

Minimum data **sampling** requirements for accurate detection of terrain-induced gait alterations change with mobile sensor position

Arshad Sher^{1,2}, OtarAkanyeti²

¹ *Department of Computer Science, Nottingham Trent University, Nottinghamshire NG11 8NS, UK,* ² *Department of Computer Science, Aberystwyth University, Ceredigion SY23 3DB, UK*

Abstract

Human gait is a key biomarker for health, independence and quality of life. Advances in wearable inertial sensor technologies have paved the way for out-of-the-lab human gait analysis, which is important for the assessment of mobility and balance in natural environments and has applications in multiple fields from healthcare to urban planning. Automatic recognition of the environment **where** walking takes place is a prerequisite for successful characterisation of terrain-induced gait alterations. A key question which remains unexplored in the field is how minimum data requirements for high terrain classification accuracy change depending on the sensor placement on the body. To address this question, we evaluate the changes in performance of five canonical machine learning classifiers by varying several data sampling parameters including sampling rate, segment length, and sensor configuration. Our analysis on two independent datasets clearly demonstrate that **a single inertial measurement unit is sufficient to recognise terrain-induced gait alterations, accuracy and minimum data requirements vary with the device position on the body**, and choosing correct data sampling parameters for each **position** can improve classification accuracy up to 40% or reduce data size by 16 times. Our findings highlight the need for adaptive data collection and processing algorithms for resource-efficient computing on mobile devices.

© 2024 Published by Elsevier Ltd.

Keywords: gait, mobile app, smartphone, wearable, accelerometer, gyroscope, IMU, sampling rate, outdoor walking, terrain recognition, machine learning, and digital health

1. Introduction

Automatic real-world human gait analysis using mobile devices is becoming more pervasive and has applications in a variety of fields including medicine, sports science, fashion and urban planning [1, 2, 3, 4, 5]. Few examples are fall detection [6], biometric authentication [7], human activity recognition [8], gait and balance assessment [9], injury prevention [10] and assessing brain health [11].

One of the most active research areas in this domain is developing adaptive data collection and analysis methods to improve gait assessment accuracy and efficiency (e.g., [12]). These methods need to take into account physiological (e.g., stress and fatigue), biomechanical (e.g., using a walking aid), contextual (e.g., fast walking to catch a bus), environmental (e.g., terrain, weather conditions and traffic) and behavioural (e.g., shopping and exercising) factors that affect gait. They also need to consider constraints related to device instrumentation (e.g., position, data resolution, noise, battery, CPU and memory).

To take a step forward in this direction, this study focuses on recognising in what environment walking takes place (i.e. walking terrain recognition) using a single inertial measurement unit (IMU). While there are many studies focusing on automatic human activity recognition, with signal processing and machine learning models differentiating walking from other activities of daily living such as sitting and standing (e.g., [13, 14]), up until now only a handful of studies have focused on walking terrain recognition [15, 16, 17, 18] (these studies are discussed in the next section).

Automatic walking terrain recognition is desired for a number of reasons. First, it will enable to improve the performance of the gait analysis algorithms through terrain-specific parameter tuning. For instance, Sher *et al.* have recently shown that in young and healthy adults step time estimation error can be lowered by up to 40% by tuning low-pass filter cut-off frequency during pre-processing the data [16]. Second, it will improve our understanding of terrain-induced gait alterations which is key while designing new interventions and evaluate their impact (e.g., balance and strength training for elderly, detecting and fixing potholes on the streets, and while designing new shoes or assistive walking devices). Third, it can be used for assessing functional independence and activity monitoring in elderly and people with chronic movement disorders (e.g., Stroke and Parkinson's patients). Here, the assumption is that individuals who are able to walk on different terrains are more independent than those who can only walk on flat surfaces or who spend most of their time in sedentary behaviour.

The current IMUs often come with tri-axial accelerometer and gyroscope sensors providing local acceleration and angular velocity measurements in three orthogonal directions. Previous studies have shown that terrain related gait changes are profound enough to create distinct acceleration and velocity patterns [19, 20, 21], although these patterns may differ from individual to individual [22, 23]. Furthermore, the quality of the IMU data changes depending on the placement of the wearable device on the body [24, 25], which should be factored in while sampling the data [26, 27]. Hence, we ask the following two questions: 1) How does IMU position affect terrain recognition accuracy, and 2) Does minimum data requirements (i.e., the minimum amount of wearable sensor data required for achieving high terrain classification accuracy) change depending on the IMU position?

The first question aims at evaluating to what extent walking terrain classification can be achieved using a single IMU and whether IMU position affects the ceiling of accuracy. The second question aims at reducing the data bandwidth (e.g., lowering sampling rate and duration of recording) for near real time data processing and freeing up the resources of the wearable device.

The remainder of the paper is organised as follows. The Section 2 describes the related work focusing on IMU-based terrain classification. It also lists the novel contributions of this study. The Section 3 details the methodology including the description of the datasets, data processing workflow, and the parameters investigated. The Section 4 presents the results. The Section 5 highlights the key findings, compares them with those presented in the previous studies, and discusses their relevance within the context of outdoor gait analysis. It also mentions the limitations of the current work and planned activities in the future. The Section 6 concludes the manuscript.

2. Related work

Previous research on IMU data processing has primarily focused on gait characterisation; i.e., mainly extraction of step-related spatial and temporal gait parameters (such as cadence, step length, step time, swing to stance ratio, double support time, gait asymmetry and variability). Most of these studies analysed data that was recorded in controlled laboratory settings [28, 29, 30, 31, 32]. In recent years, there has been attempts translating laboratory research into free-living environments [33, 34]. Although findings from free-living environment studies suggest terrain-induced gait alterations, only a handful of studies attempted for automatic terrain recognition from wearable data. A detailed review of these studies is presented below.

2.1. Previous studies on walking terrain classification

Hu *et al.* [15] collected data from 30 healthy adults (16 males and 14 females, age=23.5 ± 4.3 years, height=169.3± 21.9 cm, and weight=71.2 ±14.0 kg) while standing still and walking on four outdoor terrains: cobblestone, stairs up, stairs down, uphill, downhill, bank left and bank right. All walking terrains

including stairs were flat except uphill and downhill. Each participant performed 42 walking trials (six repetitions per terrain) while wearing daily sport shoes. All data collection was done during daylight and on dry and clear surfaces. In all trials, the participants walked 15 meters (at preferred speed) while carrying six off-the-shelf IMUs. The IMU positions were lower back, wrist, left thigh, right thigh, left shank and right shank. The IMU data were recorded at a sampling rate of 100 Hz, scaled between 0 and 1 and divided into 4-second segments with 3-second overlap. Each segment became a data object either in the training or test dataset. The study reports the performance of three deep artificial neural network classifiers (CNN - convolutional neural network, LSTM - long short time memory and Global-LSTM) when they were trained and tested using data from one IMU at a time, and using data from all IMUs combined. The highest accuracy for each IMU position were 84% (lower back, Global-LSTM), 90% (wrist, LSTM), 89% (left thigh, LSTM), 90% (right thigh, Global-LSTM), 88% (left shank, Global-LSTM), 89% (right shank, Global-LSTM) and 92% (all together, Global-LSTM). The study did not provide details on how models were trained (e.g., how data were split into train and test data, split ratio and number of repetitions). Note that Hu *et. al* made their dataset publicly available [35] and we used it in this study.

Sher *et. al* [16] collected data from 10 healthy adults (6 males and 4 females, age= 29 ± 8.7 years, height= 173 ± 7.9 cm, weight= 78.2 ± 16.2 kg) while walking on 10 outdoor terrains: pebble beach, sandy beach, forest, grass, pavement, synthetic rubber running track, stairs up, stairs down, uphill and downhill. Each participant performed 20 walking trials (two repetitions per terrain) and the duration of each walking trial was a minute. All data collection was done during daylight and surface conditions were not controlled (this was deliberate to create a more challenging dataset). In all trials, the participants walked back and forth at preferred speed between two landmarks which were 14-meters apart. The data collection was done using a mobile phone with embedded tri-axial accelerometer and gyroscope sensors (the phone position was lower back). The phone data were recorded at a sampling rate of 400 Hz, low-pass filtered using 10 Hz cut-off frequency, and divided into 2-second segments with 1-second overlap. These segments were further processed to extract features such as mean, std, skewness, kurtosis and dominant frequency. These features were used as input for terrain classification. The study reports the performance of seven machine learning classifiers (NB - naive Bayes, NN - nearest neighbour, FNN - fuzzy nearest neighbour, J48 - decision tree, JRip - a rule-based classifier, SVM - support vector machine classifier and MLP - multi layer perceptron classifier) using two different training methods: 1) training a separate classifier for each participant (i.e., personalised classifier) and 2) training one classifier for all participants (i.e., generalised classifier). The performance of the classifiers were evaluated using 10-fold cross validation with 90%-10% train-test data split. The personalised classifiers had much better performance (>90% classification accuracy) than the generalised classifier (<33%). Among personalised classifiers, the MLP classifier had the highest classification accuracy (97%). Compared to [15], this study had a richer terrain dataset based on numbers and topology). However, it had fewer participants and did not investigate how phone position would have affected the performance of the classifiers.

Dixon *et. al* [17] focused on the analysis of IMU data from running participants. They collected data from 29 healthy adults (15 males and 14 females, age= 23.3 ± 3.6 years, height= 180 ± 10 cm, weight= 63.6 ± 8.5 kg) while running on three outdoor tracks: concrete, synthetic and woodchip. All surfaces were flat. The study cohort included three groups of participants: those with no history of regular running (n=10), recreational runners who ran between 10 and 30 km every week in the last six months (n=9) and competitive runners running at least 50 km every week in the last six months (n=10). All participants performed two to four trials per track. In all trials, the participants ran 90 meters while carrying two tri-axial accelerometers (one at the lower back and one on the tibia). The accelerometer data were recorded at a sampling rate of 1024 Hz and divided into 4-second segments with 1-second overlap. Similar to [16], these segments were further processed to extract descriptive features (including few new ones such as sample entropy, autocorrelation and wavelet-derived energy) and these features were used as input to the XGBoost classifier (the full set and the reduced set including 20 best features were tested). In addition, a CNN classifier was trained using raw time series segments. Both XGBoost and CNN classifiers were trained and tested using data from one accelerometer at a time, and using data from both accelerometers combined. During training, the train-test data split ratio was 90%-10%, and training was repeated five times by randomly reshuffling the data. The performance of the XGBoost classifier were 94% (lower back), 94% (tibia) and 97% (combined)

for the full feature set, and 90% (lower back), 86% (tibia) and 88% (combined) for the reduced feature set. The performance of the CNN model was 96% (lower back), 91% (tibia) and 96% (combined).

Different from the studies mentioned above, Ahamed *et al.* [18] focused on environmental weather classification using wearable data from recreational runners. They collected data from six healthy adults while running on a pavement during winter (mean temp. $=-10^{\circ}\text{C}$) and spring seasons (mean temp. $=6^{\circ}\text{C}$). The participants included five females (age $=47\pm9.69$ years, height $=169.2\pm6.6$ cm, weight $=67.4\pm11.5$ kg) and one male (age $=29$ years, height $=170$ cm, and weight $=75$ kg). The criteria for recreational running was not specified. All participants performed two trials per season. In all trials, they ran at least 6 km while carrying an off-the-shelf IMU-based wearable device at the lower back. The device was outputting six gait parameters (pelvic drop, rotation and vertical oscillations, ground contact time, reduction in forward velocity following foot strike and cadence) at a sampling rate of 100 Hz. These parameters were used as input to RF classifier for binary classification (winter versus spring). The performance was evaluated under two training methods: 1) training a separate classifier for each participant with 70%-30% train-test data split (i.e., personalised classifiers), and 2) training one classifier for all participants (i.e., generalised classifier). One trial from each season (two trials in total) were used to train the generalised classifier and the rest of the trials were used to test it. The RF classifier had a classification accuracy of 95% (personalised classifiers) and 87% (the generalised classifier).

2.2. Novel contributions of this study

The four studies discussed in the previous section have a similar IMU data processing workflow which can be summarised in four steps. The first step is pre-processing focusing on noise filtering, data normalisation and walking onset-offset detection. The second step is segmentation where the time series data is divided into overlapping segments with fixed size using sliding window. The third step is feature extraction where each segment is processed to extract numerical values ranging from low level statistical features to actual gait parameters. The final step is terrain recognition where canonical machine learning classifiers are trained to create terrain class decision boundaries in the feature space. Note that in studies using deep learning models steps three and four are combined (i.e., segments are entered into the models directly) as feature extraction becomes part of the training process.

There are several key design choices that affect the overall terrain recognition accuracy in this workflow. These choices are the sensor configuration (do we record data from both accelerometer and gyroscope sensors or is recording from a single sensor sufficient to achieve high performance?), data sampling rate (how often do we record from the sensors?), segment length (i.e., the window size used during segmentation) and the machine learning classifier. In addition to accuracy, these design choices determine how much data is generated and how quickly it is processed.

The novel contributions of this study are:

1. To the best of our knowledge, this is the first study with systematic evaluation of how sampling rate, segment length and sensor configuration affect walking terrain recognition accuracy in relation to the IMU position.
2. It also compares the performance of five canonical machine learning classifiers in terms of accuracy and minimum data requirements. Three of these classifiers, XGBoost, SVM and RF, have been investigated separately in the literature. Here, we wish to benchmark them using the same dataset.
3. To strengthen the validity of our results, these evaluations are done using two independent datasets. The first dataset is a new dataset (see next item). The second dataset is the open access dataset [36] presented in the previous section while describing [15].
4. It presents a new dataset featuring phone IMU data which will be made publicly available after the publication of this manuscript. The detailed description of this dataset can be found in the next section (Section 3).
5. It evaluates whether IMUs embedded in mid-range smartphones are good enough to recognise outdoor walking terrains accurately. Compared to custom-built wearable solutions, creating a system that runs on a smartphone is desirable given its ubiquity in daily use, and there are increasing efforts in evaluating the reliability and validity of smartphones in gait analysis (e.g., [37]).

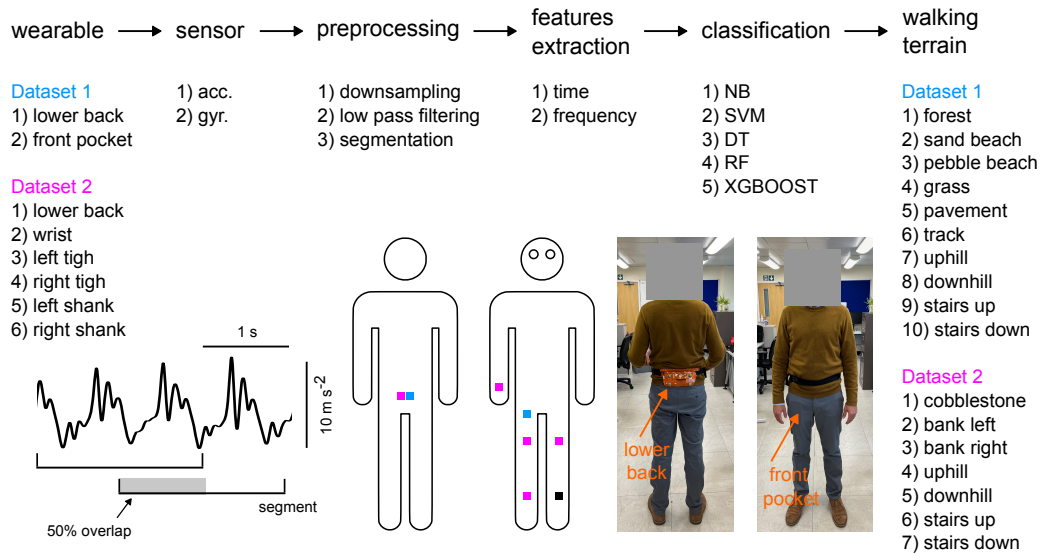


Fig. 1: Data processing workflow. Human silhouettes show device placements (filled squares) in dataset 1 (cyan) and dataset 2 (magenta). Human images show phone placements in datasets 1: the first phone was attached to the lower back using a fixation belt and phone holder (left image) and the second phone was placed in the front pocket (right image). The IMU placements in dataset 2 is depicted in Figure 2 in [15]. On the bottom left, an example of raw accelerometer data and its segmentation using sliding window (2 s) with 50% overlap is shown.

3. Methodology

Figure 1 shows the overall data collection and processing workflow proposed for automatic walking terrain recognition. Below, each part is described in detail.

3.1. Datasets

3.1.1. Dataset 1

Participants. Twenty four community dwelling young adults (15 males: 30.3 ± 8.8 years, 76.6 ± 14.8 kg, 175 ± 9 cm and 9 females: 26.2 ± 6.7 years, 70.1 ± 8.8 kg, 169 ± 3 cm) voluntarily participated in the study, including students and staff members studying and working at Aberystwyth University and are residing in the Ceredigion county, Wales. All participants were healthy with no history of falls or injuries in the past year and no apparent neurological or chronic condition that could affect their gait or compromise their safety in free-living walking conditions. The study was approved by the Aberystwyth University Ethics Committee Board and all experiments were carried out according to the Declaration of Helsinki. Informed consent was obtained from all participants.

Instrumentation. A custom-built mobile app was used to record data from mobile phone accelerometer and gyroscope sensors with an average sampling rate of 400 Hz [38]. The app was designed in Android Studio (API Level 34). Once opened, it lists all available sensors on the phone and allows the user to select one or more sensors for recording. At the beginning of each recording, the app prompts the user to enter a unique identifier including participant ID, walking terrain and trial number. The user starts and stops the recording manually. At the end of the recording, the data from selected sensors are written into separate .csv files (e.g., accelerometer.csv), which are stored in a zip file named after the unique identifier (e.g. Participant01_Forest_Trial1). The zip file is uploaded to an online server. If the phone did not have an internet connection, it is stored temporarily in the phone's local file store until the internet connection is re-established.

The experimental data was collected using two Android phones (Google Pixel 4). One phone was placed around the lower back (close to the L3 vertebra) using a phone holder and a fixation belt. The orientation

of the phone was such that the phone's x, y and z axes were corresponding to participants' vertical, medio-lateral (from left to right) and forward axes, respectively. The second phone was placed in one of the thigh pockets (front). This time the orientation of the phone was such that its x, y and z axes were corresponding to the participants' medio-lateral, vertical and forward axes, respectively. To reduce motion artefacts at the lower back, the belt was tightened as much as possible without discomforting the participants; however, the fixation of the phone in the front pocket was less controlled due to clothing of participants.

Walking terrains. Participants were asked to walk on 10 different terrains located around Aberystwyth town: i) forest, ii) sandy beach, iii) pavement (width=100 cm), iv) grass (height= 5.08 ± 1.5 cm), v) pebble beach, vi) synthetic rubber running track (width=125 cm), vii) upstairs (stairs width= 60 cm, stairs height=9 cm), viii) downstairs, ix) uphill (hill slope= 20°) and x) downhill (Fig. 2). Despite stairs descent and ascent are often considered as distinct activities from normal walking (e.g., [39]), we decided to include them because they are ubiquitous in free-living environments and lead to clearly visible changes in gait which can be considered as terrain-induced.

We tried to keep all walking experiments as real as possible and did not control for daily variations of terrain, weather and traffic conditions unless they were considered as safety hazard (e.g., slippery surface). If environmental conditions were considered unsafe for walking, the data recording session was postponed to another day and time. As a result, we were able to collect a rich dataset with gait alterations including corrective movements to avoid puddles or dried leaves on foot paths, more cautious walking on wet surfaces after rain, slower speeds while walking against the wind or easier walking on wet sand than on dry sand, and distractions caused by pedestrians walking nearby.

Data Collection. The experimental data were collected over 6 months between July 2021 (mean temp.= -17°C) and December 2021 (mean temp.= 6°C). All walking tests were carried out under the supervision of at least two experimenters. The participants walked (at a preferred speed) back and forth between two landmarks (orange cone) for one minute. The distance between the landmarks was set to 14-meters allowing 10-meter steady walk after each turn. Additional 2-meters on each side allowed participants to decelerate and accelerate safely while turning. Each measurement was repeated twice totalling 20 walking trials per participant (10 terrains $\times$ 2 repetitions). During walking trials, sensor data from two mobile phones were recorded simultaneously, and after quality control inspection, were uploaded to a secure server for offline data processing. In addition, each trial was filmed using a GoPro camera (240 frames per second). The video recordings were used to verify the surface and participant id. They also helped us keep track of external factors that might have affected gait (e.g. leaves and branches on the ground while forest walking). Note that a subset of this dataset including single phone measurements from limited number of participants ($n=10$) were used in [16].

3.1.2. Dataset 2

This open access dataset [35] included walking data from 30 healthy adults (16 males and 14 females, 23.5 ± 4.3 years, 169.3 ± 21.9 cm, and 71.2 ± 14 kg). The participants walked 15 meters on seven different terrains including: 1) cobblestone pavement, 2) bank left, 3) bank right, 4) stairs up and 5) stairs down, 6) uphill and 7) downhill. Additional measurements were taken while the participants were standing still. The participants wore six IMUs: one on the wrist, one on the lower back, two on the thighs, and two on the shanks. All sensors recorded data at 100 Hz. Each walking/standing experiment was repeated six times.

3.2. Data analysis

All data analysis were performed offline using custom-built Python scripts on a standard laptop (MSI, Intel Core i7 processor and 16 GB RAM). Unless stated otherwise, both datasets 1 and 2 were analysed using the same workflow described below.

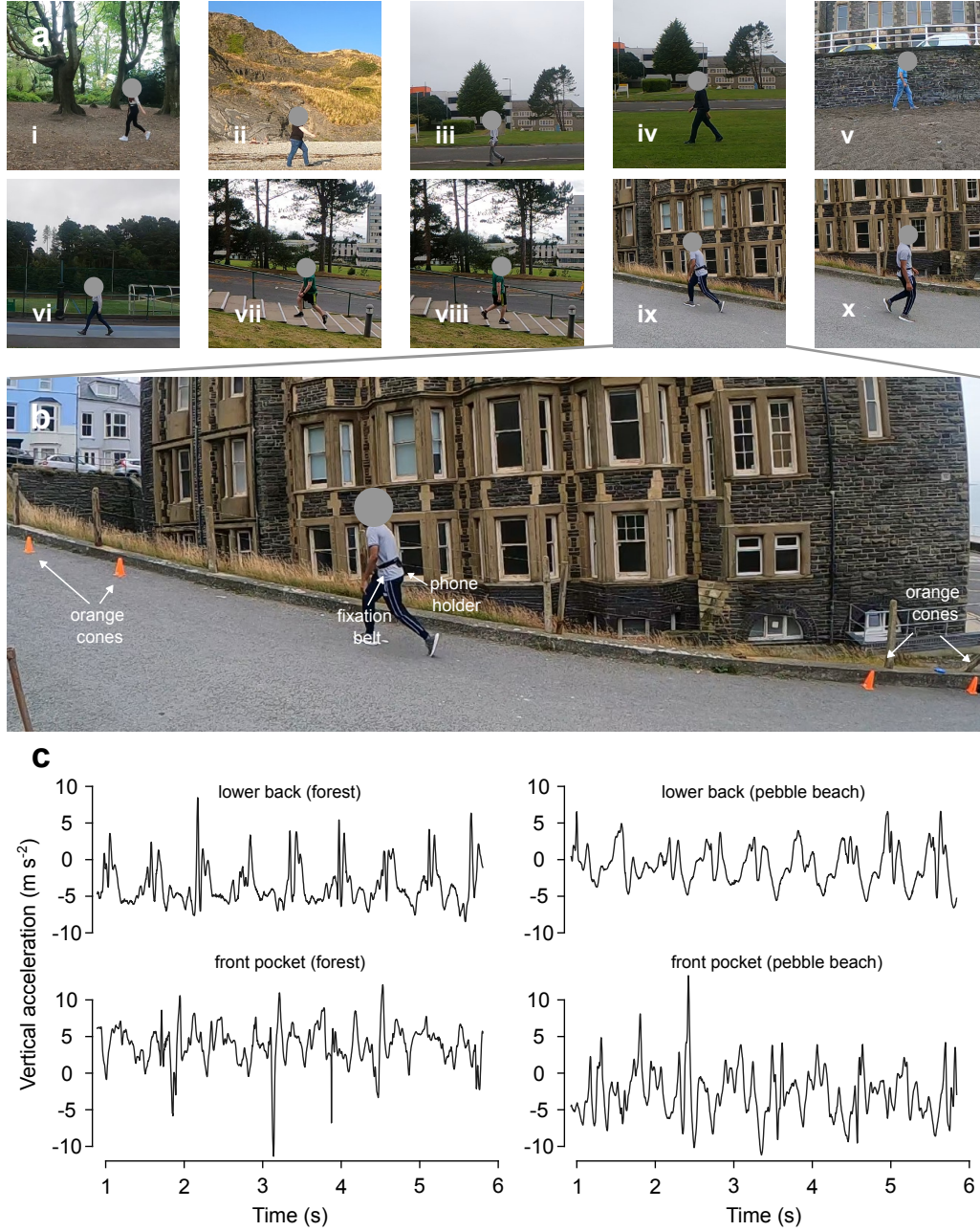


Fig. 2: Dataset 1. **a.** Outdoor walking terrains (close up snapshots): i) forest, ii) sandy beach, iii) pavement, iv) grass, v) pebble beach, vi) track, vii) upstairs, viii) downstairs, ix) uphill and x) downhill. **b.** Uphill full snapshot (ix in a). The orange cones, which were used as landmarks to mark walking path, were pointed. The distance between the inner cones was 10 m and the distance between the outer cones was 14 m. The fixation belt and phone holder, which were used to attach the phone to lower back of the subject, were also pointed. **c.** Examples of lower back and front pocket phone data (vertical acceleration) from walking trials in the forest and on the pebble beach.

3.2.1. Data preprocessing

The data was first filtered using a 2nd order Butterworth filter with a cutoff frequency of 10 Hz. This cut-off frequency allowed preserving the dominant frequency and its harmonics in the data. In addition, the first and last portions of each walking trial were trimmed to exclude portions that were not relevant to walking (five seconds in dataset 1 and two seconds in dataset 2).

The data was divided into segments using a sliding window with 50% overlap. The impact of segment length was evaluated systematically starting from 1-second segment length up to 4-second segment length with 0.5-second increments (seven segment lengths in total). Assuming that on average each step takes half second [16, 40, 41, 42, 43], the minimum segment length was chosen to ensure that it was long enough to include at least one gait cycle (i.e., one left and one right step). Similarly, 50% overlap was chosen to ensure that it generated overlapping segments with alternating gait cycles (e.g., the first segment included first left and first right steps, the second segment included first right and second left steps, third segment included second left and second right steps and so on). The maximum segment length was chosen to be consistent with the existing literature [15, 17].

To evaluate the impact of sampling rate, the data were [downsampled](#) from original 400 Hz to 200, 100 and 50 Hz (dataset 1) and from 100 Hz to 50 Hz (dataset 2). This was achieved by averaging consecutive data points (i.e., in dataset 1 two consecutive points were averaged to reduce the sampling rate to 200 Hz, four points to 100 Hz and 8 points to 50 Hz, and in dataset 2 two consecutive points were averaged to reduce the sampling rate to 50 Hz). This approach was preferred over subsampling approach to reduce the sensitivity of the data processing workflow to outliers.

3.2.2. Feature extraction

All segments were processed to extract the following time and frequency domain features: mean, std, min, max, median, skew, kurtosis, number of zero crossings, dominant frequency and dominant frequency amplitude. The feature extraction step generated 60 features per segment (10 features $\times$ six sensor channels including accelerometer and gyroscope x, y and z channels).

3.2.3. Classifier training and testing

Data segmentation and feature extraction steps created a series of two-dimensional sub-datasets which included extracted features from each segment. The size of each sub-dataset (number of segments $\times$ number of features) was determined by the segment length; the smaller the segment length, the more segments could be extracted from each walking trial (Table 1).

Table 1: Total number of segments in sub-datasets.

Segment length (s)	1	1.5	2	2.5	3	3.5	4
Dataset 1	31810	25145	19020	14720	11880	9828	8277
Dataset 2	40466	26183	19036	14741	11893	9840	8276

Five canonical machine learning models were employed using Python library scikit-learn: 1) Support Vector Machine (SVM), 2) Naive Bayes (NB), 3) Decision Tree (DT), 4) Random Forest (RF), and 5) Gradient Boosting (XGBoost). SVM, NB and DT were trained using default parameters. [Hyperparameter tuning using trial and error method was applied to optimise the performance of the RF and XGBoost classifiers. The number of trees and maximum depth in the RF was set to 100 and 10, respectively, and the node impurity was determined using the Gini index. The learning rate, booster and maximum depth in XGBoost were set to 0.3, softmax and 10, respectively.](#)

The analysis of the dataset 1 included 840 experiments investigating the effect of workflow parameters; i.e., four sampling rates $\times$ seven segment lengths $\times$ five classifiers $\times$ two device positions $\times$ three sensor configurations. Similarly, the analysis of the dataset 2 included 1260 experiments; i.e., two sampling rates $\times$ seven segment lengths $\times$ five classifiers $\times$ six device positions $\times$ and three sensor configurations (Table 2).

Table 2: Data processing workflow parameters investigated in this study. IMU position: LB - lower back; FP - front pocket; W - wrist; LT - left thigh; RT - right thigh; LS - left shin; RS - right shin. IMU sensors: A - accelerometer; G: gyroscope; A+G: accelerometer and gyroscope data combined.

Parameters	Dataset 1	Dataset 2
Sampling rate (Hz)	50, 100, 200, 400	50, 100
Segment length (s)	1, 1.5, 2, 2.5, 3, 3.5, 4	1, 1.5, 2, 2.5, 3, 3.5, 4
Classifiers	SVM, NB, DT, RF, XGBoost	SVM, NB, DT, RF, XGBoost
IMU position	LB, FP	LB, W, LT, RT, LS, RS
IMU sensors	A, G, A+G	A, G, A+G

Each experiment was repeated 10 times with randomized 80%-20% train-test data splits, and results were reported as the mean of the 10 repetitions and the standard deviation of the mean, unless stated otherwise. For each repetition, the performance of the classifiers were evaluated based on their overall surface classification accuracy. The accuracy was calculated as the percentage of correctly classified instances: i.e., number of correct predictions divided by the total number of predictions and multiplied by 100.

3.2.4. Statistical analysis

The statistical tests were performed using the statistical functions from Python library SciPy and statsmodels. The Shapiro-Wilk test was used to evaluate the normality of the performance results. In both datasets, the null hypothesis was accepted which confirmed that the data were normally distributed ($P > 0.1$). Hence, one way analysis of variance (ANOVA) followed by Tukey's HSD post hoc test were run to compare the performance of the classifier(s) within and across IMU position at a significance level of 0.05 or lower.

During statistical tests, the accelerometer and gyroscope data were used together, and the sampling rate and the segment length of each classifier were optimally set (i.e., the lowest sampling rate and segment length that achieved accuracy that was within 1% of the maximum accuracy among all pairs of sampling rate and segment length investigated).

In dataset 1, the best performing classifier(s) were further evaluated by visualising their terrain confusion matrix and by calculating their precision (true positives divided by true positives plus false positives), recall (true positives divided by true positives plus false negatives) and F1-score (two times precision times recall divided by precision plus recall).

4. Results

4.1. Dataset 1

4.1.1. Performance of the classifiers

There was statistically significant difference between the performance of the classifiers (ANOVA, $P < 0.01$) (Fig. 3b), the XGBoost classifier performed significantly better than the NB (45%, Tukey's HSD, $P < 0.01$) and SVM classifiers (30%, Tukey's HSD, $P < 0.01$) regardless of the phone position (Fig. 3a).

In general, the classification accuracy of the lower back classifiers were higher than the accuracy of the front pocket classifiers (although the difference was not statistically significant). The difference between two phone locations was 3% (for XGBoost and RF classifiers), 2% (for DT), 10% (for SVM) and 14% (for NB) (Fig 3a).

Given that the XGBoost was the best performing classifier, we had a closer look at its performance on recognising each walking terrain. The precision, recall and F1-score show that the performance of the XGBoost classifier was consistent across all terrains (Table 3). All values were above ≥ 0.9 for lower back and ≥ 0.8 for front pocket. On rare occasions, the XGBoost classifier failed to identify the correct class. Mislabelling occurred between similar terrains, for instance between pebbled and sandy beach, track and pavement, and uphill and upstairs (e.g., see the confusion matrix presented in Fig. 4).

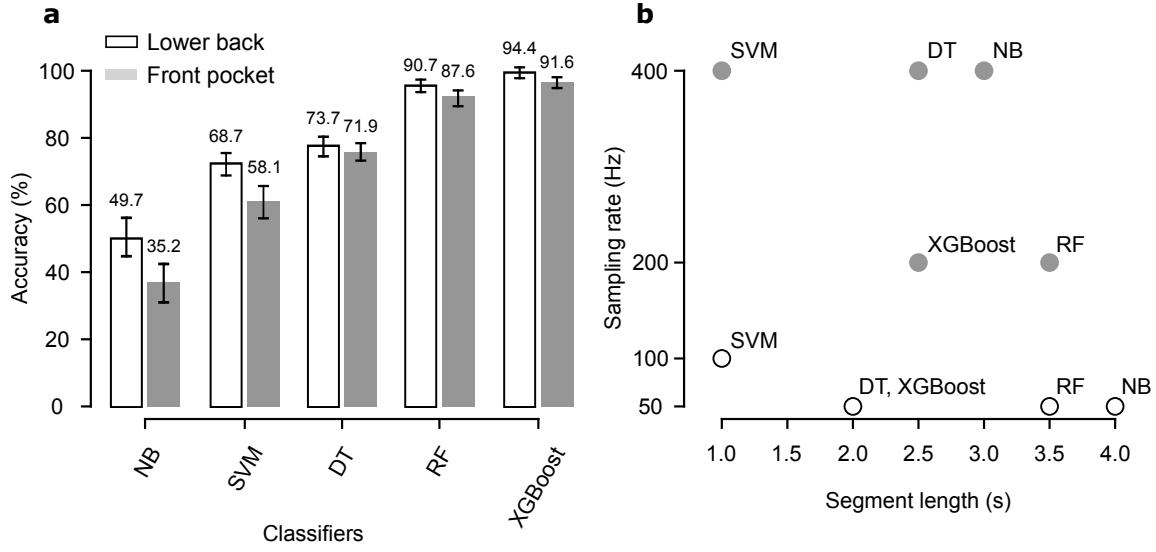


Fig. 3: Dataset 1. Comparison of classifiers when data from both accelerometer and gyroscope sensors was included in the analysis. **a.** Terrain classification accuracy for lower back (empty bar) and front pocket (grey bar) data. The error bars show to standard deviation from the mean. **b.** Optimal sampling rate and segment length for each machine learning classifier for lower back (empty circular marker) and front pocket (grey circular marker) data.

Table 3: Performance of the XGBoost classifier. n is to the number of test segments for each terrain.

Surface	Lower back				Front pocket			
	n	Precision	Recall	F1-score	n	Precision	Recall	F1-score
DownHill	86	0.95	0.96	0.95	65	0.88	0.91	0.91
DownStairs	94	0.98	0.98	0.98	67	0.91	0.96	0.93
Forest	77	0.96	0.94	0.95	71	0.92	0.99	0.95
Grass	113	0.96	0.96	0.96	87	0.95	0.92	0.94
Pavement	118	0.94	0.92	0.93	84	0.89	0.96	0.93
Pebble	111	0.94	0.93	0.94	66	0.97	0.91	0.94
Sand	123	0.93	0.95	0.94	76	0.89	0.93	0.91
Track	106	0.98	0.95	0.97	62	0.93	0.89	0.91
UpHill	97	0.96	0.95	0.95	64	0.89	0.84	0.86
UpStairs	107	0.90	0.96	0.93	53	1.00	0.81	0.90

4.1.2. Effect sampling rate, segment length and sensor configuration

The optimal sampling rate for each classifier was distinct and varied depending on the phone position (Fig. 3b). In general, the lower back classifiers required a lower sampling rate (≤ 100 Hz) than the front pocket classifiers (≥ 200 Hz). The systematic evaluation of the downsampling effect corroborated this observation (Fig. 5). The performance of the lower back classifier at 50 Hz was as good as its performance at 400 Hz (Fig. 5a). On the contrary, the performance of the front pocket classifier increased with increasing sampling rate, the accuracy at 400 Hz was 40% higher than the accuracy at 50 Hz (Fig. 5b).

The optimal segment length for each classifier was also distinct (except for the DT and XGBoost classifiers, Fig. 3b) and varied slightly with the phone position. The highest classification accuracy for the lower back classifier was achieved when the segment length was set to 2 s (Fig 5a) whereas the highest classification accuracy for the front pocket classifier was achieved when the segment length was set to 2.5 s (Fig 5b). However, the accuracy of the front pocket classifier was equally good when the segment length was set to 2 s, only 1% lower the highest accuracy (Fig 8a). Compared to the minimum segment length tested in the study (i.e., 1 s), the optimal segment length improved classification accuracy by 3% (lower back) and 4% (front pocket) (Fig 8a).

Combining accelerometer and gyroscope data improved the performance of the XGBoost classifier but

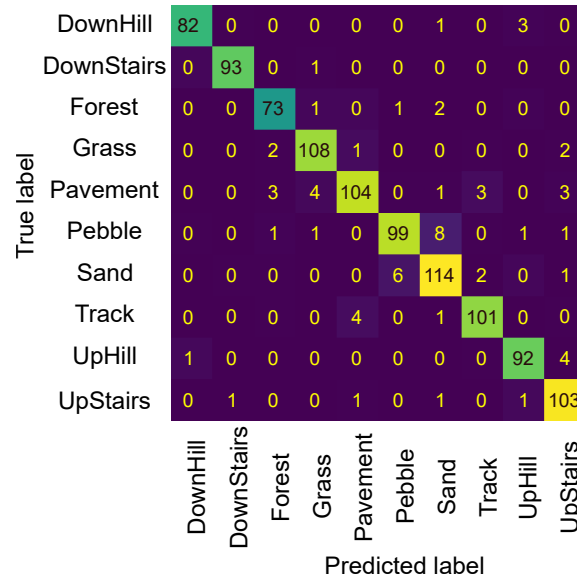


Fig. 4: Confusion matrix of the XGBoost classifier for lower back data (cumulative over 10-fold cross validation). The colour of each cell is proportional to the number of test segments evaluated; i.e., dark blue (low number of segments) and yellow (high number of segments).

not by much (Fig 5). The difference was 3% (accelerometer) and 13% (gyroscope) for the lower back and 4% (accelerometer) and 6% (gyroscope) for the front pocket.

4.2. Dataset 2

4.2.1. Performance of the classifiers

Similar to the dataset 1, there was statistically significant difference between the performance of the classifiers (ANOVA, $P < 0.01$). The XGBoost classifier was significantly better than the other four classifiers regardless of the IMU position (Tukey's HSD, $P < 0.01$) (Fig 6a). On average, the difference in performance was 13% (RF), 25% (DT), 40% (SVM) and 44% (NB).

Different from the dataset 1, there was a notable change in terrain classification accuracy when device position was considered (Fig. 6b). The difference between the best (lower back) and runner up (shank and thigh) positions and between the best (lower back) and worst (wrist) positions were 10% and 25%, respectively.

4.2.2. Effect sampling rate, segment length and sensor configuration

Downsampling to 50 Hz did not have a negative impact on performance, except for the left shank position where the accuracy dropped by 10% (Fig. 7). In contrast, the segment length had a more profound effect on performance (Fig. 7). The classification accuracy increased up to 8% when the segment length was set to 3 s or higher (Fig. 8b). The trend was similar across all IMU positions (accuracy plateauing after 3 s) suggesting that there was no benefit of varying segment length with the IMU position.

Combining data from accelerometer and gyroscope sensors was beneficial up to a certain extent. The improvement in performance was 3% (lower back), 3% (wrist), 6% (left thigh), 6% (right thigh), 5% (left shank) and 2% (right shank). If data from a single sensor were considered, the classifiers using accelerometer data performed better than the classifiers using gyroscope data at four out of six positions: 13% (lower back), 14% (wrist), 4% (left shank) and 5% (right shank). For the other two positions, the classifiers using gyroscope data were better: 5% (left shank) and 2% (right shank).

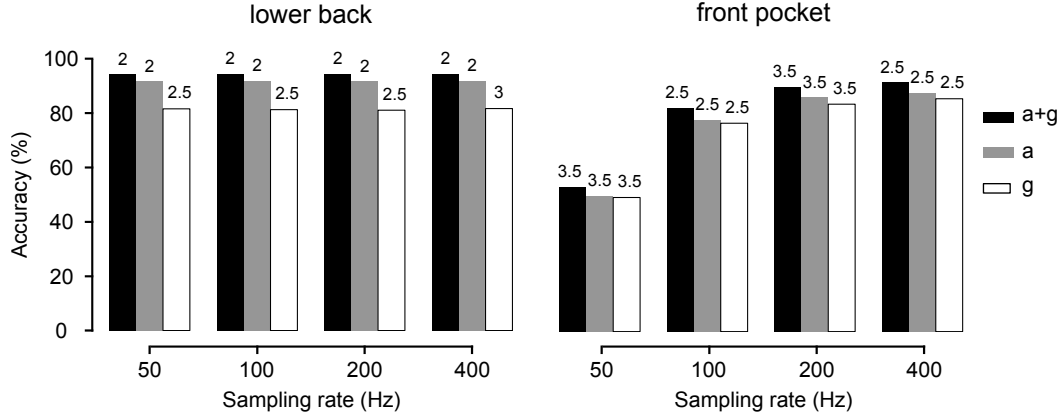


Fig. 5: Accuracy of XGBoost classifiers as a function of sampling rate and sensor configuration (dataset 1). The numbers on each bar indicates the optimal segment length (in seconds) required for the highest accuracy.

5. Discussions

5.1. Key findings

In this study, we investigated whether individuals alter their gait while walking on different terrains (i.e., terrain-induced gait alterations), and the current IMUs (found in mobile phones and other wearable devices) are sensitive and accurate enough to detect these alterations robustly. We evaluated the effectiveness of canonical machine learning classifiers in automatic terrain recognition and whether the minimum data required to achieve high accuracy vary depending on the IMU position. Five machine learning classifiers (NB, SVM, DT, RF and XGBoost) were trained and tested using two different datasets. The first dataset featured phone recordings from 24 participants, whereas the second dataset featured off-the-shelf wearable device recordings from 30 participants.

The key findings, which were consistent between two datasets, are the following:

1. The XGBoost classifier outperformed the other classifiers (4% in dataset 1 and $\geq 10\%$ in dataset 2).
2. The IMU position affected the classification accuracy with optimal position being the lower back. In dataset 1, the difference between lower back and front pocket was marginal (and statistically not significant). In dataset 2, the lower back was significantly better than the other positions ($\geq 10\%$).
3. The IMU position also affected the minimum sampling rate required for high classification accuracy. Higher sampling rate was beneficial for front pocket (dataset 1) and left shank (dataset 2) positions improving performance by 40% and 10%, respectively.
4. The minimum segment length required for high performance was 2 s in dataset 1 and 3 s in dataset 2. Optimising segment length for each IMU location resulted in small improvement in accuracy ($\leq 2\%$).
5. Compared to data from a single sensor, combining accelerometer and gyroscope data improved the classification accuracy up to 4% in dataset 1 and up to 6% in dataset 2.

5.2. Comparison with previous studies

5.2.1. XGBoost classifier

To the best of our knowledge, this is the first study evaluating the performance of the XGBoost classifier within the context of walking terrain recognition complementing [17] where authors focused on running terrain recognition. Table 4 compares the performance of the XGBoost classifier to those reported in previous terrain recognition studies when IMU was placed on the lower back.

In dataset 1, the performance of the XGBoost classifier was notably better than the generalised MLP classifier ($>60\%$) and approached the performance of the personalised MLP classifiers ($<3\%$) reported in [16].

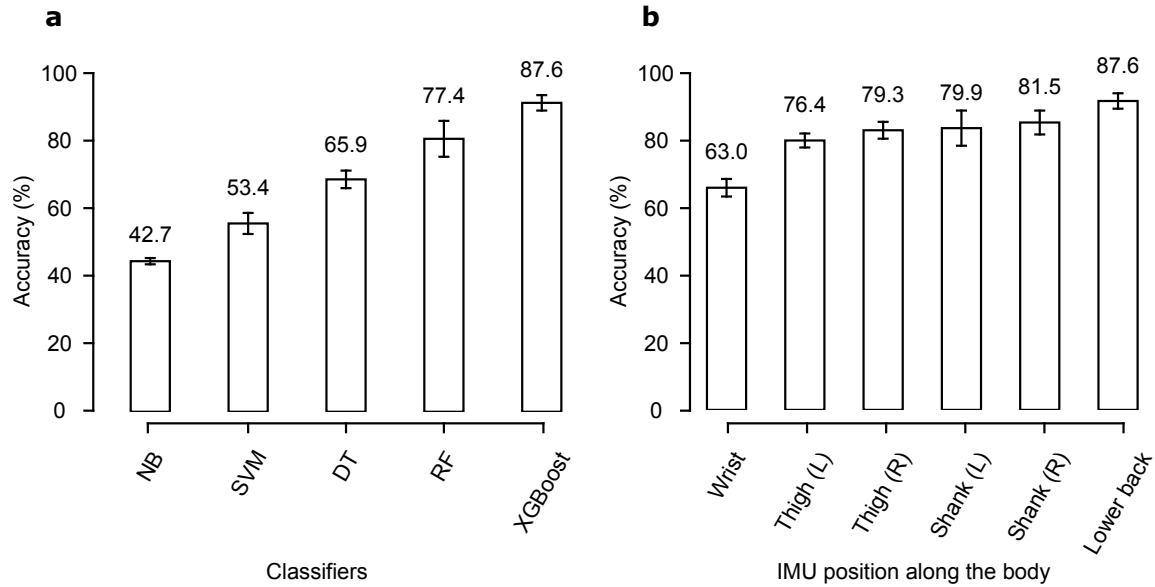


Fig. 6: Dataset 2. **a.** Comparison of classifiers when data from both accelerometer and gyroscope sensors was included in the analysis (lower back position). **b.** Accuracy of XGBoost classifiers for each IMU position. For this analysis, the sampling rate and segment length were set to 100 Hz and 2 s, and the data from accelerometer and gyroscope sensors were combined.

However, this comparison is rather inconclusive given that the low performance of the generalised MLP in [16] could be due to training on fewer participants (10 instead of 24).

In dataset 2, the performance of the XGBoost classifier was better than the Global-LSTM classifier (6%) reported in [15]. However, this was not the case for other IMU positions where the deep learning classifiers outperformed the XGBoost classifier by a wide margin: wrist (23%), left shank (17%), right shank (3%), left thigh (10%) and right thigh (7%). We discuss that this could be related to the feature extraction step in our data processing workflow. Compared to the deep learning classifiers which had the advantage of extracting their own features from each IMU position independently, we trained the XGBoost classifiers on standard time and frequency domain features. It is likely that these features were not good enough to create precise decision boundaries between terrain classes. To test this hypothesis, we need to re-train the XGBoost classifier with more complex features (e.g., [waveform length](#), [Wilson amplitude](#) and [features extracted from Wavelet decomposition](#)) and see whether its performance approaches to deep learning classifiers.

Previous research has shown that wrist IMUs can successfully differentiate walking from other activities of daily living (e.g., [44]) and estimate basic gait parameters such as step counts (e.g., [45]). However, they may fall short detecting settled gait alterations during walking because arm movements are highly variable among individuals. This may explain the lower performance of the XGBoost classifier at the wrist position. In addition, it is worth noting that IMUs on the right side had a higher performance than the IMUs on the left side. (14% - shank and 3% - thigh). We wonder if this could be related to gait symmetry and laterality of the participants (information which is not reported in the dataset). In normal gait, differences between the right and left limb movements are not uncommon (see [46] for an extended review), which may become more apparent while walking on outdoor terrains.

Overall, the superior performance of the XGBoost classifiers is not entirely surprising as previous research on fall risk estimation [47], Parkinson's gait modelling [48] and post-stroke rehabilitation assessment [49] have shown that XGBoost classifiers can be better at handling noisy and multi-dimensional time series data.

5.2.2. New dataset

The complete version of the dataset 1 is presented here for the first time. Compared to the open access dataset 2, the dataset 1 features four novel outdoor terrains (i.e., grass, forest, sandy beach and pebble beach)

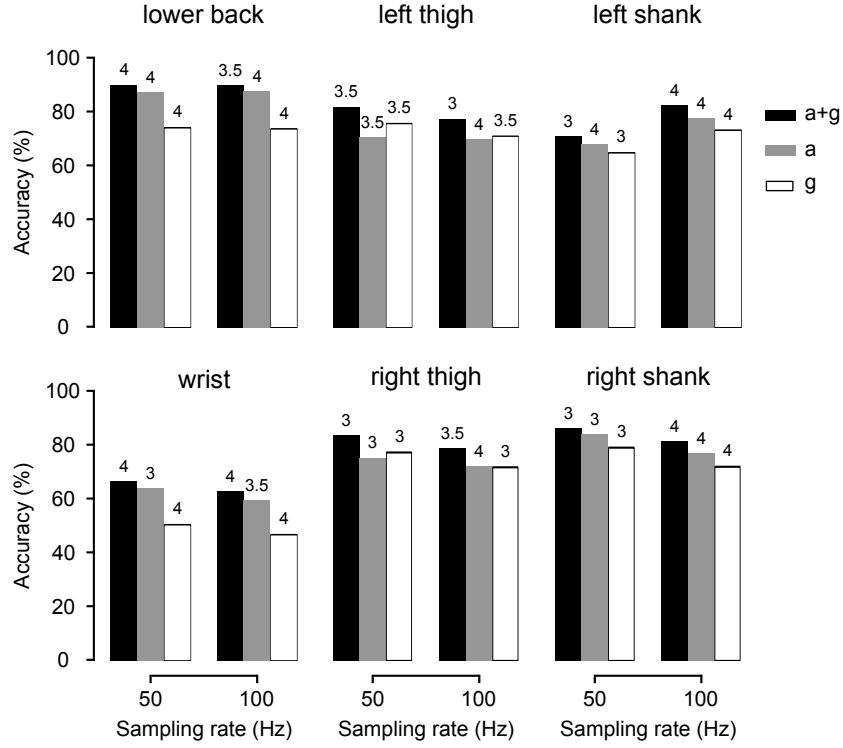


Fig. 7: Accuracy of XGBoost classifiers as a function of sampling rate and sensor configuration (dataset 2). The number on each bar indicates the optimal segment length (in seconds) required for the highest accuracy.

and one new IMU position (i.e., front pocket). It has a much higher sampling rate, and the walking trials were recorded using smartphones instead of off-the-shelf custom devices. In contrast, the dataset 2 has more participants (30 instead of 24) and IMU positions (six instead of two). The datasets 1 and 2 together form the largest outdoor walking database for wearable gait analysis.

5.3. Trade-off between accuracy and minimum data

Achieving high classification accuracy using as little data as possible is desired for improving battery, memory and CPU usage. The effect of lowering sampling rate, decreasing segment length, using data from a single sensor (instead of two) and reducing the number of features extracted from the sensor data are discussed below.

5.3.1. Sampling rate

Lowering sampling rate did not affect the performance of the IMUs' placed at the lower back (in both datasets) and on the wrist, left thigh, right thigh and right shank (in dataset 2). For these locations, the performance of the XGBoost classifier at 50 Hz was as good as, if not better than, the performance of the XGBoost classifier at the maximum available sampling rate which was 400 Hz in dataset 1 and 100 Hz in dataset 2. In these circumstances, lowering sampling rate would result in 8-fold data reduction in dataset 1 and 2-fold reduction in dataset 2. In contrast, it was necessary to use higher sampling rate for IMUs placed in the front pocket (dataset 1) and on the left shank (dataset 2). In dataset 1, this could be due to participants' clothing (e.g., large pockets). When not secured tightly in the pocket (and this was the case for most participants), phone vibrations would cause high frequency motion artefacts in the data. It is likely that these artefacts were more prominent while walking on uneven terrains (or while going up and down)

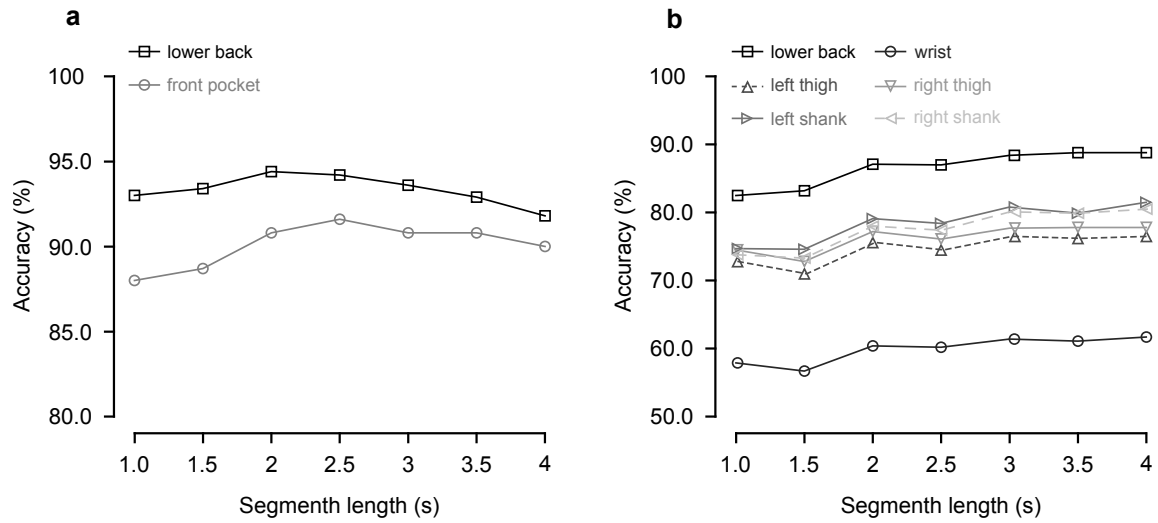


Fig. 8: Accuracy of the XGBoost classifier as a function of segment length. **a.** Dataset 1. **(b.)** Dataset 2

Table 4: Studies focusing on walking terrain classification. In the Dataset column, the number of participants and the number of walking terrains are given in parenthesis. Dataset 1* used in [16] is a subset of the dataset 1 used in this study. Also note that dataset 3 investigated gait alterations during running instead of walking (still included in the table for comparison). In the Approach column, we describe how classifiers were trained and tested: generalised (one classifier for all participants) and personalised (a separate classifier for each participant). Note that the table includes results only from the lower back as this was the only in common device position across all studies.

Study	Dataset	Activity	Approach	Classifier	Accuracy
This study	1 (24,10)	Walking	Generalised	XGBoost	94%
	2 (30,7)	Walking	Generalised	XGBoost	90%
[16]	1* (10,10)	Walking	Personalised	MLP	97%
	1* (10,10)	Walking	Generalised	MLP	28%
[15]	2 (30,7)	Walking	Generalised	Global-LSTM	84%
[17]	3 (29,3)	Running	Generalised	XGBoost	94%

providing salient cues to differentiate between terrains. In dataset 2, it was surprising that higher sampling was only necessary for the IMU placed on the left shank and not on both sides (or vice versa), which could be due to the laterality of the participants.

5.3.2. Segment length

In contrast to the sampling rate, the minimum segment length required for successful terrain classification did not change with the IMU position. We observed that in general the analysis of the dataset 2 required longer segment lengths than the analysis of the dataset 1. This could be due to the IMU sensors used in the study or walking speed of the participants (although there is no evidence for either of these postulations). We speculate that increasing segment length would be beneficial in scenarios when sensor measurements are noisy (i.e., low signal to noise ratio) or when people are walking at a slower pace so that key signatures in the data can still be captured with high fidelity.

5.3.3. Sensor configuration

Our findings suggest that combining data from accelerometer and gyroscope sensors would slightly improve the classification accuracy regardless of the device location. However, this improvement came at the cost of doubling the amount of sensor data. When combined with other data sampling parameters (e.g., sampling rate), recording data from only one sensor could drastically reduce the amount of data without dropping accuracy. For instance, in dataset 1, the performance of the lower back XGBoost classifier using accelerometer data at 50 Hz (i.e., minimal model) was still above 90%, and it was only 3% lower than the performance of the XGBoost classifier using both accelerometer and gyroscope data at 400 Hz (i.e., full model). Yet, the minimal model required 16 times less data than the full model. To put this into perspective, the memory requirement for a minute long recording using 8-byte double data type would be 72 MB for the minimal model and 1.2 GB for the full model. In addition, the size of the minimal XGBoost model was 38% smaller than the size of the full model ($2.3 \text{ MB} \leq 3.9 \text{ MB}$). Similarly, in dataset 2 the performance of the lower back XGBoost classifier using accelerometer data at 50 Hz was only 2% lower than the performance of the XGBoost classifier using both accelerometer and gyroscope data at 100 Hz. In this case, the data reduction was four times.

5.3.4. Feature selection

The preliminary results from our initial investigation suggest that the size of the machine learning classifiers could be further reduced using feature selection. For instance, in dataset 1 (lower back data, segment duration: 2 s, sampling rate: 400Hz), the accuracy of the XGBoost models with 12 features (selected using recursive feature elimination method from Python's Scikit-learn library) were only 7% (sensor configuration: a+g), 11% (a) and 7% (g) lower than the accuracy of the models with complete feature set. It is also worth noting that we may not need to record data from all three accelerometer channels. For instance, the 12 features mentioned above were calculated from accelerometer x and z channels which were mostly sensitive to movements along longitudinal (vertical) and anterior-posterior (forward) axes, respectively.

Our findings are inline with previous studies which showed that sampling rate, segment length and sensor configuration affect the performance of accelerometer-based human activity recognition systems. Khan et al. [26] used a statistical approach to determine the optimal sampling rate for a wrist-worn accelerometer during human activity recognition. They lowered the sampling rate until the distribution of the downsampled data deviated from the reference distribution at the original sampling rate. Their analysis on six different datasets showed that 50 Hz is more than enough to differentiate between the activities of daily living such as walking, sitting, going upstairs, playing soccer and vacuuming and detect gait abnormalities such as freeze (up to 10-fold data reduction was achieved through this process). Fida-et-al [50] evaluated the performance of the waist-worn accelerometer when segment length was varied between 0.5 s and 3 s with 0.5 s increments, and showed that 1.5 s was sufficient to achieve 90% classification accuracy recognising activities of daily living and transitions from one activity to another. Jain and Kanhadgan [51] investigated the importance of combining data from accelerometer and gyroscope in smartphone-based human activity recognition system which employed a data processing pipeline similar to ours shown in Figure 1. They showed that sensor fusion improved classification accuracy by 4%. They also showed that combining time and frequency domain features was beneficial.

5.4. Importance of device position

Our findings suggest that outdoor walking terrain classification using IMU data is possible but performance may vary depending on the IMU position. In a way this is not surprising given that human gait is complex and three-dimensional with each body segment playing a different role (e.g., arm movements for maintaining stability and shank/thigh movements for controlling ankle) [52, 53]. In addition, previous hill walking studies report that lower limb joint angles vary depending on the slope and walking speed [54, 55], suggesting key features contributing to terrain classification may be different for each IMU position.

The IMUs placed at the lower back was closer to the centre of mass, and the data recorded from its sensors were more symmetric (i.e., right and left steps produced similar acceleration and velocity profiles). The peaks and troughs in the forward acceleration were closely aligned with step-related temporal gait

parameters (i.e., heel strike and toe-off time points). It was also possible to differentiate between right and left steps by analysing the angular velocity along the vertical axis which were correlated to clockwise (left step) and anticlockwise (right step) hip rotations. On the other hand, the IMUs placed in the front pocket and lower limbs were more sensitive to the ipsi-lateral side than the contra-lateral side. Hence, the data were less symmetric (i.e., the steps on the ipsi-lateral side generating larger acceleration and velocity amplitudes). While the angular velocity around the medio-lateral axis was less informative in the lower back data, it was highly correlated with the leg swings in the front pocket and thigh data.

5.5. *Terrain-induced gait alterations*

The inspection of the GoPro videos in dataset 1 gave some insight into terrain-induced gait alterations. The participants had higher speed and took longer steps while walking downhill (and going down the stairs) than while walking uphill (and climbing up the stairs). They exhibited more corrective movements, while walking on the pebble beach, which resulted in smaller steps, side-stepping and reduced speed. This was presumably to maintain balance. Similarly, the participants had to exert more energy while walking on dry sand because of reduced ground reaction forces (due to the movement of the sand granules). As a result of this, they walked slower and their gait mechanics changed (e.g., back arching forward and increased foot-ground contact time). These changes were not present in participants walking on the compact sand few hours after high tide.

It is known that cognitive loading also affects gait kinematics [56]. It is possible that some of the gait changes observed in our study were induced by cognitive-motor interference [57]. For instance, the experiments involving uphill and downhill walking were performed in a relatively busier area of the town, and some participants got distracted by the passing pedestrians, dog walkers and cyclists (even though their paths did not cross). In addition, the participants walking along the forest spent more time looking at the ground presumably to avoid bumps, logs, branches, rocks and holes at varying size and dimensions. Similarly, the participants had to constantly watch where they step while walking on the pebble beach (e.g., to identify stable rocks with large surface area). In all these scenarios, it was most-likely that walking was less automatic and required executive decision making inducing gait alterations.

5.6. *Potential applications*

The study paves the way for automatic assessment of outdoor walking using mobile phone and other wearable devices. This system can be used in health and community care as an early warning system (abnormal walking may be indicative of injury or neurological movement disorders), an outcome measure (to measure quality of life and independence) or a progress monitoring tool (during and after interventions). It can also be used for designing better footwear and assistive walking devices, and in urban planning to prevent falls by creating safer walking paths especially for children and older populations. In this context, irregular behaviour in temporal and spatial gait parameters may be indicative of surface degradation or pot-holes). Finally, it can be used to detect terrain transitions and provide sensory feedback to control assistive walking devices (e.g., exoskeletons, orthotic and prosthetic devices [58]).

5.7. *Limitations and future work*

Both datasets 1 and 2 include recordings from young, healthy and well-rested adults. These recordings were also made in a relatively controlled setting where the participants were given clear instructions at the beginning of the walking trials, and they were asked to walk at a steady pace. To expand our analysis, we have started collecting data from older adults and participants with impaired walking abilities. Some of these recordings are taking place in our research laboratory to study gait alterations during unsteady walking [38], and some in more natural settings such as clinics, households and supermarkets. In particular, we are interested in neurological gait disorders (e.g., hemiplegic and Parkinsonian gait) and their rehabilitation and management using an integrated approach [5].

The proposed approach recognises walking environment indirectly from gait alterations associated with adaptations to terrain. When a change in gait pattern is detected, it cannot differentiate whether this is a terrain-induced change, pathological (e.g., injury or disease) or due to some other reasons (e.g., footwear).

This is a limitation considering alternative methods (e.g., vision based methods) capturing walking terrain directly (e.g., [59]). One approach to circumvent this issue would be GPS verification or user prompting to acquire location data. However, this approach may not be sufficient to detect terrain differences within the same geographical location.

The data analysis workflow included canonical machine learning classifiers trained on low level, hand-crafted features. The workflow can be improved in several ways: for instance, expanding feature set or using deep learning models to extract features from the raw data automatically. In particular, we are interested in features that combine information from multiple channels and sensors.

We are working on terrain-aware parameter tuning for accurate estimation of temporal and spatial gait parameters in outdoor environments. In [16], we demonstrated the utility of this approach by improving step time estimation up to 40%. Next step is to expand our analysis to other gait parameters including swing to stance ratio and symmetry between left and right sides of the body.

Finally, we acknowledge that for the most part our interpretation of the experimental results are speculative as we still lack a theoretical understanding of the neural, physiological and physical mechanisms governing terrain-induced gait alterations.

6. Conclusion

In this study, we focused on automatic walking terrain recognition in free-living environments. We presented a data processing workflow for analysing data from a single IMU, and evaluated the performance of five canonical machine learning classifiers under different data sampling configurations. Our results from two independent datasets showed that both accuracy and minimum data requirements vary with IMU position, and choosing correct data sampling parameters can improve classification accuracy up to 40 or reduce data size by 16 times. Additionally, the scarcity of benchmark datasets for outdoor walking terrain classification and gait analysis is a pressing issue in the field. To address this need to some extent, we introduce a new dataset that includes smartphone measurements from 24 participants walking on 10 different outdoor terrains.

Acknowledgement

We thank Megan Taylor Bunker for helping with the data collection. We also thank Dr. Federico Villagra Povina, Dr. Amanda Clare and Prof. Trevor Martin for their constructive feedback on the earlier version of the manuscript.

Author contributions

O.A. and A.S. designed the study. A. S. collected and analysed the data. A. S. produced the first draft of the manuscript and O. A. produced the final version. Both authors have read and agreed to the final version of the manuscript.

Funding

This research was funded by the European Commission (H2020-MSCA-RISE-2019, grant number: 873178) to O.A., Health and Care Research Wales (grant number: HS-20-42) to O.A. and the Aberystwyth University AberDoc PhD Scholarship to A.S.

Institutional review

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Aberystwyth University's Research Ethics Board.

Informed consent

Informed consent was obtained from all subjects involved in the study.

Data availability

Data will be made available through Aberystwyth University's research management portal.

Conflicts of interest

The authors declare no conflict of interest.

References

- [1] Xiuli Zhang, Max R Paquette, and Songning Zhang. A comparison of gait biomechanics of flip-flops, sandals, barefoot and shoes. *Journal of foot and ankle research*, 6(1):1–8, 2013.
- [2] Fay B Horak and Martina Mancini. Objective biomarkers of balance and gait for parkinson's disease using body-worn sensors. *Movement Disorders*, 28(11):1544–1551, 2013.
- [3] Christian A Clermont, Lauren C Benson, W Brent Edwards, Blayne A Hettinga, and Reed Ferber. New considerations for wearable technology data: changes in running biomechanics during a marathon. *Journal of applied biomechanics*, 35(6):401–409, 2019.
- [4] Erica Twardzik, Kate Duchowny, Amby Gallagher, Neil Alexander, Debra Strasburg, Natalie Colabianchi, and Philippa Clarke. What features of the built environment matter most for mobility? using wearable sensors to capture real-time outdoor environment demand on gait performance. *Gait & posture*, 68:437–442, 2019.
- [5] Meryem Şahin Erdoğan, Esra Sümer Arpak, Cemre Su Kaya Keles, Federico Villagra, Esin Öztürk Işık, Nazire Afşar, Can A Yucesoy, Luis AJ Mur, Otay Akanyeti, and Hale Saybaşı. Biochemical, biomechanical and imaging biomarkers of ischemic stroke: Time for integrative thinking. *European Journal of Neuroscience*, 2024.
- [6] Stefano Abbate, Marco Avvenuti, Francesco Bonatesta, Guglielmo Cola, Paolo Corsini, and Alessio Vecchio. A smartphone-based fall detection system. *Pervasive and Mobile Computing*, 8(6):883–899, 2012.
- [7] Guglielmo Cola, Alessio Vecchio, and Marco Avvenuti. Continuous authentication through gait analysis on a wrist-worn device. *Pervasive and Mobile Computing*, 78:101483, 2021.
- [8] Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.
- [9] Joseph Peters, Libak Abou, Ellyce Wong, Mauricette Sènan Dossou, Jacob J Sosnoff, and Laura A Rice. Smartphone-based gait and balance assessment in survivors of stroke: A systematic review. *Disability and Rehabilitation: Assistive Technology*, 19(1):177–187, 2024.
- [10] Maxwell Fordjour Antwi-Afari, Heng Li, Shah Nawaz Anwer, Sitsofe Kwame Yevu, Zezhou Wu, Prince Antwi-Afari, and Inhan Kim. Quantifying workers' gait patterns to identify safety hazards in construction using a wearable insole pressure system. *Safety science*, 129:104855, 2020.
- [11] Huining Li, Huan Chen, Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Xiaoye Qian, Dongmei Li, Ming-chun Huang, and Wenyao Xu. Neuralgait: Assessing brain health using your smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–28, 2023.
- [12] Rafael Caldas, Marion Mundt, Wolfgang Potthast, Fernando Buarque de Lima Neto, and Bernd Markert. A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait & posture*, 57:204–210, 2017.
- [13] Andrey Ignatov. Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, 62:915–922, 2018.
- [14] Shumei Zhang, Paul McCullagh, Chris Nugent, and Huiru Zheng. Activity monitoring using a smart phone's accelerometer with hierarchical classification. In *2010 sixth international conference on intelligent environments*, pages 158–163. IEEE, 2010.
- [15] B Hu, S Li, Yuhao Chen, R Kavi, and Sarah Coppola. Applying deep neural networks and inertial measurement unit in recognizing irregular walking differences in the real world. *Applied Ergonomics*, 96:103414, 2021.
- [16] Arshad Sher, Megan Taylor Bunker, and Otay Akanyeti. Towards personalized environment-aware outdoor gait analysis using a smartphone. *Expert Systems*, page e13130, 2022.
- [17] PC Dixon, KH Schütte, B Vanwanseele, JV Jacobs, JT Dennerlein, JM Schiffman, PA Fournier, and B Hu. Machine learning algorithms can classify outdoor terrain types during running using accelerometry data. *Gait & posture*, 74:176–181, 2019.
- [18] Nizam Uddin Ahamed, Dylan Kobsar, Lauren Benson, Christian Clermont, Russell Kohrs, Sean T Osis, and Reed Ferber. Using wearable sensors to classify subject-specific running biomechanical gait patterns based on changes in environmental weather conditions. *PLoS One*, 13(9):e0203839, 2018.
- [19] Matthew D Czech, Dimitrios Psaltos, Hao Zhang, Tomasz Adamusiak, Monica Calicchio, Amey Kelekar, Andrew Messere, Koene RA Van Dijk, Vesper Ramos, Charmaine Demanuele, et al. Age and environment-related differences in gait in healthy adults using wearables. *NPJ digital medicine*, 3(1):127, 2020.
- [20] Daniel B Kowalsky, John R Rebula, Lauro V Ojeda, Peter G Adamczyk, and Arthur D Kuo. Human walking in the real world: Interactions between terrain type, gait parameters, and energy expenditure. *PLoS one*, 16(1):e0228682, 2021.

- [21] Patrick Ippersiel, V Shah, and PC Dixon. The impact of outdoor walking surfaces on lower-limb coordination and variability during gait in healthy adults. *Gait & Posture*, 91:7–13, 2022.
- [22] Daniel Hamacher, NB Singh, Jaap H Van Dieën, MO Heller, and William R Taylor. Kinematic measures for assessing gait stability in elderly individuals: a systematic review. *Journal of The Royal Society Interface*, 8(65):1682–1698, 2011.
- [23] R Senden, K Meijer, IC Heyligers, HHCM Savelberg, and B Grimm. Importance of correcting for individual differences in the clinical diagnosis of gait disorders. *Physiotherapy*, 98(4):320–324, 2012.
- [24] Louis Atallah, Benny Lo, Rachel King, and Guang-Zhong Yang. Sensor positioning for activity recognition using wearable accelerometers. *IEEE transactions on biomedical circuits and systems*, 5(4):320–329, 2011.
- [25] Wei-Chun Hsu, Tommy Sugiarto, Yi-Jia Lin, Fu-Chi Yang, Zheng-Yi Lin, Chi-Tien Sun, Chun-Lung Hsu, and Kuan-Nien Chou. Multiple-wearable-sensor-based gait classification and analysis in patients with neurological disorders. *Sensors*, 18(10):3397, 2018.
- [26] Aftab Khan, Nils Hammerla, Sebastian Mellor, and Thomas Plötz. Optimising sampling rates for accelerometer-based human activity recognition. *Pattern Recognition Letters*, 73:33–40, 2016.
- [27] Eric K Allseits, Vibhor Agrawal, Abhishek Prasad, Christopher Bennett, and Kyoung Jae Kim. Characterizing the impact of sampling rate and filter design on the morphology of lower limb angular velocities. *IEEE sensors journal*, 19(11):4115–4122, 2019.
- [28] Mingjing Yang, Huiru Zheng, Haiying Wang, Sally McClean, and Nigel Harris. Assessing the utility of smart mobile phones in gait pattern analysis. *Health and Technology*, 2(1):81–88, 2012.
- [29] Weijun Tao, Tao Liu, Rencheng Zheng, and Hutian Feng. Gait analysis using wearable sensors. *Sensors*, 12(2):2255–2283, 2012.
- [30] Marco Avvenuti, Nicola Carbonaro, Mario GCA Cimino, Guglielmo Cola, Alessandro Tognetti, and Gigliola Vaglini. Smart shoe-assisted evaluation of using a single trunk/pocket-worn accelerometer to detect gait phases. *Sensors*, 18(11):3811, 2018.
- [31] Brad Manor, Wanting Yu, Hao Zhu, Rachel Harrison, On-Yee Lo, Lewis Lipsitz, Thomas Trivison, Alvaro Pascual-Leone, and Junhong Zhou. Smartphone app-based assessment of gait during normal and dual-task walking: demonstration of validity and reliability. *JMIR mHealth and uHealth*, 6(1):e8815, 2018.
- [32] Rosa Altillio, Andrea Rossetti, Qiang Fang, Xudong Gu, and Massimo Panella. A comparison of machine learning classifiers for smartphone-based gait analysis. *Medical & Biological Engineering & Computing*, 59(3):535–546, 2021.
- [33] Aner Weiss, Sarvi Sharifi, Meir Plotnik, Jeroen PP van Vugt, Nir Giladi, and Jeffrey M Hausdorff. Toward automated, at-home assessment of mobility among patients with parkinson disease, using a body-worn accelerometer. *Neurorehabilitation and neural repair*, 25(9):810–818, 2011.
- [34] Silvia Del Din, Alan Godfrey, and Lynn Rochester. Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and parkinson’s disease: toward clinical and at home use. *IEEE journal of biomedical and health informatics*, 20(3):838–847, 2015.
- [35] Yue Luo, Sarah M Coppola, Philippe C Dixon, Song Li, Jack T Dennerlein, and Boyi Hu. A database of human gait performance on irregular and uneven surfaces collected by wearable sensors. *Scientific data*, 7(1):1–9, 2020.
- [36] Tal Iluz, Eran Gazit, Talia Herman, Eliot Sprecher, Marina Brozgol, Nir Giladi, Anat Mirelman, and Jeffrey M Hausdorff. Automated detection of missteps during community ambulation in patients with parkinson’s disease: a new approach for quantifying fall risk in the community setting. *Journal of neuroengineering and rehabilitation*, 11(1):1–9, 2014.
- [37] Shu Nishiguchi, Minoru Yamada, Koutatsu Nagai, Shuhei Mori, Yuu Kajiwara, Takuya Sonoda, Kazuya Yoshimura, Hiroyuki Yoshitomi, Hiromu Ito, Kazuya Okamoto, et al. Reliability and validity of gait analysis by android-based smartphone. *Telemedicine and e-Health*, 18(4):292–296, 2012.
- [38] Arshad Sher, David Langford, Einar Dogger, Luke Lunn, Maresa Schröder, Azam Hamidinekoo, Marco Arkesteijn, Qiang Shen, Thomas Jansen, Reyer Zwiggelaar, Helen Tench, Sarah Jones, Phil Jones, Villagra Povina, Federico, and Otar Akanyeti. Automatic gait analysis during steady and unsteady walking using a smartphone. Unpublished paper, 2021.
- [39] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.
- [40] Patima Silsupadol, Paphawee Prupetkaew, Teerawat Kamnardsiri, and Vipul Lugade. Smartphone-based assessment of gait during straight walking, turning, and walking speed modulation in laboratory and free-living environments. *IEEE journal of biomedical and health informatics*, 24(4):1188–1195, 2019.
- [41] Runtong Zhong, Pei-Luen Patrick Rau, et al. A mobile phone-based gait assessment app for the elderly: Development and evaluation. *JMIR mHealth and uHealth*, 8(5):e14453, 2020.
- [42] Christopher Buckley, M Encarna Micó-Amigo, Michael Dunne-Willows, Alan Godfrey, Aodhán Hickey, Sue Lord, Lynn Rochester, Silvia Del Din, and Sarah A Moore. Gait asymmetry post-stroke: Determining valid and reliable methods using a single accelerometer located on the trunk. *Sensors*, 20(1):37, 2020.
- [43] Jeong-Woo Seo, Seul-Gee Kim, Joong Il Kim, Boncho Ku, Kahye Kim, Sangkwan Lee, and Jaeuk U Kim. Principal characteristics of affected and unaffected side trunk movement and gait event parameters during hemiplegic stroke gait with imu sensor. *Sensors*, 20(24):7338, 2020.
- [44] Sakorn Mekruksavanich and Anuchit Jitpattanakul. Smartwatch-based human activity recognition using hybrid lstm network. In *2020 IEEE SENSORS*, pages 1–4. IEEE, 2020.
- [45] Jeffrey Montes, Richard Tandy, John Young, Szu-Ping Lee, and James W Navalta. Step count reliability and validity of five wearable technology devices while walking and jogging in both a free motion setting and on a treadmill. *International journal of exercise science*, 13(7):410, 2020.
- [46] Heydar Sadeghi, Paul Allard, François Prince, and Hubert Labelle. Symmetry and limb dominance in able-bodied gait: a review. *Gait & posture*, 12(1):34–45, 2000.
- [47] Byungjoo Noh, Changhong Youm, Eunkyoung Goh, Myeounggon Lee, Hwayoung Park, Hyojeong Jeon, and Oh Yoen Kim. Xgboost based machine learning approach to predict the risk of fall in older adults using gait outcomes. *Scientific reports*,

- 11(1):12183, 2021.
- [48] James Alexander Hughes, Sheridan Houghten, and Joseph Alexander Brown. Models of parkinson’s disease patient gait. *IEEE journal of biomedical and health informatics*, 24(11):3103–3110, 2019.
 - [49] Issam Boukhenoufa, Xiaojun Zhai, Victor Utti, Jo Jackson, and Klaus D McDonald-Maier. Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review. *Biomedical Signal Processing and Control*, 71:103197, 2022.
 - [50] Benish Fida, Ivan Bernabucci, Daniele Bibbo, Silvia Conforto, and Maurizio Schmid. Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer. *Medical engineering & physics*, 37(7):705–711, 2015.
 - [51] Ankita Jain and Vivek Kanhangad. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, 18(3):1169–1177, 2017.
 - [52] Felix E Zajac, Richard R Neptune, and Steven A Kautz. Biomechanics and muscle coordination of human walking: Part i: Introduction to concepts, power transfer, dynamics and simulations. *Gait & posture*, 16(3):215–232, 2002.
 - [53] Felix E Zajac, Richard R Neptune, and Steven A Kautz. Biomechanics and muscle coordination of human walking: part ii: lessons from dynamical simulations and clinical implications. *Gait & posture*, 17(1):1–17, 2003.
 - [54] Jana R Montgomery and Alena M Grabowski. The contributions of ankle, knee and hip joint work to individual leg work change during uphill and downhill walking over a range of speeds. *Royal Society open science*, 5(8):180550, 2018.
 - [55] Javad Sarvestan, Peyman Aghaie Ataabadi, Fateme Yazdanbakhsh, Shahram Abbasi, Ali Abbasi, and Zdeněk Svoboda. Lower limb joint angles and their variability during uphill walking. *Gait & Posture*, 90:434–440, 2021.
 - [56] D Burtan, K Joyce, JF Burn, TC Handy, S Ho, and U Leonards. The nature effect in motion: visual exposure to environmental scenes impacts cognitive load and human gait kinematics. *Royal Society open science*, 8(1):201100, 2021.
 - [57] Renee Veldkamp, Mieke Goetschalckx, Hanneke E Hulst, Alice Nieuwboer, Kathleen Grieten, Ilse Baert, Carmela Leone, Lousin Moumdjian, and Peter Feys. Cognitive–motor interference in individuals with a neurologic disorder: a systematic review of neural correlates. *Cognitive and Behavioral Neurology*, 34(2):79–95, 2021.
 - [58] Rui Moura Coelho, Joao Gouveia, Miguel Ayala Botto, Hermano Igo Krebs, and Jorge Martins. Real-time walking gait terrain classification from foot-mounted inertial measurement unit using convolutional long short-term memory neural network. *Expert Systems with Applications*, 203:117306, 2022.
 - [59] Mina Nouredanesh, Alan Godfrey, Dylan Powell, and James Tung. Egocentric vision-based detection of surfaces: towards context-aware free-living digital biomarkers for gait and fall risk assessment. *Journal of neuroengineering and rehabilitation*, 19(1):79, 2022.