Med Biol Eng Comput (2017) 55:45–55
DOI 10.1007/s11517-016-1504-y

ORIGINAL ARTICLE

A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials

Omar Aziz[1,3,5] · Magnus Musngi[5] · Edward J. Park[5] · Greg Mori[4] · Stephen N. Robinovitch[1,2,3]

Received: 11 June 2015 / Accepted: 28 March 2016 / Published online: 22 April 2016

Abstract Falls are the leading cause of injury-related morbidity and mortality among older adults. Over 90 % of hip and wrist fractures and 60 % of traumatic brain injuries in older adults are due to falls. Another serious consequence of falls among older adults is the 'long lie' experienced by individuals who are unable to get up and remain on the ground for an extended period of time after a fall. Considerable research has been conducted over the past decade on the design of wearable sensor systems that can automatically detect falls and send an alert to care providers to reduce the frequency and severity of long lies. While most systems described to date incorporate threshold-based algorithms, machine learning algorithms may offer increased accuracy in detecting falls. In the current study, we compared the accuracy of these two approaches in detecting falls by conducting a comprehensive set of falling experiments with 10 young participants. Participants wore waist-mounted tri-axial accelerometers and simulated the most common causes of falls observed in older adults, along with near-falls and activities of daily living. The overall performance of five machine learning algorithms was greater than the performance of five threshold-based algorithms described in the literature, with support vector machines providing the highest combination of sensitivity and specificity.

Keywords Older adults · Falls · Wearable sensors · Machine learning · Threshold-based algorithms

1 Introduction

Falls are the leading cause of injury-related hospitalization among older adults. Almost half of older adults who fall experience a minor injury, and up to 25 % will experience a more serious injury such as a fracture [16]. A frequent and serious consequence of falls in seniors is the 'long lie,' where the faller is unable to get up on his own and remains helpless on the ground. Vellas et al. [30] reported that 70 % of older adults who had fallen at home were unable to get up unaided, and more than 20 % of patients admitted to hospital as a result of a fall had been on the ground for an hour or more. Tinetti et al. [29] found that 47 % of non-injured fallers were unable to rise from the floor without assistance. Fall-related long lies among older adults are associated with pneumonia, dehydration, hypothermia, and high mortality rates, and contribute to fear of falling and social isolation [20, 23, 24]. Half of elderly people who experienced a long lie (for an hour or more) passed away within 6 months, even if no direct injury occurred from the fall [33].

Over the past decade, there has been a great deal of research examining the role of wearable accelerometers and/or gyros for the automatic detection of falls [7–15, 17, 22, 25]. The goal of these systems was to reduce the frequency and consequences of long lies, by quickly and

✉ Omar Aziz
oaziz@sfu.ca

[1] Injury Prevention and Mobility Laboratory, Simon Fraser University, Burnaby, BC, Canada

[2] Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC, Canada

[3] School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada

[4] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

[5] School of Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC, Canada

accurately detecting the occurrence of a fall and alerting care providers to these events. While the cost and size of these sensors are rapidly decreasing, we are still at an early stage in the development of sensor systems (hardware and software combinations) that provide adequate sensitivity (ability to detect actual falls) and specificity (ability to avoid false positives, which could desensitize the recipient of a fall alarm signal).

Research to date has focused primarily on testing the accuracy of these systems with datasets gathered from young adults who simulate falls onto padded mattresses in the laboratory environment and wear the system while performing activities of daily living (ADLs). For example, Bourke et al. [7, 10] used signals from tri-axial accelerometers mounted at the trunk and thigh to distinguish falls from ADLs. They proposed a upper fall threshold (UFT) and lower fall threshold (LFT) in an attempt to optimize the balance of false positives and false negatives. The UFT showed 100 % sensitivity and 100 % specificity, while the LFT provided 100 % sensitivity and 91 % specificity. Similarly, Kangas et al. [17] attached a tri-axial accelerometer at the waist, wrist, and head of volunteers who performed simulated falls and ADLs in the laboratory. Their algorithms considered the pre-impact, impact, and post-impact phases of the fall, separately and in combination, and achieved up to 100 % specificity and 95 % sensitivity, based on a single sensor mounted at the waist.

Despite exhibiting high classification accuracy in laboratory experiments, inertial sensor-based fall detection systems have yet to achieve high market penetration. One barrier to their acceptance is the lack of evidence of their effectiveness in real-world falling scenarios in older adults. The only study, we are aware of, examining real-world accuracy was conducted recently by Bagala et al. [6], who evaluated 13 fall detection methods (including the Bourke and Kangas algorithms described above) using data from 29 real-world falls experienced by older adults that were recorded with wearable accelerometers. They found that the specificity of the thirteen algorithms averaged 83.0 % (SD = 30.3 %; maximum value = 98 %), and the sensitivity averaged 57.0 % (SD = 27.3 %; maximum value = 82.8 %), considerably lower than the values obtained with simulated falls. There could be various reasons contributing to lower accuracies of these algorithms on the real-world data; however, the primary reason could be that the fall, and ADL scenarios simulated in the laboratory experiments bear little resemblance to such real-world incidents.

In laboratory settings, the development of improved algorithms to automatically detect falls in older adults requires an understanding of the real-life common scenarios/fall mechanisms in older adults and incorporating that information in designing laboratory experiments. That being said, it is only recently that more detailed evidence has emerged indicating the common scenarios (cause and circumstances) of falls in older adults. In particular, recent findings from Robinovitch et al. [27], based on the analysis of video footage of 227 falls experienced by 130 older residents of long-term care, indicate that 48 % of falls occur while walking, and 86 % are collectively due to incorrect shift of bodyweight ([ISBW] 41 %), tripping (21 %), hit/bump (11 %), collapse/loss of consciousness (10 %), and slipping (3 %). These common fall scenarios are often found missing in the majority, if not all, of the previous laboratory-based falling experiments, and the resulting discrepancy in sensor data is, perhaps, the main cause of the poor accuracy of the fall detection algorithms, when tested on real-life fall scenarios.

Another possible reason behind low accuracy (in particular the false positives) of fall detection algorithms on real-world fall and ADL scenarios could be the lack of inclusion of near-fall trials in laboratory experiment protocol. While researchers have recreated near-fall scenarios in their laboratory experiments [3, 31], such studies were not conducted particularly to distinguish falls from non-falls; rather, they were carried out to distinguish near-falls from other gait patterns in order to evaluate fall risks. The only fall detection study that included near-falls was built upon our earlier efforts to study pre-impact fall detection using the threshold-based algorithm [21]. Other than this, there has been no fall detection study, to the best of our knowledge, which has incorporated near-fall trials in their experiment protocol, despite the high frequency of near-fall events reported in older adults [2, 28]. It can be said that the fall detection algorithms that were designed without incorporating near-falls might result in a high number of fall positives and perform rather poorly in real-life fall and non-fall scenarios.

Finally, it can be said that the previous studies have primarily used threshold-based algorithms to detect impact and posture, as the most popular method for fall detection [7, 8, 11, 14, 22]. However, due to the improvement in data processing capabilities of devices, researchers, in the recent years, have started using advanced and complex fall detection algorithms (e.g., machine learning algorithms), showing the improved results [1, 18].

To address the aforementioned limitations of the experimental designs of previous studies, we first examined a library of video sequences of 227 real-world falls experienced by 130 older adults, collected as a part of an ongoing project by our research team to study the mechanism of falls in long-term-care facilities [27]. Based on this video evidence, we then simulated the 7 most common types of falls, along with 5 types of near-falls and 8 ADLs in laboratory experiments with young adults wearing tri-axial accelerometers. We compared the accuracy in distinguishing falls from near-falls and ADLs of the previously published threshold-based algorithms by Bourke et al. [7, 10] and Kangas et al. [17] to five novel machine learning-based fall detection algorithms.

2 Materials and methods

2.1 Participants

Ten healthy young adults participated in this study, ranging in age from 22 to 32 years (Mean = 26.6 years, SD = 2.8 years). All participants were students at Simon Fraser University and recruited through advertisements and flyers on university notice boards. The experiment protocol was approved by the Research Ethics Committee at Simon Fraser University, and all participants provided informed written consent.

2.2 Experimental design

Our laboratory recently described [27] the most common causes of imbalance and activities associated with falls in older adults in long-term care, based on the collection and analysis of video sequences of 227 real-life falls experienced by 130 older adults (of mean age 78 years, SD = 10). We found that 86 % of falls were collectively due to incorrect shift of bodyweight (41 %), trip (21 %), hit/bump (11 %), collapse/loss of consciousness (10 %), and slip (3 %). Common causes of incorrect shifting of bodyweight included missteps or cross-steps during walking, imbalance when rising from a chair, and imbalance while descending from standing to sitting.

Our experimental design (Fig. 1) included the above seven types of falls, along with five types of near-falls and eight ADLs. In near-fall trials, participants successfully recovered balance after experiencing slipping, tripping, hit or bump, imbalance due to a misstep or cross-step,

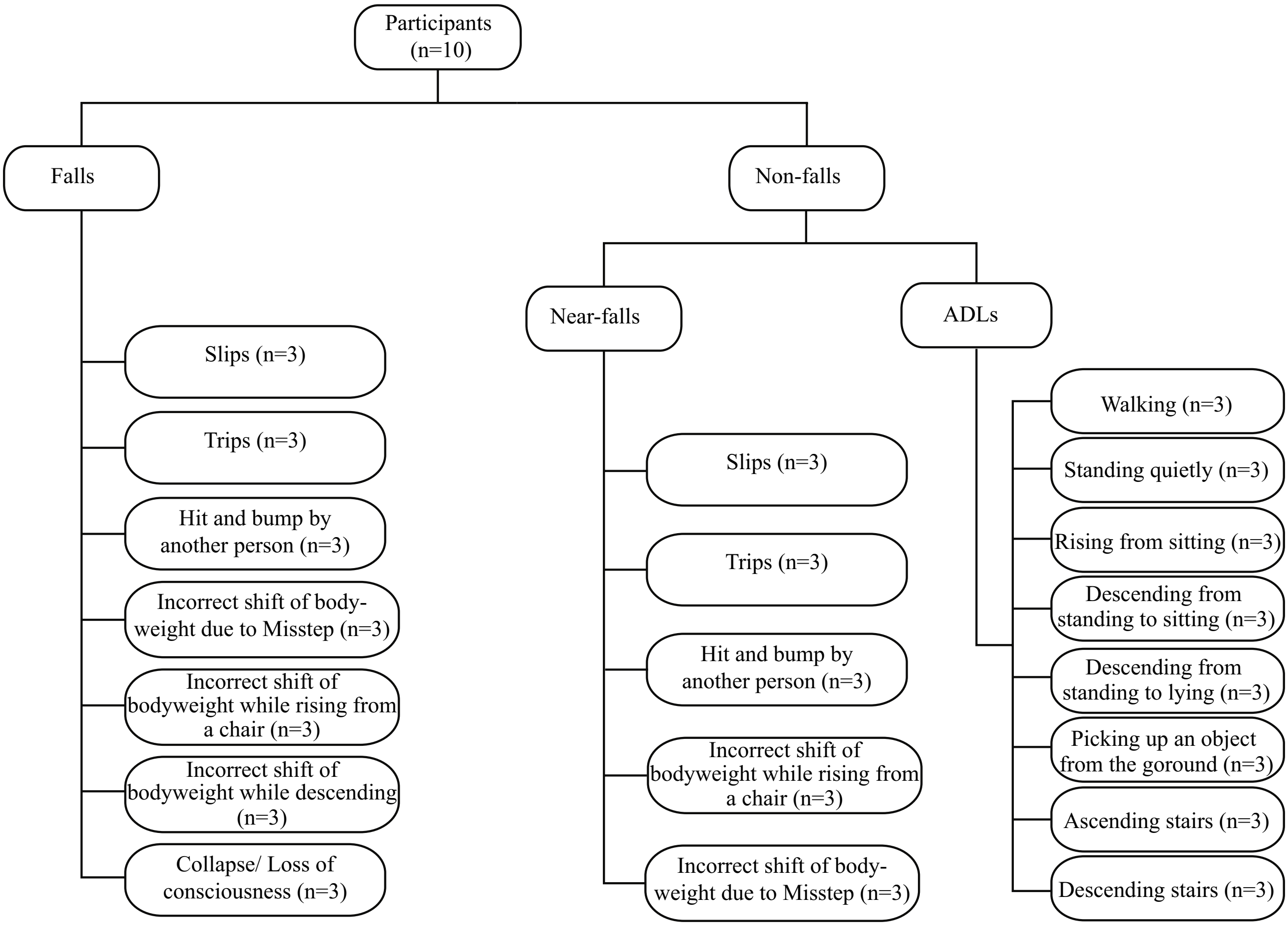

Participants (n=10)
Falls
Non-falls
Near-falls
ADLs
Slips (n=3)
Trips (n=3)
Hit and bump by another person (n=3)
Incorrect shift of body-weight due to Misstep (n=3)
Incorrect shift of bodyweight while rising from a chair (n=3)
Incorrect shift of bodyweight while descending (n=3)
Collapse/ Loss of consciousness (n=3)
Slips (n=3)
Trips (n=3)
Hit and bump by another person (n=3)
Incorrect shift of bodyweight while rising from a chair (n=3)
Incorrect shift of body-weight due to Misstep (n=3)
Walking (n=3)
Standing quietly (n=3)
Rising from sitting (n=3)
Descending from standing to sitting (n=3)
Descending from standing to lying (n=3)
Picking up an object from the goround (n=3)
Ascending stairs (n=3)
Descending stairs (n=3)

Fig. 1 Experiment protocol, indicating the 7 types of falls, 5 near-falls, and 8 activities of daily living (ADLs) simulated by each participant. Ten participants performed 3 repeated trials for each category, resulting in 210 falls, 150 near-falls, and 240 ADLs. For fall classification, near-falls and ADLs were combined in the same 'non-falls' category

and imbalance while rising from a chair. For ADLs, we included normal walking, standing quietly, rising from sitting, descending from standing to sitting, descending from standing to lying, picking up an object from the ground, ascending stairs, and descending stairs. In the experimental design, equal weight was given to each fall and non-fall scenario by having each participant simulating 3 repeated trials for each category, resulting in 210 falls, 150 near-falls, and 240 ADLs. For fall classification, near-falls and ADLs were combined into a 'non-falls' category.

2.3 Experiment protocol: simulation of falls, near-falls, and ADLs

During all fall and near-fall trials, the floor was covered with a 30-cm-thick gymnasium mattress of the type used to cushion falls in athletic activities, such as high jumps. We also inserted a 13 cm top layer of high-density ethylene vinyl acetate foam, so the composite structure was stiff enough to allow for stable standing and walking, but soft enough to reduce the forces during impact to a safe level. We also conducted training sessions with each participant, where we displayed one representative video for each real-life fall and instructed participants to fall in a fashion similar to that observed in the older adult in the video [4, 5]. For simulating near-fall and ADL events, we did not present participants with videos, but instructed them to execute non-fall scenarios as if there were frailer older adults.

In slipping trials, participants were instructed to walk up to a carpet placed over the gym mat. They were then made to slip by rapidly translating the carpet from underneath their feet. In some trials, participants were instructed to recover balance, while in others, they were asked to fall onto the gym mat. In tripping trials, participants were instructed to walk over the surface of the gym mat while wearing a tether at their ankle that suddenly became taught (and then released). Again, in some trials, participants were instructed to recover balance, while in others, they were instructed to fall onto the gym mat. During misstep/cross-step trials, participants walked forward with high variability in their step width and then took a narrow step or cross-step that caused loss of balance, followed by balance recovery or a fall. In rising from a chair trials, participants rose from sitting, then lost balance, and followed through by either recovering balance or falling onto the mat. In incorrect shift of bodyweight while descending from standing to sitting trials, participants missed sitting on a chair by glancing off the side, followed by a fall. In hit/bump trials, the investigator applied a sudden sideways force to the trunk of the participant (at the level of the T1 vertebrae) followed by balance recovery or a fall. Finally, loss of consciousness or collapse trials was simulated from an initial standing position with the participant acting out the 'legs giving way' or a 'collapse' by falling straight down. Trials that resulted in successful balance recovery were categorized as 'near-fall' events, while trials that resulted in falls were categorized as 'fall' events. For fall trials, with the exception of loss of consciousness or collapse (where participants were asked to fall straight down), no instruction was provided on the fall direction.

2.4 Data acquisition

In each trial, we recorded body kinematics using an array of seven tri-axial accelerometers (range ±6 g, Opal model, APDM Inc., Portland, OR), mounted bilaterally on the ankles and thighs, and at the anterior aspect of the waist, sternum, and head. Data were recorded at 128 Hz for a duration of 15 s per trial and streamed directly to a computer for storage and subsequent analysis.

3 Data analysis

3.1 Classification using threshold-based algorithms

We examined the accuracy of five threshold-based algorithms on our dataset. Each technique has been previously evaluated for accuracy through laboratory-based fall simulations [7, 10, 17].

The BourkeUFT [7] algorithm was implemented by thresholding the vector sum (VS—the root sum vector of the three squared accelerometer outputs) of the tri-axial accelerometer signal attached at the waist. The minimum value of the peak VS at the time of impact from all fall data was taken as the upper fall threshold (UFT). A fall was detected when the VS was over the UFT value.

The BourkeLFT [7] relies on the fact that the VS decreases during the descent phase of a fall (from its 1 g value during perfectly quiet standing) and has a zero value during pure free fall. The algorithm was implemented by evaluating the smallest value of the peak VS during the descent phase of falls (before impact). The smallest value of the VS during the descent from all fall data was taken as the lower fall threshold (LFT), and a fall was detected when the VS descended lower than the LFT value.

The Bourke4Phase [10] algorithm was based on the detection of four distinct phases of the fall (pre-fall, critical, impact, and post-fall) and the VS exceeding both the LFT and UFT thresholds [26].

The Kangas2Phase [17] fall detection considers both the impact and post-impact phases of the fall and utilizes tri-axial accelerometers placed at the wrist, waist, or head. We implemented the Kangas2Phase algorithm using signals from the waist-mounted accelerometer. A fall was registered when the sliding VS threshold (set at 2 g) was exceeded (in the impact phase) followed by confirmation of a lying posture 2 s after impact. A lying posture was detected if the average acceleration in a 0.4-s time interval was 0.5 g or lower.

Finally, the Kangas3Phase [17] fall detection algorithm considers pre-fall, impact, and post-fall posture recognition. As described by the authors, we detected the start of a fall when the VS went below 0.6 g. We then recorded the peak VS over the subsequent 1000 ms (during impact). Finally, as in the Kangas2Phase1 algorithm, we detected a lying posture 2 s after impact, if the average acceleration in a 0.4-s time interval was 0.5 g or lower.

3.2 Classification using machine learning algorithms

We examined the accuracy of five of the most common machine learning techniques used for human activity recognition, to distinguish falls from non-falls with our dataset: logistic regression (LR), decision tree (DT), naïve Bayes (NB), K-nearest neighbor (KNN), and support vector machines (SVM). The feature vector inputs to these routines were the means and variances of the X, Y, and Z accelerations from each trial acquired from the waist-mounted accelerometer, over a 2.5-s time window centered around each fall, near-fall, and ADL event (Fig. 2). Mean and variance were used as they are computationally inexpensive metrics that are commonly used as accelerometer signal features to input to machine learning algorithms for human activity classification [19]. The feature vector was then split into training and testing sets of equal size (i.e., 105 fall and 195 non-fall trials) by choosing data from the first five participants for training and the remaining five for testing. Furthermore, tenfold cross-validation was performed with the training data, and model parameters were selected that yielded the best cross-validation accuracy. The final models were then trained on the entire training set from the five participants.

3.3 Comparison of threshold-based and machine learning-based algorithms

In comparing the accuracies of the two methods (threshold-based and machine learning), we divided our analysis into two parts. In the first part, the input parameters required for the threshold-based algorithms (e.g., UFT and LFT) were derived from data recorded from falls and non-falls performed by participants 1–5. These algorithms were then tested on fall and non-fall data from the same five participants. This approach, where the algorithms were designed and tested on the same data, is identical to the approach used by previous authors [7, 10, 17]. To compare these results with the machine learning algorithms, we used the same dataset (participants 1–5) as the test dataset to calculate corresponding sensitivities and specificities. However, we trained each machine learning algorithm with fall and non-fall data recorded from participants 6–10. This is an accepted rule for an improved external validity of machine learning algorithms (that the training and learning datasets should be separate).

In the second part of the analysis, both our input parameter calculations and testing of threshold-based algorithms were conducted with data from participants 6–10. For machine learning algorithms, the roles of the training and test datasets from the 10 participants were reversed from their roles in part one. For each algorithm, we calculated sensitivities, specificities, false-positive rates, and false-negative rate as follows:

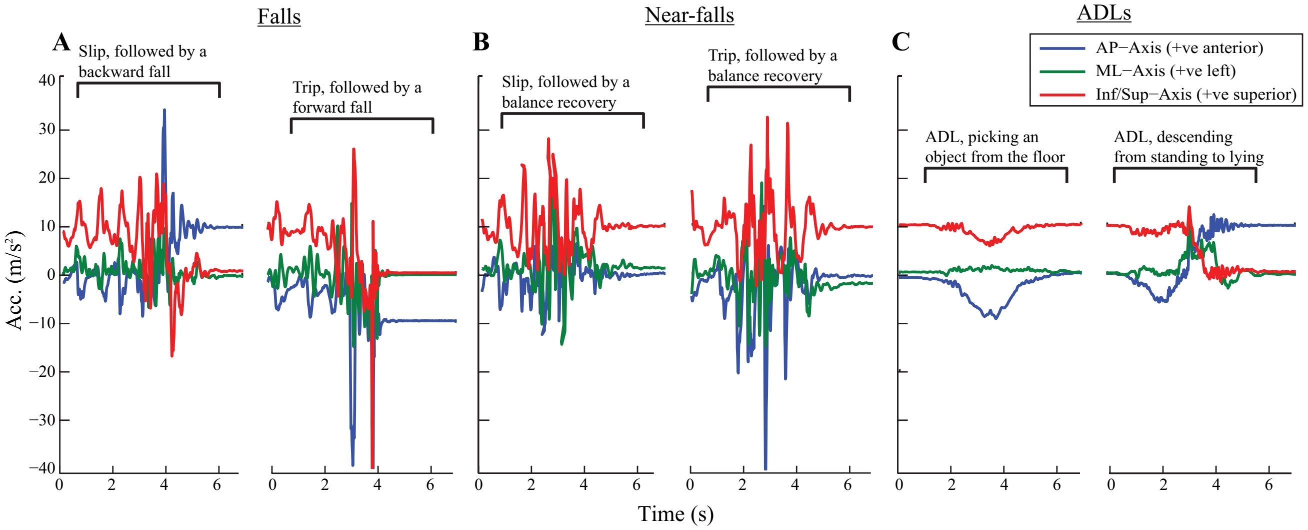


Fig. 2 Signals acquired from waist-mounted tri-axial accelerometers from a typical participant during **a** falls, **b** near-falls, and **c** activities of daily living (ADLs). *AP* acceleration in anterior/posterior direction, *ML* acceleration in medial/lateral direction, *Inf/Sup* acceleration in inferior/superior direction

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{1}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \tag{2}$$

$$\text{False positive rate} = \frac{\text{False positive}}{\text{False positive} + \text{True negative}} \tag{3}$$

$$\text{False negative rate} = \frac{\text{False negative}}{\text{False negative} + \text{True positive}} \tag{4}$$

We reported average sensitivities and specificities of each classification algorithm from the two-part analysis. We also examined rates of false-positive and false-negative errors with detection error trade-off plots. All data analyses were performed in MATLAB (R2014a, The MathWorks Inc.).

4 Results

The overall sensitivity and specificity of five threshold-based and five machine learning algorithms are shown in Fig. 3. We found that all five machine learning algorithms provided sensitivities and specificities of at least 90 %, while the sensitivities and specificities for the threshold-based algorithms varied from 0 to 100 %. Among the five machine learning algorithms, SVM provided the highest combination of sensitivity (96 %) and specificity (96 %) in distinguishing falls from non-falls. Among the threshold-based algorithms, Kangas3phase had the best classification with 94 % sensitivity and 94 % specificity.

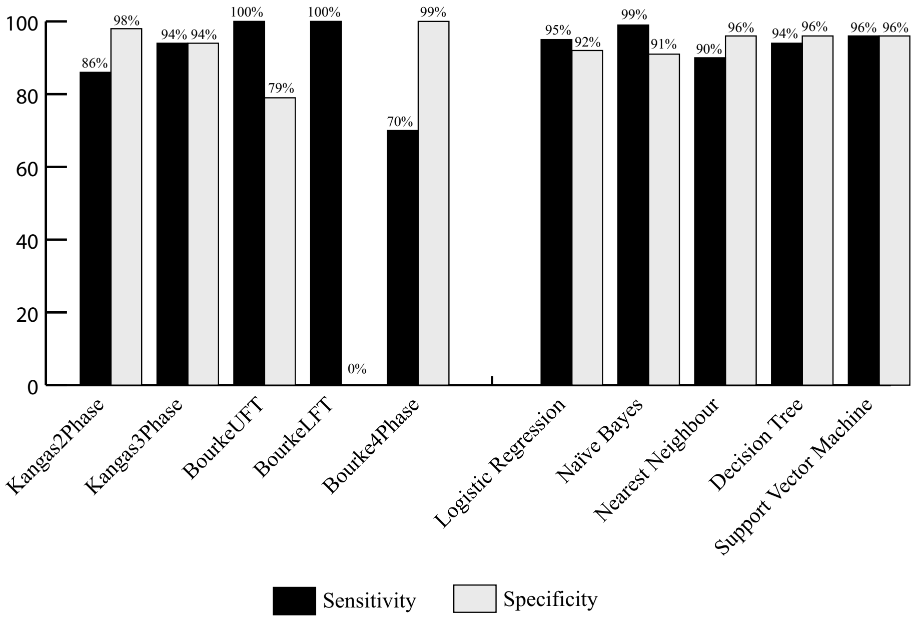
Figure 4 shows the trade-off between FP rate and FN rate for the five machine learning algorithms that we have employed in our study. If a stringent criterion was used to control the number of FPs, it was achieved at the expense of a high number of false negatives. Relaxing the criterion would reduce the number of false negatives, but would end up increasing the number of false positives. Overall, we found that the error rate was <10 % for all the five machine learning algorithms, wherein SVM, NB, and KNN performed slightly better than LR and DT.

Table 1 shows the number of trials, of the test set of 30 trials in each fall and non-fall category that were misclassified by each algorithm.

4.1 Missed falls

The number of false negatives (FNs) or missed falls by machine learning algorithms was spread across all seven types of falls. However, certain falls, such as trips, collapse/loss of consciousness, and ISBW, due to a misstep resulted in 2 or less FNs by all five machine learning algorithms, whereas slips (4 FNs by LR and 6 by DT), ISBW rising from a chair (4 FNs by KNN) and ISBW while descending (6 FNs by DT), were relatively more frequently missed. NB and SVM algorithms, which showed the highest sensitivity among all five machine learning algorithms, caused the least number of FNs (2 or less) in each of the seven types of falls. Among threshold-based algorithms, we found that BourkeUFT and BourkeLFT did not report any FNs in any of the seven fall categories. On the other hand, Bourke4Phase algorithm was the least effective in distinguishing falls and resulted in a total of 64 FNs, with 6 or more FNs in each

Fig. 3 Comparison of sensitivity and specificity between five threshold-based and five machine learning algorithms in distinguishing falls from non-falls. Overall, the machine learning algorithms performed better than the threshold-based algorithms, with SVM providing the highest sensitivity and specificity of 96 %



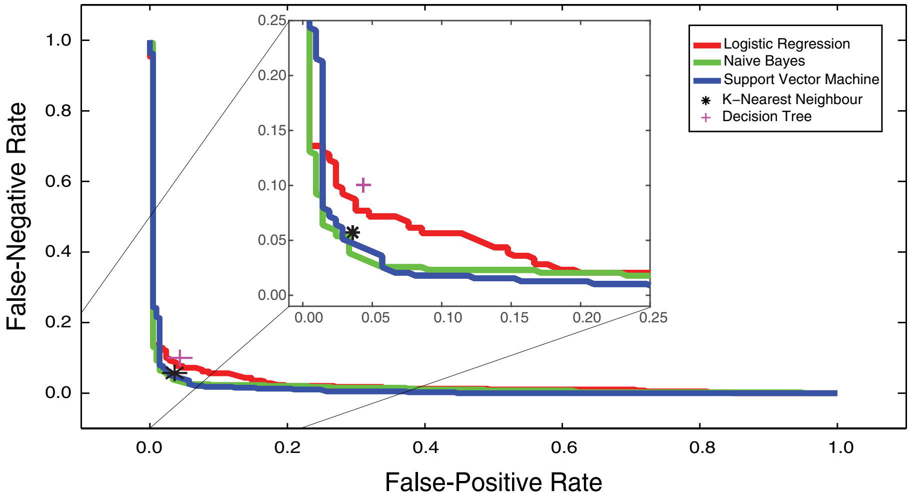


Fig. 4 Detection error trade-off plot. False-positive rate and false-negative rate of five machine learning algorithms are plotted. All algorithms show the error rate less than 10 %, wherein support vector machines (SVM), naïve Bayes (NB), and K-nearest neighbor (KNN) performed slightly better than logistic regression (LR) and decision tree (DT)

Table 1 Number and type of false negatives and false positives resulting from each of the five threshold-based and machine learning algorithms

	Threshold-based algorithms					Machine learning algorithms				
	BourkeUFT	BourkeLFT	Bourke4Phase	Kangas2Phase	Kanga3Phases	LR	NB	KNN	DT	SVM
False negatives										
Falls										
Slip	00	00	08	06	04	04	01	02	06	02
Trip	00	00	01	01	01	02	00	01	01	01
Hit/bump	00	00	06	06	00	02	02	02	03	02
Collapse	00	00	11	02	00	01	00	01	02	00
ISBW_Misstep	00	00	11	08	03	01	00	01	00	01
ISBW_Rising from a chair	00	00	13	07	01	01	00	04	03	01
ISBW_Descending	00	00	14	00	04	00	00	02	06	01
False positives										
Near-falls										
Slip	14	30	00	00	04	02	04	02	04	02
Trip	17	30	01	03	08	01	06	04	04	03
Hit/bump	18	30	01	01	01	02	03	02	02	02
ISBW_Misstep	12	30	00	01	04	04	04	04	02	03
ISBW_Rising from a chair	09	30	00	03	05	04	02	02	02	02
ADLs										
Normal walking	00	30	00	00	00	00	00	00	00	00
Standing quietly	00	30	00	00	00	00	00	00	00	00
Rising from a chair	00	30	00	00	00	00	00	00	00	00
Descending from standing to sitting	01	30	00	00	00	00	00	00	00	00
Descending from standing to lying	04	30	00	00	01	12	15	00	01	02
Picking an object from the ground	00	30	00	00	00	06	00	00	02	00
Ascending stairs	00	30	00	00	00	00	00	00	00	00
Descending stairs	09	30	00	00	00	00	00	00	00	00

ISBW incorrect shift of bodyweight, *ADLs* activities of daily living, *LR* logistic regression, *NB* naïve Bayes, *KNN* K-nearest neighbor, *DT* decision tree, *SVM* support vector machine of the fall category with the exception of trips. Between the two Kangas' algorithms, Kangas3Phase was more effective and resulted in fewer FNs as follows: 4 FNs each in slips and ISBW while descending, 3 FNs in ISBW due to a misstep, 1 FN in each in trips, and ISBW while rising from a chair. Two fall categories, hit/bump, and collapse/loss of consciousness did not result in any FNs at all by Kangas3Phase algorithm.

4.2 Near-falls as a cause of false positives

The number of false positives (FPs) due to near-falls by machine learning algorithms was also distributed across all five of the near-fall categories with no obvious trend. For example, with the exception of hip/bump, at least one of the remaining four near-fall categories resulted in at least 4 FPs for LR, NB, KNN, and DT algorithms. Among machine learning algorithms, SVM showed the least number of FPs causing 3 or less false alarms in each of the five near-fall categories. For threshold-based algorithms, we found that BourkeUFT and BourkeLFT resulted in a maximum number of FPs. BourkeLFT, in particular, was unable to distinguish between near-falls and falls from all ten participants. The remaining three threshold-based algorithms, however, showed relatively fewer number of FPs across all near-fall categories. Among these three algorithms, Bourke4phase resulted in the least number of FPs, i.e., only 1 FP each in near-fall due to trips and hit/bump.

4.3 ADLs as a cause of false positives

Two out of eight ADL categories caused FPs when analyzed by machine learning algorithms. Between the two, descending from standing to lying activity was more frequently misclassified and caused one or more false alarms by all machine learning algorithms with the exception of KNN. The second activity that was misclassified was picking an object from the ground, causing 6 FPs for NB and 2 FPs for DT. Among machine learning algorithms, KNN was found to be the most effective with zero FPs, followed by SVM and DT, which resulted in 2 FPs and 3 FPs, respectively. For threshold-based algorithms, BourkeLFT was unable to differentiate between falls and ADL trials and misclassified all of the ADL trials and categories. For remaining four algorithms, descending stairs caused the most false alarms (9 FPs) followed by descending from standing to lying (5 FPs) and descending from standing to sitting (1 FP). Among five threshold-based algorithms, both Bourke4Phase and Kangas2Phase were the most effective algorithms and did not cause any false alarm for ADLs.

5 Discussion

In this study, we conducted laboratory-based experiments and recorded accelerations from waist-mounted sensors during simulation of common falling scenarios among older adults, along with ADLs and near-falls. We then evaluated the accuracy of five threshold-based and five machine learning algorithms in distinguishing falls from non-fall trials from waist acceleration data. We found that the best-performing machine learning algorithm (SVM) provided 96 % sensitivity and 96 % specificity, while the best-performing threshold-based algorithm (Kangas3Phase) provided 94 % sensitivity and 94 % specificity.

The sensitivities and specificities we found for threshold-based algorithms were substantially lower than what were reported previously. We also found that on our laboratory database, the threshold-based algorithms (with the exception of Kangas3Phase) that provided high sensitivities showed low specificities and vice versa. For example, Bourke4Phase algorithm proposed a set of fixed threshold values for features extracted from acceleration signals and reported 100 % sensitivity and specificity in distinguishing falls from non-falls [10]. However, when the same thresholds were applied to our dataset obtained from different participants and using different data collection strategies, the accuracy of Bourke4Phase algorithm dropped to 70 % sensitivity and 99 % specificity. This issue is perhaps due to the threshold values set by the authors, which do not generalize well enough with different participants and data collection strategies. On the other hand, a fundamental goal of machine learning algorithms was to generalize a classification model beyond the examples in the training set. This is often achieved by optimizing the parameters of the machine learning algorithm through cross-validation using training data and then evaluating on the test data unseen by the classification algorithm during model training. This procedure in machine learning algorithms ensures better external validity than threshold-based algorithm even if both algorithms provide same accuracy on a given dataset.

The threshold-based algorithms, which did not incorporate post-impact fall signal analysis, such as BourkeUFT and BourkeLFT (discussed later), resulted in causing relatively greater number of false positives as opposed to the algorithms which did analyze the post-impact phase of falls, such as Kangas2Phase, Kangas3Phase, and Bourke4Phase algorithms. The number of false alarms by Bourke4Phase and Kangas2Phase algorithms was even lower than what was provided by the five machine learning algorithms. Therefore, despite the fact that machine learning algorithms provided higher combinations of sensitivity and specificity than threshold-based algorithms, further improvement in results can be achieved either by training another classification model on the signals from post-impact fall phase or by designing a fusion algorithm, i.e., combining machine learning algorithm (for descend and impact phase fall analysis) and threshold-based algorithm (for post-impact fall phase analysis).

BourkeUFT and BourkeLFT may not be the preferred algorithms for fall detection because of their very low specificity.

This is primarily due to the unusual method (as described in Sect. 3.1) employed by the authors to set threshold values to separate falls from non-fall activities. Since the threshold values in BourkeUFT and BourkeLFT were set using fall data—without analyzing the VS of the tri-axial accelerometer signals recorded from non-fall trials—the algorithms were able to detect all falls but resulted in a high number of false positives. In particular, BourkeLFT was unable to differentiate between falls and non-fall trials at all (0 % specificity). Since the LFT was set from the fall that showed largest value of VS in all falls during the descend phase (before impact), it is possible that certain falls with relatively lower vertical velocities may yield a VS that is larger in value than those resulting from normal daily activities. In cases like such, where there exists no clear boundary that can split the data, machine learning algorithms often perform better than threshold-based algorithms. For example, in non-separable data (which is often the case in fall and non-fall datasets), the SVM classification technique uses a non-negative 'slack variable' which measures the degree of misclassification of the data while choosing a boundary that splits fall and non-fall examples as clearly as possible, hence resulting in more optimized classification than by conventional threshold-based methods.

Near-fall trials, which often produced acceleration signals similar to falls before the subject recovered from loss of balance (Fig. 2), were the primary cause of FPs in threshold-based and machine learning algorithms. While the near-falls were identified more accurately by the three threshold-based algorithms that included post-fall analysis (Bourke4Phase, Kangas2Phase, and Kangas3Phase), the fall classification accuracy of these algorithms was decreased considerably. This was, perhaps, due to unoptimized threshold values, which were often set through observational analysis. These algorithms might result in reduced numbers of FPs and FNs if their threshold values are selected through an optimization method, such as cross-validation or receiver operator characteristic (ROC) curve.

Furthermore, sensitivity and specificity, along with accuracy, are the commonly used performance evaluating parameters for fall detection algorithms. However, in case of a dataset where the presence of one class is substantially more than the other, these parameters do not provide a true picture of the performance of the algorithms being evaluated. For example, in a dataset of 1200 samples (corresponding to 10 h of data windowed at 30 s) of which 50 correspond to falls, the algorithm that always predicts non-fall would yield 96 % accuracy. On the other hand, 96 % sensitivity and specificity (which are acceptable values for a test) would mean that an algorithm would miss two falls and every hour five false alarms would go off, which are far too many for care providers who may perceive such a device as ineffective. Therefore, in order to evaluate the performance of proposed algorithms, future studies should consider other performance-measuring parameters such as FP rate and FN rate.

Finally, we recognize the limitations of our study. Due to safety concerns, all fall trials were performed by young adults under controlled laboratory conditions and atop gymnasium mats. While there are important differences between falling patterns from typical laboratory studies of young participants compared to real-life falls among older adults [27], we attempted to minimize these differences by conducting training sessions with each participant, where we displayed representative videos of real-life falls and instructed them to fall in a fashion similar to that observed in the older adult in the video. Furthermore, in order to minimize the effect of surface stiffness on falling behavior, the top 13 cm layer of the mats consisted of high-density ethylene vinyl acetate foam. This provided the composite structure with a stiffness high enough to allow for stable standing and walking, but soft enough to reduce impact forces to a safe level. Another limitation of the study was that in our laboratory experiment, we gave equal weight to each type of fall and non-fall category by recording three trials in each category from every participant, whereas, in real life, occurrences of certain type of falls and ADLs are more frequent than others. Similarly, fall detection problem in real world has highly skewed dataset with a very small number of fall instances as opposed to ADLs. However, machine learning algorithms trained using such unbalanced datasets tend to generate trivial models that almost always predict the majority class [32]. Therefore, in order to balance different costs for false positive and false negative, we over-sampled the minority class by recording 210 fall trials and under-sampled the majority class by recording 390 non-fall trials (class ratio of 1:1.8). Finally, although machine learning algorithms showed in general better sensitivity and specificity trade-off with respect to threshold-based techniques, we do realize an important limitation of the study while comparing the accuracy between the two approaches. The machine learning parameters were optimized, trained, and tested on a dataset acquired in our laboratory with similar experimental settings for all subjects. On the other hand, the threshold parameter values used in this study (for Bourke4Phase, Kangas2Phase, and Kangas3Phase) were the values recommended by their respective authors obtained from different datasets and acquired under their specific laboratory settings and experimental protocols. Therefore, the comparison of accuracy between the machine learning and threshold-based algorithms might be biased toward the former approach. However, in order to minimize the possible biasness of the machine learning algorithm to our dataset, we split the entire dataset into two equal-sized partitions: one for training and the other one for testing to assess how effective machine learning algorithms were to unseen test data. A more comprehensive validation of the accuracy of the algorithms, however, will come from real-life datasets obtained in the target population as they go about their daily routine.

In summary, we evaluated five machine learning and five threshold-based algorithms in this study. Machine learning algorithms provided greater overall sensitivity and specificity than the threshold-based algorithms. Among machine learning algorithms, SVM provided the greatest sensitivity and specificity of 96 % in distinguishing falls from non-falls. Our results provide a template for further improvement in designing a robust fall detection system, which is necessary for the development of 'smart' personal emergency response system that can automatically place a call for help in case of a fall to prevent long lies. Future studies should also examine whether system accuracy in distinguishing falls from non-falls using machine learning algorithms can be improved through subsequent post-fall signal analysis, including additional sensors (gyroscopes, blood pressure monitors or pressure sensors, etc.) and taking into consideration subject anthropometric information (height, mass, age, etc.) in model training. The extent to which the accuracy of the results transfers to unexpected falls on hard surface by older adults will ultimately be addressed by testing the system with sensor signals obtained from older adults as they go about their daily activities.

Acknowledgments This work was supported by team grants from the Canadian Institutes of Health Research (Funding Reference Numbers: AMG-100487 and TIR-103945). SNR was also supported by the Canada Research Chairs program. The funding agency had no direct role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; or the preparation, review, or approval of the manuscript.

References

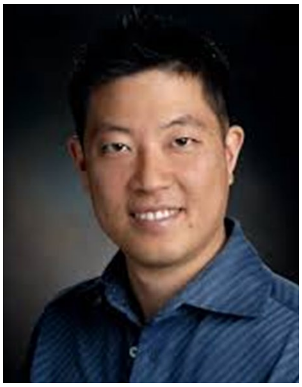
1. Albert MV, Kording K, Herrmann M, Jayaraman A (2012) Fall classification by machine learning using mobile phones. PLoS One 7(5):e36556
2. Arnold CM, Faulkner RA (2007) The history of falls and the association of the timed up and go test to falls and near-falls in older adults with hip osteoarthritis. BMC Geriatr 7:17
3. Aziz O, Park EJ, Mori G, Robinovitch SN (2012) Distinguishing near-falls from daily activities with wearable accelerometers and gyroscopes using support vector machines. In: Conference proceedings of IEEE engineering in medicine biology society 2012, pp 5837–5840
4. Aziz O, Park EJ, Mori G, Robinovitch SN (2014) Distinguishing the causes of falls in humans using an array of wearable tri-axial accelerometers. Gait Posture 39(1):506–512
5. Aziz O, Robinovitch SN (2011) An analysis of the accuracy of wearable sensors for classifying the causes of falls in humans. IEEE Trans Neural Syst Rehabil Eng 19(6):670–676
6. Bagala F, Becker C, Cappello A, Chiari L, Aminian K, Hausdorff JM, Zijlstra W, Klenk J (2012) Evaluation of accelerometer-based fall detection algorithms on real-world falls. PLoS One 7(5):e37062
7. Bourke AK, O'Brien JV, Lyons GM (2007) Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. Gait Posture 26(2):194–199
8. Bourke AK, O'Donovan KJ, Olaighin G (2008) The identification of vertical velocity profiles using an inertial sensor to investigate pre-impact detection of falls. Med Eng Phys 30(7):937–946
9. Bourke AK, Scanaill CN, Culhane KM, Brien JVO, Lyons GM (2006) An optimum accelerometer configuration and simple algorithm for accurately detecting falls. In: Proceedings of the 24th IASTED international conference on Biomedical engineering. ACTA Press, Innsbruck
10. Bourke AK, van de Ven P, Gamble M, O'Connor R, Murphy K, Bogan E, McQuade E, Finucane P, Olaighin G, Nelson J (2010) Evaluation of waist-mounted tri-axial accelerometer based fall-detection algorithms during scripted and continuous unscripted activities. J Biomech 43(15):3051–3057
11. Bourke AK, van de Ven PW, Chaya AE, GM OL, Nelson J (2008) Testing of a long-term fall detection system incorporated into a custom vest for the elderly. In: Conference Proceedings of IEEE engineering in medicine and biology society 2008, pp 2844–2847
12. Boyle J, Karunanithi M (2008) Simulated fall detection via accelerometers. In: Conference Proceedings of IEEE engineering in medicine and biology society 2008, pp 1274–1277
13. Chao PK, Chan HL, Tang FT, Chen YC, Wong MK (2009) A comparison of automatic fall detection by the cross-product and magnitude of tri-axial acceleration. Physiol Meas 30(10):1027–1037
14. Chen J, Kwong K, Chang D, Luk J, Bajcsy R (2005) Wearable sensors for reliable fall detection. In: Conference proceedings of IEEE engineering in medicine and biology society 4, pp 3551–3554
15. Diaz A, Prado M, Roa LM, Reina-Tosina J, Sanchez G (2004) Preliminary evaluation of a full-time falling monitor for the elderly. In: Conference proceedings of IEEE engineering in medicine and biology society 3, pp 2180–2183
16. Herman M, Gallagher E, Scott VJ (2006) The evolution of seniors falls prevention in British Columbia. BC Ministry of Health Services, Victoria
17. Kangas M, Konttila A, Lindgren P, Winblad I, Jamsa T (2008) Comparison of low-complexity fall detection algorithms for body attached accelerometers. Gait Posture 28(2):285–291
18. Kau LJ, Chen CS (2015) A smart phone-based pocket fall accident detection, positioning, and rescue system. IEEE J Biomed Health Inform 19(1):44–56
19. Kern N, Schiele B, Schmidt A (2007) Recognizing context for annotating a live life recording. Personal Ubiquitous Comput. 11(4):251–263
20. King MB, Tinetti ME (1995) Falls in community-dwelling older persons. J Am Geriatr Soc 43(10):1146–1154
21. Lee JK, Robinovitch SN, Park EJ (2015) Inertial sensing-based pre-impact detection of falls involving near-fall scenarios. IEEE Trans Neural Syst Rehabil Eng 23(2):258–266
22. Lindemann U, Hock A, Stuber M, Keck W, Becker C (2005) Evaluation of a fall detector based on accelerometers: a pilot study. Med Biol Eng Comput 43(5):548–551
23. Mallinson WJ, Green MF (1985) Covert muscle injury in aged patients admitted to hospital following falls. Age Ageing 14(3):174–178
24. Nevitt MC, Cummings SR, Kidd S, Black D (1989) Risk factors for recurrent nonsyncopal falls. A prospective study. JAMA 261(18):2663–2668
25. Noury N, Galay A, Pasquier J, Ballussaud M (2008) Preliminary investigation into the use of autonomous fall detectors. In: Conference proceedings of IEEE engineering in medicine and biology society 2008, pp 2828–2831
26. Noury N, Rumeau P, Bourke AK, ÓLaighin G, Lundy JE (2008) A proposal for the classification and evaluation of fall detectors. IRBM 29(6):340–349
27. Robinovitch SN, Feldman F, Yang Y, Schonnop R, Leung PM, Sarraf T, Sims-Gould J, Loughin M (2013) Video capture of the

circumstances of falls in elderly people residing in long-term care: an observational study. Lancet 381(9860):47–54

28. Srygley JM, Herman T, Giladi N, Hausdorff JM (2009) Self-report of missteps in older adults: a valid proxy of fall risk? Arch Phys Med Rehabil 90(5):786–792
29. Tinetti ME, Williams CS (1998) The effect of falls and fall injuries on functioning in community-dwelling older persons. J Gerontol A Biol Sci Med Sci 53(2):M112–M119
30. Vellas B, Cayla F, Bocquet H, de Pemille F, Albarede JL (1987) Prospective study of restriction of activity in old people after falls. Age Ageing 16(3):189–193
31. Weiss A, Shimkin I, Giladi N, Hausdorff JM (2010) Automated detection of near falls: algorithm development and preliminary results. BMC Res Notes 3:62
32. Weiss GM, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intell Res 19(1):315–354
33. Wild D, Nayak US, Isaacs B (1981) How dangerous are falls in old people at home? Br Med J (Clin Res Ed) 282(6260):266–268

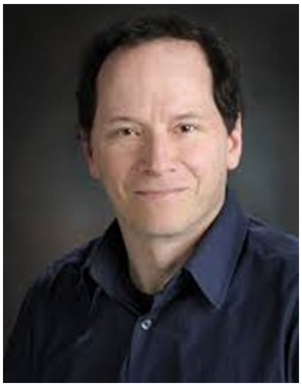
Omar Aziz received the B.Eng. degree in 2005 in Mechatronics Engineering from the National University of Science and Technology, Islamabad, Pakistan, and the MASc. degree in 2009 and the Ph.D. degree in 2015 from the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is a postdoctoral fellow at Biomechatronic Systems Laboratory at Simon Fraser University, Surrey, BC, Canada. His research interests include: wearable sensors and their applications to human movement analysis, human activity classification, injury prevention in older adults, biomechanics, robotics, and control. At Simon Fraser University, his research focuses on the development of wearable sensor systems and corresponding machine learning algorithms to monitor activity patterns, to detect falls pre- and post-impact, to determine the causes of falls, and to distinguish the events of imbalance and near-falls from activities of daily living.

Magnus Musngi received the B.Sc. degree in Computer Engineering from De La Salle University, Manila, Philippines, in 2009, and the B.ASc. degree from School of Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC, Canada, in 2015, where he is also currently pursuing a M.A.Sc. degree. He is a research assistant at Biomechatronic Systems Laboratory, Simon Fraser University, Surrey, BC, Canada. His research focuses on developing fall detection algorithms using inertial sensors. His research interests are in estimating and classifying motion using inertial sensors.



Edward J. Park received the B.A.Sc. degree in 1996 from the University of British Columbia, and the M.A.Sc. degree in 1999 and the Ph.D. degree from University of Toronto, Canada, in 1995. He is a professor at School of Mechatronic Systems Engineering and director of the Biomechatronic and Systems Laboratory. He is also a member of the Faculty of Health Sciences and the Associate Dean of Faculty of Applied Sciences at Simon Fraser University. His research interests include the following: biomechatronics and biomedical technologies for life sciences, rehabilitation and medicine, and mechatronics applied to next-generation vehicular, robotic, and space systems.

Greg Mori received the B.Sc. Hon. degree in 1999 in Computer Science and Mathematics from University of Toronto, Canada, and the Ph.D. degree in Computer Science in 2004 from the University of California, Berkeley. He is a professor and the director of School of Computing Science, Simon Fraser University, Burnaby, Canada. He was a visiting scientist at Google in Mountain View, California, in 2014–2015. His research interests include the following: computer vision, machine learning, video analysis, human activity recognition, human body pose estimation, pedestrian detection and tracking, and object recognition. He has made significant contributions toward solving the problems of human pose estimation and human action recognition.



Stephen N. Robinovitch received the B.App.Sc. degree in 1988 from the University of British Columbia, the M.S. degree in 1990 from Massachusetts Institute of Technology, and the Ph.D. degree in 1995 from Harvard/Massachusetts Institute of Technology. He is a professor and Canada Research Chair in Injury Prevention and Mobility Biomechanics at Simon Fraser University. His research focuses on improving our understanding of the cause and prevention of fall-related injuries (especially hip fracture) in older adults, through laboratory experiments, mathematical modeling, field studies in residential care facilities, and product design.