Detection of Freezing of Gait using Convolutional Neural Networks and Data from Lower Limb Motion Sensors

Bohan Shi, Arthur Tay, W.L. Au, Dawn M.L. Tan, Nicole S.Y. Chia and Shih-Cheng Yen

***Abstract*—Parkinson's disease (PD) is a chronic, non-reversible neurodegenerative disorder, and freezing of gait (FOG) is one of the most disabling symptoms in PD as it is often the leading cause of falls and injuries that drastically reduces patients' quality of life. In order to monitor continuously and objectively PD patients who suffer from FOG and enable the possibility of on-demand cueing assistance, a sensor-based FOG detection solution can help clinicians manage the disease and help patients overcome freezing episodes. Many recent studies have leveraged deep learning models to detect FOG using signals extracted from inertial measurement unit (IMU) devices. Usually, the latent features and patterns of FOG are discovered from either the time or frequency domain. In this study, we investigated the use of the time-frequency domain by applying the Continuous Wavelet Transform to signals from IMUs placed on the lower limbs of 63 PD patients who suffered from FOG. We built convolutional neural networks to detect the FOG occurrences, and employed the Bayesian Optimisation approach to obtain the hyper-parameters. The results showed that the proposed subject-independent model was able to achieve a geometric mean of 90.7% and a F1 score of 91.5%.**

***Index Terms*—Continuous wavelet transforms, convolutional neural network, freezing of gait detection, inertial measurement unit, parkinson's disease, wearable sensor.**

I. INTRODUCTION

PARKINSON'S disease (PD) is a progressive and non-reversible neurodegenerative disorder with predominantly motor impairments, such as tremor at rest, rigidity, bradykinesia, impairment of posture, and freezing of gait (FOG) [1]. Globally, PD affected about 6.1 million individuals in 2016 [2]. The estimates of the prevalence of PD range from 35.8 to 12,500 per 100,000 persons [3, 4]. The prevalence significantly increases with age, and studies have shown that the prevalence of PD for people above 65 years old is between 1.3% to 3% of that age group [5, 6]. Moreover, the number of patients diagnosed with PD increased by 31.6% from 2005 to 2015 [7], and the age-standardised prevalence of PD also increased by 21.7% from 1990 to 2016 [2]. PD is the fastest-growing neurological disease, and has become the most challenging health issue for ageing populations.

PD's pathological characteristics include the loss of dopaminergic neurons in the substantia nigra pars compacta and the accumulation of intracellular protein (α-synuclein containing Lewy bodies) inside nerve cells that lead to cell death [8, 9]. The aetiology of PD is not well understood, but past studies have revealed a moderate correlation between PD causality and the role of environmental and genetic factors [10, 11]. The abnormal degeneration of dopaminergic neurons and the death of brain cells obstruct the smooth control and coordination of voluntary movements throughout the body. When 80% of dopamine-producing cells are damaged, the cardinal motor symptoms in PD start to emerge and significantly impair the performance of simple daily tasks such as walking and static standing. These motor impairments significantly reduce a patient's quality of life, but one of the most disabling symptoms in PD is FOG, with half of all PD patients suffering the symptom [12]. The clinical definition of FOG is "brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk" [13].

FOG is more prevalent among PD patients in the advanced stages, but a study has found that it could be found in the early stage of PD [14]. It severely deteriorates PD subjects' mobility and restricts an individual's independence, often leading to falls, which are frequently associated with serious injuries. FOG is a paroxysmal and unpredictable motor anomaly, but some internal and external factors were found to induce a freezing episode, such as walking in a confined space, turning, dual-tasking, and stressful situations (e.g. inability to reach the destination) [15, 16]. FOG episodes usually continue for a few seconds, but can occasionally last for several minutes [17].

The primary symptomatic treatment for FOG is medication, and the most widely used medication is levodopa (L-dopa) therapy, which has demonstrated a positive effect on improving the dopamine-responsive type of FOG [18, 19]. Other types of treatments that tackle motor symptoms have not shown significant evidence to improve FOG, such as botulinum toxin injections and amantadine [20, 21].

Surgical treatment, such as deep brain stimulation (DBS) in

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

Bohan Shi and Arthur Tay are with the Department of Electrical and Computer Engineering, National University of Singapore. `Email: bohan.shi@u.nus.edu`

Bohan Shi is with Activate Interactive Pte. Ltd.

W.L. Au and Nicole S.Y. Chia are with the Department of Neurology, National Neuroscience Institute.

Dawn M.L. Tan is with the Department of Physiotherapy, Singapore General Hospital.

Shih-Cheng Yen is with the Innovation and Design Programme, Faculty of Engineering, National University of Singapore. `Email: shihcheng@nus.edu.sg`

the subthalamic nucleus or the globus pallidus internus area, is another approach to ease the burden of FOG [22-24]. However, DBS is not suitable for all PD patients as it is a highly invasive treatment that carries all the risks of major brain surgery [25].

Aside from treatments aimed at reducing the onset of freezing, there are also different approaches to mediate the consequences of freezing. Cueing is a movement strategy technique that supplements medication in improving the overall functional mobility of patients by assisting patients with PD to overcome FOG episodes and prevent falls [20, 26, 27]. The cueing techniques can be achieved in the form of rhythmic auditory cueing, visual assistance cues, and sensory cues. The cueing techniques's neural mechanisms are not well understood, but studies have shown that the disruptions in sensory-motor interactions might cause deficits in internal cueing for movements and movement initiation [15, 28]. The role that external cueing plays is to bypass the dysfunctional basal ganglia network and compensate for the loss of internal rhythms that results in impaired automaticity [29].

In order to deliver on-demand cueing assistance to PD patients at the most opportune moment to overcome the gait disturbance, wearable devices/systems have been proposed to monitor gait performance continuously and detect FOG events [30-32]. Technological advancements in wearable devices with small form factor single-board computers have made such a system feasible in recent years. Currently, the majority of studies have chosen to use inertial measurement unit (IMU) sensors (accelerometer and/or gyroscope) as they provide relatively accurate measurements and can be worn by patients for an extended period time without interrupting the walking pattern and normal activities of daily living. Other physiological wearable sensors, such as blood pressure and heart rate sensors and those that measure electrocardiograms and electromyograms, were also used in some studies to identify the physiological changes before the onset of FOG [33, 34]. Another way of detecting FOG is to use vision-based techniques to determine gait abnormalities [35, 36]. However, the results of vision-based methods have so far been worse than those of IMU-based approaches (discussed below), with 82.1% being the highest detection accuracy reported so far [36]. The privacy and the security of the videos will also raise barriers to the adoption of these approaches.

In order to detect FOG events, conventional machine learning approaches have required a substantial amount of domain-related expertise and tremendous efforts in pre-processing and feature engineering on the data. However, no single feature or a combination of features have been shown to detect freezing episode perfectly due to the symptom's complexity and heterogeneity. Hence, researchers have recently started to adopt deep learning (DL) models to detect FOG without generating handcrafted features. The DL models were shown to be able to learn novel and robust features of the sensor data without relying on domain experts to specify disease phenotypes [37, 38]. Furthermore, the DL models worked well with large-scale real-world data, and has been proposed as a means to improve clinical decision making by providing data-driven evidence [38, 39].

Convolutional neural networks (CNN) are a type of neural network in DL, and they are the most popular model architecture for image classification. In recent years, its practical usage has extended from identifying objects from daily life, such as dogs and cats, to discovering symptoms, identifying diseases, and predicting biological structure [37, 39, 40]. In contrast to conventional machine learning approaches, CNN require minimum pre-processing, and they capture complex and heterogeneous features from data without extensive domain knowledge. It has naturally become a favoured tool to study clinical data.

In 2018, Camps et al. [41] introduced an 8-layer 1D CNN to perform FOG detection. The model was trained using data from a group of patients in the REMPARK database where 21 PD patients wore a nine-channel tri-axial IMU (accelerometer, gyroscope and magnetometer) on the left side of the waist while performing several walking tests at home. The data collected was segmented into 2.56 seconds windows, and every window of data was transformed into the frequency domain using the short-time Fast Fourier Transform (FFT). The magnitude of each FFT window was combined with the previous window of FFT data to form a single sample. The authors further processed the data with data augmentation in order to address the data imbalance issue. The model outperformed the previous shallow ML models and achieved 91.9% sensitivity, 89.5% specificity and 90.6% geometric mean.

Later that year, Xia et al. [42] presented a simpler 5-layer CNN. The data was collected from ten subjects with accelerometers placed on three different parts of their bodies. Outlier removal and data segmentation were performed, and the raw accelerometer data with a window size of four seconds was used as the model's input. The proposed CNN model was tested with two schemes. The patient-dependent model was able to detect FOG with an accuracy of 99%. The patient-independent model was trained using the leave-one-patient-out validation and achieved an accuracy of 80.7%.

In 2020, Sigcha et al. [43] proposed to use a combination of CNN and a Long Short-Term Memory (CNN-LSTM) deep neural network model on data collected from one single accelerometer that was placed on the participant's waist. The study involved 21 participants, and the data collection was conducted at the patients' home to increase the occurrence of the FOG episodes. The authors found that stacking three previous spectral windows on the current window as the input of the CNN-LSTM model provided the best result, and they achieved mean sensitivity, specificity, and geometric mean all equal to 87.1% using the leave-one-subject-out validation.

More recently, Bikias et al. [44] proposed another CNN model that attempted to investigate the feasibility of using a wrist-based IMU sensor to detect FOG episodes. The study used the data from the CuPiD IMU dataset [45], which contained data from 18 patients. The IMU consisted of an accelerometer and a gyroscope, with a sampling rate of 128 Hz. A simple network with two CNN layers was used, and evaluation with 10-fold cross-validation achieved a mean specificity of 90% and sensitivity of 86%.

Although different DL models were used, the past studies were able to achieve excellent detection accuracy. However, these models were trained and tested with a limited number of subjects, and the gait patterns for PD patients with FOG differ significantly and can even differ significantly within a patient as the disease progresses. The performance of models might deteriorate if employed on subjects with gait characteristics different from those from subjects the models were trained on. In addition, while past studies have used either time-domain or frequency-domain data as the input, we explored in this study the possibility of leveraging time-frequency representations as the input. The use of the two-dimensional time-frequency representation as the input also demonstrated the feasibility of using computer vision techniques and architectures to detect FOG. This approach lays the groundwork for future FOG detection research to adopt and extend innovative solutions from the computer vision literature.

Inspired by the latest research in CNN models, we investigated a novel FOG detection method using CNN. The model was optimised using the sequential model-based Bayesian optimisation method. In order to evaluate the proposed model's performance, we first compared the proposed model with seven popular machine learning algorithms: 1) k-nearest neighbours (KNN), 2) Linear Regression (LR), 3) Decision Tree (DT), 4) Random Forest (RF), 5) Support Vector Machine (SVM) with linear kernels, 6) SVM with radial basis function (RBF) kernels, and 7) Extreme Gradient Boosting (XGBoost). Subsequently, the proposed model was compared against the state-of-the-art DL models: Xia's model, Camps's model, and Bikias's model. The preliminary design and results that were reported in a previous conference paper [46] were also reconstructed and examined.

II. MATERIALS AND METHODS

A. Data

Sixty-seven PD subjects who suffered from different degrees of FOG in the past agreed to participate in our study. All subjects were selected during their regular check-up and recommended by their respective neurologists from the local hospitals. The study was approved by the SingHealth Centralised Institutional Review Board of Singapore on 28th September 2016 (CIRB Ref: 2016/2743).

Each subject was instructed to perform two types of walking tests in the hospital under the observation of a physiotherapist and several researchers. The first one was the standard 7-metre Timed-Up-and-Go (7mTUG) test, and each subject conducted the test three times. As mentioned in our previous research [46], FOG subjects (freezers) often experience "white-coat syndrome" where they do not experience FOG when performing walking tests with their neurologists or physiotherapists in a hospital or a laboratory setting. On the other hand, another widely adopted FOG data collection method, home-based data collection, has a higher chance of simulating the patient's daily routine and inducing more FOG episodes. However, it also brings up significant concerns about the patient's safety and privacy. Hence, in order to reduce the "white-coat syndrome", we asked the subjects who did not experience any FOG episodes during the 7mTUG to walk freely in the clinic as the second FOG-inducing test in order to capture more occurrences of freezing. The subjects wore an IMU around the lateral malleolus area of each ankle, and a third IMU near the 7th cervical (C7) vertebra during both tests. However, the data from the third sensor was not used in the FOG detection model as it was used only to analyse the patients' sitting and standing posture and stability. Each IMU used in this study was composed of an accelerometer, gyroscope, and magnetometer, and was developed in-house at the National University of Singapore [47, 48]. The IMU data was then transmitted wirelessly over a Bluetooth connection and saved into an iPad app at a sampling rate of 50 Hz.

TABLE I: Demographics of the subjects. **PD** stands for Parkinson's Disease. **Duration of Disease** refers to the time interval between the date of PD diagnosis and the first assessment in this study. **FOG-Q** is the Freezing of Gait Questionnaire, which is currently the only validated measure to evaluate FOG subjectively [49]. **7mTUG** is the 7-meter Timed-Up-and-Go test, which is a widely used functional mobility test.

Characteristics	**PD patients (n = 63)**
Age (Year)	69.35 ± 12.4
Gender	
Male	41 (65.08%)
Female	22 (34.92%)
Duration of Disease (Years)	6.21 ± 4.83
FOG-Q Total Score	13.56 ± 4.62
Average 7mTUG (Seconds)	71.82 ± 77.43

Videos were recorded during the tests, and three experienced physiotherapists independently reviewed the videos after the tests in order to mark the FOG events. The final FOG labels were decided based on the decision of the majority. We ended up with a total of 486 FOG events in our database.

Within the sixty-seven subjects we recruited, data from four subjects who were unable to complete the tests, or encountered data loss during the tests, was excluded. Three subjects who suffered from FOG did not manifest any signs of freezing during the tests. Seven subjects demonstrated minimal or insignificant periods of FOG. However, recordings from these ten subjects were kept as examples of non-FOG gait. Four subjects faced significant challenges completing the test without walking aids, such as a walking frame or cane, so their data (which included periods of FOG) was collected with walking aids and used in our analyses. The rationale was that the use of walking aids in some environments was common in PD patients, so including this type of data would allow us to build a more robust system that could be used by PD patients to detect FOG in different environments.

The demographics of the sixty-three subjects who completed all the tests are shown in Table I.

B. Signal Pre-Processing

1) Data Filtering: Signal pre-processing, such as filtering, is usually required for classification problems using time series data. However, DL models often require minimal filtering, and introducing noise into the input data is often used in DL models to reduce generalisation error and improve model robustness [50]. Hence, in our experiments, the data was not filtered at all when training the CNN model.

When testing with the other machine learning models, the accelerometer signals were filtered with a 4th-order Butterworth band-pass filter. The cut-off frequencies were 0.2 Hz and 15 Hz. A 4th-order Butterworth low-pass filter was applied to the gyroscope signal with a cut-off frequency of 10 Hz.

2) Continuous Wavelet Transform: Previous DL FOG detection algorithms used either the time-domain raw data or frequency components obtained from the Fast Fourier Transform (FFT). Spatial and temporal domain features, such as cadence, step duration, velocity, stride length, FOG Criterion, and gait cycle duration (stride time, stance time and swing time), etc. have been shown to be effective in detecting FOG [51-54]. Frequency domain features, such as power in the freezing band (FOG episodes often occur between 3 to 8 Hz) and locomotor band (volitional activities have dominant frequencies that range from 0.5 to 3 Hz), have also been shown to be sensitive predictors in FOG detection, and can only be discovered in the frequency domain [51, 55, 56]. For example, Figure 1a shows non-FOG gait data from one of our subjects reflected in the orderly and periodic changes in the vertical axis of both the accelerometer and gyroscope signals. Figure 1b shows that most of the frequency components were distributed below 3 Hz. Figure 2a shows gait data from another one of our PD subjects suffering from a FOG episode. The data from the vertical axis of the accelerometer and gyroscope were much more random and distorted. Figure 2b shows that most of the frequency components were distributed between 3 Hz to 8 Hz.

However, based on observations of our own data and the results of past studies [32, 57], these patterns in the time or frequency domains were not always distinguishable for all patients. For the same patient, his/her FOG patterns also varied over time. This heterogeneity complicates the autonomous detection of FOG. Therefore, applying either raw time-domain data or transformed FFT data as the inputs for a CNN model can potentially lead to some critical features missing from the analysis and classification. This motivated us to make use of the wavelet transform, which would capture patterns in the time-frequency domain, to provide richer inputs to the CNN model. In Figure 1c and Figure 2c, data in the same window was transformed into the time-frequency plane using a continuous wavelet transform (CWT). The scalograms contained all the key information from the time and frequency domain analyses. Furthermore, they provided considerable additional insights into the non-stationarity of the IMU signals and the time specificity of power increases in different frequency bands.

As the CWT can provide a finer discretised scale for analysis than the discrete wavelet transform, we used the absolute value of the coefficients obtained from the Morlet mother wavelet as the input to our CNN model. The mathematical representation of the CWT for a time-series data $x(t)$ with respect to a mother wavelet ψ, is defined as :

$$CWT(s,\tau,\psi) = \int_{-\inf}^{\inf} x(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-\tau}{s})dt, s \in R^{+*}, \tau \in R \tag{1}$$

where s and τ are the scale and translation factors, respectively, used to transform the mother wavelet ψ, and $*$denotes the complex conjugate.

C. Machine Learning Models

In order to compare the performance of DL models to machine learning models, 67 features (F1 to F67, described below) that had been used in past FOG detection studies were trained using seven popular machine learning models. The features were extracted from the data in 1-second windows.

1) Frequency Domain Features: Moore et al. [58] and Delval et al. [51] pointed out that freezing of gait often occurred in the range of 3 to 8 Hz in the frequency spectra for vertical leg acceleration, while normal gait happened in the 0.5 to 3 Hz range. Therefore, we selected five widely used groups of frequency domain features (F1 to F5) described in Table II.

TABLE II: Short descriptions of the selected frequency domain features for FOG detection.

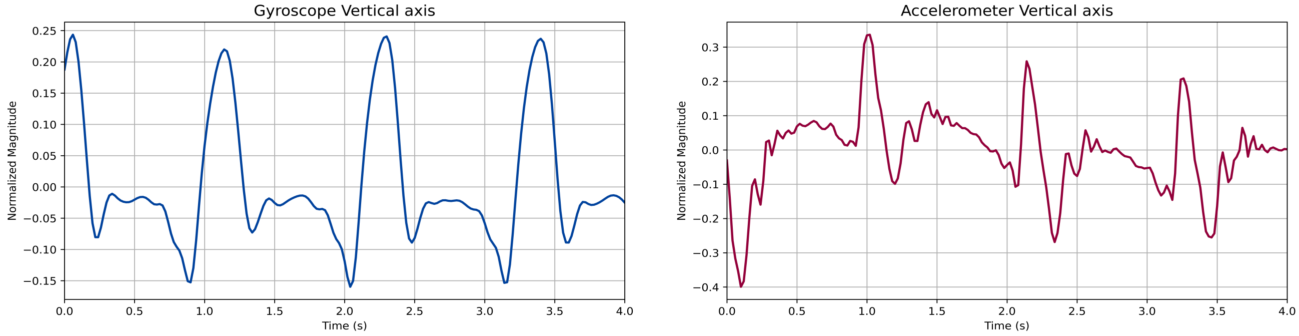
Feature Index	**Feature Name**	**Feature Description**
F1	Freeze Index (FI)	Power ratio in the freezing band (3 – 8 Hz) and the locomotor band (0.5 – 3 Hz) .
F2	Total Power	Total power in the freezing band and the locomotor band.
F3	Average Acceleration Energy	Average energy for all three axes of acceleration (X, Y, and Z).
F4	Sum of PSD	Sum of the power spectral density for vertical acceleration.
F5	Peak Frequency	Peak frequency component for vertical acceleration.

2) Entropy Features: Sample Entropy (SampEn) is an improved version of approximate entropy and is often used to evaluate a time-series data's complexity or regularity [59]. Human gait is a form of a dynamical system, and FOG is a sudden and episodic abnormality in the gait system [60, 61]. SampEn can be an effective method to analyse the regularity or stability of human gait where a higher SampEn value indicates a higher level of irregularity or randomness.

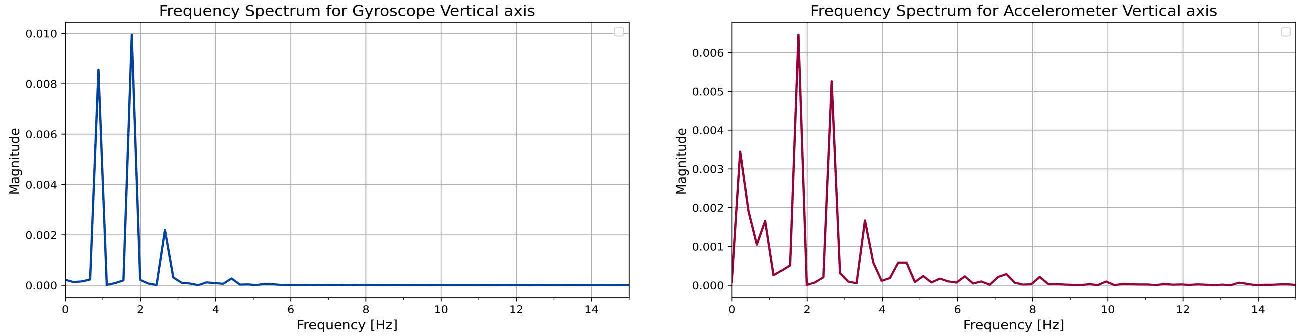
Sample Entropy calculation was performed for both accelerometer and gyroscope signals in the X, Y, and Z axes (F6 to F11), as well as the signals' magnitude (F12 to F13). This feature extraction process was performed for each window of data to form the feature vector.

3) Wavelet Features: The wavelet transform is another popular method to analyse time-series signals. For example, El-Attar et al. [62] demonstrated that features extracted from applying discrete wavelet transform (DWT) on accelerometer signals yielded a robust FOG classification model.

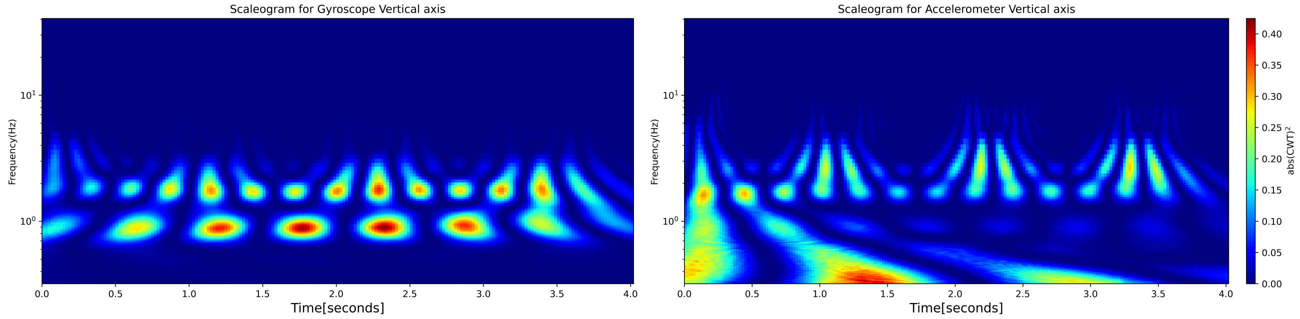
Therefore, we analysed the accelerometer signals using Daubechies orthogonal wavelets (db1) to extract 25 approximation (cA) and 25 detail coefficients (cD) from the X, Y, and Z axes, as well as the magnitude of the signals. Statistical



(a) Non-FOG gait in Time Domain



(b) Non-FOG gait in Frequency Domain



(c) Non-FOG gait in Time-Frequency Domain

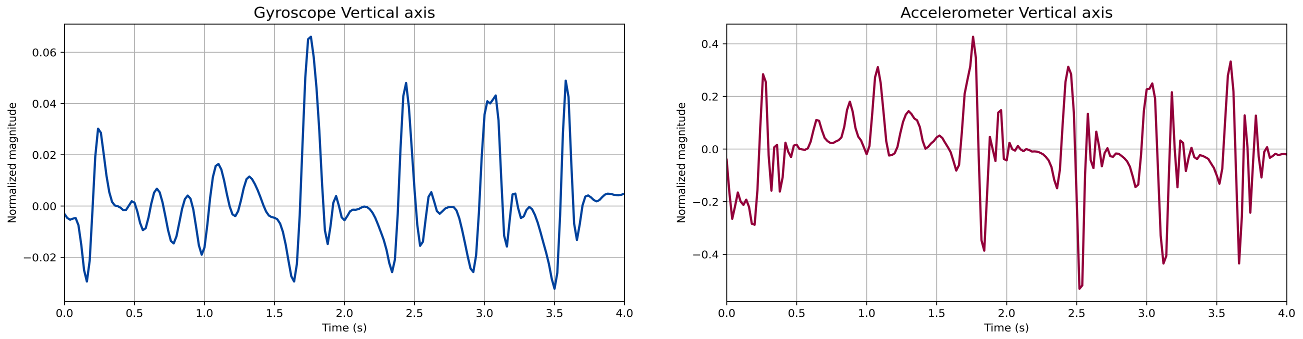
Fig. 1: Non-FOG gait signals visualised in time, frequency, and time-frequency domains.

features of the cA and cD for each of the three axes, such as 1) mean (μ), 2) standard deviation (σ), 3) median, 4) skewness (skew), 5) kurtosis (kurt), 6) minimum (min), 7) maximum (max), 8) interquartile range (iqr), and 9) median absolute deviation (MAD) were calculated for 1 second of the data to form the 27-dimension wavelet feature vector (cA features: F14 to F40, cD features: F41 to F67).

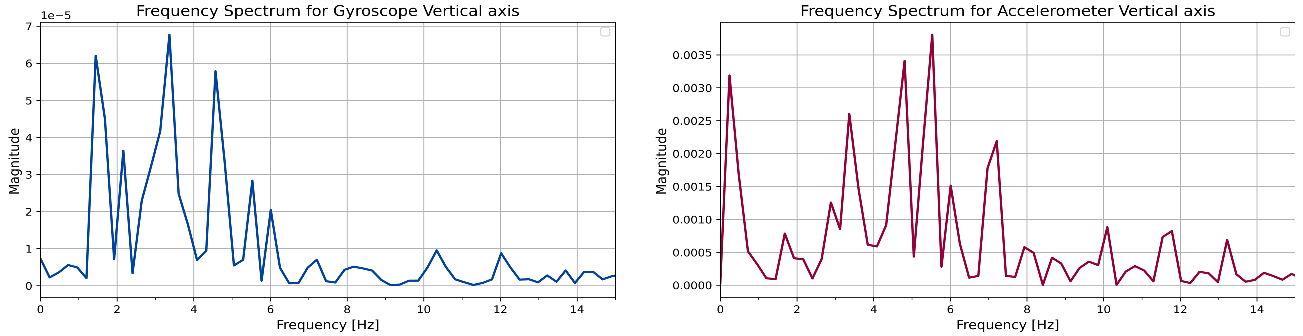
D. Data Segmentation for Deep Learning Model

Data normalisation is essential for DL models to reduce computational time and improve performance [63]. The best practice to perform data normalisation and estimate the data distribution is always to use only the training data and keep the test dataset untouched to prevent potential overfitting. In our study, the training data (consisting of the nine IMU signals) was used to fit the robust scaler, and the entire dataset was transformed using the best-fit scaler. The normalised data were then segmented into smaller 4-second windows (200 samples), with a 50% overlap. Each 4-second window was composed of two parts. The non-overlapping part of the data was 2 seconds and was defined as the current window because the window label was determined using only part of the data. The overlapping 2 seconds of data was from the previous window, and it was combined with the current window in order to capture more features and transitory patterns.

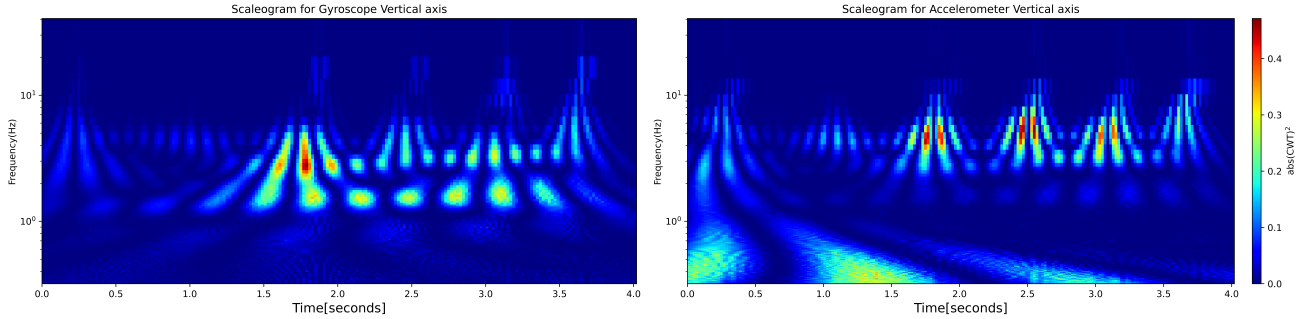
In some earlier studies, a window was labelled as a FOG window when all data in that window were FOG data [41], or more than 50% of the data in that window were FOG data [42]. However, in our data, a window that contained more than 0.2-second of FOG data (10% of the new information in each window, e.g. ten samples) was labelled as a FOG window. The shortest FOG episode in our dataset was 0.04 seconds, which was the duration of a two frames in the videos used by the therapists to identify FOG. There were only 6 FOG



(a) FoG in Time Domain



(b) FoG in Frequency Domain



(c) FoG in Time-Frequency Domain

Fig. 2: FOG signals visualised in time, frequency, and time-frequency domains.

episodes (1.2%) shorter than 0.2 seconds, and these episodes might not be detectable in real life. The rationale for labelling 0.2 seconds of FOG data as a FOG window was to minimise the detection latency and improve the detection robustness by training the model to recognise partial FOG windows and short FOG episodes. However, the obvious drawback was that a model's performance deteriorated with this approach, which might be part of the reasons why the reconstructed models in this paper exhibited reduced performance as compared to those reported in the original studies.

E. Deep Learning Model Architecture

Extensive research on CNN has shown that models can learn more complex features as it goes deeper. However, training a deeper model is challenging as it requires a large amount of computational resources and data. Therefore, we limited our network to a maximum of 8 CNN layers. An overview of the CNN architecture used is shown in Figure 3.

The first two CNN layers in the first 2D convolutional block contained 77 and 685 filters, respectively, and the kernel size was 7x7. A max-pooling layer with pool size of 5x5 and stride size of 2x2 was added after the CNN layer. The two CNN layers in the second and third blocks had identical configurations, except 128 filters were used in the second block and 464 filters were used in the third block. The kernel size in these two blocks was reduced to 5x5. The last block used a smaller 3x3 kernel and 101 filters. A 2D global average pooling layer, followed by a fully connected layer with 512 neurons, was added at the end of the convolutional block. After the fully connected layer, a sigmoid activation function was used to determine the output. All convolutional blocks ended with a batch normalisation layer, a softsign activation layer, and a dropout layer. The dropout rate was set to 0.4.

We also used the following DL techniques in the model to

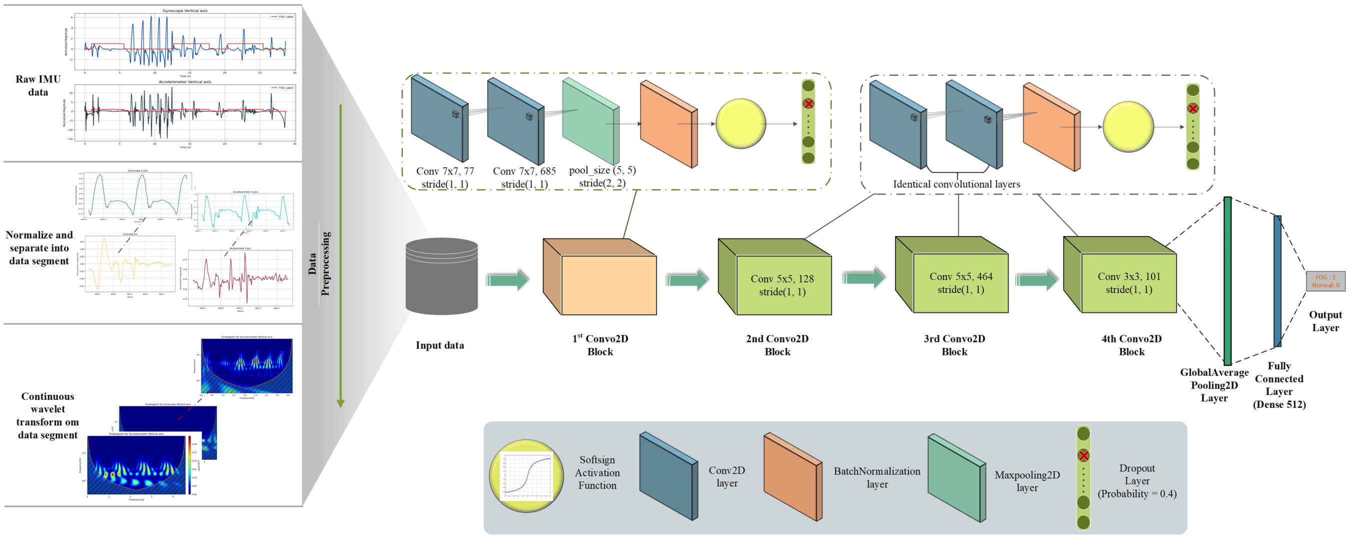

Gyroscope Vertical axis
FOG Label
Normalized Magnitude
Time (s)
Raw IMU data
Accelerometer Vertical axis
Gyroscope Z axis
Accelerometer X axis
Normalize and separate into data segment
Scaleogram for Accelerometer Vertical axis
Scaleogram for Gyroscope Vertical axis
Frequency(Hz)
Time(seconds)
Continuous wavelet transform om data segment
Data Preprocessing
Conv 7x7, 77 stride(1, 1)
Conv 7x7, 685 stride(1, 1)
pool_size (5, 5) stride(2, 2)
Identical convolutional layers
Input data
1st Convo2D Block
Conv 5x5, 128 stride(1, 1)
2nd Convo2D Block
Conv 5x5, 464 stride(1, 1)
3rd Convo2D Block
Conv 3x3, 101 stride(1, 1)
4th Convo2D Block
GlobalAverage Pooling2D Layer
Fully Connected Layer (Dense 512)
FOG : 1
Normal: 0
Output Layer
Softsign Activation Function
Conv2D layer
BatchNormalization layer
Maxpooling2D layer
Dropout Layer (Probability = 0.4)

Fig. 3: Proposed CNN architecture.

improve model robustness and prevent overfitting:

1) Regularization: Overfitting is a common modelling issue, and it often occurs with CNN models. This error happens when the model fits the training data too well but fails to generalise to the test data. A few regularisation techniques to overcome overfitting were implemented in our model.

A large weight in CNNs will typically amplify noise in the input data, causing the error to increase further while propagating through the network. Hence, it is often an indicator of overfitting. Maximum normalised weight constraints were applied to all our convolution layers to ensure that the magnitude of weights did not exceed a given threshold during training.

A dropout layer is another regularisation technique to prevent overfitting. It randomly sets the output of some hidden neurons to zero during training at the given retaining probability. Dropout layers typically work well with a maximum normalised weight constraint. We tested different combinations of values for the maximum normalised weight and the dropout probability of retention, and the optimal result is discussed in the next section.

Early stopping was another regularisation method we used to prevent overfitting, where training was stopped when there were no significant decreases in validation loss over 20 epochs.

2) Global Average Pooling: Another technique used to reduce overfitting was introducted by Lin et al. [64] who introduced a global average pooling layer to replace the fully connected layer and enhance model discriminability within the receptive fields. Our model used a 2D global average pooling layer, followed by a fully connected layer at the end of the layer. The implementation has been used in recent advanced CNN architectures, such as EfficientNet [65] and MobileNet [66], as it reduces the computational cost of using two fully connected layers and maintains the performance of the model at the same time.

3) Batch Normalisation: Recent deep learning models have tended to increase their depth with multiple layers that are combined sequentially, with the inputs to each neural network layer coming from the activity of the previous layer. During the training process, the parameters in each neural layer will be updated with each mini-batch of data, and this change of parameters will create a constant shift in the distribution of inputs [67, 68]. When these inputs propagate through the network, this small distribution change is amplified and causes a slowdown in the network convergence. This phenomenon has been described as an internal covariate shift [67]. We used batch normalisation in our network to normalise each layer's inputs to reduce the training time and improve the model's robustness [69].

F. Model Optimisation

With the thousands of combinations of all the hyper-parameters, it would have been very time-consuming to identify the optimal hyper-parameters if we used the entire dataset. As such, we used only 30% of the training dataset to search for the potential candidates and obtained an estimate of the optimal model performance. Furthermore, as the network structure was relatively complex, conventional hyper-parameter tuning approaches, such as the exhaustive grid search or random search, could not be performed using the full scale of the network as the computational cost would have been enormous. Hence, we adopted the "Taking the Human Out of the Loop" concept [70] and chose the Bayesian Optimisation approach to select the optimal combination of hyper-parameters, such as activation functions, dropout rate, kernel initialisers, weight constraints, optimisers, loss functions, and the number of filters in each layer. A hyper-parameter tuning library, scikit-optimize [71], was adopted in the fine-tuning process, and the Gradient Boosted Regression Trees technique was used to minimise the negative G-mean.

The Bayesian Optimisation (BO) function tried to find a new hyperparameter sample X_n bounded by the given options χ (shown in Table III). For each iteration, the BO selected the best X_n by optimising the acquisition function, α, with the surrogate function obtained from the previous iteration D_{n-1}:

$$X_n = argmax_{X \in \chi}\, \alpha(X \mid D_{n-1}) \tag{2}$$

where the acquisition function α, was defined as :

$$\alpha(X) = \mathbb{E}_{max}(f(X_n) - f(X^{+})) \tag{3}$$

The $f(X^{+})$ and $f(X_n)$ symbolised, respectively, the highest validation G-mean derived from the objective function so far, and the current validation G-mean from the current set of hyperparameters. D_{n-1} represented the surrogate function that was estimated from the last set of hyperparameters, and the objective function $f(X_n)$. The surrogate function was an approximation of the objective function used to infer the posterior of the optimisation process. The BO was a sequential optimisation process, where the initial set of X_n was determined randomly from the options χ. The surrogate function D_{n-1} was updated by the initial set of hyperparameters or the last set of values that had been evaluated. The updated D_{n-1} was then fed to Equation 2 to find an improved set of hyperparameters. We set the maximum number of iterations to 256, but the validation G-mean exhibited no significant improvements after 154 evaluations in our experiment, which suggested that the sequential optimisation process had converged to the global optimum. The final set of hyper-parameters are summarised in Table III.

To speed up the fine-tuning process further, the model was trained using synchronous distributed training where each GPU ran a replica model with a local batch size of 10. We trained each combination of parameters for 50 epochs with a global batch size of 80 (the local batch size * the number of GPUs). Theoretically, this training strategy would have increased the training speed by eight times.

G. Model Evaluation

Six evaluation metrics were used to evaluate the proposed model's performance and compare it to other state-of-the-art algorithms Table IV. The Shapiro-Wilk normality test was performed on all evaluation metrics from all classification models. The one-tailed Student's t-test ($\alpha = 0.05$) was used to test for statistically significant improvements if the distributions from both evaluation metrics passed the normality test, and the Hedge's g was reported for the effect size estimation. Otherwise, statistical significance was assessed using the nonparametric Mann–Whitney U test ($\alpha = 0.05$) and the Rank-Biserial Correlation was applied to determine the effect size.

For most PD patients, FOG events constitute only a small part of their regular walking experience. The gait data collected for FOG studies will always be imbalanced with FOG incidents being the minority class. We used the geometric mean (G-mean) and F1 score (harmonic mean of the precision and recall) as they are generally the better metrics to evaluate model performance by taking into account data imbalances [72, 73].

TABLE III: CNN hyperparameters explored using Bayesian Optimisation and the optimal parameters obtained.

Parameter	**Options**	**Optimal Parameters**
Activation Function	1. relu 2. softplus 3. softsign 4. selu 5. swish	softsign
Dropout Rate	*U (0.1, 0.9)*	0.4
Kernel Initializer	1. lecun uniform 2. lecun normal 3. normal 4. glorot normal 5. glorot uniform 6. he normal 7. he uniform 8. orthogonal 9. variance scaling	normal
Layer Weight Constraint	logU (0.6, 5)	1.8
Optimizer	1. SGD 2. RMSprop 3. Adagrad 4. Adadelta 5. Adam 6. Adamax 7. Nadam 8. Ftrl	Adamax
Loss Function	1. binary crossentropy 2. sparse categorical crossentropy 3. hinge 4. squared hinge	binary crossentropy

III. RESULTS AND DISCUSSION

All experiments were conducted using Python, Tensorflow, and other relevant python libraries. The model was trained on an Nvidia Tesla V100 GPU using an Amazon Web Services Elastic Compute Cloud (EC2) cluster.

The data was split into 80% training data and 20% test data, i.e., 50 subjects in the training dataset and 13 subjects in the test dataset. During the training and validation process, the 10-fold cross-validation was performed using only the training data, and the test set was held out for final evaluation.

A. ML Classification Results

Seven popular machine learning models (KNN, LR, DT, RF, SVM, SVM-RBF and XGBoost) were selected to evaluate the classification performance using conventional handcrafted features. All models were fine-tuned using a grid search to determine the best set of hyper-parameters. Each model was trained with Stratified 10-fold validation. The mean performance of models over the 10-fold validation is shown in Table V and Figure 4. The XGBoost showed the best performance in accuracy (80.65%) , G-mean (81.03%), and F1

TABLE IV: Classification Evaluation Metrics. The true positive rate (***TP***) indicated the proportion of FOG episodes correctly classified. The false positive rate (***FP***) showed the proportion of non-FOG data windows misclassified as FOG episodes. The true negative rate (***TN***) computed the proportion of non-FOG episodes that were classified accurately. The false-negative rate (***FN***) was the proportion of FOG episodes misclassified as non-FOG episodes.

Evaluation Metrics	**Mathematical Expression**	**Explanation**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	The total percentage of correctly classified windows.
Sensitivity / Recall	$\frac{TP}{TP+FN}$	The true positive rate, which corresponded to the ratio of FOG windows that were correctly classified as FOG windows.
Specificity	$\frac{TN}{TN+FP}$	The true negative rate, which indicated the proportion of non-FOG windows that were correctly considered as non-FOG.
Precision	$\frac{TP}{TP+FP}$	The ratio of correctly classified FOG windows to the total proportion of classified FOG windows.
Geometric Mean	$\sqrt[2]{Sensitivity * Specificity}$	The G-Mean, the square root of the product of the sensitivity and specificity is a performance measurement that helps to balance the result among different classes.
F1-Score	$2 * \frac{Precision*Recall}{Precision+Recall}$	The harmonic mean of precision and recall is an evaluation metric that assesses the classification performance by evenly weighting recall and precision.

score (77.41%). Other models showed a mean classification accuracy below 80%. The XGBoost model also exhibited statistical improvement over KNN and SVM (RBF) models for four evaluation metrics, with effect sizes (Hedge's g) above 0.8.

B. Deep Learning Classification Results

1) Baseline CNN Model and Reconstructed Models: We retrained our previous model [46] with 10-fold cross-validation to evaluate the model performance. Furthermore, we reconstructed some of the state-of-the-art models mentioned in the first section as a comparison. As the dataset was different, and because of the increased heterogeneity in our data because of the much larger number of subjects, some of the reconstructed models (like Camps's model and Bikias's model) did not achieve the performance reported in the original articles. In contrast, our baseline model and the reconstructed Xia's models exhibited very comparable results to their reported subject-independent models. The performance differences in the reconstructed models could also have been due to differences in data collection conditions (e.g. data collection in hospital versus home), sensor data (e.g. inclusion of magnetometer data), window sizes, and labeling. The results are shown in Table VI and Figure 5.

2) Proposed CNN Model: We trained the models with a 10-fold cross-validation scheme and evaluated them with the hold-out test set to demonstrate that the proposed CNN model outperformed previous models. Table VI shows that the proposed CNN model exhibited a statistically significant improvement from the ML models for the average accuracy, precision, G-mean, and F1 score. All metrics except the sensitivity were significantly improved compared to Xia's model. However, the proposed model displayed significant improvements from Camps's model only in specificity, precision, and G-mean. The average accuracy, specificity, G-mean, and F1 score were also significantly improved from the CNN baseline model. Compared with the latest CNN FOG detection model, Bikias's model, the proposed model manifested significant improvement for accuracy, sensitivity, G-mean and F1 score. All the significant improvements in the metrics showed substantial effect size (Hedge's g > 0.8 or the Rank-Biseriall correlation > 0.5), indicating large practical differences [74, 75]. The only three exceptions were when the CNN baseline model's F1 score and Bikias's model's G-mean were compared with the corresponding metrics from the best ML model, and the proposed model's F1 score was compared with the F1 scores from the CNN baseline model and Xia's model. These metrics failed the normality test, and the correlation was below 0.5. Furthermore, the proposed model showed significant improvement over G-mean against all the models with substantial effect size.

The best performance obtained from the proposed model achieved a 90.7% G-mean and 91.5% F1 score. Furthermore, the proposed model and the reconstructed Bikias's and Camps's model displayed minor performance fluctuations throughout the 10-fold cross-validation, indicating that these models provided similar performance levels with data from new subjects. No significant improvement for sensitivity was found in the statistical analysis, mainly because the proposed model was optimised to find the best G-mean performance.

IV. CONCLUSION

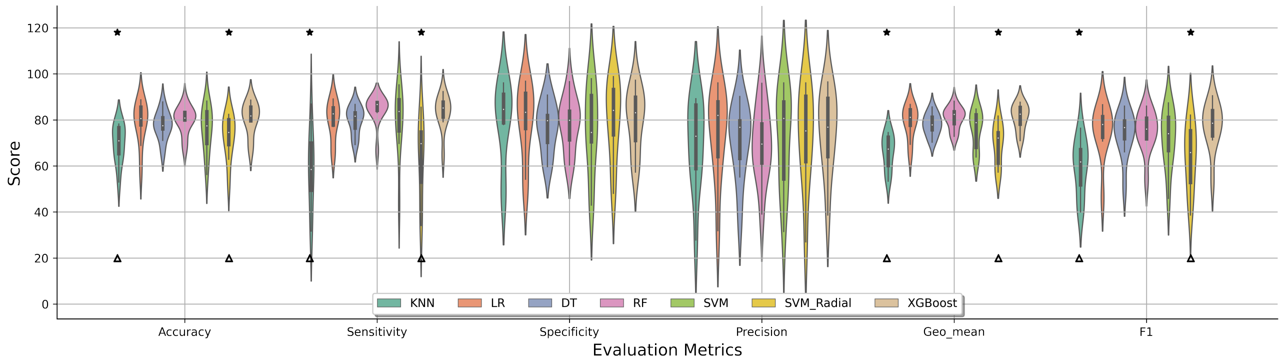
In the last two decades, FOG detection algorithms have slowly changed from classical feature extraction and statistical analysis methods toward adopting various machine learning (ML) and deep learning (DL) algorithms and techniques to discover data characteristics. However, FOG detection still remains a challenging problem because of the complexity and heterogeneity of the symptoms. Another challenge is the amount of high-quality data available to develop a reliable and robust deep learning model.

In this study, we proposed a novel method to analyse IMU data using time-frequency analysis and proposed a robust model structure that was trained and validated on a relatively large cohort of PD patients who suffered from FOG. Using the "Taking the Human Out of the Loop" concept, we employed Bayesian Optimisation to determine the optimal hyperparameters and the final model design. Our proposed design is also a subject-independent model, and it is immune to the fluctuation in gait patterns when PD patient mobility deteriorates over time. The empirical results also indicated that the model can provide consistent performance and excellent FOG detection accuracy. Moreover, the statistically significant improvement of G-mean in the proposed model compared against all other models demonstrated that the Bayesian optimisation approach could effectively determine the hyperparameter over desired evaluation metrics or objective functions.

The proposed model used the two-dimensional time-frequency representations as inputs, demonstrating the feasi-

TABLE V: The ML prediction results with selected FOG features. * denotes results that were significantly poorer than the best ML model, XGBoost ($p < 0.05$). Statistical significance required either that the effect size measurement Hedge's g was greater than 0.8, or the Rank-Biserial Correlation was greater than 0.5.

ML Classifier	**Accuracy**		**Sensitivity**		**Specificity**		**Precision**		**Geo_mean**		**F1**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
K Nearest Neighbour	69.68*	8.33	58.06*	17.2	79.75	17.57	69.8	19.7	66.07*	7.59	59.59*	12.07
Logistic Regression	78.88	9.38	80.09	8.35	79.69	16.1	74.81	19.33	79.15	7.19	75.27	11.43
Decision Trees	77.69	6.23	79.84	5.99	76.78	10.77	71.07	16	77.96	4.79	73.96	9.93
Random Forest	79.92	5.99	84.38	6.44	78.2	11.5	70.14	16.3	80.83	4.8	75.15	9.04
SVM - Linear	75.72	9.97	79.36	14.88	75.64	18.9	71.87	21.54	75.87	8.32	71.84	12.6
SVM - RBF	72.76*	9.44	63.67*	17.57	80.21	17.3	72.65	21.6	69.69*	9	64.04*	14.39
XGBoost	80.65	7.06	83.31	7.62	79.89	13.37	75.25	17.76	81.03	5.67	77.41	10.32



Statistical Significance Level (For all tests : p < 0.05):
★ denotes statistically significant poorer result than the best ML model (XGBoost).

Effect size:
△ denotes the effect size measurement Hedge's g was greater than 0.8 or the Rank-Biserial Correlation was greater than 0.5.

Fig. 4: ML classification results with the selected FOG features.
1) K-nearest neighbours (KNN), 2) Linear Regression (LR), 3) Decision Tree (DT), 4) Random Forest (RF), 5) Support Vector Machine (SVM) with linear kernels, 6) SVM with radial basis function (RBF) kernels, and 7) Extreme Gradient Boosting (XGBoost).

TABLE VI: Classification performance for the baseline and re-constructed DL models. * denotes significant improvement from XGBoost. † indicates that the proposed model exhibited a significant improvement from the corresponding model for the specific metric. Statistical significance meant that the effect size measurement Hedge's g was greater than 0.8, or the Rank-Biserial Correlation was greater than 0.5.

DL Model	**Accuracy**		**Sensitivity**		**Specificity**		**Precision**		**Geo_mean**		**F1**	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
CNN Baseline Model	83.65†	2.55	85.67	6.34	79.94†	9.35	86.51*	4.3	82.42†	2.45	85.81	1.67
Xia's Model	79.33†	1.92	84.36	7.14	71.83†	8.74	77.49†	2.38	82.09†	3.76	82.9	2.4
Camps's Model	85.87*	1.49	89.78*	3.51	81.34†	2.3	84.78†	1.36	85.42*†	1.34	87.17*	1.58
Bikias's Model	83.99†	0.18	81.65†	0.53	86.91	0.69	88.63	0.49	88.62†	0.19	84.99*†	0.18
Proposed Model	**87.06***	2	87.75	2.9	86.39	3.7	**88.7***	3.24	**88.7***	2.08	**88.17***	1.96

bility of using computer vision techniques and architecture to detect FOG. This approach lay the groundwork for future FOG detection research to adapt and develop more innovative solutions from the computer vision domain. However, the main limitations of this model are that the model does not reduce the computational cost and relies heavily on GPUs to process the data during the training phase of the model. Nevertheless, we did not find a significant increase in computational cost for the inference process compared with other models. Another limitation of this study is that the data collection was performed in the hospital, where patients pay extra attention to their movements. This approach might not represent sufficiently all

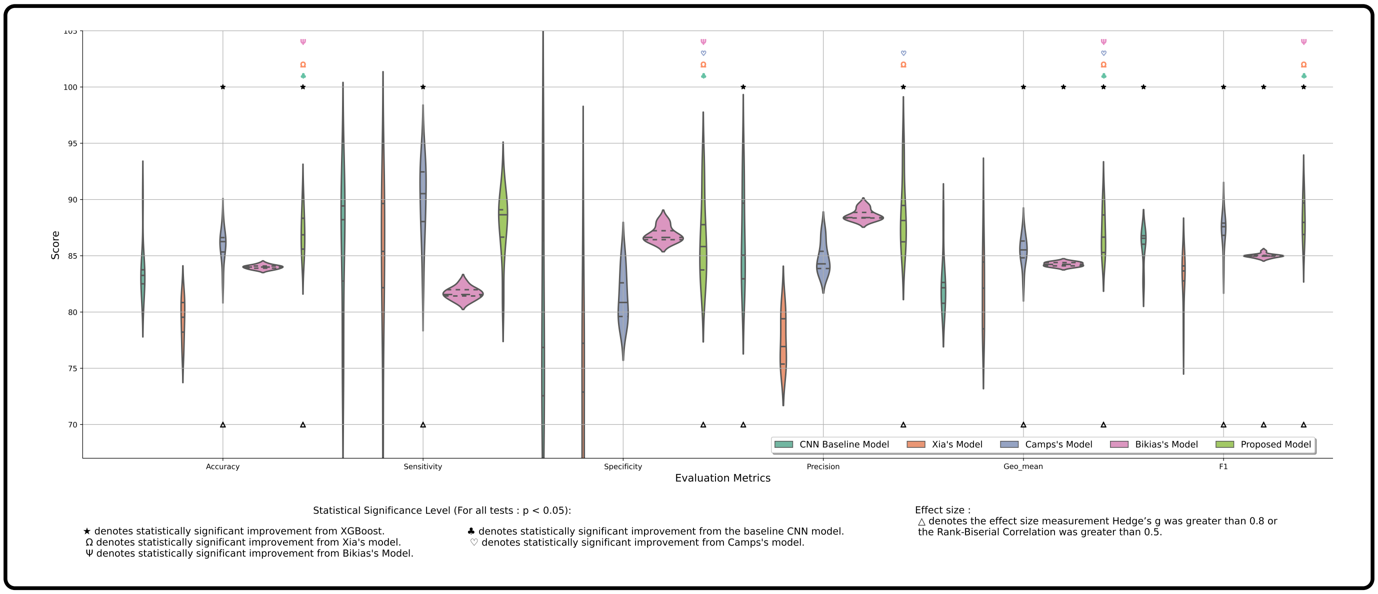


Fig. 5: Classification performance for the baseline and re-constructed DL models.

the real-world environments that trigger FOG and the patients' daily difficulties at home.

References

[1] L. V. Kalia and A. E. Lang, "Parkinson's disease," *The Lancet*, vol. 386, pp. 896–912, 8 2015, doi: 10.1016/S0140-6736(14)61393-3.

[2] E. R. Dorsey *et al.*, "Global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 17, pp. 939–953, 11 2018, doi: 10.1016/S1474-4422(18)30295-3.

[3] S. V. Campenhausen *et al.*, "Prevalence and incidence of parkinson's disease in europe," *European Neuropsychopharmacology*, vol. 15, pp. 473–490, 8 2005.

[4] W. Muangpaisan *et al.*, "Systematic review of the prevalence and incidence of parkinson's disease in asia," *Journal of Epidemiology*, vol. 19, pp. 281–293, 2009.

[5] O. Riedel *et al.*, "Estimating the prevalence of parkinson's disease (pd) and proportions of patients with associated dementia and depression among the older adults based on secondary claims data," *International Journal of Geriatric Psychiatry*, vol. 31, pp. 938–943, 8 2016.

[6] W. Poewe *et al.*, "Parkinson disease," *Nature Reviews Disease Primers*, vol. 3, pp. 1–21, 3 2017.

[7] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, pp. 1545–1602, 10 2016.

[8] G. Alves *et al.*, "Epidemiology of parkinson's disease," *Journal of Neurology*, vol. 255, pp. 18–32, 9 2008.

[9] O. B. Tysnes and A. Storstein, "Epidemiology of parkinson's disease," *Journal of Neural Transmission*, vol. 124, pp. 901–905, 8 2017.

[10] A. H. Schapira, "Etiology of parkinson's disease," *Neurology*, vol. 66, pp. S10–S23, 5 2006.

[11] K. Wirdefeldt *et al.*, "Epidemiology and etiology of parkinson's disease: A review of the evidence," *European Journal of Epidemiology*, vol. 26, 6 2011.

[12] N. Giladi *et al.*, "Freezing of gait in pd: Prospective assessment in the datatop cohort," *Neurology*, vol. 56, pp. 1712–1721, 6 2001.

[13] J. G. Nutt *et al.*, "Freezing of gait: Moving forward on a mysterious clinical phenomenon," *The Lancet Neurology*, vol. 10, pp. 734–744, 8 2011.

[14] N. Giladi *et al.*, "Freezing of gait in patients with advanced parkinson's disease," *Journal of Neural Transmission*, vol. 108, pp. 53–61, 2001.

[15] A. Nieuwboer and N. Giladi, "The challenge of evaluating freezing of gait in patients with parkinson's disease," *British Journal of Neurosurgery*, vol. 22, pp. S16–S18, 1 2008.

[16] Y. Okuma and N. Yanagisawa, "The clinical spectrum of freezing of gait in parkinson's disease," *Movement Disorders*, vol. 23, pp. S426–30, 2008.

[17] A. H. Snijders *et al.*, "Clinimetrics of freezing of gait," *Movement Disorders*, vol. 23, pp. S468–S474, 2008.

[18] J. D. Schaafsma *et al.*, "Characterization of freezing of gait subtypes and the response of each to levodopa in parkinson's disease," *European Journal of Neurology*, vol. 10, pp. 391–398, 7 2003.

[19] J. L. McKay *et al.*, "Freezing of gait can persist after an acute levodopa challenge in parkinson's disease," *npj Parkinson's Disease*, vol. 5, p. 25, 12 2019.

[20] J. Nonnekes *et al.*, "Freezing of gait: A practical approach to management," *The Lancet Neurology*, vol. 14, pp. 768–778, 7 2015.

[21] L.-L. Zhang *et al.*, "Freezing of gait in parkinsonism and its potential drug treatment," *Current Neuropharmacology*, vol. 14, pp. 302–306, 4 2016.

[22] G. Deuschl *et al.*, "A randomized trial of deep-brain stimulation for parkinson's disease," *New England Journal of Medicine*, vol. 355, pp. 896–908, 8 2006.

[23] V. Ricchi *et al.*, "Transient effects of 80 hz stimulation on gait in stn dbs treated pd patients: A 15 months follow-up study," *Brain Stimulation*, vol. 5, pp. 388–392, 7 2012.

[24] T. Xie *et al.*, "Low-frequency stimulation of stn-dbs reduces aspiration and freezing of gait in patients with pd," *Neurology*, vol. 84, pp. 415–420, 2015.

[25] T. B. Stoker *et al.*, "Emerging treatment approaches for parkinson's disease," *Frontiers in Neuroscience*, vol. 12, 10 2018.

[26] L. Rochester *et al.*, "The effect of external rhythmic cues (auditory and visual) on walking during a functional task in homes of people with parkinson's disease," *Archives of Physical Medicine and Rehabilitation*, vol. 86, pp. 999–1006, 5 2005.

[27] M. E. Morris *et al.*, "Striding out with parkinson disease: Evidence-based physical therapy for gait disorders," *Physical Therapy*, vol. 90, pp. 280–288, 2 2010.

[28] S. Ghai *et al.*, "Effect of rhythmic auditory cueing on parkinsonian gait: A systematic review and meta-analysis," *Scientific Reports*, vol. 8, p. 506, 12 2018.

[29] S. J. Spaulding *et al.*, "Cueing and gait improvement among people with parkinson's disease: A meta-analysis," *Archives of Physical Medicine and Rehabilitation*, vol. 94, pp. 562–570, 3 2013.

[30] P. Ginis *et al.*, "External input for gait in people with parkinson's disease with and without freezing of gait: One size does not fit all," *Journal of Neurology*, vol. 264, pp. 1488–1496, 7 2017.

[31] M. Gilat *et al.*, "Freezing of gait: Promising avenues for future treatment," *Parkinsonism and Related Disorders*, vol. 52, pp. 7–16, 7 2018.

[32] V. Mikos *et al.*, "A wearable, patient-adaptive freezing of gait detection

system for biofeedback cueing in parkinson's disease," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, pp. 503–515, 6 2019.

[33] A. Nieuwboer *et al.*, "Abnormalities of the spatiotemporal characteristics of gait at the onset of freezing in parkinson's disease," *Movement Disorders*, vol. 16, pp. 1066–1075, 11 2001.

[34] I. Maidan *et al.*, "Heart rate changes during freezing of gait in patients with parkinson's disease," *Movement Disorders*, vol. 25, pp. 2346–2354, 10 2010.

[35] R. Sun *et al.*, "Convolutional 3d attention network for video based freezing of gait recognition." Institute of Electrical and Electronics Engineers Inc., 1 2019.

[36] K. Hu *et al.*, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 1215–1225, 4 2020.

[37] R. Miotto *et al.*, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, pp. 1236–1246, 11 2017.

[38] A. Esteva *et al.*, "A guide to deep learning in healthcare," pp. 24–29, 1 2019.

[39] O. Faust *et al.*, "Deep learning for healthcare applications based on physiological signals: A review," pp. 1–13, 7 2018.

[40] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706–710, 1 2020.

[41] J. Camps *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, 1 2018.

[42] Y. Xia *et al.*, "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, 9 2018.

[43] L. Sigcha *et al.*, "Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors," *Sensors*, vol. 20, p. 1895, 3 2020.

[44] T. Bikias *et al.*, "Deepfog: An imu-based detection of freezing of gait episodes in parkinson's disease patients via deep learning," *Frontiers in Robotics and AI*, vol. 0, p. 117, 5 2021.

[45] M. Bächlin *et al.*, "Wearable assistant for parkinsons disease patients with the freezing of gait symptom," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, pp. 436–446, 3 2010.

[46] B. Shi *et al.*, "Convolutional neural network for freezing of gait detection leveraging the continuous wavelet transform on lower extremities wearable sensors data," vol. 2020-July. Institute of Electrical and Electronics Engineers Inc., 7 2020, pp. 5410–5415.

[47] W. W. Lee *et al.*, "A smartphone-centric system for the range of motion assessment in stroke patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 1839–1847, 11 2014.

[48] Y. Kumar *et al.*, "Wireless wearable range-of-motion sensor system for upper and lower extremity joints: a validation study," *Healthcare Technology Letters*, vol. 2, pp. 12–17, 2 2015.

[49] M. H. Nilsson *et al.*, "Development and testing of a self administered version of the freezing of gait questionnaire," *BMC Neurology*, vol. 10, p. 85, 12 2010.

[50] I. Goodfellow *et al.*, "Dataset augmentation," p. 237, 2016.

[51] A. Delval *et al.*, "Objective detection of subtle freezing of gait episodes in parkinson's disease," *Movement Disorders*, vol. 25, pp. 1684–1693, 8 2010.

[52] C. A. Coste *et al.*, "Detection of freezing of gait in parkinson disease: Preliminary results," *Sensors*, vol. 14, pp. 6819–6827, 4 2014.

[53] P. Tahafchi *et al.*, "Freezing-of-gait detection using temporal, spatial, and physiological features with a support-vector-machine classifier." Institute of Electrical and Electronics Engineers Inc., 9 2017, pp. 2867–2870.

[54] M. Zago *et al.*, "Gait evaluation using inertial measurement units in subjects with parkinson's disease," *Journal of Electromyography and Kinesiology*, vol. 42, pp. 44–48, 10 2018.

[55] J. M. Hausdorff *et al.*, "Time series analysis of leg movements during freezing of gait in parkinson's disease: akinesia, rhyme or reason?" *Physica A: Statistical Mechanics and its Applications*, vol. 321, pp. 565–570, 4 2003.

[56] H. G. MacDougall and S. T. Moore, "Marching to the beat of the same drummer: The spontaneous tempo of human locomotion," *Journal of Applied Physiology*, vol. 99, pp. 1164–1173, 9 2005.

[57] V. Mikos *et al.*, "Optimal window lengths, features and subsets thereof for freezing of gait classification," vol. 2018-Janua. Institute of Electrical and Electronics Engineers Inc., 2 2018, pp. 1–8.

[58] S. T. Moore *et al.*, "Ambulatory monitoring of freezing of gait in parkinson's disease," *Journal of Neuroscience Methods*, vol. 167, pp. 340–348, 1 2008.

[59] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate and sample entropy," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 278, pp. 2039–2049, 2000.

[60] J. M. Hausdorff, "Gait dynamics in parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling," *Chaos*, vol. 19, 2009.

[61] N. G. Pozzi *et al.*, "Freezing of gait in parkinson's disease reflects a sudden derangement of locomotor network dynamics," *Brain*, vol. 142, pp. 2037–2050, 7 2019.

[62] A. El-Attar *et al.*, "Discrete wavelet transform-based freezing of gait detection in parkinson's disease," *Journal of Experimental and Theoretical Artificial Intelligence*, pp. 1–17, 9 2018.

[63] N. M. Nawi *et al.*, "The effect of data pre-processing on optimized training of artificial neural networks," *Procedia Technology*, vol. 11, pp. 32–39, 1 2013.

[64] M. Lin *et al.*, "Network in network," 12 2014.

[65] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," vol. 2019-June. International Machine Learning Society (IMLS), 5 2019, pp. 10 691–10 700.

[66] A. Howard *et al.*, "Searching for mobilenetv3," vol. 2019-Octob. Institute of Electrical and Electronics Engineers Inc., 5 2019, pp. 1314–1324.

[67] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," vol. 1. International Machine Learning Society (IMLS), 2 2015, pp. 448–456.

[68] S. Santurkar *et al.*, "How does batch normalization help optimization?" vol. 2018-Decem, 5 2018, pp. 2483–2493.

[69] G. Yang *et al.*, "A mean field theory of batch normalization." International Conference on Learning Representations, ICLR, 2 2019.

[70] B. Shahriari *et al.*, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, pp. 148–175, 1 2016.

[71] T. Head *et al.*, "scikit-optimize/scikit-optimize," 9 2020.

[72] M. Kubat and S. Matwin, "Addressing the curse of imbalanced data sets: One-sided sampling," 1997, pp. 179–186.

[73] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, 9 2009.

[74] C. J. Ferguson, "An effect size primer: A guide for clinicians and researchers." pp. 301–310, 12 2015.

[75] D. C. Funder and D. J. Ozer, "Evaluating effect size in psychological research: Sense and nonsense," *Advances in Methods and Practices in Psychological Science*, vol. 2, pp. 156–168, 6 2019.