# Multimodal Freezing of Gait Detection: Analyzing the Benefits and Limitations of Physiological Data

Po-Kai Yang, Benjamin Filtjens, Pieter Ginis, Maaike Goris, Alice Nieuwboer, Moran Gilat, Peter Slaets, and Bart Vanrumste, *Senior Member, IEEE*

***Abstract*— Freezing of gait (FOG) is a debilitating symptom of Parkinson's disease (PD), characterized by an absence or reduction in forward movement of the legs despite the intention to walk. Detecting FOG during free-living conditions presents significant challenges, particularly when using only inertial measurement unit (IMU) data, as it must be distinguished from voluntary stopping events that also feature reduced forward movement. Influences from stress and anxiety, measurable through galvanic skin response (GSR) and electrocardiogram (ECG), may assist in distinguishing FOG from normal gait and stopping. However, no study has investigated the fusion of IMU, GSR, and ECG for FOG detection. Therefore, this study introduced two methods: a two-step approach that first identified reduced forward movement segments using a Transformer-based model with IMU data, followed by an XGBoost model classifying these segments as FOG or stopping using IMU, GSR, and ECG features; and an end-to-end approach employing a multi-stage temporal convolutional network to directly classify FOG and stopping segments from IMU, GSR, and ECG data. Results showed that the two-step approach with all data modalities achieved an average F1 score of 0.728 and F1@50 of 0.725, while the end-to-end approach scored 0.771 and 0.759, respectively. However, no significant difference was found compared to using only IMU data in both approaches (p-values: 0.466 to 0.887). In conclusion, adding physiological data did not provide a statistically significant benefit in distinguishing between FOG and stopping. The limitations may be specific to GSR and ECG data, and may not generalize to other physiological modalities.**

***Index Terms*— Freezing of gait assessment, multimodal, inertial measurement unit, galvanic skin response, electrocardiogram, Parkinson's disease.**

Received 11 October 2024; revised 15 January 2025; accepted 20 February 2025. Date of publication 25 February 2025; date of current version 4 March 2025. This work was supported in part by the Development of the Freezing of Gait Interactive Tagging (FOG-IT) Project from KU Leuven under Grant C3/20/109; in part by the Flemish Government (Flanders AI Research Program); in part by the Federal Public Service for Policy and Support (project AidWear); and in part by the Flemish Supercomputer Center (VSC), funded by the Research Foundation-Flanders and the Flemish Government, for resources and services. The work of Po-Kai Yang was supported by the Ministry of Education (KU Leuven-Taiwan) Scholarship. The work of Benjamin Filtjens was supported in part by the KU Leuven Internal Funds Postdoctoral Mandate under Grant PDMT2/22/046, in part by the Data Sciences Institute at the University of Toronto under Grant DSI-PDFY3R1P13, and in part by the Strategic Basic Research Project RevalExo funded by the Research Foundation Flanders under Grant S001024N. The work of Maaike Goris was supported by the Research Foundation Flanders under Grant 1SHEK24N. *(Corresponding author: Po-Kai Yang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee Research 211 UZ/KU Leuven under Application No. S65059.

Po-Kai Yang is with the eMediaResearch Laboratory/STADIUS, Department of Electrical Engineering (ESAT), and the Intelligent Mobile Platforms Research Group, Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium (e-mail: po-kai.yang@kuleuven.be).

Benjamin Filtjens is with the eMediaResearch Laboratory/STADIUS, Department of Electrical Engineering (ESAT), and the Intelligent Mobile Platforms Research Group, Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium, also with the KITE Research Institute, University Health Network, Toronto, ON M5G 2A2, Canada, also with University Health Network, Toronto, ON M5G 2C4, Canada, also with the Data Sciences Institute, University of Toronto, Toronto, ON M7A 2S4, Canada, and also with the Vector Institute, Toronto, ON M5G 0C6, Canada (e-mail: benjamin.filtjens@kuleuven.be).

Pieter Ginis, Maaike Goris, Alice Nieuwboer, and Moran Gilat are with the Neurorehabilitation Research Group (eNRGy), Department of Rehabilitation Sciences, KU Leuven, 3001 Leuven, Belgium (e-mail: pieter.ginis@kuleuven.be; maaike.goris@kuleuven.be; alice.nieuwboer@kuleuven.be; moran.gilat@kuleuven.be).

Peter Slaets is with the Intelligent Mobile Platforms Research Group, Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium (e-mail: peter.slaets@kuleuven.be).

Bart Vanrumste is with the eMediaResearch Laboratory/STADIUS, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium (e-mail: bart.vanrumste@kuleuven.be).

Digital Object Identifier 10.1109/TNSRE.2025.3545110

## I. Introduction

PARKINSON'S disease (PD) affects over six million people globally. Many people with PD experience a symptom called freezing of gait (FOG), which is defined as a sudden inability to move forward despite the intention to walk, posing risks of falls and reduced quality of life [1], [2], [3]. Although commonly referred to as a singular phenomenon, FOG is notably heterogeneous, manifesting in various forms: 1) rapid shuffling characterized by very short steps and insufficient foot clearance; 2) trembling in place, marked by alternating tremulous leg oscillations with minimal or no forward progression; and 3) pure akinesia, where there is little to no visible movement in the lower limbs [4]. Current FOG treatments primarily involve dopaminergic medications and

physical therapy, adjusted based on FOG severity, which is typically assessed using subjective measures such as the New Freezing of Gait Questionnaire (NFOG-Q) [5]. These tools, however, are prone to recall bias and lack sensitivity [6]. Semi-objective evaluation of FOG severity typically involves standardized tasks such as the Timed-Up and Go (TUG) [7] and 360 degree turning tasks [8] in clinical centers, with post-hoc visual analysis of video recordings serving as the gold standard for FOG assessment [9]. Nonetheless, this manual annotation process is labor-intensive and time-consuming, prompting the exploration of automatic approaches leveraging machine learning (ML) and deep learning (DL) models [10], [11], [12], [13], [14], [15], [16].

Monitoring FOG severity in daily life face two main challenges. First, while video recordings are commonly used, they raise privacy concerns and are limited by the camera's field of view [17], [18], [19]. Wearable sensors, such as inertial measurement units (IMU), offer a privacy-respecting alternative but require extensive training on diverse datasets to ensure reliable accuracy [10], [12], [13], [14], [15], [16], [20]. Second, data collected in controlled laboratory settings often lacks the complexity of real-life scenarios, such as the variability in gait and volitional stopping periods [21]. These stopping periods, characterized by the absence of lower limb movements, pose challenges in differentiating from akinetic freezing [22]. Recent research has incorporated these stopping periods into training regimes to enhance detection performance in realistic conditions [12], [21], [23], [24]. Although such adaptations improve detection accuracy [12], relying solely on IMU signals may be insufficient for distinguishing between volitional stopping and FOG [22], [25].

Factors such as stress and anxiety are recognized contributors to the occurrence of FOG [26], [27], influencing physiological signals such as electrocardiogram (ECG) and galvanic skin response (GSR), which are indicators of emotional states [28], [29]. ECG signals have been investigated for their potential to enhance FOG detection, as heart rate (HR) and heart rate variability increase during FOG-episodes [22], [25], [30], [31]. However, these ECG-related metrics also rise during physical activity including gait [22], [31] and are highly influenced by cognitive load [26], [32], limiting their specificity for FOG detection when using traditional methods such as multivariate gaussian distribution [31]. In contrast, GSR features show potential for distinguishing FOG from other gait activities, as demonstrated by Mazilu et al. [31]. Despite this, most studies have focused on statistical testing of GSR and ECG features [22], [30], [31] rather than leveraging them in ML-based clasification models. While multimodal approaches integrating IMU data with physiological signals such as GSR, electroencephalography, and electromyography have been investigated [33], [34], the integration of ECG with IMU and GSR data for FOG detection remains unexplored. Additionally, these approaches may not reflect everyday scenarios where FOG and stopping frequently occur in close succession. Moreover, while HR alone may not reliably differentiate FOG from normal gait [22], [31], its integration with IMU signals could potentially improve accuracy [22]. Thus, Cockx et al. suggested using IMU data to identify movement reduction first and then utilize ECG characteristics to determine whether the reduction is due to FOG or stopping, yet this approach still requires validation [22]. Additionally, when monitoring FOG severity in daily life, patients' movements are often accompanied by other tasks that place a high cognitive load, such as walking through doorways or when performing a concurrent dual-task (DT) when walking [35], [36], [37], [38], [39]. These cognitively demanding tasks could potentially influence the patients' stress and anxiety levels [40], [41], which, in turn, may affect the utility of physiological data such as HR in distinguishing FOG from volitional stopping [22]. However, previous FOG detection studies [25], [31], [33], [34] have not thoroughly evaluated the impact of DT on the effectiveness of FOG detection under such conditions.

To the best of our knowledge, this study is the first to investigate the benefits of integrating motor (IMU) and physiological (GSR, ECG) data for FOG and stopping event detection using ML and DL. Unlike previous studies designed to investigate the mechanisms of GSR and ECG on FOG [22], [31], our study aims to clarify their added value in FOG detection during typical scenarios, such as TUG tests that include stopping periods and cognitive load DTs. Based on previous studies showing significant differences in ECG and GSR data between FOG and other gait activities, this study hypothesized that adding physiological data would improve FOG detection and evaluated this hypothesis with two proposed approaches explained next. First, we developed a two-step approach, building on Cockx et al. [22], which employs a traditional white-box ML classifier to categorize segments of reduced forward progression as FOG or stopping based on multimodal features. We identified these segments either manually, to evaluate the contribution of each modality, or automatically, using a DL-based segmentation model. Second, we introduced an end-to-end approach that directly detects FOG and stopping from raw multimodal signals using the segmentation model.

## II. Methods

This study explored the added value of combining physiological data (GSR, ECG) with motor data (IMU) for FOG detection. We introduced two approaches: a two-step approach, where an action segmentation model forms as a segmentation block detects forward movement reduction segments followed by a differentiation block classifying these as FOG or stopping (Figure 1a); and an end-to-end approach, where the model directly detects FOG and stopping segments (Figure 1b).

We conducted three experiments: (1) model selection for the blocks within each approach, (2) evaluation of the best-performing model for the differentiation block using various modality combinations (IMU, GSR, ECG; IMU+GSR, IMU+ECG, GSR+ECG; IMU+GSR+ECG) and an analysis of the contribution of individual features, and (3) a comparison of the two approaches using IMU data alone versus combined IMU, GSR, and ECG data.

### A. Problem Definition

An IMU trial can be represented as $X_{imu} \in \mathbb{R}^{T \times C_{in}}$, where $T$ is the number of samples, and $C_{in}$ is the input feature

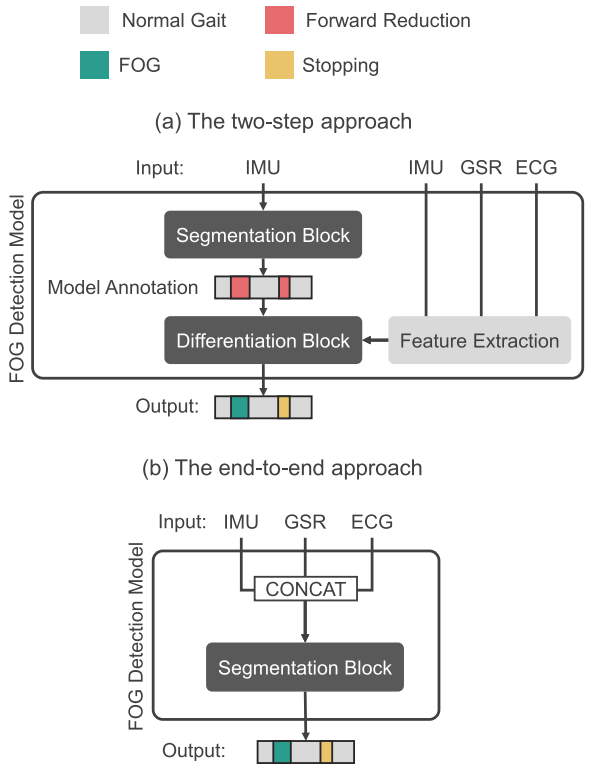


Fig. 1. An overview of our two-step and end-to-end approach: (a) Two-step approach: The segmentation block identifies forward movement reduction segments, which are then classified as FOG or stopping by the differentiation block. (b) End-to-end approach: The segmentation block directly classifies segments as FOG or stopping.

dimension. For IMU signals, $C_{in}$ equals to 12 for acceleration and gyroscope data from two IMUs. Each IMU trial is associated with a GSR signal, which can be represented as $X_{gsr} \in \mathbb{R}^{T \times C_{in}}$, where $C_{in} = 2$ for C0 and C1. Additionally, the IMU and GSR trial is associated with an ECG trial $X_{ecg} \in \mathbb{R}^{T \times C_{in}}$, where $C_{in} = 1$ and $T$. HR extracted from ECG is denoted as $X_{hr} \in \mathbb{R}^{T \times C_{in}}$, where $C_{in} = 1$. Each input trial $X_{imu}$, $X_{gsr}$, $X_{ecg}$, and $X_{hr}$ is associated with a ground truth label vector $Y^{T \times L}$, where the label $L$ represents the manual annotation of FOG by the clinical experts.

*1) The Two-Step Approach:* First, an action segmentation model identifies segments with reduced forward movement. These segments are then classified by a traditional ML model to differentiate between FOG and stopping.

In this approach, the IMU data is processed by the segmentation model to produce prediction of shape $T \times \hat{L}$, where $\hat{L} = 2$ (0 for non-forward reduction and 1 for forward reduction segments, including FOG and stopping). Features from IMU, GSR, and ECG data are extracted for each predicted segment. The differentiation block subsequently classifies these forward reduction segments as either FOG or stopping events. To train the differentiation block, segments manually annotated by experts were used (as shown in Figure5), while during inference, segments identified by the segmentation block were employed (as shown in Figure1a). To determine the best model for the differentiation block, we compared four ML models: XGBoost [42], Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree. For the segmentation block, we evaluated four action segmentation models: Bi-LSTM [43], Dilated TCN [44], MS-TCN [45], and ASFormer [46].

*2) The End-to-End Approach:* The segmentation model directly classifies segments into FOG, stopping, or non-forward reduction movements, bypassing the intermediate segmentation step. This aligns with recent studies that treat FOG detection as an action segmentation task [11], [12].

In the end-to-end approach (as shown in Figure1b), IMU, GSR, and HR data are combined along the feature dimension into a $T \times 15$ input (with HR replacing ECG for aligned sampling rates). This input is processed by an action segmentation model, producing $T \times \hat{L}$ annotations, where $\hat{L} = 3$ (0 for non-forward reduction, 1 for FOG, 2 for stopping). The best-performing model among Bi-LSTM, Dilated TCN, MS-TCN, and ASFormer was used for this approach.

TABLE I
PARTICIPANT AND DATA CHARACTERISTICS

| Characteristics | | Average ± STD | Total |
|---|---|---|---|
| Age | | 67.33 ± 6.71 | - |
| Disease Duration | | 12.39 ± 5.01 | - |
| NFOG-Q (Range 0-28) | | 19.11 ± 3.53 | - |
| MDS-UPDRS | Part-I (Range 0-52) | 15.17 ± 7.41 | - |
| | Part-II (Range 0-52) | 21.61 ± 6.98 | - |
| | Part-III (Range 0-132) | 36.56 ± 11.81 | - |
| | Part-IV (Range 0-24) | 6.06 ± 4.39 | - |
| | Total | 79.22 ± 24.07 | - |
| MoCA | | 26.00 ± 2.52 | - |
| #Trials | | 21.22 ± 4.13 | 382 |
| #FOG-Trials | | 5.33 ± 5.56 | 96 |
| Duration (minutes) | | 7.03 ± 3.70 | 126.63 |
| %TF | | 7.98 ± 13.95 % | 13.00 % |
| #FOG | | 17.44 ± 33.34 | 314 |
| #Stops | | 14.22 ± 4.59 | 256 |

This table shows average and standard deviations for metrics: number of trials, FOG trials, duration of the tasks in total, %TF, #FOG, and #Stops, along with scores from the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS), New Freezing of Gait Questionnaire (NFOG-Q), and Montreal Cognitive Assessment (MoCA). For MDS-UPDRS and NFOG-Q, higher scores indicate worse conditions, while for MoCA, higher scores are better.

## B. Data Collection

This study included 18 participants with PD (4 female, 14 male) with FOG, each with at least one self-reported daily FOG episode lasting a minimum of five seconds in the previous month. All subjects provided informed consent, and the study was approved by the Ethics Committee Research UZ/KU Leuven, with protocol number S65059. The average age was 67.33 ± 6.71 years with an average disease duration of 12.39 ± 5.01 years. Participant characteristics are detailed in Table I. The IMU portion of the dataset was previously utilized in our other studies [12], [47].

Participants underwent TUG tests to provoke FOG. They were instructed to rise from a chair, walk 2.5 meters, turn, return, and sit. Tests were conducted under both OFF-medication (12 hours after the last dose) and ON-medication (one hour post dose) states, both with and without self-generated, where the participant decided when to stop themselves, or researcher-imposed stopping, where verbal instructions to "stop" were given by the researcher. The inclusion of both stopping conditions aimed to capture stopping

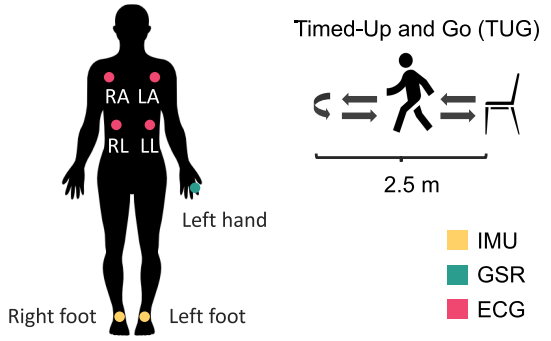


Fig. 2. An illustration of sensor positions and TUG test protocol.

events with prior awareness on the part of the participant (self-generated) and those that occurred unintentionally (researcher-imposed). However, we did not perform a separate analysis to determine if the model could differentiate between these two types of stopping. The primary objective remains to improve FOG detection and ensure the model minimizes false positives (detecting stopping as FOG, or vice versa). Additionally, they performed the normal tasks (i.e., without stopping) with a DT, the auditory Stroop task [48], with the order of testing randomized to control for fatigue and learning. Each subject completed 24 TUG trials: two normal, two with DT, four normal with stopping in the turn, and four normal with stopping in a straight line, totaling 12 trials OFF and 12 trials ON medication.

Data collection utilized Shimmer3 IMU sensors (64 Hz) positioned on top of both feet, alongside a Shimmer3 GSR+ sensor (64 Hz) monitoring GSR. ECG sensors (512 Hz) were placed at torso positions: the left arm (LA), right arm (RA), left leg (LL), and right leg (RL). All sensor placements (Figure 2) and settings followed the manual of Shimmer [49]. RGB videos were recorded at 30 frames-per-second for offline FOG annotation performed iteratively by two expert raters using the Elan software [9]. FOG episodes were defined as periods where no effective forward steps could be made, starting with an ineffective step and ending after at least two effective steps, which were not included in the episode duration [1], [9].

The dataset included 382 trials, totaling 126.63 minutes with 13.00% of the time annotated as FOG. It contained 314 FOG episodes (204 in single task (ST) trials, 110 in DT trials) and 256 stopping episodes (255 in ST trials, 1 in DT trials). Fifteen of the eighteen participants experiencing FOG, averaging 17.44 $\pm$ 33.34 episodes each (median = 4). Details are presented in Table I.

### C. Data Preprocessing

IMU signals were mean-centered to remove bias [12]. GSR signals were filtered with a third-order low-pass filter at 0.9 Hz, producing the filtered signal (C0) and its first derivative (C1) [31], [50], [51]. Following the procedure in [22], ECG signals were processed by selecting the highest-quality lead (i.e., lead II), applying a fifth-order bandpass filter (1-40 Hz), and detecting R-peaks using the Pan Tompkins algorithm [52] implemented via NeuroKit2 [53]. The average ECG quality score was 0.743 [53], with 86% of the signals rated as excellent [54]. Based on these metrics, we considered the ECG quality in our dataset suitable for further analysis [55]. HR was calculated using NeuroKit2 [53] from ECG and resampled to 64 Hz to synchronize with IMU and GSR signals. Figure 3 shows an example of synchronized IMU, C0, C1, ECG, and HR data.

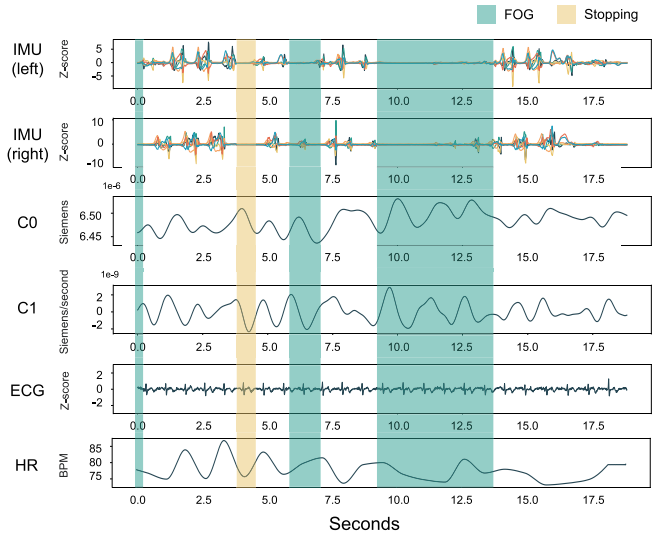


Fig. 3. An illustration of the processed IMU, GSR (C0 and C1), ECG, and HR. IMU signals were z-normalized for better visualization. Experts' annotations of FOG and stopping episodes are highlighted in green and yellow, respectively.

TABLE II
FEATURES EXTRACTED FOR EACH WINDOW

| Signal | Feature | Description |
|---|---|---|
| IMU | Freezing Index | Power ratio between the freezing band (3 - 8 Hz) and the locomotor band (0.5 - 3 Hz) |
| | Total power | Total power in the freezing band and the locomotor band |
| GSR | Mean | The average value over the signal |
| | Median | The median over the signal |
| | STD | The standard deviation value |
| ECG | VLF Power | Power on very-low frequencies (VLF) [0.01,0.04] Hz |
| | LF Power | Power on low frequencies (LF) [0.04, 0.15] Hz |
| | HF Power | Power on high frequencies (HF) [0.15, 0.4] Hz |
| | LF/HF Ratio | The ratio between the power on LF and HF bands |
| HR | Mean | Mean over the HR values in the window |
| | Median | Median over the HR values in the window |
| | STD | Standard deviation over the HR values in the window |
| | HRV | The heart rate variability (HRV = $\frac{\text{STD(HR)}\times 100}{\text{mean(HR)}}$) |

An overview of 18 features from IMU, GSR, and ECG data, with IMU and GSR features duplicated for left/right IMUs and C0/C1 GSR. Features were calculated for pre-event, event, and post-event windows, and differences between these periods, yielding 90 features per FOG or stopping segment.

*1) Feature Generation:* For the differentiation block, each FOG and stopping segment was associated with three time windows: pre-event (3 seconds before), event (the segment itself), and post-event (3 seconds after). The 3-second window size, chosen to capture an average of three heartbeats while avoiding multiple gait events within the window [22], [31], [34]. From these windows, 18 features listed in Table II were extracted per window. Additionally, features were derived from the differences between the pre-event and event windows, and between the event and post-event windows, leading to a total of 90 features per segment, as shown in Figure 4. These features were used to classify each segment in the differentiation block.

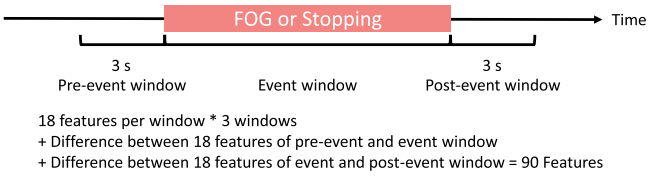


Fig. 4. An illustration of feature generation for FOG and stopping segments: features are calculated within three windows: 3s before the event (pre-event), during the event, and 3s after the event (post-event).

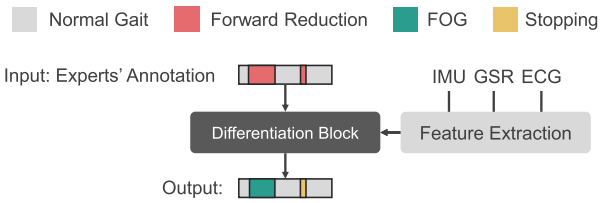


Fig. 5. An illustration of the evaluation process for the differentiation block. To ensure an unbiased comparison of models and assess feature importance, expert-annotated segments are used as input during the evaluation of the differentiation block.

## D. Experimental Setting

The experimental settings and metrics used in the three experiments are detailed in the following subsections.

*1) Experiment 1 (Model Selection):* We used leave-one-subject-out cross-validation (LOSO-CV) to compare four ML models for the differentiation block. Each subject was tested individually, with the remaining subjects used for training and hyperparameter tuning. Validation data consisted of 20% of the training set, randomly selected from 4 of 17 subjects. To ensure that evaluation was not influenced by the segmentation block's performance, the differentiation block was fed with experts' manual annotations (as shown in Figure 5). For the differentation block, the hyperparameters for XGBoost, SVM, KNN, and Decision Tree models were tuned, with Table III listing the specific parameters for each model. Optimal settings were selected based on validation performance. For the segmentation block, Bi-LSTM, Dilated TCN, MS-TCN, ASFormer, were evaluated using LOSO-CV, with hyperparameters based on original studies [43], [44], [45], [46].

To evaluate the differentiation block, AUROC was computed across 570 segments. Sample-wise F1 scores were assessed per trial, classifying each sample as True Positive (TP), False Positive (FP), or False Negative (FN). Segment-wise F1@50 score was used to address over- and under-segmentation, where a segment is TP if its intersection over union with a ground-truth segment exceeds 0.5; otherwise, it is FP. Unmatched ground-truth segments are labeled FN. The F1 score is averaged across trials for each subject:

$$F1 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{J_n}\sum_{j=1}^{J_n}\frac{F1_{n,j}(FOG) + F1_{n,j}(Stopping)}{2} \quad (1)$$

where $N$ is the number of subjects, and $J_n$ is the number of trials for subject $n$. For non-FOG/stopping trials with no detected FOG/stopping episodes, an F1 score of 1 was assigned, indicating correct recognition of no FOG/stopping.

TABLE III
HYPERPAMARETER TUNED FOR THE ML MODELS

| ML Model | Hyperparameter | Value |
|---|---|---|
| XGBoost | max_depth | 1, 3, 5, 7, 9 |
| | learning_rate | 0.1, 0.01, 0.001 |
| | subsample | 0.5, 0.7, 1.0 |
| | min_child_weight | 1, 3, 5, 7, 9 |
| | n_estimators | 1, 3, 5, 7, 9 |
| SVM | C | 0.1, 1, 10, 100, 1000 |
| | gamma | 1, 0.1, 0.01, 0.001, 0.0001 |
| | kernel | 'rbf', 'linear' |
| | class_weight | 'balanced', None |
| KNN | n_neighbors | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 |
| | leaf_size | 20, 25, 30, 35, 40 |
| | p | 1, 2 |
| | weights | 'uniform', 'distance' |
| | metric | 'minkowski', 'chebyshev' |
| Decision Tree | max_features | 'auto', 'sqrt', 'log2' |
| | ccp_alpha | 0.1, 0.01, 0.001 |
| | max_depth | 5, 6, 7, 8, 9 |
| | criterion | 'gini', 'entropy' |

This table presents the hyperparameters tuned for the ML models. The optimal settings were determined based on validation performance using the LOSO-CV approach.

In the two-step approach, the segmentation block combines FOG and stopping as forward reduction. F1 and F1@50 scores were calculated and averaged per trial as follows:

$$F1 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{J_n}\sum_{j=1}^{J_n}F1_{n,j}(ForwardReduction) \quad (2)$$

For non-forward reduction trials with no detected episodes, an F1 score of 1 was set, indicating correct detection of no forward reduction. All DL models were trained for 50 epochs with the Amsgrad optimizer, starting at a learning rate of 0.0005, reduced by 5% per epoch, using cross-entropy loss.

In the end-to-end approach, FOG and stopping were treated as positive classes, and F1 and F1@50 scores were computed similarly to the differentiation block using Equation 1.

*2) Experiment 2 (Feature Comparison for Differentiation Block):* Next, we assessed the best model for the differentiation block from Experiment 1, testing it with various modality combinations (IMU, GSR, ECG; IMU+GSR, IMU+ECG, GSR+ECG; IMU+GSR+ECG) to determine the contribution of each modality. We further analyzed the impact of individual features on the model trained with all multimodal data (IMU+GSR+ECG). To ensure an unbiased evaluation, the differentiation block was provided with expert annotations, as illustrated in Figure 5. All experiments were conducted using LOSO-CV, with hyperparameters tuned according to the settings from Experiment 1.

Model performance was assessed using AUROC and F1 scores, calculated as formula 1. Moreover, for feature importance analysis, we used the SHapley Additive exPlanations (SHAP) method [56] to quantify each feature's impact, which employs Shapley values from cooperative game theory to determine the impact of each feature on the model's predictions. SHAP decomposes a prediction $f(x)$ into contributions from each feature $x_i$:

$$\theta_i = \sum_{S\subseteq M\setminus\{i\}}\frac{|S|!(M-|S|-1)!}{M!}[f(S\cup\{i\}) - f(S)]$$

Here, $S$ is a subset of features excluding $i$, and $f(S)$ is the model prediction for subset $S$. The overall prediction

is represented as:

$$f(x) = \theta_0 + \sum_{i=1}^{M} \theta_i$$

where $\theta_0$ is the average output across the dataset, and $\theta_i$ are the SHAP values for feature $i$. A beeswarm plot visualizes these SHAP values, where a SHAP value of $-1$ indicates a strong impact on predicting FOG, and 1 indicates a strong impact on predicting stopping.

*3) Experiment 3 (FOG Detection Performance Comparison):* Finally, we compared both approaches using IMU data alone and all data. In the two-step approach, IMU data was used for segmentation, and the differentiation block was trained on either all features or just IMU features generated from the model-predicted segments. In the end-to-end approach, models were fed with either IMU, GSR, and HR signals or IMU data alone, with HR replacing raw ECG to align sampling rates with IMU and GSR.

All experiments used LOSO-CV, with hyperparameters tuned and training strategies consistent set as in experiments 1 and 2. Model performance was evaluated using F1 scores, calculated using Equation1. FOG severity was assessed from a clinical perspective using the percentage of time frozen (%TF) and the number of FOG episodes (#FOG). Consistency between model predictions and expert annotations was measured with the intra-class correlation coefficient (ICC(2,1)).

### E. Statistics

We used Repeated Measures ANOVA to test for significant differences in F1 scores between models. Pairwise comparisons were performed with paired Student's t-tests (N = 18, corresponding to the number of subjects), adjusted for multiple comparisons using the Li correction [57]. Levene's and Shapiro-Wilk tests assessed variance homogeneity and normality. All analyses were conducted at a 0.05 significance level, using Python libraries (SciPy 1.7.11, statsmodels 0.13.2, pingouin 0.3.12) and the R package scmamp 0.2.55.

## III. Results

### A. Model Selection

In the first experiment, we aimed to conduct a model selection study for both the two-step approach and the end-to-end approach to ensure that subsequent experiments were evaluated using the best-performing models.

We began by evaluating models for both blocks in the two-step approach, namely the segmentation block and the differentiation block. As shown in Table IV, ASFormer emerged as the top-performing model for detecting forward movement reduction based on IMU signals, making it the preferred choice for the segmentation block. For the differentiation block, XGBoost achieved the highest F1, F1@50, and AUROC scores for classifying forward movement reduction segments, as indicated in Table V, leading to its selection for this block.

Next, we compared models for the end-to-end approach. Table VI indicates that MS-TCN excelled in detecting FOG and stopping using IMU, GSR, and HR signals, and was therefore selected for the segmentation block.

TABLE IV
RESULTS OF DIFFERENT ML MODELS AS DIFFERENTIATION BLOCK

| Model | F1 (↑) | p-value | F1@50 (↑) | p-value | AUROC (↑) |
|---|---|---|---|---|---|
| XGBoost | **0.855** | - | **0.853** | - | **0.900** |
| SVM | 0.832 | 0.552 | 0.838 | 0.686 | 0.884 |
| KNN | 0.810 | 0.354 | 0.813 | 0.475 | 0.813 |
| Decision Tree | 0.779 | 0.002 | 0.779 | 0.001 | 0.746 |

This table compares four ML models for classifying expert-annotated forward movement reduction segments as FOG or stopping, highlighting the statistical significance of XGBoost versus the others. An upward arrow (↑) indicates that higher values represent better performance.

TABLE V
RESULTS OF DIFFERENT DL MODELS AS SEGMENTATION BLOCK

| Model | F1 (↑) | p-value | F1@50 (↑) | p-value |
|---|---|---|---|---|
| ASFormer [46] | **0.772** | - | **0.752** | - |
| MS-TCN [45] | 0.771 | 0.930 | **0.752** | 0.965 |
| Dilated TCN [44] | 0.664 | 0.012 | 0.513 | <0.001 |
| Bi-LSTM [43] | 0.548 | <0.001 | 0.472 | <0.001 |

This table compares four DL models for detecting forward reduction segments, with ASFormer showing significant improvement over all models except MS-TCN. An upward arrow (↑) indicates that higher values represent better performance.

TABLE VI
RESULTS OF DIFFERENT DL MODELS FOR THE END-TO-END APPROACH

| Model | F1 (↑) | p-value | F1@50 (↑) | p-value |
|---|---|---|---|---|
| MS-TCN [45] | **0.771** | - | **0.759** | - |
| ASFormer [46] | 0.719 | 0.028 | 0.705 | 0.151 |
| Dilated TCN [44] | 0.636 | 0.002 | 0.564 | 0.002 |
| Bi-LSTM [43] | 0.729 | 0.247 | 0.707 | 0.627 |

This table compares four DL models for detecting non-FOG and FOG segments, with p-values indicating the statistical significance of MS-TCN compared to the others. An upward arrow (↑) indicates that higher values represent better performance.

### B. Feature Comparison for Differentiation Block

In the second experiment, we aimed to evaluate the best-performing model for the differentiation block of the two-step approach identified in the first experiment (i.e., XGBoost) with different modality combinations (IMU, GSR, ECG; IMU+GSR, IMU+ECG, GSR+ECG; IMU+GSR+ECG). Table VII shows that the model trained with all IMU+GSR+ECG features (all-modality) achieved the highest F1 and F1@50 and the second-highest AUROC score. It classified FOG episodes and stopping (for all trials in both ST and DT) with the fewest errors, showing significantly higher F1 and F1@50 scores (all $p < 0.001$) compared to models without IMU features (GSR, ECG, GSR + ECG). While the all-modality model also achieved higher F1 scores and AUROC than those with IMU features (IMU, IMU + GSR, IMU + ECG), the differences were not statistically significant. Note that models using only GSR or ECG data performed only marginally better than random guessing, as indicated by AUROC scores just above 0.5, suggesting limited effectiveness in differentiating FOG from stopping episodes on these signals alone. When classification results were separated for ST and DT trials, the all-modality model improved correct identification of FOG episodes in ST trials from 75.98% to 79.90%, with fewer FNs and FPs, but decreased performance in DT trials from 85.45% to 84.55%. This indicates that physiological data improves performance slightly for ST trials but not for DT trials.

Feature importance for the all-modality XGBoost model is visualized in Figure 6 through a beeswarm plot, highlighting

TABLE VII
COMPARISON OF THE DIFFERENTIATION BLOCK (XGBOOST) TRAINED USING VARIOUS FEATURE COMBINATIONS

| Input Modality | F1 (↑) (mean ± STD) | p-value | F1@50 (↑) (mean ± STD) | p-value | AUROC (↑) | FOG as ST | FOG (↑) DT | FOG as ST | Stop (↓) DT | Stop as ST | Stop (↑) DT | Stop as ST | FOG (↓) DT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IMU + GSR + ECG | **0.855 ± 0.096** | - | **0.853 ± 0.100** | - | 0.900 | **163** | 93 | **34** | 17 | **217** | **1** | **31** | **0** |
| IMU + GSR | 0.834 ± 0.102 | 0.504 | 0.832 ± 0.105 | 0.375 | **0.906** | 161 | 95 | 37 | 15 | 204 | **1** | 43 | **0** |
| IMU + ECG | 0.840 ± 0.133 | 0.964 | 0.840 ± 0.131 | 0.946 | 0.869 | 155 | **96** | 42 | **14** | 216 | **1** | 32 | **0** |
| GSR + ECG | 0.623 ± 0.142 | <0.001 | 0.625 ± 0.139 | <0.001 | 0.579 | 115 | 63 | 80 | 46 | 134 | 0 | 116 | 1 |
| IMU | 0.829 ± 0.133 | 0.913 | 0.831 ± 0.129 | 0.884 | 0.879 | 155 | 94 | 41 | 16 | 213 | **1** | 36 | **0** |
| GSR | 0.592 ± 0.148 | <0.001 | 0.595 ± 0.145 | <0.001 | 0.511 | 87 | 51 | 107 | 58 | 136 | 0 | 115 | 1 |
| ECG | 0.592 ± 0.115 | <0.001 | 0.593 ± 0.111 | <0.001 | 0.570 | 108 | 66 | 87 | 44 | 126 | **1** | 124 | **0** |

This table compares XGBoost models with different feature combinations, showing significant differences in F1 and F1@50 ($p<0.001$), with p-values for the best model (IMU+GSR+ECG) provided. Results are detailed for correctly and incorrectly classified FOG and stopping episodes for ST and DT trials. The discrepancy in total counts is due to segments potentially containing both FOG and stopping, but each segment is classified into only one category. Additionally, the total number of stopping episodes in DT trials is 1, as participants were not asked to stop during DT, though one did so once. An upward arrow (↑) indicates that higher values represent better performance, while a downward arrow (↓) indicates that lower values represent better performance.

TABLE VIII
RESULTS OF THE TWO-STEP AND END-TO-END APPROACH

| Approach | F1 (↑) | p-value | F1@50 (↑) | p-value | ICC (↑) %TF | ICC (↑) #FOG |
|---|---|---|---|---|---|---|
| Two-step (IMU) | 0.731 | 0.792 | 0.727 | 0.887 | 0.893 | 0.913 |
| Two-step (IMU+GSR+ECG) | 0.728 | | 0.725 | | 0.869 | 0.919 |
| End-to-end (IMU) | 0.780 | 0.607 | 0.770 | 0.466 | 0.938 | 0.893 |
| End-to-end (IMU+GSR+ECG) | 0.771 | | 0.759 | | 0.829 | 0.701 |

This table compares the two-step and end-to-end approaches using IMU data alone versus all three modalities. Pairwise comparisons revealed no significant differences between IMU alone and IMU+GSR+ECG in either approach, nor among the four approaches overall (F1: p=0.561; F1@50: p=0.794).

the top 25 contributing features (calculated based on absolute SHAP values using the SHAP method) in differentiating between FOG and stopping. The plot shows that most influential features are derived from the IMU, with only six from GSR and five from ECG. This indicates that IMU features substantially contribute to the model's predictive capability, while the contribution of the physiological features is relatively limited. Moreover, the beeswarm plot highlights key feature influences on the model's predictions. For example, the model is more likely to predict an episode as FOG when it detects a higher freezing index both before and after the episode, as indicated by negative SHAP values. Furthermore, increases in the standard deviation of GSR (C0) and heart rate (HR) from the period before the episode to during the episode also influence the model towards predicting FOG.

### C. FOG Detection Performance Comparison

Finally, we compared the performance of both the two-step and end-to-end approaches using IMU data alone versus the full multimodal data (IMU+GSR+ECG). In the two-step approach, ASFormer was utilized for the segmentation block, and XGBoost for the differentiation block. For the end-to-end approach, the MS-TCN model was employed to directly detect FOG and stopping events from the multimodal signals.

*1) The Two-Step Approach:* The two-step approach was evaluated (Table VIII), with the all-modality model achieving an F1 score of 0.728 and an F1@50 score of 0.725, showing high agreement with expert annotations (ICC(%TF) = 0.869, ICC(#FOG) = 0.919). The IMU-only model scored similarly (F1 = 0.731, F1@50 = 0.727) with no significant difference between the two models ($p = 0.792$ and 0.887), indicating a limited impact from including GSR and ECG data.

Optimally, the two-step approach would have the segmentation block accurately detect the exact onset and offset of all forward movement reduction episodes, aligning with expert annotations. This would potentially achieving F1 and F1@50 scores of 0.855 and 0.853 (as shown in Table VII).

*2) The End-to-End Approach:* We also evaluated the end-to-end approach with MS-TCN for segmentation. Table VIII shows that MS-TCN trained with all three modalities achieved an F1 score of 0.771 and an F1@50 of 0.759, slightly lower than the IMU-only model (F1 = 0.780; F1@50 = 0.770). These differences were not statistically significant ($p = 0.607$ and 0.466), indicating a marginal decrease in detection performance with the inclusion of physiological signals. Moreover, no significant differences were observed between the two-step and the end-to-end approaches in terms of F1 ($p = 0.561$) and F1@50 ($p = 0.794$). While the end-to-end model shows higher F1 scores than the two-step approach, it is less interpretable, as the two-step approach allows for greater exploration of the features used by the differentiation block.

Additionally, we conducted a subject-based analysis of the end-to-end approach to examine whether the physiological data was more effective for specific participants. Figure 7 visualizes the differences in F1 scores between the model trained with all modalities and the model trained solely on IMU data. The plot indicates that, for most subjects, the performance differences between the two models varied around zero. However, the model trained with all modalities performed worse for S13 and S15, who rank among the top three subjects with the highest frequency of FOG episodes (S13: %TF = 49.7, #FOG = 128; S15: %TF = 24.58, #FOG = 36). Conversely, the model trained with all modalities performed better for S07, though no clear pattern or explanation could be identified for this improvement.

## IV. DISCUSSION

This study investigated FOG detection using multimodal data (IMU, GSR, ECG) to assess the possible benefits of integrating physiological signals with IMU data. Unlike previous research that focused on statistical differences between FOG and stopping events [22], [25], we evaluated the performance of an XGBoost model trained to classify FOG and stopping segments annotated by experts. The model performed best with features from all three data modalities, correctly

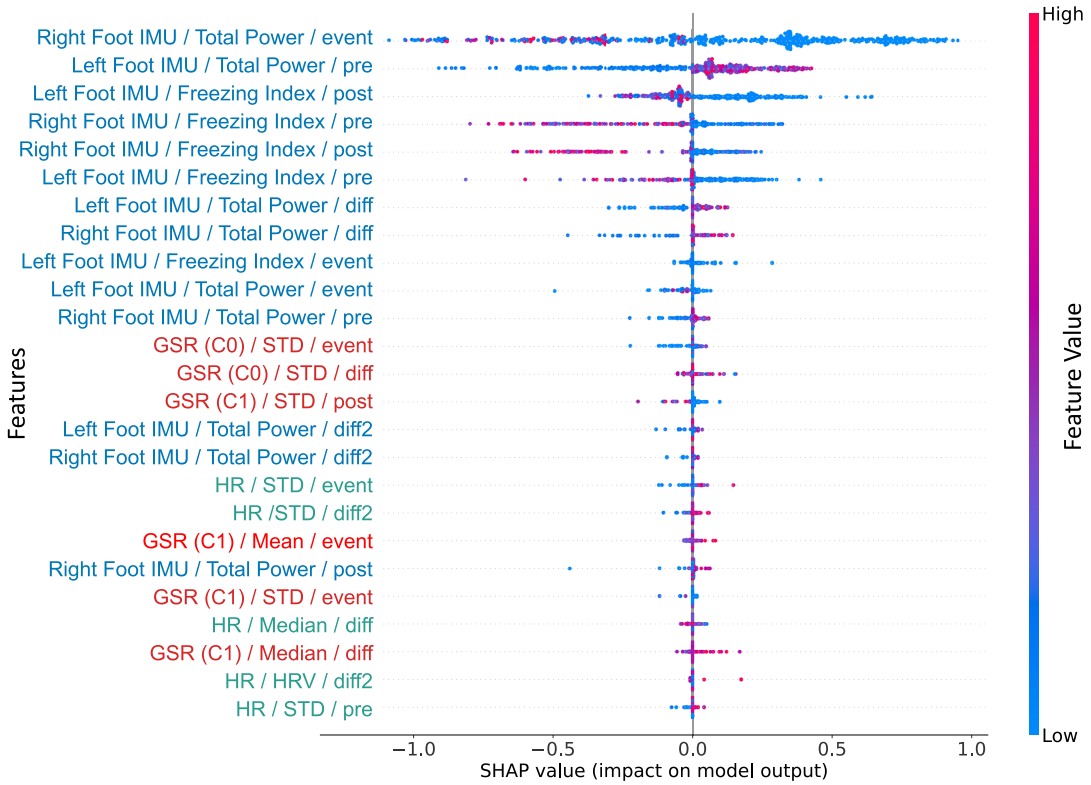


Fig. 6. Beeswarm plot of the top 25 contributing features in the XGBoost model. This plot displays SHAP values for each feature, listed vertically, across all input episodes. Points intensify in red with higher feature values and in blue with lower values, highlighting how feature values contribute to the prediction. The naming of the features is organized as follows: Modality / Feature Name / Extracted Window. For the extracted window, "diff" represents the difference between the event and pre-event window, while "diff2" represents the difference between the post-event and event window.

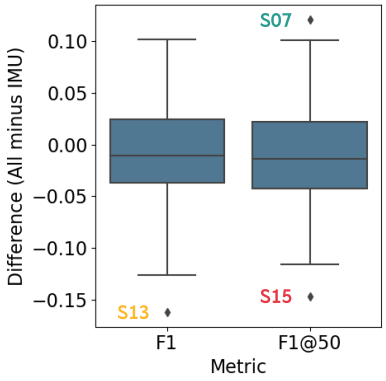


Fig. 7. Boxplot illustrating the differences in F1 scores between the end-to-end model trained with all data and the model trained only with IMU data. Positive values indicate better performance for the model trained with all data, while negative values indicate better performance for the model trained solely on IMU data. Outlier subjects, i.e., S07, S13, and S15, are marked on the plot.

identifying 81.53% of FOG episodes and 85.16% of stopping episodes, with misclassification rates of 16.24% and 11.72%, respectively. Physiological features improved detection more in ST than DT trials, and SHAP analysis revealed that IMU features were the most influential (14 of the top 25 features). This study also introduced two approaches for FOG detection: a two-step approach using ASFormer for segmentation followed by XGBoost for classification, which offers higher interpretability, and an end-to-end approach using MS-TCN for direct classification, functioning as a uninterpretable model. However, adding physiological features did not significantly improve performance and even led to slightly lower results in terms of F1 and F1@50 in either approach. This suggests the limited utility of physiological data for improving FOG detection. However, the improvement observed in the differentiation block when using expert-annotated episodes as a perfect segmentation model implies that the benefit of physiological data arises primarily when the onset and offset of reduced forward movement are clearly identified. This indicates that the time windows selected for analyzing physiological features are critical for maximizing their effectiveness in FOG detection. Additionally, although the inclusion of physiological data did not significantly improve performance at the population level, we observed performance improvements on a subject-specific basis.

Previous studies have identified significant differences in physiological features among normal gait, FOG, and stopping episodes using specialized criteria for statistical testing [22], [26], [30], [31]. Our findings suggested limited added value of these signals for FOG detection. We interpret the limited discriminatory effectiveness of physiological data with regard to the following four factors. Firstly, physiological signals change slowly, requiring longer window periods (typically 3 seconds) that can introduce noise from multiple short FOG or stopping episodes. Signals with greater temporal resolution, such as IMU [12] and EEG [33], [34], may be better suited for

future studies. Secondly, rapid sequences of movements, such as turning and sitting to standing, introduce motion artifact that complicates distinguishing physiological changes between consecutive FOG episodes [22], [31]. Thirdly, variations in physiological responses to environmental stimuli make uniform responses difficult [31]. Fourthly, cognitive load from DT complicates the differentiation between FOG and stopping using physiological signals, as DT may also influence patients' stress levels [22], [32], [58]. Finally, the study was conducted in a controlled laboratory setting, ensuring consistent factors such as room temperature and lighting, which is impractical to replicate in real-life scenarios. These limitations suggest that while physiological data provides valuable insights into the mechanisms of FOG, its utility for FOG detection is constrained by inherent variability and signal dynamics.

This study faces several limitations. First, this study used only early fusion for the end-to-end approach. Exploring middle fusion [59] or models that handle different temporal resolutions could better integrate IMU and physiological data, as physiological signals vary more slowly and do not need the 64 Hz sampling rate as for the IMU data. Secondly, the variability in FOG frequency among participants mirrors the range of FOG severity typically seen in clinical settings but could impact the evaluation of physiological data's usefulness. For participants with fewer FOG episodes, the limited number of data segments might make it harder to assess the contribution of physiological signals. On the other hand, participants with more frequent episodes may generate noisier data due to the close succession of episodes, which could reduce the effectiveness of physiological data in distinguishing FOG from non-FOG events. Thirdly, our study aimed to identify the benefits of physiological data for FOG detection as an initial step toward real-world applications. Although the TUG task with volitional stopping allows for investigating the distinction between stopping and FOG during straight-line walking and turning, it does not consider specific activities related to gait initiation, such as standing up from a chair or turning in place before starting to walk. Our dataset was not designed to capture these transitional movements, which could present valuable opportunities for future multimodal data collection. However, incorporating such scenarios introduces challenges in identifying robust features that are both specific and sensitive enough to function reliably at the individual patient level across diverse conditions. Finally, our dataset (N = 18, #FOG = 314) is comparable to studies such as Daphnet (N = 10, #FOG = 237) [60], O'Day et al. (N = 16, #FOG = 211) [10], CuPid (N = 18, #FOG = 237) [14], Rempark (21 subjects, 1321 episodes) [61], and Zhang et al. (N = 12, #FOG = 334) [34]. However, a larger and more diverse cohort is needed to fully evaluate the impact of physiological data and enhance generalizability. In addition, future studies should include measures of state and trait anxiety of the cohort recruited.

## V. Conclusion

Advancing free-living FOG detection requires accurately distinguishing FOG from stopping episodes. This study proposed two approaches for analyzing the possible benefits of physiological data over IMU data for FOG detection. The two-step approach used ASFormer to identify reduced forward movement segments, which were then classified as FOG or stopping by an XGBoost model. The end-to-end approach utilized MS-TCN to directly detect FOG and stopping. Evaluations showed that adding physiological signals had limited benefit on performance in both approaches.

## References


[1] J. G. Nutt, B. R. Bloem, N. Giladi, M. Hallett, F. B. Horak, and A. Nieuwboer, "Freezing of gait: Moving forward on a mysterious clinical phenomenon," *Lancet Neurol.*, vol. 10, no. 8, pp. 734–744, Aug. 2011.

[2] B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi, "Falls and freezing of gait in Parkinson's disease: A review of two interconnected, episodic phenomena," *Movement Disorders*, vol. 19, no. 8, pp. 871–884, 2004.

[3] N. Giladi and A. Nieuwboer, "Understanding and treating freezing of gait in parkinsonism, proposed working definition, and setting the stage," *Movement Disorders*, vol. 23, no. S2, pp. S423–S425, Jul. 2008.

[4] J. D. Schaafsma, Y. Balash, T. Gurevich, A. L. Bartels, J. M. Hausdorff, and N. Giladi, "Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease," *Eur. J. Neurol.*, vol. 10, no. 4, pp. 391–398, Jul. 2003.

[5] A. Nieuwboer et al., "Reliability of the new freezing of gait questionnaire: Agreement between patients with Parkinson's disease and their carers," *Gait Posture*, vol. 30, no. 4, pp. 459–463, Nov. 2009.

[6] F. Hulzinga et al., "The new freezing of gait questionnaire: Unsuitable as an outcome in clinical trials?" *Movement Disorders Clin. Pract.*, vol. 7, no. 2, pp. 199–205, Feb. 2020.

[7] M. Mancini, K. C. Priest, J. G. Nutt, and F. B. Horak, "Quantifying freezing of gait in Parkinson's disease during the instrumented timed up and go test," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 1198–1201.

[8] M. Mancini, K. Smulders, R. G. Cohen, F. B. Horak, N. Giladi, and J. G. Nutt, "The clinical significance of freezing while turning in Parkinson's disease," *Neuroscience*, vol. 343, pp. 222–228, Feb. 2017.

[9] M. Gilat, "How to annotate freezing of gait from video: A standardized method using open-source software," *J. Parkinson's Disease*, vol. 9, no. 4, pp. 821–824, Oct. 2019.

[10] J. O'Day et al., "Assessing inertial measurement unit locations for freezing of gait detection and patient preference," *J. NeuroEng. Rehabil.*, vol. 19, no. 1, p. 20, Dec. 2022.

[11] B. Filtjens, P. Ginis, A. Nieuwboer, P. Slaets, and B. Vanrumste, "Automated freezing of gait assessment with marker-based motion capture and multi-stage spatial–temporal graph convolutional neural networks," *J. NeuroEng. Rehabil.*, vol. 19, no. 1, p. 48, Dec. 2022.

[12] P.-K. Yang et al., "Freezing of gait assessment with inertial measurement units and deep learning: Effect of tasks, medication states, and stops," *J. NeuroEng. Rehabil.*, vol. 21, no. 1, p. 24, Feb. 2024.

[13] B. Shi, A. Tay, W. L. Au, D. M. L. Tan, N. S. Y. Chia, and S.-C. Yen, "Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 7, pp. 2256–2267, Jul. 2022.

[14] S. Mazilu et al., "Online detection of freezing of gait with smartphones and machine learning techniques," in *Proc. 6th Int. Conf. Pervasive Comput. Technol. Healthcare (PervasiveHealth) Workshops*, Jul. 2012, pp. 123–130.

[15] Y. Zhang and D. Gu, "A deep convolutional-recurrent neural network for freezing of gait detection in patients with Parkinson's disease," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–6.

[16] B. Li, Z. Yao, J. Wang, S. Wang, X. Yang, and Y. Sun, "Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors," *Electronics*, vol. 9, no. 11, p. 1919, Nov. 2020.

[17] D. Sweeney, L. Quinlan, P. Browne, M. Richardson, P. Meskell, and G. ÓLaighin, "A technological review of wearable cueing devices addressing freezing of gait in Parkinson's disease," *Sensors*, vol. 19, no. 6, p. 1277, Mar. 2019.

[18] N. M. de Vries et al., "Exploring the Parkinson patients' perspective on home-based video recording for movement analysis: A qualitative study," *BMC Neurol.*, vol. 19, no. 1, pp. 1–6, Dec. 2019.

[19] K. W. Park, M. S. Mirian, and M. J. McKeown, "Artificial intelligence-based video monitoring of movement disorders in the elderly: A review on current and future landscapes," *Singap. Med. J.*, vol. 65, no. 3, pp. 141–149, 2024.

[20] M. Mancini, B. R. Bloem, F. B. Horak, S. J. G. Lewis, A. Nieuwboer, and J. Nonnekes, "Clinical and methodological challenges for assessing freezing of gait: Future perspectives," *Movement Disorders*, vol. 34, no. 6, pp. 783–790, Jun. 2019.

[21] A. R. John et al., "Predicting the onset of freezing of gait using EEG dynamics," *Appl. Sci.*, vol. 13, no. 1, p. 302, Dec. 2022.

[22] H. Cockx, J. Nonnekes, B. R. Bloem, R. van Wezel, I. Cameron, and Y. Wang, "Dealing with the heterogeneous presentations of freezing of gait: How reliable are the freezing index and heart rate for freezing detection?" *J. NeuroEng. Rehabil.*, vol. 20, no. 1, p. 53, Apr. 2023.

[23] T. Bikias, D. Iakovakis, S. Hadjidimitriou, V. Charisis, and L. J. Hadjileontiadis, "DeepFoG: An IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning," *Frontiers Robot. AI*, vol. 8, p. 117, May 2021.

[24] T. Krasovsky et al., "Bilateral leg stepping coherence as a predictor of freezing of gait in patients with Parkinson's disease walking with wearable sensors," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 798–805, 2023.

[25] B. Heimler et al., "Heart-rate variability as a new marker for freezing predisposition in Parkinson's disease," *Parkinsonism Rel. Disorders*, vol. 113, Aug. 2023, Art. no. 105476.

[26] K. Economou, D. Quek, H. MacDougall, S. J. G. Lewis, and K. A. E. Martens, "Heart rate changes prior to freezing of gait episodes are related to anxiety," *J. Parkinson's Disease*, vol. 11, no. 1, pp. 271–282, Feb. 2021.

[27] S. Rahman, H. Griffin, N. Quinn, and M. Jahanshahi, "The factors that induce or overcome freezing of gait in Parkinson's disease," *Behavioural Neurol.*, vol. 19, no. 3, pp. 127–136, 2008.

[28] M. Kusserow, O. Amft, and G. Tröster, "Monitoring stress arousal in the wild," *IEEE Pervasive Comput.*, vol. 12, no. 2, pp. 28–37, Apr. 2013.

[29] C. Kappeler-Setz, F. Gravenhorst, J. Schumm, B. Arnrich, and G. Tröster, "Towards long term monitoring of electrodermal activity in daily life," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 261–271, Feb. 2013.

[30] I. Maidan, M. Plotnik, A. Mirelman, A. Weiss, N. Giladi, and J. M. Hausdorff, "Heart rate changes during freezing of gait in patients with Parkinson's disease," *Movement Disorders*, vol. 25, no. 14, pp. 2346–2354, Oct. 2010.

[31] S. Mazilu, A. Calatroni, E. Gazit, A. Mirelman, J. M. Hausdorff, and G. Tröster, "Prediction of freezing of gait in Parkinson's from physiological wearables: An exploratory study," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1843–1854, Nov. 2015.

[32] H. Johansson, U. Ekman, L. Rennie, D. S. Peterson, B. Leavy, and E. Franzén, "Dual-task effects during a motor-cognitive task in Parkinson's disease: Patterns of prioritization and the influence of cognitive status," *Neurorehabilitation Neural Repair*, vol. 35, no. 4, pp. 356–366, Apr. 2021.

[33] L. Mesin et al., "A multi-modal analysis of the freezing of gait phenomenon in Parkinson's disease," *Sensors*, vol. 22, no. 7, p. 2613, Mar. 2022.

[34] W. Zhang et al., "Multimodal data for the detection of freezing of gait in Parkinson's disease," *Sci. Data*, vol. 9, no. 1, p. 606, 2022.

[35] J. Spildooren, S. Vercruysse, K. Desloovere, W. Vandenberghe, E. Kerckhofs, and A. Nieuwboer, "Freezing of gait in Parkinson's disease: The impact of dual-tasking and turning," *Movement Disorders*, vol. 25, no. 15, pp. 2563–2570, Nov. 2010.

[36] F. Pieruccini-Faria, J. A. Jones, and Q. J. Almeida, "Motor planning in Parkinson's disease patients experiencing freezing of gait: The influence of cognitive load when approaching obstacles," *Brain Cognition*, vol. 87, pp. 76–85, Jun. 2014.

[37] J. Nonnekes, A. H. Snijders, J. G. Nutt, G. Deuschl, N. Giladi, and B. R. Bloem, "Freezing of gait: A practical approach to management," *Lancet Neurol.*, vol. 14, no. 7, pp. 768–778, Jul. 2015.

[38] M. Mancini et al., "Measuring freezing of gait during daily-life: An open-source, wearable sensors approach," *J. NeuroEng. Rehabil.*, vol. 18, no. 1, pp. 1–13, Jan. 2021.

[39] K. Mejia, "Dual task in the daily life of patients with Parkinson's disease," in *Movement Disorders*, vol. 33. Hoboken, NJ, USA: Wiley, 2018, pp. S167–S168.

[40] Z. Nodehi et al., "Anxiety and cognitive load affect upper limb motor control in Parkinson's disease during medication phases," *Ann. New York Acad. Sci.*, vol. 1494, no. 1, pp. 44–58, Jun. 2021.

[41] K. A. E. Martens, C. R. Silveira, B. N. Intzandt, and Q. J. Almeida, "Overload from anxiety: A non-motor cause for gait impairments in Parkinson's disease," *J. Neuropsychiatry Clin. Neurosci.*, vol. 30, no. 1, pp. 77–80, 2018.

[42] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[43] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1961–1970.

[44] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.

[45] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3575–3584.

[46] F. Yi, W. Hong-Yu, and T. Jiang, "ASFormer: Transformer for action segmentation," in *Proc. Brit. Mach. Vis. Conf.*, Jan. 2021, pp. 1–19.

[47] P.-K. Yang et al., "Automatic detection and assessment of freezing of gait manifestations," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 2699–2708, 2024.

[48] K. Kestens, S. Degeest, M. Miatton, and H. Keppler, "An auditory stroop test to implement in cognitive hearing sciences: Development and normative data," *Int. J. Psychol. Res.*, vol. 14, no. 2, pp. 37–51, Oct. 2021.

[49] A. Burns et al., "SHIMMER—A wireless sensor platform for non-invasive biomedical research," *IEEE Sensors J.*, vol. 10, no. 9, pp. 1527–1534, Sep. 2010.

[50] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, pp. 940–943.

[51] F. Gravenhorst, A. Muaremi, A. Gruenerbl, B. Arnrich, and G. Troester, "Towards a mobile galvanic skin response measurement system for mentally disordered patients," in *Proc. 8th Int. Conf. Body Area Netw.*, 2013, pp. 432–435.

[52] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[53] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021.

[54] Z. Zhao and Y. Zhang, "SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation," *Frontiers Physiol.*, vol. 9, p. 727, Jun. 2018.

[55] C. Gao, N. Gisolfi, and A. Dubrawski, "Signal quality auditing for time-series data," 2024, *arXiv:2402.00803*.

[56] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp 4768–4777.

[57] J. (David) Li, "A two-step rejection procedure for testing multiple hypotheses," *J. Stat. Planning Inference*, vol. 138, no. 6, pp. 1521–1527, Jul. 2008.

[58] K. A. E. Martens, D. S. Peterson, Q. J. Almeida, S. J. G. Lewis, J. M. Hausdorff, and A. Nieuwboer, "Behavioural manifestations and associated non-motor features of freezing of gait: A narrative review and theoretical framework," *Neurosci. Biobehavioral Rev.*, vol. 116, pp. 350–364, Sep. 2020.

[59] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.

[60] M. Bachlin et al., "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, Mar. 2010.

[61] D. Rodríguez-Martín et al., "Home detection of freezing of gait using support vector machines through a single Waist–Worn triaxial accelerometer," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171764.