Automatic Detection and Assessment of Freezing of Gait Manifestations

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 an episodic and highly disabling symptom of Parkinson's disease (PD). Although described as a single phenomenon, FOG is heterogeneous and can express as different manifestations, such as trembling in place or complete akinesia. We aimed to analyze the efficacy of deep learning (DL) trained on inertial measurement unit data to classify FOG into both manifestations. We adapted and compared four state-of-the-art FOG detection algorithms for this task and investigated the advantages of incorporating a refinement model to address oversegmentation errors. We evaluated the model's performance in distinguishing between trembling and akinesia, as well as other forms of movement cessation (e.g., stopping and sitting), against gold-standard video annotations. Experiments were conducted on a dataset of eighteen PD patients completing a FOG-provoking protocol in a gait laboratory. Results showed our model achieved an F1 score of 0.78 and segment F1@50 of 0.75 in detecting FOG manifestations. Assessment of FOG severity was strong for trembling (ICC=0.86, [0.66,0.95]) and moderately strong for akinesia (ICC=0.78, [0.51,0.91]). Importantly, our model successfully differentiated FOG from other forms of movement cessation during 360-degree turning-in-place tasks. In conclusion, our study demonstrates that DL can accurately assess different types of FOG manifestations, warranting further investigation in larger and more diverse verification cohorts.**

***Index Terms*— Freezing of gait assessment, detection, manifestations, phenotypes, Parkinson's disease, deep learning.**

Manuscript received 31 January 2024; revised 27 May 2024 and 2 July 2024; accepted 17 July 2024. Date of publication 19 July 2024; date of current version 31 July 2024. This work was supported by the Development of the Freezing of Gait Interactive Tagging (FOG-IT) Project from KU Leuven under Grant C3/20/109. The work of Po-Kai Yang was supported by the Ministry of Education (KU Leuven-Taiwan) Scholarship. The work of Benjamin Filtjens was supported by KU Leuven Internal Funds Post-Doctoral Mandate under Grant PDMT2/22/046. The work of Maaike Goris was supported by Fonds Wetenschappelijk Onderzoek (FWO) 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 Medical Ethical Committee Research of UZ/KU Leuven under Application No. S65059.

Po-Kai Yang and Benjamin Filtjens are with the KU Leuven, Department of Electrical Engineering (ESAT), eMedia Research Laboratory/STADIUS, 3001 Leuven, Belgium, and also with the KU Leuven, Department of Mechanical Engineering, Intelligent Mobile Platforms Research Group, 3001 Leuven, Belgium (e-mail: po-kai.yang@kuleuven.be; benjamin.filtjens@kuleuven.be).

Pieter Ginis, Maaike Goris, Alice Nieuwboer, and Moran Gilat are with the KU Leuven, Department of Rehabilitation Sciences, Neurorehabilitation Research Group (eNRGy), 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 KU Leuven, Department of Mechanical Engineering, Intelligent Mobile Platforms Research Group, 3001 Leuven, Belgium (e-mail: peter.slaets@kuleuven.be).

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

Digital Object Identifier 10.1109/TNSRE.2024.3431208

I. Introduction

PARKINSON'S disease (PD) is a neurodegenerative disorder that already affects over six million people worldwide with a prevalence that is rising [1]. One of the most debilitating symptoms associated with PD is freezing of gait (FOG), which has been defined as a "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk" [1], [2], [3]. The unpredictable nature and the inability of patients to take corrective steps after losing their balance during FOG poses a significant risk of falls and related injuries for PD patients [4], [5], [6], and a lower quality of life [7]. Although described as a single phenomenon, FOG is very heterogeneous and can be expressed as different manifestations, namely: 1) episodic rapid shuffling with very short steps and poor clearance of the feet, 2) trembling in place visible as alternating tremulous oscillations in the legs with minimal or no forward progression, and 3) complete akinesia with minimal or no visible movement in the lower limbs [8]. However, whether or not shuffling should be included in the definition of FOG is being debated given that there is still forward progression of the feet [9]. FOG episodes could exhibit various manifestations, with some episodes showing only one type while others may encompass multiple types. Akinetic FOG is more likely to occur during tasks with high cognitive load [10], despite being less common than other manifestations [8], [11]. Given that the etiology of the different manifestations likely differs and that akinetic and trembling features may respond differently to therapy [11], it is of interest to develop an objective assessment of FOG manifestations. Consequently, a better understanding of these complex phenomena will help to guide appropriate treatment [8].

The standard method for assessing FOG severity during standardized tasks involves labor-intensive visual analysis of post-task video footage by clinical experts [12]. This approach necessitates frame-by-frame labeling of FOG episodes to calculate semi-objective measures, in particular the percentage of time spent frozen (%TF) [13]. To mitigate this challenge,

researchers have proposed automatic annotation methods of video data [14], [15] or of data from wearable sensors, such as inertial measurement units (IMUs) [9], [16], [17], [18], [19], [20], [21], or from 3D motion capture [22]. However, fixed-camera video data collection poses challenges, particularly for at-home monitoring, and 3D motion capture is constrained to an in-lab setting. Therefore, IMU-based methods are preferred in this context. Despite the popularity of IMU-based approaches, there are currently no studies proposing automatic detection of FOG manifestations using IMU data.

The current study is the first attempt to automatically quantify different FOG manifestations using deep learning (DL) and lower limb movement characteristics measured by IMUs. We proposed a FOG manifestation detection model that consists of two components: an initial detection block and a subsequent annotation refinement block. The former aims to assign initial probabilities to distinct FOG manifestations for each temporal sample, while the latter seeks to mitigate the issue of oversegmentation inherent in predictions [23]. We adapted and assessed four state-of-the-art FOG detection algorithms [9], [18], [19], which will be further discussed in section II-A, for the initial detection block, aiming to select the most effective model for detecting FOG manifestations. Next, the multi-stage temporal convolutional neural network (MS-TCN) [23] was utilized for refinement [9]. To quantify FOG manifestation severity, we calculated the %TF as per previous work [12], [13] and the percentage time frozen of each manifestation. Given the lack of overt movement in the legs during particularly akinetic FOG episodes, it is important to verify that the model does not simply detect FOG in the absence of motion and is indeed able to distinguish such FOG events from other forms of volitional movement cessation. As such, to determine the robustness of our approach, we further investigated whether our DL algorithm could distinguish between FOG manifestations and other forms of volitional cessation (e.g., stopping and sitting) [24]. This involved explicit model training to detect five classes: normal gait, trembling FOG, akinetic FOG, stopping, and sitting.

II. Related Work

Various methods have been proposed to automatically detect FOG using wearable sensor data obtained through IMUs [9], [16], [17], [18], [19], [20], [21]. IMUs record the movement of the associated body segment as a time series of 3-axis acceleration and angular velocity. The raw signals themselves or features extracted from them have been employed to train various FOG detection models. Typically, these models segment sensor data into multiple windows of a predefined size (e.g., 1 second). To determine the granularity of prediction, a stride size is specified for generating FOG annotations. Conventional models often aim to produce FOG annotations with a stride size equal to half of the window size [19], resulting in downsampled FOG annotation. However, as illustrated in Figure 1, generating predictions by sliding windows with a stride size equal to half of the window size may not be the most optimal for defining the exact onsets and offsets of FOG episodes. Therefore, to annotate FOG episodes frame by frame, matching how clinical experts annotate videos, a stride of one sample is more appropriate [9].

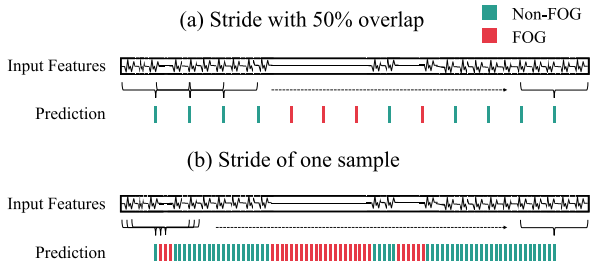

Fig. 1. This example illustrates the impact of different stride sizes for a sliding window-based model. For each window, the model predicts a single label (e.g., non-FOG or FOG). The x-axis represents the timeline for the annotations. This example shows that generating predictions with a 50% overlap between consecutive windows results in a downsampled prediction. Additionally, the first FOG episode which was shorter than the stride size was ignored.

However, generating frame by frame FOG annotations with DL often leads to oversegmentation errors [23], wherein long FOG episodes are annotated as multiple short FOG episodes, thereby impacting the FOG detection performance of the models [22]. The MS-TCN stands as one of the current state-of-the-art DL models, and was initially designed for frame-by-frame sequence mapping in computer vision tasks [23]. The MS-TCN first generates an initial prediction using multiple temporal convolution layers and subsequently refines this prediction over multiple TCN stages. To address oversegmentation errors in FOG detection, DL models with a refinement stage were used for FOG detection with 3D Motion Capture [22] and IMU data [9].

In this section, we delve into the IMU-based FOG detection models previously advocated in the literature, which have potential to be extended for FOG manifestation detection.

A. FOG Detection Models for Initial Detection

Automated FOG detection models segment an IMU sequence into fixed-length windows using a sliding-window scheme [16], [17], [18], [19], [20], [21]. Within each window, a single label is predicted for all the samples as either FOG or non-FOG. Since each window can contain multiple labels at FOG and non-FOG transitions, the ground-truth label is typically established through majority voting [17], [18], [19] or the label at the center of the window [9].

Such earlier approaches also relied on manual feature engineering to distinguish between FOG and non-FOG. For instance, Moore et al. developed a thresholding algorithm based on the Freeze Index to distinguish between FOG and non-FOG [25]. They defined the Freeze Index as the power in the freezing band (0.5-3 Hz) divided by the power in the locomotor band (3-8 Hz), which others have subsequently applied as well [16]. Moreover, other studies introduced modifications like an energy threshold [26], stride features [27] and number of turns [28], which were combined with the Freeze Index to identify FOG episodes.

Going beyond the aforementioned threshold-based methods, previous studies also employed traditional machine learning (ML) models on hand-engineered features to detect FOG. For example, Tsipouras et al. employed decision trees and random

forests on the mean entropy calculated from the acceleration of six IMUs (i.e., right/left wrist, right/left leg, chest, and waist) and the angular velocity from two IMUs (chest and waist) [29]. Moreover, Mazilu et al. tested eleven ML models (e.g., random forests, k-nearest neighbor, and AdaBoost) on seven hand-engineered acceleration features (i.e., mean, standard deviation, variance, entropy, energy, Freeze Index, and power) [30]. Additionally, Shi et al. combined all the aforementioned features with wavelet features to form a set of 67 expressive features to characterize FOG [19]. They compared seven popular ML algorithms (e.g., k-nearest neighbors, support vector machines, and extreme gradient boosting (XGBoost)) and concluded that XGBoost enabled the best FOG detection performance [19].

However, manually engineered features run the risk of not being fully generalizable to all patients, given that PD and FOG are highly heterogeneous. Recent studies have thus shifted towards end-to-end DL models [17], [18], [19]. Due to their large parametric space, DL techniques can directly infer relevant features from raw input data. For example, Zhang et al. used raw acceleration and spectrograms of one waist IMU as input for a DeepCNN-LSTM model trained to detect FOG [31]. Li et al. proposed a DL model using a TCN and long-short-term-memory network for FOG detection using acceleration signals from three IMU sensors [32]. O'Day et al. fed raw acceleration and angular velocity data from one to eleven IMUs into a convolutional neural network (CNN) to detect FOG [18]. Lastly, Shi et al., besides proposing the feature-based model, also introduced an improved method that used the continuous wavelet transform (CWT) as a pre-processing step on each acceleration and angular velocity signal to generate scalograms which were used as input for a CNN [19]. Their results showed that CWT, in combination with a CNN, is state-of-the-art in FOG detection.

In our recent study [9], we compared three traditional ML models with two DL models for FOG detection using raw IMU data. Instead of generating downsampled FOG annotations, we generate FOG annotations on the sample level by sliding windows with a one-sample stride size. Our results indicated that a TCN from [33] gave the best performance in frame by frame FOG detection [9].

III. Methods

In this study, we proposed a FOG manifestation detection model that contains an initial annotation detection block and an annotation refinement block. We adapted four previously proposed FOG detection models for the task of FOG manifestation detection. Our aim was to compare these models and identify the most effective one for initial FOG manifestation detection. Subsequently, we assessed the performance of the best-performing model when integrated with an annotation refinement block.

A. Initial FOG Manifestation Detection Block

We compared four models for initial FOG manifestation detection. A feature-based approach based on XGBoost was selected based on a study with extensive predefined features [19]. Two signal-based models were selected: a well-used open-source tool in clinical settings that encompassed a 1D CNN model [18] and a 1D TCN model that was recently proposed for fine-grained IMU-based FOG detection [9]. Additionally, we included a 2D CNN model trained with scalograms derived from IMU signals [19]. The data preprocessing steps, training strategy, and hyperparameter settings of all models, followed those proposed and described in their original studies. However, for model inference, we established a uniform setting for all models, which will be explained in section IV-B.

In the following subsections, we first explain the task of detecting FOG manifestations for the initial detection block. Following this, we provide detailed insights into the implementation specifics of the four selected models.

1) Problem Definition: An IMU trial can be represented as $X \in \mathbb{R}^{T \times C_{in}}$, where T is the number of samples, and C_{in} is the input feature dimension. Each IMU trial X 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. To generate predictions for each sample, each IMU trial was split from $X \in \mathbb{R}^{T \times C_{in}}$ into multiple windows with a fixed number of samples equal to the window size k. The model learns a function $f : X^i \rightarrow \hat{Y}^i$ that transforms a given input sequence $X^i = x_0^i, \ldots, x_{k-1}^i$ into an output label $\hat{y}^i$ that closely resembles the ground truth label for window i.

2) XGBoost: This study used the feature-based model proposed by Shi et al. [19], which applied the XGBoost [34] algorithm on 67 features generated from the IMU on the left tibia, including five frequency domain features, six entropy features, and 54 wavelet features. Two features calculated from magnetometer signals were removed as our dataset does not include magnetometers. The accelerometer signals were filtered with a 4th-order Butterworth band-pass filter (0.2-15 Hz), and the angular velocity signals were filtered with a 4th-order Butterworth low-pass filter (10 Hz), at a sampling frequency of 50Hz. The window size was set to one second with 50% overlap between consecutive windows [19].

3) 1D CNN: The second model is a 1D CNN model proposed by O'Day et al. [18]. The IMU data was split into windows of two seconds with 50% overlap between consecutive windows. Each window was normalized to zero mean and unit variance and augmented with random rotations about the individual IMU axes to simulate variation in sensor placement [18]. Notably, their original study incorporated a post-processing step to smooth oversegmented FOG episodes. However, in our study, we omitted this step when adapting the model for FOG manifestation detection, as its original approach is not suitable for multiclass classifications.

4) 1D TCN: Regarding the 1D TCN network, we applied the same 1D TCN selected from our previous study [9], specifically, we used the TCN architecture from [33]. The TCN architecture has a single TCN block comprising five temporal convolution layers. Employing a kernel size of 3, dimensionality of 32, and dilation rates designed to cover the input window size. This TCN utilized valid convolutions, directly transforming the input sequence of shape $k \times C_{in}$ into an output of shape 1×32. The output was passed through a linear layer with a softmax activation function, generating

probabilities for output L classes. For multi-class classification tasks in this study, we applied an unweighted cross-entropy loss.

5) 2D CNN: Lastly, this study used the 2D CNN model proposed by Shi et al. [19]. The raw IMU signals were first normalized and split into multiple windows with a window size of four seconds and 50% overlap between consecutive windows. The normalized signals in each window were used to generate scalograms with CWT.

B. Refinement Block

Given an IMU sequence of length T, the output of the initial detection block consists of the initial probabilities of each target class for each time sample within the sequence with a shape of $T \times L$. These initial probabilities serve as the input for the refinement block. For the refinement block, we employed a model derived from the MS-TCN architecture [23]. This refinement model is structured with four ResNet-style TCN stages. In each stage, the input sequence is initially processed through a 1×1 convolutional layer, adjusting the input dimension from $T \times L$ to $T \times C$, where C represents the number of filters. These modified features then pass through eight TCN layers, each comprising a dilated temporal convolution, Batch Normalization layer, Rectified Linear Unit function, and a residual connection. Subsequently, the final layer of each stage is a 1×1 convolutional layer equipped with a softmax activation function. This final layer outputs refined probabilities for the L classes for each sample in time. The training procedure and hyperparameters employed remained consistent with those detailed in [9]. In summary, the model was trained end-to-end for 50 epochs using two optimizers, both employing cross-entropy loss for the initial detection and the refinement block. Both blocks employed the Amsgrad optimizer with a learning rate of 0.0005, which decayed by a factor of 0.95 for each epoch.

IV. Evaluation

A. Dataset

We utilized an existing IMU dataset [9], and expanded it from twelve to eighteen subjects in this study. The testing setup for the extension remains identical to that outlined in [9]. The dataset includes eighteen PD patients, all recruited if they subjectively reported having at least one FOG episode per day with a minimum duration of five seconds. All subjects provided informed consent, and the study was approved by the Ethics Committee Research UZ/KU Leuven, with protocol number S65059. Subjects varied in their age (Range = 54 - 76; Mean = 67.33 ± 6.71 years), disease duration (Range = 3 - 23; Mean = 12.39 ± 5.01 years), and self-reported FOG severity with the New Freezing of Gait questionnaire score (19.11 ± 3.53) [35] and Movement Disorders Society Unified Parkinson's Disease Rating Scale (total score = 79.22 ± 24.07) [36]. Seven of the subjects underwent deep brain stimulation (DBS), and two subjects (S10 and S11) utilized mobility aids during the experiments.

The dataset was recorded with five Shimmer3 IMU sensors on all subjects, attached to the pelvis, both sides of the tibia and talus. All IMUs recorded at a sampling frequency of 64 Hz during the measurements. Synchronously, RGB videos were captured at 30 frames per second for offline FOG annotation purposes. FOG events were visually annotated at a frame-based resolution by a clinical expert, after which all FOG events were verified by another clinical expert, using Elan annotation software [12]. Annotators used the definition of FOG as a brief episode with the inability to produce effective steps, and the episode ended at the foot off that was followed by at least two effective steps [1], [12], which adopts a stricter definition of FOG that distinguishes shuffling and festination as non-FOG events, and only trembling in place and complete akinesia as FOG events. The definition of shuffling was based on [8], namely small steps with minimal forward progression, while festination was defined as a tendency to move forward with increasingly rapid but ever smaller steps [2].

The dataset featured the timed up-and-go (TUG) test, with turning in both directions, and the 360 turning in-place (360Turn) test [37], with alternating 360-degree turning for one minute. The tasks were measured with and without a dual task, namely the auditory Stroop task [37], and with and without a self-generated or researcher-imposed stopping. Stopping in the TUG was performed four times, twice with a stop in the straight walking part and twice with a stop in the turning part of the TUG; while stopping in the 360Turn was performed once. All pre-mentioned tasks were done first in the clinical Off-medication state (approximately 12 hours after the last PD medication intake) and repeated in the same order during the On-medication state (approximately one hour after medication intake). Based on the measurement protocol, 32 trials were collected for each subject. Subjects with more than 32 trials underwent repeated measurements, while subjects with fewer than 32 trials encountered technical difficulties.

B. Experimental Setting

The FOG detection models proposed in [18] and [19] exhibit several limitations. Firstly, they rely on majority voting, which alters expert annotations and impacts the determination of FOG severity outcome values, as depicted in Figure 2a. Secondly, these models lack the granularity necessary for precise identification of the onset and cessation points of each FOG episode, due to sliding windows with 50% overlap during inference time. Lastly, these models were trained and assessed without padding on both ends. Consequently, they tend to generate predictions shorter than the original input sequence. Illustrated in Figure 2b, even when generating predictions with a one-sample stride size, they still predict a shorter sequence of length $T - k + 1$ from an input sequence of length T and a window size of k.

To overcome these issues, we defined a uniform evaluation setting for comparing the models. Firstly, we addressed the issue of adapting the experts' annotations by defining the ground truth label as the center label of the window. This consistent ground truth labeling approach was employed during model training and inference, ensuring coherence and comparability. Secondly, to achieve consistent granularity in

TABLE I
DATASET CHARACTERISTICS

	DBS	#Trials	Duration in minutes	#FOG-Trials	#Trembling-Trials	#Akinetic-Trials	%TF	%TF-T	%TF-A	#FOG	#FOG-T	#FOG-A
S1	No	29	17.10	16	8	15	19.56	1.50	18.06	35	10	26
S2	No	29	13.90	9	7	8	12.64	5.64	7.00	34	16	18
S3	No	31	13.22	6	5	1	7.10	6.93	0.17	37	36	1
S4	No	27	10.48	12	12	0	7.89	7.89	0.00	30	30	0
S5	No	33	13.87	9	9	2	7.04	6.30	0.74	39	38	2
S6	No	32	12.84	1	1	0	0.10	0.10	0.00	1	1	0
S7	No	32	17.64	22	22	2	14.10	13.82	0.27	106	104	2
S8	No	33	13.48	0	0	0	0.00	0.00	0.00	0	0	0
S9	No	31	14.51	15	15	0	4.49	4.49	0.00	61	61	0
S10	Yes	21	25.59	21	20	13	52.86	35.53	17.32	111	103	25
S11	Yes	31	15.61	17	17	5	12.27	10.65	1.62	74	66	8
S12	Yes	34	20.40	10	6	7	2.07	0.79	1.28	19	9	10
S13	No	30	19.95	25	25	10	48.27	47.67	0.6	151	138	14
S14	Yes	29	12.34	5	2	3	0.98	0.42	0.56	5	2	3
S15	Yes	33	14.77	22	21	8	35.89	30.51	5.39	101	82	19
S16	Yes	28	11.69	8	5	7	2.25	0.94	1.32	31	12	19
S17	No	33	14.98	10	10	6	4.90	3.93	0.96	56	45	11
S18	Yes	29	12.78	8	8	4	26.18	25.22	0.95	44	40	4
Overall		545	275.18	216	193	91	16.81	12.96	3.85	935	793	162

Overview of the dataset for each subject, including the number of trials, total duration in minutes, the number of FOG trials (#FOG-trials), the number of trembling trials (#Trembling-Trials), the number of akinetic trials (#Akinetic-Trials), the percentage of time frozen (%TF), the percentage of time trembling (%TF-T), the percentage of time akinetic (%TF-A), the number of FOG episodes (#FOG), the number of trembling episodes (#FOG-T), and the number of akinetic episodes (#FOG-A). The #FOGs are not the sum of #FOG-T, and #FOG-A, as a FOG episode could contain both manifestations. Whether the subject underwent deep brain stimulation (DBS) was noted in the table.

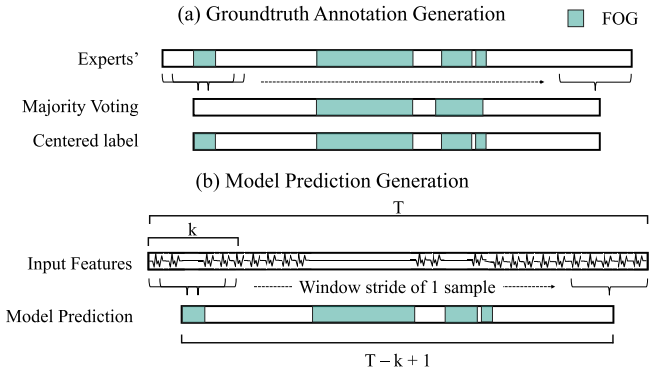

Fig. 2. Visual representation highlighting how sliding window-based FOG detection models alter the ground-truth experts' annotation. Figure (a) shows that majority voting results in minor temporal shifts and the removal of short segments. In contrast, using each window's centered label as ground truth maintains the experts' annotation. (b) Shows that shifting each prediction window with a stride of one sample enables frame by frame sample-wise predictions, but still reduces the sequence from length T to $T-k+1$, given a window of duration k.

predictions, we slid the windows with a stride of one sample during inference (as illustrated in Figure 1b). Meanwhile, during model training, we maintained a 50% overlap approach. Thirdly, each model was trained explicitly for multiple different window sizes, i.e., 1, 2, and 4 seconds (k = 64, 128, 256 samples), to identify the optimal window size for individual models. This comprehensive evaluation accounted for varying temporal contexts and allowed a more thorough analysis of the model's performance. Nevertheless, the window size impacts the predicted sequence length as illustrated in Figure 2b. To ensure a fair comparison among these models, we set the value of k to 4 seconds (256 samples) for all models. Conversely, while examining the overall model's performance (initial FOG manifestation detection block + refinement block) in distinguishing the FOG manifestations from other forms of volitional movement cessation, the entire T predicted sequence was evaluated. Lastly, except for the feature-based model, i.e., XGBoost, all models were trained using data from all five IMUs. The XGBoost exclusively employed features derived from the left tibia IMU (denoted as "leg" in [19]).

All 545 trials in the dataset were used to train and evaluate the models. The labels for the FOG manifestation detection task were three ($L = 3$), with $l = 0$ for non-FOG (i.e., walking, turning, sit-to-stand, stand-to-sit, and other volitional movement cessations), $l = 1$ for trembling in place, and $l = 2$ for complete akinesia. Next, we evaluated the model performance in discerning FOG manifestations from other types of volitional movement cessation, such as stopping and sitting. The model was trained with five target classes ($L = 5$), where $l = 1$ represents trembling in place, $l = 2$ represents complete akinesia, $l = 3$ represents stopping, $l = 4$ represents sitting, and $l = 0$ represents all other events (i.e., walking, turning, sit-to-stand, and stand-to-sit). All other events are hereinafter simply referred to as "normal gait".

All experiments employed a leave-one-subject-out cross-validation approach on all 18 subjects. The dataset was split into training (17 subjects) and testing (the left-out subject) sets. This procedure was repeated iteratively until each subject had been used for testing.

C. Metrics

This paper assessed FOG severity from a clinical perspective, primarily focusing on the percentage time-frozen (%TF) [13]. To further quantify the FOG manifestations, this study proposed the percentage time of trembling in place (%TF-T) and percentage time of complete akinesia (%TF-A), inspired by previous studies [38], [39]. The (%TF-T) was calculated as the total duration of trembling in place divided by the total duration of all tasks. The %TF-A was calculated as the total duration of complete akinesia divided by the total duration of all tasks. Table I summarizes the FOG severity for each subject in the dataset. To assess the agreement between model-predicted FOG severity and expert-annotated FOG severity, the intra-class correlation coefficient (ICC(2,1))

was used. The ICC value indicates the agreement between the model and the experts. A higher ICC value suggests higher agreement. As the clinical metrics are a summary of FOG severity per subject and insufficiently sensitive for model comparison [9], the F1 score was used to compare the performance of the different models.

The F1 score is a widely used metric for evaluating the accuracy of binary classification models. For sample-wise predictions, the comparison is performed at the individual sample level. Each prediction of the sample is classified as True Positive (TP), False Positive (FP), or False Negative (FN) based on the correspondence between the predicted and ground truth labels. The F1 score is calculated under the formula:

$$F1 = \frac{2 \times TP}{2 \times TP + (FP + FN)}$$

For the tasks of multi-class manifestation classification (non-FOG, trembling FOG, and akinetic FOG) and multi-class manifestation and volitional movement cessation classification (normal gait, trembling FOG, akinetic FOG, stopping, and sitting), we calculated an F1 score for each class individually in a one vs. all manner. This means that when computing the F1 score for a specific class, that class is considered positive, while all other classes are treated as negative. In our study, the non-FOG and normal gait classes were regarded as background classes, consistently treated as negative classes throughout the analysis. Specifically, for the multi-class manifestation classification, the F1 score was calculated under the formula for all subjects s:

$$\text{F1-Total}_s = (\text{F1-Trembling}_s + \text{F1-Akinetic}_s)/2$$

For the multi-class manifestation and volitional movement cessation classification, the F1 score was calculated under the formula:

$$\begin{aligned}\text{F1-Total}_s = (&\text{F1-Trembling}_s + \text{F1-Akinetic}_s \\ &+ \text{F1-Stopping}_s + \text{F1-Sitting}_s)/4\end{aligned}$$

These individual F1 scores were then averaged (F1-Total) for all eighteen subjects:

$$\text{F1-Total} = \frac{1}{18}\sum_{s=1}^{18} \text{F1-Total}_s$$

To assess the potential benefits of integrating a refinement block aimed at reducing oversegmentation errors, the F1-score at an intersection over union of 50% (F1@50) was employed [40]. Furthermore, we computed the F1@50 for both manifestations and derived an averaged F1@50-Total, calculated using the same formula as for F1-Total. All F1 scores were calculated for each subject by averaging scores across all trials. For non-FOG trials where the model detected no FOG episodes, an F1 score of 1 was assigned, indicating correct recognition in the absence of FOG [9].

D. Statistics

The Repeated Measures ANOVA test was used to investigate whether the differences between the models in the F1 scores were statistically significant. Post hoc paired Student's t-tests were applied to investigate significant differences between pair-wise models. Post hoc hypotheses were corrected for multiple comparisons with the Li correction [41]. Additionally, the paired Student's t-test was applied to examine significant differences between models trained with and without a refinement block. Homogeneity of variances across subjects was verified with Levene's tests. The Shapiro-Wilk test was used to determine whether the variables were normally distributed across subjects. The Bland-Altman plot was used to investigate systematic bias between FOG severity outcomes (i.e., %TF, %TF-T, and %TF-A) predicted by the model and the experts' annotation. The significance level for all tests was set at 0.05. Analyses employed SciPy1.7.11, bioinfokit2.1.0, statsmodels0.13.2, and pingouin0.3.12 in Python 3.7.11. Post hoc tests used scmamp0.2.55 in R 4.0.3.

TABLE II
COMPARISON OF THE FOUR MODELS IN TERMS OF THE F1 SCORE

Model	Size (s)	F1-Tremblnig	F1-Akinetic	F1-Total
XGBoost	1	0.48	0.77	0.63
	2	0.56	0.80	0.68
	4	**0.63**	**0.81**	**0.72**
1D CNN	1	0.28	0.15	0.22
	2	0.32	0.22	0.27
	4	**0.37**	**0.30**	**0.34**
2D CNN	1	0.39	0.26	0.32
	2	**0.67**	0.69	0.68
	4	0.65	**0.81**	**0.73**
1D TCN	1	0.65	0.77	0.71
	2	0.73	0.76	0.74
	4	**0.75**	**0.79**	**0.77**

We modified four FOG detection models to detect two FOG manifestations and compared the performance of these models with three different window sizes (i.e., 1, 2, and 4 seconds).

V. RESULTS

A. FOG Manifestation Detection: Initial Detection Model Comparison

We compared the four models for detecting the initial manifestation of FOG. Models trained with a four-second window size achieved the highest F1-Total and F1-Trembling scores (Table II). ANOVA tests revealed significant differences between all F1 scores (all $p<0.005$). Post hoc tests in Figure 3 confirmed that the 1D TCN outperformed XGBoost in terms of F1-Trembling and F1-Total, 1D CNN in terms of all three F1 scores, and 2D CNN in terms of F1-Trembling.

As depicted in Figure 4, both the 1D CNN and XGBoost exhibited numerous instances of oversegmentation, with the 1D CNN particularly prone to false positives for akinetic FOG. Although the 2D CNN model demonstrated fewer errors, it misclassified trembling as akinetic and failed to detect short FOG episodes. Conversely, the 1D TCN model exhibited robustness in distinguishing trembling from akinetic FOG episodes and displayed fewer instances of oversegmentation compared to other models.

B. FOG Manifestation Detection: Initial Detection + Refinement Block

Subsequently, we explored the advantages of incorporating a refinement block into the best-performing initial detection model, namely the 1D TCN. As shown in Table III, concerning

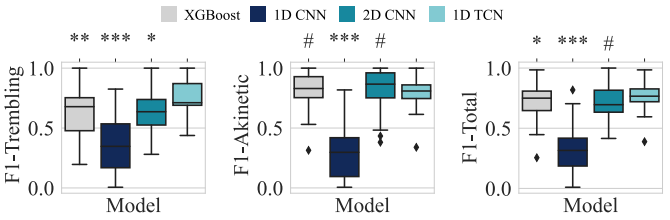

Fig. 3. The spread of the F1-Score across subjects. The ANOVA test showed a significant difference between all F1-Scores of the four models. The significance levels of the post hoc tests with respect to the 1D TCN model (corrected for three pairwise comparisons) are visualized above their respective box plot. Significance levels were visualized as: p≤0.005 (***), p≤0.01 (**), p≤0.05 (*), and p>0.05 (#).

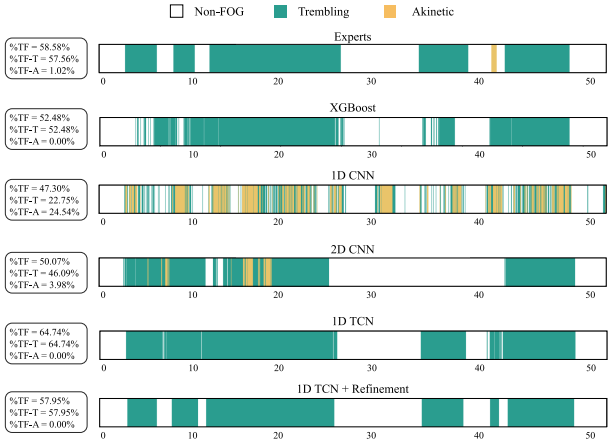

Fig. 4. Representative example of a FOG trial with predictions of the best four models and the 1D TCN model with a refinement block compared with the experts' annotation. The figures visualize the over-segmentation of the models. The x-axis indicating trial time in seconds.

sample-wise F1 scores, the F1-Akinetic and F1-Total exhibited a significant improvement in the model with the refinement block in comparison to the model without it. Moreover, the model with a refinement block exhibited significantly higher F1@50 scores than the model without it. This enhancement indicates a reduction in oversegmentation errors, emphasizing the evident advantage of incorporating such a refinement block for precise detection of FOG manifestations. A visual comparison between the 1D TCN with and without a refinement block is illustrated in Figure 4.

C. FOG Manifestation Severity Assessment

Next, we assessed the overall performance of the model, which consists of the 1D TCN as the initial detection block followed by a refinement block, in terms of FOG manifestation severity outcomes. The results indicated a strong agreement between the model and experts for both %TF (ICC = 0.92, CI=[0.81,0.97]) and %TF-T (ICC = 0.86, CI=[0.66,0.95]), as well as a moderately strong agreement for %TF-A (ICC = 0.78, CI=[0.51,0.91]). Bland–Altman plots (Figure 5) revealed no systematic error between our model and expert annotations, with a mean bias of 2.17% (CI=[−0.94,5.28]) for %TF, 2.91% (CI=[−0.52,6.33]) for %TF-T, and −0.74% (CI=[−3.29,1.82]) for %TF-A. Notably, three outliers: S10, S14, and S15, were identified, all of whom underwent DBS, with S10 also using a mobility aid during experiments. Misclassifications occurred, such as trembling being mistakenly

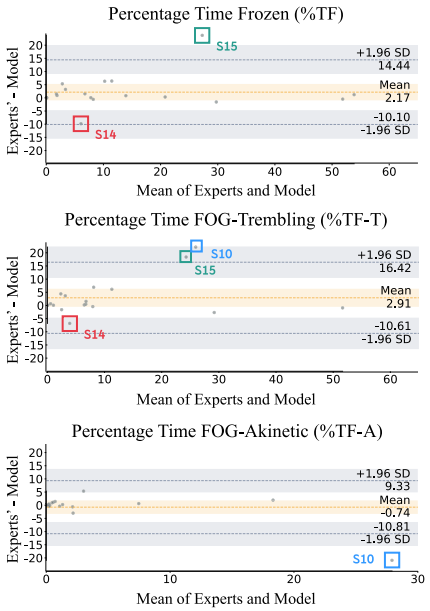

Fig. 5. The Bland-Altman plot compares three clinical metrics between the model and experts. Each dot represents score differences per patient against the mean score. Orange shows 95% CI for mean bias, gray for upper and lower LOA CI. Three outliers are marked (S10: blue, S14: red, S15: green).

labeled as akinetic for S10, likely due to the mobility aid, while festination and continuous shuffling were frequently misinterpreted as FOG for S14. Additionally, oversegmentation of numerous long FOG episodes during Off-medication occurred for S15, a characteristic not present in other training subjects. Additionally, as shown in Bland–Altman plots, the limits of agreement (LOA) for %TF-T ranged from −10.61% to 16.42%, while for %TF-A, the LOA ranged from −10.81% to 9.33%. The lower ICC for akinetic FOG compared to trembling suggests lower consistency with experts. However, the narrower LOA for %TF-A indicates greater confidence in predicting %TF-A, likely due to reduced variability compared to %TF-T.

D. FOG Manifestations Versus Other Forms of Volitional Movement Cessation

Next, we investigated the proposed model's ability (i.e., 1D TCN combined with a Refinement Block) to distinguish FOG manifestations from volitional movement cessation by training the model with five target classes, i.e., normal gait, trembling, akinetic, stopping, and sitting. The model achieved an overall F1 score of 0.65 and an overall F1@50 score of 0.63. The dataset consists of 275.18 minutes, with 75.49% normal gait samples, 12.96% trembling samples, 3.85% akinetic samples, 4.02% stopping samples, and 3.67% sitting samples. As seen in the confusion matrix (Figure 6), the model correctly predicted 94% of the normal gait samples, 71% of the akinetic samples, 65% of the stopping samples, and 82% of the sitting samples. However, the model struggled to accurately identify trembling samples, with only 56% of them correctly classified, while 28% were classified as normal gait and 14% as akinetic. To investigate the model's ability to distinguish between stopping and FOG manifestations, we split up the evaluation for non-FOG and FOG trials. When evaluating

TABLE III
COMPARISON OF THE MODELS TRAINED WITH AND WITHOUT A REFINEMENT BLOCK

Model	F1-Tremblnig	F1@50-Trembling	F1-Akinetic	F1@50-Akinetic	F1-Total	F1@50-Total
1D TCN	0.75	0.63	0.79	0.76	0.77	0.70
1D TCN + Refinement block	**0.76**	**0.72**	**0.81**	**0.79**	**0.78**	**0.75**
Statistics (t-value, p-value)	1.120, p=0.278	5.868, p<0.005	2.162, p=0.045	3.291, p<0.005	2.250, p=0.038	5.688, p<0.005

We assessed the model performance in detecting two FOG manifestations by comparing models trained with and without a refinement block. We reported the F1 and F1@50 for the two manifestations, along with an overall score. To determine if the differences in scores were statistically significant, we employed a paired t-test for evaluation.

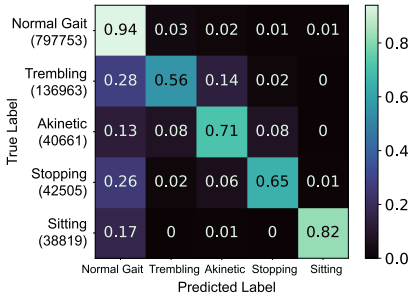

Fig. 6. The normalized confusion matrix visualizes the model's ability to distinguish between the five classes. Each row shows the true label and the total number of samples for that label, while each column shows the predicted label. The matrix is normalized by dividing each element by the total number of samples in the true class, indicating the ratio of correct predictions for each class.

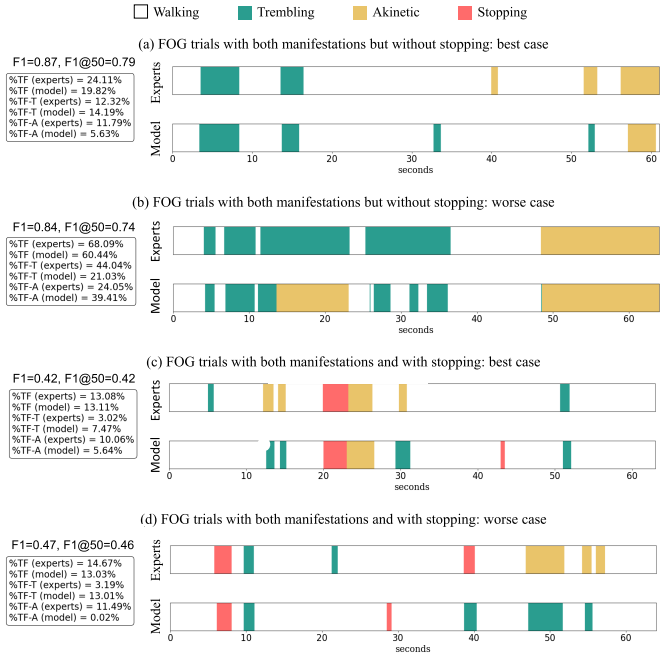

Fig. 7. Overview of the annotations for four IMU trials (selected from S2 for 360Turn tasks) featuring FOG trials with both manifestations but without stopping, and FOG trials with both manifestations and with stopping. For each condition, the trial with the best model prediction and the worst model prediction was selected. The figures compare the experts annotation (top) with the model prediction (bottom). The color codes are: white=normal gait, green=trembling, yellow=akinetic, red=stopping, and blue=sitting. The x-axis denotes the trial time in seconds.

non-FOG trials, the model could accurately annotate 75% of stopping samples as stopping. In contrast, when evaluating FOG trials, the model could only correctly annotate 51% of stopping samples, with 31% as normal gait, and 12% as akinetic. Within these FOG trials, the model correctly annotated 71% of stopping in the TUG tasks, but only 37% in the 360Turn tasks (38% as normal gait and 17% as akinetic). Additionally, in FOG trials, the model correctly annotated 48% of akinetic episodes in the TUG tasks (21% as normal gait and 21% as stopping), while it correctly annotated 81% of akinetic episodes in the 360Turn tasks. These phenomena are demonstrated in the qualitative results presented in Figure 7, which shows FOG trials with both manifestations, both with and without stopping. In Figures 7a and 7b, the model struggled to distinguish between trembling and akinetic when both manifestations were present, resulting in lower F1 scores. Similarly, Figures 7c and 7d illustrate the model's difficulty in differentiating between trembling and stopping, as well as between trembling and akinetic, also leading to lower F1 scores.

VI. DISCUSSION

Previous FOG assessment studies [9], [18], [19], [22] combined various types of FOG into a single category. However, FOG can have different manifestations, which may have other pathophysiologic origins [8]. Therefore, objectively detecting these different FOG manifestations is crucial to tailor future FOG treatment approaches. To address this bottleneck, this study proposed a DL model to support the detection of two FOG manifestations, i.e., trembling and akinetic FOG. Our model comprises an initial detection block and a refinement block. We adapted and compared four state-of-the-art FOG detection models to identify the best model for initial FOG manifestation detection. Furthermore, the MS-TCN model was applied for the refinement block. Our results exhibited that 1D TCN [33] model statistically outperformed XGBoost [19], 1D CNN [18], and 2D CNN [19] in detecting FOG manifestations, particularly regarding the F1-Trembling. Moreover, the incorporation of a refinement block significantly reduced oversegmentations, resulting in higher F1@50 scores.

To quantify the severity of FOG manifestations, previous studies calculated the percentage of each FOG manifestation with respect to the total duration of FOG [38]. However, this metric does not measure the duration of each specific FOG manifestation directly. Instead, it indicates the distribution of different manifestations within the total duration of FOG episodes. Therefore, using the percentage of each manifestation within observed FOG as a metric to quantify the severity of FOG manifestations may not be reliable. As a result, inspired by previous studies [38], [39], we proposed two metrics extended from %TF, i.e., %TF-T and %TF-A, to quantify FOG manifestation severity. Our proposed model showed a strong agreement with the experts' observations for %TF-T (ICC=0.86) and a moderately strong agreement for %TF-A (ICC=0.78). The ICC for FOG manifestation severity between independent raters was reported as 0.31 (CI=[0.11,0.49]) for the percentage of trembling and 0.44 (CI=[0.35,0.54]) for the percentage of akinetic [38]. Although [38] showed that

annotating FOG manifestations is challenging, which would result in a low inter-rater agreement, our model prediction showed a moderate to strong agreement with our experts' annotation, showing its ability to learn how our experts' annotated the trials.

Next, we investigated the model performance in distinguishing FOG manifestations from other forms of volitional movement cessation, i.e., stopping and sitting, by evaluating the model trained explicitly for the five classes: normal gait, trembling, akinetic, stopping, and sitting. Results showed that our model could correctly detect sitting from FOG manifestations. However, stopping could only be accurately detected in TUG trials or trials that did not contain FOG. In 360Turn trials with FOG, short stopping events were often mislabeled as normal gait. Akinetic episodes were accurately detected only in 360Turn trials, while in TUG trials, trembling was misclassified as akinetic, and akinetic as stopping. Hence, motor signals alone may be insufficient to distinguish stopping from FOG, particularly during complex motion sequences that are likely to be encountered in everyday life. A promising avenue is to amalgamate motor and physiological signals (e.g., heart rate), which have recently shown potential in distinguishing between FOG and stopping, but lack the expressivity to distinguish between FOG and gait [24], which was highly distinguishable in our approach. Therefore, including physiological signals in our method seems a promising future improvement to disentangle the different FOG manifestations.

Furthermore, the results showed that the agreement between our model and the experts in terms of %TF-A was lower than %TF-T. This finding shows different results than [38] with previously reported lower inter-rater ICC values for trembling compared to akinetic FOG. Trembling FOG (i.e., alternating tremulous oscillations with no forward progression) and akinetic FOG (i.e., no visible movement in the lower limbs) are determined based on observable leg motion. There are several potential explanations: Firstly, some trembling movements may not be observable in the videos by the experts, especially if the movements are very small. Although our study procedure had participants wearing tight-fitting shorts, this may become even more challenging in clinical practice where patients with FOG are wearing their own comfortable long-legged pants. Secondly, as FOG manifestations may shift within one episode, it becomes very challenging and time-consuming for the experts to label it to the highest detail. Therefore, they resort to labeling the episode (or larger blocks of the episode) to the manifestation that is dominantly present. Moreover, when annotating FOG, experts may observe no leg motion, making it challenging for them to discern whether the subject is hesitating during tasks, stopping volitionally, or experiencing akinetic FOG.

Several limitations should be considered. Firstly, we adapted and compared four FOG detection models for initial FOG manifestation detection, using window sizes of 1, 2, and 4 seconds. The best performance was at 4 seconds. Due to limited GPU memory, we couldn't use an 8-second window. However, differences in F1 scores between models were more significant than those between window sizes. Secondly, the dataset used in this study consisted of videos annotated sequentially by two clinical experts, with the second expert verifying and correcting the annotations made by the first expert. Due to our sequential annotation process, there was no opportunity to measure inter-rater agreement in terms of %TF-T and %TF-A to compare against our models' annotations. The third limitation involves adapting four state-of-the-art FOG detection models from binary FOG detection to three-class FOG manifestation detection. To address this, future research could focus on fine-tuning or developing alternative DL models for multi-class classification tasks, enhancing comparative analysis. The fourth limitation is the limited FOG manifestations in the dataset: 35.66 minutes of trembling and 10.59 minutes of akinetic episodes out of 275.18 total minutes. Given the infrequency of FOG and the rarity of akinetic FOG [37], [38], this ratio is in line with the literature. Our dataset uniquely includes detailed FOG manifestation annotations and is one of the larger datasets available (N=18, #FOG=935). For comparison, other public domain datasets include Daphnet (N=10, #FOG=237) [26], O'Day et al. (N=16, #FOG=211) [18], CuPid (N=18, #FOG=237) [30], and Rempark (N=21, #FOG=1321) [42]. Further validation will be required in larger and less heterogeneous cohorts. The fifth limitation is that gait demographics (such as height, weight, and leg length) were not collected in this study. Consequently, the potential influence of gait demographics on FOG manifestation detection was not investigated.

VII. Conclusion

The current study is the first attempt to automatically quantify FOG manifestations using DL. Our approach demonstrated a strong agreement with experts' annotations on %TF and %TF-T and a moderately strong agreement for %TF-A. Future work is now possible to establish whether these results hold for a larger and more varied verification cohort.

Acknowledgment

The authors would like to thank the participants for their willingness to participate.

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] M. Rudzińska et al., "Causes and consequences of falls in Parkinson disease patients in a prospective study," *Neurologia I Neurochirurgia Polska*, vol. 47, no. 5, pp. 423–430, 2013.

[5] P. H. S. Pelicioni, J. C. Menant, M. D. Latt, and S. R. Lord, "Falls in Parkinson's disease subtypes: Risk factors, locations and circumstances," *Int. J. Environ. Res. Public Health*, vol. 16, no. 12, p. 2216, Jun. 2019.

[6] S. S. Paul, C. G. Canning, C. Sherrington, S. R. Lord, J. C. T. Close, and V. S. C. Fung, "Three simple clinical tests to accurately predict falls in people with Parkinson's disease," *Movement Disorders*, vol. 28, no. 5, pp. 655–662, May 2013.

[7] S. Perez-Lloret et al., "Prevalence, determinants, and effect on quality of life of freezing of gait in Parkinson disease," *JAMA Neurol.*, vol. 71, no. 7, p. 884, Jul. 2014.

[8] 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.

[9] 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, pp. 1–22, Feb. 2024.

[10] P. Ginis et al., "Contribution of cognitive load to akinetic and trembling freezing of gait manifestations in patients with Parkinson's disease," *Movement Disorders*, vol. 37, p. S652, Jan. 2022.

[11] M. Falla, G. Cossu, and A. Di Fonzo, "Freezing of gait: Overview on etiology, treatment, and future directions," *Neurolog. Sci.*, vol. 43, no. 3, pp. 1627–1639, Mar. 2022.

[12] 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.

[13] T. R. Morris et al., "A comparison of clinical and objective measures of freezing of gait in Parkinson's disease," *Parkinsonism Rel. Disorders*, vol. 18, no. 5, pp. 572–577, Jun. 2012.

[14] K. Hu et al., "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 1215–1225, Apr. 2020.

[15] R. Sun, Z. Wang, K. E. Martens, and S. Lewis, "Convolutional 3D attention network for video based freezing of gait recognition," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–7.

[16] 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.

[17] 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.

[18] 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, pp. 1–15, Dec. 2022.

[19] 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.

[20] S. Pardoel, G. Shalin, J. Nantel, E. D. Lemaire, and J. Kofman, "Early detection of freezing of gait during walking using inertial measurement unit and plantar pressure distribution data," *Sensors*, vol. 21, no. 6, p. 2246, Mar. 2021.

[21] B. Sijobert, J. Denys, C. A. Coste, and C. Geny, "IMU based detection of freezing of gait and festination in Parkinson's disease," in *Proc. IEEE 19th Int. Funct. Electr. Stimulation Soc. Annu. Conf. (IFESS)*, Sep. 2014, pp. 1–3.

[22] 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, pp. 1–14, Dec. 2022.

[23] 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. 3570–3579.

[24] 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, pp. 1–15, Apr. 2023.

[25] S. T. Moore, H. G. MacDougall, and W. G. Ondo, "Ambulatory monitoring of freezing of gait in Parkinson's disease," *J. Neurosci. Methods*, vol. 167, no. 2, pp. 340–348, Jan. 2008.

[26] 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.

[27] A. Delval et al., "Objective detection of subtle freezing of gait episodes in Parkinson's disease," *Movement Disorders*, vol. 25, no. 11, pp. 1684–1693, Aug. 2010.

[28] M. Mancini, N. Hasegawa, D. S. Peterson, F. B. Horak, and J. G. Nutt, "Digital measures of freezing of gait across the spectrum of normal, non-freezers, possible freezers and definite freezers," *J. Neurol.*, vol. 270, no. 9, pp. 4309–4317, Sep. 2023.

[29] M. G. Tsipouras et al., "On assessing motor disorders in Parkinson's disease," in *Proc. Int. Conf. Wireless Mobile Commun. Healthcare*, vol. 55, 2011, pp. 35–38.

[30] 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*, 2012, pp. 123–130.

[31] 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.

[32] 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.

[33] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7745–7754.

[34] 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.

[35] 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.

[36] S. Fahn and R. Elton, "The unified Parkinson's disease rating scale," in *Recent Developments in Parkinson's Disease*, S. Fahn, C. D. Marsden, D. Calne, and M. Goldstein, Eds., Florham Park, NJ, USA: MacMillan HealthCare Information, 1987, pp. 153–163.

[37] N. D'Cruz et al., "Dual task turning in place: A reliable, valid, and responsive outcome measure of freezing of gait," *Movement Disorders*, vol. 37, no. 2, pp. 269–278, Feb. 2022.

[38] Y. Kondo et al., "Measurement accuracy of freezing of gait scoring based on videos," *Frontiers Hum. Neurosci.*, vol. 16, p. 309, May 2022.

[39] K. Kompoliti, C. G. Goetz, S. Leurgans, M. Morrissey, and I. M. Siegel, "'On' freezing in Parkinson's disease: Resistance to visual cue walking devices," *Movement Disorders*, vol. 15, no. 2, pp. 309–312, 2000.

[40] 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.

[41] 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.

[42] 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.