MDPI

Article

An explainable spatial-temporal graphical convolutional network to score freezing of gait in parkinsonian patients

Hyeokhyen Kwon[1], Gari D. Clifford[1], Imari Genias[2], Doug Bernhard[3], Christine D. Esper[3], Stewart A. Factor[3], and J. Lucas McKay[1,3]*

[1] Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, USA; hyeokhyen.kwon@dbmi.emory.edu; gari@dbmi.emory.edu; dbernha@emory.edu
[2] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA; imari.genias@emory.edu
[3] Jean and Paul Amos Parkinson's disease and Movement Disorders Program, Department of Neurology, School of Medicine, Emory University, Atlanta, GA, USA; cedoss@emory.edu; sfactor@emory.edu
* Correspondence: lucas@dbmi.emory.edu

Abstract: Freezing of gait (FOG) is a poorly understood heterogeneous gait disorder seen in patients with parkinsonism which contributes to significant morbidity and social isolation. FOG is currently measured with scales that are typically performed by movement disorders specialists (ie. MDS-UPDRS), or through patient completed questionnaires (N-FOG-Q) both of which are inadequate in addressing the heterogeneous nature of the disorder and are unsuitable for use in clinical trials The purpose of this study was to devise a method to measure FOG objectively, hence improving our ability to identify it and accurately evaluate new therapies. We trained interpretable deep learning models with multi-task learning to simultaneously score FOG (cross-validated F1 score 97.6%), identify medication state (OFF vs. ON levodopa; cross-validated F1 score 96.8%), and measure total PD severity (MDS-UPDRS-III score prediction error $\leq$ 2.7 points) using kinematic data of a well-characterized sample of N=57 patients during levodopa challenge tests. The proposed model was able to identify kinematic features associated with each FOG severity level that were highly consistent with the features that movement disorders specialists are trained to identify as characteristic of freezing. In this work, we demonstrate that deep learning models' capability to capture complex movement patterns in kinematic data can automatically and objectively score FOG with high accuracy. These models have the potential to discover novel kinematic biomarkers for FOG that can be used for hypothesis generation and potentially as clinical trial outcome measures.

Keywords: Deep Learning; Motion Capture; Multi-task Learning; Parkinson's Disease

Citation: Kwon, H.; Clifford, G. D.; Genias, I.; Bernhard, D.; Esper, D. C.; Factor, S. A.; McKay, J. L. Explainable spatial-temporal graphical convolutional network to score freezing of gait. *Sensors* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Parkinson's Disease (PD) is a slowly progressive neurodegenerative disorder that predominantly affects dopamine-producing neurons in the brain, and individuals with PD exceed more than 10 million people worldwide [1,2]. One of the most disabling features of PD and one of the greatest unmet needs is freezing of gait (FOG), which unfortunately is not always clearly treatable medically and/or surgically. FOG is described as brief arrests of stepping when initiating gait, turning, or walking straight ahead [3–5]. When a person freezes, they feel like their feet are "glued" to the floor. FOG is a frequent cause of falls and serious injuries, and represents a significant public health burden (∼86% of patients fall each year) [6–8].

One critical factor limiting our ability to treat FOG is that clinicians measure it relatively coarsely, primarily with expert rater observations as part of the Movement Disorder Society-Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS-III) scale [9]. This scale requires specially trained raters who have typically completed movement disorders training. In addition, despite being resource-intensive, FOG is only quantified with a single item

Table 1. Clinical and demographic features of study participants.

	PD-FOG	PD-NoFOG	PP-FOG
N	35	17	5
Age, y	69 ± 7	67 ± 12	66 ± 6
Sex, M/F	30/5	11/6	2/3
Disease duration, y	10.5 ± 6.7	6.0 ± 3.6	6.0 ± 3.3
LED, mg	1429 ± 673	833 ± 303	1258 ± 640
MDS-UPDRS-III (OFF)	34.0 ± 10.6	30.8 ± 13.2	39.4 ± 7.8
MDS-UPDRS-III (ON)	20.7 ± 8.7	18.4 ± 14.5	31.6 ± 9.0
NFOG-Q	20.1 ± 4.9	0.0 ± 0.0	17.8 ± 7.5

on an ordinal scale from 0 to 4, which may be too insensitive to detect small beneficial effects. The most established self-reporting scale used in research settings, the N-FOG-Q is acknowledged to be insufficiently sensitive for clinical trial use [10]. Previous work have shown that FOG may be associated with non-dopaminergic system changes [3,11,12], which suggests the potential for new treatments beyond dopaminergic medications like carbidopa-levodopa [4]. However, developing a novel drug that is effective in treating FOG requires accurately quantifying FOG to increase the precision for clinical trials.

Multiple studies have proposed methods to phenotype and rate FOG from kinematic data during walking. For example, those include capturing impaired gait patterns from lower back motion [13], describing gait complexity as a topological nonlinear dynamics system [14], or exploring combinations of sensor locations (shank, thigh, waist), axes (orthogonal, mediolateral, and antero-posterior), window lengths, and features (statistical, frequency, and time-series) to find the best setting that captures FOG characteristics.

Much of the prior work is characterized by a few substantial limitations [15], including 1) a small number of body-worn sensor locations, 2) small sample sizes with mostly early-stage PD patients lacking of severe FOG cases, 3) little consensus on proposed methods across studies, and 4) a relative paucity of studies conducted in the ON- and OFF-medication states, which is necessary to develop technology that will work over the entire medication cycle.

More importantly, most prior studies rely on hand-crafted features for identifying FOG, which may neglect important latent features within the data. For example, relative power in a "freeze band" of accelerometry or other signals [16–18], peak detection or similar methods applied to body segment motion [19,20], cycle-to-cycle variation in gait parameters [21], or a combination of the above were used in a support vector machine or other shallow machine learning models [22]. Due to the variability and complexity of FOG behavior, it is unlikely that manually designed spectral features will capture all the characteristics of FOG phenotypes. The popular "freeze band" analysis cannot capture pure akinetic freezing, which does not present with tremulousness.

Here, we use a deep learning approach to capture complex patterns in kinematic data and automatically score FOG, as well as identifying medication state and measure total MDS-UPDRS-III score during a rigorous levodopa challenge paradigm [3]. We analyzed over 30 hours of 3D motion capture data of 57 patients with varying PD disease duration and FOG severity, including 5 patients with primary progressive FOG, a distinct condition in which FOG presents without parkinsonian features [5]. This dataset is among the largest samples seen in the FOG literature (in which the average sample size was recently estimated as 18 ± 15 [15]). To our knowledge, this work is the first application of interpretable deep learning to solve such a multi-task problem in PD.

2. Materials and Methods

We trained an interpretable deep learning model on whole-body 3D kinematic data taken from behavioral motor tasks in N=57 patients with and without FOG. Clinical,

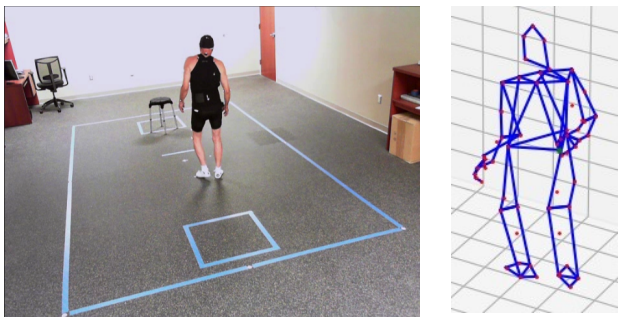

Figure 1. Motion capture recording during timed-up-and-go testing. Left: clinical motion capture laboratory. Right: example of kinematic marker data. Participants were instructed to rise from the stool, walk to the taped box, and return three times during each test.

imaging, and cerebrospinal fluid analysis results from patients in this sample have been reported previously [3,23].

2.1. Behavioral testing

2.1.1. Study participants

Although this was an observational study for which registration was not required, it was registered through clinicaltrials.gov (NCT02387281). Participants were recruited from the Emory Movement Disorders Clinic and provided written informed consent according to procedures approved by Emory University IRB. The inclusion criteria included: Age $\geq$ 18 years; PD diagnosis according to United Kingdom Brain Bank criteria [24]; Hoehn & Yahr stage I-IV in the OFF state; ability to sign a consent document and willing to participate in all aspects of the study. Participants with FOG were additionally required to have FOG noted in medical history and confirmed visually by examiner. Exclusion criteria included: vascular parkinsonism and drug-induced parkinsonism as well as the presence of cerebrovascular disease or extensive white matter disease; prior treatment with medications that cause parkinsonism; neurological or orthopedic disorders interfering with gait; dementia or other medical problems precluding completion of the study protocol. Demographic and clinical characteristics of study participants are presented in Table 1.

2.1.2. Levodopa challenge paradigm

Each participant was assessed twice using an identical testing protocol: first, in the practically defined "OFF" state $>$ 12 hours after the last intake of all antiparkinsonian medications, and second, after a levodopa equivalent dose of $\sim$150% of the typical morning dose sufficient to elicit a full "ON" state. Additional details of the levodopa testing procedure have been presented previously [3]. In each state, they were assessed with the MDS-UPDRS-III motor exam [9] and with timed-up-and-go (TUG) tests in the motion capture laboratory [25] in normal and cognitive dual-task conditions [26], with three replicates each. Patients were instructed to turn left on all TUG tests, consistent with our clinical testing paradigm. Performance was scored in person, and scores were confirmed from video if necessary.

2.1.3. Motion capture

TUG tests were recorded using 3D optical motion capture (Motion Analysis Corporation, Santa Rosa, CA). The motion capture facility is located in our clinical center and measures 5.8m $\times$ 9.0m with a capture area of 3.0m $\times$ 4.6m, and is equipped with 14 Osprey cameras with a resolution of 640 $\times$ 480 running at 120 Hz. During the testing session, patients wore tight-fitting clothes and were instrumented with reflective adhesive markers as recommended by the motion capture system manufacturer, configured as a superset of the Helen-Hayes kinematic marker set [27], incorporating additional markers on the hands. An example of the kinematic marker data is shown in Figure 1. Prior to analysis, all

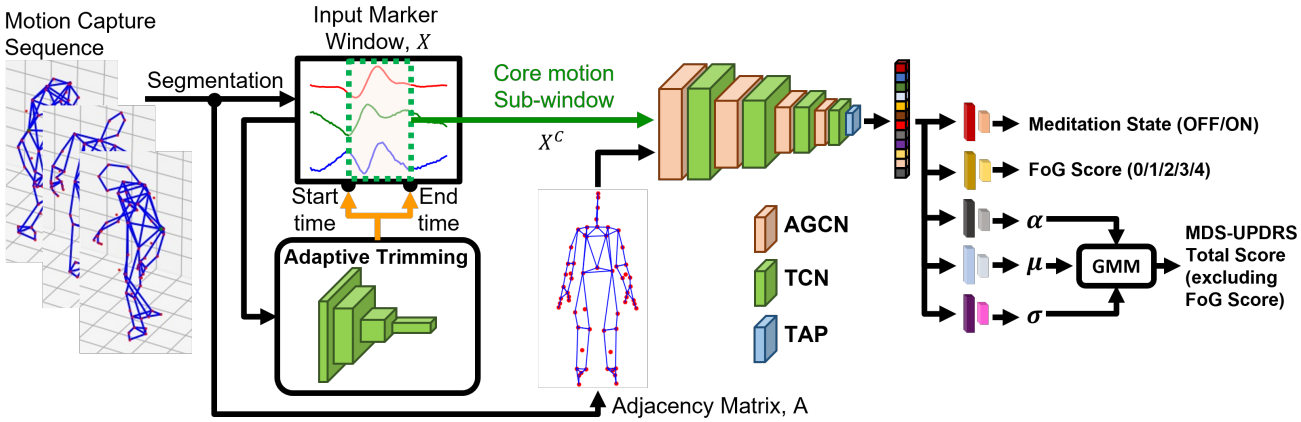

Figure 2. Overall model architecture. The recorded motion capture sequence is segmented into 4-second analysis windows, which are first processed with the Adaptive Trimming (AT) model. AT model, which uses a 4-layer temporal convolutional network (TCN), predicts the start and end index of the core motion segment that is most relevant for the prediction task. The core motion segment is processed with a 4-layer adaptive temporal-spatial graphical convolutional network (AGCN), which automatically learns the attention map for the most relevant joint and limb motion for the prediction task. The feature representation from the final layer of AGCN is processed with temporal average pooling (TAP) to summarize temporal information, which is then used to predict medication state, FOG score, and MDS-UPDRS-III total score (excluding FOG score) at the same time. Specifically for regressing the MDS-UPDRS-III total score (excluding FOG score), Gaussian Mixture Model (GMM)-based regressor is used to take account of the non-Gaussian distribution of the target values.

kinematic data were projected to a hip-centered coordinate system and normalized to zero mean and unit standard deviation.

2.2. Modeling

2.2.1. Model Overview

Our proposed model is an attention-based adaptive graphical convolutional network (AGCN, [28]) with adaptive trimming [29]. The overall model architecture is shown in Figure 2. We process the 3D motion capture data following a common deep learning-based human activity recognition paradigm [30,31]. Motion capture data from each testing sequence is comprised of three channels (x, anterior/posterior; y, lateral; z, vertical) for each of 60 kinematic markers, for a total of 180 independent channels. The data from each sequence is segmented into analysis windows of 4 seconds, $N \times C \times T$, where $N = 60$, $C = 3$, and $T = 480$ for 120 Hz signals, with 1 second intervals. A 4-second analysis window is chosen to capture a sufficient duration of FOG episodes while patients are walking [32].

Each 4-second analysis window is labeled with medication state (OFF/ON), FOG score (0, 1, 2, 3, or 4, from MDS-UPDRS-III item 3.11), and MDS-UPDRS-III total score, excluding the FOG score. The proposed model is trained to predict the labels for each analysis window based on the 3D kinematic data.

For extracting kinematic features from each 4-second window, the proposed model considers two aspects: 1) the core motion segment, which corresponds to the most relevant section of time within each window, and 2) the most relevant kinematic marker (joint) and edge between markers (bone) for the given multiple prediction tasks. The model uses adaptive trimming (AT) to identify the core motion segment within each 4-second analysis window and trims the given input signal for further analysis [29]. The trimmed core motion segment is processed to automatically identify the most relevant joint and limb parts for making predictions by using the AGCN model [28]. The AGCN model extracts feature representation by treating a given core motion sequence of 60 markers as a graphical model representing a human skeleton, where each node is marker position (joint)

and the edge is connectivity between markers (bone) of ongoing kinematic sequence. The AGCN automatically learns the most important joint and bone motions across all samples (domain-dependent attention weights) and specific to given samples (input-specific attention weights) for predicting medication state, FOG score, and MDS-UPDRS-III total score, excluding FOG score.

2.2.2. Trimming Core Motion Segment

Adaptive Trimming (AT) enables the model to identify core motion segments and to flexibly trim the signal that is most useful for specific prediction tasks of interest. From a previous study, AT was very effective at detecting gym exercise classification task [29]. In this work, the AT is fully trained with a given kinematics dataset to predict the start and end time of the core motion segment from a 4-second analysis window, $X \in \mathbb{R}^{N \times C \times T}$.

$$c = sigmoid(F^{center}(F_{at}(X))) \tag{1}$$
$$w = \exp(F^{width}(F_{at}(X))) \tag{2}$$

F_{at} is a four-layer convolutional network for extracting feature to predict core motion locations. F^{center} and F^{width} are two-layer fully connected models to predict center location, $0 < c < 1$ and width of core motion segment, $0 < w < 1$, which are further processed to derive start, s, and end, e, indices of given window, where $0 < s, e < T$.

$$s = T \times sigmoid(c - \frac{w}{2}) \tag{3}$$
$$e = T \times sigmoid(c + \frac{w}{2}) \tag{4}$$
$$X^C = X[s:e] = F_{crop}(X, s, e) \tag{5}$$
$$= F_{sampler}(F_{grid\_gen}(X), s, e) \tag{6}$$

The cropping operation adapts grid generator, F_{grid_gen}, and sampler, $F_{sampler}$ that is used in spatial transformer network (STN) [33] to learn differentiable geometric manipulator function for cropping 2D images for the most salient object in the scene for image recognition. For AT, F_{grid_gen} generates 1D temporal grid with detected start, s, and end, e, indices of core motion signals and the temporal segment, $X[s:e] \in \mathbb{R}^{N \times C \times T'}$ is sampled with $F_{sampler}$, where $T' = e - s + 1$. This cropping operation resembles an interpolation process, which makes the whole AT model differential that can be trained with gradient back-propagation operation.

2.2.3. Adaptive Graph Convolution

The trimmed core motion segment is represented as temporal graphical sequence, $G = (V, E)$, where the node set, $V = \{v_{ti} | t = 1, \cdots, T', i = 1 \cdots\}$, includes markers (joints) in a skeleton sequence. The edge set is composed of two subsets, in which the first edge subset is the intra-skeleton connectivity (limbs) $E_S = \{v_{ti}v_{tj} | (i,j) \in H\}$, where H is the set of connected joints defined by motion capture system, and second edge subset is the inter-frame edges, which connect the same joints in consecutive frames $E_F = \{v_{ti}v_{(t+1)i}\}$.

Given a temporal-spatial graph representation of motion segment, we first encode spatial dimension by using an AGCN [28] with K_v kernel size, which is defined as follows,

$$f_{out} = \sum_{k}^{K_v} W_k f_{in}(A_k + B_k + C_k) \tag{7}$$

where, $f_{in} \in \mathbb{R}^{C_{in} \times T' \times N}$ and $f_{out} \in \mathbb{R}^{C_{out} \times T' \times N}$ are input and output feature map and $W_k \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1}$ is weight vector of the 1×1 convolution operation. $A_k = \Lambda^{-\frac{1}{2}}(\bar{A}_k)\Lambda_k^{-\frac{1}{2}}$ is a normalized $N \times N$ adjacency matrix of defined skeleton structure from our motion

capture system, where $\bar{A}_k$ is a binary $N \times N$ adjacency matrix indicating the connectivity between the joints and $\Lambda_k^{ii} = \sum_j (\bar{A}_k^{ij}) + \alpha$ is the normalized diagonal matrix using $\alpha = 0.001$ to avoid empty rows.

The attention maps of each node (joint) and edge (limb) are encoded in B_k and C_k, which are learned fully data-driven manner. $B_k \in \mathbb{R}^{N \times N}$ is an attention graph that encodes the underlining node and limb importance considering the entire samples of the task domain. B_k is fixed once the parameters are trained and used for the inference. C_k is an input-dependent attention graph to determine the strength of the connection between any two nodes in a given input graph sequence. Specifically, we applied embedded Gaussian Affinity [34] to calculate self-similarity between two nodes, v_i and v_j in a given input feature map, f_{in}.

$$C_k^{ij} = f(v_i, vj) = \frac{e^{\theta_k(v_i)^T \phi_k(v_j)}}{\sum_{j=1}^{N} e^{\theta_k(v_i)^T (\phi_k(v_j))}} \tag{8}$$

Compared to A_k and B_k, C_k can flexibly attend to more important joint and limb motions according to changing inputs at inference time. Combining A_k (predefined skeletal connectivity), B_k (domain-specific connectivity), and C_k (input-specific connectivity) helps the model to fully adjust the graphical structure of the input sample to only focus on the motion signals that are useful for jointly predicting medication state, FOG score, and MDS-UPDRS-III total score excluding FOG score. Additionally, we did not restrict the learned B_k and C_k to be left and right body symmetric to take into account the potential for asymmetric symptoms [35,36].

To further encode temporal dimension, $K_t \times 1$ temporal convolution is applied to spatial feature, f_{out}, extracted from the above mentioned attention-based graph convolution model, thereby, deriving spatial-temporal graphical representation, $f_{out}^{ST} = conv_{K_t \times 1}(f_{out})$. In this study, we use a four-layer temporal-spatial graphical convolutional network (TGCN) with 64 feature maps to encode core motion in the given 4-second analysis window. Temporal Average Pooling (TAP) [37] is applied to the output of the last layer to summarize the feature across the temporal axis.

2.2.4. Multi-task Prediction

The feature representation from the last TGCN layer is used to simultaneously predict medication state, FOG score, and MDS-UPDRS-III total score excluding FOG score. *i)* Medication state is a binary classification problem, either OFF or ON state. The feature representation is processed with two-layer fully connected model and a two-way softmax classifier, which is trained with binary cross-entropy loss. *ii)* FOG score has 5 levels, from 0 (absent) to 4 (severe) FOG. For FOG score prediction, the feature representation is processed with two-layer fully connected model and five-way softmax classifier, which is trained with multi-class cross-entropy loss. *iii)* MDS-UPDRS-III total score excluding FOG score is a positive integer ranging between 0 to 120. Before the model training, we apply Z-score normalization to marginalize the impact of outliers to bias the model prediction behaviors. To additionally consider the non-Gaussian distribution of the MDS-UPDRS-III total score excluding FOG score, we processed the feature representation with Gaussian Mixture Model (GMM) regression model [38].

$$p(y|x) = \sum_{i=1}^{m} \alpha_i(x) \mathcal{N}(\mu_i(x), \sigma_i^2(x)) \tag{9}$$

where $x \in \mathbb{R}^D$ is feature representation from the last TGCN layer, $\alpha = softmax(f_\alpha(x))$ is mixing coefficients for Gaussian distributions and $\mu = f_\mu(x)$ and $\sigma = exp(f_\sigma(x))$ are mean and standard deviation of each Gaussian distribution. For projection functions, $f_\alpha, f_\mu, f_\sigma$, two-layer fully connected models were used. In our experiment, the naive regression with a two-layer fully connected model and mean square error having a single Gaussian distribution assumption did not converge when training.

Table 2. Summary of timed-up-and-go testing sessions stratified by medication state and FOG score.

Medication state	FOG Score 0	1	2	3	4
OFF	21	15	9	8	7
ON	38	11	7	3	1

3. Experiment Setting

3.1. Model Hyperparameter, Training and Evaluation

i) AT: Temporal kernel size and feature map were 3×1 and 64, respectively, for all four layers of the temporal convolutional model, F_{at}. Max pooling with $\times\frac{1}{2}$ was used at each output layer for aggregating temporal dimension. For predicting center location and width size of core motion segment, F^{center} and F^{width}, two-layer fully connected layer model was used with 128 units and ReLU activation function [39]. *ii) AGCN*: Four layers of the temporal graphical convolutional model were used, and kernel sizes of $K_v = 3$ and $K_t = 5$, respectively, for graphical and temporal convolution. Across all layers and convolutions, we used 64 feature maps, ReLU activation function, and max pooling with $\times\frac{1}{2}$ to aggregate along the temporal dimension. *iii) Multi-task Prediction*: For two-layer fully connected models to predict medication state, FOG score, and GMM regression parameters, we used 256 and 128 units with ReLU activation function.

For the training model, we used a learning rate fixed at 1×10^{-3} with Adam optimizer and used a batch size of 16. Model training was stopped when no decrease in loss is observed from the validation set, which model is also used for evaluating the test set.

For evaluating the proposed method, we used 10-fold cross-validation. At each fold, 50%, 20%, and 30% of the dataset was used for the training, validation, and testing sets, respectively. We avoided placing adjacent analysis windows in different folds to avoid pairwise similarity biasing the cross-validation results [40].

3.2. Performance Metrics

For performance metrics, we used binary F1 score and mean F1 score for medication state and FOG score prediction, respectively, which is widely used for evaluating prediction performance in the presence of label imbalance. As shown in Table 2, most participants had FOG scores ≤ 2 for both OFF and ON medication states. The mean F1 score is an average of per-class F1 score, which is the harmonic mean of precision and recall of each class.

$$Precision^c = \frac{TP^c}{TP^c + FP^c} \tag{10}$$

$$Recall^c = \frac{TP^c}{TP^c + FN^c} \tag{11}$$

$$F1\ score^c = 2 \times \frac{Precision^c \times Recall^c}{Precision^c + Recall^c} \tag{12}$$

$$Mean\ F1\ score = \frac{1}{C}\sum_{c}^{C} F1\ score^c \tag{13}$$

where C is the number of classes and $C = 5$ for FOG Score classification. For a class c, TP^c is a true positive that represents the total of successfully classified class windows, FP^c is a false positive that represents the total misclassified class windows, and FN^c is a false negative that represents the total misclassified non-class windows.

For evaluating the regression performance for MDS-UPDRS-III total score excluding FOG score, we used root mean square error (RMSE).

3.3. Comparison with Baseline Models

We compared the proposed model to: *i)* shallow models with hand-crafted features, and *ii)* deep learning models including convolutional networks and graphical convolutional networks. We compared classifier performance across models using 95% Wilson score

confidence intervals [41] for Medication State and FOG Score and using standard normal approximation based 95 % confidence intervals for MDS-UPDRS-III.

Shallow Baseline Models. The first baseline models we considered were shallow models, such as Random Forest (RF) and Support Vector Machine (SVM) with radial basis function (RBF) kernel, with FOG-related hand-crafted features. Following previous work [13,42, 43], we extracted various time, frequency, and distribution features, including freezing index [44], variance, sample entropy [45], central frequency, dominant frequency, and wavelet mean [46] features from the acceleration signals at multiple on-body locations. We used second-order Savitsky-Golay differentiation to derive acceleration traces from joint marker kinematics.

To investigate whether the lateralization of parkinsonian symptoms would impact model performance, We iterated RF and SVM models using markers from the left side of the body only (RF-L, SVM-L) and using markers from both sides (RF-LR, SVM-LR). We focused on lower body parts and independently trained RF and SVM for each task separately, following previous work [13,42].

Deep Baseline Models. We also compared the proposed model to several deep learning models for processing human skeleton time-series, including Temporal convolutional network (TCN) [47], Graphical convolutional network (GCN) [48], GCN with attention model (AGCN) [28], and AGCN with Adaptive Trimming (AT+AGCN). We used identical hyperparameters for model architecture and training wherever possible in order to make the fairest possible comparisons between deep learning models. All models were 4-layer with 64 feature maps and $\times\frac{1}{2}$ max pooling. For deep learning models, and for the classification and regression, we used two-layer fully connected layer with 256 and 128 units with ReLU activation functions.

3.4. Comparison with Single-Task Prediction

Since the proposed model is constrained to learn features relevant to three simultaneous prediction tasks, we reasoned that the identified features might be sub-optimal for single task prediction, leading to decreased performance. Therefore, we re-trained the deep learning models (with the exception of AT+AGCN+GMM) on the FOG score prediction task only and assessed changes in performance. We did not include the AT+AGCN+GMM model in this analysis as without the MDS-UPDRS-III prediction task it is identical to the AT+AGCN model.

3.5. Model Interpretability

We considered it critical to assess the clinical relevance of features identified by the model as relevant to medication state, FOG score, or total MDS-UPDRS-III score. These included individual kinematic markers (often referred to as "joints" in the computer vision literature) and segments ("limbs") with high attention scores, and kinematic marker trajectories with high relevance to particular labels.

To derive overall model attention to individual segments or limbs, we aggregated attention maps across all samples in the dataset by averaging the learned attention maps and graphical structure over all M samples:

$$E_A = \frac{1}{M \times K} \sum_{i}^{M} \sum_{k}^{K_v} (A_k + B_k + C_k^i) \in \mathbb{R}^{N \times N} \tag{14}$$

where A_k, B_k, and $C_k(x_i)$ are the normalized $N \times N$ adjacency matrix of predefined skeleton structure, the domain-wise $N \times N$ attention map, and the input-dependent $N \times N$ attention map at each kernel, respectively. The attention weights for joints and segments are then defined as the diagonal components E_A^{jj} and off-diagonal components $E_A^{ij,j\neq i}$ of E_A, respectively.

Table 3. Prediction performance of the proposed model (AT+AGCN+GMM) and comparison to baseline models. Performance metrics are presented as mean ± 95% confidence interval. Deep learning models are indicated by italics. Abbreviations are described in text. [a]Total score with FOG item (3.11) subtracted. †P<0.05, improvement in RF vs. SVM. ‡P<0.05, improvement in -LR vs. -L. *P<0.05, improvement in deep learning models vs preceding row. §P<0.05, improvement in multi-task vs. single task prediction.

Model	Medication State (F1)	FOG score (F1)	MDS-UPDRS-III[a] (RMSE)
SVM-L	0.540 ± 0.016	0.429 ± 0.026	9.346 ± 0.138
RF-L	$0.594 \pm 0.012^{\dagger}$	$0.553 \pm 0.038^{\dagger}$	9.189 ± 0.301
SVM-LR	$0.616 \pm 0.017^{\ddagger}$	$0.608 \pm 0.031^{\ddagger}$	$8.714 \pm 0.101^{\ddagger}$
RF-LR	$0.657 \pm 0.019^{\dagger,\ddagger}$	$0.684 \pm 0.040^{\dagger,\ddagger}$	$7.918 \pm 0.427^{\dagger,\ddagger}$
TCN [47]	$0.875 \pm 0.017^{*}$	$0.851 \pm 0.020^{*,\S}$	$4.551 \pm 0.276^{*}$
GCN [48]	$0.913 \pm 0.015^{*}$	$0.929 \pm 0.021^{*,\S}$	$4.023 \pm 0.373^{*}$
AGCN [28]	$0.949 \pm 0.010^{*}$	$0.948 \pm 0.018^{*,\S}$	$3.703 \pm 0.300^{*}$
AT+AGCN	0.955 ± 0.021	$0.955 \pm 0.026^{\S}$	$3.555 \pm 0.394^{*}$
AT+AGCN+GMM	0.975 ± 0.018	0.967 ± 0.022	$2.753 \pm 0.440^{*}$

To identify individual kinematic marker trajectories and core motion segments with high relevance to particular labels, we visualized individual analysis windows and core motion segments that the model predicted with high confidence, as measured by the entropy of the class prediction distribution. We visualized these data and discussed the interpretation with clinician experts within our project team, within the movement disorders group at our center, and at a regional forum in the Atlanta area hosted by the study sponsor in order to assess whether the identified features were consistent with the features that movement disorders specialists are trained to identify as characteristics of freezing.

3.6. Model Performance and Potential Bias

After evaluating the proposed model against other candidate models, we assessed the potential for bias in model performance associated with participant demographics. After computing individual F1 score for each participant, we compared model performance across age and sex with linear models. Linear models used FOG study group (PD-FOG, PD-NoFOG, PP-FOG), dichotomized age, and sex as predictors of individual F1 score. Statistical significance was assessed with Wald tests at P=0.05.

4. Results

4.1. Overall Model Perfomance

Here, we report the overall prediction performance of the proposed model (AT+AGCN+GMM in Table 3), compared with the performance of baseline models for predicting medication state, FOG score, and MDS-UPDRS-III total score excluding FOG item. In general, the proposed model's performance was very high for both classification and regression tasks: Medication State, 97.6% cross-validated F1 score; FOG Score, 96.8% cross-validated F1 score; and MDS-UPDRS-III, 2.7 point RMSE, which is within the minimal clinically-important difference [49] for the instrument. In particular, the addition of the GMM regression component — which learns non-Gaussian distributions flexibly — to the second-best performing model architecture (AT+AGCN) significantly improved MDS-UPDRS-III performance. Performance of all models is summarized numerically in Table 3.

The prediction performance of the proposed model on MDS-UPDRS-III score excluding FOG item is shown in Figure 3. The overall RMSE was 2.7 ± 0.4 points. As expected, overall, ON medication sessions have lower and OFF medication sessions have higher MDS-UPDRS-III total scores, as indicated by the higher prevalence of red points to the left of the plot and the higher prevalence of blue points to the right of the plot. We noted

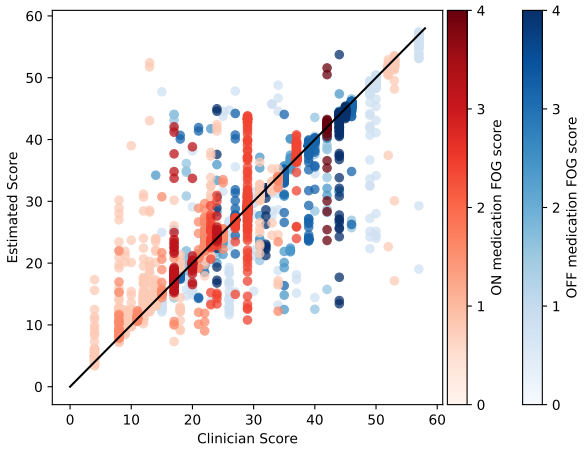

Figure 3. Scatter plot comparing clinician-rated versus model-estimated MDS-UPDRS-III total score, excluding the FOG item (3.11). Unity line is shown for reference. Each dot represents a single 4-second analysis window. Colors are used to represent the FOG item scores corresponding to each analysis window, with darker colors indicating more severe FOG in the OFF (blue) and ON (red) medication states.

that the model tended to overestimate lower scores and under-estimate higher scores, as indicated by datapoints in the upper left and lower right.

4.2. Comparison to Baseline Models

For comparison to the previous state-of-the-art models in FOG analysis, we started analysis with shallow models (RF and SVM) using only lower body parts. We tested the use of left only and both left and right lower body parts (RF-LR and SVM-LR). Using both sides of body significantly improved performance on all three tasks; increasing F1 score by 12% and 33% for Medication State and FOG Score, respectively, and decreasing MDS-UPDRS-III RMSE by 10%. Among the shallow models (RF-LR and SVM-LR), RF significantly outperformed SVM, increasing F1 score by 7% and 13% for Medication State and FOG Score, respectively, and decreasing MDS-UPDRS-III RMSE by 9%, presumably due to its ability to learn non-linear decision boundaries.

Deep learning models also substantially outperformed shallow ML models, providing evidence that learning FOG representations that capture complex patterns may be more effective than using existing hand-crafted FOG features. Compared to the best performing shallow model (RF-LR), the TCN model, which mainly captures temporal patterns of each joint movement sequence, improved F1 score by 33% and 24% for Medication State and FOG Score, respectively, and decreased MDS-UPDRS-III RMSE by 43%.

Among the deep learning models, we also noted significant performance improvements in F1 scores among graph-based models vs. the more traditional TCN, as graph-based models can additionally capture positional relations between joints with a graphical data structure defined as a human skeleton. The simplest graph-based model significantly outperformed the TCN on all three tasks (4%, 9%, and 11% improvements on medication state, FOG score, and MDS-UPDRS-III, respectively). Further significant improvements were noted with the addition of attention mechanisms which enable the model to adaptively concentrate its representation powers for the most relevant joint depending on the given input (4%, 2%, and 8%). The additions of adaptive trimming and the Gaussian mixture model prediction did not significantly improve F1 scores, but significantly reduced MDS-UPDRS-III RMSE (4% and 23%, respectively). We speculate that the flexibility of the GMM model stabilized the gradient backpropagated from the regression branch to help find a more effective feature representation for all tasks.

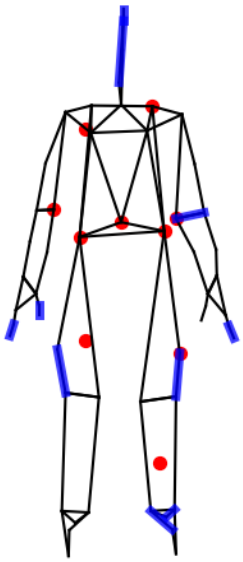

Figure 4. Kinematic markers (referred to as "joints" in pose estimation literature, red) and segments (referred to as "limbs" in pose estimation literature, (blue) with the top 10 attention weights in the prediction tasks.

4.3. Comparison to Single-Task Prediction

All four deep learning models tested showed significantly improved performance on FOG score prediction when trained on the multi-task problem (medication state, FOG score, and MDS-UPDRS-III) rather the single task problem (FOG Score only). When the models were trained on the single task problem, the TCN, GCN, AGCN, AT+AGCN demonstrated F1 scores of 0.825 ± 0.016, 0.892 ± 0.033, 0.903 ± 0.047, and 0.925 ± 0.012, respectively, a 3.8% *decrease* in performance on average compared to the multi-task problem. We speculate that additional information provided to the models by predicting medication states and MDS-UPDRS-III total scores helped to learn representations that are more targeted and personalized to discriminate detailed differences in FOG phenotypes in varying PD conditions, which eventually helped improve overall FOG score classification performance.

4.4. Model Interpretability

4.4.1. Most Relevant Joints and Limbs

We visualized markers and segments with the top ten largest attention weights to assess which body parts were most salient to the prediction task (Figure 4). Attention weights were concentrated on the head, chest, waist, hands, and (particularly left) legs. We suggest that the attention paid to markers on essentially all body segments reflects the fact that FOG is a full-body phenomenon, and suggests that the model may be attending to en-bloc turns [50] — which tend to be maintained across medication states [51] — or other elements of impaired intersegmental coordination. We noted that in particular, the model attended closely to segments on the left foot, which had been suggested previously by a clinical expert on our team as relevant to FOG in this testing condition, which requires left turns. Interestingly, the model also attended to the fingers and elbows. Although these body parts are not typically attended to during clinical FOG examination, patients with FOG can also freeze during upper limb movements [52], leaving open the possibility that the model was attending to hand movements characteristic of freezing.

4.4.2. Most Relevant Motion Segments

We also visualized patterns of left heel movement that were predicted as relevant to particular medication states and FOG scores with high confidence. Figure 5 shows the

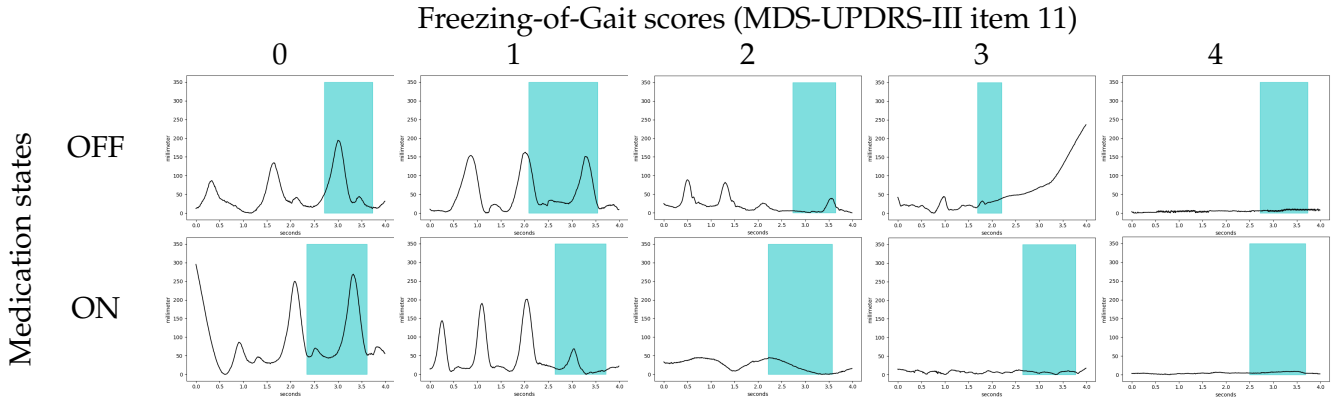

Figure 5. Examples of detected core motion segments of left heel motions from the adaptive trimming model for different medication states and FOG scores. The horizontal and vertical axes of plots are in seconds and millimeters, respectively, with the vertical axis indicating the height of the left heel marker above the laboratory floor. All plots depict four seconds of recorded movement. Core motion segments detected by the adaptive trimming model are depicted in blue.

detected core motion segments of left heel movements from the adaptive trimming model during the timed-up-and-go trials of patients with ON and OFF medication states and different FOG scores. In general, the adaptive trimming model automatically captured approximately a single step cycle within each 4 second analysis window with movement patterns especially related to FOG.

The identified kinematic associated with detected core segments were highly consistent with the features that movement disorders specialists are trained to identify as characteristics of freezing. This analysis shows that the model focuses on periods with regular gait activity for epochs corresponding to FOG scores of 0 and 1, and periods of interrupted gait activity or pure akinesia for epochs with higher scores. For the samples with a FOG score of 0 (first column), the model considered normal stepping gait and used a walking cycle motion for making predictions. For the samples with a FOG score of 1 (second column), the models detected decreasing step length from the motion automatically. As the FOG score became higher, the model tended to detect more FOG-related gait motions. For the samples with a FOG score of 2 (third column), the model detected onset of gait signal related to festination (tendency to speed up in parallel with a loss of normal amplitude of repetitive movements) [53]. For the samples with FOG score of 3 and 4 (fourth and fifth columns), the model detected freezing gait, akinesia, and trembling signals as core motion signal that is relevant for predicting FOG scores.

4.5. Classifier Performance and Potential Bias

After computing individual F1 score for each participant, we compared model performance across study and demographic groups to assess potential bias. Linear models found no significant differences in F1 score across study groups or sex, but found significantly decreased performance (reduction in F1 score of 17%, $P<0.01$) among older participants (age $\geq$ 69 years) compared to younger participants.

5. Discussion and Conclusion

In our experiment, we designed a deep neural network model to simultaneously predict levodopa medication state (ON/OFF), FOG score (0-4), and MDS-UPDRS-III total score (less FOG score) from full-body kinematics data of 57 patients, including 5 patients with atypical parkinsonism, assessed with TUG tests in the off and on medication state. As compared to formal clinical assessments by a movement disorders specialist, our AGCN model classified levodopa medication state and FOG score with 96.4% and 96.2% F1 scores respectively, and regressed MDS-UPDRS-III total score with root mean square error (RMSE) of 2.7 points.

To the best of our knowledge, this is the first work that applies an interpretable deep learning model with full-body kinematics for classifying FOG. This model detects time segments having characteristic movements of FOG during walking, e.g. small shuffling steps, akinesia, and tremulousness. Additional findings demonstrated that FOG is not limited to the lower extremity, and also significantly involves movements in the upper body, further supporting that FOG requires phenotyping using whole-body kinematics. Findings from our analysis may lead to novel hypotheses to define more granular FOG phenotypes, or potentially to technologies that enable continuous monitoring of FOG severity in order to test new therapies with improved precision.

Overall, while the current study uses 3D kinematic data, we believe that the underlying approach will generalize to motion estimates obtained through pose estimation or body-worn sensors, enabling future applications in clinical and home settings with 2D video. The patterns of body motion recorded here result from fundamental principles of physics and biomechanics, which are likely to hold regardless of the method used to measure motion. For example, the laws of motion and principles of energy conservation apply regardless of whether motion is measured using 3D kinematic data, pose estimation, or body-worn sensors. This is likely why it is feasible to estimate virtual IMU signals from video data [29].

The study has three main limitations. First, we did not attempt to identify freezing of gait (FOG) at the millisecond level, which would be necessary for use in assistive technology. Second, we did not attempt to measure FOG severity as a continuous outcome, which could increase precision in clinical trials. Finally, the study sample was predominantly white and had fewer females than would be representative of the Parkinson's disease (PD) population [54], so the generalizability of the results to the entire PD population may be limited.

One primary contribution of this work is the application of deep learning to the problem of scoring FOG, which has primarily been examined with hand-crafted and engineered features such as spectral power in a prespecified "freeze band" [16] calculated from a prespecified set of body segments. Indeed, despite the typical notion that FOG is an interruption of walking — leg movements — our results indicate that scoring FOG with high accuracy may require attention to body parts across the body, including the hands and head. We believe that adopting a data-driven approach with explainable deep learning models represents an important way forward in modeling kinematics from walking and turning motions of parkinsonian patients.

We hope that using deep learning to discover data-driven kinematic features will lead to the development of a more fine-grained and objective FOG severity scales, which could provide valuable information to clinicians and researchers, help to improve diagnosis, treatment, and overall management of FOG (cf. [55]).

Author Contributions: Conceptualization, H.K., S.F., and L.M.; methodology, H.K.; software, H.K.; validation, H.K., I.G., and L.M.; formal analysis, H.K. and L.M.; investigation, H.K. and L.M.; resources, S.F., L.M., G.D., and C.E.; data curation, D.B. and L.M.; writing—original draft preparation, H.K.; writing—review and editing, L.M., G.D., I.G., C.E., S.F.; visualization, H.K.; supervision, G.D., L.M., and S.F.; project administration, L.M. and S.F.; funding acquisition, S.F. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The McCamish Center for Parkinson's Disease Innovation, the by Curtis Family Fund, Sartain Lanier Family Foundation (S.A.F.), and NIH K25HD086276 (J.L.M.).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Emory University (IRB Number 2688 approved on 2 June 2021 and IRB Number 73518 approved on 10 December 2014)."

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study according to procedures approved by the Emory University Institutional Review Board.

Data Availability Statement: The deidentified raw data and code for supporting the evidences in this work will be made available by the corresponding author upon reasonable request.

Acknowledgments: The computational needs for this research are supported in part by Oracle Cloud credits and related resources provided by the Oracle for Research program.

Conflicts of Interest: J.L.M. performs paid consulting work for Biocircuit technologies. None of these interests are directly related to the outcomes of this study. S.A.F. has the following competing interests: Honoraria: Lundbeck, Teva, Sunovion, Biogen, Acadia, Neuroderm, Acorda, CereSpire. Grants: Ipsen, Medtronics, Boston Scientific, Teva, US World Meds, Sunovion Therapeutics, Vaccinex, Voyager, Jazz Pharmaceuticals, Lilly, CHDI Foundation, Michael J. Fox Foundation, NIH, Royalties: Demos, Blackwell Futura for textbooks, Uptodate, Other Bracket Global LLC, CNS Ratings LLC. None of these interests are directly related to the outcomes of this study. The other authors declare no other interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A Wilson Score Interval

Binomial proportion confidence interval calculates the outcome of series of Bernoulli trials to estimate the confidence interval for the probability of success. Wilson score interval is a asymmetric approximation of binomial confidence interval, which tackles two problems that rises when using naive symmetric normal approximated confidence interval [56], which are overshoot and zero width intervals [41]. Moreover, wilson score interval is robust with small samples and skewed observations as in our dataset (Table 2), which is common in human behavior analysis problems [29,57]. Wilson score interval can be calculated as follows:

$$p \approx \frac{1}{1+\frac{z^2}{n}}\left(\hat{p}+\frac{z^2}{2n}\right) \pm \frac{z}{1+\frac{z^2}{n}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}} \tag{A1}$$

where $\hat{p}$ is the success probability and n is the number of experiments. For 95% confidence interval, $z = 1.96$.

References

1. Pringsheim, T.; Jette, N.; Frolkis, A.; Steeves, T. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Movement disorders* **2014**, *29*, 1583–1590.
2. Dorsey, E.; Sherer, T.; Okun, M.S.; Bloem, B.R. The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's disease* **2018**, *8*, S3–S8.
3. Lucas McKay, J.; Goldstein, F.C.; Sommerfeld, B.; Bernhard, D.; Perez Parra, S.; Factor, S.A. Freezing of Gait can persist after an acute levodopa challenge in Parkinson's disease. *NPJ Parkinson's disease* **2019**, *5*, 1–8.
4. Nonnekes, J.; Snijders, A.H.; Nutt, J.G.; Deuschl, G.; Giladi, N.; Bloem, B.R. Freezing of gait: a practical approach to management. *The Lancet Neurology* **2015**, *14*, 768–778.
5. Factor, S.A.; Jennings, D.L.; Molho, E.S.; Marek, K.L. The natural history of the syndrome of primary progressive freezing gait. *Archives of neurology* **2002**, *59*, 1778–1783.
6. Haddad, Y.K.; Bergen, G.; Florence, C. Estimating the economic burden related to older adult falls by state. *Journal of public health management and practice: JPHMP* **2019**, *25*, E17.
7. Florence, C.S.; Bergen, G.; Atherly, A.; Burns, E.; Stevens, J.; Drake, C. Medical costs of fatal and nonfatal falls in older adults. *Journal of the American Geriatrics Society* **2018**, *66*, 693–698.
8. Pelicioni, P.H.; Menant, J.C.; Latt, M.D.; Lord, S.R. Falls in Parkinson's disease subtypes: risk factors, locations and circumstances. *International journal of environmental research and public health* **2019**, *16*, 2216.
9. et al., G. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society* **2008**, *23*, 2129–2170.
10. Hulzinga, F.; Nieuwboer, A.; Dijkstra, B.W.; Mancini, M.; Strouwen, C.; Bloem, B.R.; Ginis, P. The New Freezing of Gait Questionnaire: Unsuitable as an outcome in clinical trials? *Mov. Disord. Clin. Pract.* **2020**, *7*, 199–205.
11. Hatcher-Martin, J.; McKay, J.; Sommerfeld, B.; Howell, J.; Goldstein, F.; Hu, W.; Factor, S. Cerebrospinal fluid Aβ42 and fractalkine are associated with Parkinson's disease with freezing of gait. *medRxiv* **2020**.
12. Forsaa, E.; Larsen, J.; Wentzel-Larsen, T.; Alves, G. A 12-year population-based study of freezing of gait in Parkinson's disease. *Parkinsonism & related disorders* **2015**, *21*, 254–258.
13. Zhang, Y.; Yan, W.; Yao, Y.; Bint Ahmed, J.; Tan, Y.; Gu, D. Prediction of freezing of gait in patients with Parkinson's disease by identifying impaired gait patterns. *IEEE transactions on neural systems and rehabilitation engineering* **2020**, *28*, 591–600.
14. Yan, Y.; Liu, Y.; Li, C.; Wang, J.; Ma, L.; Xiong, J.; Zhao, X.; Wang, L. Topological Descriptors of Gait Nonlinear Dynamics toward Freezing-of-Gait Episodes Recognition in Parkinson's Disease. *IEEE Sensors Journal* **2022**.

15. Silva de Lima, A.; Evers, L.J.; Hahn, T.; Bataille, L.; Hamilton, J.; Little, M.; Okuma, Y.; Bloem, B.; Faber, M. Freezing of gait and fall detection in Parkinson's disease using wearable sensors: a systematic review. *Journal of neurology* **2017**, *264*, 1642–1654.
16. Yungher, D.; Morris, T.; Dilda, V.; Shine, J.; Naismith, S.; Lewis, S.J.; Moore, S.T. Temporal characteristics of high-frequency lower-limb oscillation during freezing of gait in Parkinson's disease. *Parkinson's Disease* **2014**, *2014*.
17. et al., R.M. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PloS one* **2017**, *12*, e0171764.
18. Tahafchi, P.; Molina, R.; Roper, J.A.; Sowalsky, K.; Hass, C.J.; Gunduz, A.; Okun, M.S.; Judy, J.W. Freezing-of-Gait detection using temporal, spatial, and physiological features with a support-vector-machine classifier. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017, pp. 2867–2870.
19. Reches, T.; Dagan, M.; Herman, T.; Gazit, E.; Gouskova, N.A.; Giladi, N.; Manor, B.; Hausdorff, J.M. Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. *Sensors* **2020**, *20*, 4474.
20. Ferrari, A.; Ginis, P.; Hardegger, M.; Casamassima, F.; Rocchi, L.; Chiari, L. A mobile Kalman-filter based solution for the real-time estimation of spatio-temporal gait parameters. *IEEE transactions on neural systems and rehabilitation engineering* **2015**, *24*, 764–773.
21. Diep, C.; O'Day, J.; Kehnemouyi, Y.; Burnett, G.; Bronte-Stewart, H. Gait Parameters Measured from Wearable Sensors Reliably Detect Freezing of Gait in a Stepping in Place Task. *Sensors* **2021**, *21*, 2661.
22. Mancini, M.; Shah, V.V.; Stuart, S.; Curtze, C.; Horak, F.B.; Safarpour, D.; Nutt, J.G. Measuring freezing of gait during daily-life: an open-source, wearable sensors approach. *Journal of NeuroEngineering and Rehabilitation* **2021**, *18*, 1–13.
23. Hatcher-Martin, J.M.; McKay, J.L.; Pybus, A.F.; Sommerfeld, B.; Howell, J.C.; Goldstein, F.C.; Wood, L.; Hu, W.T.; Factor, S.A. Cerebrospinal fluid biomarkers in Parkinson's disease with freezing of gait: an exploratory analysis. *NPJ Parkinsons Dis.* **2021**, *7*, 105.
24. Hughes, A.J.; Daniel, S.E.; Kilford, L.; Lees, A.J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *Journal of neurology, neurosurgery & psychiatry* **1992**, *55*, 181–184.
25. et al., N. Using the Timed Up & Go test in a clinical setting to predict falling in Parkinson's disease. *Archives of physical medicine and rehabilitation* **2013**, *94*, 1300–1305.
26. Shumway-Cook, A.; Baldwin, M.; Polissar, N.L.; Gruber, W. Predicting the probability for falls in community-dwelling older adults. *Physical therapy* **1997**, *77*, 812–819.
27. Kadaba, M.P.; Ramakrishnan, H.; Wootten, M. Measurement of lower extremity kinematics during level walking. *Journal of orthopaedic research* **1990**, *8*, 383–392.
28. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12026–12035.
29. Kwon, H.; Abowd, G.; Plötz, T. Complex Deep Neural Networks from Large Scale Virtual IMU Data for Effective Human Activity Recognition Using Wearables. *Sensors* **2021**, *21*, 8337.
30. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115.
31. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In Proceedings of the IJCAI, 2016.
32. Bachlin, M.; Plotnik, M.; Roggen, D.; Maidan, I.; Hausdorff, J.M.; Giladi, N.; Troster, G. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine* **2009**, *14*, 436–446.
33. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Advances in neural information processing systems* **2015**, *28*.
34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
35. Toth, C.; Rajput, M.; Rajput, A.H. Anomalies of asymmetry of clinical signs in parkinsonism. *Movement disorders: official journal of the Movement Disorder Society* **2004**, *19*, 151–157.
36. Djaldetti, R.; Ziv, I.; Melamed, E. The mystery of motor asymmetry in Parkinson's disease. *The Lancet Neurology* **2006**, *5*, 796–802.
37. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv preprint arXiv:1312.4400* **2013**.
38. Bishop, C.M. Mixture density networks **1994**.
39. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the Icml, 2010.
40. Hammerla, N.Y.; Plötz, T. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In Proceedings of the Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, 2015, pp. 1041–1051.
41. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **1927**, *22*, 209–212.
42. Naghavi, N.; Wade, E. Prediction of freezing of gait in Parkinson's disease using statistical inference and lower–limb acceleration data. *IEEE transactions on neural systems and rehabilitation engineering* **2019**, *27*, 947–955.
43. Palmerini, L.; Rocchi, L.; Mazilu, S.; Gazit, E.; Hausdorff, J.M.; Chiari, L. Identification of characteristic motor patterns preceding freezing of gait in Parkinson's disease using wearable sensors. *Frontiers in neurology* **2017**, *8*, 394.
44. Moore, S.T.; MacDougall, H.G.; Ondo, W.G. Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of neuroscience methods* **2008**, *167*, 340–348.

45. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* **2000**, *278*, H2039–H2049.
46. Baby, M.S.; Saji, A.; Kumar, C.S. Parkinsons disease classification using wavelet transform based feature extraction of gait data. In Proceedings of the 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT). IEEE, 2017, pp. 1–6.
47. Du, Y.; Fu, Y.; Wang, L. Skeleton based action recognition with convolutional neural network. In Proceedings of the 2015 3rd IAPR Asian conference on pattern recognition (ACPR). IEEE, 2015, pp. 579–583.
48. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-second AAAI conference on artificial intelligence, 2018.
49. Horváth, K.; Aschermann, Z.; Ács, P.; Deli, G.; Janszky, J.; Komoly, S.; Balázs, É.; Takács, K.; Karádi, K.; Kovács, N. Minimal clinically important difference on the Motor Examination part of MDS-UPDRS. *Parkinsonism & related disorders* **2015**, *21*, 1421–1426.
50. Khobkhun, F.; Hollands, M.; Tretriluxana, J.; Srivanitchapoom, P.; Richards, J.; Ajjimaporn, A. Benefits of task-specific movement program on en bloc turning in Parkinson's disease: A randomized controlled trial. *Physiother. Res. Int.* **2022**, *27*, e1963.
51. Hong, M.; Earhart, G.M. Effects of medication on turning deficits in individuals with Parkinson's disease. *J. Neurol. Phys. Ther.* **2010**, *34*, 11–16.
52. Heremans, E.; Nackaerts, E.; Vervoort, G.; Vercruysse, S.; Broeder, S.; Strouwen, C.; Swinnen, S.P.; Nieuwboer, A. Amplitude manipulation evokes upper limb freezing during handwriting in patients with Parkinson's disease with freezing of gait. *PLoS One* **2015**, *10*, e0142874.
53. Imai, H. Festination and freezing. *Rinsho Shinkeigaku= Clinical Neurology* **1993**, *33*, 1307–1309.
54. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **2018**, *17*, 939–953.
55. Hulzinga, F.; Nieuwboer, A.; Dijkstra, B.; Mancini, M.; Strouwen, C.; Bloem, B.; Ginis, P. The new freezing of gait questionnaire: unsuitable as an outcome in clinical trials? *Movement disorders clinical practice* **2020**, *7*, 199–205.
56. Wallis, S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* **2013**, *20*, 178–208.
57. Murahari, V.S.; Plötz, T. On attention models for human activity recognition. In Proceedings of the Proceedings of the 2018 ACM international symposium on wearable computers, 2018, pp. 100–103.