Subject-Independent Freezing of Gait (FoG) Prediction in Parkinson's Disease Patients

Toygun Basaklar, Yigit Tuncel, and Umit Y. Ogras
Department of Electrical and Computer Engineering
University of Wisconsin-Madison, Madison, WI, 53706, USA
E-mail: basaklar@wisc.edu; tuncel@wisc.edu; uogras@wisc.edu

***Abstract*—Freezing of gait (FoG), which implies a brief absence or reduction of ability to walk, is one of the most common symptoms of Parkinson's Disease (PD). Predicting FoG episodes in time can prevent their onset by providing specific cues to the patients. This paper presents a deep learning approach to predict FoG episodes using a Long Short-Term Memory network (LSTM). It also identifies key issues and concepts which have not been dwelled upon before in the existing literature, paving the way to a more systematic methodology for future work. We evaluate our approach using a publicly available dataset that includes accelerometer readings from 10 PD patients. We achieve up to 89% prediction accuracy with an average prediction time of $1.42\ s$ *using a subject-independent model*.**

I. Introduction

Parkinson's Disease (PD) is one of the most common age-related neurodegenerative diseases. It results in muscular rigidity, tremor, bradykinesia, slowness in movement, and postural instability [1–4]. Over 80% of PD patients develop freezing of gait (FoG) [5], a brief absence of the ability to walk despite the intention of moving the feet [6]. Consequences of FoG include anxiety, loss of mobility, and fall, which significantly deteriorate patients' quality of life.

Clinical studies suggest that auditory, visual, or tactile cues synchronized with the gait help patients to exit the FoG state and resume walking [6]. Therefore, FoG-related problems can be avoided by systems that can predict FoG episodes and activate an appropriate cueing mechanism. FoG prediction is a challenging preposition since FoG episodes are rare events that are closely intertwined with daily life activities. For example, FoG episodes occur during turning, walking through doorways, or dual-tasking, and 90% of them last less than 20 seconds [5]. Specific experimental sessions are designed to simulate the daily life activities in clinical settings and collect data from PD patients. However, the duration of the sessions and the frequency of FoG occurrence are limited.

FoG detection, which implies classifying FoG episodes *after patients experience them*, is heavily studied for over a decade. Over 50 studies related to FoG detection have been published by late 2019 [7]. However, there is a strong need for techniques to *predict potential FoG episodes and prevent their onset* by providing preemptive cueing. These approaches can prevent FoG episodes and increase the patients' quality of life. Despite this potential, only a handful of studies related to FoG prediction have been published before 2020 [7]. The number of studies that address FoG prediction has doubled in recent years, thanks to the evolving sensor technology. However, there is still a lack of a clear consensus on the definition of prediction of FoG episodes. This problem arises because the methodology to identify the samples used to evaluate the prediction accuracy and time differs substantially for different studies. There is an ambiguity regarding the proper prediction performance of these studies. There are two main underlying reasons: i) Overestimation of the performance of the classifier due to the partial overlap between training and test set, ii) Inclusion of the segments/samples right before the actual onset to the evaluation process although correctly classifying those segments does not facilitate *prediction* since it cannot enable preemptive cueing. Furthermore, most studies focus on subject-specific modeling using methods that employ costly feature extraction techniques. However, the frequency of FoG occurrence is low and the classifier cannot learn the general characteristics of pre-FoG and FoG samples with limited data.

This paper presents a *subject-independent deep learning approach using only time-series data from wearable sensors.* Since raw sensor data has strong temporal dependency between samples, we employ Long-Short Term Memory (LSTM) network and evaluate our approach on a publicly available dataset (Daphnet Freezing of Gait Dataset [8]). We identify key points that are missed by the existing literature regarding the methodology of FoG prediction. Using this insight, we refine the prediction accuracy and average prediction time to facilitate a more intuitive and systematic methodology for future work. The major contributions of this work are as follows:

- Achieve up to 89% prediction accuracy with average prediction time of $1.42\ s$ *using a subject-independent model*.
- Identify the key issues regarding the definition of prediction and propose a simpler and intuitive alternative.
- Provide a comprehensive analysis of the effects of the percentage overlap between each consecutive segment, pre-FoG duration, segment size, and prediction horizon on the prediction accuracy and time.

II. Related Work

Prior studies showed that the sensor data preceding the actual FoG onset has distinct characteristics [3, 4, 6, 9]. Thus, it is crucial to identify pre-FoG samples correctly. The pre-FoG period is still a debate topic. A single value may not be correct

This research was funded in part by NSF CAREER award CNS-2114499, DARPA Young Faculty Award (YFA) Grant D14AP00068.

for every patient and FoG instance since the pre-FoG duration varies similar to FoG episode durations [5]. Therefore, the methodology for processing and labeling sensor samples is critical for achieving accurate FoG prediction.

The conventional methodology followed by most of the studies begins with data pre-processing (e.g., filtering, outlier detection, and correction). Then, data is segmented into windows of t-seconds for further processing. These segments are generally obtained by sliding window approach with t_s seconds of step size or overlap of w-percent of the previous window. Since the goal is to predict FoG episodes, segments preceding the actual FoG onset are labeled as "*FoG*" (2-class problem [1–3, 10]) or as "*pre-FoG/transition*" (3-class problem [9–11]). Once the segments are identified and labeled, the raw data is used as time series [2, 12] or transformed into features using feature extraction algorithms [1, 9, 11]. Then, the data is classified as "*FoG*", "*pre-FoG*", and "*no-FoG*" segments using either a conventional machine learning algorithm (e.g., support vector machine, linear discriminant analysis [10, 13]) or neural-network/autoregressive modeling approach [1–3, 13].

Recent studies that use the Daphnet dataset mainly focus on subject-specific modeling since gait characteristics, sensor placements, and cognitive tasks during sessions are user-specific. However, FoG occurrence is rare, and data collected from one subject may not represent the general characteristics of pre-FoG and FoG episodes. It even may not be sufficient to train a reliable classifier for the same subject. One idea to remedy this problem is using transfer learning between the trained models for different subjects. Two recent studies [2, 12] employ the transfer learning approach by first training the neural network with data from all subjects but one and add a final layer for the target subject's data. In [2], an overall accuracy, which is over 90%, is reported for all samples instead of prediction accuracy for transition samples, and in [12], the definition of correct prediction of an FoG episode may overestimate the performance of the proposed approach. Despite the transfer learning approach, both studies show that their approach is not generalizable among all subjects. Recent work by Kleanthous et al. [9] proposes subject-independent modeling using the data from all subjects to train and test multiple machine learning models using different feature sets. It introduces the "*transition*" class and investigates the effect of three different segment sizes without overlap. Transition segments include only the samples right before the onset. Hence, correctly classifying those segments should not be considered as *prediction*. A more detailed comparison is presented in Section IV-B.

The definition of the prediction needs to be precise. In addition, the methodology should consider the total energy consumption since FoG prediction targets wearable systems with stringent energy budgets considering the need to collect data in the home environment. In contrast to prior work, this paper focuses on a precise definition of the prediction of FoG episodes. We employ subject-independent modeling using time-series data to predict FoG episodes and investigate the effects of the overlap between segments, pre-FoG duration, segment size, and prediction horizon on the prediction accuracy.

III. Subject-Independent FoG Prediction

A. Preliminaries

The input to the proposed technique is data streaming from wearable sensors, such as inertial measurement units. For example, our study employs the Daphnet Freezing of Gait Dataset [8], designed to benchmark the algorithms to detect FoG in real-time using wearable accelerometers. 3D accelerometers are placed above the ankle, above the knee, and on the lower back of 10 PD patients, all with FoG history (mean age of 66.5 ± 4.8 $s.d$). The sampling rate for the sensors is 64 Hz. Subjects performed two sessions, each consisting of three basic walking tasks, more specifically, (i) walking back and forth in a straight line, including several 180 degrees turn; (ii) random walking with a series of initiated stops and several 360 degrees turn and (iii) walking tasks simulating the activities of daily living which includes getting something to drink, entering and leaving rooms, walking to kitchen or restrooms. Physiotherapists annotated the FoG occurrences both during and after the experiments by analyzing the video recordings. Each sample from accelerometers is labeled as "0" for "*not part of the experiment*", "1" for "*no-FoG*", and "2" for "*FoG*". Eight out of ten patients exhibited FoG during the experiments, and in total, 237 FoG (29 minutes) instances were recorded out of 500 minutes of data.

B. Data Pre-processing and Preparation

We first remove all the samples labeled as "0" since they are not part of the experiments. The dataset also has "no activity/stop" samples for each subject, as shown in Figure 1. Since these samples have similar characteristics to the "*FoG*" samples (highlighted in light red in Figure 1), they can create false positives for the "*FoG*" class. To reduce this risk, we remove these samples from the dataset by manually inspecting the data for each subject. Upon inspection of data, we identify artifacts/outliers (e.g., abnormal spikes). These outliers are detected using a Hampel identifier with a window length of 32 and a threshold of 10 times the median absolute deviation. Then, they are replaced using a 2^{nd} order polynomial interpolation of neighboring values. The data is then filtered using a 4^{th} order Butterworth band-pass filter with cutoff frequencies of 0.5 Hz and 20 Hz. We choose 20 Hz as the high cutoff frequency since the human walking motion and the information for FoG event detection is in the low-frequency range. Once the filtering is done, the data is normalized in the $[-1, 1]$ interval using the values only from the training set.

Since the dataset is heavily imbalanced with "*no-FoG*" samples being the majority class, we follow the 2-class problem by labeling the samples preceding the actual FoG onset by *pre-FoG* duration (τ_{pf}) as "2" (FoG). In Figure 1, the actual FoG samples are highlighted with light red, while the preceding samples τ_{pf} seconds before the actual onset are highlighted with gray for illustration. Next, we segment the data into t-second windows with w-percent overlap between consecutive segments. For example, Figure 1 shows two consecutive 3-second segments. Since the overlap is $w = 25\%$, Segment

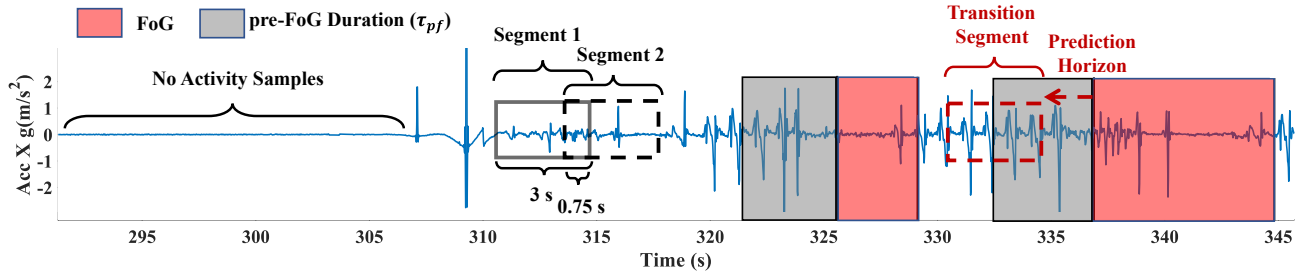

Fig. 1: Visualization of our data preparation approach. Accelerometer data from S5.

2 includes last 0.75 s of Segment 1. If more than half of the samples in a segment are labeled as "*FoG*", it is labeled as "*FoG*". In summary, each segment has $64 \times t$ timesteps and 9 channels (three 3D accelerometer is used in the dataset). Apart from "*no-FoG*" and "*FoG*" segments, we define transition segments as the segments which have any samples within the *pre-Fog* duration, τ_{pf}. We identify each transition segment and its distance in time to the FoG onset.

All the steps above are applied to each subject's data separately. Then, we combine the segments from all subjects and randomly choose 80% of segments as the training and 20% as the test set while keeping the ratio of number of segments from each class the same in both sets and propose a *generalizable subject-independent model* to predict FoG.

C. Proposed Deep Learning Approach for FoG Prediction

Since the real-time sensor data has a strong temporal dependency between samples, we implement an LSTM network to model this dependency. Our network consists of two bidirectional LSTM layers with timesteps of size $64 \times t$ with nine features, where each timestep has 32 hidden states. Hyperbolic tangent and sigmoid are used as the state and gate activation functions, respectively. Two fully connected layers with ReLU activation follow the LSTM layers. The first one consists of 32 neurons, while the second one consists of 16 neurons. We employ a dropout layer with a rate of 0.5 between the LSTM and fully connected layers to prevent overfitting. We also perform a batch normalization layer after each hidden layer in the network to improve the learning performance. Finally, a softmax layer is used at the output layer. The dataset is heavily skewed towards the "*no-FoG*" class, where the ratio between class samples is approximately 0.2. Therefore, we use sample weighting in the loss function. We set the learning rate as 5×10^{-4} and the loss function as categorical cross-entropy. We also define early stopping criteria based on validation loss. The training batch size and the maximum number of epochs are set to 64 and 1000, respectively. The proposed network is implemented in Python using Keras API.

D. Performance Metrics

To evaluate the overall performance of the proposed approach, we use accuracy, sensitivity, and specificity metrics defined as follows:

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN}, \ Specificity = \frac{TN}{TN + FP} \quad (2)$$

where TP, FP, TN, and FN represent the true positives, false positives, true negatives, and false negatives, respectively. Also, "*no-FoG*" is the positive class, "*FoG*" is the negative class.

We define the prediction accuracy and time in their simplest form. Recall that we identify the transition segments in the test set and their distance in time to the actual FoG onset. We define a period of time called the prediction horizon as shown in Figure 1 to choose which transition segments should be considered to calculate prediction accuracy and time. We identify the transition segments that are closest to the prediction horizon. Since the dataset is segmented to t-second windows, these transition segments may end before or after the prediction horizon. We define *prediction accuracy* as the ratio of the transition segments identified as "*FoG*" to the overall number of transition segments. We also calculate the *average prediction time* by taking the average time difference between correctly identified transition segments and the corresponding FoG onset.

IV. Experimental Results

This work investigates the effects of the percentage overlap (w) between each consecutive segment, pre-FoG duration (τ_{pf}), the segment size (t-seconds), and the prediction horizon on the prediction accuracy using the sets $[0\%, 25\%, 50\%]$, $[1, 3, 5]$, $[0.5, 1, 2, 3]$, and $[0, 1, ..., \tau_{pf}]$ respectively. We also provide the number of transition segments included in calculating prediction accuracy and time. *This information has a major impact on the evaluation, which is overlooked in the existing literature.* The results given in this section are averages of five runs with different random seeds.

A. FoG Prediction Evaluation

This section evaluates the performance of the proposed approach using the performance metrics described in Section III-D.

1) Impact of Percentage Overlap: The main objective of this work is to predict the FoG episodes. To this end, for each parameter setting, we identify the transition segments and calculate their classification accuracy, which is defined as *prediction accuracy* in Section III-D. Figure 2 shows the prediction accuracy of the proposed approach for $\tau_{pf} = 1\ s$

using various segment sizes, overlap percentages, and prediction horizons. *The prediction accuracy increases with the overlap duration*, which goes up to 91.17% using 3-second segment size as shown in Figure 2. A similar behavior is observed for other parameter settings, as illustrated in Figure 3 and Figure 4. The dataset is highly skewed to the "*no-FoG*" class. Overlap provides additional temporal dependent segments, which improve the learning performance of the model for the under-represented class. However, it may also introduce partial data leakage between training and test set and cause an overestimation of the model's performance, which is neglected by most studies in the literature.

The overall accuracy, sensitivity, and specificity of the model increase as the percentage overlap changes from 0% to 50% for all τ_{pf} values and segment sizes, as shown in Table I. For instance, for $\tau_{pf} = 3$ s and 2-second segment size, the accuracy, sensitivity, and specificity values goes from 82.04%, 82.27%, and 81.51% to 93.05%, 92.76%, and 93.71% for $w = 0\%$, $w = 25\%$, and $w = 50\%$ respectively. Similar results are obtained for other parameter settings suggesting that there is no overfitting towards one class, which are given in Table I. Note that the overall accuracy in this table does not represent the prediction accuracy.

2) Impact of pre-FoG Duration (τ_{pf}): The increase in τ_{pf} results in a more balanced dataset since the number of "*no-FoG*" segments decreases as the number of "*FoG*" segments increases. Larger τ_{pf} does not mean the model improves with a balanced dataset since sample weighting is used in the loss function. On the contrary, it may result in false positives, which degrades the overall performance. However, we cannot observe any significant performance differences between models using different τ_{pf} values as the pre-FoG gait characteristics depend on the subjects [12] while our model is subject-independent.

3) Impact of Prediction Horizon and Segment Size: As the prediction horizon increases, the prediction accuracy decreases or remains similar (see Figures 2, 3, and 4). This pattern is observed by other studies in the literature as well [9]. There is a clear trend in Figure 2, Figure 3, and Figure 4 except for one salient case. For $\tau_{pf} = 5$ s, our model achieves up to 69.03%, 70.56%, 76.73%, 84.24%, 77.92%, and 97.62% prediction accuracy with average prediction time of 4.56, 4.16, 3.17, 2.16, 1.27, and 0.17 seconds before the actual onset respectively using 0.5-second segment size and $w = 50\%$. Although the overall prediction accuracy decreases as the prediction horizon increases in this setting, the accuracy for the 2-second prediction horizon is much larger than the accuracy for the 1-second prediction horizon. This shows that the transition segments for these two cases have different characteristics, suggesting that the split of training and test set may have a major impact on the obtained results. We deduct that the methodology for FoG prediction should follow a random split technique with a large number of runs (> 10) to be averaged.

Segment size being larger than or equal to the τ_{pf} mainly results in the same transition segments to be evaluated for different prediction horizons. Hence, the prediction horizon parameter loses its importance and the prediction accuracy and time for those configurations become very similar. As it is shown in Figure 2, different prediction horizons yield the same prediction accuracy for the $\tau_{pf} = 1$ s and for $t \geq 1s$. It may also result in negative prediction time values since the transition segments may include samples after the actual FoG onset. For example, 3-second segment size yields an average prediction time of -1.5 s for $\tau_{pf} = 1$ s (see Table I).

4) Best Parameter Setting: We suggest that the FoG should be predicted at least 1 second before the actual FoG onset to enable preemptive cueing or fall preventive mechanism. Considering this, we identify the best parameter configuration for each τ_{pf} values. The best parameter pairs for $\tau_{pf} = 1$ s, $\tau_{pf} = 3$ s, and $\tau_{pf} = 5$ s are (0.5-second segment size, $w = 50\%$), (2-second segment size, $w = 50\%$), and (1-second segment size, $w = 50\%$) respectively. Using these pairs, our model achieves up to 89.42%, 88.34%, and 88.52% prediction

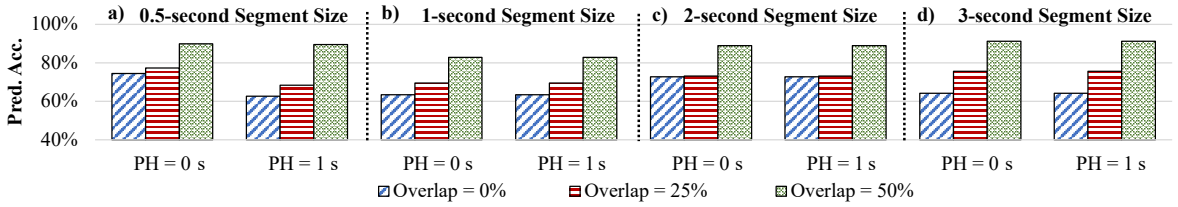

Fig. 2: Prediction Accuracy for $\tau_{pf} = 1$ s. *PH: Prediction Horizon

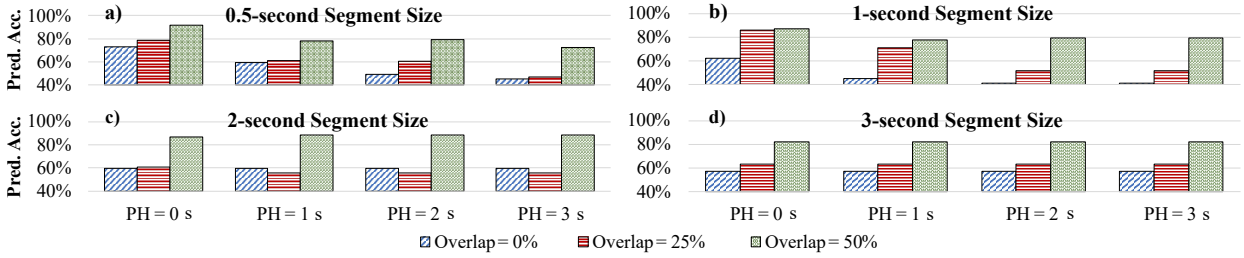

Fig. 3: Prediction Accuracy for $\tau_{pf} = 3$ s. *PH: Prediction Horizon

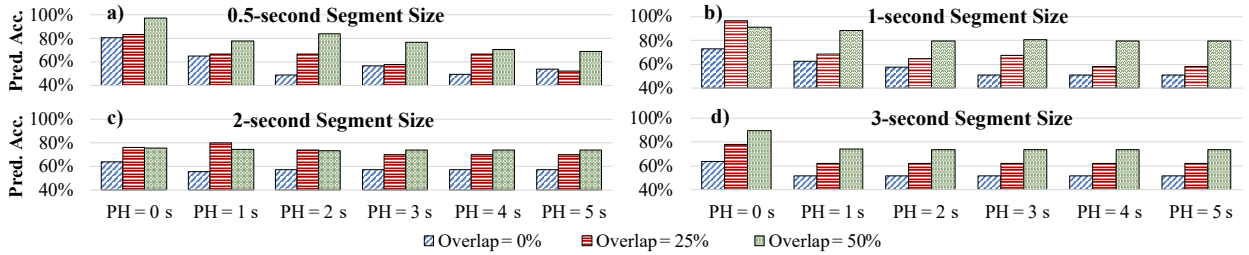

Fig. 4: Prediction Accuracy for $\tau_{pf} = 5\ s$. *PH: Prediction Horizon

TABLE I: Additional results for $\tau_{pf} = 1\ s$, $\tau_{pf} = 3\ s$, and $\tau_{pf} = 5\ s$. *PH: Prediction Horizon

	0.5-second Segment Size			1-second Segment Size			2-second Segment Size			3-second Segment Size		
	$w = 0\%$	$w = 25\%$	$w = 50\%$	$w = 0\%$	$w = 25\%$	$w = 50\%$	$w = 0\%$	$w = 25\%$	$w = 50\%$	$w = 0\%$	$w = 25\%$	$w = 50\%$
$\tau_{pf} = 1\ s$												
Overall Accuracy	82%	85%	91%	83%	85%	90%	84%	85%	92%	82%	87%	92%
Sensitivity	84%	86%	91%	84%	86%	91%	86%	86%	92%	85%	88%	92%
Specificity	78%	82%	92%	78%	83%	88%	79%	81%	92%	72%	85%	92%
Avg. Pred. Time (s) (P.H=0)	0.06	0.19	0.15	0.09	0.09	0.08	-0.73	-0.73	-0.71	-1.52	-1.42	-1.46
# of Transition Segments (P.H=0)	39	39	49	47	49	45	37	44	38	42	45	41
Avg. Pred. Time (s) (P.H=1)	0.57	0.57	0.57	0.09	0.09	0.08	-0.73	-0.73	-0.71	-1.52	-1.42	-1.46
# of Transition Segments (P.H=1)	43	46	52	47	49	45	37	44	38	42	45	41
$\tau_{pf} = 3\ s$												
Overall Accuracy	82%	84%	88%	82%	87%	93%	82%	87%	93%	81%	87%	92%
Sensitivity	84%	86%	88%	84%	88%	95%	82%	89%	93%	84%	89%	94%
Specificity	76%	80%	88%	76%	84%	89%	82%	82%	94%	76%	83%	88%
Avg. Pred. Time (s) (P.H=0)	0.24	0.25	0.16	0.63	0.47	0.37	0.27	0.10	0.26	0.45	0.41	0.43
# of Transition Segments (P.H=0)	34	33	37	33	35	42	35	41	35	38	30	39
Avg. Pred. Time (s) (P.H=1)	1.24	1.21	1.16	1.10	1.36	1.37	1.27	1.22	1.26	0.60	0.54	0.60
# of Transition Segments (P.H=1)	37	42	38	36	38	43	35	39	34	38	30	39
Avg. Pred. Time (s) (P.H=2)	2.08	2.19	2.16	2.08	2.07	2.09	1.27	1.22	1.26	0.45	0.41	0.43
# of Transition Segments (P.H=2)	39	40	34	36	35	36	35	39	34	38	30	39
Avg. Pred. Time (s) (P.H=3)	2.55	2.56	2.56	2.06	2.06	2.10	1.23	1.20	1.26	0.43	0.44	0.45
# of Transition Segments (P.H=3)	33	41	39	36	35	36	35	39	34	38	30	39
$\tau_{pf} = 5\ s$												
Overall Accuracy	81%	83%	89%	81%	85%	92%	80%	86%	90%	80%	83%	91%
Sensitivity	82%	84%	90%	83%	86%	92%	81%	86%	92%	80%	84%	92%
Specificity	78%	80%	87%	79%	84%	92%	80%	85%	88%	81%	81%	90%
Avg. Pred. Time (s) (P.H=0)	0.26	0.22	0.17	0.75	0.33	0.39	1.26	0.89	0.72	0.11	0.19	0.54
# of Transition Segments (P.H=0)	34	34	43	35	32	37	36	33	31	34	34	30
Avg. Pred. Time (s) (P.H=1)	1.28	1.20	1.17	1.69	1.56	1.42	1.30	1.77	1.70	1.18	1.14	1.15
# of Transition Segments (P.H=1)	36	33	37	34	30	30	36	32	32	29	33	28
Avg. Pred. Time (s) (P.H=2)	2.29	2.25	2.16	2.74	2.46	2.42	2.27	2.11	2.31	2.18	2.21	2.30
# of Transition Segments (P.H=2)	31	39	39	33	35	30	28	29	30	29	31	30
Avg. Pred. Time (s) (P.H=3)	3.28	3.23	3.17	3.11	3.32	3.40	3.27	3.26	3.34	2.24	2.30	2.47
# of Transition Segments (P.H=3)	33	35	34	33	35	36	28	30	29	29	31	30
Avg. Pred. Time (s) (P.H=4)	4.06	4.18	4.16	4.10	4.11	4.08	3.27	3.26	3.34	2.18	2.21	2.30
# of Transition Segments (P.H=4)	29	31	40	33	34	37	28	30	29	29	31	30
Avg. Pred. Time (s) (P.H=5)	4.54	4.55	4.56	4.10	4.11	4.08	3.27	3.26	3.34	2.18	2.21	2.30
# of Transition Segments (P.H=5)	30	38	31	33	34	37	28	30	29	29	31	30

accuracy with an average prediction time of 0.57, 1.26, and 1.42 seconds for $\tau_{pf} = 1\ s$, $\tau_{pf} = 3\ s$, and $\tau_{pf} = 5\ s$, respectively. In summary, *our approach achieves satisfactory prediction accuracy and time values using subject-independent modeling*. The obtained results for various parameter configurations are in line with the literature *with the addition of a precise definition of prediction and metrics related to transition segments*.

B. Comparison to the State-of-the-Art Approaches

This section compares our results against recent studies on FoG prediction using the Daphnet dataset. It also discusses key issues regarding their approach and the definition of prediction.

Kleanthous et al. [9] propose a 3-class problem by introducing the "*transition*" class as the period between "*no-FoG*" and "*FoG*" labels. They investigate 2, 3, and 4-second windows using different feature sets and classifiers. Since the "*transition*" class is formed from the "*no-FoG*" class and no overlap is applied between segments, the total number of segments from each class becomes around 200 segments, which is significantly small for most machine learning algorithms. They use subject-independent modeling and an SVM classifier with radial basis kernel function. Their approach achieves a sensitivity of 72.34%, 91.49%, 75.00% and specificity of 87.36%, 88.51%, and 93.62% for "*FoG*", "*transition*", and "*no-FoG*" classes respectively, with 3-second windows. Sensitivity values of the best performing classifier show that it favors the transition class the most while performing poorly on the other two classes. Moreover, their prediction horizon is effectively zero seconds since the classifier uses the transition segments, which are adjacent to the actual FoG onset.

Arami et al. [11] choose a distinctive subject-specific approach where they first predict the features ahead of time and classify the predicted features as "*no-FoG*" and "*FoG*". First, 75 different features are extracted using 4-second sliding windows with a step size of 0.5 seconds (leading to 87.5% overlap). Then, they train autoregressive models using the autocorrelation information of all features for each subject.

Finally, the predicted features are classified using SVM with radial basis function and a simple neural network. Their study achieves up to 94% correctly predicted FoG onsets with an expected prediction horizon of 1.72 seconds. However, considering that there is considerable overlap between windows (87.5%), there is a significant probability of data leakage between training and test set, which may artificially increase the performance of the proposed approach.

Yuan and Chakraborty [12] define the classes as "*WALK*" and "*ALARM*", where the alarm state indicates that the subject is either in pre-FoG or FoG state. Their model predicts every accelerometer sample. However, they aggregate the prediction of 30 samples ($\sim$0.5 s) into one prediction and call it a time unit. If more than half of the samples are classified as "*ALARM*", then the time unit is identified as "*ALARM*". The authors also state that once a patient is to be predicted in the "*ALARM*" state, their method requires two "*WALK*" predictions to switch back to the "*WALK*" state. Considering that their study is subject-specific, averaging across the best configurations from all patients, the reported prediction accuracies for 1, 3, and 5 seconds of prediction horizon are 94.7%, 82.9%, and 68.1% respectively. However, they define the correct prediction of an FoG episode as the first correct "*ALARM*" time unit that leads to the FoG onset. This may overestimate the performance of their approach considering that their method requires two "*WALK*" predictions to switch back to the "*WALK*" state. They also do not report the number of FoG events in the test set for each patient, which is important since some patients have only a few FoG instances ($<$5) in their test set.

The existing literature on FoG prediction uses varying definitions for prediction, predicted segments, accuracy, and prediction horizon. In this work, we propose a subject-independent approach with a precise definition of prediction accuracy and time. Overall, we achieve 89.42%, 88.34%, and 88.52% prediction accuracy with an average prediction time of 0.57, 1.26, and 1.42 seconds for $\tau_{pf} = 1\ s$, $\tau_{pf} = 3\ s$, and $\tau_{pf} = 5\ s$ respectively.

The Daphnet dataset [8], which is the only publicly available dataset for FoG analysis until the year 2021, has several limitations. The dataset is small, especially for analysis of rare events such as FoG. Besides, the dataset has many cases where there is no activity recorded. We attribute this to the laboratory environment in which the patients felt exhausted and overwhelmed during the experiments. In addition, although there are different types of FoG [10], the labeling in the dataset does not differentiate between these different types (i.e., labeling is done only according to the FoG presence). These limitations may degrade the learning process and overall performance of any proposed approach using this dataset. Therefore, there is a critical need for an FoG dataset which i) has a high number of FoG occurrences (obtained experimentally or generated synthetically using Generative Adversarial Networks [14]), ii) employs a long-term experimental procedure in a home environment, and iii) has detailed labeling both for the types of FoG and for different types of normal activities.

V. Conclusion

FoG affects the quality of life of PD patients, causing anxiety, loss of mobility, and falls. Studies show that external cues help patients to overcome the freezing state and resume walking. Hence, predicting an FoG episode can enable preemptive cueing and avoids the FoG episodes. This paper presented a subject-independent, generalizable deep learning approach, which can achieve up to 89% prediction accuracy with an average prediction time of $1.42\ s$ using a publicly available dataset. As future work, we plan to design a long-term experimental procedure that uses our own low-power wearable device prototype [15], and collect reliable and goal-oriented data for freezing of gait from Parkinson's Disease patients.

References

[1] S. Masiala, W. Huijbers, and M. Atzmueller, "Feature-set-engineering for detecting freezing of gait in parkinson's disease using deep recurrent neural networks," *arXiv preprint arXiv:1909.03428*, 2019.

[2] V. G. Torvi, A. Bhattacharya, and S. Chakraborty, "Deep domain adaptation to predict freezing of gait in patients with parkinson's disease," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1001–1006.

[3] N. Naghavi and E. Wade, "Prediction of freezing of gait in parkinson's disease using statistical inference and lower–limb acceleration data," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 27, no. 5, pp. 947–955, 2019.

[4] R. Deb *et al.*, "Trends in technology usage for parkinson's disease assessment: A systematic review," *medRxiv*, February 2021.

[5] M. Mancini *et al.*, "Measuring freezing of gait during daily-life: an open-source, wearable sensors approach," *J. Neuroeng. Rehabilitation*, vol. 18, no. 1, pp. 1–13, 2021.

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

[7] 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.

[8] M. Bachlin *et al.*, "Wearable assistant for parkinson's disease patients with the freezing of gait symptom," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, 2009.

[9] 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.

[10] 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.

[11] 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.

[12] A. M. Yuan and S. Chakraborty, "A study of deep learning for predicting freeze of gait in patients with parkinson's disease," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 1324–1331.

[13] N. Naghavi, A. Miller, and E. Wade, "Towards real-time prediction of freezing of gait in patients with parkinson's disease: addressing the class imbalance problem," *Sensors*, vol. 19, no. 18, p. 3898, 2019.

[14] S. An *et al.*, "Mgait: Model-based gait analysis using wearable bend and inertial sensors," *ACM Transactions on Internet of Things*, 2021. [Online]. Available: https://arxiv.org/abs/2102.11895

[15] G. Bhat, R. Deb, and U. Y. Ogras, "Openhealth: open-source platform for wearable health monitoring," *IEEE Design & Test*, vol. 36, no. 5, pp. 27–34, 2019.