Smartphone- and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis During the Two-Minute Walk Test

Andrew P. Creagh, Cedric Simillion, Alan K. Bourke, Alf Scotland, Florian Lipsmeier, Corrado Bernasconi, Johan van Beek, Mike Baker, Christian Gossens, Michael Lindemann, and Maarten De Vos

***Abstract*—Leveraging consumer technology such as smartphone and smartwatch devices to objectively assess people with multiple sclerosis (PwMS) remotely could capture unique aspects of disease progression. This study explores the feasibility of assessing PwMS and Healthy Control's (HC) physical function by characterising gait-related features, which can be modelled using machine learning (ML) techniques to correctly distinguish subgroups of PwMS from healthy controls. A total of 97 subjects (24 HC subjects, 52 mildly disabled (PwMSmild, EDSS [0–3]) and 21 moderately disabled (PwMSmod, EDSS [3.5–5.5]) contributed data which was recorded from a Two-Minute Walk Test (2MWT) performed out-of-clinic and daily over a 24-week period. Signal-based features relating to movement were extracted from sensors in smartphone and smartwatch devices. A large number of features ($n = 156$) showed fair-to-strong ($R > 0.3$) correlations with clinical outcomes. LASSO feature selection was applied to select and rank subsets of features used for dichotomous classification between subject groups, which were compared using Logistic Regression (LR), Support Vector Machines (SVM) and Random Forest (RF) models. Classifications of subject types were compared using data obtained from smartphone, smartwatch and the fusion of features from both devices. Models built on smartphone features alone achieved the highest classification performance, indicating that accurate and remote measurement of the ambulatory characteristics of HC and PwMS can be achieved with only one device. It was observed however that smartphone-based performance was affected by inconsistent placement location (running belt versus pocket). Results show that PwMSmod could be distinguished from HC subjects (Acc. 82.2 ± 2.9%, Sen. 80.1 ± 3.9%, Spec. 87.2 ± 4.2%, F_1 84.3 ± 3.8), and PwMSmild (Acc. 82.3 ± 1.9%, Sen. 71.6 ± 4.2%, Spec. 87.0 ± 3.2%, F1 75.1 ± 2.2) using an SVM classifier with a Radial Basis Function (RBF). PwMSmild were shown to exhibit HC-like behaviour and were thus less distinguishable from HC (Acc. 66.4 ± 4.5%, Sen. 67.5 ± 5.7%, Spec. 60.3 ± 6.7%, F_1 58.6 ± 5.8). Finally, it was observed that subjects in this study demonstrated low intra- and high inter-subject variability which was representative of subject-specific gait characteristics.**

***Index Terms*—Gait, machine learning, multiple sclerosis, sensor-based measure, smartphone, smartwatch.**

I. Introduction

MULTIPLE Sclerosis (MS) is a progressive neurodegenerative disease that is typically diagnosed in young adults, causing varied and unpredictable physical and mental disability and neurological deterioration over time [1]. Ambulatory function have been perceived as the most prominent physical impairments in people with multiple sclerosis (PwMS) [2], who often have postural instability [3], gait abnormalities [4] and pronounced gait variability [5] that can manifest at different stages of diseases progression. Many studies hint at the strong predictive nature in alterations during ambulation (gait) due to MS [4], [6], [7]. Some commonly used measures for assessing the disease state of PwMS are a combination of clinician-administered rating scales, such as the Expanded Disability Status Scale (EDSS) [8] and patient-reported outcomes such as the Multiple Sclerosis Impact Scale-29 (MSIS-29) and Multiple Sclerosis Walking Scale-12 (MSWS-12) [9]. The Timed 25-Foot Walk (T25FW), developed as part of the Multiple Sclerosis Functional Composite score [10], [11], and the Two-Minute Walk Test (2MWT) are used to assess physical gait function and fatigue in PwMS. The 2MWT outcome is typically reported as distance travelled [12], [13]. These clinically administered measures however have a number of limitations, such as: low intra- and inter-rater reliability [14], in addition to an infrequent

Manuscript received November 17, 2019; revised March 23, 2020 and May 11, 2020; accepted May 19, 2020. Date of publication May 28, 2020; date of current version March 5, 2021. *(Michael Lindemann and Maarten De Vos are co-last authors.) (Corresponding author: Andrew P. Creagh.)*

Andrew P. Creagh is with the Institute of Biomedical Engineering, University of Oxford, Oxford OX1 2JD, U.K., and also with F. Hoffmann-La Roche Ltd., 4070 Basel, Switzerland (e-mail: andrew.creagh@eng.ox.ac.uk).

Cedric Simillion, Alan K. Bourke, Alf Scotland, Florian Lipsmeier, Corrado Bernasconi, Johan van Beek, Mike Baker, Christian Gossens, and Michael Lindemann are with F. Hoffmann-La Roche Ltd., 4070 Basel, Switzerland (e-mail: cedric.simillion@roche.com; alan.bourke@roche.com; alf.scotland@roche.com; florian.lipsmeier@roche.com; corrado.bernasconi@roche.com; johan.van_beek@roche.com; mike.baker.mb1@roche.com; christian.gossens@roche.com; michael.lindemann@roche.com).

Maarten De Vos is with the Institute of Biomedical Engineering, University of Oxford, Oxford OX1 2JD, U.K., with the Department of Electrical Engineering, KU Leuven 3000, Leuven, Belgium, and also with the Department of Development and Regeneration, KU Leuven 3000, Leuven, Belgium (e-mail: maarten.devos@eng.ox.ac.uk).

Digital Object Identifier 10.1109/JBHI.2020.2998187

TABLE I
POPULATION DEMOGRAPHICS

	HC (n = 24)	PwMSmild (n = 52)[a]	PwMSmod (n = 21)[b]	P-value HC vs PwMSmild[1]	P-value HC vs PwMSmod[1]
Age	35.6 ± 8.9	39.3 ± 8.3	40.5 ± 6.9	0.12	0.07
Sex (M/F)	18/6	16/36	7/14	< 0.001[2]	< 0.01[2]
EDSS		1.7 ± 0.8	4.2 ± 0.7		
EDSS (amb.)		0.1 ± 0.3	1.9 ± 1.5		
T25FW [s]	5.0 ± 0.9	5.3 ± 0.9	7.9 ± 2.2	0.26	< 0.001
Total # smartphone tests	1926	5498	2024		
Total # smartwatch tests	1436	4258	1502		
Total # linked tests	1362	3921	1452		
# tests per subject	57.6 ± 47.6	76.4 ± 45.7	70.1 ± 48.2	0.12	0.36
Total # running belt tests	905	2424	1296		
# tests per subject	37.7 ± 42.7	46.6 ± 41.8	61.7 ± 48.8	0.41	0.08
Total # pocket tests	457	1497	156		
# tests per subject*	30.5 ± 34.4	36.5 ± 37.3	11.1 ± 11.5	0.62	0.18

Clinical scores taken as the average per subject over the entire study, where the mean ± standard deviation across population are reported; EDSS, Expanded Disability Status Scale; T25FW, the Timed 25-Foot Walk; EDSS (amb.) refers to the ambulation sub-score as part of the EDSS; [s], indicates measurement in seconds;
[a]PwMS with average EDSS [0–3]; [b]PwMS with average EDSS [3.5–5.5];
*HC (n = 15), PwMSmild (n = 41), PwMSmod (n = 14);
[1]Mann-Whitney U Test; [2]Chi-squared (χ^2) test.

administration, which can miss episodic manifestations of disease. In recent years there has been a shift towards the adoption of body worn inertial sensors to more objectively evaluate gait performance [7], [15]–[18]. Upper and lower body characteristics related to dynamic balance during ambulation have been captured from inertial sensors affixed to the wrists, shank and trunk, which were found to significantly differentiate PwMS and HCs ($p < 0.05$) compared to standard stop-watch timed tests such as the T25FW and Timed-Up-and-Go (TUG) test [7]. It has also been shown that PwMS have higher gait feature variability than HC [6], [18]. Greene *et al.* have demonstrated that PwMS can be distinguished from HC by modelling gait features from shank mounted inertial sensors using a cross-sectional analysis of the TUG test [17]. Many of these studies however asses ambulatory ability using multiple inertial sensors during fixed lengths of controlled walking, in-clinic. Consumer wearable sensors (such as smartphone and smartwatches embedded with inertial sensors) offer a unique opportunity to monitor physical function ubiquitously, more subtly and remotely in PwMS [19]. Furthermore, high-frequency monitoring assessments may be more accurate than conventional outcomes recorded at periodic visits in detecting subtle progressive sub-clinical changes that may predict disease activity or long-term disability in PwMS [20]. Earlier identification of changes in PwMS impairment are important to identify and provide better therapeutic strategies [21]. The "Monitoring of Multiple Sclerosis (MS) Participants With the Use of Digital Technology (Smartphones and Smartwatches) - A Feasibility Study" (NCT02952911) was a study to assess the feasibility of remote patient monitoring using smartphones and smartwatch devices applying a range of testing modalities in PwMS and HC [22], [23]. This paper applies feature-based approaches to characterise gait function in PwMS using remotely captured sensor-data from the 2MWT. Machine learning (ML) techniques are then used to distinguish subgroups of PwMS and HC as dichotomous classification tasks.

II. METHODS

A. Dataset

PwMS and HC enrolled in this study were requested to perform the 2MWT daily over a 24-week period. Subjects were assessed clinically during site-visits at baseline, week 12 and week 24. Further information on NCT02952911, including 2MWT instructions[1], adherence results and more detailed demographics can be found at [22]. Each subject was also provided with a waist-worn running belt and instructed to attach the smartphone to the anterior of their waist. The smartwatch can be worn on either wrist. Subjects with both smartphone and smartwatch data available ($n = 97$) are presented in Table I. To allow comparisons between smartphone and smartwatch devices, only test instances where subjects have used both devices during their 2MWT were included in this study. MS is a heterogeneous disease, and in order to differentiate subjects with presumed gait symptoms, subjects were divided into subgroups (mild and moderate) based on their mean EDSS: PwMSmild ($n = 52$, EDSS [0–3]), and PwMSmod ($n = 21$, EDSS [3.5–5.5]), using a similar threshold to other MS gait studies [18]. EDSS is considered a primary outcome for assessing the disease state of PwMS [8], [24]. By definition, gait disorders begin to become prominent in subjects with EDSS ≥ 3.5 and subjects with EDSS < 3.5 are mildly impaired [8]. Note: the entire range of subjects' pooled EDSS scores in this study was [0–7]. Differences in clinical characteristics were analysed using the Mann-Whitney U Test, except categorical differences in sex which were investigated using a Chi-squared (χ^2) test.

B. Feature Extraction

1) Pre-Processing: Subjects were provided with a Samsung Galaxy S7 smartphone and Motorola 360 Sport smartwatch.

[1]While the instructions given were analogous to those of a 2MWT, this was not a controlled and clinically assessed 2MWT during site-visits and therefore the outcome of walking distance was not measured.

Both smartphone and smartwatch devices contain 3-axis accelerometer $(\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z)$ and gyroscope $(\mathbf{g}_x, \mathbf{g}_y, \mathbf{g}_z)$ sensors which were sampled at 50 Hz. Signals were filtered with a 4^{th} order butterworth filter with a cut-off frequency at 17 Hz [17], [25]. Orientation of the smartphone can be determined by assessing through which axis the mean component of gravity is incident upon during the 2MWT. Prior to windowing, the sensor coordinate frame was aligned with the global reference frame using the technique described in [26] and thus the anterior-posterior axis ($\mathbf{x}$–) was aligned with the direction of motion, which was orthogonal to the vertical ($\mathbf{y}$–) and medial-lateral ($\mathbf{z}$–) axis. Features were computed on all sensor axes and also on the orientation invariant signal magnitude, for example $\|\mathbf{a}\| = (\mathbf{x}^2 + \mathbf{y}^2 + \mathbf{z}^2)^{\frac{1}{2}}$, where $\mathbf{x} = (a_{x_1}, a_{x_2}, \ldots, a_{x_T})$ and so forth. Subjects' whole 2MWT tests were then windowed into non-overlapping 30 second epochs to help minimise potential signal artefacts, including subject turns during the test. Bouts of non-gait were filtered using methods described in [25]. Features were extracted on each epoch and the mean value and standard deviation per 2MWT were taken. The 2MWT was not clinically assessed in this study during site-visits and as such the outcome of walking distance [12], [13] is unavailable. Step counts have been proposed to approximate the distance travelled [27], and smartphone step count has also been shown to estimate the walking distance in HC over a fixed length during the six-minute walk test (6MWT) [28]. Subsequently, this study implemented a step count, using methods described by Lee *et al.* [29], to roughly approximate the 2MWT walking distance for comparative purposes.

2) Energy Features: The continuous wavelet transform (CWT) can measure the similarity between a discrete signal and an analysing function, providing a precise time-frequency representation of a signal [30]. It has been shown that the Morlet wavelet effectively captures gait-related spatio-temporal characteristics from acceleration signals obtained from different body locations [31]. A sparse representation of gait signals was also obtained using the Discrete Wavelet Transform (DWT) where the signal was decomposed into a number of different bandwidths expressed by approximation and detail coefficients on which features were computed. We extracted the wavelet coefficients experimenting with three wavelet families (Daubechies, Symlets, Coiflets) [32]. At each decomposition level, or bandwidth, we computed the energy, entropy (using both Shannon's and the log energy definitions), and the Teager-Kaiser Energy Operator (TKEO) on approximation (cA) and detail (cD) coefficients. Wavelet Energy (both using the discrete CWT and DWT representation) is defined as:

$$E(x) = \sum_{i=1}^{N} |x_i|^2 \tag{1}$$

Wavelet (non-normalised) Shannon Entropy is defined as:

$$H(x) = -\sum_{i=1}^{N} x_i^2 \log(x_i^2) \tag{2}$$

where $x = cD_j$; $x = cA_j$ are the detail and approximation coefficients at level $j = 1, 2, 3, \ldots, L$;

Empirical Mode Decomposition (EMD) has been used previously to characterise the frequency range distributions of gait rhythms from accelerometer signals [33]. Classical EMD decomposes a signal into a small finite number of intrinsic mode functions (IMFs) using the Hilbert-Huang transform (HHT) to encode instantaneous frequency and amplitude information [34]. IMFs offer a data driven approach to analyse non-linear, non-stationary signals. The energy (1) of each IMF is computed, where the first IMF represents the "high-frequency (noise)" components with the latter IMFs capturing the relatively "low-frequency (signal)" components of gait rhythm. The relative signal-to-noise ratio (SNR) is then computed as a feature to characterise the ratio of gait to higher frequency perturbations in the sensor signal.

3) Statistical and Entropy Features: A number of statistical features were also computed on the sensor signals such as the mean, standard deviation, skewness, kurtosis, zero-crossing rate and auto-correlation coefficients.

Multiscale entropy (MsEn) calculates the sample entropy (SampEn) of a signal at increasingly coarser grains (scales) [35]. MsEn is advantageous to entropy alone in that calculating the entropy of a signal at multiple time scales discriminates long-range correlations in complex systems from completely random signals. Costa *et al.* [35], for example, found that faster and unconstrained walking had more complex dynamics than slower walking, as captured through greater SampEn at different scales. Further to calculating raw MsEn over the first 20 time scales, higher order MsEn-based statistical features were also computed. Similar entropy parameters of embedding dimension $m = 2$, and tolerance $r = 0.2$ were used [35]. Recurrence period density entropy (RPDE) is a method used to characterise the deviations from exact periodicity and stochasticity within a signal [36], proposed here to capture the ability to maintain consistent gait rhythm.

Supplementary material, including a full description and implementation of all of the features extracted in this study can be found at: https://github.com/apcreagh/MS-GAIT_feature_extraction.

C. Feature Analysis

Univariate feature differences, taken as the median feature value per subject over all available test observations, were investigated using the Mann-Whitney U Test for each paired subject group combination (HC, PwMSmild, and PwMSmod). Associations between mean clinical metrics (EDSS and T25FW) and median feature values per subject were investigated using Spearman's correlation (R_s). Differences in feature distributions between smartphone locations identified as running belt or pocket were investigated using a Mann-Whitney U Test, for subjects who contributed both placements only. In order to reduce our highly dimensional feature space (features derived from 2 devices, with 3 axes per sensor) prior to model building all redundant features were first removed based on R_s to subject group (HC, PwMSmild, PwMSmod) < 0.3 ($n = 93$ and $n = 63$ respective smartphone and smartwatch features retained). P-values were corrected for multiple hypothesis testing using

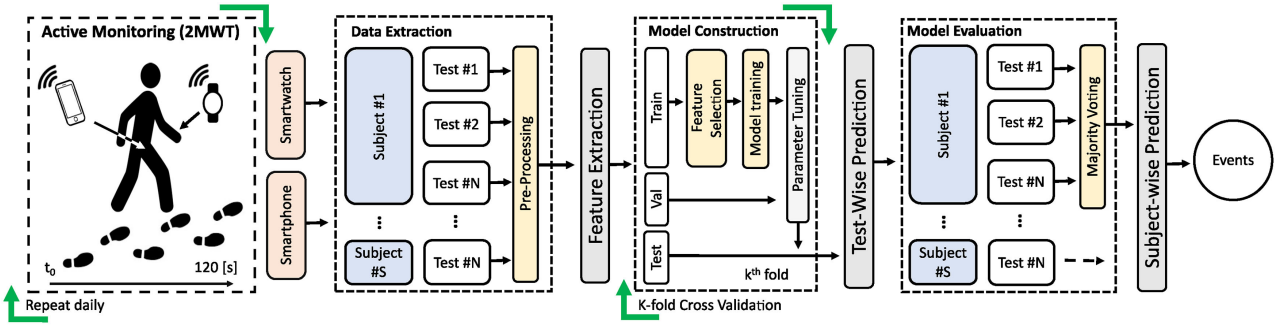


Fig. 1. Schematic of gait processing pipeline. Users are requested to perform a remote 2MWT daily, for up to 24 weeks, with a smartphone affixed within a waist-worn running belt on the anterior of their waist and smartwatch on their wrist. Sensor data is extracted from both devices independently and signal-based features are computed on each 2WMT per subject. Classification models are then constructed, tuned and evaluated using subject-wise k-fold cross-validation. Individual 2MWT predictions per subject are then majority voted to generate a single prediction per subject, which are used to distinguish subgroups of PwMS and HC as binary classification tasks.

methods described in [37]. Exploratory analysis of data and feature structure was performed using principal component analysis (PCA) [38]. Feature reproducibility was investigated using the intraclass correlation coefficient (ICC) metric (see appendix A for more details).

D. Model Construction

A number of machine learning (ML) techniques were explored in order to asses the ability to discriminate HC from PwMS sub-groups as binary classification tasks. Model generalisability was determined using 5-fold, subject-wise, stratified, cross-validation (CV). After partitioning the data into training sets, observations were randomly re-sampled to balance class distributions, as subject each contributed unequal quantities of 2MWT observations. CV was repeated 10 times to reduce biases in re-sampling and dataset splitting.

Logistic Regression (LR) was compared to Support Vector Machines (SVM) and a Random Forest classifier (RF) [39]. LASSO regularisation for generalised linear models (*lassoglm*) was employed in order to reduce the dimensions of the extracted feature space into a ranked parsimonious set [39]. In this case $lassoglm$ is an extension of LASSO which uses a logit link function: $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta^{\intercal}x$, yielding a posterior probability mapping binomial responses. Features were ranked per CV fold by increasing shrinkage regularisation parameter λ, and cumulatively presented to LR and SVM classifiers. A top feature ranking table was deduced by interrogating the feature subsets selected by $lassoglm$ at each fold and repetition. The relative stability of features selected was assessed by recording the percentage of time that the feature was selected in the top 5 and top 25 features at each fold and repetition.

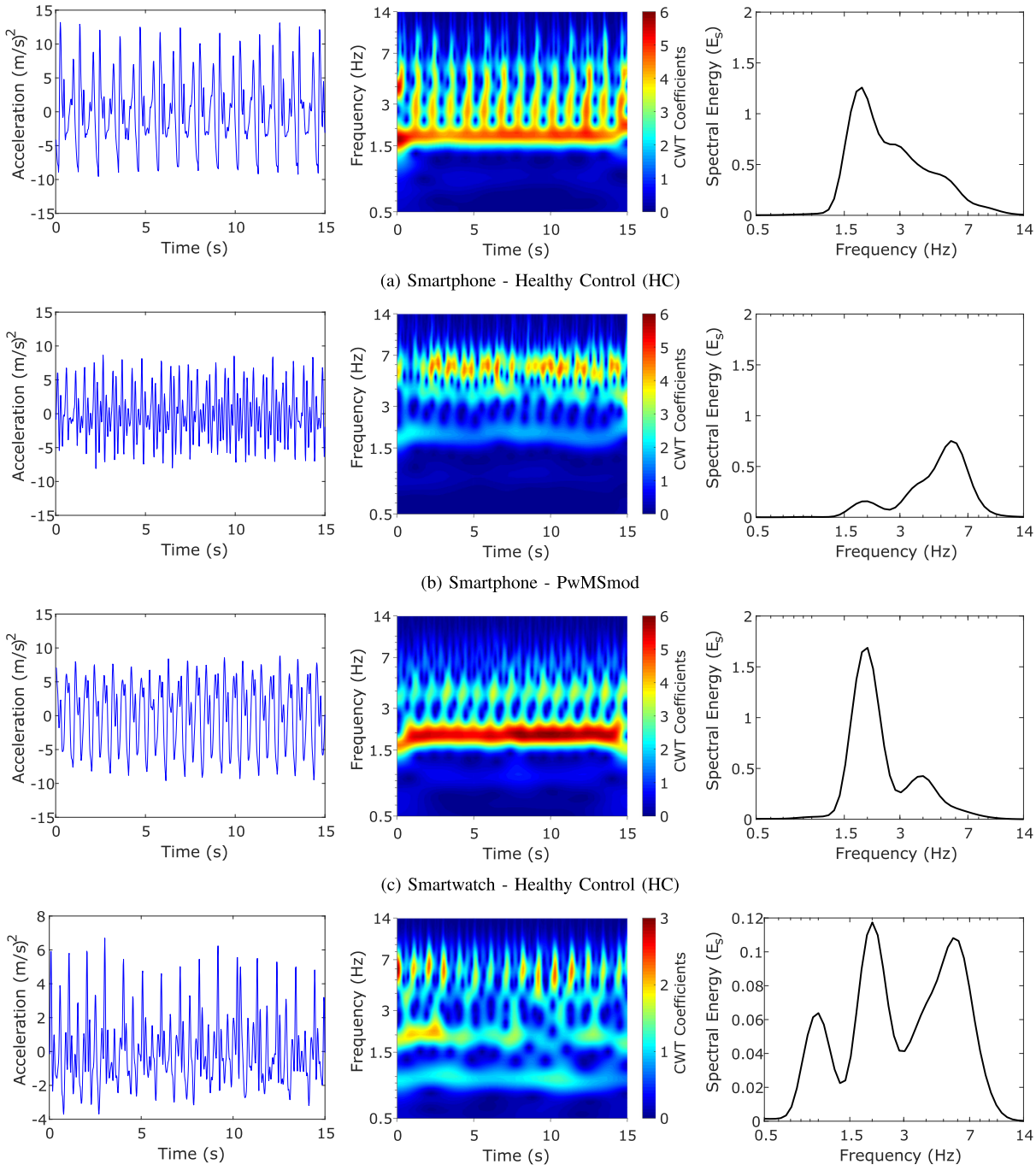
Instead of using the the raw *lassoglm* coefficients (β) for regression problems, it has been suggested that bias or prediction error can be reduced by performing a separate regression post-lasso [40]. Observations were assigned to the class yielding the largest posterior probability, where in the case of the SVM, posterior probabilities were first obtained using methods described by Platt [41]. SVM tuning was performed for each fold via grid-search over internal CV to determine optimal values of the Gaussian radial bias function (RBF) kernel parameter γ and the penalty parameter C. We selected the pair that gave the lowest CV misclassification error for each added feature to the classifier [39]. A selection of RF classifiers were built (using 1500 trees) and trained with a split criterion based on Gini impurity. Classifier performance was examined by varying the number of input variables chosen at each node (denoted as $mtry$). Values of $mtry$ used were tested as the square root of the number of features ($n = 13$); double and half this value was also investigated as suggested in [42]. Classification models were built using individual test observations and metrics based on majority voting of individual test predictions per subject are reported in order to increase prediction robustness. Classification performance metrics such as accuracy (acc), sensitivity (sen) and specificity (spec) are computed, where the more diseased class is the positive case. In order to account for the imbalance in the number of subjects within each sub-group, we also report the macro-average of the F_1 score for each class [43], [44]. Distribution differences in performance results calculated based on feature sets (smartphone, smartwatch, smartphone & smartwatch) and classification models (SVM, LR, RF) built across CV repetitions were tested using a Wilcoxon signed-rank test.

All data processing and analysis was performed using MATLAB vR2018a (The MathWorks, Natick, MA, USA). Fig. 1 schematically illustrates the entire gait processing, model construction and evaluation pipeline.

III. Results

A. Feature Analysis

Examples of raw sensor signals illustrate visual differences between HC and PwMSmod for both smartphone (Fig. 2(a) and 2(b)) and smartwatch devices (Fig. 2(c) and 2(d)). The CWT spectral energy density distribution (Fig. 3) demonstrated that PwMSmod had less power than PwMSmild and HC ($p < 0.01$). It was additionally observed that the ratio of total (AUC) spectral energy in the gait domain (0.5–3 Hz) to higher frequency "noise" energy per subject was lower in PwMSmod than in HC or PwMSmild smartphone tests (Fig. 3(a), $p < 0.001$). This study further focused on subjects



(a) Smartphone - Healthy Control (HC)

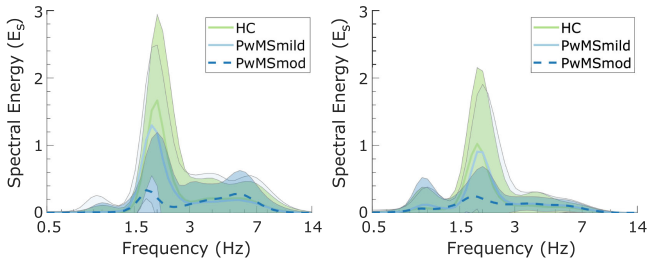
(b) Smartphone - PwMSmod

(c) Smartwatch - Healthy Control (HC)

(d) Smartwatch - PwMSmod

Fig. 2. Typical examples of accelerometer data recorded by smartphone device carried in a running belt for representative. (a) HC and (b) PwMSmod subjects, and their respective linked accelerometer data recorded by smartwatch device for the same, (c) HC, and (d) PwMSmod subject. The first column represents of raw magnitude acceleration signals $\|\mathbf{a}\|$. The second column shows the top view of the CWT scalogram, which is the absolute value of the CWT as a function of time and frequency. The third column corresponds to the scale-dependent (spectral) energy density (E_s) distribution of the CWT coefficients. (HC: T25FW, 3.6 $\pm$ 0.4 [s]); (PwMSmod: EDSS, 4 $\pm$ 0; T25FW, 8.1 $\pm$ 1.3 [s]); Note the axis scales for figure (d).

who exhibit clinically moderate disease symptoms, within the gait domain (PwMSmod, EDSS [3.5–5.5]). Table II depicts the top features between HC and PwMSmod as selected by $lassoglm$, the percentage of time chosen in the top 5 and 25 selected features, along with associated statistics and correlation to clinically administered metrics. A number of the top features selected derive from energy and entropy in the frequency bands for gait and movement [15]. Smartphone device features contributed most to the number of top features in the top 15 ranking (11 overall). The top-ranked features show high stability and consistency with good to excellent ICC values. Furthermore, these features significantly discriminated HC from



(a) Smartphone (b) Smartwatch

Fig. 3. The scale-dependent (spectral) energy density E_s distribution of the CWT coefficients for HC, PwMSmild and PwMSmod groups for (a) smartphone and (b) smartwatch devices. Bold lines and shaded region corresponds to the median and standard deviation in spectral energy amplitude per group.

PwMSmod using the median feature values per subject. Many top features demonstrated significant correlation ($p < 0.05$) to MS clinical measures in PwMSmod, especially T25FW [11] and EDSS [8].

The top features selected by $lassoglm$ were also compared to total step count [29]. Step count was significantly correlated with EDSS (R_s: 0.64, $p < 0.01$) and T25FW (R_s: 0.52, $p < 0.001$) for PwMSmod groups only. Total step count did not significantly distinguish HC from PwMSmild ($p = 0.09$) or PwMSmod ($p = 0.08$) groups. However, a significantly lower step count was observed in PwMSmod versus PwMSmild ($p < 0.01$).

B. Classification Analysis

To gather an understanding of added smartphone and smartwatch feature performance we computed the out-of-sample classification accuracy (HC vs. PwMSmod) as we varied the number of features added into an SVM and LR classifier.

While subjects were instructed to preferably carry the smartphones in the provided running belt, analysis has found that some participants wore the smartphone on either the running belt or in a pocket during the 24-week testing period (HC, $n = 15$; PwMSmild, $n = 41$; PwMSmod, $n = 14$). Smartphone orientations captured in landscape orientation were deduced to have come from the running belt in the anterior waist location (HC $n = 905$; PwMSmild, $n = 2424$; PwMSmod $n = 1296$), whereas portrait orientations were labelled as pocket locations (HC, $n = 457$; PwMSmild, $n = 1497$; PwMSmod, $n = 156$). No subject in this analysis contributed only pocket locations.

SVM classification accuracy using running belt tests rather than any location (either pocket or the running belt) (Fig. 4(a)) yielded improved accuracy; significantly so ($p < 0.05$) beyond 3 features thereafter (besides 7–8 features added p = 0.09 and p = 0.23 respectively). Classification accuracy plateaued after 15 features are added to the model. Many smartphone features indicated significantly different distributions between pocket and running belt locations ($n = 22, p < 0.05$). Given the smaller number of subjects and highly skewed number of pocket observations contributed per subject, where few subjects contributed the majority of pocket tests, we were unable to test the classification performance of pocket locations alone.

Classification performance was compared using features derived from either smartphone or smartwatch devices or with features derived from both devices (Fig. 4(b) and Table II). Maximum out-of-sample CV subject classification performance was reached using an SVM for smartphone devices with 23 features (Acc. 82.2 ± 2.9, Sen. 80.1 ± 3.9, Spec. 87.2 ± 4.2, F_1: 84.3 ± 3.8), compared with 19 features for smartwatch devices (Acc. 71.3 ± 3.6, Sen. 71.8 ± 6.8, Spec. 71.3 ± 3.7, F_1. 71.1 ± 3.5). Furthermore, smartphone devices showed significantly better accuracy ($p < 0.05$) with at least 8 features added to the classifier.

Additional classification models were also constructed to explore the separability between PwMSmild and PwMSmod groups, and HC and PwMSmild groups separately shown in Table III. It was observed that classification of PwMSmild and PwMSmod groups performed similarly: subject classification was maximised (Acc. 82.3 ± 1.9, Sen. 71.6 ± 4.2, Spec. 87.0 ± 3.2, F_1: 75.1 ± 2.2) using an SVM classifier with 21 features. Separation between HC and PwMSmild sub-groups however was less visible; where maximum classification accuracy was achieved (Acc. 66.4 ± 4.5, Sen. 67.5 ± 5.7, Spec. 60.3 ± 6.7, F_1: 58.6 ± 5.8) with 19 features modelled using an SVM.

Univariate and multivariate modelling of top-ranked signal-based complexity features achieved improved classification performance compared to using total step count alone as a feature for all classification outcomes based on EDSS grouping (Fig. 4(a)) and Table IV.

Classification was compared across different classifiers (LR, SVM and RF) as depicted in Table III. SVM models performed best at distinguishing HC from PwMSmild and PwMSmod, whereas the RF was marginally better at separating PwMSmild vs. PwMSmod. Accuracy was not significantly different between LR, SVM and RF models for all binary classification tasks expect at distinguishing HC from PwMSmod, with both LR and SVM performing significantly better than the RF ($p < 0.05$). Maximal subject classification was obtained for all classifiers using ≤ 26 features. It was observed that majority voting improved all classifier's performance.

IV. Discussion

The present study examined gait and physical function in PwMS and HC using remotely captured smartphone and smartwatch sensor data, while subjects performed a 2MWT. The aim of this analysis was twofold: (1) can meaningful features be derived from the remotely performed 2MWT that correlate with clinical assessments, and (2) can multivariate modelling of these performance metrics correctly distinguish groups of HC and PwMS with mild and moderate disability.

A. Feature Evaluation

The 2MWT assesses walking distance in PwMS [12], which can be indirectly approximated by the number of steps taken [27]. Studies have found that cadence (steps/minute) during the 6MWT and daily step count to be significantly different between MS subgroups [45], [46]. This simple biomechanical metric, as computed based on Lee *et al.* [29], showed less

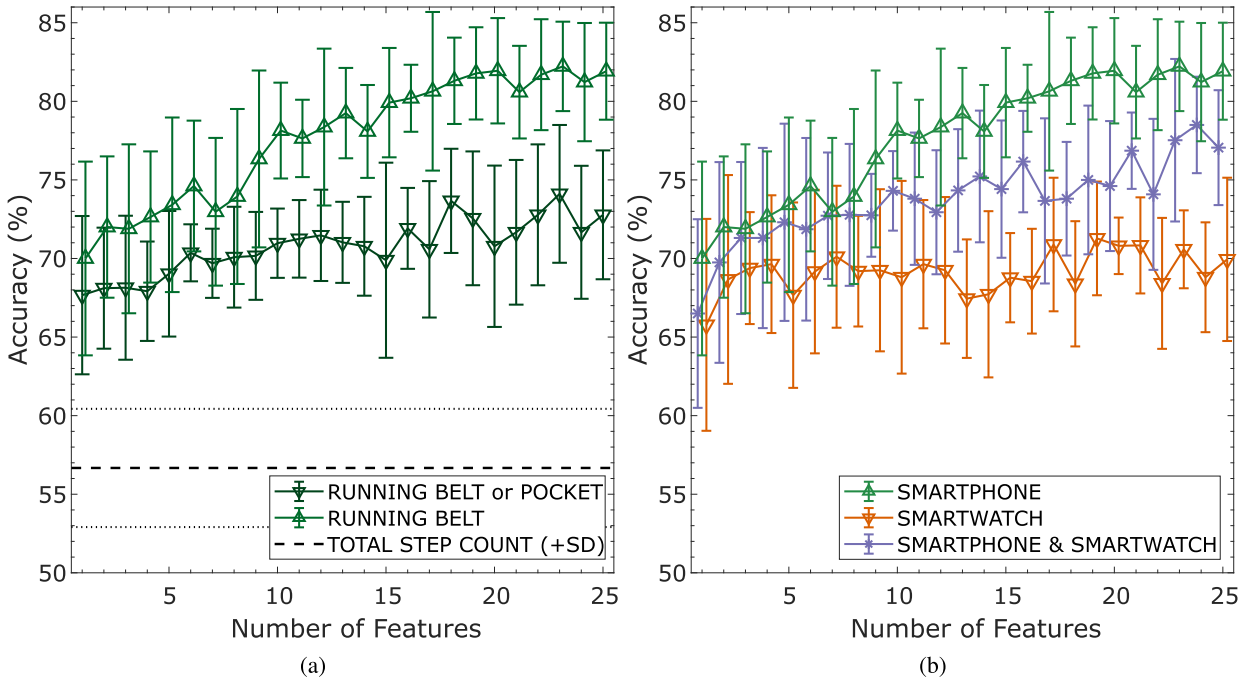


Fig. 4. Comparison of out-of-sample subject classification performance (HC vs. PwMSmod) as features are added to an SVM classifier using 5-fold cross validation with 10 repetitions. Figure (a) compares classification accuracy between smartphone tests performed using the running belt only versus tests using either the running belt or the pocket. Total step count represents the total step count over the whole 2MWT if used as a feature for classification. Figure (b) compares classification performance using features from smartphone, smartwatch and both devices. Confidence intervals denote one standard deviation (SD) around the quoted mean performance. For clarity, we present here only the first 25 steps.

discriminatory power between groups compared to signal-based complexity features. However, it must be considered that a recent comparative study calculating step counts in PwMS has also found considerable variability in the precision and accuracy of the algorithms and devices [19]. Ideally, step length would be used to calculate distance travelled during the 2MWT. Methods to derive step length however are erroneous or highly complex [47], [48], and their use in this study would also require validation in more controlled (non-remote) settings. Therefore, we opted to approximate the outcome of distance travelled for the remote 2MWT using a simple step count and its inclusion as a classifying feature is presented here primarily for comparative purposes. Classification performance, benchmarked against total step count, improved for all classification outcomes based on EDSS group stratification through the usage of these features (Tables III and IV). A number of the top features characterise the energy, frequency and variability of sensor signals recorded in the gait domain (Table II). In particular, percentage energy, multiscale and Shannon's entropy appeared prominently for both devices. Fig. 3 demonstrated that HC had more power in the gait domains (0.5–3 Hz). Also, the entropy computed in the gait band $H(cD_5(\|\mathbf{a}\|))$ was higher in PwMSmild and PwMSmod, i.e., the predictability of the gait signal in PwMS was lower than HC, with PwMS exhibiting lower energy and slower movements in their gait. Mapping clinical meaning to this could attribute PwMS, especially PwMSmod, to be less predictable with increased gait variability. Those with more severe MS could move more erratically as suggested in the overall top features selected from both devices that characterise the signal-to-noise ratio (SNR) and accelerometer skewness, which could capture noise and jerk-like movements.

B. Classification Evaluation

It was observed that PwMSmod were distinguishable from both HC and PwMSmild (Acc. 82.3 $\pm$ 2.9% and 82.3 $\pm$ 1.9% respectively), whereas PwMSmild were relatively less distinguishable from HC (Acc. 66.4 $\pm$ 4.5%). One of the primary means of PwMS disease assessment is the EDSS [8], hence it was used to stratify sub-groups of PwMS in this study. While those subjects with EDSS $\geq$ 3.5 are considered to have some gait related impairment, analysis of EDSS ambulation sub-scores demonstrated lower levels of ambulatory impact [49]. The heterogeneity of MS as a disease and its effect on symptom manifestation must also be considered when stratifying sub-groups based on clinical assessments and analysis thereafter, especially for those with lower (milder) severity scores [4], [8]. For example, examination of the mean PCA value per subject (Fig. 5(d)) using the first 2 components demonstrated that some PwMSmod can appear like HCs and vice-versa, where PwMSmild bisected regions between the two groups. This is an outcome to be expected from such a model considering that PwMSmild subjects experience very little gait abnormalities, and T25FW times were not significantly different between PwMSmild and HC ($p = 0.26$).

LASSO is a common ML technique with integrated feature selection and regression functionality (in this case LR) which is robust for reducing large numbers of features [39]. LR, in

TABLE II
COMPARISON OF STEP COUNT AND TOP FEATURES[1] BETWEEN HC AND PWMSMOD AS SELECTED BY $lassoglm$ ACROSS 5-FOLD CV WITH 10 REPETITIONS

	Device	Feature	Description	Top 5 (%)	Top 25 (%)	ICC	p	R_s EDSS	R_s T25FW
	Smartphone	# $Steps$	Total number of steps counted over entire 2MWT [29].			0.91(0.81-0.95)	0.08	0.64**	0.52***
1	Smartphone	$H(cD_5(\|\boldsymbol{a}\|))$	Entropy (2) of the 5th DWT coefficient detail quantifies the predictability of the gait signal roughly corresponding to frequency range 1.5-3.3 Hz (faster gait).	78	80	0.96(0.93-0.98)	***	0.52*	0.47**
2	Smartwatch	$skew(E_s(\|\boldsymbol{a}\|))$	Skewness as a measure of the asymmetry of the scale-dependent energy density distribution of the CWT coefficients E_s calculated from the the acceleration signal. This measures the relative magnitude of how far the distribution deviates from the 'normal', which was used as a proxy for smooth stable gait movements.	72	74	0.81(0.61-0.91)	***	0.56**	0.40**
3	Smartwatch	$\frac{max_1(E_s(\|\boldsymbol{a}\|))}{max_2(E_s(\|\boldsymbol{a}\|))}$	The ratio of the maximum scale-dependent energy peak density E_s to the next highest peak, corresponding to the frequencies computed using a CWT, over the 2MWT.	38	70	0.81(0.61-0.91)	**	0.45*	0.31*
4	Smartphone	$std(MsEn(\|\boldsymbol{a}\|)$	Std. deviation in multiscale entropy over all scales. This quantifies the variation in the predictability of a gait signal over multiple temporal scales, where $MsEn$ characterises dynamic complexity of gait within a signal [35].	30	100	0.80(0.59-0.90)	***	0.05	0.19
5	Smartphone	$E(cD_5(\|\boldsymbol{a}\|))$	The energy (1) contained in the 5th DWT coefficient detail roughly corresponding to frequency range 1.5-3.3 Hz. This could be a proxy for (faster) gait power.	24	86	0.94(0.88-0.97)	***	0.53*	0.49**
6	Smartphone	$RPDE(\|\boldsymbol{a}\|)$	Recurrence period density entropy characterises the periodicity of a gait signal and the ability to maintain consistent gait rhythm [36].	28	54	0.87(0.73-0.93)	**	0.21	0.32*
7	Smartwatch	$std(SNR(\|\boldsymbol{a}\|))$	Variability in the SNR between epochs, characterised by IMFs and computed using EMD, with the signal sampled every 0.5 [s]. This is analogous for the variability in the ratio (amount) of gait to higher- frequency perturbations over the 2MWT.	16	82	0.74(0.45-0.87)	*	0.22	0.30*
8	Smartphone	$std(SNR(\|\boldsymbol{a}\|))$	(see above)	16	84	0.81(0.60-0.90)	***	0.20	0.40**
9	Smartphone	$skew(\|\boldsymbol{a}\|)$	Skewness as a measure of the asymmetry of the probability distribution of the acceleration signal values. The relative magnitude of how far a distribution deviates from the normal which used as a proxy for smooth stable ambulatory movements.	4	92	0.40(0.20-0.71)	n.s.	0.30	0.06
10	Smartphone	$std(zcr(\boldsymbol{a}_z))$	The std. deviation in the zero-crossing rate calculates the rate of sign-changes along the medial-lateral plane over the 2MWT, which can measure the dynamic sway during ambulation.	10	76	0.68(0.32-0.84)	*	0.53*	0.47**
11	Smartphone	$zcr(\|\boldsymbol{a}\|)$	The zero-crossing rate calculates the rate of sign-changes in a signal, roughly capturing the static-to-dynamic transitions within gait.	24	68	0.82(0.65-0.92)	*	0.15	0.33*
12	Smartphone	$std(\boldsymbol{a}_z)$	The standard deviation in the acceleration values in the medial-lateral plane for the 2MWT, which can measure the dynamic sway during gait.	4	95	0.86(0.65-0.94)	**	0.15	0.10
13	Smartphone	$SNR(\|\boldsymbol{a}\|)$	The mean SNR, where the signal and noise are characterised by IMFs and computed using EMD, with the signal sampled every 0.5 [s]. This is analogous for the ratio (amount) of gait to higher- frequency perturbations over the 2MWT.	10	76	0.73(0.46-0.87)	*	0.10	0.32*
14	Smartwatch	$\frac{MsEn(\|\boldsymbol{a}\|)_{(:10)}}{MsEn(\|\boldsymbol{a}\|)_{(11:)}}$	The ratio in multi-scale entropy over first 10 to the last 10 scales captures the dynamic complexity of gait versus that of random fluctuations within the gait signal [35].	10	64	0.76(0.53-0.87)	**	0.38	0.50**
15	Smartphone	$kurt(\|\boldsymbol{a}\|)$	Kurtosis as a measure of how outlier prone the distribution of the magnitude of acceleration signal values are to quantify gait-related perturbations.	0	76	0.81(0.61-0.90)	*	0.55**	0.61***

[1]See supplementary material for a full description of all of the features extracted in this study.

ICC, Intraclass correlation coefficient (95% CI); Other statistics calculated on median feature value per subject (HC, $n = 24$; PwMSmod $n = 21$): P, Mann Whitney U Test between groups; R_s EDSS, Spearman's correlation to EDSS in PwMSmod; R_s T25FW, Spearman's correlation to Timed-25 ft. walk test; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; n.s., not significant.

essence, describes a linear relationship between the predictors with a non-linear mapping to response variables. LR models performed similarly to the SVM and RF classifier, suggesting perhaps a simple linear relationship exists between the combination of gait features. However there may be other non-linear feature selection techniques (such as Relieff or mRMR), which may select more optimal features in a more optimal ranking [39]. RF models built in this study, which internally use non-linear feature selection, did however perform significantly worse than LR and SVM models for HC vs. PwMSmod classification tasks ($p < 0.05$). A limitation of this work is that each subject's tests were considered independent and identically distributed

TABLE III
SMARTPHONE-BASED SUBJECT CLASSIFICATION OUTCOMES BY EDSS GROUPING FOR VARIOUS CLASSIFIERS

Classifier	Acc.	Sen.	Spec.	F_1
HC vs. PwMSmild				
LR	64.5 ± 5.8	68.3 ± 7.1	56.5 ± 5.7	58.3 ± 5.5
SVM	**66.4 ± 4.5**	**67.5 ± 5.7**	**60.3 ± 6.7**	**58.6 ± 5.8**
RF	63.2 ± 4.0	73.3 ± 4.1	41.4 ± 8.8	50.4 ± 5.6
PwMSmild vs. PwMSmod				
LR	83.7 ± 2.4	71.7 ± 6.4	88.9 ± 2.3	77.1 ± 2.7
SVM	82.3 ± 1.9	71.6 ± 4.2	87.0 ± 3.2	75.1 ± 2.2
RF	**84.0 ± 1.9**	**75.7 ± 1.5**	**87.8 ± 2.2**	**78.1 ± 2.3**
HC vs. PwMSmod				
LR	80.4 ± 5.3	76.7 ± 7.8	84.2 ± 5.6	80.0 ± 5.4
SVM	**82.2 ± 2.9**	**80.1 ± 3.9**	**87.2 ± 4.2**	**84.3 ± 3.8**
RF	76.2 ± 3.5	70.9 ± 5.9	81.1 ± 3.9	76.5 ± 3.9

Mean and standard deviation across CV repetitions (%); LR - Logistic Regression; SVM - Support Vector Machines with a RBF; RF - Random Forest.

TABLE IV
SMARTPHONE-BASED STEP COUNT CLASSIFICATION OUTCOMES BY EDSS GROUPING USING A SVM WITH A RBF

Classifier (Step Count†)	Acc.	F_1
HC vs. PwMSmild	59.1 ± 3.2	54.9 ± 3.4
PwMSmild vs. PwMSmod	68.6 ± 3.7	59.8 ± 3.4
HC vs. PwMSmod	56.7 ± 3.8	60.0 ± 5.1

Mean and standard deviation across CV repetitions (%). †Total step count over the whole 2MWT used as a single feature for classification.

(i.i.d), where subjects each contributed a varying number of tests. In reality, test observations per subject will be dependent and may be better suited to sequence modelling approaches. For example, there may be alternatives to RF in this application such as Mixed Effect Trees [50], which consider repeated measures (tests) for classification and could help overcome potential model biases related to the varying number of observations per subject.

Comparing the classification accuracy of HC vs. PwMSmod using smartphone and smartwatch devices showed comparative prediction accuracy for a smaller number of features added (<8) to our models (Fig. 4(b)). However, beyond this smartphone features demonstrated significantly improved performance ($p < 0.05$) over smartwatch features. Surprisingly, drawing from both the smartphone and smartwatch feature space did not lead to improved classification performance and maximal accuracy was achieved using smartphone features only (smartphone: Acc. 82.2 ± 2.9%; smartwatch: Acc. 71.3 ± 3.6%). Further investigations revealed a high variability in the type of feature (calculated from smartphone vs. smartwatch) selected at each fold. Interrogating the feature distributions within CV folds highlighted poor feature generalisability between training and testing sets in some cases. Table II indicated for example the top feature $H(cD_5(\|\mathbf{a}\|))$ was picked 80% of the time, but when the feature was picked, it was nearly always in the top 5 features per fold — in other folds the distributions were changed so dramatically they would not be picked. Combining features from both devices amplified these problems and did not lead to improved results. Finally, the added smartphone classification accuracy beyond 10–15 features was minimal (Fig. 4(b)), demonstrating that accurate classification can be achieved with a small number of features.

C. Considerations for the Remote Characterisation of Ambulation in HC & PwMS Using Smartphones and Smartwatches

Although remote monitoring has many advantages such as unobtrusive, high-frequency assessment of disease, a number of confounding factors must be considered when taking measurements in real-world non-laboratory scenarios. Some examples encountered included the differences in subjects' adherence during the study, some subjects not following the prescribed protocols (instances of smartphones in pocket locations or the use of only one device during testing), along with the many degrees of freedom associated with self-generated patient data from real-world testing. While the instructions given (see [22]) were standardised, analogous to that of an in-clinic performed 2MWT [12], [13], the 2MWT in this study was a remotely executed out-of-clinic assessment. The performance of the 2MWT can be highly influenced by the testing environment such as the length of the hallways, the number and frequency of subjects' turns, or other factors which we cannot determine remotely.

The number of unique HC ($n = 24$), PwMSmild ($n = 52$) and PwMSmod ($n = 21$) in this study was relatively few. Sampling sufficiently sized data from a more diverse cohort should also be considered in order to build robust and generalizable models. For example, biases may exist related to the mismatch in the male to female ratio between HC and PwMS groups (Table I). It was also acknowledged that there was a high standard deviation in the number of tests contributed by each subject, however no subject group contributed significantly more tests than another (Table I). Besides individual subjects' adherence rates, this variability was also partly due to the exclusion criteria imposed on the data used in this analysis, where only linked smartphone and smartwatch tests, and of those, only smartphone tests performed with the running belt were considered for fair comparisons between the devices. As such, performing CV on subject-wise splits with each subject contributing varying numbers of tests can cause the distributions of features to vary between CV partitions. This helped attribute to the high classification variance between folds and variability in the type and number of features needed for maximal classification. Furthermore, subject heterogeneity can also highly influence model robustness when data is sparse. For example, Fig. 5 illustrated the first two principal components computed from PCA plotted against each other for all HC and PwMS test observations. The top 25 smartphone features were used to perform the PCA and hence represents the overall structure of the features which are mostly selected for classification. It was observed that intra-subject variability appeared low, while inter-subject variability was distinctly high as subjects clustered within themselves. Subject-wise CV is hence confounded not only by low subject numbers and sparse, heterogeneous data, but in this case was also heavily exacerbated by subject's gait feature patterns, which were uniquely associated to the individual. This manifested as problems exhibited in feature selection and generalisability across CV partitions. There has been much

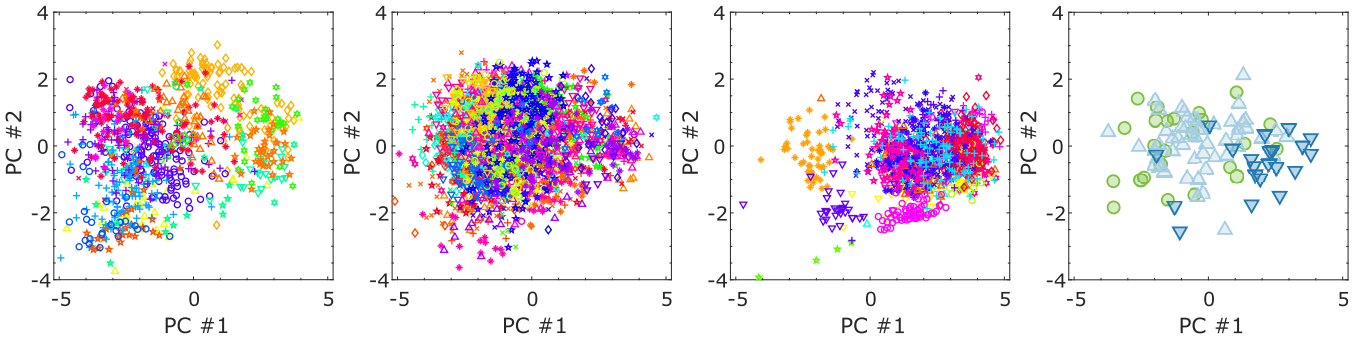


Fig. 5. Example of PCA performed on top 25 smartphone features. The first two components have been plotted per subject and their respective test observations for HC (a) PwMSmild, (b) PwMSmod, and (c) subjects. Figure (d) represents the mean component value per subject coded by subject group, HC (green circles), PwMSmild (light blue triangles) and PwMSmod (dark blue inverted triangles). Intra-subject variability appears low, while inter-subject variability is distinctly high as subjects cluster within themselves.

discussion within the academic community as to the advantages and disadvantages of subject-wise versus observation-wise CV approaches [51]–[53]. Subject-wise CV has been criticised in that it may break assumptions to consistently estimate generalisation error in heterogeneous data and lead to model under-fitting and larger classification error [52]. However, in subject-wise CV it is argued that there is no "leakage" of subject information between training and testing partitions where models can learn individual subjects' characteristics [51]. Further recent analysis by Neto *et al.* [53], using and real-world mobile data, demonstrated how various examples of heterogeneity across subjects can lead to the identification of subjects' characteristics rather than disease in observation-wise CV approaches. As such, this study adopted the conservative subject-wise CV method to eliminate identity-confounding factors, but acknowledges that sub-optimal classification and model under-fitting can occur given the low number of subjects contributing multiple observations in this study.

Finally, a high inter-dependence was observed within the feature space. A number of the top features within a source (device) and features between sources were highly correlated with each other. This indicates that some features may represent the same information. Fig. 6 in Appendix B shows the inter-source and intra-source feature correlation. This could attribute to predictor redundancy and explain the marginal classification performance beyond 10 features added for all sources (Fig. 4(b)). However, smartwatch features reached a plateau in added information before smartphone features, suggesting that this device may contain less information and hence more feature redundancy. Every top smartwatch feature (HC vs. PwMSmod) characterises the relative power in gait to non-gait frequency domains derived from CWT and EMD. It should be considered that the location of smartwatch (wrist) sensor to smartphone (running belt) may have a profound effect on the depth of information that can be recorded about gait function. This may be manifested by a larger and more varied number of useful features (as discussed, the top smartwatch features chosen only characterise the relative power in gait to non-gait frequency domains).

Despite applying an orientation transformation during pre-processing, it was found that some smartphone-based features were also location dependent (running belt versus pocket) and classification based on features from only one pose increased the accuracy of our models (Fig. 4(a)). As such, smartphone-based tests performed with the running belt were considered different to the smartphone in the pocket.

D. Future Work

This study demonstrated the feasibility of characterising gait function in PwMS remotely using body-worn inertial sensors embedded in consumer smartphones and smartwatches.

It was observed that classification performance was affected by inconsistent placement location (i.e., some 2MWTs were performed using the running belt versus others where the smartphone was placed within a pocket). The results from this study therefore emphasise the importance of a standardised approach to remote sensor monitoring and advocates the use of a consistent sensor location such as a running belt for future studies. Advantages of running belts are that they offer a fixed and standard placement location to capture gait characteristics, and avoid the need for participants to have pockets or another means to carry their device.

Adherence to the prescribed protocol was an issue observed in this study however. It is acknowledged that instructing participants to regularly carry their smartphone and smartwatch for a daily 2MWT, and to affix their smartphone using a running belt is both obtrusive and inconvenient. As such, perhaps the use of a smartphone within a pocket, or even a single smartwatch during passively collected free-living gait may be a better option for the unobtrusive, long-term monitoring of PwMS subjects. Future work is needed first however to explore in greater detail: (1) the effect of placement location of smartphone devices, and (2) to investigate the differences in information captured by smartphone and smartwatch devices for quantifying gait dysfunction in PwMS, particularly in more controlled settings. These further studies should also aim to compare the outcome

measures investigated in this work to clinically administered 2MWTs and in-clinic gait measurement systems. This suggested analysis would allow the further evaluation, understanding and improvement of the most optimal protocols designed to explore how sensor data can represent PwMS impairment remotely and out-of-clinic.

As MS symptoms fluctuate periodically, the real value of remote monitoring PwMS may ultimately lie in investigating test performance as a function of time. To sufficiently capture the time-varying nature of MS ambulatory impairments it will require both robust outcome measures and the design and standardisation of feasible and unobtrusive protocols for objective assessments — that can be delivered remotely and administered frequently — such as those introduced in this work.

V. Conclusion

This study demonstrates the benefits of ML and multivariate feature modelling in the identification of the signs of ambulatory function impairment in PwMS from remotely captured smartphone and smartwatch inertial sensor data. A combination of statistical- and signal-based features calculated from both devices performed better than simple biomechanical metrics such as step count, which was used to approximate the standard 2MWT outcome of walking distance. Many previous studies probing the characteristics and separability of PwMS and HC have used multiple standalone inertial sensors affixed to the body at various locations during controlled in-clinic assessments [6], [7], [17], [18]. In this study it can be seen that sufficient information for accurate MS symptom characterisation may be captured in relatively few features ($\leq$26) obtained from an out-of-clinic using only one device. It was found that PwMSmod, who experience gait-related dysfunction, could be distinguished with a high accuracy from PwMSmild and HC, whom the latter two groups were more difficult to differentiate from each other. The work presented here, with on-going future work, helps establish a methodological foundation to construct models that can identify patterns of PwMS ambulatory impairment from remote gait assessments. MS is a heterogeneous, mutable disease and subjects may experience symptoms in various domains which hard thresholds on infrequently administered clinical scales may fail to capture. Key advantages of objective assessments like those in this study are that they can be administered at high-frequency and longitudinally in out-of-clinic environments.

Appendix A
Calculation of ICC

The intra-class correlation coefficient (ICC) is a widely used metric to quantify the test-retest reliability of test observations in the biomedical field [54]. The reliability of the feature values can be inferred if we consider each feature value over repeated test observations per subject. ICC (A, k) was calculated for the the 14-day session median across subjects. To be included in the analysis, subjects needed to have a minimum of 3 measurements per window. Reliability was categorised as either poor (ICC $<$ 0.5), moderate (ICC $=$ 0.5–0.75), good (ICC $=$ 0.75–0.9) or excellent (ICC $>$ 0.9).

Appendix B

Inter-Source and Intra-Source Correlation

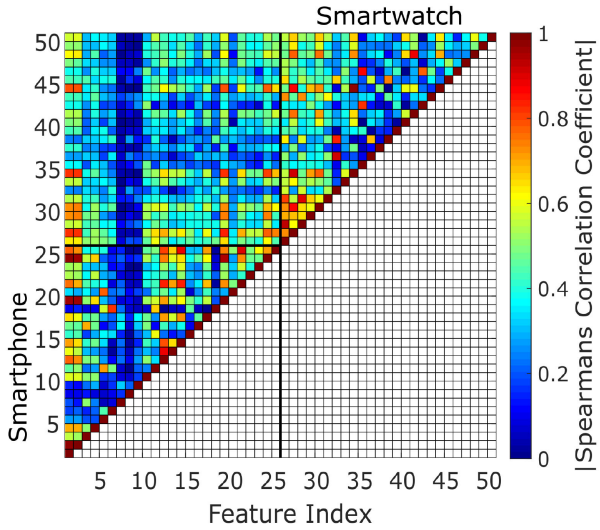


Fig. 6. Pairwise correlation matrix showing the intra- and inter-source correlation for the top 25 ranked smartphone (index 1–25) and smartwatch (index 26–50) features respectively. The top 10 smartwatch features are highly correlated with each other, whereas the top 10 smartphone features exhibit much less inter-correlation. The inter-source correlation is strong in the top 5 features between smartphone and smartwatch devices.

Acknowledgment

The author would like to thank all staff and participants involved in capturing test data. This study was sponsored by F. Hoffmann-La Roche Ltd. This research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). A. P. Creagh is a Ph.D. student at the University of Oxford and acknowledges the support of F. Hoffmann-La Roche Ltd.; C. Simillion and F. Lipsmeier are employees of F. Hoffmann-La Roche Ltd.; A. K. Bourke and C. Bernasconi are contractors for F. Hoffmann-La Roche Ltd.; A. Scotland and M. Lindemann are consultants for F. Hoffmann-La Roche Ltd. via Inovigate; M. Baker, J. van Beek and C. Gossens are employees and shareholders of F. Hoffmann-La Roche Ltd.; M. De Vos has nothing to disclose.

References

[1] M. M. Goldenberg, "Multiple sclerosis review," *Pharm. Ther.*, vol. 37, no. 3, pp. 175–184, 2012.
[2] C. Heesen, J. Böhm, C. Reich, J. Kasper, M. Goebel, and S. Gold, "Patient perception of bodily functions in multiple sclerosis: Gait and visual function are the most valuable," *Mult. Scler. J.*, vol. 14, no. 7, pp. 988–991, 2008.
[3] C. L. Martin *et al.*, "Gait and balance impairment in early multiple sclerosis in the absence of clinical disability," *Mult. Scler. J.*, vol. 12, no. 5, pp. 620–628, 2006.
[4] J. J. Sosnoff, B. M. Sandroff, and R. W. Motl, "Quantifying gait abnormalities in persons with multiple sclerosis with minimal disability," *Gait Posture*, vol. 36, no. 1, pp. 154–156, 2012.
[5] S. Crenshaw, T. Royer, J. Richards, and D. Hudson, "Gait variability in people with multiple sclerosis," *Mult. Scler. J.*, vol. 12, no. 5, pp. 613–619, 2006.
[6] J. M. Huisinga, M. Mancini, R. J. S. George, and F. B. Horak, "Accelerometry reveals differences in gait variability between patients with multiple sclerosis and healthy controls," *Ann. Biomed. Eng.*, vol. 41, no. 8, pp. 1670–1679, 2013.

[7] R. Spain *et al.*, "Body-worn motion sensors detect balance and gait deficits in people with multiple sclerosis who have normal walking speed," *Gait Posture*, vol. 35, no. 4, pp. 573–578, 2012.

[8] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[9] V. Khurana, H. Sharma, N. Afroz, A. Callan, and J. Medin, "Patient-reported outcomes in multiple sclerosis: A systematic comparison of available measures," *Eur. J. Neurol.*, vol. 24, no. 9, pp. 1099–1107, 2017.

[10] R. Rudick, G. Cutter, and S. Reingold, "The multiple sclerosis functional composite: A new clinical outcome measure for multiple sclerosis trials," *Mult. Scler. J.*, vol. 8, no. 5, pp. 359–365, 2002.

[11] R. W. Motl *et al.*, "Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis," *Mult. Scler.*, vol. 23, no. 5, pp. 704–710, 2017.

[12] R. W. Bohannon, Y. C. Wang, and R. C. Gershon, "Two-minute walk test performance by adults 18 to 85 years: Normative values, reliability, and responsiveness," *Arch. Phys. Med. Rehabil.*, vol. 96, no. 3, pp. 472–477, 2015.

[13] D. A. Scalzitti, K. J. Harwood, J. R. Maring, S. J. Leach, E. A. Ruckert, and E. Costello, "Validation of the 2-minute walk test with the 6-minute walk test and other functional measures in persons with multiple sclerosis," *Int. J. MS Care*, vol. 20, no. 4, pp. 158–163, 2018.

[14] E. Willoughby and D. Paty, "Scales for rating impairment in multiple sclerosis a critique," *Neurology*, vol. 38, no. 11, pp. 1793–1793, 1988.

[15] D. Jarchi, J. Pope, T. K. Lee, L. Tamjidi, A. Mirzaei, and S. Sanei, "A review on accelerometry based gait analysis and emerging clinical applications," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 177–194, 2018.

[16] A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin, "Direct measurement of human movement by accelerometry," *Med. Eng. Phys.*, vol. 30, no. 10, pp. 1364–1386, 2008.

[17] B. R. Greene *et al.*, "Assessment and classification of early-stage multiple sclerosis with inertial sensors: Comparison against clinical measures of disease state," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1356–1361, Jul. 2015.

[18] R. I. Spain, M. Mancini, F. B. Horak, and D. Bourdette, "Body-worn sensors capture variability, but not decline, of gait and balance measures in multiple sclerosis over 18 months," *Gait Posture*, vol. 39, no. 3, pp. 958–964, 2014.

[19] J. M. Balto, D. L. Kinnett-Hopkins, and R. W. Motl, "Accuracy and precision of smartphone applications and commercially available motion sensors in multiple sclerosis," *Mult. Scler. J.–Exp. Transl. Clin.*, vol. 2, 2016.

[20] R. Bove *et al.*, "Evaluating more naturalistic outcome measures," *Neurol. Neuroimmunol. Neuroinflamm.*, vol. 2, no. 6, p. e162, 2015.

[21] G. Comi *et al.*, "Effect of early interferon treatment on conversion to definite multiple sclerosis: A randomised study," *Lancet*, vol. 357, no. 9268, pp. 1576–1582, 2001.

[22] L. Midaglia *et al.*, "Adherence and satisfaction of smartphone-and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study," *J. Med. Internet Res.*, vol. 21, no. 8, 2019, Art. no. e14863.

[23] A. Creagh *et al.*, "Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test," *Physiol. Meas.*, vol. 41, no. 5, pp. 054002, Jun. 2020, doi: 10.1088/1361-6579/ab8771.

[24] S. Meyer-Moock, Y.-S. Feng, M. Maeurer, F.-W. Dippel, and T. Kohlmann, "Systematic literature review and validity evaluation of the expanded disability status scale (EDSS) and the multiple sclerosis functional composite (MSFC) in patients with multiple sclerosis," *BMC Neurol.*, vol. 14, p. 58, 2014, doi: 10.1186/1471-2377-14-58.

[25] H. Aodhán, D. Silvia Del, R. Lynn, and G. Alan, "Detecting free-living steps and walking bouts: Validating an algorithm for macro gait analysis," *Physiol. Meas.*, vol. 38, no. 1, pp. N1–N15, 2017.

[26] M. Gadaleta and M. Rossi, "IDNet: Smartphone-based gait recognition with convolutional neural networks," *Pattern Recognit.*, vol. 74, no. Supplement C, pp. 25–37, 2018.

[27] J. Bassett *et al.*, "Accuracy of five electronic pedometers for measuring distance walked," *Med. Sci. Sports Exerc.*, vol. 28, no. 8, pp. 1071–1077, 1996.

[28] N. A. Capela, E. D. Lemaire, and N. Baddour, "Novel algorithm for a smartphone-based 6-minute walk test application: Algorithm, application development, and evaluation," *J. Neuroeng. Rehabil.*, vol. 12, no. 1, p. 19, 2015.

[29] H.-h. Lee, S. Choi, and M.-J. Lee, "Step detection against the dynamics of smartphones," *Sensors*, vol. 15, no. 10, pp. 27 230–27 250, 2015.

[30] P. S. Addison, J. Walker, and R. C. Guido, "Time–frequency analysis of biosignals," *IEEE Eng. Med. Biol. Mag.*, vol. 28, no. 5, pp. 14–29, Sep./Oct. 2009.

[31] S. Khandelwal and N. Wickström, "Novel methodology for estimating initial contact events from accelerometers positioned at different body locations," *Gait Posture*, vol. 59, pp. 278–285, 2018.

[32] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York, NY, USA: Academic, 2008.

[33] P. Ren *et al.*, "Gait rhythm fluctuation analysis for neurodegenerative diseases by empirical mode decomposition," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 52–60, Jan. 2017.

[34] N. E. Huang *et al.*, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A: Math., Phys. Eng. Sci.*, vol. 454, pp. 903–995, 1998.

[35] M. Costa, C. K. Peng, A. L. Goldberger, and J. M. Hausdorff, "Multiscale entropy analysis of human gait dynamics," *Phys. A: Stat. Mech. Appl.*, vol. 330, no. 1, pp. 53–60, 2003.

[36] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. Online*, vol. 6, no. 1, p. 23, 2007.

[37] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc.: Ser. B (Methodol.)*, vol. 57, no. 1, pp. 289–300, 1995.

[38] I. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2011, pp. 1094–1096.

[39] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer Series in Statistics, vol. 1, New York, NY, USA: Springer, 2001.

[40] A. Belloni and V. Chernozhukov, "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, vol. 19, no. 2, pp. 521–547, 2013.

[41] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[43] H. G. He and A. Edwardo, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[44] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[45] R. W. Motl *et al.*, "Evidence for the different physiological significance of the 6-and 2-minute walk tests in multiple sclerosis," *BMC Neurol.*, vol. 12, p. 6, 2012, doi: 10.1186/1471-2377-12-6.

[46] A. Neven, A. Vanderstraeten, D. Janssens, G. Wets, and P. Feys, "Understanding walking activity in multiple sclerosis: Step count, walking intensity and uninterrupted walking activity duration related to degree of disability," *Neurol. Sci.*, vol. 37, no. 9, pp. 1483–1490, 2016.

[47] R. C. González, D. Alvarez, A. M. López, and J. C. Alvarez, "Ambulatory estimation of mean step length during unconstrained walking by means of cog accelerometry," *Comput. Methods Biomech. Biomed. Eng.*, vol. 12, no. 6, pp. 721–726, 2009.

[48] Q. Wang, L. Ye, H. Luo, A. Men, F. Zhao, and C. Ou, "Pedestrian walking distance estimation based on smartphone mode recognition," *Remote Sens.*, vol. 11, no. 9, p. 1140, 2019, doi: 10.3390/rs11091140.

[49] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[50] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed effects regression trees for clustered data," *Stat. Probab. Lett.*, vol. 81, no. 4, pp. 451–459, 2011.

[51] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, pp. 1–9, 2017.

[52] M. A. Little *et al.*, "Using and understanding cross-validation strategies. Perspectives on Saeb *et al.*" *GigaScience*, vol. 6, no. 5, pp. 1–6, 2017.

[53] E. C. Neto *et al.*, "Detecting the impact of subject characteristics on machine learning-based diagnostic applications," *NPJ Digit. Med.*, vol. 2, no. 1, p. 99, 2019, doi: 10.1038/s41746-019-0178-x.

[54] J. P. Weir, "Quantifying test-retest reliability using the intraclass correlation coefficient and the sem," *J. Strength Cond. Res.*, vol. 19, no. 1, pp. 231–240, 2005.