Subtype-Specific Freezing of Gait Detection Strategies via Feature-Masked-Based CNN

Xinyue Yu, Kaylena Ehgoetz Martens, and Arash Arami, *Member, IEEE*

***Abstract*— Freezing of gait (FOG), a disabling symptom of Parkinson's disease, varies in presentations and motion context. Its heterogeneity motivates subtype categorization such as manifestation-specific subtypes (akinesia, trembling or shuffling) or motion-specific subtypes (gait-initiation, walking or turning). Despite numerous deep learning FOG detection models, few consider FOG heterogeneity. Thus, it remains unclear whether subtype-specific models could enhance detection generalizability. This paper clusters FOG data into manifestation- or motion-specific subtypes and encodes their detection strategies in feature masks. It is found that while motion-specific subtypes share a common detection strategy, manifestation-specific subtypes require distinct strategies. A feature-mask-based CNN is proposed to train general and subtype-specific models using waist-mounted 3D accelerometer data. Manifestation models exhibit enhanced generalizability across subtypes over a general model, boosting the overall average FOG sensitivity by 24.95% ± 9.80 and non-FOG specificity by 18.29% ± 8.71. Differently, motion models show no such improvement. This suggests that detection strategy is driven by manifestation composition of data. The general model favors the dominant manifestation-specific subtype(s), a bias corrected by tailored manifestation-specific strategy. No comparable benefit arises from motion models due to their similar manifestation compositions.**

***Index Terms*— Freezing of gait (FOG) Detection, FOG Subtypes, Convolutional Neural Network (CNN), Feature(Time-Domain, Frequency-Domain) Masking.**

I. Introduction

PARKINSON'S disease (PD) is a neurodegenerative disorder that leads to various motor and non-motor deficits [1]. Freezing of Gait (FOG) is a common motor symptom typically in people with advanced PD [2], though possible early on [3]. This phenomenon is defined as a temporary, episodic lack or deterioration of forward movement, even with the intention to walk [4]. Typical pathophysiologies of FOG exhibit disturbed temporal and frequency patterns, as evidenced by leg trembling in 4 to 5 Hz [5], impaired gait timing regulation [6], and asymmetrical and uncoordinated gait [7]. A vivid depiction of this debilitating episode by patients is 'feeling as though their feet are glued to the floor' [8]. Beyond its impact on quality of daily life, FOG is the leading cause of falls in patients with PD [9] .

Given its debilitating impact, FOG detection is actively studied. Robust detection algorithms can aid both understating and treatment of FOG. Since FOG relates to PD progression [2] and is harder to provoke in clinical setting [10], free-living FOG monitoring by detection algorithms embedded in wearables offer valuable insights for patients, clinicians and researchers. Automatic FOG detection also reduces the burden of manual video labeling by expert clinicians [11] in standard assessments. In addition, it enables real-time and on-demand activation of nonpharmaceutical FOG treatments, such as cueing via vibration or audiovisual stimuli [12] [13], or control of assistive devices such as lower limb exoskeletons [14]. For real-time implementation of FOG detection algorithm, sensors that capture kinematics are the most popular wearable choices. According to a recent review [15], 72% of the related studies adopt either inertial measurement unit (IMU) (consists of accelerometers and gyroscope) or solely accelerometers. Therefore, this study will focus on kinematics-based detection algorithms for its significant application potentials.

The freeze index (FI) is one of the important kinematic features used for FOG detection, defined as the ratio of spectral power in the freeze band (3–8 Hz) to the locomotor band (0.5–3 Hz) of vertical leg acceleration [16]. The empirical threshold of FI is an early and widely-used FOG detection method, but may misclassify voluntary stops as FOG due to similar spectral shifts that increase FI by decreasing locomotor band power [17]. Machine learning models such as support vector machines (SVM) [18] and decision trees [19] demonstrate

X. Yu is with the Department of Systems and Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada(e-mail: thea.yu@uwaterloo.ca).

K. Martens is with the Department of Kinesiology and Health Sciences, University of Waterloo, Waterloo, ON N2L 3G1, Canada(e-mail:kaylena.ehgoetz.martens@uwaterloo.ca).

A. Arami is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. He is also with the Kite Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Canada(e-mail:aarami@uwaterloo.ca).

Manuscript received April 25, 2025; revised XX Month, 2025.

enhanced detection performance by incorporating diverse features [20]. However, those features rely on hand-engineered extraction based on prior knowledge and thus may overlook some useful though less interpretable data-driven patterns. Recently, popular deep learning methods such as feedforward probabilistic neural network [21] , [22] and recurrent neural network [23] that well capture temporal changes have been applied to FOG detection problem. They have shown impressive detection accuracy with their ability to automatically capture complicated and various data features. However, their success depends on large, diverse training datasets [24], which are challenging to collect due to FOG's episodic nature.

FOG episodes vary in their responsiveness to dopaminergic medication, manifestation, and context, leading to the definition of subtypes within each aspect. Pharmacological subtypes categorize FOG as either responsive or unresponsive to levodopa [25], a common dopamine therapy that serves as the cornerstone of PD management [1]. Based on manifestations, FOG has been categorized into gait [26] or phenomenological [27] subtypes, that include shuffling-forward, trembling in-place, and akinesia (rarest [26]). Contextually, FOG stems from internal (e.g., affective state, cognitive load) and external factors. For instance, FOG may occur with or without heightened emotional arousal, as found in an investigation 48% of 29 freezers reported panic attacks during FOG [28]. Similarly, FOG may occur with or without added cognitive load (e.g., a dual task color classification during 360° turning) [29]. As an example of environmental-driven external factor, doorway width is inversely related to frequency of FOG episodes [30]. As an example of interaction-driven external factor, motion context also influences FOG, with its episodes observed during gait initiation [31], walking [32] [24], and turning [33] (most frequent [26]).

Despite promising results from machine learning and deep learning methods in detecting FOG using kinematics data [21] [34] [35] [36] [37] [38], few approaches account for FOG heterogeneity in their design. Mo and Chan trained a transformer to classify non-FOG, FOG at walking, FOG at turning, or FOG at gait initiation [39]. Yang et al. evaluated different deep models to differentiate trembling-in-place FOG, akinesia FOG, and non-FOG using lower-body IMU records (shuffling treated as non-FOG events) [40]. These studies show that FOG under different motion contexts or with different manifestations are distinguishable, supporting the feasibility of subtype-specific detection. Shaban et al. trained separate similar Convolutional Neural Networks (CNN), focusing on FOG at walking, turning or gait initiation, and found the turning CNN performed worst, suggesting greater complexity in the patterns of FOG at turning which could not be well modeled by similar CNN structure [41]. Since there is no comparison between those subtype-specific models and the general model, it cannot be further inferred whether that subtype training has led to any enhancement from the general approach.

Considering FOG heterogeneity is essential for designing effective detection algorithms, as subtype importance varies between patients. For instance, some may find akinesia the most disruptive for its mobility impact, while others may be more concerned with frequent trembling episodes. Without evaluating and maintaining detection performance across different subtypes, detection algorithms may fail to generalize to diverse patient needs. Addtionally, deep learning models trained on imbalanced data are especially prone to bias, overfitting to common FOG subtypes while neglecting rare ones such as akinesia. A balanced approach that detects all subtypes reliably can offer unbiased support for future research including the investigations on whether different FOG subtypes are related to different pathologies [27].

This study focused on motion context and manifestations—FOG heterogeneity directly linked to kinematics—when designing a kinematics-based detection algorithm. For motion context, motion-specific subtypes are defined as gait initiation, walking, and turning subtypes. For FOG manifestation, manifestation-specific subtypes are defined as shuffling, trembling, and akinesia subtype. Viewing the heterogeneity according to either subtype definition, this study investigates whether each subtype requires a unique detection strategy (i.e., specific features) and whether subtype-tailored detection models can enhance detection generalizability across subtypes compared to a general model.

II. Materials and Methods

We used tDCS FOG dataset [42], consisting of 833 trials collected from 62 PD patients performing Ziegler FOG-provoking protocol [43]. Each trial began with quite standing, followed by a 1 m walk to a test spot for two 360° in-place turns (clockwise and counter-clockwise), then a 2 m walk through an open doorway and back. The 3D acceleration of the participants' lower back was acquired at 128 Hz along with expert-annotated labels of non-FOG and FOG events of turning, walking, or start-hesitation (motion-specific subtypes).

After preprocessing (Sec. II-A) and feature extraction (Sec. II-B), we split the data using subtype-specific clustering (Sec .II-C). Then, we obtained feature masks representing desired detection strategies for the general (i.e. all FOG VS all non-FOG) and subtype-specific (i.e. a subtype VS all non-FOG or a subtype VS a non-FOG subset most easily mistaken for that subtype) detection purposes (Sec II-D). Lastly, CNNs embedding those feature masks were trained. The detection performances

between the general CNN (general model) and subtype-specific CNNs (subtype models) were compared.

A. Data Preprocessing

The raw data were first segmented into 4s sliding windows with 0.25s overlap. A relatively large window size was chosen to catch gait changes while a small overlap was used to capture short FOG episodes. 3D (acc_{total}) as well as 2D acceleration norms in the sagittal (acc_{sag}), transverse (acc_{trans}), and frontal (acc_{front}) planes were used to reduce the potential influence from sensor placement variations and provide more targeted insights. The non-FOG and FOG label as well as motion-specific subtype label of a window were determined using Algorithm 1, a flexible labeling method aims for faster detection than classical majority voting.

Algorithm 1 Dominant-Frequency-Based Labeling

1) Define multiple frequency ranges:
   - $f_range_1 = [0.5, 0.75]$ Hz
   - $f_range_2 = (0.75, 1.25]$ Hz
   - $f_range_3 = (1.25, 2.25]$ Hz
   - $f_range_4 = (2.25, 4.25]$ Hz
2) Obtain the dominant frequencies of a window:
   - $f_1 = \arg\max(\text{FFT}(\text{acc}_{sagittal}))$
   - $f_2 = \arg\max(\text{FFT}(\text{acc}_{transverse}))$
   - $f_3 = \arg\max(\text{FFT}(\text{acc}_{frontal}))$
3) Determine the label span of the window:
   - If the majority of $\{f_1, f_2, f_3\}$ falls in one f_range, or at least one of $\{f_1, f_2, f_3\}$ falls in $[0.5, 4.25]$ Hz and $\min(f_1, f_2, f_3)$ falls in one f_range, then:
     - $label_span = \text{round}\left(\frac{2}{f_range_{\text{lower bound}}}\right) s$
   - Otherwise:
     - $label_span = 0.5$ s.
4) FOG or motion-specific subtype label = majority label in its last $label_span$ samples of the window.

Lastly, the dataset was split into 10 trial-independent folds using Algorithm 2, ensuring inter-fold consistency, allowed every fold to contain both FOG and non-FOG samples with diverse and representative characteristics. For 10-fold cross-validation used later in model training, a 80%-10%-10% train-valid-test split was achieved by each time setting 8 folds as training set, one fold as validation set, and one fold as test set. The roles of the folds rotate sequentially across iterations.

B. Feature Extractions

16 time-domain (TD) and 6 frequency-domain (FD) features commonly used for FOG detection algorithms [22] [44] (Table I) were computed for each window, separately for each of the four acceleration norms. Since majority of the spectral energy of human motion is below 15 Hz [45], the FD features were obtained from the 0.5-15 Hz frequency range of FFT spectrum with a Hanning window. Instead of the four acceleration norms, these extracted TD and FD features were used in the following analysis and model training.

Algorithm 2 Balanced Fold Assignment

1) Sort trials by sample count (descending).
2) For each trial, iterate through each fold:
   - Place the trial in the fold.
   - Calculate for every fold:
     - $n_samples$: Total sample count.
     - $short_fog$, $long_fog$: Sample count in short (<4s) / long FOG episodes.
     - gi_fog, t_fog, w_fog: Sample count in gait-initiation, turning, walking FOG.
     - p_nf: Sample count in trials with no FOG.
   - Compute the sum of standard deviations of $n_samples$, $short_fog$, $long_fog$, gi_fog, t_fog, w_fog, and p_nf across 10 folds.
3) Assign the trial to the fold that results in the smallest sum of standard deviations.
4) Repeat until all trails are assigned.

C. Subtype-specific Clustering

While the dataset included motion-specific labels for FOG, it lacked such labels for non-FOG and manifestation-specific labels entirely. To enable subtype-specific modeling, clustering was first used to assign each FOG window to a manifestation- or motion-specific subtype, though original motion-specific labels from the dataset were used for training. The same clustering algorithms were then applied to non-FOG windows, assigning each to a subtype it most closely resembled, i.e., the subtype in which it was most likely misclassified as FOG. The inclusion of both FOG and non-FOG data in each subtype cluster enables subtype-specific classifier development. Example FOG and non-FOG samples of each subtype cluster are shown in Supplementary Material. Clustering methods, detailed below, were tuned using only the training set.

1) Manifestation-specific Subtypes: Akinesia, characterized by a complete lack of movement [26], was distinguished from the other two motion-heavy subtypes using spectral entropy, which measures the complexity and orderliness of a signal. To further differentiate between trembling and shuffling cases, we utilized FI, which quantifies the ratio of power in shuffling-focused locomotion band (0.5 to 3 Hz) over the power in trembling-

TABLE I
SUMMARY OF TIME-DOMAIN AND FREQUENCY-DOMAIN FEATURES

Time Domain (TD)	Description
Mean	Average value
Median (Med)	Middle value of the data when sorted in ascending order
Trimmed median (TremMed)	Median value after removing 5% of the lowest and highest data
Standard deviation (Std)	Variation from the average
Min, Max	Minimum and maximum values
Range	Difference between the maximum and minimum values
Root mean square (RMS)	Square root of the average of the squared values
Mean absolute deviation (MAD)	Average of the absolute differences from the mean value
25th percentile (P25), 75th percentile (P75)	The value below which 25% or 75% of the data fall into
Interquartile Range (IQR)	Difference between P75 and P25
Skewness (Skew)	Asymmetry of the data distribution
Kurtosis (Kurt)	Sharpness of the peak, with high Kurtosis corresponding to more outliers.
Normalized Signal Magnitude Area (NormSMA)	Sum of the absolute values, normalized by window length
Mean Crossing Rate (MCR)	Number of times that the data crosses its average
Frequency Domain (FD)	**Description**
Spectral Entropy (SpecEnt)	Randomness and complexity of the frequency components of the data
Locomotion Band Power (LBPr)	Energy in 0.5 to 3 Hz frequency range
Freeze Band Power (FBPr)	Energy in 3 to 8 Hz frequency range
Freezing Index (FI)	Ratio of FBPr over LBPr
Peak Frequency (PeakFreq)	Frequency associated with the maximum power spectrum value
Average Energy (AvgEng)	Mean value of the energy of a window

focused freezing band (3 to 8 Hz) [16]. Although FI was originally defined for vertical acceleration, extending FI to other axes has also been found useful in representing FOG characteristics [21]. We employed Expectation Maximization to fit two Gaussian Mixture Models (GMM) separately on the average spectral entropy and FI across four acceleration norms. Data windows were first clustered with the entropy-based GMM: those in the low-entropy cluster are labeled as akinesia. The remaining are further clustered using the FI-based GMM, with low-FI windows labeled as shuffling, and the rest as trembling.

2) Motion-specific Subtypes: A lightweight and interpretable Extreme Gradient Boosting (XGBoost) model was trained, using the dataset labels, to classify each FOG window into a motion-specific subtype. The top features relied by the XGBoost were found to be RMS and TremMed of transverse-plane acceleration. The XGBoost was then applied to the non-FOG data to assign them motion-specific labels.

D. Feature Mask and Feature-mask-based CNN

1) Effect-size-based Feature Mask: To mitigate sample size effects, effect size, calculated as the mean difference between groups divided by the standard deviation across all data, was used instead of p-value to evaluate feature differences between FOG and non-FOG groups [46]. A 4 by 22 feature mask is obtained by min-max normalizing an effect size matrix, where each element corresponds to a Table I feature. The feature mask is later embedded in a detection model as a weighted matrix applied to the input features, with a higher importance assigned to a feature differing more between FOG and non-FOG groups. Therefore, a feature mask is also a direct visualization of the detection strategy of a model.

All feature masks were generated on the training set. For the general model, a general feature mask was obtained using all FOG and all non-FOG data. For each subtype model, a subtype-specific feature mask was obtained with data of the corresponding subtype cluster. For analysis, feature masks were also generated for each FOG subtype and all non-FOG. Spearman's rank and Pearson correlations were used to assess feature ranking consistency and relative feature importance, respectively.

2) Model Structure: Fig. 1 shows the structure of the feature-mask-based CNN. After being multiplied to the corresponding feature mask, the extracted TD and FD features fed to three convolutional layers to catch relations across features and across acceleration norms. Batch normalization, residual connection, and dropout layer between each convolutional layer were used to stabilize training and avoid overfitting. The outputs of last convolutional layer was flattened and fed to two fully connected layers, with the last layer providing the probability of FOG and non-FOG (binary classification).

3) Model Training and Evaluation: A learning rate of 0.001, cross entropy loss, and Adam optimizer were used. Down-sampling was applied to the majority class on training and validation sets to address class imbalance. Early stopping was used to avoid overfitting, with a patience (the number of epochs to continue training after validation loss stops improving) of 20 for general model and 40 for subtype-model training (due to smaller dataset size). The general model was trained on all FOG and non-FOG data, while subtype models used only the corresponding FOG subtype and the same-cluster non-FOG samples from the training and validation sets.

For evaluation, test set was split into subtype-specific test sets (subtype test set), each containing a FOG subtype and its same-cluster non-FOG. Each subtype model was tested on its corresponding subtype test set, while the general model was evaluated on all subtype test sets. With FOG as the positive class, true positive rate (TPR), i.e., sensitivity to FOG, and true negative rate (TNR), i.e., specificity to non-FOG, were evaluated

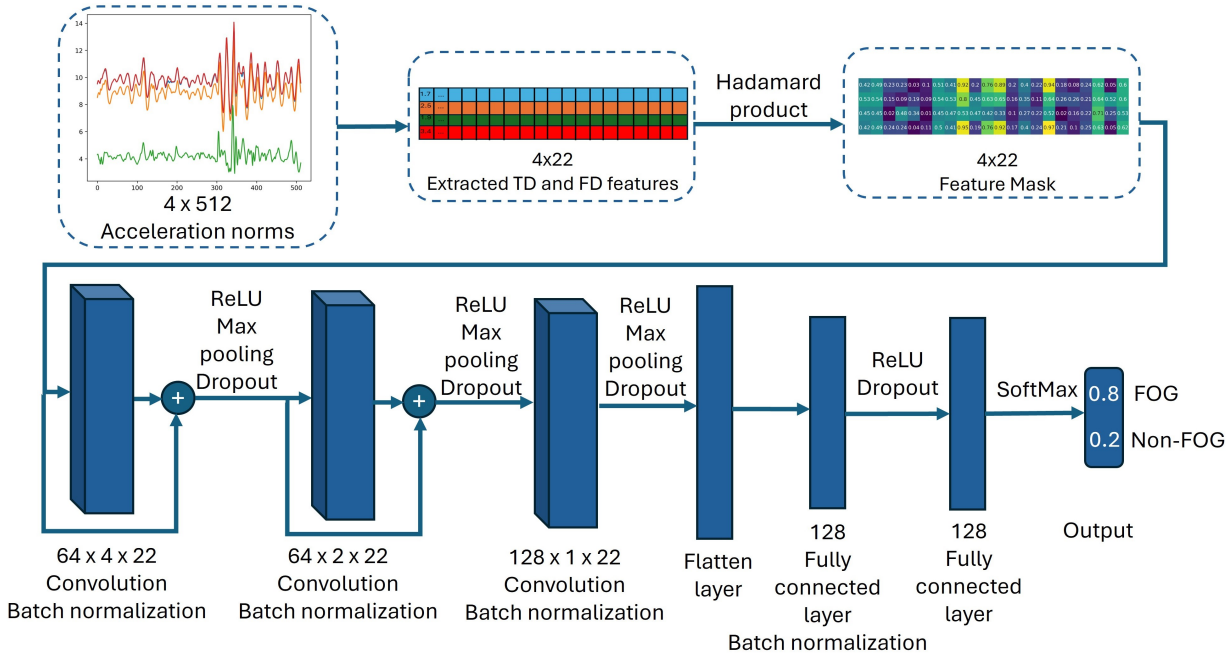

Fig. 1. Feature-mask-based CNN structure.

in 10-fold cross validation. Wilcoxon signed-rank test assessed the statistical significance of performance differences between models with a Bonferroni correction ($\alpha = \frac{0.05}{3} \approx 0.0167$). Separately for manifestation- and motion-specific test sets, an Aligned Rank Transform Analysis of Variance (ART ANOVA) test (Table II) was performed on TPRs and TNRs with respect to the influence of model and subtype test sets to determine whether further comparisons within test sets are warranted.

III. Results

A. Feature Mask Visualization

The average feature masks over 10 folds are visualized in Fig. 2, 3, and Supplementary Material. All showed low variance indicated by the within or below moderate (0.2-0.3) coefficient of variation (CV) of all features of notable significance (normalized effect size > 0.1).

Manifestation feature masks in Fig. 2 (a) show distinct pattern from each other and from the general feature mask in Fig. 3. Specifically, while the general feature mask attributes higher importance to FD features, the akinesia-specific feature mask values TD features more than FD features. The shuffling-specific feature mask exhibits a more uniform feature importance distribution across TD and FD. Trembling-specific feature mask pays high attention to TD features related to the transverse plane. In contrast, motion-specific feature masks in Fig. 2 (b) present a similar distribution to each other and the general feature mask. These feature masks rely heavily on FD features, with similar ranked preferences in peak frequency, FI, locomotion-band power, and spectral entropy as the top features. The only noticeable, albeit small, difference is the higher reliance on mean crossing rate by the walking-specific feature mask. These observations hold when extending the non-FOG coverage to all non-FOG (see Supplementary Material),which shows that the same-cluster non-FOG is the main non-FOG focus of a subtype-specific detection strategy.

B. Feature Mask Similarity

The 10-fold average of Spearman's rank and Pearson correlations between each combination of feature masks are shown in Fig. 4. These average correlations have a low variance across 10 folds with highest CV of 0.17.

Correlations between the general feature mask and the manifestation feature masks (Spearman: 0.19 to 0.49; Pearson: 0.24 to 0.49) are lower than those for motion-specific feature masks (Spearman: 0.6 to 0.96; Pearson: 0.66 to 0.79). Among manifestation feature masks, shuffling has the highest Spearman correlation with the general feature mask. Shuffling and trembling feature masks both show the highest Pearson correlation with the general mask. Between motion-specific feature masks, turning shows the highest Spearman and Pearson correlation with the general mask.

Correlations between manifestation feature masks (Spearman: 0.11 to 0.15; Pearson: 0.07 to 0.16) are lower than those between motion-specific ones (Spearman:

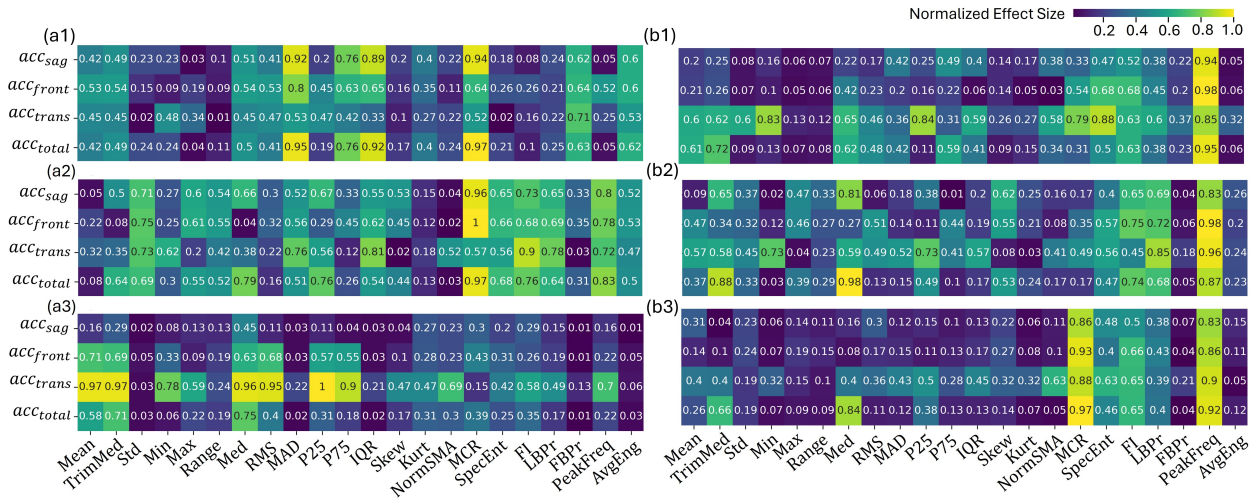

Fig. 2. Average feature masks obtained with FOG subtypes and same-cluster Non-FOG subsets for (a) manifestation-specific subtypes ((a1) Akinesia, (a2) Shuffling, (a3) Trembling) and for (b) motion-specific subtypes ((b1) Gait-initiation, (b2) Turning, (b3) Walking).

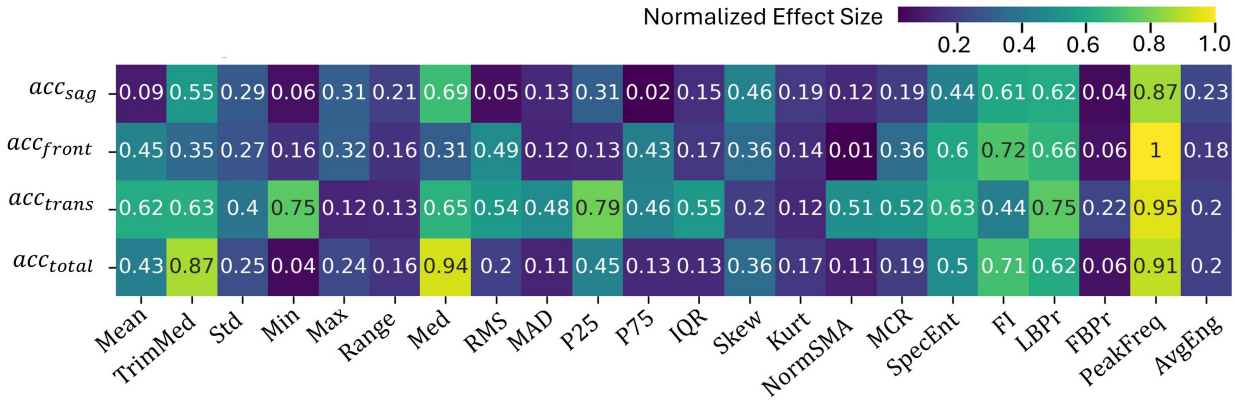

Fig. 3. Average feature mask obtained with all FOG and all Non-FOG.

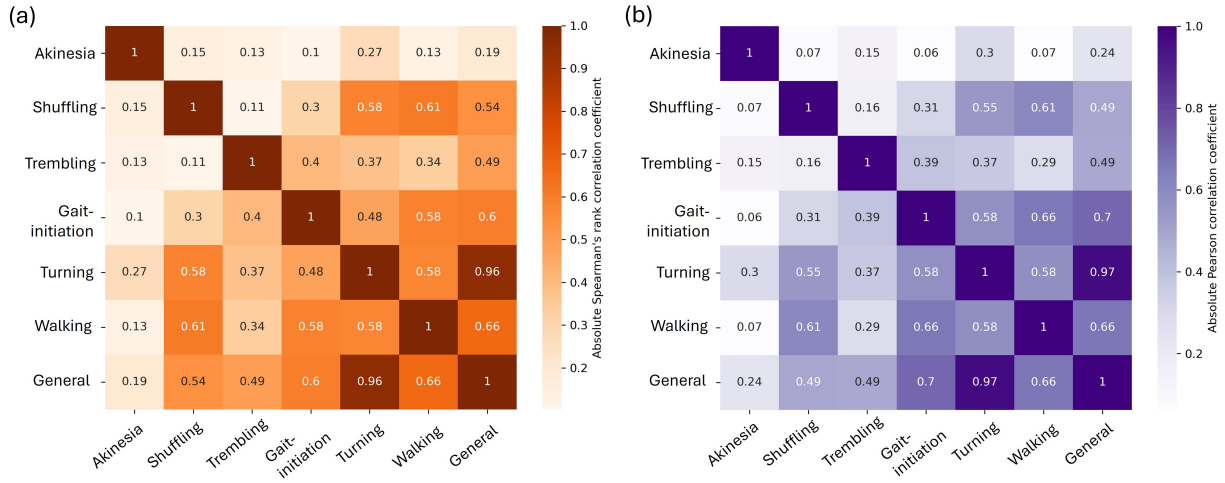

Fig. 4. Spearman's rank correlation (a) and Pearson correlation (b) between feature masks.

TABLE II
ART ANOVA TEST RESULTS

Manifestation-specific Subtype Test Setss				
Source of Variation	**TNR**		**TPR**	
	F value	**p value**	**F value**	**p value**
Model	9.414	0.003	8.148	0.006
Test Set	20.036	< 0.001	76.347	< 0.001
Motion-specific Subtype Test Sets				
Source of Variation	**TNR**		**TPR**	
	F value	**p value**	**F value**	**p value**
Model	0.082	0.775	0.321	0.574
Test Set	5.150	0.009	18.605	< 0.001

0.48 to 0.58; Pearson: 0.58 to 0.66).

C. Detection Model Performances

Fig. 5 visualizes the 10-fold TPRs and TNRs of FOG and non-FOG class on each subtype test set by the corresponding subtype model and the general model. Fig. 6 presents the average TPR and TNR differences between the models.

1) ART ANOVA Test: Based on p values, the TPRs and TNRs on manifestation-specific test sets are significantly influenced by the model (the general model or the subtype model) and used subtype test set, which justifies further comparisons of models within each subtype test set. The TPRs and TNRs on motion-specific test sets are only effected by the test set.

2) Manifestation-specific Model Performances: For model performances in manifestation-specific clusters, shown in Fig. 5 (a) and (b), the general model is better at FOG detection than non-FOG detection in trembling and akinesia. Subtype models significantly improved non-FOG detection, in these clusters. The general model has high specificity for shuffling-like non-FOG, moderate sensitivity for akinesia-like non-FOG, and low effectiveness on trembling-like non-FOG (typically below 50%). Akinesia- and trembling-specific models significantly improve non-FOG specificity for their corresponding non-FOG clusters, particularly for trembling, which arises to about 90%). This comes with a drop of trembling-specific model TPR, though it remains above 80%. No significant trade-off is observed for the akinesia-specific model. In terms of FOG detection, the general model performs perfectly on akinesia and trembling subtypes data but shows weakness in shuffling subtype. The shuffling-specific model effectively resolves this, enhancing the median TPR from 60% to over 80%, at the cost of a moderate drop in its TNR. Fig. 6 (a) further quantifies these effects, showing that the improvements (above-zero bars) outweigh the performance decline (below-zero bars) for all test sets. To evaluate the overall performance changes by the subtype models, the last column aggregates the TPR and TNR differences across all subtype test sets, treating each as equally important. It shows a noticeable net improvement on both FOG and non-FOG by 18.29% ± 8.71 and 24.95% ± 9.80, respectively.

3) Motion-specific Model Performances: For model performances in motion-subtype clusters, illustrated in Fig. 5 (c) and (d), the general model has a moderate median TPR and TNR (above 75%) on all test sets, although being more variant for walking FOG. Subtype models show no significant performance improvements compared to the general model except a small enhancement in turning FOG, with similar amount of specificity decline in turning-FOG-like non-FOG. According to Fig. 6 (b), there are more degradations than improvements by the subtype models while the extent of changes are small. The aggregated performance shifts by the motion-specific models decrease average TPR by 1.89% ± 8.74 and TNR by 5.17% ± 10.76.

IV. DISCUSSION

A. Feature masks are shaped by dominant subtype-specific cluster(s)

Subtype-specific feature masks reflect the dominant discriminative characteristics within their corresponding clusters, highlighting the potential for developing subtype-specific algorithms to further enhance FOG classification. Due to the used definition for manifestations and the clustering approach, the FOG and non-FOG samples within each subtype cluster are expected to have similarities in either FI or spectral entropy (two of the FD features with medium, 0.44, to large, 0.63 and 0.72, normalized effect size in the general model). Akinesia-specific feature mask aims to distinguish completely still freezing from mainly quite standing (akinesia-like non-FOG). Both cases are characterized by low, uniform frequency components, making FD features less helpful. Therefore, the attentions are shifted to TD features, such as MAD and IQR. Shuffling-specific feature mask differentiates short, dragging steps from normal walking and turning (shuffling-like non-FOG), a challenge given their similar underlying gait patterns. Consequently, more TD and FD features are required to encode the subtle differences. For instance, higher dependency on FD features such as MCR helps capture gait rhythm differences. Increased attention to TD features such as range and Std describe gait speed and variability. Trembling-specific feature mask needs to separate tremor from sensor noise or jolting normal motion (trembling-like non-FOG), which also shares similar dominant FD characteristics and thus leads to increased attention to TD. Among added TD features, transverse-plane ones

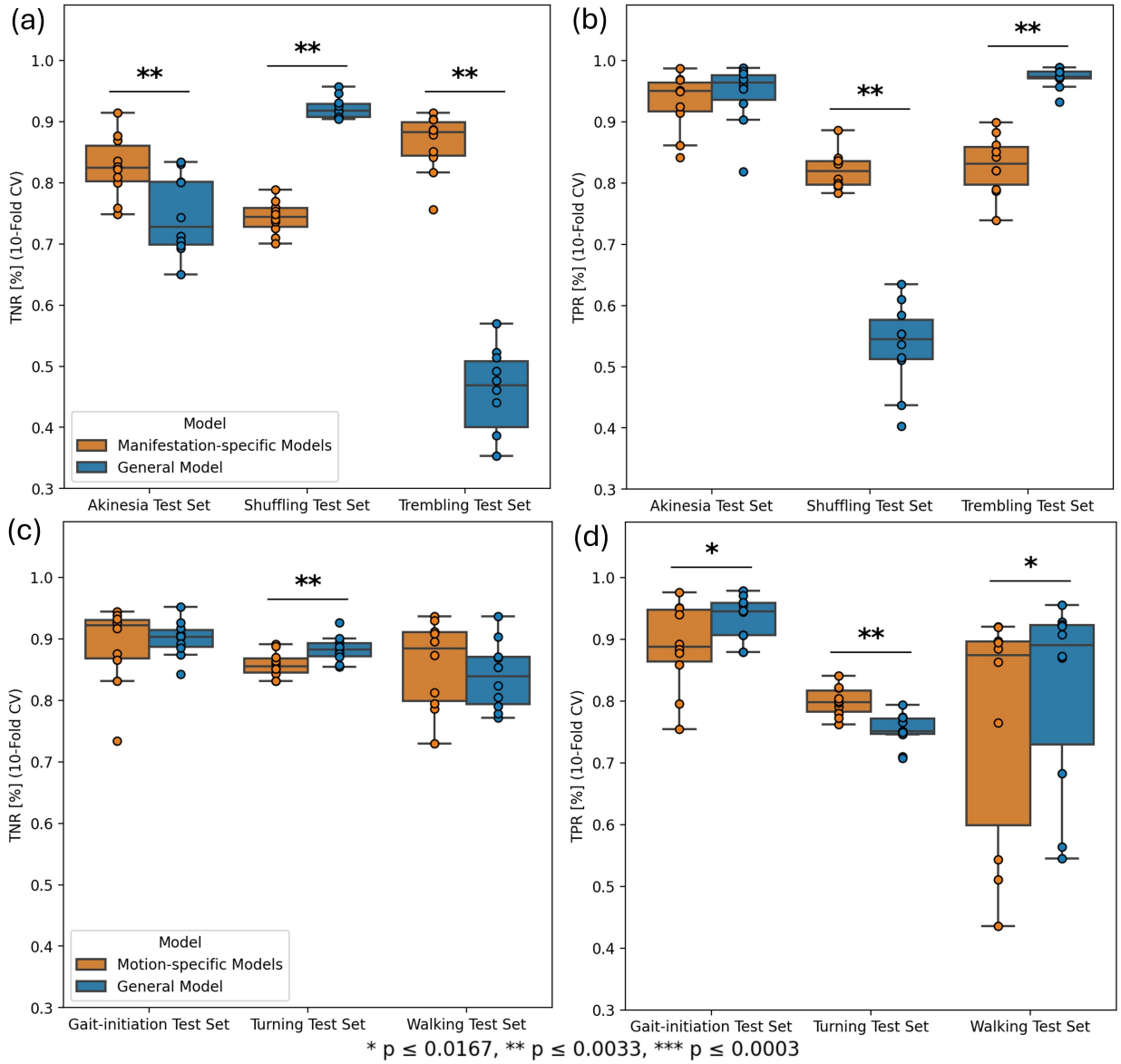

Fig. 5. Performance comparison between manifestation-specific models vs general model on (a) TNR (b) TPR, and motion-specific models vs general model on (c) TNR (d) TPR.

are favored, as they reflect directional changes and movement scale, key for distinguishing true movement in noisy non-FOG windows.

The three motion-specific-subtype feature masks have similar FD-dominated feature preferences, with minor difference for walking case with the addition of MCR, which potentially captures gait rhythm and oscillations linked to the tremors of walking FOG.

The general feature mask is also primarily correlated with its dominant subtype cluster(s), suggesting a potential bias. It has the highest Spearman and (or) Pearson correlations with shuffling and trembling feature masks within all manifestation feature masks. The non-FOG case of shuffling cluster, i.e., walking and turning, dominates normal motion, while trembling is the most frequent manifestation. Similarly, it is most correlated with turning feature mask across motion feature masks, as turning is the dominant motion context in the dataset.

B. Manifestation-specific models result in improved generalizability

Confirming the issue raised in Sec. IV-A, the general model biased towards the dominant FOG (trembling subtype) and non-FOG manifestation (shuffling-like non-FOG), while ignoring minor manifestations like shuffling subtype and other non-FOG scenarios.

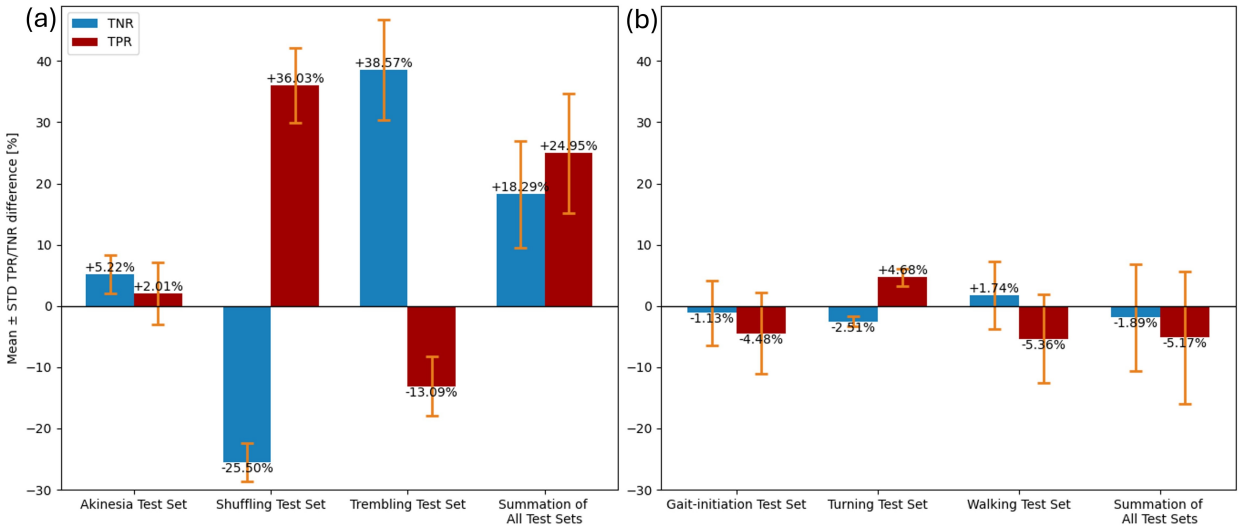

Fig. 6. Average TPR and TNR differences between the general model and (a) manifestation-specific models and (b) motion-specific models.

Resolving this, manifestation-specific models significantly improves sensitivity in shuffling and reduces false alarms in akinesia-like and trembling-like non-FOGs. As a common behavior of binary classifier due to slight adjustment in decision boundary, improvement in one class may associate with degradation in the other class. Thus, those improvements are associated with decrease in shuffling-like specificity and sensitivity for trembling FOG. However, the sensitivity improvements outweigh the decrease of specificity. Consequently, the overall TPR and TNR improves, reflecting enhanced generalizability across the FOG subtypes and their same cluster non-FOG subsets. These improvements typically arise where the general model performs near chance level (median TPR or TNR$<$0.6), while the corresponding performance drops in subtype models remain modest (median TPR or TNR$>$ 0.7) of the subtype model, justifying the trade-off. The observed enhancements also highlight the effectiveness of the proposed subtype-specific training of feature-mask based CNNs, which improves the performance on underrepresented subtype clusters ignored by the general model.

In contrast, motion models only show statistically better sensitivity in turning FOG, albeit being small and leading to a similar specificity reduction for the same-cluster. Overall performance shifts by motion-subtype models also shows no improvements compared to the general model. This is expected since those subtype models and the general model share the same detection strategy as indicated by feature masks, making improvements via motion-specific tailoring unlikely.

C. The manifestation composition of a dataset drives the detection strategy

Tailored training is effective only if the targeted subset is sufficiently distinct from the general data. Since the tailored training was effective with manifestation- but not motion-specific models, it suggests that waist acceleration differ remarkably across manifestation clusters but remain similar across motion clusters. Thus, manifestation clusters act as building blocks of a detection strategy: a specific manifestation composition (i.e. proportion of samples in each manifestation cluster) of a dataset leads to a specific desired detection strategy that maximizes the overall accuracy.

Overall accuracy naturally favors dominant manifestation clusters without accounting for cluster-specific importance. This explains why a general model, optimized for the dataset's overall manifestation distribution, tends to favor dominant subtype. In contrast, motion clusters exhibit similar manifestation composition to the full dataset, as also reflected in their shared detection strategy between motion-specif models and the general model. Since the general model is already tuned to this distribution, it performs reasonably well across all motion-specific test sets, making motion-specific tailoring generally unnecessary. However, this does not rule out the possibility of further improving generalization across motion subtypes, especially in presence of atypical events with distinct manifestation compositions. In such cases, a motion subtype can be further split to represent the underrepresented manifestation composition. For instance, motivated by the high TPR variability of the general and the walking-specific model in walking test set, we can further scrutinize that motion subtype. Given the low

sensitivity of the general model on shuffling data, those walking test folds with low FOG TPR likely contain more shuffling-dominated walking FOG. Being under-represented, these atypical compositions fail to influence the detection strategy and thus remain undetected. To address this, training a model specifically tailored to such less frequent cases with special manifestation profiles may enhance detection performance.

D. Limitations and Future Directions

The utilized dataset has a minimal sensor setting of only one waist-placed accelerometer. Additional FOG-modulated bio-signals collected by other sensors such as skin conductance reflecting anxiety state [47] and foot plantar pressure reflecting both spatial and temporal gait patterns [48] [49] may provide additional insights into the subtype-specific features and differences in FOG detection strategies. Collecting motion data, using inertial measurement units or accelerometers, from other body segments, for instance, ankles, may help further improve deep learning model accuracy while keeping a user-friendly simple sensor setup [35]. In addition, the manifestation-specific subtype labels are based on our implemented clustering result instead of a direct expert labeling. It will be beneficial to involve expert labels to improve the quality of dataset and to obtain a more rigorous clustering specific to the subtypes. An enriched dataset and/or automatic clustering of samples with an improved technique can lead to potentially a more targeted subtype training. Future work may also focus on the same verification method but on prediction instead of detection problem enabling possible interventions.

V. Conclusion

This paper reveals that unique detection strategies (represented by feature masks), tailored on subtype-specific signal characteristics, are required for manifestation-specific subtypes. Manifestation-specific models significantly improve the generalizability of FOG detection across both FOG subtypes and their corresponding non-FOG subsets compared to the general model. This demonstrates the effectiveness of the proposed subtype-specific tailoring and feature-mask-based CNN for subtype-specific tailoring. In contrast, motion-specific subtypes benefit less from tailored strategies, as their subtype-specific models share similar detection strategies. These motion-specific models do not show prominent generalizability enhancement from the general model. This suggests that the underlying manifestation composition, rather than motion type, primarily drives the optimal detection strategy. While the general model is biased toward the dominantly represented FOG, trembling FOG in used dataset, and tends to overlook the less frequent shuffling FOG, manifestation-specific models overcome this limitation through tailored strategies. Since motion subtypes generally share similar manifestation compositions with each other and the overall data, further model tailoring for motion is both unnecessary and unlikely to yield meaningful gains.

References

[1] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, and A. E. Lang, "Parkinson disease," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[2] N. Giladi, T. Treves, E. Simon, H. Shabtai, Y. Orlov, B. Kandinov, D. Paleacu, and A. Korczyn, "Freezing of gait in patients with advanced parkinson's disease," *Journal of neural transmission*, vol. 108, pp. 53–61, 2001.

[3] W.-S. Zhang, C. Gao, Y.-Y. Tan, and S.-D. Chen, "Prevalence of freezing of gait in parkinson's disease: a systematic review and meta-analysis," *Journal of neurology*, vol. 268, pp. 4138–4150, 2021.

[4] E. Heremans, A. Nieuwboer, and S. Vercruysse, "Freezing of gait in parkinson's disease: where are we now?" *Current neurology and neuroscience reports*, vol. 13, pp. 1–9, 2013.

[5] Y. Okuma, "Freezing of gait in parkinson's disease," *Journal of neurology*, vol. 253, pp. vii27–vii32, 2006.

[6] J. Hausdorff, J. Schaafsma, Y. Balash, A. Bartels, T. Gurevich, and N. Giladi, "Impaired regulation of stride variability in parkinson's disease subjects with freezing of gait," *Experimental brain research*, vol. 149, pp. 187–194, 2003.

[7] M. Plotnik, N. Giladi, Y. Balash, C. Peretz, and J. M. Hausdorff, "Is freezing of gait in parkinson's disease related to asymmetric motor function?" *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 57, no. 5, pp. 656–663, 2005.

[8] A. H. Snijders, M. J. Nijkrake, M. Bakker, M. Munneke, C. Wind, and B. R. Bloem, "Clinimetrics of freezing of gait," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. S2, pp. S468–S474, 2008.

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

[10] A. Nieuwboer, W. d. Weerdt, R. Dom, and E. Lesaffre, "A frequency and correlation analysis of motor deficits in parkinson patients," *Disability and rehabilitation*, vol. 20, no. 4, pp. 142–150, 1998.

[11] M. Gilat, "How to annotate freezing of gait from video: a standardized method using open-source software, j. parkinsons dis. 9 (2019) 821–824."

[12] M. Piper, G. M. Abrams, and W. J. Marks Jr, "Deep brain stimulation for the treatment of parkinson's disease: overview and impact on gait and mobility," *NeuroRehabilitation*, vol. 20, no. 3, pp. 223–232, 2005.

[13] P. J. McCandless, B. J. Evans, J. Janssen, J. Selfe, A. Churchill, and J. Richards, "Effect of three cueing devices for people with parkinson's disease with gait initiation difficulties," *Gait & posture*, vol. 44, pp. 7–11, 2016.

[14] C. Changcheng, Y.-R. Li, and C.-T. Chen, "Assistive mobility control of a robotic hip-knee exoskeleton for gait training," *Sensors*, vol. 22, no. 13, p. 5045, 2022.

[15] T. Huang, M. Li, and J. Huang, "Recent trends in wearable device used to detect freezing of gait and falls in people with parkinson's disease: A systematic review," *Frontiers in aging neuroscience*, vol. 15, p. 1119956, 2023.

[16] S. T. Moore, H. G. MacDougall, and W. G. Ondo, "Ambulatory monitoring of freezing of gait in parkinson's disease," *Journal of neuroscience methods*, vol. 167, no. 2, pp. 340–348, 2008.

[17] 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?" *Journal of neuroengineering and rehabilitation*, vol. 20, no. 1, p. 53, 2023.

[18] P. Tahafchi, R. Molina, J. A. Roper, K. Sowalsky, C. J. Hass, A. Gunduz, M. S. Okun, and J. W. Judy, "Freezing-of-gait detection using temporal, spatial, and physiological features with a support-vector-machine classifier," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 2867–2870.

[19] N. Kleanthous, A. J. Hussain, W. Khan, and P. Liatsis, "A new machine learning based approach to predict freezing of gait," *Pattern Recognition Letters*, vol. 140, pp. 119–126, 2020.

[20] S. Pardoel, J. Kofman, J. Nantel, and E. D. Lemaire, "Wearable-sensor-based detection and prediction of freezing of gait in parkinson's disease: a review," *Sensors*, vol. 19, no. 23, p. 5141, 2019.

[21] A. Arami, A. Poulakakis-Daktylidis, Y. F. Tai, and E. Burdet, "Prediction of gait freezing in parkinsonian patients: A binary classification augmented with time series prediction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 9, pp. 1909–1919, 2019.

[22] T. Ashfaque Mostafa, S. Soltaninejad, T. L. McIsaac, and I. Cheng, "A comparative study of time frequency representation techniques for freeze of gait detection and prediction," *Sensors*, vol. 21, no. 19, p. 6446, 2021.

[23] J. Camps, A. Sama, M. Martín, D. Rodriguez-Martin, C. Pérez-López, J. M. M. Arostegui, J. Cabestany, A. Catala, S. Alcaine, B. Mestre *et al.*, "Deep learning for freezing of gait detection in parkinson's disease patients in their homes using a waist-worn inertial measurement unit," *Knowledge-Based Systems*, vol. 139, pp. 119–131, 2018.

[24] S. K. Bansal, B. Basumatary, R. Bansal, and A. K. Sahani, "Techniques for the detection and management of freezing of gait in parkinson's disease–a systematic review and future perspectives," *MethodsX*, vol. 10, p. 102106, 2023.

[25] S. Factor, M. Scullin, A. Sollinger, J. Land, C. Wood-Siverio, L. Zanders, A. Freeman, D. Bliwise, and F. Goldstein, "Freezing of gait subtypes have different cognitive correlates in parkinson's disease," *Parkinsonism & related disorders*, vol. 20, no. 12, pp. 1359–1364, 2014.

[26] J. Schaafsma, Y. Balash, T. Gurevich, A. Bartels, J. M. Hausdorff, and N. Giladi, "Characterization of freezing of gait subtypes and the response of each to levodopa in parkinson's disease," *European journal of neurology*, vol. 10, no. 4, pp. 391–398, 2003.

[27] A. H. Snijders, K. Takakusaki, B. Debu, A. M. Lozano, V. Krishna, A. Fasano, T. Z. Aziz, S. M. Papa, S. A. Factor, and M. Hallett, "Physiology of freezing of gait," *Annals of neurology*, vol. 80, no. 5, pp. 644–659, 2016.

[28] A. Lieberman, "Are freezing of gait (fog) and panic related?" *Journal of the neurological sciences*, vol. 248, no. 1-2, pp. 219–222, 2006.

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

[30] D. Cowie, P. Limousin, A. Peters, M. Hariz, and B. L. Day, "Doorway-provoked freezing of gait in parkinson's disease," *Movement disorders*, vol. 27, no. 4, pp. 492–499, 2012.

[31] R. G. Cohen, J. G. Nutt, and F. B. Horak, "Recovery from multiple apas delays gait initiation in parkinson's disease," *Frontiers in human neuroscience*, vol. 11, p. 60, 2017.

[32] M. Mancini, V. V. Shah, S. Stuart, C. Curtze, F. B. Horak, D. Safarpour, and J. G. Nutt, "Measuring freezing of gait during daily-life: an open-source, wearable sensors approach," *Journal of neuroengineering and rehabilitation*, vol. 18, pp. 1–13, 2021.

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

[34] L. L. Chan, S. Yang, M. Aswani, L. Kark, E. Henderson, S. R. Lord, and M. A. Brodie, "Development, validation, and limits of freezing of gait detection using a single waist-worn device," *IEEE Transactions on Biomedical Engineering*, 2024.

[35] J. O'Day, M. Lee, K. Seagers, S. Hoffman, A. Jih-Schiff, Ł. Kidziński, S. Delp, and H. Bronte-Stewart, "Assessing inertial measurement unit locations for freezing of gait detection and patient preference," *Journal of neuroengineering and rehabilitation*, vol. 19, no. 1, p. 20, 2022.

[36] N. Naghavi and E. Wade, "Towards real-time prediction of freezing of gait in patients with parkinson's disease: a novel deep one-class classifier," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1726–1736, 2021.

[37] F. Demrozi, R. Bacchin, S. Tamburin, M. Cristani, and G. Pravadelli, "Toward a wearable system for predicting freezing of gait in people affected by parkinson's disease," *IEEE journal of biomedical and health informatics*, vol. 24, no. 9, pp. 2444–2451, 2019.

[38] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, and B. Vanrumste, "Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, no. 1, p. 24, 2024.

[39] W. T. Mo and J. H. Chan, "Predicting three types of freezing of gait events using deep learning models," *arXiv preprint arXiv:2310.06322*, 2023.

[40] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, and B. Vanrumste, "Automatic detection and assessment of freezing of gait manifestations," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[41] M. Shaban, "A novel variational mode decomposition based convolutional neural network for the identification of freezing of gait intervals for patients with parkinson's disease," *Machine Learning with Applications*, vol. 16, p. 100553, 2024.

[42] A. Howard, amit salomon, eran gazit, J. Hausdorff, L. Kirsch, Maggie, P. Ginis, R. Holbrook, and Y. F. Karim, "Parkinson's freezing of gait prediction," https://kaggle.com/competitions/tlvmc-parkinsons-freezing-gait-prediction, 2023, kaggle.

[43] K. Ziegler, F. Schroeteler, A. O. Ceballos-Baumann, and U. M. Fietzek, "A new rating instrument to assess festination and freezing gait in parkinsonian patients," *Movement Disorders*, vol. 25, no. 8, pp. 1012–1018, 2010.

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

[45] E. Cavita-Huerta, J. Reyes-Reyes, H. M. Romero-Ugalde, G. L. Osorio-Gordillo, R. F. Escobar-Jiménez, and V. M. Alvarado-Martínez, "Human activity recognition from accelerometry, based on a radius of curvature feature," *Mathematical and Computational Applications*, vol. 29, no. 5, p. 80, 2024.

[46] G. M. Sullivan and R. Feinn, "Using effect size—or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.

[47] S. Mazilu, A. Calatroni, E. Gazit, A. Mirelman, J. M. Hausdorff, and G. Tröster, "Prediction of freezing of gait in parkinson's from physiological wearables: an exploratory study," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1843–1854, 2015.

[48] K. Hu, S. Mei, W. Wang, K. A. E. Martens, L. Wang, S. J. Lewis, D. D. Feng, and Z. Wang, "Multi-level adversarial spatio-temporal learning for footstep pressure based fog detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 4166–4177, 2023.

[49] G. Shalin, S. Pardoel, E. D. Lemaire, J. Nantel, and J. Kofman, "Prediction and detection of freezing of gait in parkinson's disease from plantar pressure data using long short-term memory neural-networks," *Journal of neuroengineering and rehabilitation*, vol. 18, pp. 1–15, 2021.