full FoG episode suggests that even earlier prediction may be possible. This has spurred growing interest in identifying a pre-FoG stage, where gait abnormalities are present but a full freeze has not yet occurred.

## 2.3 Pre-FoG stages

Shifting the predictive horizon earlier, recent studies have explored the potential of redefining the pre-FoG stage and utilizing predictive models to anticipate the onset of FoG episodes [5, 6, 8, 24, 29]. By predicting the pre-FoG stage, these approaches aim to provide earlier warnings, thereby alleviating the delay inherent in binary classification-based FoG prediction.

Mazilu et al. defined the pre-FoG area as 3 seconds before the onset of FoG [8]. They developed a decision tree-based model to extract relevant features from ECG and SC signals and achieved an F1-score of 56% in FoG prediction. However, their approach was associated with a prediction delay of up to 2 seconds limiting its advantage. Palmerini et al. defined the pre-FoG stage as 2 seconds before the FoG episode [24]. They employed a linear discriminator to classify pre-FoG and FoG states, attaining a sensitivity of 0.83 and a specificity of 0.67. While this method demonstrated promising results, it faced limitations in accurately distinguishing the normal walking state (non-FoG) from the pre-FoG stage.

Similarly, Naghavi et al also defined the pre-FoG stage as the first 2 seconds before the onset of FoG. Their approach achieved an 88.5% sensitivity for FoG detection across 10 participants using ACC sensors data from DAPHNet dataset [9], but with a minimum delay of 2 seconds [5].

Aiming to further refine the pre-FoG prediction, Zhang et al took a personalized approach. They calculated the slope of each step rhythm in the acceleration signal for individual participants and determined a personalized pre-FoG window length for each. They employed the AdaBoost method to analyze the ACC signals collected from the lower backs of participants at Ruijin Hospital in China. Their analysis achieved an accuracy of 77.9% based on data from 12 patients, resulting in a latency reduction of 0.93 seconds [6].

Recently, we explored the use of an autoencoder-based method to predict the pre-FoG stage, considering various pre-FoG window lengths from 1 to 3 seconds. the model achieved a sensitivity of 0.9 in predicting pre-FoG on a rotating FoG independent dataset from PDTURN dataset [7]. However, the model employed in this method is relatively complex, which may lead to substantial memory consumption.

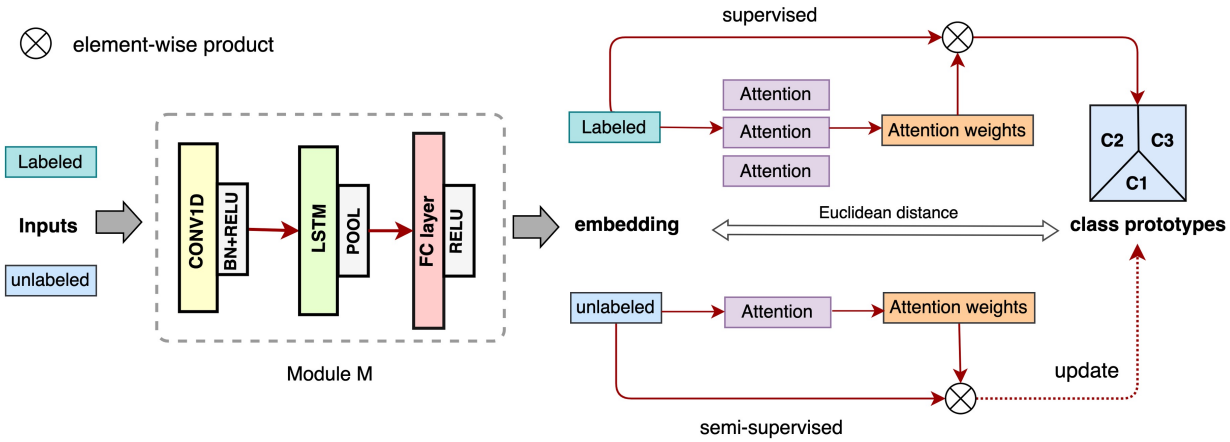


Fig. 1. Model architecture, which encompasses both supervised and unsupervised modes. The dashed line indicates updating the prototype representation using unlabeled data only in a semi-supervised mode.

While the existing research has made significant progress, several limitations persist that require attention. The studies previously reviewed primarily tested their methods on a single dataset, resulting in a lack of external dataset verification, which hinders the generalizability of their findings. To obtain additional dataset verifications, Borzi et al. evaluate their CNN-based method on three distinct ACC datasets and utilizing differential segmentation to mitigate the issue of class imbalance. They successfully detected 52.3% of FoG attacks within 3 seconds prior to onset on an independent test set [23]. However, while their method demonstrates promise, its accuracy requires further improvement.

## 2.4 Semi-supervised learning FoG Detection Methods

Mikov et al. employed a self-labeling strategy combined with online learning and batch selection to iteratively update model parameters, aiming to adapt to the gait patterns of new patients incrementally. However, this approach heavily relies on self-generated labels, which introduces the risk of error accumulation and potential model degradation over time. Moreover, their method is limited to binary classification (FoG vs. non-FoG) and does not support more granular pre-FoG detection [30]. In a recent study, Xia et al. proposed a self-supervised learning approach for predicting early-stage FoG, reporting a sensitivity of 85% and specificity of 95% on the DAPHNet dataset. However, the model's performance in pre-FoG detection remains limited, with specificity dropping to 81.56% and 69.5% in different settings—indicating insufficient robustness for reliable early warning. Moreover, the Transformer-based architecture adopted in their work is computationally complex and resource-intensive, posing challenges for practical deployment on resource-constrained wearable devices [29].

Therefore, to address these limitations, we aim to propose an effective lightweight FoG detection system that enables timely detection of the onset of the pre-FoG stage, so that early intervention and relief of FoG attacks can be achieved.

# 3 Methods

## 3.1 Overview

Our proposed method is primarily based on the prototype network [31], which is designed to learn prototype embedding representations corresponding to three distinct classes: pre-FoG, FoG, and non-FoG. For feature extraction, we employ a hybrid CNN-LSTM network to capture short-term local patterns and their long-term temporal evolution. This architecture was chosen over Transformers due to its superior computational efficiency and lower data requirements. Building upon this foundation, we developed a semi-supervised framework that estimates the class of each new sample by calculating the distance between the sample and the corresponding class prototype, utilizing unlabeled samples for this analysis. This framework operates under two primary modes: the fully supervised mode and the semi-supervised mode. In the fully supervised mode, the model is trained and tested exclusively on labeled data, ensuring a direct evaluation of its performance based on known outcomes. In contrast, the semi-supervised mode incorporates unlabeled data from the test set to enhance the training process, allowing the model to leverage additional information that may improve its predictive capabilities. The following sections will detail the methodologies employed within each of these modes.

## 3.2 Supervised learning

We now have a labeled dataset $D = (\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), (\mathbf{X}_3, y_3), \ldots, (\mathbf{X}_N, y_N)$, where $N$ represents the number of labeled samples and each $\mathbf{X}_k$ is a time series segment. $y_k$ is the corresponding class label for that segment. Specifically, each time series segment $\mathbf{X}_k$ can be represented as $\mathbf{X}_k = [\mathbf{X}_{k,1}, \mathbf{X}_{k,2}, \mathbf{X}_{k,3}, \ldots, \mathbf{X}_{k,N}]$, where $\mathbf{X}_{k,i} \in \mathbb{R}^{L \times F}$. Here, $L$ is the sequence length, and $F$ is the number of features for each time step, different datasets containes different features dimnension as shown in Table. 3.

In the proposed approach, as shown in Fig. 1, a neural network module $M$ is applied to the input data $\mathbf{x}_k$ to extract different features. The module $M$ consists of a convolution layer followed by batch normalization and a ReLU activation function [32], and then process the features through the LSTM [33] to capture temporal dependencies in sequential data to get the feature embeddings $\mathbf{e}_{k,i} \in \mathbb{R}^{L \times 1}$ for each class $k$ and $i^{\text{th}}$ sample, where

$$\mathbf{e}_{k,i} = M(\mathbf{X}_{k,i}). \tag{1}$$

The prototype embedding $\mathbf{c}_k \in \mathbb{R}^{L \times 1}$ is a weighted sum of $\mathbf{e}_{k,i}$:

$$\mathbf{c}_k = \sum_i s_{k,i} \times \mathbf{e}_{k,i}, \tag{2}$$

where the weight $s_{k,i}$ is the weight of the $i^{th}$ sample for class $k$. These weights are trainable parameters, which are learned within the attention block illustrated in Fig. 1. For each class $k$, the process can be represented by:

$$s_{k,i} = \text{softmax}(\mathbf{W}_{2k}^T \tanh(\mathbf{W}_{1k}\mathbf{e}_{k,i})), \tag{3}$$

where $\mathbf{W}_{1k} \in \mathbb{R}^{u \times L}$ and $\mathbf{W}_{2k} \in \mathbb{R}^{u \times K}$ represent trainable parameters that are designed to assign appropriate weights to the feature embeddings corresponding to each class $k$. Here, $u$ denotes the size of the hidden dimension for both parameters, while $K$ signifies the total number of classes. In our method, $K$ is equal to 3.

The classification probability $p_\Theta(y = k \mid \mathbf{x})$ is then calculated based on the distance between $\mathbf{e}_k$ and $\mathbf{c}_k$ for each class $k$:

$$p_\Theta(y = k \mid \mathbf{x}) = \frac{\exp\left(-D(\mathbf{e}, \mathbf{c}_k)\right)}{\sum_i^K \exp\left(-D(\mathbf{e}, \mathbf{c}_i)\right)} \tag{4}$$

where $D(\cdot, \cdot)$ is an Euclidean distance that measures the similarity between the input feature embedding and the class prototype embedding. The closer the input feature embedding is to the prototype embedding of a class, the higher the probability of that class.

### 3.3 Semi-supervised learning

In the semi-supervised setting, we now assume to have a labeled dataset $D = (\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), (\mathbf{X}_3, y_3), \ldots, (\mathbf{X}_N, y_N)$ as well as an unlabeled dataset $D' = \mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3, \ldots, \mathbf{X}'_H$, where $H$ represents the number of unlabeled samples, where the samples do not include the corresponding class labels.

To leverage the information in the unlabeled dataset, we first apply the same feature extraction module $M$ to obtain the embeddings $\mathbf{e}'_j$ for the unlabeled samples:

$$\mathbf{e}'_j = M(\mathbf{X}'_j) \tag{5}$$

Next, we aim to update $\mathbf{c}_k$ for each class $k$ by incorporating the valuable information from the unlabeled data. The updated $\mathbf{c}_k$ can be expressed as:

$$\mathbf{c}_k = \frac{\sum_i s_{k,i} \times \mathbf{e}_{k,i} + \sum_j s'_j \times \mathbf{e}'_j}{2} \tag{6}$$

In this equation, the first term, $\sum_i \mathbf{s}_{k,i} \times \mathbf{e}_{k,i}$, represents the weighted sum of the feature embeddings $\mathbf{e}_{k,i}$ from the labeled data as shown in Equation 2. The second term, $\sum_j s'_j \times \mathbf{e}'_j$, incorporates the feature embeddings $\mathbf{e}'_j$ from the unlabeled data, weighted by $s'_j$, which are also learned during the training process.

The weights $s'_j$ are calculated using a similar attentional structure as in the supervised setting without the class $k$ information:

$$s'_j = \text{softmax}\left(\mathbf{W}_2^{'\top} \tanh\left(\mathbf{W}'_1 \mathbf{e}'_j\right)\right) \tag{7}$$

where $\mathbf{W}'_{1k} \in \mathbb{R}^{u \times L}$ and $\mathbf{W}'_{2k} \in \mathbb{R}^{u \times K}$ are additional trainable parameters that learn to assign appropriate weights to the feature embeddings from the unlabeled data for each class $k$. Then we apply the Equation **??** to predict the probability over class $k$.

### 3.4 Loss

In the proposed semi-supervised learning framework, the overall loss function to be minimized during the training process is defined as a combination of two terms:

$$\mathcal{L}(\Theta) = -\log p_\Theta(y = k \mid \mathbf{x}) - \lambda \cdot \mathcal{L}_{\text{metric}} \tag{8}$$

The first term, $-\log p_\Theta(y = k \mid \mathbf{x})$, represents the negative log-likelihood of the predicted class label $k$ for the given input $\mathbf{x}$. This term encourages the model to learn a set of parameters $\Theta$ that can accurately classify the input samples.

The second term captures the contribution from a metric learning component, which aims to enhance the separability of the class prototype representations. $\lambda$ is a hyperparameter that controls the influence of the metric learning component. $\mathcal{L}_{\text{metric}}$ captures the contribution from the prototype distance, which can be expressed as:

$$\mathcal{L}_{\text{metric}} = \frac{1}{N} \sum_{i=1}^{N} D(\mathbf{c}_i, \mathbf{c}_j) \tag{9}$$

where $N$ is the total number of prototype pairs, and $D(\cdot, \cdot)$ represents the Euclidean distance calculated between different class prototypes $\mathbf{c}_i$ and $\mathbf{c}_j$.

The hyperparameter $\lambda$ controls the relative importance of the metric learning component in the overall loss function, the default value of $\lambda$ is 0.0001. By incorporating this metric learning term, the model is encouraged to learn prototype representations that are more separated in the feature space, leading to improved classification performance.

Therefore, this combined loss leverage both the class prediction likelihood and the metric learning objectives to guide the model towards learning robust and discriminative prototype representations. The negative log-likelihood term drives the model to correctly classify the input samples, while the metric learning component further enhances the separability of the learned prototype embeddings.

### 3.5 Evaluation metrics

The accuracy of a classification model is defined as the ratio of correctly predicted instances to the total instances in the dataset. It can be calculated from the confusion matrix as follows:

$$\text{Accuracy} = \frac{TP}{TP + FN + FP} \tag{10}$$

where $TP$ represents the number of instances where the model correctly identified a FoG episode. $FN$ represents the number of instances where the model failed to detect an actual FoG episode. $FP$ represents the number of instances where the model incorrectly identified a FoG when there was no actual FoG episode present.

Sensitivity measures the ability of the FoG detection system to correctly identify actual episodes of FoG,

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

Specificity assesses the ability of the FoG detection system to correctly identify when FoG is not occurring,

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

where $TN$ represents the number of instances where the model correctly identified a non-FoG event. $FP$ represents the number of instances where the model incorrectly predicted a FoG event when the actual state was non-FoG.

The F1-score is a harmonic mean of precision and recall, which provides a balanced measure of the model's performance. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

where Precision is the fraction of true positive predictions among all positive predictions, and Recall (which is the same as Sensitivity) is the fraction of true positive predictions among all actual positive instances.

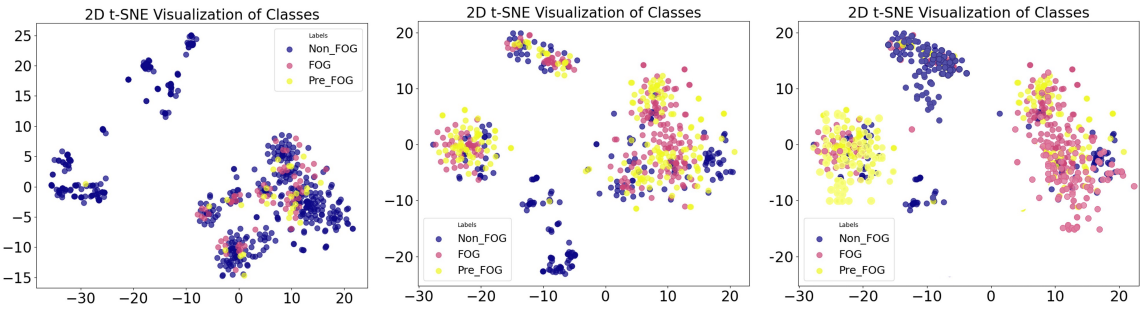


Fig. 2. This figure illustrates the data transformation pipeline, sequentially presenting the raw data, the class-balanced embeddings after SMOTE oversampling, and the final t-SNE visualization of our model's output.

## 4 Experiments and Results

### 4.1 Datasets

This section focuses on the utilization of four distinct public datasets. The first dataset is DAPHNet [9]. The second dataset, refered to as the Charite Parkinson's Gait Dynamics (CPGDD) dataset, was collected by Dvorani et al. at the Department of Neurology, Charité—Universitätsmedizin Berlin [11]. The third dataset, known as the Parkinson's Disease during Turning-In-Place Dataset (PDTURN), compiled by Ribeiro et al., which specifically focuses on recording data during the turning transition period of individuals diagnosed with PD [10]. Finally, the BXHC dataset, collected at Beijing Xuanwu Hospital (BXHC) in China [13].

*4.1.1 DAPHNet.* This dataset consists of ACC signals collected from three anatomical locations: the ankle, knee, and hip, across 10 subjects (7 males and 3 females) at a sampling frequency of 64 Hz. Testing was conducted in the morning during the OFF phase of the drug cycle, which was defined as more than 12 hours post-administration of the last anti-Parkinson's medication. Each session included three distinct walking tasks designed to capture various aspects of daily mobility: (a) walking back and forth in a straight line along the laboratory corridor, (b) walking randomly within the reception hall, and (c) performing activities that simulate daily living (ADL) [9].

*4.1.2 CPGDD.* This dataset focuses on gait phase detection using foot motion data recorded by inertial sensors positioned on the dorsum of the foot, with a sampling frequency of 200 Hz. This dataset includes 16 patients aged between 50 and 82 years old (13 males and 3 females). The experimental protocol began with patients in a seated position, who were instructed to stand up, walk 1 meter, and then execute two complete 360° rotations in both directions within a marked square. After completing these movements, participants walked to the door, opened it, and exited the room [11].

*4.1.3 PDTURN.* This dataset involved 35 PD patients, comprising 16 women and 19 men, aged 44 to 84 years. Each participant wore an IMU device on the calf of the most affected side of the body while performing three in-situ rotation tests at a sampling frequency of 128 Hz. During each test, participants stood and alternated 360° turns to the right and left for a duration of 2 minutes, at a self-selected speed [10].

*4.1.4 BXHC.* This dataset collected data from 12 PD patients (6 males and 6 females) in a drug-free state. Each patient was asked to complete four tasks: a quarter turn, a U-turn, and navigating around an obstacle. Data collection included EEG, EMG, ECG, skin conductance, and ACC measurements, with an sampling frequency of

Table 2. Datasets Description

| Dataset | Frequency (Hz) | Subjects | Attacks | Location | Activity | Sensor |
|---|---|---|---|---|---|---|
| DAPHNet [9] | 64 | 10 | 237 | ankle, knee, hip | straight walking, random walking, ADL | ACC |
| CPGDD [11] | 128 | 16 | 282 | foot | standing, walking, 360° turning, door opening | ACC, GYRO |
| PDTURN [10] | 200 | 35 | 173 | ankle | repeated 360° turning | ACC, GYRO |
| BXHC [13] | 500 | 12 | 324 | arm, shank, waist | Quarter turn, U-turn, obstacle navigation | ACC, GYRO |

Table 3. Model Performance on Independent Datasets

| Dataset | Feat_dim | pre-FoG (s) | Sensitivity | Specificity | F1 Score | Avg. Accuracy | Semi | Gain | Time(ms) /sample |
|---|---|---|---|---|---|---|---|---|---|
| DAPHNet [9] | 9 | 3s | 0.96 | 0.99 | 0.87 | 0.91 | ✓ | - | $0.14_{\pm 0.01}$ |
| | | | 0.97 | 0.98 | 0.86 | 0.91 | × | - | |
| | | 2s | 0.99 | 0.99 | 0.88 | 0.89 | ✓ | - | |
| | | | 0.97 | 0.99 | 0.88 | 0.89 | × | - | |
| CPGDD [11] | 6 | 3s | 0.87 | 0.96 | 0.86 | **0.85** | ✓ | ↑ 0.02 | $0.3_{\pm 0.05}$ |
| | | | 0.81 | 0.88 | 0.85 | 0.83 | × | - | |
| | | 2s | 0.92 | 0.96 | 0.84 | **0.85** | ✓ | ↑ 0.01 | |
| | | | 0.89 | 0.96 | 0.83 | 0.84 | × | - | |
| PDTURN [10] | 6 | 3s | 0.92 | 0.97 | 0.81 | **0.83** | ✓ | ↑ 0.02 | $0.23_{\pm 0.01}$ |
| | | | 0.92 | 0.97 | 0.83 | 0.81 | × | - | |
| | | 2s | 0.87 | 0.98 | 0.83 | **0.82** | ✓ | ↑ 0.01 | |
| | | | 0.86 | 0.97 | 0.80 | 0.81 | × | - | |
| BXHC [13] | 24 | 3s | 0.94 | 0.97 | 0.87 | **0.82** | ✓ | ↑ 0.01 | $0.74_{\pm 0.01}$ |
| | | | 0.93 | 0.97 | 0.87 | 0.81 | × | - | |
| | | 2s | 0.96 | 0.97 | 0.86 | **0.83** | ✓ | ↑ 0.01 | |
| | | | 0.95 | 0.97 | 0.85 | 0.82 | × | - | |

500 Hz. Four ACCs were strategically placed on the lateral tibia of both legs, the fifth lumbar vertebra of the waist, and the left arm to capture comprehensive movement data.

Table 2 provides a detailed summary of the four datasets, including the number of FoG attacks, the locations and types of different sensors. It is important to highlight that the datasets used in various studies differ in terms of collection locations and sensor types, resulting in varying input feature dimensions. For instance, Zhang et al. [13] utilized 3D-ACC and 3D-GYRO data from four anatomical locations—left shank, right shank, arm, and waist—yielding a total input dimension of 24 (4 locations × 6 dimensions each). In contrast, Bachlin et al. [9] focused solely on accelerometer data collected from three different locations, resulting in an input dimension of

9 (3 locations × 3 dimensions). Similarly, Dvorani et al. and Ribeiro et al. collected data from only two sensors located in a single anatomical region, which resulted in a uniform input dimension of 6 across the datasets [10, 11].

### 4.2 Data Preprocessing

Each original dataset includes binary labels, with a value of 1 indicating the presence of FoG and a value of 0 denoting the absence of FoG, thereby distinguishing between the state of FoG occurrence and the normal state. To enhance the predictive power of pre-FoG onset, we specify a pre-FoG stage on these existing datasets. Some previous studies define the pre-FoG stage as a time interval of 1 to 5 seconds [5, 7, 8, 24, 29], and Zhang et al. adopted personalized annotations with different lengths for each participant [6]. To ensure consistency in data processing, we standardized the pre-FoG labels across the four datasets. And for each dataset, we established two distinct datasets, designating the intervals of 2 seconds and 3 seconds prior to the onset of FoG as pre-FoG.

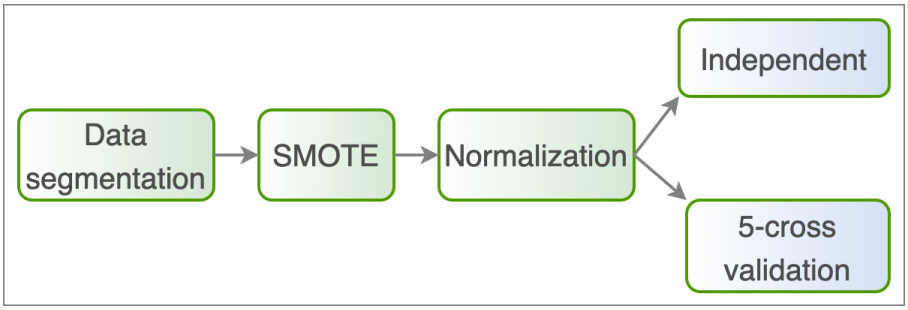


Fig. 3. Data processing flow

After relabeling the dataset, we proceeded with a comprehensive data preprocessing phase, as illustrated in Fig 3. Initially, we performed a segmentation step for each dataset, setting the segmentation window size to match the sampling frequency. For instance, with a sampling frequency of 128 Hz, we utilized a window size of 128. The step size was defined as half of the window length, and the data was segmented continuously by sliding the window across the dataset. For each segment, the label attribute was assigned based on the label with the most time points within that segment. For example, if more than 50% of the time points in a segment were labeled as FoG, the segment was then labeled as FoG.

However, after data segmentation, the distribution of different categories of clips is extremely imbalanced, with non-FoG usually accounting for the majority. Given the temporal dynamics associated with the onset of FoG and the imbalanced distribution of pre-FoG relative to non-FoG instances, we employed the Synthetic Minority Oversampling Technique (SMOTE) [34] to generate synthetic samples for the minority class. Fig. 2 presents a t-SNE visualization of the data categories before and after the application of SMOTE. The left panel illustrates the pre-SMOTE sampling distribution, where non-FoG instances dominate, while the right panel demonstrates a more balanced data distribution post-SMOTE. Additionally, to address the larger category of non-FoG, we conducted appropriate downsampling to preserve the essential characteristics of the data distribution while facilitating the learning of the corresponding time series features.

To enhance the model's robustness to real-world variability and potential sensor noise, we employed a comprehensive data augmentation strategy during the training phase. The time-series windows were subjected to several transformations, including jittering by adding small random perturbations, scaling through multiplication by random factors, and time warping achieved via random cropping or slight temporal stretching. These operations

expanded the diversity of the training data and encouraged the model to learn representations that remain stable under minor temporal misalignments and amplitude fluctuations. Following these foundational preprocessing steps, similar to the approach outlined by Zhang et al. [6], we implemented two primary evaluation schemes: (1) an independent verification using 20% of the data, where cross-validation is conducted for each individual; and (2) a 5-fold cross-validation approach applied across all subjects. Importantly, we conducted these experiments in both fully supervised settings, where all training data was labeled, as well as semi-supervised settings, where the test data was treated as unlabeled data and incorporated into the model learning process. This comprehensive evaluation approach provided a robust and reliable assessment of the FoG detection capabilities of our proposed methods. Prior to these experiments, we also applied normalization to the datasets, ensuring that normalization procedures were conducted separately on the training and test sets to prevent any potential information leakage.

## 4.3 Results

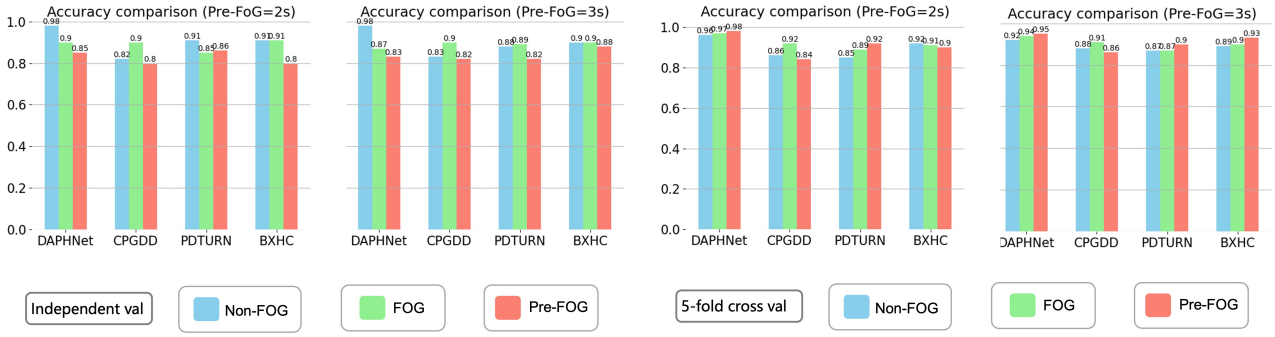


Fig. 4. The accuracy distribution for non-FoG, FoG, and pre-FoG across independent test experiments (left) and 5-fold cross-validation (right). In each experiment, the cases of pre-FoG are categorized separately for durations of 2 seconds and 3 seconds.

*4.3.1 Independent validation.* Table 3 presents a summary of the performance metrics of our model, as evaluated on four independent datasets. Key indicators highlighted include sensitivity, specificity, and F1 score for the pre-FoG segments, as well as the average accuracy across all classes and the inference time per sample. Experiments were conducted with pre-FoG durations of 2 seconds and 3 seconds, respectively.

It is noteworthy that all datasets exhibit high sensitivity and specificity for pre-FoG detection in both fully supervised and semi-supervised scenarios, underscoring the robustness of our method. We further investigate the impact of semi-supervised learning on the accuracy across different datasets. Specifically, we observe that for three datasets: CPGDD, PDTURN, and BXHC, the semi-supervised model enhances mean accuracy by 0.01 to 0.02 compared to the fully supervised mode, while the pre-FoG F1 score improves by 0.1 to 0.3. While the absolute improvement in accuracy may appear modest, it is important to emphasize that this gain was achieved in a semi-supervised setting—where the model utilizes both limited labeled data and additional unlabeled samples—compared to a fully-supervised baseline with the same architecture. This demonstrates the ability of our approach to effectively extract useful information from unlabeled data, a challenging yet clinically meaningful scenario. Notably, this improvement is more pronounced when the length of pre-FoG is 3 seconds, highlighting the significance of our semi-supervised approach in leveraging information from unlabeled data.

In contrast, the results for the DAPHNet dataset do not demonstrate a significant enhancement through the semi-supervised method, as the performance indicators for both fully supervised and semi-supervised modes

Table 4. Cross-validation Performance Metrics

| Dataset | Method | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| DAPHNet | Bachlin et al. [9] | 0.87 | 0.96 | 0.95 |
| | Naghavi et al. [5] | 0.73 | 0.82 | - |
| | Kleanth et al. [35] | 0.78 | 0.89 | 0.89 |
| | Orphaned et al. [36] | 0.83 | 0.91 | 0.88 |
| | Xia et al. [29] | 0.85 | 0.95 | 0.94 |
| | **Ours** | **0.96** | **0.99** | **0.98** |
| CPGDD | Dvorani et al. [11] | 0.86 | 0.80 | 0.85 |
| | **Ours** | **0.92** | **0.97** | **0.85** |
| PDTURN | Yang et al. [37] | 0.81 | 0.87 | - |
| | Dimoudis et al. [38] | 0.97 | 0.97 | - |
| | **Ours** | **0.97** | **0.99** | **0.92** |

remain comparable. It is important to note that the overall accuracy of the DAPHNet dataset is considerably higher than that of the other datasets, achieving a peak F1 score of 0.87 and a peak mean accuracy of 0.91. Nevertheless, our method still exhibits advantages within this dataset.

These findings indicate that our simplified network architecture effectively learns complex features and accurately detects pre-FoG events in a fully supervised setting. Additionally, we observed processing speeds ranging from 0.2 to 0.8 milliseconds per sample on a 32GB V100 GPU, suggesting that the actual inference time of the algorithm is brief and negligible. The overall time delay associated with the application of the algorithm is contingent upon the size of the gait window, which is typically half the total window duration, averaging around 0.5 seconds. Consequently, our model is capable of responding to FoG warnings in real time with a delay of less than 0.5 seconds, enabling the system to provide warnings 2 to 3 seconds in advance.

Furthermore, we assessed the algorithm's recognition capabilities for both non-FoG and FoG events. As illustrated in Fig. 4, when the pre-FoG duration is set to 2 seconds, the recognition results across three categories for different datasets reveal that the algorithm demonstrates a high level of accuracy in distinguishing between non-FoG and FoG states. Specifically, the accuracy for non-FoG classifications is at least 0.83, while the accuracy for FoG classifications reaches a minimum of 0.88. Despite the inherent class imbalance, and even after employing the SMOTE technique to address this issue, the algorithm's ability to identify non-FoG and FoG events remains significantly higher than its capacity to recognize pre-FoG occurrences.

In addition, we evaluated the algorithm's ability to identify non-FoG and FoG events. As shown in Figure 4, when the pre-FoG duration is set to 2 seconds, the recognition results of the three categories of different datasets show that the algorithm exhibits high accuracy in distinguishing non-FoG and FoG states. Specifically, the accuracy of non-FoG classification is at least 0.9, while the accuracy of FoG classification reaches a minimum of 0.85. When the pre-FoG duration is 3 seconds, the accuracy of non-FoG reaches at least 0.83, while the accuracy of FoG reaches a minimum of 0.87. Despite the inherent class imbalance, even after adopting the SMOTE technique to address this issue, the algorithm's ability to identify non-FoG and FoG events is still significantly higher than its ability to identify pre-FoG events.

These results provide robust evidence for the accuracy of our method when applied to an independent validation set. The consistent performance observed across various metrics underscores the reliability and effectiveness of our approach in accurately detecting FoG-related events.

Table 5. Parameters Used for Each Dataset

| Parameter | DAPHNet | CPGDD | PDTURN | BXHC |
|---|---|---|---|---|
| Window/Step | 64/32 | 200/100 | 128/64 | 500/250 |
| Epochs | 50 | 30 | 40 | 50 |
| Batch Size | 32 | 32 | 32 | 32 |
| Feature Dim | 9 | 6 | 6 | 24 |
| Learning Rate | 0.001 | 0.001 | 0.001 | 0.001 |
| Optimizer | Adam | Adam | Adam | Adam |

*4.3.2 Cross validation.* Furthermore, we conducted a 5-fold cross-validation on the entire subjects to evaluate the robustness and generalizability of our model. This process involved dividing the dataset into five equal subsets. The model was trained on four of these subsets, while the remaining subset served as the test set. This training and testing procedure was repeated five times, such that each subset was utilized as the test set exactly once. The final accuracy was calculated by averaging the accuracies obtained from all five test sets.

The results of this cross-validation are presented in the right half of Fig. 4. These findings indicate that our model does not exhibit signs of overfitting. Moreover, the performance metrics are superior to those derived from the scheme independent experiment, which was developed based on patient-specific cross validation. This suggests that our model operates independently of particular patient characteristics. Therefore, the model demonstrates efficacy in accurately identifying non-FoG, FoG, and pre-FoG events. Such a level of performance underscores the model's potential for real-world applications, where it could substantially enhance the understanding and management of gait disorders.

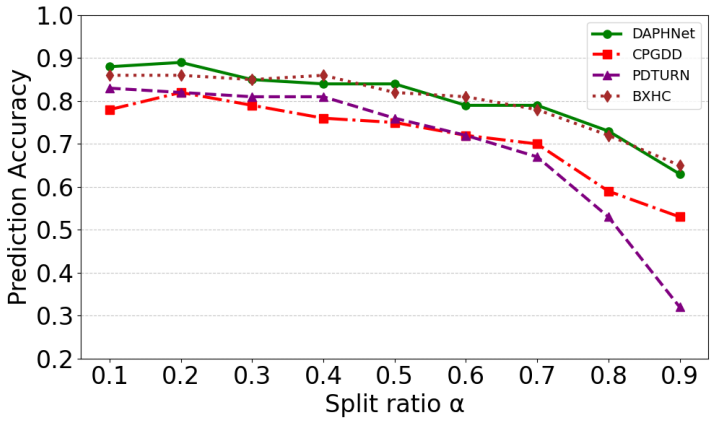


Fig. 5. The impact of different split ratios ($\alpha$) on the average accuracy across four distinct datasets. Specifically, an $\alpha$ value of 0.1 corresponds to a scenario in which 10% of the dataset as unlabeled data

*4.3.3 Baseline comparison.* Comparing our method to existing studies presents challenges due to the reliance on numerous non-public datasets and codes in the literature [15, 20, 23]. Additionally, most studies focus exclusively on binary classification, distinguishing only between FoG and non-FoG. While some studies utilize the publicly available Cupid dataset, they incorporate input signals such as ECG and SC, which fall outside the scope of

Table 6. Daphnet personalized pre-FoG window length

| ID | Gender | Age | Personalized Pre-FoG (s) |
|---|---|---|---|
| 01 | M | 66 | 2.4 |
| 02 | M | 67 | 1.3 |
| 03 | M | 59 | 1.7 |
| 04 | M | 62 | 3.1 |
| 05 | M | 75 | 1.7 |
| 06 | F | 63 | 2.3 |
| 07 | M | 66 | 1.4 |
| 08 | F | 68 | 2.2 |
| 09 | M | 73 | 2.7 |
| 10 | F | 65 | 3.0 |
| **Mean± STD** | - | **66.4 ±4.8** | **2.2 ± 0.65** |

our analysis [8, 15, 23]. Consequently, we primarily compare the detection of FoG segments with methods that employ k-fold cross-validation on these public datasets, as few approaches specifically predict pre-FoG events [5, 29]. Table 4 illustrates that our method demonstrates significant advantages across various datasets using cross validation across all the subjects, highlighting its efficacy.

In relation to BXHC dataset, which is a recently released multimodal dataset, Xia et al. conducted 5-fold cross-validation tests and reported sensitivities of 87% and specificities of 86%. However, these results are lower than our findings of 96% sensitivity and 97% specificity. It is important to note that their study utilized a multimodal input comprising EEG and SC data, whereas our analysis is based solely on IMU sensor data from BXHC. Therefore, we refrained from making direct comparisons with the results from BXHC.

## 4.4 Experimental settings

For the independent validation, the dataset was randomly divided into training and test sets for each subject in an 8:2 ratio. The overall accuracy was calculated as the average across all subjects in the independent test dataset. In the cross-validation experiments, classification accuracy was determined as the average from ten distinct test participants. All models were trained on a V100 GPU with 32 GB of VRAM. Given the use of four different public datasets, we did not employ unified parameters across all datasets; variations in sensor sampling frequencies necessitated adjustments in our window size settings. Detailed usages of the parameters utilized for each of the four datasets are presented in Table 5.

Table 7. Comparison of model performance between uniform pre-FoG windows and personalized windows on the Daphnet.

| Metric | Uniform (2s) | Uniform (3s) | Personalized | Difference | p-value) |
|---|---|---|---|---|---|
| Sensitivity | 0.98 | 0.96 | 0.97 | -0.01 | 0.05 |
| Specificity | 0.99 | 0.99 | 0.99 | 0.0 | 0.02 |
| F1-Score | 0.88 | 0.87 | 0.88 | 0.0 | 0.22 |
| Average Accuracy | 0.89 | 0.91 | 0.90 | +0.01 | 0.50 |

Note: The *Difference* and *p-value* columns quantify the change between the Personalized Windows and the 2-s uniform window result for accuracy, P-values are obtained from a paired t-test.

## 4.5 Ablation Study

*4.5.1 Pre-Fog Labeling.* A central question in FoG prediction is whether a uniform window can effectively capture pre-FoG patterns across a diverse population, or if a personalized approach is necessary to account for individual behavioral differences. To this end, we compared our fixed window approach (2s and 3s) against a personalized window strategy on the Daphnet dataset. The personalized windows were determined on a per-subject basis following the methodology established by Zhang et al. [6]. Specifically, we determined personalized pre-FoG lengths based on the vertical axis data from the waist-mounted sensor for each subject. This process involved calculating the slope of the step rhythm from the accelerometer signal for each individual participant, followed by applying an adaptive threshold to identify the subject-specific duration of gait deterioration that most consistently precedes a FoG episode. The detailed personalized pre-FoG window lengths for each subject are documented in Table 6.

The results of this comparative analysis are presented in Table 7. Our findings show that the use of personalized pre-FoG window lengths did not lead to statistically significant improvements over fixed windows. Sensitivity with personalized windows was slightly lower than with the 2s window and only marginally higher than with the 3s window, while specificity and F1-score remained essentially unchanged. Notably, the fixed 3s window consistently achieved competitive performance across all metrics (sensitivity, specificity, F1-score, and average accuracy), highlighting its robustness and stability across the cohort.

Table 8. Model performance (Average Accuracy) under extremely low labeling rates.

| Dataset | 100% Labeled | 10% Labeled | 5% Labeled | 1% Labeled |
|---|---|---|---|---|
| DAPHNet | 0.96 | 0.60 | 0.61 | 0.60 |
| CPGDD | 0.92 | 0.53 | 0.51 | 0.53 |
| PDTURN | 0.92 | 0.35 | 0.29 | 0.22 |
| BXHC | 0.89 | 0.61 | 0.58 | 0.59 |

*4.5.2 Testing size.* The previous experimental assumptions were based on a test set ratio of 0.2, representing both labeled and test data. However, in practical applications, the proportion of unlabeled data is often much larger than 0.2, as significant amounts of unlabeled data are generated in real-life scenarios.

Therefore, we tested our model across a range of test set ratios, from 0.1 to 0.9 as shown in Fig. 5. The average accuracy for non-FoG, pre-FoG, and FoG categories was calculated on the $\alpha$ test set, where $\alpha$ represents the percentage of the unlabeled data.

To further stress-test the robustness of our semi-supervised framework, we evaluated its performance under extremely low labeling rates (10%, 5%, and 1%) across all datasets. This scenario simulates a highly challenging yet practical setting where very few annotated samples are available. The results revealed a notable divergence in model resilience. On the PDTURN dataset, which focuses almost exclusively on 360-degree turning tasks, the model's accuracy dropped sharply to only 22% at the 1% labeling rate—the lowest among all datasets under this extremely low-resource setting. We attribute this significant performance drop to the dataset's inherent lack of movement diversity; its singular focus on a single, repetitive motion pattern severely limits the amount of transferable knowledge the model can learn from the very small labeled subset, thereby hindering its ability to effectively leverage the unlabeled data.

In stark contrast, our method demonstrated strong robustness on the more movement-heterogeneous datasets such as DAPHNet, CPGDD, and BXHC. Even at the extreme 1% labeling rate, the model retained average accuracy above 50%, specifically achieving 60%, 57%, and 52%, respectively. This maintained performance highlights a key

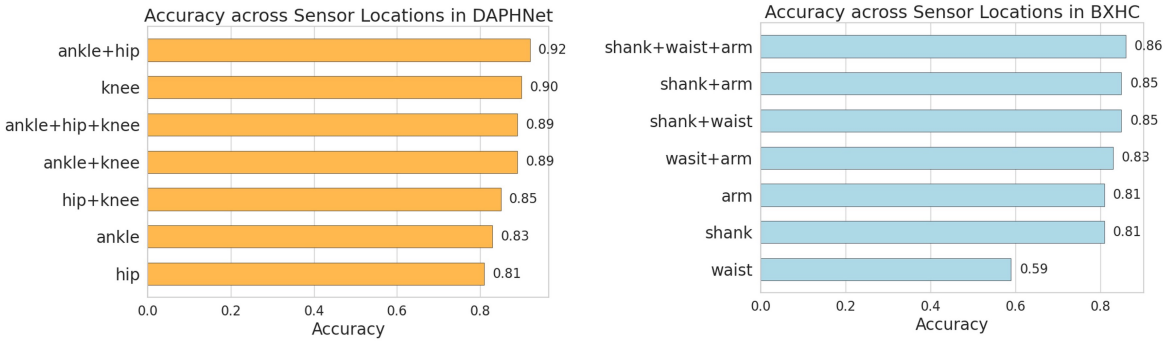


Fig. 6. The average accuracy of different sensor position combinations in the two datasets is compared. The yellow one represents DAPHNet and the blue one represents BXHC.

advantage of our prototype-based semi-supervised approach: its capacity to extract meaningful and generalizable patterns from very few labeled examples when presented with diverse movement types. These results strongly affirm the efficacy and practical applicability of our method in real-world scenarios where acquiring large annotated datasets is infeasible.

Our findings indicate that as the proportion of unlabeled data increases, the model experiences varying degrees of accuracy decline across the four datasets. Notably, the accuracy trends for some datasets, such as DAPHNet and BXHC, stabilize at an average accuracy of over 0.6, regardless of the increasing proportion of unlabeled data. This suggests the potential for effectively incorporating unlabeled data in practical applications.

In the case of the CPGDD dataset, the accuracy remains above 0.5. Conversely, the accuracy for the PD Truning dataset shows a significant decline once the ratio exceeds alpha = 0.7; however, it maintains an accuracy of over 0.6 for values below this threshold. Overall, these experiments demonstrate the effectiveness of our method for evaluating unlabeled data in large proportions, highlighting its practical applicability.

*4.5.3 Sensor locations.* Due to the variety of body positions used in Datasets 1 and 4, we performed an ablation study to explore the impact of various sensor combinations, and the results are shown in Fig. 6. Notably, the lowest accuracy was achieved using either the hip or waist sensors alone, likely due to the relatively far distances between these locations from the limbs, which led to poor performance. Both datasets consistently demonstrate that multi-position sensor configurations outperform single sensors, with BXHC clearly following this trend; in particular, the combination of three positions yielded the highest accuracy, highlighting the significant advantages of multimodal position information. This enhancement can be attributed to the complementary data captured from various body parts.

In contrast, DAPHNet revealed an unexpected performance advantage for knee sensors when used alone, suggesting that this particular sensor may be more sensitive or relevant for detecting specific types of movements or conditions. Furthermore, the ankle and hip combinations achieved notable success, outperforming even the three-position configuration. This observation suggests that the effectiveness of sensor arrangements may be influenced by the specific characteristics and distribution of the data within each dataset. The findings motivate important considerations for real-world applications, where maximizing accuracy using a minimum number of sensors can lead to more efficient and effective systems. This approach not only reduces costs and complexity but also minimizes the potential for measurement errors and data processing challenges associated with excessive sensor data.

## 5 Discussions

This study builds upon the concept of pre-FoG introduced in previous research, focusing on the identification and prediction of pre-FoG states. By integrating semi-supervised learning techniques, we aim to leverage unlabeled data for real-time predictions of FoG occurrences. Notably, our approach standardizes pre-FoG marking, designating all FoG stages as occurring either 2 or 3 seconds prior to actual onset. While this uniformity simplifies the labeling process, it overlooks the individualized characteristics of FoG manifestations. As highlighted by [6], the pre-FoG experience can vary significantly among individuals and across different FoG stages. Although personalized labeling could enhance the accuracy of our predictions, it is a labor-intensive process that poses challenges as the dataset expands. Therefore, future research must explore methods that balance personalized labeling with standardized approaches to improve the overall efficacy of pre-FoG stage identification.

Throughout this study, we assume that the model runs online within a cloud infrastructure, with real-time predictions transmitted to the patient's device via the cloud, as is the case with most sensor network-based applications [37]. However, practical implementations may face latency issues arising from communication delays or hardware system outages, which can negatively impact the timeliness of predictions. Given these considerations, our approach actually favors a soft real-time framework. One potential solution is to adapt the model for offline use by integrating it into mobile devices such as smartphones and smartwatches. While this approach mitigates dependence on network connectivity, it introduces challenges related to environmental noise and high computational memory requirements. Addressing these factors will be crucial for the successful deployment of our model.

Despite the encouraging results, our study did not explicitly examine variability across different patient populations, where heterogeneity in disease progression and gait manifestations may substantially affect prediction accuracy. Second, the current framework assumes the continuous availability of wearable sensor data, yet real-world applications often encounter scenarios of partial sensor failure or missing modalities. Finally, our experiments were primarily conducted under relatively controlled conditions, which may not fully capture the complexities of daily living environments, such as crowded spaces, uneven terrains, or external distractions. These limitations highlight the need for future research to evaluate the model's robustness across diverse cohorts, to design strategies for handling incomplete sensor inputs, and to validate performance under more ecologically valid conditions.

Moreover, it is important to emphasize that the implications of this work extend beyond FoG prediction in Parkinson's disease. The underlying semi-supervised framework for analyzing pre-motor impairment patterns holds significant promise for broader gait analysis and other neurological disorders. The core capability of our model—to learn informative representations from limited labels and abundant unlabeled data—can be translated to the detection of various gait abnormalities and the monitoring of continuous changes in gait quality over time, which is crucial for assessing overall mobility and rehabilitation efficacy. Furthermore, the methodology is inherently disorder-agnostic and could be adapted to other conditions where gait is a key biomarker, such as Huntington's disease, cerebral palsy, muscular dystrophy, and post-stroke recovery. By demonstrating a practical, data-efficient framework for deriving clinical insights from continuous sensor data, this work contributes to the burgeoning field of digital biomarkers, helping to pave the way for scalable and personalized neurological health monitoring.

In summary, our future work will concentrate on the practical implementation of both online and offline systems, striving to enhance the model's adaptability and performance in diverse real-world settings.

## 6 Conclusion

In this study, we introduced a novel semi-supervised method leveraging a prototype network that effectively harnesses valuable information derived from unlabeled data to enhance the detection of pre-FOG events across

various public datasets. By employing a window size of 1 second and a moving step of half a second, our system achieves an impressive reasoning speed of less than 1 millisecond per sample. The predictive accuracy of our model exceeds 0.8 up to 2 to 3 seconds in advance of actual FOG events, underscoring its potential for real-time detection capabilities. This advancement signifies a significant departure from fully supervised methodologies, as our system is designed to operate in an online detection mode, enabling continuous monitoring and timely intervention. Furthermore, our proposed method not only enhances the model's robustness by incorporating unlabeled data but also alleviates the reliance on extensive labeled datasets, which are often difficult and costly to obtain. As we advance this research, we envision further enhancements to the offline potential applications in wearable technology, which could facilitate the deployment of our system in everyday life.

## References


[1] John G Nutt, Bastiaan R Bloem, Nir Giladi, Mark Hallett, Fay B Horak, and Alice Nieuwboer. Freezing of gait: moving forward on a mysterious clinical phenomenon. *The Lancet Neurology*, 10(8):734–744, 2011.

[2] Ana Lígia Silva de Lima, Luc JW Evers, Tim Hahn, Lauren Bataille, Jamie L Hamilton, Max A Little, Yasuyuki Okuma, Bastiaan R Bloem, and Marjan J Faber. Freezing of gait and fall detection in Parkinson's disease using wearable sensors: a systematic review. *Journal of neurology*, 264:1642–1654, 2017.

[3] Rachel Chee, Anna Murphy, Mary Danoudis, Nellie Georgiou-Karistianis, and Robert Iansek. Gait freezing in Parkinson's disease and the stride length sequence effect interaction. *Brain*, 132(8):2151–2160, 2009.

[4] Robert Iansek, Frances Huxham, and Jennifer McGinley. The sequence effect and gait festination in Parkinson disease: Contributors to freezing of gait? *Movement disorders: official journal of the Movement Disorder Society*, 21(9):1419–1424, 2006.

[5] Nader Naghavi and Eric Wade. Prediction of freezing of gait in Parkinson's disease using statistical inference and lower–limb acceleration data. *IEEE transactions on neural systems and rehabilitation engineering*, 27(5):947–955, 2019.

[6] Yuqian Zhang, Weiwu Yan, Yifei Yao, Jamirah Bint Ahmed, Yuyan Tan, and Dongyun Gu. Prediction of freezing of gait in patients with parkinson's disease by identifying impaired gait patterns. *IEEE transactions on neural systems and rehabilitation engineering*, 28(3):591–600, 2020.

[7] Luyao Yang, Osama Amin, and Basem Shihada. Real-time freezing of gait detection: Harnessing advanced AI for better mobility. In *Proc. IEEE Int. Conf. E-health Networking, Application and Services*, 2024.

[8] Sinziana Mazilu, Alberto Calatroni, Eran Gazit, Anat Mirelman, Jeffrey M Hausdorff, and Gerhard Tröster. Prediction of freezing of gait in Parkinson's from physiological wearables: an exploratory study. *IEEE journal of biomedical and health informatics*, 19(6):1843–1854, 2015.

[9] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2009.

[10] Caroline Ribeiro De Souza, Runfeng Miao, Júlia Ávila De Oliveira, Andrea Cristina De Lima-Pardini, Débora Fragoso De Campos, Carla Silva-Batista, Luis Teixeira, Solaiman Shokur, Bouri Mohamed, and Daniel Boari Coelho. A public data set of videos, inertial measurement unit, and clinical scales of freezing of gait in individuals with Parkinson's disease during a turning-in-place task. *Frontiers in Neuroscience*, 16:832463, 2022.

[11] Ardit Dvorani, Vivian Waldheim, Magdalena CE Jochner, Christina Salchow-Hömmen, Jonas Meyer-Ohle, Andrea A Kühn, Nikolaus Wenger, and Thomas Schauer. Real-time detection of freezing motions in Parkinson's patients for adaptive gait phase synchronous cueing. *Frontiers in Neurology*, 12:720516, 2021.

[12] Hantao Li. Multimodal dataset of freezing of gait in Parkinson's disease. *Mendeley Data*, 3:2021, 2021.

[13] Wei Zhang, Zhuokun Yang, Hantao Li, Debin Huang, Lipeng Wang, Yanzhao Wei, Lei Zhang, Lin Ma, Huanhuan Feng, Jing Pan, et al. Multimodal data for the detection of freezing of gait in Parkinson's disease. *Scientific data*, 9(1):606, 2022.

[14] Alka Rachel John, Zehong Cao, Hsiang-Ting Chen, Kaylena Ehgoetz Martens, Matthew Georgiades, Moran Gilat, Hung T Nguyen, Simon JG Lewis, and Chin-Teng Lin. Predicting the onset of freezing of gait using EEG dynamics. *Applied Sciences*, 13(1):302, 2022.

[15] Nader Naghavi, Aaron Miller, and Eric Wade. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: Addressing the class imbalance problem. *Sensors*, 19(18):3898, 2019.

[16] Yiwen Dong, Sung Eun Kim, Kornél Schadl, Peide Huang, Wenhao Ding, Jessica Rose, and Hae Young Noh. In-home gait abnormality detection through footstep-induced floor vibration sensing and person-invariant contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7054–7067, 2024.

[17] Renfei Sun, Zhiyong Wang, Kaylena Ehgoetz Martens, and Simon Lewis. Convolutional 3d attention network for video based freezing of gait recognition. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2018.

[18] Yuki Kondo, Kyota Bando, Ippei Suzuki, Yuta Miyazaki, Daisuke Nishida, Takatoshi Hara, Hideki Kadone, and Kenji Suzuki. Video-based detection of freezing of gait in daily clinical practice in patients with parkinsonism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2250–2260, 2024.

[19] Scott Pardoel, Julie Nantel, Jonathan Kofman, and Edward D Lemaire. Prediction of freezing of gait in Parkinson's disease using unilateral and bilateral plantar-pressure data. *Frontiers in neurology*, 13:831063, 2022.

[20] Scott Pardoel, Gaurav Shalin, Julie Nantel, Edward D Lemaire, and Jonathan Kofman. Early detection of freezing of gait during walking using inertial measurement unit and plantar pressure distribution data. *Sensors*, 21(6):2246, 2021.

[21] Kenneth Koltermann, John Clapham, GinaMari Blackwell, Woosub Jung, Evie N Burnet, Ye Gao, Huajie Shao, Leslie Cloud, Ingrid Pretzer-Aboff, and Gang Zhou. Gait-guard: Turn-aware freezing of gait detection for non-intrusive intervention systems. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 61–72. IEEE, 2024.

[22] Thomas Bikias, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, and Leontios J Hadjileontiadis. DeepFoG: an IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. *Frontiers in Robotics and AI*, 8:537384, 2021.

[23] Luigi Borzì, Luis Sigcha, Daniel Rodríguez-Martín, and Gabriella Olmo. Real-time detection of freezing of gait in Parkinson's disease using multi-head convolutional neural networks and a single inertial sensor. *Artificial intelligence in medicine*, 135:102459, 2023.

[24] Luca Palmerini, Laura Rocchi, Sinziana Mazilu, Eran Gazit, Jeffrey M Hausdorff, and Lorenzo Chiari. Identification of characteristic motor patterns preceding freezing of gait in Parkinson's disease using wearable sensors. *Frontiers in neurology*, 8:394, 2017.

[25] Qianhong Chen, Zhonglue Chen, Fei Zhang, Shengdi Chen, Kang Ren, and Nannan Lu. Dual-level freezing of gait recognition. *IEEE Sensors Journal*, 2025.

[26] Yu-Sun Min, Tae-Du Jung, Yang-Soo Lee, Yonghan Kwon, Hyung Joon Kim, Hee Chan Kim, Jung Chan Lee, and Eunhee Park. Biomechanical gait analysis using a smartphone-based motion capture system (opencap) in patients with neurological disorders. *Bioengineering*, 11(9):911, 2024.

[27] Yuzhu Guo, Debin Huang, Wei Zhang, Lipeng Wang, Yang Li, Gabriella Olmo, Qiao Wang, Fangang Meng, and Piu Chan. High-accuracy wearable detection of freezing of gait in parkinson's disease based on pseudo-multimodal features. *Computers in Biology and Medicine*, 146:105629, 2022.

[28] Kenneth Koltermann, Woosub Jung, GinaMari Blackwell, Abbott Pinney, Matthew Chen, Leslie Cloud, Ingrid Pretzer-Aboff, and Gang Zhou. 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*, pages 22–33, 2023.

[29] Yi Xia, Hua Sun, Baifu Zhang, Yangyang Xu, and Qiang Ye. Prediction of freezing of gait based on self-supervised pretraining via contrastive learning. *Biomedical Signal Processing and Control*, 89:105765, 2024.

[30] Val Mikos, Chun-Huat Heng, Arthur Tay, Nicole Shuang Yu Chia, Karen Mui Ling Koh, Dawn May Leng Tan, and Wing Lok Au. Real-time patient adaptivity for freezing of gait classification through semi-supervised neural networks. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 871–876. IEEE, 2017.

[31] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[32] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[33] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[34] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[35] Natasa Kleanthous, Abir Jaafar Hussain, Wasiq Khan, and Panos Liatsis. A new machine learning based approach to predict freezing of gait. *Pattern Recognition Letters*, 140:119–126, 2020.

[36] Natasa K Orphanidou, Abir Hussain, Robert Keight, Paulo Lishoa, Jade Hind, and Haya Al-Askar. Predicting freezing of gait in Parkinsons disease patients using machine learning. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.

[37] Luyao Yang, Osama Amin, and Basem Shihada. Intelligent wearable systems: Opportunities and challenges in health and sports. *ACM Computing Surveys*, 56(7):1–42, 2024.

[38] Dimitris Dimoudis, Nikos Tsolakis, Christoniki Magga-Nteve, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. InSEption: a robust mechanism for predicting FoG episodes in PD patients. *Electronics*, 12(9):2088, 2023.