Freezing of Gait Detection in Parkinson's Disease: A Subject-Independent Detector Using Anomaly Scores

Thuy T. Pham*, Steven T. Moore, Simon J. G. Lewis, Diep N. Nguyen, Eryk Dutkiewicz, Andrew J. Fuglevand, Alistair L. McEwan, Philip H.W. Leong

***Abstract*—Freezing of gait (FoG) is common in Parkinsonian gait and strongly relates to falls. Current clinical FoG assessments are patients' self-report diaries and experts' manual video analysis. Both are subjective and yield moderate reliability. Existing detection algorithms have been predominantly designed in subject-dependent settings. In this work, we aim to develop an automated FoG detector for subject-independent. After extracting highly relevant features, we apply anomaly detection techniques to detect FoG events. Specifically, feature selection is performed using correlation and clusterability metrics. From a list of 244 candidates, 36 candidates were selected using saliency and robustness criteria. We develop an anomaly score detector (*ASD*) with adaptive thresholding to identify FoG events. Then, using accuracy metrics, we reduce the feature list to seven candidates. Our novel multi-channel freezing index was the most selective across all window sizes, achieving sensitivity (specificity) of 96% (79%). On the other hand, freezing index from the vertical axis was the best choice for a single input, achieving sensitivity (specificity) of 94% (84%) for ankle and 89% (94%) for back sensors. Our subject-independent method is not only significantly more accurate than those previously reported, but also uses a much smaller window (e.g., $3s$ vs. $7.5s$) and/or lower tolerance (e.g., $0.4s$ vs. $2s$).**

***Index Terms*—Anomaly score, gait freezing, feature selection.**

I. Introduction

Gait is one of the most affected motor characteristics in Parkinson's disease (PD). Freezing of gait (FoG), defined as a motor block of movement (especially before gait initiation) during turns or when meeting obstacles [1], is one of the most common symptoms (e.g., reference [2] reported that 47% of more than six thousand subjects had 28% FoG events every day). Moreover, there is a strong relationship between FoG and falls [1], [3], [4]. Current clinical FoG assessment methods are self-report diaries from patients (e.g. the Unified Parkinson's Disease Rating Scale (UPDRS) [5], Freezing of Gait Questionnaire [6]) and manual video analysis of walking [7], [8]. These methods are unfortunately subjective. UPDRS has poor agreement with expert labels (the kappa statistic only ranged from 0.49 to 0.78) [9]. The reliability of existing manual video assessment is not robust (within or across multiple participant recruitment sites); and the intra-rater reliability is remarkably low [10]. An additional difficulty lies in provoking FoG during routine clinical examinations [11].

Objective FoG detection is very much desirable, especially out-of-lab deployment with wearable devices [12][13]. Compared with kinematic and electrophysiological data (e.g. electromyographic and electroencephalogram), acceleration data have been widely adopted thanks to the small size of accelerometers, making them suitable for wearable systems. An early effort was reported in [14] with two accelerometers at both ankles. The authors of [14] found that freezing gait has high frequency components ($6 \rightarrow 8Hz$) compared with normal gait ($2Hz$). Wavelet analysis [15] has been used to classify normal and freezing gait (including the ratios of each level's power to discriminate the freezing and resting states) [14]. A freezing index (*FI*), defined as the power in the *freeze* band ($3 \rightarrow 8Hz$) divided by the power in the *locomotor* band ($0.5 \rightarrow 3Hz$) [12], has been used to build FoG detectors [12], [13], [16], [17], [18], [19], [20].

Recently, to detect all FoG and festination episodes stride length and cadence were suggested rather than FI [21]. Another work using Pearson's correlation introduced a rule-based 5-class classifier for strides including two classes for FoG with tremor and complete motor block [22]. As these reports were based on separate specific channels and several contexts, we compare our work with following similar studies: simple thresholding techniques [12], [13], [16], [19], [20] and supervised/semi-supervised learning classifiers [17], [18].

To extract features, two types of inputs can be used: single input (e.g., single channels from single sensors (SCSS), the sum of squares of all three channels of single sensors (MCSS)) and multiple inputs (i.e., multiple channels of multiple sensors, MCMS). While SCSS and MCSS have been well studied, MCMS is for the first time considered in this work. Note that reference [19] examined one case of using seven sensors (single axis each sensor) that was the majority votes of seven outputs, we categorize that into the SCSS group. We refer MCMS to a case where feature values are computed from a

*: correspondence thuy.pham@sydney.edu.au. T. Pham, A. McEwan and P. Leong are with the Dept. of Electrical and Information Engineering, The University of Sydney, NSW, Australia. D. Nguyen and E. Dutkiewicz are with School of Computing and Communications, University of Technology Sydney, NSW, Australia. A. J. Fuglevand is with Dept. of Physiology, University of Arizona, AZ, USA. S.T. Moore is with the Human Aerospace Laboratory, Neurology Department, Icahn School of Medicine at Mount Sinai, New York, NY, USA and also with School of Engineering and Technology, Central Queensland University, Rockhampton, Qld, Australia, 4702. S.J.G. Lewis is with Parkinson's Disease Clinic, Brain and Mind Research Institute, The University of Sydney, 94 Mallett Street, Camperdown, Sydney, NSW 2050, Australia.

matrix of inputs.

Recently, apart from *FI*, several features from accelerometer data (e.g., average, standard deviation, variance, median, entropy, energy, and power) have been proposed for FoG detectors [12], [13], [14], [16], [17], [18], [19], [20], [23]. Advanced statistical techniques to assess gait of human in general (e.g., postural control) can be found in a comprehensive feature investigation [24], however the work was concerned with 3D motion analysis for trajectory data using a single accelerometer at the lumbar. The authors concluded that no measure in their study was able to discriminate the gait patterns of individuals within clinical groups of PD and peripheral neuropathy. Furthermore, freezing of gait data was not collected in that study.

On the other hand, we explore the new combinations of inputs. We investigate three new computation methods: the spectral coherence [25], *multi-channel FI* (FI_{MC}), and Koopman spectral analysis [26] (FI_K). FI_{MC} and FI_K, are applicable only to MCMS inputs.

With regard to feature selection, FI was compared with several other features [23]. These include statistical and zero crossing rate (SCSS group), sum of the Euclidean norm of magnitude, eigenvalues of the covariance matrix, the mean energy, and principal component analysis over the three axes of the sensor (MCSS group). Nevertheless, the authors of [23] solely relied on mutual information (MI) which measures the correlation of features with labels (Shannon's information theory) [27]. This selection could not guarantee the clusterability [28] of the selected features. Thus, in our work we introduce two additional saliency criteria for feature ranking: the variance ratio of clusters [28] and the separability calculated by Euclidean distances from an instance to a *near-hit* and *near-miss* [29]. These two criteria help finding more discriminative relevant features to increase the performance of classification.

Several techniques have been recently proposed for subject-independent FoG detection. Using the vertical axis of an ankle sensor, a *global threshold FI* of 2.3 with $6s$ windowing was suggested in [12], then another global FI of 3 with $7s$ windowing was reported in [19]. By examining the same three locations of sensors with [19], reference [20] selected a different choice for the global FI of 1.4 ($2s$ windows and the dorsoventral direction of the lumbar sensor). Model learning based classifiers have worked well for subject-dependent or group-dependent settings [17], [18]. Nevertheless in order to achieve subject-independent settings these automatic techniques only address global parameters. We suggest to use an universal technique that can avoid subjective parameters.

A primary reason hindering subject-independent performance lies in the generalization of parameters. One example could be a strong *context* dependence of parameters in conjunction with large subject-variability [19]. We hypothesize that an anomaly score detector (*ASD*) can significantly improve subject-independent performance. Anomaly detection is a technique to identify patterns in data that are not similar to previous behaviors. Inspired by observations of an increase in FI during a FoG event (vesus a locomotor activity) [14][12], we investigate if this is also the case for other features. When the current feature value of a data window is lower than the on-the-fly threshold, we consider the window a *potential* non-FoG epoch. During detection, the threshold is the average of all previous values from *potential* non-FoG epochs. Thus, ASD adapts itself based on previous data, rather than seeking a universal fixed threshold. Furthermore, ASD can address diurnal variation. In other words, ASD is inherently independent of subject variability.

The main contributions of this work are:

- This is the first reported feature selection technique based on voting process with not only mutual information criterion but also clusterability for FoG detection.
- We report new features that are more relevant and discriminative than those previously employed.
- We propose a better model of detection in subject-independent settings using anomaly scores which, to the best of our knowledge, achieves the best reported performance (about 10% more accurate).

II. Methods

A. Data Sets

We first developed our algorithm with a dataset from the Daphnet project [16]. Then we deployed out-of-sample tests with a different dataset that recorded independently as one part of a larger project for FoG studies [30]. FoG annotation/labels were assessed on the Movement Disorder Society Unified Parkinson's Disease Rating Scale Section III (MDS-UPDRS-III) [31] and Hoehn and Yahr stage score [32].

1) Development Set: Seven male and three female advanced PD patients who could walk unassisted in the OFF period were recruited at Tel Aviv Sourasky Medical Center (TASMC) in Israel as a part of the EU FP6 Daphnet project (a collaboration with ETH Zurich, Switzerland) [16]. These ten participants (66.5 ± 4.8 years old) have been diagnosed with PD for 13.7 ± 9.67 years (Hoehn and Yahr score [32] (H&Y) is 2.6 ± 0.65). As illustrated in Fig. 1, three tri-axial accelerometers were attached at the shank, thigh, and lower back using elasticized straps. Data was recorded at $64Hz$ and transmitted via a Bluetooth link.

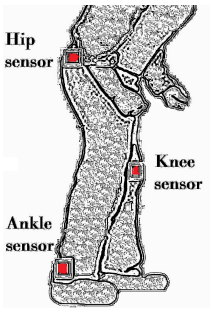


Figure 1. Three tri-axial accelerometers were attached at the shank, the thigh, and the lower back.

Three walking tasks ($10-15$ minutes each) were conducted: walking a straight line, with numerous turns, and a daily living activity (e.g., fetching coffee, opening doors); more details

are given in [16]. Three tri-axial accelerometers were attached at the shank, thigh, and lower back using elasticized straps. To prevent aliasing, data points were sampled at $64Hz$ and transmitted via a Bluetooth link. Annotation and simultaneous video taping were used by physiotherapists to determine the start/end times of FoG episodes. A FoG event label started when the gait pattern (i.e., alternating left–right stepping) was arrested and ended when the pattern was resumed [16]. The study was approved by the local Human Subjects Review Committee, and was performed in accordance with the ethical standards of the Declaration of Helsinki.

A total of five hundred minutes of data were collected. Eight participants had FoG while two did not. A total of 237 freezing events ($0 \rightarrow 66$ per subject, 23.7 ± 20.7) were recognized using video analysis by physiotherapists. This is used as the *ground truth* in our accuracy evaluations. For algorithm development (i.e., ranking features and tuning parameters), we take a random sample of 70% (five) participants who had FoG events (66 ± 5.9 years old, with PD for 16.2 ± 10.15 years, H&Y score: 2.3 ± 0.44). For out-of-sample tests, we use the remaining subjects (66.8 ± 4.1 years old, with PD for 11.2 ± 9.6 years, H&Y score: 2.9 ± 0.74). Specifically, the test set consists of 30% (three) of participants who had FoG and the others with no FoG.

2) Test Set:

We employed an independent data set for out-of-sample tests from a larger FoG study project [30]. This set included 24 patients (mean (SD) age: 69 (8.41) with advanced PD (mean (SD) Hoehn and Yahr: 2.66 (0.53); UPDRS III: 40.24 (11.06)) at Parkinson's Disease Research Clinic (the Brain and Mind Research Institute, University of Sydney, NSW Australia). These participants had severe self-reported freezing behavior and satisfied UKPDS Brain Bank criteria [33]. The subjects were deemed unlikely to have dementia or major depression according to DSM-IV criteria (by consensus rating of a neurologist and a neuropsychologist) and had a mean (SD) Mini-Mental State Examination (MMSE) [34] score of 28.57 (1.61). The study was approved by The Human Research and Ethics Committee at the University of Sydney and written consents from participants obtained.

Participants were recorded in the practically-defined 'off' state following overnight withdrawal of dopaminergic therapy. Six patients also had Deep Brain Stimulation (five Subthalamic Nuclei and one Pedunculopontine Nuclei), which were turned off for one hour prior to assessment. None of the patients described any increase in freezing behavior following the administration of their usual dopaminergic therapy.

Walking tasks were described in details at the previous work [30] that were designed to best provoke FoG during data collection. Participants, started from a sitting position, walked along a corridor about five meters meeting a marked square on the floor (size of 0.6 m) then made a turn (180^o or 540^o to the left or right of the subject). Procedure of each task were introduced to a participant at the beginning of the trial, if the subject had failed to meet the procedure, the measurement was abandoned. Each trial started by a signal from the investigator and was completed on return to the beginning position.

Data from accelerometer were acquired by seven tri-axial sensors attached to each subject at the back, foot, thigh and/or knee (further details as in the previous work [19]). These sensors were inertial measurement units (IMUs - Xsens MTx, Enschede, Netherlands) that were $38 \times 53 \times 21mm$ and 30 g. Data transmitted via a wireless link to a computer (sampling frequency of 50 Hz). Clocks of computer for data acquisition and of the video camera were used to synchronize the timing between clinical annotations and acceleration measurement.

Manual assessment of FoG made by clinicians (neurologist/neuropsychologist experienced in FoG) using video taped during each trial. These annotations were converted to binary labels ("0" for non-FoG or "1" for FoG each time instance). Each trial was assessed by two clinicians. The official label was determined FoG if at least one clinician marked as such. The agreement of these two raters were previously reported with high intraclass correlation coefficient (0.82 for number of FoG epochs and 0.99 for percent time frozen) [30][19]).

For a better comparison with the development stage, we selected data from all three tri-axial channels at three sensor locations of back, left thigh, and left shank. There were total of 71 trials across 15 subjects with six different walking procedures.

B. Feature Extraction

1) New features: We study four new types of feature extraction. The first two use single input data channels: the maximum and number of peaks in the spectral coherence [25] (called C_{XYNpks} and C_{XYmax}). The others use multiple inputs: FI_{MC} and FI_K.

Let x and y be two consecutive data windows. The spectral coherence C_{XY} between x and y using the Welch method [25] is $C_{XY}(\omega) = \frac{P_{XY}(\omega)}{\sqrt{P_{XX}(\omega).P_{YY}(\omega)}}$where ω is frequency, $P_{XX}(\omega)$ is the power spectrum of signal x, $P_{YY}(\omega)$ is the power spectrum of signal y, and $P_{XY}(\omega)$ is the cross-power spectrum for signals x and y. When $P_{XX}(\omega) = 0$ or $P_{YY}(\omega) = 0$, then $P_{XY}(\omega) = 0$ and we assume that $C_{XY}(\omega)$ is zero. To estimate the power and the cross spectra, let $\mathfrak{F}_x(\omega)$ and $\overline{\mathfrak{F}_x(\omega)}$ denote the Fourier transform and its conjugate of signal x, respectively, i.e. $\mathfrak{F}_x(\omega) = \int_{-\infty}^{+\infty} x(t).e^{-j\omega t}dt$. The power spectrum is then: $P_{XX}(\omega) = \mathfrak{F}_x(\omega).\overline{\mathfrak{F}_x(\omega)}$; $P_{YY}(\omega) = \mathfrak{F}_y(\omega).\overline{\mathfrak{F}_y(\omega)}$; and $P_{XY}(\omega) = \mathfrak{F}_x(\omega).\overline{\mathfrak{F}_y(\omega)}$.

Let a matrix $\mathbf{X}$ of size $N \times M$ represent a N-channel recording session with M regularly spaced time samples. Similar to the single input FI, FI_{MC} is the ratio of powers P_H to P_L (i.e., for the *freeze* and *locomotor* bands) that are summations of single powers over N channels. Specifically:

$$P_H = \frac{1}{2f_s}\sum_{n=1}^{N}[\sum_{i=H_1+1}^{H_2}[P_{XXn}(i)] + \sum_{i=H_1}^{H_2-1}[P_{XXn}(i)]] \quad (1)$$

$$P_L = \frac{1}{2f_s}\sum_{n=1}^{N}[\sum_{i=L+1}^{H_1}[P_{XXn}(i)] + \sum_{i=L}^{H_1-1}[P_{XXn}(i)]] \quad (2)$$

$$FI_{MC} = \frac{P_H}{P_L} \quad (3)$$

where N is number of inputs, f_s is sampling frequency, $H_1 = \frac{3N_{FFT}}{f_s}$, $H_2 = \frac{8N_{FFT}}{f_s}$, $L = \frac{0.5N_{FFT}}{f_s}$.

We also extract another type of freeze index from $\mathbf{X}$, called FI_K, that results from a spectral analysis using the Koopman operator [26]. This operator was introduced to study the spectrum of Hamiltonian systems by using linear transformations on Hilbert space. Dynamic Mode Decomposition [35] is a technique to estimate a linear model with Koopman eigenfunctions and eigenvalues. Inspired by a feature extraction application in [36], Koopman eigenvalues and eigenfunctions are considered *frequencies* (λ) and the *power* (K(λ)); details of equations and algorithms as in [36]. Hence, we define FI_K as follows, $FI_K = \frac{\sum_{\lambda=H_1+1}^{H_2} K(\lambda)}{\sum_{\lambda=L+1}^{H_1} K(\lambda)}$ where $L = 0.5 \times 2\pi$, $H_1 = 3 \times 2\pi$, $H_2 = 8 \times 2\pi$.

2) Exploratory Pool: We construct a feature pool that consists of 244 features (Appendix A: Table III). The first half of the pool are 122 candidates, extracted using seven previously published features (i.e., average, standard deviation, variance, median, entropy, energy, power and *FI* as found in [12], [13], [14], [16], [17], [18], [19], [20], [23]) and our four aforementioned new features. We apply these eleven extraction functions to single and multiple inputs. Specifically, FI_{MC} and FI_K are applied to MCMS while the other functions are to SCSSs and the sum square of all three channels of single sensors. The second half of the pool consists of 122 anomaly score vectors (details as in the next section) of the above 122 features.

C. FoG Detector

We consider FoG events to be anomalies while the other events are normal data. ASD, a detector based on anomaly scores, is a simple way to combine features and produce an anomaly detector. If the feature value extracted from a window is higher than the current threshold, the window is labelled as a FoG event. The threshold can be calculated as in Eq. 4.

Let $\phi(n)$ be a value of feature vector of length N at time n. We define its anomaly score, $A(n)$, as follows:

$$A(n) = \text{sign}(\phi(n) - \frac{\alpha}{|n-1|}\sum_{m=1}^{n-1}[\phi(m)A(m)]) \quad (4)$$

where $A(1)$=1, $n \in [2, N]$, $\alpha > 0$ is a scale factor, and sign(x) is 1 if $x > 0$ else 0.

Intially, the first data window is assumed to be normal behaviour. If this assumption is wrong, we expect the averaging effect of Eq. 4 will low pass filter FoG events and eventually converge to a normal value. This work reports a simple case of Eq. 4 where $\alpha = 1$ (i.e no scaling deviation, other scales will be examined in another report specifically for that)

D. Feature Selection

We introduce a voting process to select the best feature from the large exploratory pool. This process uses three levels of selection: saliency, robustness, and accuracy; called *Round1, Round2, Round3* respectively (Fig. 2). After each level, selected candidates become more favourable. Specifically, *Round1* suggests the most salient and discriminative subset. Then, *Round2* examines if the candidates are robust across window sizes. Finally, *Round3* tests the detection performance of these features using our ASD.

In *Round1*, we rank feature candidates according to three saliency criteria, i.e., mutual information (MI), separability calculated using Euclidean distances (DIS), and the variance ratio of clusters (VarRatio). This step is implemented across 7 window sizes ($2 \rightarrow 8$ s in steps of $1s$), creating 21 lists of ranking scores. The range for window sizes is based on the minimum and maximum values currently suggested in the literature (e.g., 2 s in [20] and $7.5s$ in [19]). After finding a subgroup of high saliency score, we examine the robustness in *Round2*. Secondly, we identify salient candidates that are shared in more than one list across window sizes or criteria (i.e., robustness). Finally, we use accuracy metrics to find the subset for our ASD.

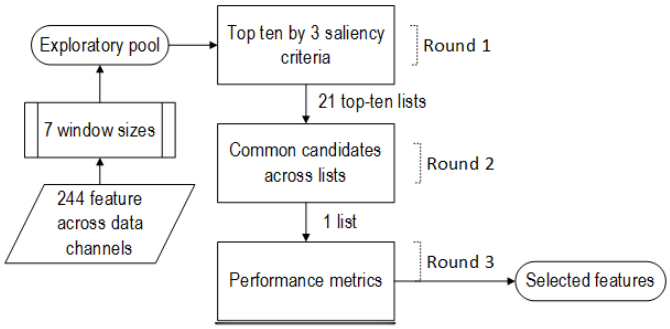


Figure 2. Feature selection process. 244 features as described in Table III. 7 window sizes are $2 \rightarrow 8$ s in steps of $1s$. Three saliency criteria are DIS, MI, VarRatio scores. Common candidates are entries that are shared by more than one list of *Round1*.

1) Saliency Criteria: Let X be a discrete random variable $X \in \mathbb{X}$ and C be a target variable ($c \in \mathbb{C}$, class label set). The *entropy* $H_b(X)$ of X measures its uncertainty [27]. The mutual information between X and C, $I(X;C)$, measures the relevance of X to C [27].

$$I(X;C) \stackrel{\text{def}}{=} H(X) - H(X|C) \quad (5)$$

$$= \sum_{x\in\mathbb{X}}\sum_{c\in\mathbb{C}} p(xc)\log\frac{p(xc)}{p(x)p(c)} \quad (6)$$

To compute the clusterability, we use the *RELIEF* algorithm [29] to calculate *DIS* scores (i.e., Euclidean distances between features and a *near-hit* or *near-miss* instance) [37]. The variance ratio of a feature X is the ratio of the between-cluster variance ($B_C(X)$) to the within-cluster variance ($W_C(X)$),

$V(X) \stackrel{\text{def}}{=} \frac{B_C(X)}{W_C(X)}$. The higher $V(X)$ implies that it is easier to cluster X [28], therefore the feature is more desirable.

2) Performance Metrics: In the literature, automatic techniques have been evaluated using different measures such as confusion matrices and/or intra-class correlations (ICCs) [38]. For instance, authors of [16], [18], [23] used timing-instance-based confusion matrices (i.e., counting FoG time frames and often involving a tolerance of milliseconds or seconds); and authors of [12], [13], [19], [20] used event-based confusion matrices (i.e., counting continuous FoG epochs) and ICCs on the number of FoG events or percentage of freezing time over a trial. With regard to real-time applications using wearable FoG detectors, the timing-based method is of most interest, whereas event-based is important in clinical FoG assessments. We utilize both types during feature selection as extra criteria (apart from saliency scores).

In our work, ICCs are used as supplemental criteria during *Round3* to select features rather than in performance comparisons with other works due to several limitations of ICC usages. First, low intra-rater reliability was reported for FoG number (0.44 (CI 0.18)). Secondly, at least two observers are recommended to analyse task videos [10]. In this work, information regarding the reliability for manual ratings were not available (nor were the number of raters). Thirdly, walking tasks were designed to have a single recording session per subject (about 30 minutes) rather than several short trial recordings (around one minute each). Hence, because in our data set the number of individual recordings is relatively small, thus, we group the data into one-minute segments. We assume that the segmentation is close to the multi-trials settings. Therefore, our estimation of ICC is a non-deceasing relationship with the reported ICC in the literature. Given two vectors of an automatic detection result and manual labels, we calculate the estimated intraclass correlation as in [39]; specifically we use the ICC(A-1) designation (two-way random effects) for the degree of absolute agreement among measurements.

With respect to the timing-based metrics, in confusion matrices, we refer to *ground truth* as the manual video analysis, and *positives* for FoG windows. True Positives (TP) are windows which were marked as FoG by both a test algorithm and the label. False Positives (FP) are windows labelled as FoG but did not agree with the *ground truth*. Windows that we failed to label as FoG but were annotated as such, are defined as False Negatives (FN). When the test method and the human agree a window was non-FoG, it is counted as a True Negative (TN). Please note that the reference labels used in this work were made by human thus are subjective. Likewise the literature works [16], [18], we investigate a tolerance, tol. Let t be the time instance an automated method decides it is FoG. If within the range of $[t - tol,\ t + tol]$, there is at least one instance where the reference (i.e., manual method) says it is FoG, we count this agreement is a true positive. Otherwise it is a false positive. Similarly for negative cases. The tolereance will be determined during the experiments using the performance curves (ROC).

Sensitivity and specificity are $\frac{TP}{TP+FN}$ and $\frac{TN}{TN+FP}$, respectively. F1-score, which is the harmonic mean of precision and sensitivity, with best value at 1 and worst at 0 [40], is calculated as $\frac{2TP}{(2TP+FP+FN)}$.

III. Results

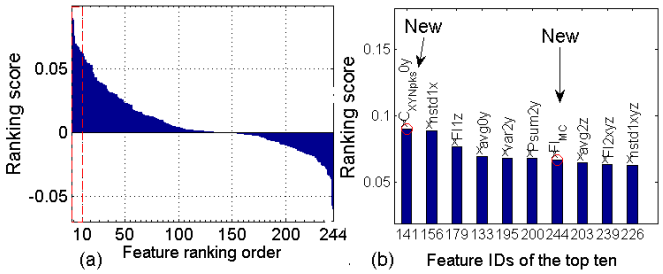
A. Selection by Saliency (Round1)

Fig. 3 illustrates feature ranking results (*Round1*) using window size of $2s$ and DIS saliency criterion. Details of results for other window sizes and saliency criteria can be found in the appendix B Fig. 5.

As can be seen, scores outside the top ten rank (the dotted vertical line) dropped quickly. Therefore, we selected these top ten candidates to create 210 input entries from 21 short-lists for the *Round2*. However, we also noticed that only 64 distinct features in the output of *Round1* (out of 244 candidates, See the appendix B Fig. 5 for the sharing of selected features among outputs).

In Figs. 3,5, new features were indicated with circle markers and labelled horizontal axes with feature identifications (IDs). Description of IDs can be found in the appendix A Table III. Specifically, the output of *Round1* includes *FI*0y (i.e., freezing index from ankle at vertical axis [12], [16], [18]) and previously proposed features (e.g., *FI*2y [19], *FI*2x [20], energy, sum of power $Psum$ [16], and their standard deviation, mean, variance [23]). Among the 64 distinct features, our new candidates, C_{XYNpks}, C_{XYmax}, FI_K, and FI_{MC}, were listed in more top-ranking lists than the existing ones.

Figure 3.
Example of feature ranking (a) and the shortlists (b) using DIS and window size of 2 s. Vertical axes: saliency scores. Horizontal axes: ranking order. The top-ten lists are in the dotted boxes. Features with circle markers are new while others are have been currently used in literature. The top ten identifications (IDs) of features are detailed in Table III. E.g., *FI*0y is the current popular existing feature. Continued at the appendix B Fig. 5.



B. Selection by Robustness (Round2)

In the second round, we consider candidates of *Round1* that are selected as the top ten in more than one list (across window sizes and/or criteria) robust features. There were 33 entries in *Round2* (i.e., about half of *Round1*). Interestingly, FI_{MC} is one of the most robust candidates in terms of being selective across window sizes (Table IV). Other new or popular existing candidates are also added in the table for comparison purposes.

C. FoG Detection Performance (Round3)

Fig. 4 and Table III present performance observations using a simple form of ASD (Section II-C). Our new features in this work and popular existing features (details in Table IV) were also included for comparison purposes.

Fig. 4 only showed seven candidates that had at least one report of $ICC > 0.2$ for freezing time percentage and number of FoG (suggestion from [10]). Specifically, these candidates are FI_{MC}, FI2y, FI2x, FI1z, FI0y, Mean 0z, and Mean 1z (Table III).

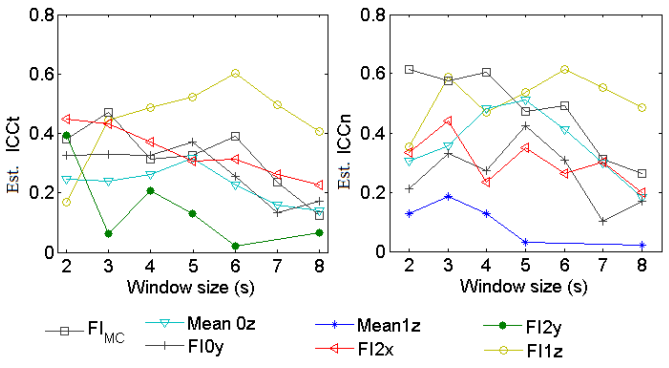
In terms of receiver operating characteristic (ROC), for each window size of each feature extraction with a timing tolerance range from $0 \rightarrow 1s$ in steps of $0.1s$, we observed that configurations FI_{MC} (3), $FI0y$(2s *or* 7s), $FI1z$(6), FI2y ($3s$ or $8s$), called *Round3*, had excellent results (Appendix C Fig. 6).

D. Tests and comparisons with the same cohort set

We then applied unseen test sets (five subjects who have been with PD for 11.2 ± 9.6 years; H&Y score: 2.9 ± 0.74) to validate ASD. Three subjects had FoG during data collection while the other two had no FoG. We noticed that, during the validation, FI_{MC} ($3\ s$) and FI2y ($3s$) had high accuracies with lowest deviation between development and out-of-sample tests (Table I). FI0y, a popular feature in existing detectors ($7\ s$ windows), achieved a sensitivity of 79% (specificity 79.5%) at a tolerance of $0.3s$. On the other hand, FI0y scores the highest F1-score of 84% with $2\ s$ window size (tolerance of $0.9\ s$). FI2y with $8\ s$ windows and $0.9\ s$ tolerance can achieve sensitivity of 87.5% (specificity of 84.5%).

Hence, we propose an optimized configuration for ASD as follows: window size as small as $3\ s$, tolerance for performance measurements of $0.4\ s$, freezing index is used for feature extraction. There was slightly preference of sensor locations between ankle and hip in terms of further performance improvement.

Figure 4. ICCs for feature selection. Markers are for different features. Left: Estimated ICC for the freezing time percentage. Right: Estimated ICC for number of FoG events.



E. External Validation Tests

Finally, using independent test sets that were from a different cohort to the one we used for development (Sec. II-A2), we validated our proposed ASD-based method (i.e., online ASD detector, freezing index feature, window size of $3\ s$). Though the performance improvement between ankle and hip sensor locations was not signifincant during the development stage, for a better comparison with existing works that used both types of inputs: single channel and multiple channels, such cases were still included in our report. Table II shows its high accuracy performance comparing with earlier works across several configurations of inputs.

IV. Discussion

During the development stage, we observed that beside the existing FI extracted from ankle sensor at vertical axis, our new feature with multiple channels, FI_{MC}, is one of the top features in saliency, clusterability, and robustness. Only seven out of 244 candidates met requirements of our three-round selection procedure. To detect FoG, we implemented an anomaly score based detector, ASD. With ASD, our features outperformed existing works with a small window and/or low tolerance. Specifically, FI2y, the freezing index from vertical data at a hip sensor, was found to be the best choice for performance; achieving sensitivity (specificity) of 87.5% (84.5%). FI_{MC}, is also a promising candidate. For example, FI_{MC} has high ICC and is the most robust candidates across window sizes during feature selection by saliency. FI_{MC} achieved a sensitivity of 81% (specificity of 77%) at the smallest tolerance of $0.2s$ ($3s$ windows).

During the test stage, we reported out-of-sample test outcomes in as many similar configurations as suggested from compared works. Our ASD that performed better than current methods can use only one type of feature extraction (freezing index) from a single channel. It is flexible and convenient to choose a sensor location between ankle and hip.

Regarding the system design, to the best of our knowledge, references [17], [18] achieved the best published performance to date for subject-independent settings. Specifically, with different reported configurations, these two methods used a *context recognition network* [16] and a Random Forest [18] with leave one out cross validation techniques (LOOCV). Other works used various *global* $\overline{FI}$ values with different channel selection. Note that, sensitivities and specificities in [12], [19], [20] were for event-based calculation that may differ from the others in Table II. Our algorithm can operate in an online manner, has low computational cost and latency is one window (3 sec for the best configuration during development). Furthermore, we demonstrated excellent subject-independent, unsupervised anomaly detection accuracy. As presented, our performance is significantly higher than the one of compared automatic detectors while using a much smaller window and/or lower tolerance. However, to assess the performance of our proposed algorithm against less advanced subjects (e.g., lower H & Y scores), we may need to recruit a more diverse population test set to further validate the method

Table I
DEVELOPMENT PERFORMANCE OF ASD USING FEATURES IN *Round3*. 'WIN': WINDOW SIZE. 'TOL': TOLERANCE. 'SD': STANDARD DEVIATION OF DEVELOPMENT AND OUT-OF-SAMPLE TEST. PERFORMANCE IN %.

Feature ID	Name Channel	Parameter		Development (%)		Out-of-sample (%)		Average ±SD (%)		
		Win.	Tol.	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	F1-score
244	$\boldsymbol{FI}_{MC}$	3s	0.2s	85	74.0	77.0	80.0	81.0 ±6	77.0±4	74.5± 6
139	*FI 0y*	2s	0.9s	88.0	81.0	86.0	63.0	87.0 ±1	72.0±13	84.0± 10
		7s	0.3s	71.0	93.0	87.0	66.0	79.0 ±11	79.5±19	82.5± 11
179	FI 1z	6s	0.1s	80.3	80.0	82.0	58.0	81.0 ±1	69.0±16	78.0± 6
199	FI2y	3s	0.4s	75.0	80.0	83.0	92.0	79.0 ±6	86.0±8	76.5± 12
		8s	0.9s	76.0	74.0	99.0	95.0	87.5 ±16	84.5±15	82.0± 24

Table II
OUT-OF-SAMPLE DETECTION PERFORMANCE OF ASD (VERUS EXISTING METHODS [16] [a] [18] [b] [19][20] [c] [d]) ACROSS CONFIGURATIONS [e] AND DATASETS [f] . PERFORMANCE IN %.

Method	Settings			Performance (%)		
	Input	*Win*	*Tol*	Sensitivity	Specificity	F1
CNR [16]	FI0y, Psum0y	4s	*2s*	73.1[a]	81.6[a]	-
Learning [18]	Mean0y, Std0y, FI0y, Energy0y	4s	1s	66.25 [b]	95.38 [b]	-
Global [19]	FI012y[d], $\overline{FI} = 3$	7.5s	-	84.3 [c]	78.4 [c]	-
Global [20]	FI2x, $\overline{FI} = 1.47$	2s	-	75.0 [c]	76.0 [c]	-
Online ASDs (proposed), external validation [f]						
ASD multi-inputs	$\boldsymbol{FI}_{MC}$	3s	0.4s	96 ±17	79 ±41	99 ±7
ASD ankle y-axis	FI0y	3s	0.4s	94 ±23	84 ±36	99 ±4
ASD hip y-axis	FI2y	3s	0.4s	89 ±32	82 ±39	96 ±18
ASD hip x-axis	FI2x	3s	0.4s	89 ±32	94 ±23	97 ±17

[a] as reported in [16] using CNR classifier and LOOCV.
[b] as reported in [18] using Random Forest classifier and LOOCV.
[c] for event-based calculation while others were for timing-based.
[d] the majority vote of *seven sensors* [19].
[e] *Input*: features, sensors, and axes. '*Tol*': tolerance. '*Win*': window size.
[f] 71 trials of 15 subjects; different cohort to the development set (same to the work [19]) .

V. CONCLUSION

In this work, we studied the important task of FoG detection. We reported new features and a detection scheme that are more relevant in identification of freezing occurrences in subject-independent settings. Freezing index feature from single channel (x or y axis) at ankle or hip sensor location can be used for an anomaly detection based scheme to detect FoG events. Our proposed method is objective and significantly outperforms (e.g., mean (±SD) of sensitivity, specificity are 94% ($\pm 23\%$) and 84% ($\pm 36\%$) for *ASD ankle y-axis*) other automated methods in the literature.
In future work, a combination of these two candidates should be further evaluated. A more elaborate technique for the ASD threshold is also worthy of further study. These findings form a further step towards subject-independent out-of-lab FoG detectors.

ACKNOWLEDGMENT

Funding resources: Endeavour/Prime Minister's Australia Scholarship; the Faculty Research Cluster Program at The University of Sydney; and NHMRC-ARC Dementia Research Development Fellowship 1110414.

APPENDIX A
LIST OF FEATURES

The feature pool consists of 244 features (Table III). There are eleven extraction functions: seven previously published and our four new methods. We apply these eleven extraction functions to single and multiple inputs.

APPENDIX B
DETAILS OF FEATURE RANKING

Fig. 5 illustrates three types of ranking scores (i.e., MI, DIS, and VarRatio) across window sizes for each feature candidate (sorted from high to low scores). The order of ranking is from 1 to 244 (high to low); a higher saliency score indicated the higher ranking order. The other window sizes shared a similar trend.

APPENDIX C
PERFORMANCE AGAINST TOLERANCES AND WINDOWS

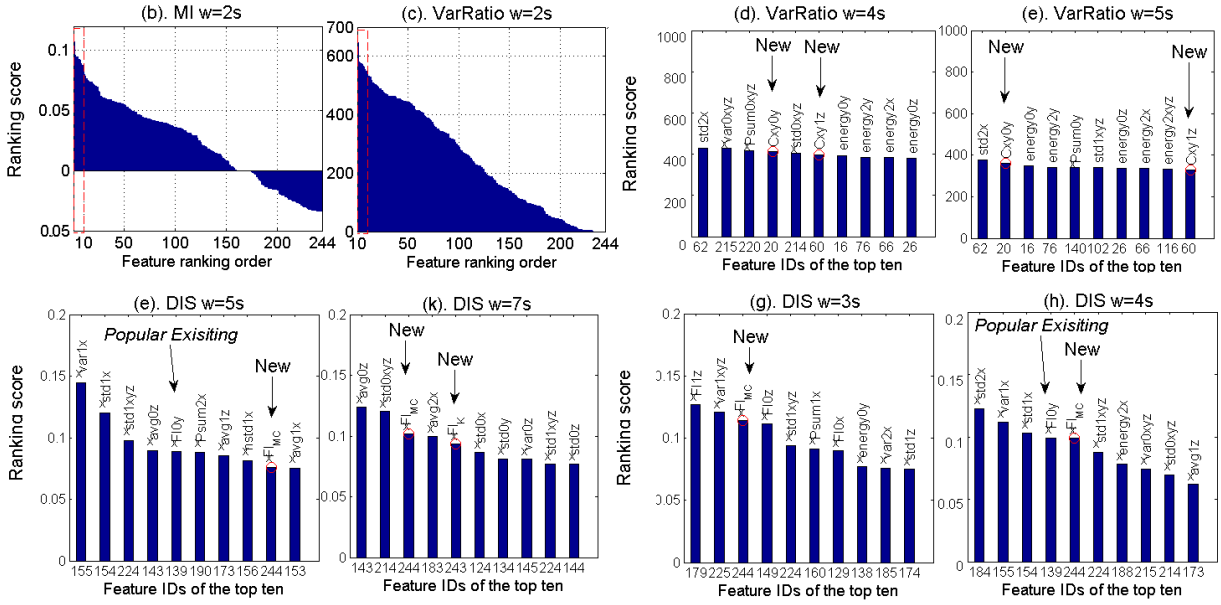
Fig. 6 presents performance metrics across window sizes of each feature extraction with a timing tolerance range from $0 \rightarrow 1s$ in steps of $0.1s$. Due to the difficulty of displaying ROCs across many variables, F1-scores were presented.

Table III

FEATURE POOL EXAMINED IN THIS WORK. THREE TYPES OF DATA INPUTS: SCSS, MCSS, MCMS. IDENTIFICATIONS OF SENSORS AND AXES DESCRIBE FOR EACH CHANNEL USED TO EXTRACT FEATURES. EIGHT EXISTING FUNCTIONS AND TWO NEW OF C_{XY} ARE FOR SCSS AND MCSS. FOR MCMS, TWO NEW FEATURES ARE FI_K (BY KOOPMAN SPECTRAL ANALYSIS) AND FI_{MC} (BY FOURIER TRANSFORM EQ. 1). FEATURE IDS 123 $\rightarrow$ 244 ARE CORRESPONDING ANOMALY SCORE BASED FEATURES OF THE ABOVE 1 $\rightarrow$ 122.

	SCSS									MCSS			MCMS	
Sensor	0 (ankle)			1 (knee)			2 (back)			0	1	2	0,1,2	0,1,2
Axis	x	y	z	x	y	z	x	y	z	$\sqrt{x^2+y^2+z^2}$			x,y,z	x,y,z
Extraction	Average, standard deviation, variance, median, entropy, energy, power, FI, C_{XYNpks}, and C_{XYmax}												FI_K	FI_{MC}
IDs	1:10	11:20	21:30	31:40	41:50	51:60	61:70	71:80	81:90	91:100	101:110	111:120	121	122

Figure 5. Example of feature ranking and the shortlists. (a,b,c) are ranking scores for the feature pool; (a). DIS, (b). MI, (c). VarRatio scores. Vertical axes: saliency scores. Horizontal axes: ranking order. The top-ten lists are in the dotted boxes. The others, (d-k), illustrate the sharing among shortlists across window sizes and criteria. Features with circle markers are new while others are have been currently used in literature. The top ten identifications (IDs) of features are detailed in Table III. E.g., $FI0y$ is the current popular existing feature.



REFERENCES

[1] B. Bloem *et al.*, "Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena", *Mov disord*, vol. 19, no. 8, pp. 871–884, Aug. 2004.

[2] M. Macht *et al.*, "Predictors of freezing in Parkinson's disease: a survey of 6,620 patients", *Mov disord*, vol. 22, no. 7, pp. 953–956, May 2007.

[3] M. Latt *et al.*, "Clinical and physiological assessments for elucidating falls risk in Parkinson's disease", *Mov disord*, vol. 24, no. 9, pp. 1280–1289, Jul. 2009.

[4] S. Paul *et al.*, "Three simple clinical tests to accurately predict falls in people with Parkinson's disease", *Mov disord*, vol. 28, no. 5, pp. 655–662, May 2013.

[5] S. Fahn and R. Elton, "Unified rating scale for Parkinson's disease", *Recent developments in Parkinson's disease. florham park. new york: Macmillan*, pp. 153–163, 1987.

[6] N. Giladi *et al.*, "Validation of the freezing of gait questionnaire in patients with Parkinson's disease", *Movement disorders*, vol. 24, no. 5, pp. 655–661, 2009.

[7] A. H. Snijders *et al.*, "Obstacle avoidance to elicit freezing of gait during treadmill walking", *Movement disorders*, vol. 25, no. 1, pp. 57–63, 2010.

[8] C. Moreau *et al.*, "Externally provoked freezing of gait in open runways in advanced parkinson's disease results from motor and mental collapse", *Journal of neural transmission*, vol. 115, no. 10, pp. 1431–1436, 2008.

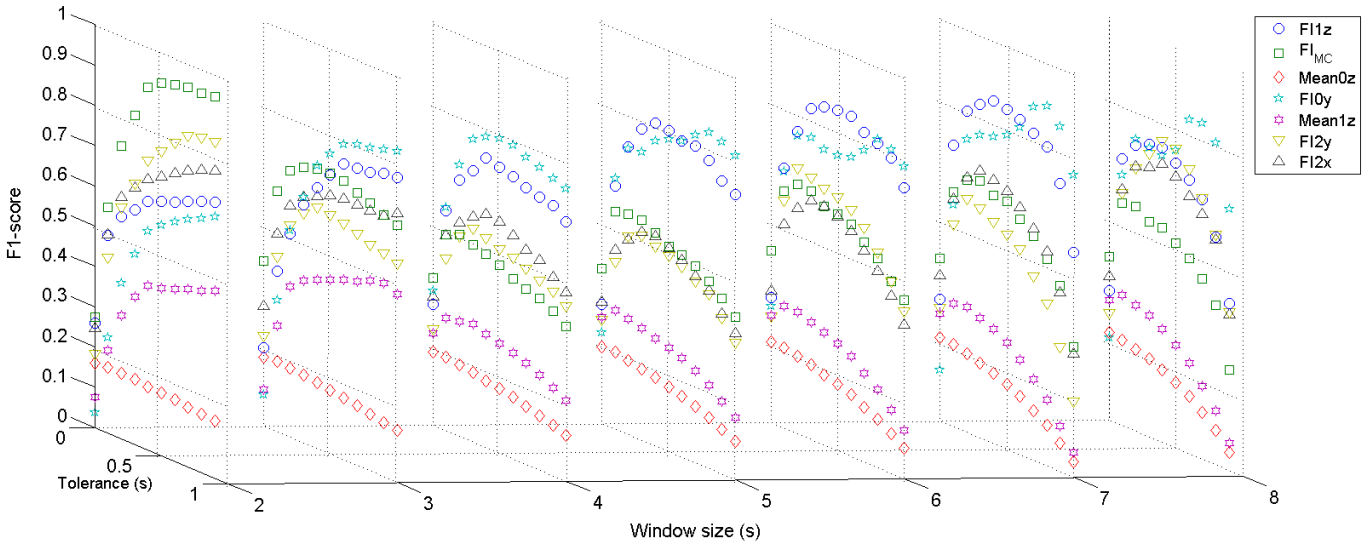
[9] J. Reimer *et al.*, "Use and interpretation of on-off diaries in Parkinson's disease", *J neurol neurosurg psychiatr*, vol. 75, no. 3, pp. 396–400, Mar. 2004.

[10] T. R. Morris *et al.*, "A comparison of clinical and objective measures of freezing of gait in Parkinson's disease", *Parkinsonism & related disorders*, vol. 18, no. 5, pp. 572–577, 2012.

[11] J. Schaafsma *et al.*, "Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease", *Eur j neurol*, vol. 10, no. 4, pp. 391–398, Jul. 2003.

[12] S. Moore *et al.*, "Ambulatory monitoring of freezing of gait in Parkinson's disease", *J neurosci methods*, vol. 167, no. 2, pp. 340–348, Jan. 2008.

Figure 6. Effects of window sizes and tolerances on F1-scores of ASD. Tolerance from $0s \rightarrow 1s$. Three dimensional view for planes of window sizes from $2s \rightarrow 8s$.Markers are for different features.



[13] E. Gazit *et al.*, "Assessment of Parkinsonian motor symptoms using a continuously worn smartwatch: Preliminary experience", *Movement disorders*, vol. 30, S272–S272, 2015.

[14] J. Han *et al.*, "Gait analysis for freezing detection in patients with movement disorder using three dimensional acceleration system", *Engineering in medicine and biology society, proceedings of the 25th annual international conference of the ieee*, vol. 2, 1863–1865 Vol2, 2003.

[15] I. Daubechies and B. J. Bates, "Ten lectures on wavelets", *The journal of the acoustical society of america*, vol. 93, no. 3, pp. 1671–1671, 1993.

[16] M. Bachlin *et al.*, "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom", *Information technology in biomedicine, ieee transactions on*, vol. 14, no. 2, pp. 436–446, 2010.

[17] B. Cole *et al.*, "Detecting freezing-of-gait during unscripted and unconstrained activity", *Engineering in medicine and biology society,embc, annual international conference of the ieee*, pp. 5649–5652, 2011, ISSN: 1557-170X.

[18] S. Mazilu *et al.*, "Online detection of freezing of gait with smartphones and machine learning techniques", *Pervasive computing technologies for healthcare (pervasivehealth), 6th international conference on*, pp. 123–130, 2012.

[19] S. T. Moore *et al.*, "Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry", *Journal of neuroengineering and rehabilitation*, vol. 10, no. 1, p. 1, 2013.

[20] H. Zach *et al.*, "Identifying freezing of gait in Parkinson's disease during freezing provoking tasks using waist-mounted accelerometry", *Parkinsonism & related disorders*, vol. 21, no. 11, pp. 1362–1366, 2015.

[21] C. Azevedo Coste *et al.*, "Detection of freezing of gait in parkinson disease: Preliminary results", *Sensors*, vol. 14, no. 4, pp. 6819–6827, 2014.

[22] M. D. Djurić-Jovičić *et al.*, "Automatic identification and classification of freezing of gait episodes in parkinson's disease patients", *Ieee transactions on neural systems and rehabilitation engineering*, vol. 22, no. 3, pp. 685–694, 2014.

[23] S. Mazilu *et al.*, "Feature learning for detection and prediction of freezing of gait in Parkinson's disease", in, Springer, 2013, pp. 144–158.

[24] E. Sejdić *et al.*, "A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains", *Ieee transactions on neural systems and rehabilitation engineering*, vol. 22, no. 3, pp. 603–612, 2014.

[25] R. Challis and R. Kitney, "Biomedical signal processing (part 3 of 4):the power spectrum and coherence function", *Medical and biological engineering and computing*, vol. 28, no. 6, pp. 509–524, 1990.

[26] B. O. Koopman, "Hamiltonian systems and transformation in hilbert space", *Proceedings of the national academy of sciences of the united states of america*, vol. 17, no. 5, p. 315, 1931.

[27] C. Shannon, "A mathematical theory of communication", *Bell system technical journal, the*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[28] M. Ackerman and S. Ben-David, "Clusterability: A theoretical study", *International conference on artificial intelligence and statistics*, pp. 1–8, 2009.

[29] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm", *Proceedings of the tenth national conference on artificial intelligence*, AAAI'92, pp. 129–134, 1992.

Table IV
33 TOP SALIENT AND ROBUST FEATURES (*Round2*) AND FOUR OTHERS OF INTEREST. IDS ARE IDENTIFICATIONS OF FEATURES. 'STD': STANDARD DEVIATION. DIS, MI, AND VARRATIO ARE CRITERIA.

Feature ID	Name	Sensor, Channel	Window sizes (second)		
			DIS	MI	VarRatio
244	$\boldsymbol{FI}_{MC}$	all	**all**	-	-
194	Std	2,y	-	all	2
124	Std	0,x	8	all	3
214	Std	0,xyz	4,8	all	4,6,7,8
134	Std	0,y	8	all	6,8
154	Std	1,x	4,5	all	-
174	Std	1,z	3	all	-
184	Std	2,x	4, 5, 6	2	all
138	Energy	0,y	3	-	all
98	Psum	0xyz	-	$3 \rightarrow 7$	-
224	Std	1xyz	$3 \rightarrow 8$	-	-
155	Variance	1x	$4 \rightarrow 7$	-	-
198	Energy	2y	6	-	$2 \rightarrow 5$,7
26	Energy	0z	-	-	$2 \rightarrow 5$,7
102	Std	1xyz	-	$2 \rightarrow 5$	5,7
164	Std	1y	7	$2 \rightarrow 5$,8	-
215	Variance	0xyz	-	-	$4 \rightarrow 8$
188	Energy	2x	4	-	$2 \rightarrow 5$
112	Std	2xyz	6	$2 \rightarrow 5$	-
175	Variance	1z	7	6,7,8	3,6
158	Energy	1x	7	-	6,7,8
22	Std	0z	7,8	6,7,8	-
116	Energy	2xyz	-	-	2,3,5
135	Variance	0y	-	-	6,8
93	Variance	0xyz	-	6,7,8	-
185	Variance	2x	3,6	-	-
139	*FI* [12]	*0y*	*4,5*	-	-
173	Mean	1z	4,5	-	-
179	FI	1z	2,3	-	-
225	Variance	1xyz	3,7	-	-
143	Mean	0z	5,8	-	-
20	*Cxymax*	*0y*	-	-	*4,5*
60	*Cxymax*	*1z*	-	-	*4,5*
195	Variance	2y	2,6	-	2
Other new or existing features for comparison purposes					
141	*CxyNpks*	*0y*	*2*	-	-
243	FI_K	all	*8*	-	-
199	FI [19]	2y	6	-	-
189	FI [20]	2x	-	-	-

[30] J. Shine *et al.*, "Assessing the utility of freezing of gait questionnaires in parkinson's disease", *Parkinsonism & related disorders*, vol. 18, no. 1, pp. 25–29, 2012.

[31] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results", *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[32] M. M. Hoehn, M. D. Yahr, *et al.*, "Parkinsonism: Onset, progression, and mortality", *Neurology*, vol. 50, no. 2, pp. 318–318, 1998.

[33] W. Gibb and A. Lees, "A comparison of clinical and pathological features of young-and old-onset parkinson's disease", *Neurology*, vol. 38, no. 9, pp. 1402–1402, 1988.

[34] M. F. Folstein *et al.*, ""mini-mental state": A practical method for grading the cognitive state of patients for the clinician", *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[35] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data", *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.

[36] J.-C. Hua *et al.*, "Using dynamic mode decomposition to extract cyclic behavior in the stock market", *Physica a: Statistical mechanics and its applications*, vol. 448, pp. 172–180, 2016.

[37] G. Brown *et al.*, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection", *The journal of machine learning research*, vol. 13, no. 1, pp. 27–66, 2012.

[38] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability", *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[39] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients.", *Psychological methods*, vol. 1, no. 1, p. 30, 1996.

[40] C. J. V. Rijsbergen, *Information retrieval*, 2nd. Newton, MA, USA: Butterworth-Heinemann, 1979.