Journal of Biomedical Informatics 131 (2022) 104116

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Improved screening of fall risk using free-living based accelerometer data

D. Kelly [a,*], J. Condell [a], J. Gillespie [a], K. Munoz Esquivel [a], J. Barton [b], S. Tedesco [b], A. Nordstrom [c], M. Åkerlund Larsson [c], A. Alamäki [d]

[a] Ulster University, Northern Ireland, United Kingdom
[b] Tyndall National Institute, University College Cork, Ireland
[c] Umeå University, Sweden
[d] Karelia University of Applied Sciences, Finland

ARTICLE INFO

Keywords:
Fall risk
Accelerometer

ABSTRACT

Falls are one of the most costly population health issues. Screening of older adults for fall risks can allow for earlier interventions and ultimately lead to better outcomes and reduced public health spending. This work proposes a solution to limitations in existing fall screening techniques by utilizing a hip-based accelerometer worn in free-living conditions. The work proposes techniques to extract fall risk features from periods of free-living ambulatory activity. Analysis of the proposed techniques is conducted and compared with existing screening methods using Functional Tests and Lab-based Gait Analysis. 1705 Older Adults from Umea (Sweden) were assessed. Data consisted of 1 Week of hip worn accelerometer data, gait measurements and performance metrics for 3 functional tests. Retrospective and Prospective fall data were also recorded based on the incidence of falls occurring 12 months before and after the study commencing respectively. Machine learning based experiments show accelerometer based measures perform best when predicting falls. Prospective falls had a sensitivity and specificity of 0.61 and 0.66 respectively while retrospective falls had a sensitivity and specificity of 0.61 and 0.68 respectively.

1. Introduction

Approximately 28–35% of people aged 65 and over fall each year increasing to 32–42% for those over 70 years of age [1]. Falls can have a number of negative effects on fall victims resulting in decreased quality of life due to reduced activities of daily living, physical deterioration and social isolation and death. In 2017, in the Western European region, 8.4 million adults age 70 and older sought medical attention due to a fall, with case fatality rate between 0.4% and 1.1% reported [2].

Among the most serious injuries resulting from falls are hip fractures and traumatic brain injury. Falls also have a significant economic burden on national health care services with falls costing between 0.85% and 1.5% of the total health care expenditures [3].

Accurate assessment of fall risk can allow for earlier fall-risk reduction interventions such as physical therapy, home modification and medication withdrawal. Early interventions have potential to reduce fall occurrence, fall-related costs, fear of falling and negative effects on quality of life [4]. Clinical fall risk assessment tools often utilize questionnaires and/or functional assessments of posture, gait, cognition and other risk factors. However, clinical assessments have a number of limitations due to the subjective and qualitative nature of the assessment methodologies [5].

It has been recommended that screening of fall risk should be conducted for older adults at least annually by physicians [6]. However, effective fall risk assessment remains underutilized in clinical practice due to unreliable subjective measures, lack of cost-effective technology and clinical time constraints [7]. In order for fall risk screening to be practically integrated into typical clinical practice, fall risk assessment techniques must be developed that meet 3 key criteria: (1) accurate assessment of fall risk, (2) use of inexpensive technology and (3) easy to administer [7].

As an alternative to clinical assessment tools, technology based fall risk assessment tools have been proposed in the literature in order to provide more objective and quantitative measures of fall risk. Inertial sensors are the most commonly used sensor for fall risk assessment and recent reviews indicate that inertial sensors have the potential to provide quantitative, objective and reliable indications of fall risk [8,5,9]. This paper therefore aims to utilize a hip worn accelerometer in free-

* Corresponding author.
E-mail address: d.kelly@ulster.ac.uk (D. Kelly).

https://doi.org/10.1016/j.jbi.2022.104116
Received 24 November 2021; Received in revised form 7 April 2022; Accepted 5 June 2022
Available online 8 June 2022

living conditions to assess ambulatory activity and identify future fallers. In this work, free-living conditions refers to conditions whereby participants continue their normal daily life routines.

1.1. Related work

Several reviews have been published in recent years related to Sensor-based Falls Risk Testing (SFRT) [8,10–12,7,13,5]. An analysis of these reviews identified 3 key issues that future research should focus on. These issues relate to 1) fall classification criterion 2) data acquisition methodology and 3) validation protocols used. Each issue was raised in at least 2 of the recent review articles.

1.1.1. Issue 1: fall classification criterion

The classification criterion is the baseline measure used to compare a proposed fall risk measurement technique against. The method used to classify fall risk differs among the literature. Participants are commonly classified as fallers or non-fallers based on one of several different methods including Clinical Assessment, Prospective Falls and Retrospective Falls [8]:

- Clinical Assessment: A person is classified as a faller or non-faller based on their performance of an assessment, or set of assessments, in clinical settings such as the Timed Up and Go (TUG) test.
- Retrospective Falls: A person is classified as a faller or non-faller based on self-reported fall history denoting the presence or absence of fall occurrences in the past.
- Prospective Falls: A person is classified as a faller or non-faller based on self-reported fall occurrence within a follow-up period from the assessment (commonly 1 year).

Clinical assessments use functional tests such as the Tinetti Test [14], Grip Strength [15] or Timed Up and Go (TUG) [16]. However, there is conflicting evidence supporting functional tests in predicting future falls. For example, two different studies by, Kojima et al. [17] and Moller et al. [18], report significantly different sensitivity and specificity results. Both studies assessed the predictive validity of TUG as a predictor of future falls (6 months follow up) in community dwelling older adults (+65 years). Kojima et al. and Moller et al. reported a similar optimal TUG cut-off point of 12.6 and 13 s respectively. However, while Kojima et al. reported a sensitivity and specificity of 0.305 and 0.895 respectively, Moller et al. reported a sensitivity and specificity of 0.667 and 0.5 respectively. Thus, sensitivity and specificity differed by 36% and 39.5% respectively.

Basing the criterion measure of a study on clinical assessment tests will introduce false positives and false negatives into the ground truth classification criterion [5]. A false positive occurs when a participant performs a clinical assessment with a score that meets the fall risk threshold criteria but subsequently does experience a fall. Conversely, a false negative occurs when a participant performs a clinical assessment with a score that does not meet the fall risk threshold criteria but subsequently does experience a fall.

Retrospective falls act as a proxy measure for fall risk since it is known that a faller has a higher risk of falling again [19]. However, this method requires that a fall has already occurred in order to identify a person as at risk and, similar to clinical assessments, will introduce false negatives and false positives into the ground truth. Thus, when comparing a proposed SFRT technique with clinical assessment or retrospective ground truth, evaluations will not reflect true performance of future faller and non-faller classification.

Recent reviews have stressed the problematic fact that Clinical fall risk assessment and Retrospective falls are the two most commonly used criterion measures in the literature [13]5. Since the goal of fall risk assessment is to predict the likelihood of future falls, prospective falls is the preferred criterion. However, only a small number of studies have employed prospective falls as the criterion method.

By conducting a detailed literature review of recent SFRT papers, aided by two review papers [5,7], we identified a total of 15 papers that utilized prospective falls for SFRT [20–34].

1.1.2. Issue 2: free-living/community based data acquisition

Previous work has shown that fall risk predictors can be extracted from accelerometer signals recorded during periods of steady state walking [35]. For example, Hua et al. showed that features extracted from vertical acceleration signals had good discriminatory power in separating high risk fallers from low risk fallers.

However, the study by Hua et al. captures data during the performance of standardised gait test in a controlled lab setting. Capture of data in controlled lab/clinic conditions is the most commonly used approach in the literature for SFRT [10,7,11]. However, this approach has a number of disadvantages. Firstly, there are potential issues relating to participants' awareness of being observed (Hawthorne Effect) during the performance of gait and balance tests in controlled settings. Research has shown that gait performance can differ when participants are being observed [10,36]. Thus, gait based measures in lab/clinic settings may not reflect naturalistic behaviour. Secondly, controlled lab-based measures require costly staff time to administer the test and often use expensive sensor equipment.

As an alternative, gait based measures can be extracted from free-living behaviour where sensors are worn by participants in their natural daily living environments [10]. Participants are thus not required to attend specialist clinics or lab settings and participants do not need to be monitored or supervised by health care professionals. Furthermore, behaviour measured in free-living conditions should be more representative of natural behaviour. It has been shown that accelerometers worn in free-living conditions have the potential to identify fallers [10].

Of the 15 previously described prospective falls based studies, 2 were focused on free-living based data acquisition [32,33] while the remaining were based on lab based measures. The 2 studies were based on the same data set of 319 older adults. Van Schooten et al. [32] reported Area Under the Curve (AUC) performance of 0.66–0.72 while Aicha et al. [33] report AUC performance of 0.61–0.7. Cross Validation (CV) was used to measure performance. However, it was difficult to assess if performance would be maintained in real-world conditions due to some ambiguities in the evaluation protocols which we discuss in the next section.

1.1.3. Issue 3: modeling and validation

Recent research has highlighted some troubling trends in relation to the presentation of over-optimistic SRFT results [13]. In particular, concerns have been raised in relation to sample size, questionable modeling and problematic validation methodologies. One of the biggest challenges in SFRT is acquiring a large enough sample size to ensure sufficient study power. This is particularly challenging and costly for prospective based falls studies where there is a requirement for a 6–12 month follow up with each participant. However, from the set of 15 reviewed prospective falls papers, we found that the largest number of participants used was 319 and the average number of participants used was 127($\pm$86). Most studies have therefore been too small to gain any real statistical insight into the effectiveness of techniques if applied to a larger population.

Another issue relates to the misuse of model validation methodologies. A fundamental component of machine learning is that one should separate the problem of model selection from that of evaluating the final performance of the predictor. To evaluate performance, it is important to set aside an independent test set referred to as a "holdout" test set [37]. The remaining data should be used both for training and performing model selection.

A problem identified by Shany et al. was that testing data is commonly used in some of the model training pipeline steps, such as feature selection, model selection, parameter tuning [13]. Use of test data for model selection can lead to models that are biased towards the

available data and thus can produce models that are over-trained and produce inflated accuracy scores that are unlikely to maintain their reported performance during real-world use. In order to build an unbiased model, it is vital that testing data should never be used to inform the feature selection, model selection or parameter tuning.

While it was difficult to ascertain the exact model training methodology used in the 15 previously described prospective studies due to ambiguous descriptions in some papers, we identified 7 papers which appear to perform feature selection on training and testing data [22,20,28,25,30,31,34] and 2 papers which appear to perform model selection using training and testing data [26,34].

Research also highlights issues with the use of CV as the primary evaluation technique [38]. Specifically, there is no way to know how reliable CV based performance measures are. CV has been shown to have a large uncertainty for small sample sizes common in health and medical research. It is therefore advised, for small datasets with $N < 1000$, that performance be reported using a single holdout test set.

In the 15 previously described prospective falls based studies, we identified 1 study that utilized a holdout test set [29]. Nine studies utilized CV (or a variation of CV) as the primary means of evaluating the accuracy of predictions [23,22,25,30,32,33,26,34]. The remaining 6 studies used resubstitution where the model is trained on all data and then tested by resubstituting the same data as the test data [21,20,24,28,31,27]. Resubsitution clearly has a number of issues and will result in performance estimates that will significantly overestimate performance.

2. Methods

2.1. Data acquisition

1705 Participants, all aged exactly 70 years old and from Umeå Sweden, took part in the study (817 Female and 888 Male). Participants had an average weight of 76.9 kg (±14.1 Kg) and an average Body Mass Index of 26.5 (±4.08).

All participants attended an examination session conducted by a research nurse. During the examination session, participants were asked to perform a set of standardized functional tests; TUG, Gait Velocity during a 6 Meter Walk Test and Non-Dominant Hand Grip Strength. Participants also performed a gait assessment on a pressure sensitive walkway (GAITRite, CIR Systems Inc, USA). Participants were asked to walk the length of the walkway (6 Meters) and 65 gait based measures were calculated for each participant using proprietary software (GAITRite). Gait measures include Step Time, Cycle Time, Step Length, Heel to Heel Base Support Distance, Single Leg Support Time, Double Leg Support Time, Swing Time, Stance Time, Step Extremity Ratio and Toe In/out angle.

History of falls in the 12 months prior to the study commencing was also recorded based on participant recall. A fall was defined as an event which results in a person coming to rest inadvertently on the ground or floor or other lower level.

After the examination session, participants were provided with a hip mounted tri-axial accelerometer (GT9X Actigraph,Actigraph LLC, USA) which they were asked to wear for 7 consecutive days. Acceleration for x, y and z axis were recorded for the duration of the 7 days at 30 Hz. A 7 day duration was chosen based on likelihood of that period representing the participants normal gait, while also considering logistics, availability of devices and ethics. At the end of the 7 days, the device was returned by the participant and acceleration data was retrieved for each participant (average of 911 MB per participant). Six and twelve months after the examination session, follow-up telephone interviews were conducted to ask whether participants have experienced a fall since their examination session. It is worth noting that some participants may experience a fall in the previous 12 months and also experience a fall in the 12 month follow up period.

2.2. Ethical considerations

Prior to participating, all participants gave oral and written consent. The study follows the ethical principles of the Declaration of Helsinki and has been approved by the Regional Ethics Review Board in Umea (dnr 2012-85-32-M, supplement to dnr 07-031-M).

2.3. Free-living accelerometer data processing

The majority of previous SFRT research has performed data collection in controlled lab settings with participants performing a set of standardised physical activities. Switching to data collection in unsupervised free-living conditions would make the assessment procedure more accessible for patients, less expensive to administer and would overall increase the feasibility of deploying the screening tool for all older adults.

Previous work has shown promising results where features based on gait quality, extracted from free-living conditions, could be used as predictors of fall risk [32,33]. The technical novelty of this work is that it builds on this premise to calculate gait quality based features from accelerometer data retrieved from the tri-axial accelerometer worn by participants. Prior to extracting gait based measures from free-living based data, automatic detection of periods of steady state ambulatory activity is first performed.

2.3.1. Signal processing

Signal processing and data analysis was performed using the Python programming language and the Python libraries: SciPy, Pandas, SciKitLearn, AutoSKLearn and TSfresh.

With the accelerometer worn around the waist, the aim was that participants would wear the device such that the accelerometer axis aligned with anatomical axis, where the accelerometer x, y and z axis aligned with sagittal, longitudinal and frontal axis respectively. Fig. 1 provides an illustration of the accelerometer and anatomical axis. It was observed that participants consistently aligned the accelerometer y axis with the anatomical longitudinal axis due to the constraints imposed by the sensor belt mounting mechanism. However, due to the potential to mount the sensor at any position on the waist between the left hip and right hip, alignment of the sensor x and z axis with the anatomical frontal and sagittal axis was performed inconsistently by participants. The x and z axis were therefore combined into a single horizontal acceleration magnitude $A_{horiz} = \sqrt{A_x^2 + A_z^2}$ to ensure measurements were consistent across all participants. Overall acceleration magnitude is also calculated $A_{mag} = \sqrt{A_x^2 + A_y^2 + A_z^2}$.

2.3.2. Ambulatory activity detection

Prior to extracting gait features, periods of ambulatory activity were automatically identified from a filtered accelerometer signal. Candidate steps were first identified by performing peak detection on a vertical acceleration signal $A_y^{'}$ filtered using a 4th order Butterworth bandpass filter (0.25–2.5 Hz). All identified peaks were evaluated and defined as a candidate step only if the signal had a zero-crossing and crossed both a positive and negative threshold of ±0.8 on either side of the zero-crossing. All candidate steps were then grouped into clusters based on temporal proximity to one another.

Clusters of candidate steps that met a set of pre-defined criteria were classified as periods of ambulatory activity. While it is important that a large number of periods of ambulatory activity are correctly identified (i.e. true-positives), it is more important that other periods of activity, that are not ambulatory in nature, are not included for further analysis (i.e. false-positives). Previous work indicates that fall predictors can be extracted from steady state walking patterns [35]. The criteria was therefore designed to detect steady state ambulatory activity to minimize the number of false positive periods of ambulatory activity

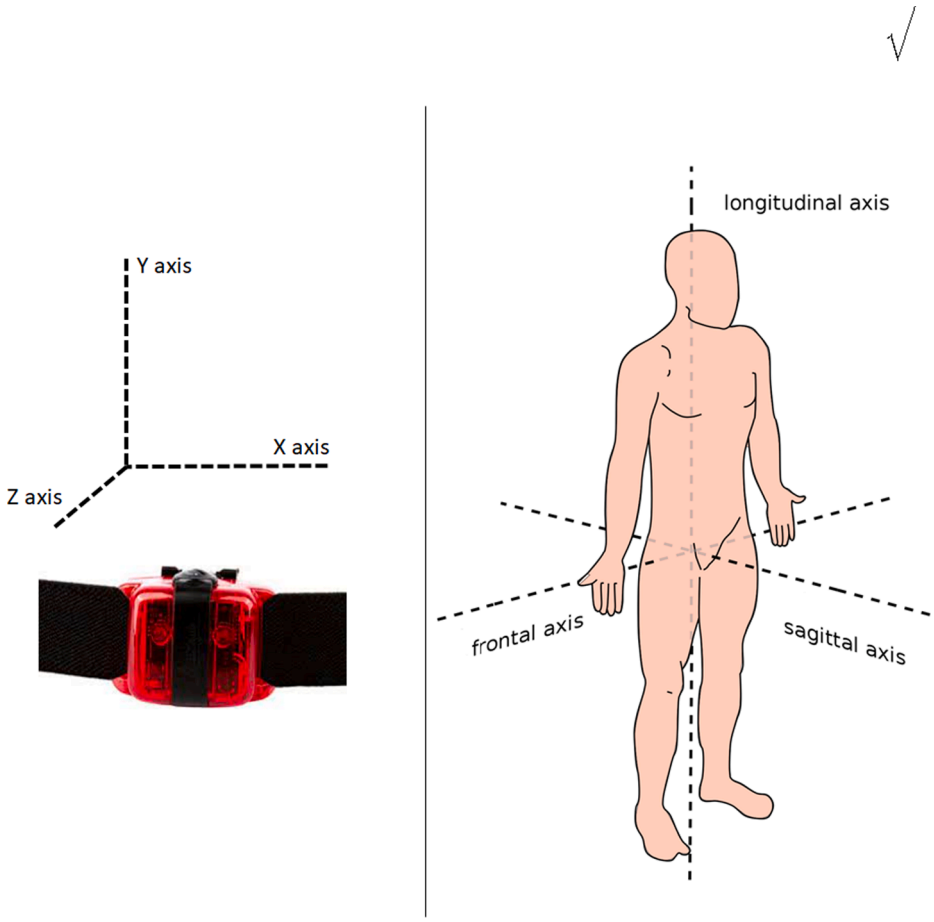

Fig. 1. (Left) Actigraph GT9X Axis Configuration (Right) Anatomical Axis.

detected. The criteria implemented is defined as follows:

- Time between each step peak should be between 0.2 and 3 s
- Standard Deviation of time between all step peaks should be within ±0.8 seconds
- There should be a minimum of 25 steps within the cluster.

2.3.3. Feature extraction

For each period of detected ambulatory activity, raw accelerometer data for that period was processed and features were extracted. A 4th order butterworth bandpass filter (0.05–3.0 Hz) was applied to the three raw accelerometer signals, A_y, A_{horiz} and A_{mag}. A single feature vector F was computed for each participant from the median of all feature vectors f_i, where f_i is the feature vector computed from the i_{th} bout of detected ambulatory activity. Each feature vector consists of a step based feature vector s_i and a frequency based feature vector p_i such that $f_i = \{s_i, p_i\}$.

The step based feature vector s_i, is comprised of a total of 180 statistical based features computed for each filtered signal $(A_y, A_{horiz}, A_{mag})$. A time series feature extraction library, TSfresh, was utilized to extract 60 'simple' features for each filtered signal [39]. Step features include maximum, minimum, index of maximum, index of minimum, variance, signal mass center, number of peaks, absolute eneregy, auto-correlation mean and sample entropy. Step features are computed for each step, within the i_{th} bout of ambulatory activity, between the local minima preceding and succeeding the step peak. This was performed for all peaks within the i_{th} bout of ambulatory activity and the overall step feature s_i was calculated as the median of all step features within the i_{th} bout of ambulatory activity.

The frequency based feature vector p_i, is comprised of a total of 500 Fast Fourier Transform (FFT) and Wavelet features which are calculated from a fixed size 10 s window within the bout of ambulatory activity. For each bout of ambulatory activity, FFT Phase and Continuous Wavelet Transform (CWT) coefficients are calculated for the 10 s A_{mag} signal. The feature vector p_i, is therefore comprised of 150 FFT phase coefficients and 350 CWT coefficients, where CWT is calculated for 3 different window lengths (3, 5 and 10).

For bouts of ambulatory activity that last longer than 10 s, a sliding window approach was used to find a 10 s period that minimizes the standard deviation of time between step peaks. Thus, the goal was to identify a 10 s period of ambulatory activity with the most consistent step cadence in order to calculate FFT and Wavelet based features from. Fig. 2 provides a visualisation of candidate steps and how periods of ambulatory activity are identified. Fig. 2 also shows an example of a 10 s window being select for FFT and Wavelet coefficients to be extracted from.

2.4. Model training

As discussed in Section 1.1.3, model training should be considered as a pipeline of steps comprising feature pre-processing, feature selection, training, parameter tuning and model selection. Model training, and any of the individual pipelines steps, will only be performed on the training set.

The three data types (FT, PSW, FLA) are available for all 1705 participants. Three categories of prediction models will therefore be trained and evaluated to directly compare the prediction performance of the 3 different data types. Fall History, defined as a fall occurring in the 12 months prior to the study, will also be assessed as predictor of future falls and as a feature to complement each of the three data types features.

The scope of this paper is not to develop or propose novel machine

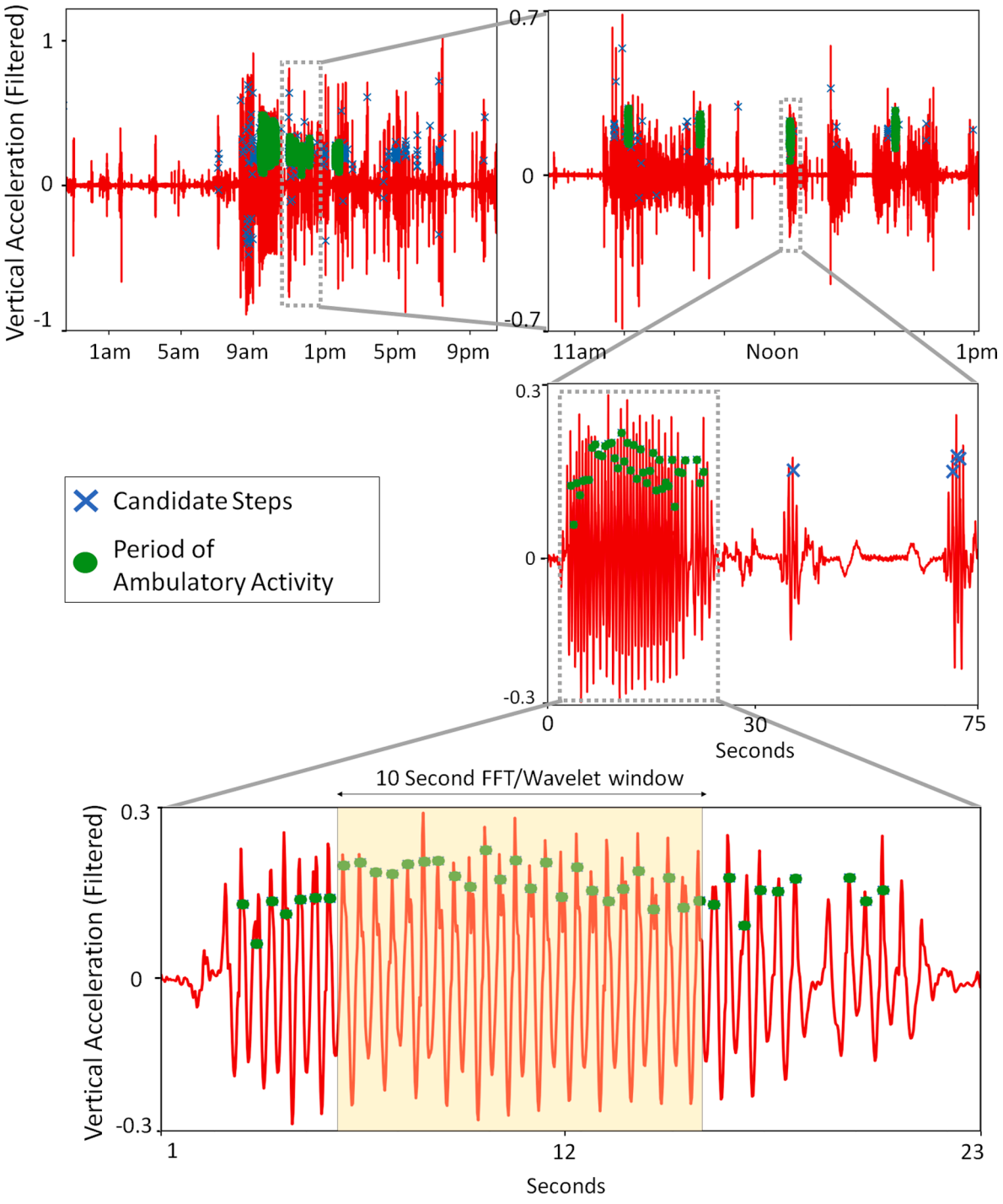

Fig. 2. Filtered Accelerometer Signal for 1 Day (with multiple zoom levels) showing detection of candidate steps and periods of ambulatory activity.

learning models or configurations. The aim is, however, to evaluate the proposed FLA features, and compare them to more commonly used FT and PSW features, in predicting prospective falls. In order to build predictive models, consideration needs to be given to the type of feature preprocessing and learning algorithms that will be used. In addition, appropriate hyper-parameters need to be set for the chosen algorithms. To remove any potential bias that could be introduced, by choosing preprocessing techniques, learning algorithms or hyperparamers that favours one of the 3 data types over another, a systematic and objective methodology to select algorithms and hyperparamaters is implemented. This method uses an automated machine learning methodology based on Bayesian optimization methods to select from 15 classifiers and 14 feature preprocessing algorithms [40]. The 15 classifiers were: Decision Tree, Adaboost Decision Tree, Extra Trees, Random Forest, Gradient Boosting, Bernoulli Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, K-nearest neighbour, Linear Discriminant Analysis, Linear Support Vector Machine (SVM), Radial Basic Function SVM, Neural Network, Stochastic Gradient Descent classifier and Quadratic Discriminant Analysis. The 14 pre-processing algorithms were: Independent Component Analysis, Principle Component Analysis (PCA), Kernel PCA, Nystroem Sampler, densifier, feature agglomeration, feature selection (ANOVA), feature selection (chi2), feature selection (Extra-Trees), feature selection (random trees), feature selection (SVM), polynomial features, truncated SVD and no pre-processing.

The final prediction model, including algorithms, models, features and hyperparamaters, is selected based on performance calculated from 10-fold cross-validation on the training set. Overall model performance of the final prediction model is evaluated using the holdout test.

For each of the three data types (FT,PSW, FLA) the automated Bayesian Optimization based machine learning methodology was utilized to configure a machine learning pipeline and train models [40]. Models were configured and trained using training set data only. The training set was split into a cross-validation set (N = 1087, 163 Fallers) and a validation set (N = 192, 28 Fallers). The Bayesian Optimization system was implemented to maximize the average g-mean, over 10 folds, computed from the cross-validation set. Using early stopping, the validation set was utilized to reduce the models over-fitting on the cross-validation set. After each epoch of the Bayesian Optimization process, g-mean was calculated for the cross-validation set and the validation set. Early stopping of the Bayesian Optimization process was performed when validation set performance began to diverge from cross-validation set performance.

3. Results

Research by Haagsma et al. [2] reported a fall incidence rate of 14,835 per 100,000 in Sweden in 2017 and an average fall incidence rate of 13,979 (±3706) per 100,000 for 22 Western European Countries. The fall incidence rate for participants in this study was comparable to national and international rates with an incidence rate of 14,956 per 100,000. Of the 1705 participants in this study, 255 participants reported at least one fall within 12 months after the study commenced with 16.4% (n = 134) of females and 13.6% (n = 121) of males reporting a fall. Of the 525 participants reported having a fall prior to the study commencing. Of the 525 participants reporting a fall in the 12 months prior to the study, only 19% (n = 98) of those reported a fall in the 12 month follow up period. The remaining (n = 151) prospective falls were from 'new fallers' who did not report a fall in the 12 months prior to the study commencing.

Experiments were conducted to address the 3 issues discussed in the previous section. A novel risk assessment technique, based on free-living accelerometer data is proposed. This technique is evaluated and compared with 2 other commonly used risk assessment techniques.

Results show the classification capability of 3 different data types in predicting prospective and retrospective falls:

- (FT) **F**unctional **T**est scores from Grip Strength, TUG tests, Gait Velocity
- (PSW) Gait measures from supervised lab-based **P**ressure **S**ensitive **W**alkway
- (FLA) Proposed gait measures from **F**ree-**L**iving **A**ccelerometer data.

A holdout test set of participants, stratified for occurrences of prospective and retrospective falls, was created using 25% of the data (N = 428: 64 Prospective Fallers, 130 Retrospective Fallers). A training set of participants was created using remaining participants, not in the holdout test set (N = 1279: 191 Prospective Fallers, 393 Retrospective Fallers).

Sections 3.1 and 3.2 discuss the results of statistical analysis and machine learning experiments respectively.

3.1. Statistical analysis

3.1.1. Functional tests

This section presents the results of statistical univariate analysis performed on variables in the training set only (N = 1279). Timed up and Go (TUG), Gait Velocity (GV) and Grip Strength (GS) are all common functional tests used in the literature to assess falls risk. Table 1 shows the mean and standard deviation of scores from the 3 different functional tests for prospective and retrospective fallers compared with prospective and retrospective non-fallers respectively and p values calculated using two-sample t-tests. Distributions are compared for all participants as well as for male and female specific sub-groups.

Results show that GV and GS scores were significantly different for prospective fallers compared to non-fallers while TUG times was not. Results did show a significant difference for all three functional tests for retrospective fallers compared to non-fallers. None of the other gender specific distributions were significantly different when assessing prospective falls.

Using Receiver Operating Characteristics (ROC) curves, we evaluate different threshold values for each of the 3 functional tests to assess discriminative ability. Table 2 shows results of ROC analysis with Area Under the Curve (AUC) for each of the functional tests shown for Prospective and Retrospective Falls. Sensitivity and Specificity are also shown in Table 2 for threshold values that achieved the maximum g-mean where g-mean is defined as $\sqrt{Sensitivity \times Specificity}$. G-mean is implemented in this work as a balanced singular assessment metric to measure performance on an imbalanced data-set where there is significantly more non-fallers than fallers [41].

Results of ROC analysis indicate that prediction of prospective falls using each of the 3 functional tests independently is no better than random guessing. Similarly, the classification of fall history is similar to random guessing for TUG and GV. However, GS shows moderate discriminative ability in identifying retrospective fallers with an AUC of 0.604 ***(see Fig. 3).

3.1.2. Lab gait measures

Gait measures from a lab-based pressure sensitive walkway were also recorded for all participants (GAITRite, CIR Systems Inc, USA). A two-sample t-test was performed on each of the 65 gait parameters extrac-

Table 1
Comparison of functional test distributions for fallers and non-fallers.

	Prospective			Retrospective		
	Faller *n = 191*	Non-Faller *n = 1087*	p	Faller *n = 393*	Non-Faller *n = 885*	p
TUG All [s]	9.9(±6.7)	9.8(±3.9)	.39	10.0(±4.1)	9.7(±4.4)	**.04**
TUG Male [s]	10.5(±10.7)	9.8(±3.5)	**.004**	10.0(±3.5)	9.8(±4.6)	.41
TUG Female [s]	9.5(±3.0)	9.8(±4.3)	.13	10.0(±4.6)	9.6(±3.8)	**.02**
GV All [cm/s]	114.8(±359.6)	117.8(±344.7)	**.04**	115.5(±334.5)	118.2(±351.8)	**.02**
GV Male [cm/s]	115.2(±362.9)	118.3(±324.3)	.13	117.0(±255.6)	118.2(±354.7)	.49
GV Female [cm/s]	114.4(±360.2)	117.2(±367.0)	.18	114.4(±388.4)	118.2(±349.1)	**.02**
GS All [kg]	33.1(±114.5)	34.9(±115.5)	**.03**	32.1(±115.5)	35.8(±111.7)	< **.001**
GS Male [kg]	42.8(±56.2)	43.2(±55.8)	.64	42.4(±68.2)	43.4(±51.6)	.11
GS Female [kg]	24.9(±16.2)	25.8(±22.1)	.06	24.8(±22.0.5)	26.1(±20.1)	< **.001**

Table 2
ROC Area Under the Curve (AUC), Specificity and Sensitivity for functional test thresholds

	Prospective				Retrospective			
	ROC AUC	Threshold	Sens	Spec	ROC AUC	Threshold	Sens	Spec
TUG All [s]	0.478	9	0.64	0.36	0.535	9.5	0.53	0.52
TUG Male [s]	0.55	9.5	0.55	0.51	0.526	9.5	0.52	0.52
TUG Female [s]	0.524	9.5	0.53	0.43	0.547	9.5	0.54	0.53
GV All [cm/s]	0.500	120	0.59	0.46	0.533	115	0.46	0.58
GV Male [cm/s]	0.523	120	0.58	0.48	0.504	112.5	0.38	0.63
GV Female [cm/s]	0.48	120	0.59	0.45	0.553	115	0.52	0.58
GS All [kg]	0.501	30	0.44	0.60	0.604	30	0.52	0.648
GS Male [kg]	0.504	44	0.57	0.48	0.540	45	0.63	0.44
GS Female [kg]	0.515	27	0.59	0.43	0.577	26	0.55	0.55

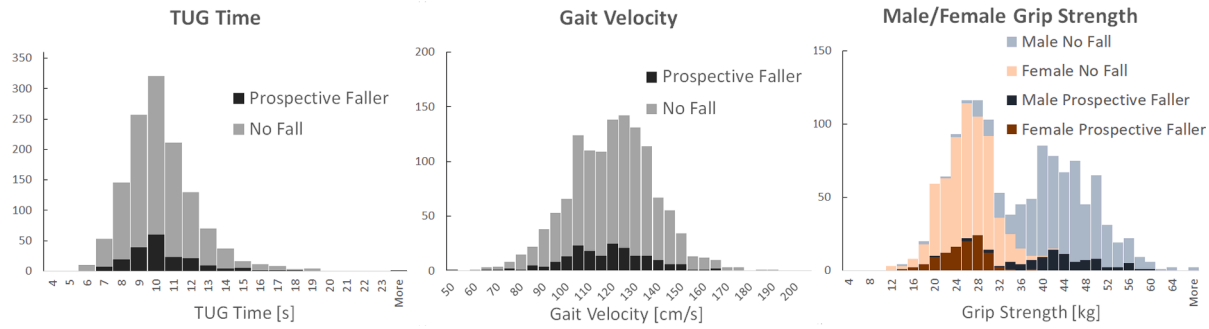

Fig. 3. Physical Function test score distributions for Prospective Falls.

ted by the Gaitrite proprietary software, comparing measurement distributions between fallers and non-fallers. Out of the 65 gait measures, only 2 measures showed a statistically significant difference ($p < 0.05$) between prospective fallers and non-fallers, as shown in Table 3. For retrospective fallers, 5 measures showed a statistically significant difference ($p < 0.05$) between fallers and non-fallers, as shown in Table 4. T-tests based on male and female specific subsets resulted in no statistically significant difference ($p < 0.05$) for any of the 65 measures between prospective fallers and non-fallers or between retrospective fallers and non-fallers.

ROC analysis of the statistically significant measures, as predictors of prospective and retrospective falls, was also conducted. Tables 3 and 4 also show details of ROC AUC, and optimal threshold data, for each of the measures as well as t-test results.

Results show that fallers performed gait activities with a reduced stride velocity compared to non-fallers. Stride Velocity is defined as stride length divided by the stride time, where stride length is the distance between the heel points of two consecutive footprints of the same foot. However, ROC analysis results indicate that gait measures have poor discriminative ability in identifying prospective and retrospective fallers.

3.1.3. *Free-living accelerometer data*

This work proposes improved measurement of fall risk using accelerometer based features comprising descriptors of average step patterns, and movement frequency components, during steady state ambulatory activity. Fig. 4 illustrates the mean step acceleration signals for fallers and non-fallers in the training set. Steps were detected for each participant using methods described in Section 2.3.2. All steps within all detected periods of ambulatory activity were normalized to 1 s using time series interpolation and then averaged to compute a participant specific average step. Mean, standard deviation and 99% Confidence Intervals (CI) were then computed from all participants specific average steps. Analysis of step signals for fallers compared with non-fallers resulted in significant differences in the horizontal acceleration before and after vertical acceleration step peaks.

In addition to step based analysis, Fast Fourier Transform (FFT) signals are computed from 10 s ambulatory activity windows for each participant. FFT summary features are then computed by calculating the median of all ambulatory FFT signals for a given participant. Fig. 5 illustrates the mean FFT Magnitude and Phase signals for fallers versus non-fallers. Mean FFT signals were calculated by averaging the FFT summary features (computed from A_{mag} signal) for all fallers and non-fallers respectively. The FFT magnitude signal shows ambulatory activity is commonly composed of hip movement frequencies of ~1.9hz, ~3.75 Hz and ~5.5 Hz. It can be seen that the average FFT magnitude is almost identical between fallers and non-fallers. However, significant differences were identified in the FFT Phase signal. FFT Phases occurring at ~1.9 Hz and ~5.5 Hz show significant differences between fallers and non-fallers as labeled by regions A and B in Fig. 5. FFT Phase represents how frequency components align in time, therefore, this result likely means that the main frequency component of 3.75 Hz is aligned similarly for fallers and non-fallers. However, the two less prominent frequency components of 1.9 Hz and 5.5 Hz are aligned differently in time

Table 3
Results of T-test and ROC Analsis of Statistically Signifigant Gaitrite Measurements ($p < 0.05$) for Prospective Falls

Gait Measure	Mean (SD) Faller	Mean (SD) Non-faller	p	ROC AUC	Threshold	Sens	Spec
Left Stride Velocity [cm/s]	115.3 (±19.1)	118.4 (±18.6)	0.036	0.543	115	0.48	0.59
Right Stride Velocity [cm/s]	115.3 (±19.1)	118.4 (±18.6)	0.041	0.544	115	0.47	0.59

Table 4
Results of T-test and ROC Analsis of Statistically Signifigant Gaitrite Measurements ($p < 0.05$) for Retrospective Falls.

Gait Measure	Mean (SD) Faller	Mean (SD) Non-faller	p	ROC AUC	Threshold	Sens	Spec
Left Swing [% of Cycle]	38.0 (±2.1)	38.3 (±1.8)	0.02	0.53	38%	0.46	0.60
Left Stance [% of Cycle]	61.9 (±2.1)	61.7 (±1.8)	0.041	0.532	62%	0.46	0.60
Left Stride Velocity [cm/s]	116.1 (±18.4)	118.8 (±18.8)	0.02	0.531	115	0.45	0.59
Right Stride Velocity [cm/s]	116.1 (±18.4)	118.8 (±18.9)	0.02	0.532	115	0.46	0.59
Double Support Time [secs]	.265 (±.056)	.259 (±.052)	0.036	0.531	0.27	0.41	0.65

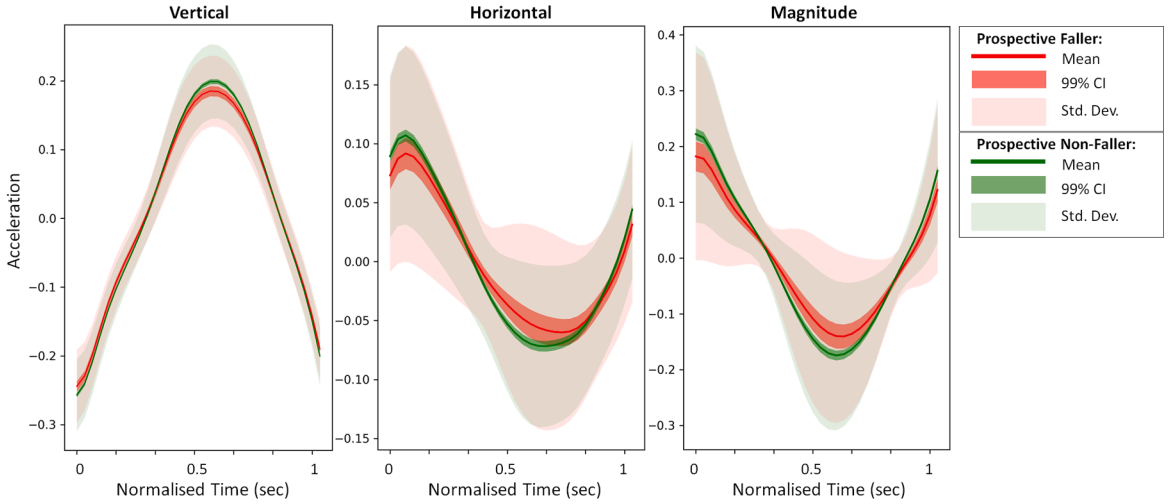

Fig. 4. Average Step Acceleration patterns for fallers and non-fallers.

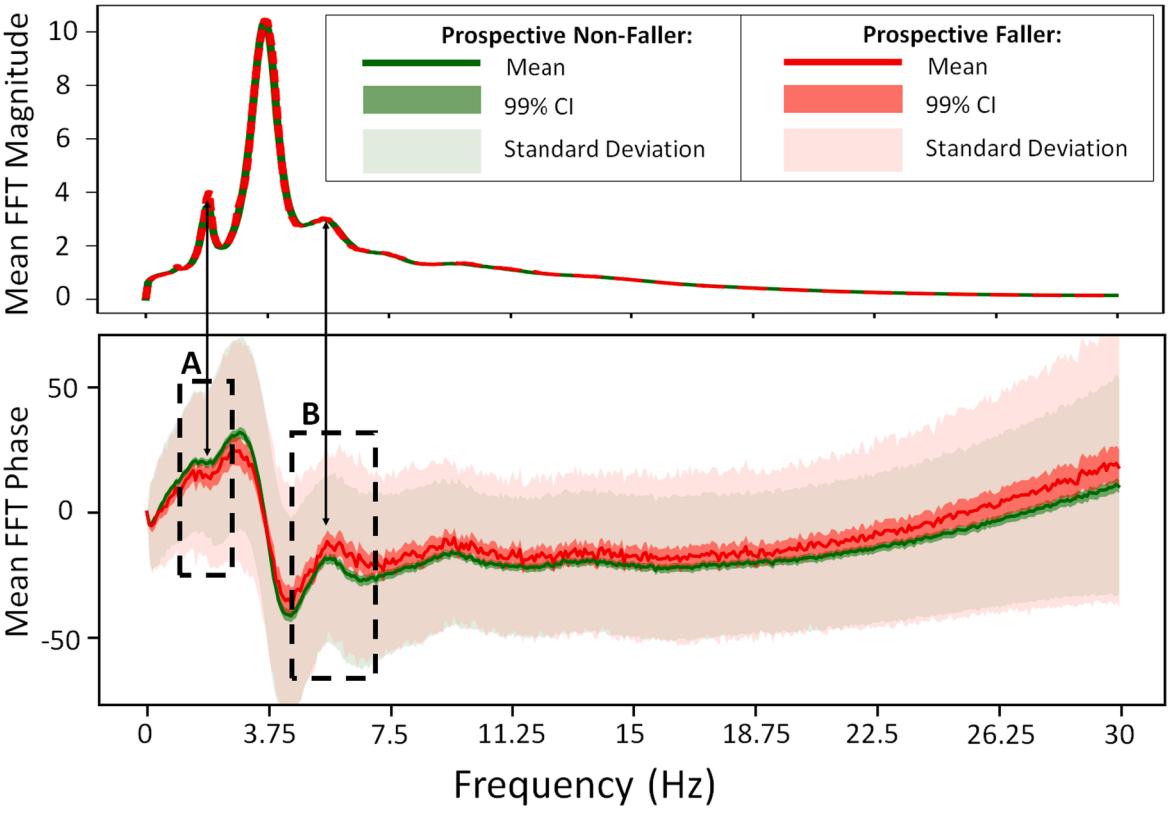

Fig. 5. Mean FFT Magnitude and Phase Signals computed from 10 s FFT windows for all participants using A_{mag}.

between fallers and non-fallers.

Fig. 6 shows FFT phase summary features for 8 different participants where four participants were prospective fallers and four were not. It can be seen that Participants E, F and G, who experienced prospective falls, have a lower phase value at ∼1.9 Hz and a larger phase at ∼5.5 Hz when compared to Participants A, B and C respectively. However, FFT

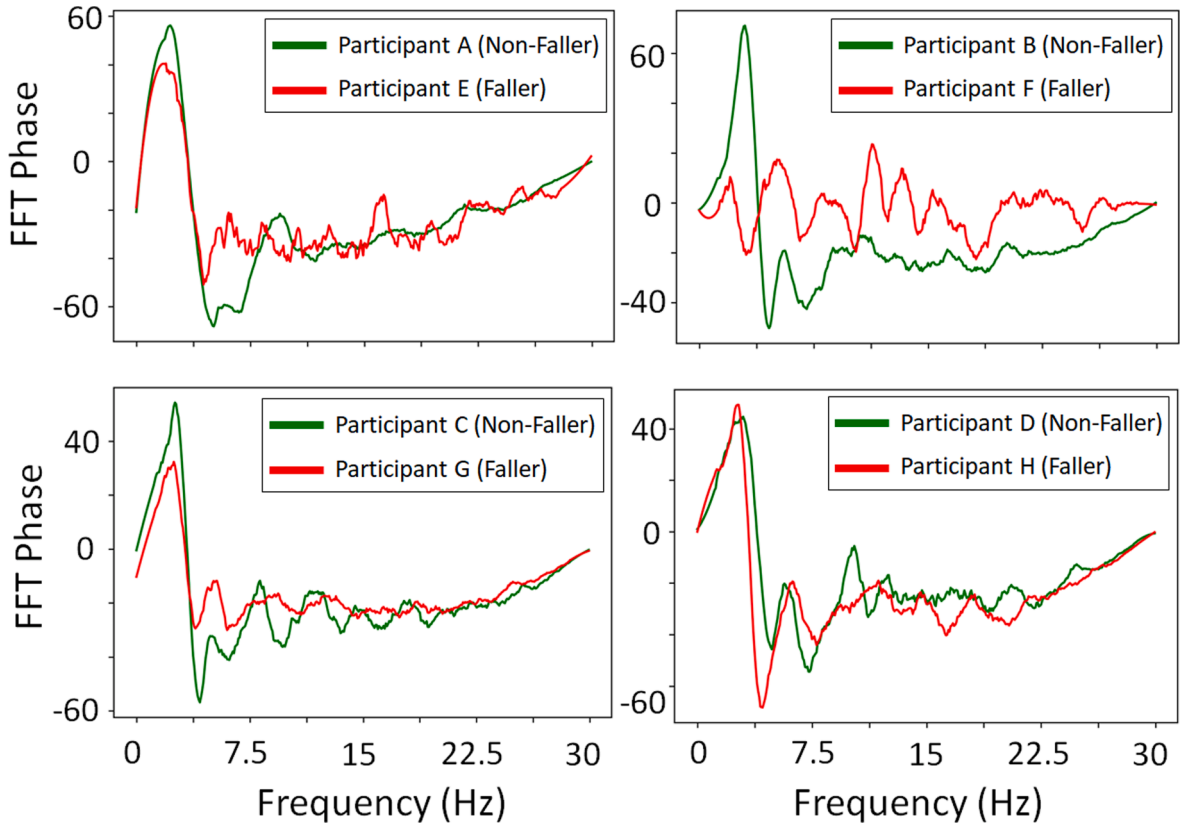

Fig. 6. Mean FFT Phase Signals for 8 different participants (4 non-fallers and 4 fallers).

phase summary features for Participant D and H do not follow the same pattern.

3.2. Machine learning classification

After model configuration and training was performed for each data type (FLA, PSW, FT), testing was performed on each model using the holdout test set (N = 428, 64 Fallers). Results include independent assessment of FLA, FT and PSW data types as well as results evaluating potential complementary information provided by combining data types. Tables 5 and 6 show fall classification performance scores for retrospective and prospective fall respectively for the different data types. Sensitivity and Specificity scores are shown, in both tables, for the cross-validation set and the holdout test set when early stopping was and was not implemented. Confidence intervals were calculated using the Normal Approximation (Wald) method. Results show that ML models trained on FLA data performed best for prospective falls with sensitivity and specificity of 0.61 and 0.66 respectively. Similarly, FLA data also performed best for retrospective falls with sensitivity and specificity of 0.61 and 0.68 respectively.

The best performing model for retrospective falls using FLA data, as configured and selected by the Bayesian Optimization algorithm, was based on a Naive Bayes classifier. Feature vectors were prepossessed by first standardizing features then processed using a Nystroem kernel map using a polynomial kernel. The best performing model for prospective falls using FLA data, as configured and selected by the Bayesian

Table 5
Performance on Retrospective Falls. (F-select: feature selection, where subset of features are chosen based on statistical analysis or on structure of a trained ML model. d = polynomial degree. n = number of components. c = number of clusters).

		CV		Test Set		Selected ML Algorithms	
	Data	Sens	Spec	Sens (95% CI)	Spec (95% CI)	ML Model	Feature processing
No Early Stopping	FLA	0.65	0.63	0.52 (0.43–0.61)	0.58 (0.56–0.67)	Naive Bayes	Nystroem sampler(n = 150)
	PSW	0.59	0.62	0.50 (0.41–0.59)	0.56 (0.50–0.62)	SVM(rbf)	F-Select(SVM)
	FT	0.60	0.58	0.58 (0.49–0.67)	0.6 (0.54–0.66)	SGD(squared hinge)	F-Select(Extra Trees)
	FLA + PSW	0.61	0.67	0.55(0.46–0.64)	0.63 (0.57–0.68)	SVM(sigmoid)	F-Select(SVM)
	FLA + FT	0.59	0.70	0.52 (0.43–0.61)	0.65 (0.59–0.71)	SGD(squared-hinge)	K-PCA(n = 100)
	PSW + FT	0.60	0.58	0.58 (0.49–0.67)	0.63 (0.57–0.68)	SVM(RBF)	F-Select(SVM)
	FLA + FT + PSW	0.67	0.64	0.50 (0.41–0.59)	0.60 (0.54–0.65)	SGD(hinge)	Agglomeration(c = 15)
Early Stopping	FLA	**0.61**	**0.67**	**0.61 (0.53–0.70)**	**0.68 (0.62–0.73)**	Naive Bayes	Nystroem sampler(n = 150)
	PSW	0.58	0.61	0.57 (0.48–0.65)	0.60 (0.54–0.65)	SVM(rbf)	F-Select(SVM)
	FT	0.59	0.60	0.60 (0.51–0.68)	0.61 (0.54–0.65)	SGD(squared hinge)	PCA(97% Variance)
	FLA + PSW	0.64	0.6	0.64 (0.55–0.72)	0.58 (0.52–0.64)	SVM(sigmoid)	F-Select(SVM)
	FLA + FT	0.6	0.66	0.58 (0.49–0.66)	0.68 (0.62–0.73)	SGD(squared-hinge)	K-PCA(n = 100)
	PSW + FT	0.60	0.60	0.60 (0.52–0.63)	0.61 (0.54–0.65)	SVM(RBF)	F-Select(SVM)
	FLA + FT + PSW	0.60	0.63	0.59 (0.50–0.68)	0.65 (0.59–0.70)	SVM(linear)	F-Select(Extra Trees)

Table 6
Performance on Prospective Falls. (F-select: feature selection, where subset of features are chosen based on statistical analysis or on structure of a trained ML model. d = polynomial degree. n = number of components. c = number of clusters)

		CV		Test Set		Selected ML Algorithms	
	Data	Sens	Spec	Sens (95% CI)	Spec (95% CI)	ML Model	Feature processing
No Early Stopping	FLA	0.67	0.69	0.5 (0.37–0.63)	0.64 (0.59–0.69)	SGD(Modified Huber)	Agglomoration(n = 28)
	PSW	0.55	0.57	0.51 (0.39–0.64)	0.55 (0.50–0.60)	Random Forest	F-select(ANOVA best 20)
	FT	0.52	0.57	0.42 (0.30–0.55)	0.58 (0.53–0.63)	Naive Bayes	F-select(extra trees)
	FLA + PSW	0.59	0.68	0.44 (0.31–0.57)	0.64 (0.59–0.69)	SVM(RBF)	Nystroem sampler(n = 150)
	FLA + FT	0.63	0.72	0.47 (0.34–0.60)	0.69 (0.64–0.74)	SGD(squared hinge)	Kernel-PCA(n = 100)
	PSW + FT	0.5	0.64	0.5 (0.37–0.63)	0.64 (0.59–0.69)	Extra-Trees(n = 100)	F-select(Extra Trees)
	FLA + FT + PSW	0.66	0.66	0.45 (0.33–0.58)	0.61 (0.56–0.66)	SVM(RBF)	Agglomeration(c = 25)
Early Stopping	FLA	**0.64**	**0.63**	**0.61 (0.49–0.71)**	**0.66 (0.61–0.71)**	SGD(hinge)	PCA(98% Variance)
	PSW	0.54	0.61	0.55 (0.42–0.67)	0.61 (0.56–0.66	Extra-Trees	PCA(97% Variance)
	FT	0.55	0.6	0.56 (0.43–0.68)	0.62 (0.57–0.67)	Naive Bayes	F-select(random trees)
	FLA + PSW	0.58	0.61	0.58 (0.45–0.70)	0.63 (0.58–0.68)	SVM(linear)	Polynomial Features (d = 2)
	FLA + FT	0.62	0.63	0.63 (0.50–0.74)	0.64 (0.59–0.69)	SGD(modified huber)	F-select(ANOVA best 100)
	PSW + FT	0.54	0.61	0.53 (0.40–0.66)	0.62 (0.57–0.67)	SVM(RBF)	Polynomial Features(d = 2)
	FLA + FT + PSW	0.6	0.61	0.59 (0.46–0.71)	0.62 (0.57–0.67)	SVM(sigmoid)	PCA(97% Variance)

Optimization algorithm, was based on a Stochastic Gradient Descent (SGD) classifier. Feature vectors were prepossessed by first standardizing features then reducing the dimension using PCA to only include 98% of variance.

Confidence Interval (CI) ranges, in Tables 5 and 6, overlap for a number of different data type models. Thus, it is difficult to determine, based on CI alone, whether there is a statistically significant difference in performance between a given pair of models. To further investigate performance difference between pairs of models, a McNemars test was performed on all pairs of early stopping models to evaluate if there was a significant difference in the dichotomous predictions made by each pair of models. Table 7 and 8 show results of the McNemar tests for retrospective and prospective falls respectively using a 95% confidence level. The tests show that most models perform with no significant difference in sensitivity. The best performing model (FLA) performs with a statistically significant difference in specificity compared to a number of other models for both retrospective and prospective falls.

3.2.1. *Effect of incorrect modeling/validation methodologies*

This section briefly describes two additional experiments conducted to evaluate the effect two problematic machine learning modeling/validation methodologies, described in Section 1.1.3, have on evaluation results.

Exp1: investigates the effect of feature selection being performed on the training set and test set. Only FLA data was used and one condition within the protocol was changed such that feature selection was performed on both the training set and the holdout test set before initiating the Bayesian Optimization process. A univariate statistical test was performed on all FLA features and the 100 features with the lowest p values were selected. Results of the experiment for prospective and retrospective falls are shown in Table 9. When compared with FLA results reported in the previous Section, 'Exp1' performance shows an artificial increase in performance of 7% for the holdout test set as a result of bias, towards the test set, being introduced into the model. When predicting prospective falls, for example, sensitivity and specificity was originally 0.61 and 0.66 respectively. Sensitivity and specificity increase to 0.68 and 0.74 respectively when the test set is used during feature selection. The results for Exp1, however, do not reflect real-world performance. Results do illustrate the problem of over optimistic metrics being reported as a result of the test set being used in model development.

Table 7
Matrix showing comparison of retrospective fall model pairs using McNemars test (P = Difference in Sensitivity between model pair, N = Difference in Specificity between model pair)

	FLA	PSW	FT	FLA+ PSW	FLA+ FT	PSW + FT
PSW	-N					
FT	-N	- -				
FLA + PSW	-N	- -	-N			
FLA + FT	- -	-N	-N	-N		
PSW + FT	-N	- -	- -	-N	-N	
FLA + FT + PSW	- -	-N	-N	-N	-N	-N

Table 8
Matrix showing comparison of prospective fall model pairs using McNemars test (P = Difference in Sensitivity between model pair, N = Difference in Specificity between model pair)

	FLA	PSW	FT	FLA+ PSW	FLA+ FT	PSW + FT
PSW	-N					
FT	-N	- -				
FLA + PSW	-N	- -	- -			
FLA + FT	- -	- -	-N	- -		
PSW + FT	P N	- -	- -	- -	-N	
FLA + FT + PSW	-N	- -	- -	- -	- -	- -

Exp2: investigates the appropriateness of CV as a methodology for evaluating performance in small SFRT data sets. As discussed in Section 1.1.3, there are potential issues with the use of CV as a method of evaluating SFRT models, particularly in terms of large uncertainty for small sample sizes. A training subset was therefore created to represent an average SFRT data-set (N = 150, 50 Fallers) using random subsampling. Only FLA data was used and one protocol condition was changed such that the full training set was not used and instead the smaller training subset was used. Results of the experiment for prospective and retrospective falls are shown in Table 9. The difference between CV performance and holdout test set performance is of particular interest for this experiment. CV is often used in the literature without an external data-set or consideration for over-fitting on the CV data-set, therefore particular attention should also be paid to results not implementing early stopping. When no early stopping is implemented, results show significantly over-optimistic CV performance, with prospective falls having a sensitivity and specificity of 0.84 and 0.79 respectively compared to 0.45 and 0.57 for the holdout test set.

The Bayesian Optimization process aims to select optimal ML pipeline configuration parameters (e.g. learning algorithm, model hyperparameters) by maximizing CV performance. This mimics the process, commonly employed by researchers, of performing CV on different ML pipelines configurations through manual trial and error or by using techniques such as grid search or random search in order to optimize the

Table 9
Results of 2 Experiments to evaluate the effect of Incorrect Modeling Methodologies (Exp1 = Train and Test used for feature selection. Exp2 = Small data-set used with CV as main evaluation. ES = Early Stopping. F-select = feature selection, where subset of features are chosen based on statistical analysis or on structure of a trained ML model. d = polynomial degree. n = number of components. c = number of clusters)

			CV		Test Set		Selected ML Algorithms	
	Classification	ES	Sens	Spec	Sens (95% CI)	Spec (95% CI)	ML Model	Feature processing
Exp 1	Prospective	N	0.66	0.75	0.63 (0.54–0.71)	0.68 (0.62–0.73)	SGD(squared hinge)	Agglomeration(c = 26)
	Prospective	Y	0.65	0.71	0.68 (0.59–0.76)	0.74 (0.68–0.78)	SGD(squared hinge)	Agglomeration(c = 40)
	Retrospective	N	0.65	0.72	0.60 (0.46–0.71)	0.64 (0.59–0.69)	SGD(squared hinge)	Agglomeration(c = 100)
	Retrospective	Y	0.63	0.69	0.67 (0.54–0.78)	0.70 (0.65–0.75)	SGD(log)	Agglomeration(c = 29)
Exp 2	Prospective	N	0.84	0.79	0.45 (0.36–0.54)	0.57 (0.51–0.63)	SGD(squared hinge)	Agglomeration(c = 26)
	Prospective	Y	0.59	0.62	0.52 (0.43–0.61)	0.60 (0.54–0.66)	Naive Bayes	Agglomeration(c = 17)
	Retrospective	N	0.78	0.76	0.46 (0.33–0.58)	0.59 (0.54–0.64)	SGD(squared-hinge)	Agglomeration(c = 100)
	Retrospective	Y	0.52	0.61	0.50 (0.37–0.63)	0.60 (0.55–0.65)	SGD(perceptron)	F-Select(ANOVA best 100)

performance metric produced by CV. However, this process results in a classifier that is over-fitted on training data. While individual models within each fold have not been over-fitted, the over-fitting in this instance is due to the ML pipeline configurations being selected to maximize performance on the training set only with no regard for data external to the training set. CV is therefore more appropriate in situations where model selection and configuration has been performed using data that is not used in the CV evaluation process or when techniques to counteract over-fitting of pipeline configurations has been implemented.

4. Discussion

The aim of SFRT systems is to improve the fall risk screening process for identifying people at high risk of falling in the future. However, statistical analysis of commonly used screening techniques showed that, on their own, the tools have limited ability to correctly identify fallers. Univariate statistical analysis of commonly used functional tests (TUG, GV and GS) showed poor discriminative ability to correctly identifying participants who experienced falls in the future. While statistical t-tests showed significant differences between fallers and non-fallers for the different functional tests, ROC analysis showed poor fall prediction performance for TUG, GV and GS. Previous work has reported fall prediction AUC results higher than the results reported in this work. For example, Kojima et al. [17] report an AUC of 0.58 for the TUG assessment compared to an AUC of 0.48 in this work. Investigating this further, while the average TUG time for non-fallers were similar for both studies, the average TUG time for fallers in our study (9.9 sec) was significantly faster than that of the time reported by Kojima et al. (11.4 sec). It is possible that the poor performance of the clinical assessments in this work is in part due to the group of prospective fallers having a higher than average physical capacity compared to fallers in other studies.

Univariate analysis of gait measurements, extracted from a lab-based pressure sensitive walkway, showed that stride velocity was significantly different between fallers and non-fallers. However, similar to the functional tests, ROC analysis results showed limited ability to classify prospective or retrospective fallers. It is likely, however, that combining multiple lab based assessment tools together could increase the accuracy of predicting future falls.

Statistical analysis of accelerometer based measures suggests that gait quality is potentially useful in screening for fall risk. Analysis of step profiles, captured from an accelerometer worn in free-living conditions, showed significant differences in the horizontal acceleration before and after vertical acceleration step peaks when comparing fallers and non-fallers. Horizontal acceleration signals had a smaller range between maximum and minimum points during ambulatory activity while also exhibiting larger variability. This suggests fallers are more likely to exhibit reduced and/or unstable trunk sway. Similarly, FFT phase signals, computed from Accelerometer magnitude during ambulatory activity, showed significant differences at 1.5 Hz and 5.5 Hz when comparing fallers and non-fallers. This suggests that ambulatory movements performed at frequencies of 1.5 Hz and 5.5 Hz are more closely aligned in time for fallers when compared to non-fallers. While it is difficult to determine the exact reason for this, we postulate that it is likely due to participants with gait difficulties developing new gait patterns to overcome the specific difficulties. For example, participants experiencing leg pain may reduce stance phase duration to reduce overall pain. These gait modifications are likely to result in a more regular time aligned gait.

Machine learning based experiments, using features constructed from different combinations of FLA, FT and PSW data types showed that the FLA data type performs best for retrospective falls with a sensitivity and specificity of 0.61 and 0.68 respectively. Similarly, the FLA data type also performs best for prospective falls with a sensitivity and specificity of 0.61 and 0.66 respectively. Interestingly, while performance scores were marginally higher for retrospective falls, there was no significant difference between the 2 best performing models for retrospective and prospective falls. Similar to results of the statistical analysis, FT and PSW data did not perform well in predicting prospective falls. Both FT and PSW performed with moderate performance when classifying retrospective fallers.

Prediction of falls in the future is an extremely challenging problem. Previously reported performance measures of prospective falls prediction in the literature are difficult to interpret due to the use of problematic modelling and validation methodologies and likelihood of results being over-optimistic. Direct comparison between results reported in this work and results in the literature are therefore difficult. Liu et al. [29], however, appear to use robust modelling and validation methodology with feature selection performed on training data only and the use of a holdout test set. Participants (N = 95) performed an instrumented Alternate Step Test and regression models, trained on half of the data, achieved sensitivity and specificity of 69% and 68% respectively when tested on the other half of the data.

As previously discussed in Section 1.1.3, it vital that performance of an SFRT system be evaluated such that results reflect real-world performance. Results show that over-optimistic performance results can be produced if modelling and evaluation methodologies are not appropriate. Results show that over fitting occurs on the full cross-validation set when early stopping is not implemented. When early stopping was not implemented, FLA performance on the holdout test set dropped by 3%-6% compared to performance on the cross-validations set. This performance drop increased significantly to 22%-39% when a smaller CV training set (N = 150) was used. No significant difference in performance was seen between cross-validation and holdout test set when early stopping is implemented. Relying on CV as the only evaluation methodology, particularly when data is limited, can result in ML pipeline configurations that are over-fitted on the training set unless methodologies to counteract overfitting are implemented. Performance on a holdout test set is, therefore, the most appropriate evaluation methodology as it is more representative of performance in real-world settings. Over-optimistic results were also shown to occur when the test set is

included in the feature selection process. Results indicated that bias, towards the test set, was introduced into the model resulting in an over-optimistic performance increase of 7%.

There are some limitations to the study. First, recording of prospective falls using telephone interviews, 6 and 12 months after observation, can result in inaccuracies due to participants no recognizing or remembering a fall. Secondly, the machine learning experiments used a Bayesian optimization algorithm based on 15 classifiers and 14 feature pre-processing algorithms. It is possible that other classification or pre-processing algorithms may have produced improved performance for different experiment conditions. Future work could, for example, investigate deep learning based models to classify fallers using FLA data.

5. Conclusion

This study illustrates the potential of accelerometers, worn in free-living conditions, as an unobtrusive and cost effective way of performing continuous fall risk screening. In addition, the study also highlighted the importance of correct evaluation methodologies when reporting fall risk prediction performance showing that it is vital that future works report performance that will reflect real-world performance to ensure continued advancement of the SFRT field.

A recent theoretical modelling analysis concluded that the maximal accuracy of a fall prediction model, attempting to identify people with at least one fall incident over the course of a year would not exceed 0.81 [13]. While the results achieved are not perfect, when compared to the theoretical maximum of 0.81, sensitivity and specificity in the ranges of 0.61–68 are good. When compared with other commonly used fall risk indicators, FLA performs with a significantly higher accuracy and at a fraction of the cost of lab-based measures. By enabling risk assessment to be conducted at home with low cost technology, the potential of this technology to have real-world impact is strong. Risk screening could be performed on significantly more people, therefore increasing the potential to provide earlier fall-risk reduction interventions for more people. While gait quality is a strong predictor of fall risk, it is not the only risk factor. Falls are a multicausal phenomenon with a complex interaction between participant characteristics and environmental factors. The scope of this work was to focus on gait quality. However, it is possible that additional complementary screening tools could increase sensitivity and specificity even further.

Data availability

The data that support the findings of this study are available from Umea University. Restrictions, related to ethical approval for the data collection, apply to the availability of these data and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of from Umea University.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the European Union Interreg Northern Periphery and Arctic 2014-2020 program.

We are grateful for access to the Tier 2 High Performance Computing resources provided by the Northern Ireland High Performance Computing (NI-HPC) facility, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/T022175/1.

References

[1] World Health Organization, WHO global report on falls prevention in older age, tech. rep., 2008.
[2] J.A. Haagsma, B.F. Olij, M. Majdan, E.F. Van Beeck, T. Vos, C.D. Castle, Z.V. Dingels, J.T. Fox, E.B. Hamilton, Z. Liu, N.L. Roberts, D.O. Sylte, O. Aremu, T.W. Bärnighausen, A.M. Borzì, A.M. Briggs, J.J. Carrero, C. Cooper, Z. El-Khatib, C.L. Ellingsen, S.M. Fereshtehnejad, I. Filip, F. Fischer, J.M. Haro, J.B. Jonas, A.A. Kiadaliri, A. Koyanagi, R. Lunevicius, T.J. Meretoja, S. Mohammed, A. Pathak, A. Radfar, S. Rawaf, D.L. Rawaf, L.S. Riera, I. Shiue, T.J. Vasankari, S.L. James, S. Polinder, Falls in older aged adults in 22 European countries: incidence, mortality and burden of disease from 1990 to 2017, Injury Prevention 26 (2020) i67-i74.
[3] S. Heinrich, K. Rapp, U. Rissmann, C. Becker, H.H. König, Cost of falls in old age: A systematic review, Osteoporosis Int. 21 (2010) 891–902.
[4] S.S. Rao, Prevention of Falls in Older Patients - American Family Physician, Am. Family Phys. 72 (2005) 81–88.
[5] J. Howcroft, J. Kofman, E.D. Lemaire, Review of fall risk assessment in geriatric populations using inertial sensors, J. NeuroEng. Rehab. 10 (1) (2013) 91.
[6] J.A. Stevens, E.A. Phelan, Development of STEADI: A Fall Prevention Resource for Health Care Providers, Health Promotion Pract. 14 (2013) 706–714.
[7] R. Sun, J.J. Sosnoff, Novel sensing technology in fall risk assessment in older adults: A systematic review, BMC Geriatr. 18 (2018) 1–10.
[8] L. Montesinos, R. Castaldo, L. Pecchia, Wearable inertial sensors for fall risk assessment and prediction in older adults: A systematic review and meta-analysis, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2018) 573–582.
[9] M. Olson, T. Lockhart, Predicting Fall Risk Through Automatic Wearable Monitoring, Int. J. Progn. Health Manage. 12 (2021) 8.
[10] M. Nouredanesh, A. Godfrey, J. Howcroft, E.D. Lemaire, J. Tung, Fall risk assessment in the wild: A critical examination of wearable sensors use in free-living conditions, Gait Posture 5 (2020).
[11] P. Bet, P.C. Castro, M.A. Ponti, Fall detection and fall risk assessment in older person using wearable sensors: A systematic review, Int. J. Med. Informat. 130 (2019) 103946.
[12] S.-H. Park, Tools for assessing fall risk in the elderly: a systematic review and meta-analysis, Aging Clin. Exp. Res. 30 (2018) 1–16.
[13] T. Shany, K. Wang, Y. Liu, N.H. Lovell, S.J. Redmond, Review: Are we stumbling in our quest to find the best predictor? Over-optimism in sensor-based models for predicting falls in older adults, Healthcare Technol. Lett. 2 (2015) 79–88.
[14] M.W. Rivolta, M. Aktaruzzaman, G. Rizzo, C.L. Lafortuna, M. Ferrarin, G. Bovi, D. R. Bonardi, A. Caspani, R. Sassi, Evaluation of the Tinetti score and fall risk assessment via accelerometry-based movement analysis, Artif. Intell. Med. 95 (2019) 38–47.
[15] M. Arvandi, B. Strasser, K. Volaklis, K.-H. Ladwig, E. Grill, R. Matteucci Gothe, A. Horsch, M. Laxy, U. Siebert, A. Peters, B. Thorand, C. Meisinger, Mediator Effect of Balance Problems on Association Between Grip Strength and Falls in Older Adults: Results From the KORA-Age Study, Gerontol. Geriatric Med. 4 (2018) 233372141876012.
[16] M. Marschollek, K.H. Wolf, M. Gietzelt, G. Nemitz, H.M. Zu Schwabedissen, and R. Haux, "Assessing elderly persons' fall risk using spectral analysis on accelerometric data - A clinical evaluation study", in Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - Personalized Healthcare through Technology, pp. 3682–3685, IEEE Computer Society, 2008.
[17] G. Kojima, T. Masud, D. Kendrick, R. Morris, S. Gawler, J. Treml, and S. Iliffe, "Does the timed up and go test predict future falls among British community-dwelling older people? Prospective cohort study nested within a randomised controlled trial," BMC Geriatrics, vol. 15, pp. 1–7, 4 2015.
[18] U. Olsson Möller, J. Kristensson, P. Midlöv, C. Ekdahl, U. Jakobsson, "Predictive Validity and Cut-Off Scores in Four Diagnostic Tests for Falls – A Study in Frail Older People at Home," https://doi.org/10.3109/02703181.2012.694586, vol. 30, pp. 189–201, 9 2012.
[19] P. Bet, P.C. Castro, M.A. Ponti, Fall detection and fall risk assessment in older person using wearable sensors: A systematic review, 10 2019.
[20] T. Doi, S. Hirata, R. Ono, K. Tsutsumimoto, S. Misu, H. Ando, The harmonic ratio of trunk acceleration predicts falling among older people: results of a 1-year prospective study, J. Neuroeng. Rehabil. 10 (2013) 7.
[21] U. Laessoe, H.C. Hoeck, O. Simonsen, T. Sinkjaer, M. Voigt, Fall risk in an active elderly population - Can it be assessed? J. Negative Res. BioMed. 6 (2007) 2.
[22] M. Marschollek, A. Rehwald, K.H. Wolf, M. Gietzelt, G. Nemitz, H. Meyer zu Schwabedissen, R. Haux, Sensor-based fall risk assessment - an expert 'to go', Methods Inform. Med. 50 (2011) 420–426.
[23] K. Paterson, K. Hill, N. Lythgo, Stride dynamics, gait variability and prospective falls risk in active community dwelling older women, Gait Posture 33 (2011) 251–255.
[24] R. Schwesig, D. Fischer, A. Lauenroth, S. Becker, S. Leuchte, Can falls be predicted with gait analytical and posturographic measurement systems? A prospective follow-up study in a nursing home population, Clin. Rehabil. 27 (2013) 183–190.
[25] B.R. Greene, E.P. Doheny, C. Walsh, C. Cunningham, L. Crosby, R.A. Kenny, Evaluation of Falls Risk in Community-Dwelling Older Adults Using Body-Worn Sensors, Gerontol. 58 (2012) 472–480.
[26] J. Howcroft, J. Kofman, E.D. Lemaire, Prospective Fall-Risk Prediction Models for Older Adults Based on Wearable Sensors, IEEE Trans. Neural Syst. Rehab. Eng. 25 (2017) 1812–1820.
[27] B.C. Kwok, R.A. Clark, Y.H. Pua, Novel use of the Wii Balance Board to prospectively predict falls in community-dwelling older adults, Clin. Biomech. 30 (2015) 481–484.

[28] M. Schwenk, K. Hauer, T. Zieschang, S. Englert, J. Mohler, B. Najafi, S. Hauer, Zieschang, Englert, Mohler, Najafi, Sensor-Derived Physical Activity Parameters Can Predict Future Falls in People with Dementia, Gerontology 60 (2014) 483–492.
[29] Y. Liu, S.J. Redmond, T. Shany, J. Woolgar, M.R. Narayanan, S.R. Lord, N.H. Lovell, Validation of an accelerometer-based fall prediction model, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, pp. 4531–4534, Institute of Electrical and Electronics Engineers Inc., 11 2014.
[30] M. Gietzelt, F. Feldwieser, M. Gövercin, E. Steinhagen-Thiessen, M. Marschollek, A prospective field study for sensor-based identification of fall risk in older people with dementia, Informat. Health Soc. Care 39 (2014) 249–261.
[31] M.J. Mohler, C.S. Wendel, R.E. Taylor-Piliae, N. Toosizadeh, B. Najafi, Motor Performance and Physical Activity as Predictors of Prospective Falls in Community-Dwelling Older Adults by Frailty Level: Application of Wearable Technology, Gerontology 62 (2016) 654–664.
[32] K.S. Van Schooten, M. Pijnappels, S.M. Rispens, P.J. Elders, P. Lips, A. Daffertshofer, P.J. Beek, J.H. Van Dieën, Daily-life gait quality as predictor of falls in older people: A 1-year prospective cohort study, PLoS ONE 11 (2016) 7.
[33] A. Nait Aicha, G. Englebienne, K. van Schooten, M. Pijnappels, B. Kröse, Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry, Sensors 18 (2018) 1654.
[34] J. Howcroft, E.D. Lemaire, J. Kofman, Prospective elderly fall prediction by older-adult fall-risk modeling with feature selection, Biomed. Signal Process. Control 43 (2018) 320–328.
[35] A. Hua, Z. Quicksall, C. Di, R. Motl, A.Z. LaCroix, B. Schatz, D.M. Buchner, Accelerometer-based predictive models of fall risk in older women: a pilot study, npj Digital Med. 1 (2018) 25.
[36] V. Robles-García, Y. Corral-Bergantiños, N. Espinosa, M.A. Jácome, C. García-Sancho, J. Cudeiro, P. Arias, Spatiotemporal Gait Patterns During Overt and Covert Evaluation in Patients With Parkinson's Disease and Healthy Subjects: Is There a Hawthorne Effect?, J. Appl. Biomech. 31 (2015) 189–194.
[37] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[38] A. Isaksson, M. Wallman, H. Göransson, M.G. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, Pattern Recogn. Lett. 29 (2008) 1960–1965.
[39] M. Christ, N. Braun, J. Neuffer, A.W. Kempa-Liehr, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package), Neurocomputing 307 (2018) 72–77.
[40] M. Feurer, A. Klein, K.E. Jost, T. Springenberg, M. Blum, F. Hutter, Efficient and Robust Automated Machine Learning, tech. rep., 2015.
[41] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263–1284.