Convolutional Neural Network for Freezing of Gait Detection Leveraging the Continuous Wavelet Transform on Lower Extremities Wearable Sensors Data

Bohan Shi[1,5], Shih Cheng Yen[1,2], Arthur Tay[1], Dawn M.L. Tan[3], Nicole S.Y. Chia[4] and W.L. Au[4]

Abstract— **Freezing of Gait is the most disabling gait disturbance in Parkinson's disease. For the past decade, there has been a growing interest in applying machine learning and deep learning models to wearable sensor data to detect Freezing of Gait episodes. In our study, we recruited sixty-seven Parkinson's disease patients who have been suffering from Freezing of Gait, and conducted two clinical assessments while the patients wore two wireless Inertial Measurement Units on their ankles. We converted the recorded time-series sensor data into continuous wavelet transform scalograms and trained a Convolutional Neural Network to detect the freezing episodes. The proposed model achieved a generalisation accuracy of 89.2% and a geometric mean of 88.8%.**

I. INTRODUCTION

Freezing of Gait (FoG) is the most disabling motor symptom in Parkinson's disease (PD), with 50% of PD patients suffering the symptom. The symptom severely deteriorates the mobility of PD subjects, restricts the independence of an individual, and often leads to falls that are frequently accompanied by fall-induced injuries [1, 2]. FoG is described as a clinical phenomenon of "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk [3]". FoG is usually observed in the advanced stage of PD, but research has found that it can long antedate the middle and late stages of PD [1]. FoG is paroxysmal and arbitrary in nature, but some internal and external factors can often induce a freezing episode, such as walking in a constrained space, making turns, dual-tasking, and stressful situations (trying to reach the destination or encountering an obstacle) [4, 5]. An FoG episode usually lasts a few seconds, but can sometimes extend to minutes [6].

The most widely accepted way to identify and evaluate FoG is through clinical observations and questionnaire-based evaluation that are often subjective and susceptible to assessment bias [7–9]. Employing wearable sensors and applying deep learning techniques to detect and access FoG objectively has been gaining increasing interest over the last decade [10], and this article takes a similar approach to investigate FoG in PD.

Currently, there are a few papers on using a deep learning (DL) approach to detect FoG. Camps and colleagues [11] performed the Fast Fourier Transform (FFT) on the data collected from a single inertial measurement unit (IMU) with three axial sensors (accelerometer, gyroscope, and magnetometer). Adopting a sliding window strategy and a shallow Convolutional Neural Network (CNN) architecture, their model was able to detect FoG with a geometric mean of 90%. Xia et al. [12] attached three accelerometers on three different body parts, and after pre-processing the data, used the time-series data directly as the input for a 5-layer CNN network. The group achieved 99% accuracy with a patient dependent model, and 80.7% for a patient independent model. Another recent study conducted by Ashour et al. [13] adopted the Long Short Term Memory network to detect FoG events. They collected accelerometer data from ten subjects, and developed a patient-dependent model with an accuracy of 83.38%.

All the studies mentioned above covered a small sample of PD subjects. Camps and colleagues recruited 21 subjects, while Xia's and Ashour's studies both had only ten participants. Moreover, some of the participants did not experience FoG during the data collection, which resulted in insufficient FoG data to assess the true generalisation capability of DL approaches. The results in these studies might have have been good as the abnormal gait signatures were less challenging to generalise because of their small sample size. However, the model might not be robust, and the features used might not be representative, to detect freezing events in a larger PD population because of the heterogeneous nature of PD.

II. METHODS

In this paper, we propose a two dimensional (2D) CNN architecture, using wavelet transform to convert the time series IMU data into a time-frequency representation that acts as the input of the CNN model. The goal is to implement the algorithm in a wearable device in the future to detect FoG episodes accurately so that clinicians can evaluate the frequency and duration of the FoG efficiently and objectively. Hence, the algorithm will also need to be computationally tractable.

*This work was supported in part by NUS-NNI 2016 Grant R263000C36133 and in part by NMRC/CISSP/2014/2015.

[1]B.H. Shi, S.C. Yen and A. Tay are with the Department of Electrical and Computer Engineering, National University of Singapore. Email: bohan.shi@u.nus.edu

[2]S.C. Yen is with the Innovation and Design Programme, Faculty of Engineering, National University of Singapore.

[3]D.M.L. Tan is with the Department of Physiotherapy, Singapore General Hospital.

[4]N.S.Y. Chia and W.L. Au are with the Department of Neurology, National Neuroscience Institute.

[5]B.H. Shi is with Activate Interactive Pte. Ltd.

A. Dataset

Sixty-seven PD patients were recruited through their primary physicians during their routine visits at the National Neuroscience Institute (NNI) outpatient clinics at Tan Tock Seng Hospital (TTSH) and Singapore General Hospital (SGH). Subject characteristics are summarised in Table I. The Freezing of Gait Questionnaire (FoG-Q) rating score is the only validated clinical assessment to evaluate the severity of FoG and overall gait performance for PD patients [14]. FoG-Q has excellent reliability and consistency, and is more sensitive in identifying freezers than the Unified Parkinson Disease Rating Scale (UPDRS) [7, 8]. In our study, the participants exhibited a mean FoG-Q score of 13.46 (SD = 4.96), which meant that all participants were PD freezers, and suffered from moderate to severe levels of freezing frequency, duration, as well as the severity of the gait disturbance.

TABLE I
DEMOGRAPHICS OF THE PARTICIPANTS.

Characteristics	PD patients (n = 67)
Age (Year)	69.58 ± 12.09
Gender	
Male	44 (34.33 %)
Female	23 (65.67 %)
Duration of Disease (Year)	6.02 ± 4.74
FOG-Q score	13.46 ± 4.96

Each subject performed the 7-metre Timed-Up-and-Go (7mTUG) task three times, while wearing three IMU sensors, each with three axial sensors (accelerometer, gyroscope, and magnetometer). Two of them were positioned at the lateral malleolus area of the ankles, while the third was positioned at the 7th cervical vertebra (C7) of the spine. PD patients with FoG often experience a phenomenon that is the opposite of the white coat effect, which is when blood pressure readings in patients increase in clinical settings compared to other environments. In contrast, previous findings showed that PD patients with FoG manifest fewer gait abnormalities in hospitals and research laboratories than in their daily lives during the off periods of their medication [4, 5, 9]. In order to simulate non-laboratory settings, the subjects were additionally asked to walk around the clinics at NNI and SGH while wearing the sensors to increase the number of freezing occurrences.

Video recordings were performed throughout the assessments to capture the freezing events. After each evaluation, three experienced health professionals with expertise in FoG independently labelled the videos frame-by-frame. The freezing events in the final data were annotated based on majority vote among the raters. The frame rate for the video recordings was 50 frames per second (fps), and the shortest possible FoG event labelled was 20 ms.

To the best of our knowledge, our dataset is the largest FoG-centric and IMU-based dataset in terms of the number of participants. The large sample size and FoG-focused data allowed us to provide more definitive evidence of the performance of DL models as DL techniques tend to overfit or demonstrate a lack of generalisation when trained and tested with a small dataset.

B. Data Insights

An exploratory data analysis (EDA) was performed to understand the characteristics of the dataset and the freezing events. The EDA was also an essential step in the model design to determine some of the hyper-parameters.

The duration of a FoG episode in our dataset ranged from less than 1 second to 126 seconds, and 88.6% of the FoG episodes were less than 10 seconds Fig. 1. The majority of FoG events (96%) were less than 30 seconds, which was consistent with previous findings [15].

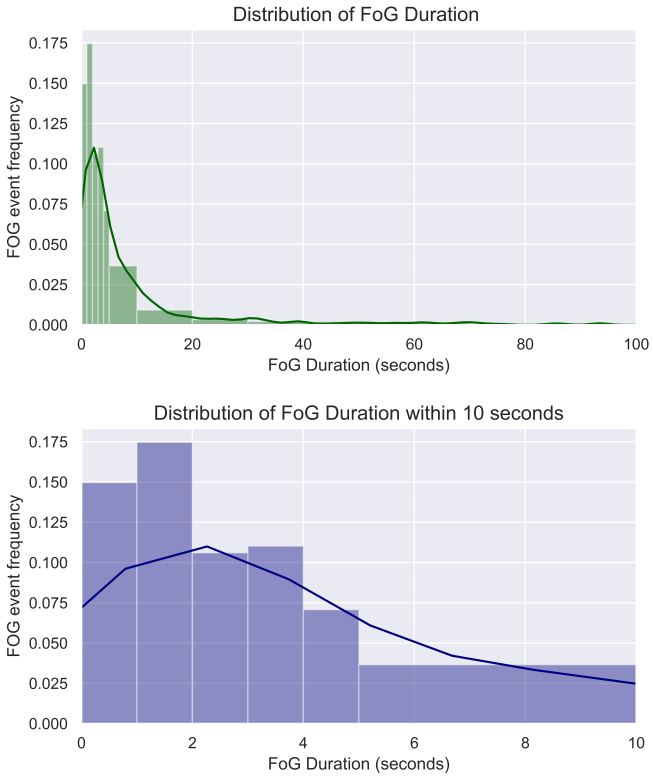

Fig. 1. Distribution of the freezing event duration.

From our dataset, FoG events predominantly occurred during turning (56.8%). 28% of freezing episodes happened while walking towards the destination, and 12.5% took place when the subjects tried to initiate movement (Fig. 2).

C. Model Development

1) Optimal Window Length: In our previous study [16], a window length of 4 seconds was determined to be the optimal window length for our dataset. In this study, we segmented the data into 4 seconds windows using the overlapping sliding window method. The overlapping size was 3 seconds; hence most of the windows contained 75% information from the previous windows and 25% new information. Adopting

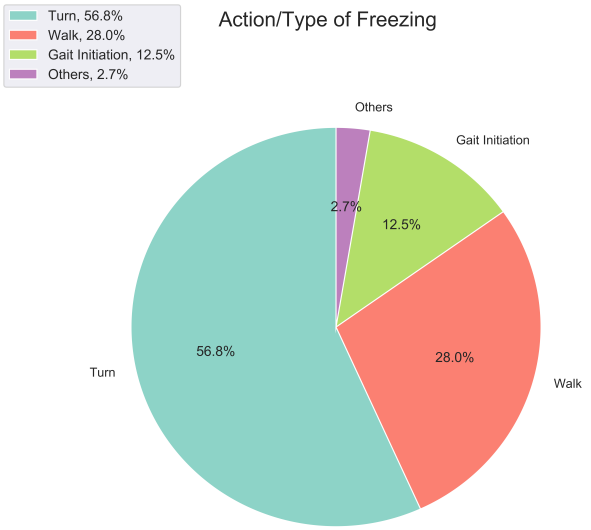

Fig. 2. Percentage of the FOG inducing actions.

Fig. 3. Class distribution of the training and test dataset.

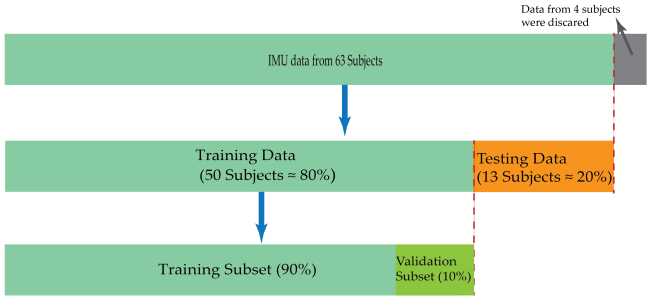

Fig. 4. Illustration of the data splitting strategy.

the overlapping sliding method allowed us to include the events during the transition period among windows [17].

2) Time-frequency Representation: DL is a promising method to detect anomalies in various types of biomedical signals. Currently, FoG research has been mainly focused on adopting the DL method with either time domain or frequency domain IMU signal as the input. However, there are shortcomings in using time or frequency domain inputs on their own. Hence, we implemented a novel method of using a time-frequency wavelet representation as the input.

We performed the Morlet continuous wavelet transform (CWT) on the segmented time-series windows to analyse the frequency components of the data that varied over time. The absolute coefficients of the CWT were then used to generate a scalogram for that window. The scalograms were used as the final image input of the CNN model.

3) Data Splitting and Validation Strategies: The 63 subjects in our database (data from 4 subjects were discarded as the patients' physical disabilities prevented them from completing the assessments) were split into training and test datasets with a ratio of 80% and 20%, respectively. As our study design was FoG-centric, and all the subjects had self-reported freezing episodes prior to the assessment, there were no significant imbalances in freezing and non-freezing data in both the training and test data sets (Fig. 3). Furthermore, our proposed model was a subject-independent model that learnt the common and generalised features from 50 subjects and predicted the FoG occurrences on the completely untouched test data of 13 PD subjects. During the training process, 10% of the training data was used as validation data (the data splitting strategy is illustrated in Fig. 4). The 20% test data was kept untouched during the training process. The model evaluation metrics were accuracy and geometric mean (G-Mean).

D. Model Architecture

As mentioned above, the computational cost was one of the primary considerations. Hence, the model training time was a way to estimate the computational cost. For example, implementing an Inception module could significantly improve model performance and allow the network to go deeper, but it required more than 72 hours to train and more memory to handle the data based on our experiments. On the contrary, the proposed model with FFT or CWT scalogram as the input took less than 2 hours to train. Hence, complex model structures and recent sophisticated DL techniques, such as Inception module [18] and Residual blocks [19], were not used in our model at this stage. The model was trained on a Nvidia Tesla V100 GPU in the High-Performance Computing Cluster at the National University of Singapore.

A 2D convolutional neural network (CNN) model was developed for this study, and the overview of the model structure is presented in Fig. 5. This model employed a cross-entropy loss and a sigmoid activation function in the last layer for binary classification. This model avoided using more advanced loss and activation functions to ensure that the model could be easily compiled into a micro-controller or a mobile device in the near future. From the empirical results, we chose to use Adam as the model optimiser.

1) Convolution Layers: There are a total of 8 convolution layers in our network. In the first 2D convolution block, two 5×5 convolutional kernel filters were applied to reduce the dimensions of the input and the overall computational cost. The large filter size increased the receptive field size and

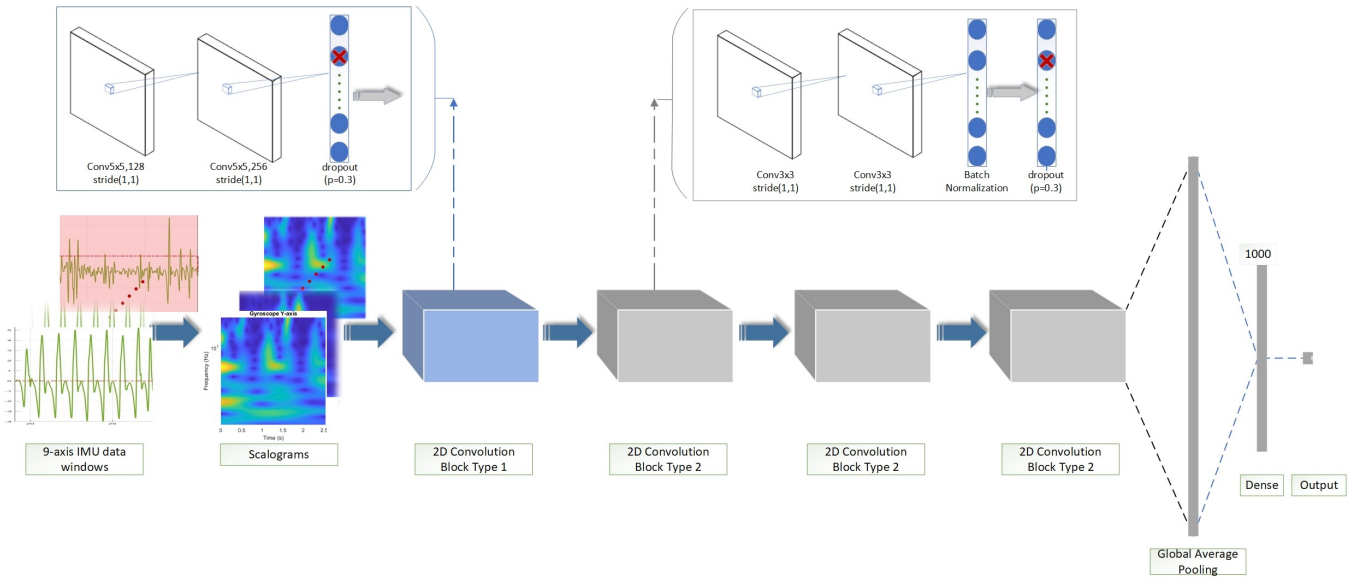

Fig. 5. The architecture of the proposed model.

captured the more generic features. In the convolution block type 2, smaller 3×3 kernel filters were employed to extract the local features.

Stride size of 1 was used in all convolutional layers. No padding was added to any of the layers.

2) Global Average Pooling Layer: At the end of the network, we utilised a Global Average Pooling (GAP) layer followed by a fully connected layer. This network structure was inspired by the ResNet-50 [19] and it helped to reduce computational cost and overfitting [20].

E. Model Regularisation

1) Layer Regularisation: In each convolution layer, the L2 kernel and bias regulariser were employed to penalise the weight matrices of the neurons during the optimisation. Regularisation was crucial in reducing the risk of overfitting.

2) Early Stopping Condition: In order to further reduce overfitting, the early stopping strategy was applied, and the training would stop when the validation loss from the validation dataset showed no significant improvement over a set of consecutive epochs.

3) Dropout: Dropout regularisation was implemented after each convolution block to prevent overfitting. We used a dropout rate of 0.3.

III. RESULTS

A. Evaluation Metrics

In order to evaluate the performance of the FoG detection algorithm, the accuracy (i.e. proportion of correct detection) was calculated. However, accuracy alone is not an appropriate measurement to assess the algorithm's efficiency as it performs poorly for unbalanced data because the predictions are biased towards the majority class [21].

Another three widely used evaluation metrics for medical data, sensitivity, specificity, and the G-Mean, were employed to provide in-depth analysis of the algorithm's performance. Sensitivity is the probability of the true positive detection rate, which determines how accurate the algorithm can detect the freezing condition, and specificity is the true negative rate that determines the ability of the algorithm to reject the events without FoG [22, 23]. The G-Mean is the root of the product of the sensitivity and specificity, which is a performance measurement to balance the result among different classes [21, 24]. In our model, the highest G-Mean was used to determine the optimal model and the best combination of hyper-parameters.

The formulae for all the metrics are given by:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

$$G-Mean = \sqrt{Sensitivity * Specificity} \tag{4}$$

where :

- True positive (***TP***) : The number of occurrences when the event was a **FoG** episode and the model predicted the event as a **FoG** episode.
- False positive (***FP***) : The number of occurrences when the event was a **Non-FoG** episode and the model predicted the event as a **FoG** episode.
- True negative (***TN***) : The number of occurrences when the event was a **Non-FoG** episode and the model predicted the event as a **Non-FoG** episode.
- False negative (***FN***) : The number of occurrences when the event was a **FoG** episode and the model predicted the event as a **Non-FoG** episode.

B. Model Evaluation

In order to evaluate the performance of our model, we first compared the performance of the proposed model with a simple classic Hidden Markov Model (HMM) using Viterbi's algorithm. The HMM performed poorly in identifying both the FOG and the Non-FoG episodes as it was always biased towards one of the classes.

Next, we reconstructed all three above-mentioned state-of-the-art models [11–13], and summarised the results into Table II. Although we used the same model structures as described in the literature, the performance of the reconstructed 1D CNN (FFT) and LSTM model significantly deteriorated when tested on a larger population of PD patients with FoG. Interestingly, Xia's 1D CNN model using our time series raw data performed better than the results reported in their original article.

Our proposed model using the wavelet input achieved an accuracy of 89.2% and a G-mean of 88.8% using our FoG dataset. In addition, applying the same model structure on FFT data (i.e. using only frequency domain information) also outperformed the state-of-the-art models, which indicated that our proposed model was more robust and had a model structure that was more suited to detecting FoG in IMU data.

TABLE II
MODEL PERFORMANCE COMPARISON WITH RELATED METHODS

Model	Accuracy	Sensitivity	Specificity	G-Mean
Hidden Markov model (Viterbi Algorithm) - 1D CNN - Time Series Raw Data	47.2	42.3	56.2	34.6
Reconstructed Xia's Model - 1D CNN - Time Series Raw Data	83.2	80.4	88.5	84.3
Reconstructed Camps's Model - 1D CNN - FFT	78.9	77.4	80.4	78.9
Reconstructed Ashour's Model - LSTM - Time Series Raw Data	74.4	77.3	69	73.1
Proposed Model - 1D CNN - Time Series Raw Data	74.3	80.2	66.3	73
Proposed Model - 1D CNN - FFT	83.8	**83.7**	83.8	83.7
Proposed Model - 2D CNN - Wavelet	**89.2**	82.1	**96**	**88.8**

IV. DISCUSSION

This paper presented a novel method to convert time-series IMU data into CWT scalogram images and using a 2D CNN model to detect the onset of freezing episodes from the input images. Since computational devices, such as GPUs and Google's TPUs, are becoming more powerful and are optimised to manipulate images, using the scalograms as the image input helps to achieve significantly better performance than using only time-domain or frequency-domain information.

In this study, we chose to train a generic model that might not perform as well as models trained on individual subjects because a large number of the subjects in our study exhibited fairly heterogeneous and unique gait characteristics, which increased the difficulty of finding features and patterns common to all subjects. However, the generic model is a better option when implementing the model into an embedded system because it will work for the majority of patients and scenarios. The performance can be improved further if the patient continues to use the system, and an online learning algorithm can be applied to learn personalised features and fine-tune the parameters to achieve optimal performance.

REFERENCES

[1] N. Giladi, T. A. Treves, E. S. Simon, H. Shabtai, Y. Orlov, B. Kandinov, D. Paleacu, and A. D. Korczyn, "Freezing of gait in patients with advanced Parkinson's disease," *Journal of Neural Transmission*, vol. 108, no. 1, pp. 53–61, 2001.

[2] N. Giladi, M. P. McDermott, S. Fahn, S. Przedborski, J. Jankovic, M. Stern, and C. Tanner, "Freezing of gait in PD: Prospective assessment in the DATATOP cohort," *Neurology*, vol. 56, no. 12, pp. 1712–1721, jun 2001.

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

[4] A. Nieuwboer and N. Giladi, "The challenge of evaluating freezing of gait in patients with Parkinson's disease," *British Journal of Neurosurgery*, vol. 22, no. SUPPL. 1, pp. S16–S18, jan 2008.

[5] Y. Okuma and N. Yanagisawa, "The clinical spectrum of freezing of gait in Parkinson's disease," *Movement Disorders*, vol. 23, no. SUPPL. 2, pp. S426–30, 2008.

[6] A. H. Snijders, M. J. Nijkrake, M. Bakker, M. Munneke, C. Wind, and B. R. Bloem, "Clinimetrics of freezing of gait," *Movement Disorders*, vol. 23, no. SUPPL. 2, pp. S468–S474, 2008.

[7] Giladi, Shabtai, Simon, Biran, Tal, and Korczyn, "Construction of freezing of gait questionnaire for patients with Parkinsonism." *Parkinsonism & related disorders*, vol. 6, no. 3, pp. 165–170, jul 2000.

[8] N. Giladi, J. Tal, T. Azulay, O. Rascol, D. J. Brooks, E. Melamed, W. Oertel, W. H. Poewe, F. Stocchi, and E. Tolosa, "Validation of the freezing of gait questionnaire in patients with Parkinson's disease," *Movement Disorders*, vol. 24, no. 5, pp. 655–661, apr 2009.

[9] C. Barthel, E. Mallia, B. Debû, B. R. Bloem, and M. U. Ferraye, "The Practicalities of Assessing Freezing of Gait," *Journal of Parkinson's Disease*, vol. 6, no. 4, pp. 667–674, 2016.

[10] M. Gilat, A. Lígia Silva de Lima, B. R. Bloem, J. M. Shine, J. Nonnekes, and S. J. Lewis, "Freezing of gait: Promising avenues for future treatment," *Parkinsonism and Related Disorders*, vol. 52, pp. 7–16, jul 2018.

[11] J. Camps, A. Samà, M. Martín, D. Rodríguez-Martín, C. Pérez-López, J. M. Moreno Arostegui, J. Cabestany, A. Català, S. Alcaine, B. Mestre, A. Prats, M. C. Crespo-Maraver, T. J. Counihan, P. Browne, L. R. Quinlan, G. Ó. Laighin, D. Sweeney, H. Lewy, G. Vainstein, A. Costa, R. Annicchiarico, À. Bayés, and A. Rodríguez-Molinero, "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, jan 2018.

[12] Y. Xia, J. Zhang, Q. Ye, N. Cheng, Y. Lu, and D. Zhang, "Evaluation of deep convolutional neural networks for detection of freezing of gait in Parkinson's disease patients," *Biomedical Signal Processing and Control*, vol. 46, pp. 221–230, sep 2018.

[13] A. S. Ashour, A. El-Attar, N. Dey, H. A. El-Kader, and M. M. Abd El-Naby, "Long short term memory based patient-dependent model for FOG detection in Parkinson's disease," *Pattern Recognition Letters*, vol. 131, pp. 23–29, 2020.

[14] M. H. Nilsson, G.-M. Hariz, K. Wictorin, M. Miller, L. Forsgren, and P. Hagell, "Development and testing of a self administered version of the Freezing of Gait Questionnaire," *BMC Neurology*, vol. 10, no. 1, p. 85, dec 2010.

[15] 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, jan 2008.

[16] V. Mikos, C. H. Heng, A. Tay, N. S. Y. Chia, K. M. L. Koh, D. M. L. Tan, and W. L. Au, "Optimal window lengths, features and subsets thereof for freezing of gait classification," in *ICIIBMS 2017 - 2nd International Conference on Intelligent Informatics and Biomedical Sciences*, vol. 2018-Janua. Institute of Electrical and Electronics Engineers Inc., feb 2018, pp. 1–8.

[17] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors," *Sensors (Switzerland)*, vol. 19, no. 22, nov 2019.
[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, 2015, pp. 1–9.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem. IEEE Computer Society, dec 2016, pp. 770–778.
[20] M. Lin, Q. Chen, and S. Yan, "Network in network," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, dec 2014.
[21] J. S. Akosa, "Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data," in *SAS Global Forum*, vol. 942, 2017, pp. 1–12.
[22] D. G. Altman and J. M. Bland, "Statistics Notes: Diagnostic tests 1: Sensitivity and specificity," *Bmj*, vol. 308, no. 6943, p. 1552, jun 1994.
[23] T. W. Loong, "Understanding sensitivity and specificity with the right side of the brain," *British Medical Journal*, vol. 327, no. 7417, pp. 716–719, sep 2003.
[24] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, aug 2018.