



OPEN

XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes

Byungjoo Noh¹, Changhong Youm²✉, Eunkyong Goh³, Myeounggon Lee², Hwayoung Park², Hyojeong Jeon⁴ & Oh Yoen Kim⁵

This study aimed to identify the optimal features of gait parameters to predict the fall risk level in older adults. The study included 746 older adults (age: 63–89 years). Gait tests (20 m walkway) included speed modification (slower, preferred, and faster-walking) while wearing the inertial measurement unit sensors embedded in the shoe-type data loggers on both outsoles. A metric was defined to classify the fall risks, determined based on a set of questions determining the history of falls and fear of falls. The extreme gradient boosting (XGBoost) model was built from gait features to predict the factor affecting the risk of falls. Moreover, the definition of the fall levels was classified into high- and low-risk groups. At all speeds, three gait features were identified with the XGBoost (stride length, walking speed, and stance phase) that accurately classified the fall risk levels. The model accuracy in classifying fall risk levels ranged between 67–70% with 43–53% sensitivity and 77–84% specificity. Thus, we identified the optimal gait features for accurate fall risk level classification in older adults. The XGBoost model could inspire future works on fall prevention and the fall-risk assessment potential through the gait analysis of older adults.

Falls are among the most common causes of injury, severe health problems, and even death in older adults¹. Numerous studies have revealed a relationship between falls and risk factors such as advanced age², declined cognitive function³, strength deficit, gait abnormalities, and reduced balance⁴. In particular, gait abnormalities in aging, including slow walking speed, greater gait variability, and shorter steps, are considered one of the greatest risk factors for falls^{5–8}. Furthermore, gait abnormalities or decreased gait ability decisively imply a reduced physical fitness as a result of aging⁹, which may cause a falls. Thus, the underlying causes of fall must be identified to predict their risk. In addition, it is necessary to identify in advance the influential predictor factors affecting falls through a gait performance test and use them as fundamental data to prevent falls. Furthermore, novel methods are required to overcome the limitations of existing studies.

The machine learning (ML) techniques have gained attention for addressing the clinically relevant spatiotemporal gait parameters for disease classification¹⁰. The ML techniques use features extracted from a set of clinically relevant data, allowing computer algorithms to form a predictive model¹¹. In addition, the ML algorithm can be used to extract the optimal features affecting the risk of falls from the gait features, which measured a more continuative state for longer durations using the inertial measurement unit (IMU) sensors. In this study, we applied the ML technique using the extreme gradient boosting (XGBoost) algorithm¹², a decision tree-based ensemble ML technique. The XGBoost minimizes the residuals of the models and increases the predictive power by combining weak learners¹³. Using XGBoost, we expect to distinguish the spatiotemporal gait parameters from high and low fall-risk level subjects. To our knowledge, only a few prediction researchers have studied the risk of falls and extracted the essential factors using ML techniques^{14,15}. A model using support vector machine with parameter tuning was proposed in¹⁴. The model was developed to discriminate the balancing problems in older adults; thus, the model does not predict the risk of falling. Moreover, models based on artificial neural networks were developed to examine the efficiency in classifying with or without recurrent falling utilizing a set of clinical characteristics corresponding to risk factors of falls in the older adults¹⁵. However, the developed models did not consider extracting the essential gait parameters. Furthermore, no studies have used the XGBoost algorithm to classify high and low fall risk levels objectively based on their gait spatiotemporal features.

¹Department of Kinesiology, Jeju National University, Jeju, Republic of Korea. ²Department of Health Sciences, The Graduate School of Dong-A University, Busan, Republic of Korea. ³Human Life Research Center, Dong-A University, Busan, Republic of Korea. ⁴Department of Child Studies, Dong-A University, Busan, Republic of Korea. ⁵Department of Food Science and Nutrition, Dong-A University, Busan, Republic of Korea. ✉email: chyoun@dau.ac.kr

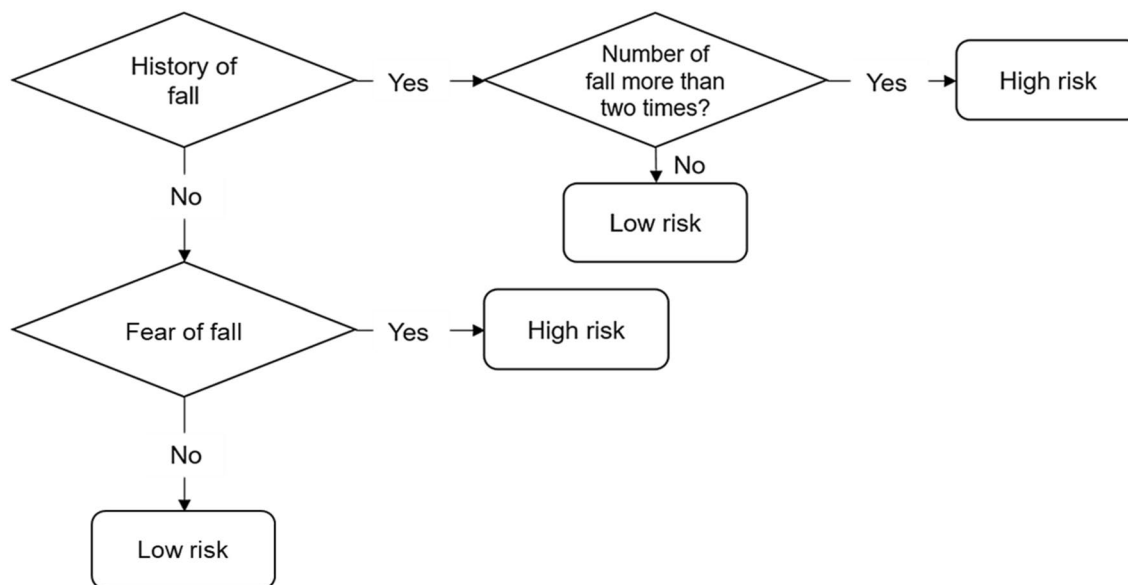


Figure 1. Fall levels definition. This metric was defined to classify the high and low fall risk based on the history of falls and fear of falls questionnaire²¹.

Spatiotemporal parameters of gait have been used for the classification the falls using ML algorithms¹⁶. However, previous studies have compared ML models' performances considering relatively fewer steps in their gait assessment, such as the Timed up and go test. The gait assessment with relatively fewer steps may not yield similar results to an actual walking environment. Research on gait with more continuative states for longer durations strengthens the reliability of spatiotemporal variables using the wearable sensor technology¹⁷. IMU sensors can measure the gait outside of a laboratory and real-world at low cost than motion capture system with continuative states for longer steps¹⁸. Thus, gait analysis with numerous consecutive steps is necessary for improving the reliability of gait variables. Additionally, this study was focused on identifying the optimal features of gait parameters to predict the fall risk level in older adults. Therefore, we used the XGBoost algorithm of ML on gait performance tests with speed modification to identify fall risk levels in older adults and define optimal gait parameters.

Methods

Participants. Participants were recruited from a community-wide survey in Busan Metropolitan City. Participants satisfied the following exclusion criteria: (1) they were unable to walk without any support, and (2) they have a history of musculoskeletal injuries or neurophysiological problems in the past six months. In total, 746 older adults with ages ranging from 63 to 89 years participated in the study. All methods were performed in accordance with the relevant guidelines and regulations. All participants signed their informed consent after reading all the study details. This study was approved by the Institutional Review Board of Dong-A University (IRB number: 2-104709-AB-N-01-201808-HR-023-02).

Instrumentation. Gait performance data were collected as previously described by Noh et al.¹⁸ and Lee et al.¹⁹. Gait performance tests were evaluated using a gait analysis system (DynaStab, JEIOS, Busan, South Korea), including shoe-type data loggers (Smart Balance SB-1, JEIOS, Busan, South Korea) and embedded IMU sensors (IMU-3000, InvenSense, San Jose, CA, USA) on both outsoles. Gait performance data was collected by triaxial accelerations with up to ± 6 g and triaxial angular velocities with up to $\pm 500^\circ\text{s}^{-1}$ along three orthogonal axes. Gait performance data were collected at a sampling frequency of 100 Hz using a data acquisition system (Smart Balance version 1.5, JEIOS, Busan, South Korea). Various sizes of shoe-type data loggers were available for each participant. The international physical activity questionnaire-short form was used to estimate their habitual physical activity (PA) levels with respect to the metabolic equivalents (METs/week)²⁰.

Assessment of fall level. The fall was defined as “an unexpected event in which the person comes to rest on the ground, floor, or lower level.”²¹. Silva et al.²² defined a metric to classify the fall risks. In this regard, participants were asked questions concerning their history of falls (Q. Have you fallen in the last 6 months?), number of falls (Q. How many times did you fall in the last 6 months?), and fear of falls (Q. Are you afraid of falling?). Subsequently, a fall risk level was classified along with a metric of fall levels definition, which indicates if the person shows more or less probability to fall (Fig. 1). The low-risk group represented 62% of the dataset, composed of 456 participants. The high-risk group represented approximately 38% of the dataset, including the remaining 290 participants. This distribution follows a similar agreement with the fall incidence in older adults, which is less than 30%²³.

Assessment of gait performance. Three gait performance tests were performed on a straight 20-m over-ground walkway with gait speed modification (slower (−20% of preferred), self-preferred, and faster (+20% of preferred) speeds)²⁴. Before the gait performance test, the preferred walking speed was defined using a metronome (beats/min). Participants were asked to perform the walking as close as possible to the targeted slower and faster-walking speed by a metronome. Verbal or visual instructions were provided to perform overground walking. They practiced walking at three-speed conditions using the metronome as a familiarization session for approximately 10 min.

Data analysis. The gait data were filtered using a second-order Butterworth low-pass filter with a cutoff frequency of 10 Hz^{19,25}. Heel strikes and toe-off of gait events were detected when the linear acceleration along the anteroposterior axis and the vertical axis reached its maximum value, respectively^{19,25}. We excluded the acceleration and deceleration step periods of the gait performance test to analyze in the steady-state condition. We calculated the spatiotemporal parameters [i.e., walking speed, stride length, cadence, stance phase, stride time, and gait asymmetry (GA)]. The walking speed was calculated using the formula: walking speed = walking distance (m)/walking duration (s). In addition, the cadence, stride time, and stride length were calculated using the formula²⁶. The n is a total number of heel strike (HS) events.

$$\text{Cadence (steps/min)} = \frac{\text{Step counted} \times 60}{\text{Walking duration}} \quad (1)$$

$$\text{Stride time (s)} = \frac{1}{n-1} \sum_{n=1}^{n-1} \frac{HS_{n+1} - HS_n}{100} \quad (2)$$

$$\text{Stride length (m)} = \frac{1}{n-1} \sum_{n=1}^{n-1} (HS_{n+1} - HS_n) \times \text{walking speed} \quad (3)$$

The stance phase was also calculated according to the two kinds of formulas²⁶. If the $TO_n > HS_n$, then the stance phase was calculated as follow:

$$\text{Stance phase (\%)}_{\text{case1}} = \frac{1}{n-1} \sum_{n=1}^{n-1} \left(\frac{TO_n - HS_n}{HS_{n+1} - HS_n} \times 100 \right) \quad (4)$$

The n is the total number of toe off (TO) events. For instance, if the total number of TO is lower than HS events (e.g., 14 vs. 15), the n is a total number of TO events ($n = 14$), whereas if the total number of TO is equal to the HS events (e.g., 14), then the n is a total number of HS or TO events ($n = 14$). Alternatively, if the $TO_n > HS_n$, then the stance phase was calculated as follow:

$$\text{Stance phase (\%)}_{\text{case2}} = \frac{1}{n-1} \sum_{n=1}^{n-1} \left(\frac{TO_{n+1} - HS_n}{HS_{n+1} - HS_n} \times 100 \right) \quad (5)$$

If the total number of TO is greater than HS events (e.g., 13 vs. 12), the n is a total number of HS events ($n = 12$), whereas if the total number of TO is equal to the HS events (e.g., 12), then the n is a [total number of HS or TO events − 1, $n = 11$]. Walking speed and stride length were normalized by the height of each participant. Moreover, the variability of the stride length, stride time, and stance phase were quantified as the coefficient of variance (CV; standard deviation/mean × 100). The GA was measured according to the bilateral differences between the left and the right limbs during walking²⁷, and the formula is as follow:

$$\text{Gait asymmetry (\%)} = \left| \ln \left(\frac{\text{Short swing time}}{\text{Long swing time}} \right) \right| \times 100 \quad (6)$$

The swing time was calculated for each left and right side, and the Long swing time is defined greater mean value between the left and right sides, whereas the short swing time is a relatively lower mean value²⁷. The proposed ML model considered five demographic variables (age, sex, body mass index (BMI), total PA, and education levels) and nine gait variables (walking speed, stride length, cadence, stance phase, stride time, CV of stride length, CV of stance phase, CV of stride time, GA) as predictor variables.

Statistical analysis. The gait data was evaluated through the Shapiro–Wilk test for normality. The gait variables were normalized into max–min scores for all variables. An independent sample t -test was used to determine significant differences in characteristics between all participants (high-risk and low-risk group) and demographics.

To predict the factors affecting the risk of falls by the spatiotemporal parameters of gait at three different speeds, we derived the ML technique using the XGBoost algorithm. The ML technique aims to find a relationship between the input $X = \{x_1, x_2, \dots, x_N\}$ and the output Y . As described above, relying on the fall-levels definition, the risk of falls was classified into high- and low-risk groups.

For a given dataset with n samples and m features, K additive functions are used in the XGBoost model to predict the output through the following estimation¹²:

Variables	All participants (n = 746)	High-risk (n = 290)	Low-risk (n = 456)	p-value
Age (years)	73.0 ± 5.2	74.2 ± 5.4	72.2 ± 4.9	< 0.001
Height (cm)	157.4 ± 8.1	158.0 ± 8.2	156.9 ± 8.1	0.070
Body weight (kg)	61.3 ± 8.8	61.4 ± 8.9	61.2 ± 8.7	0.729
Body mass index (kg/m ²)	24.7 ± 2.9	25.1 ± 3.1	24.5 ± 2.7	0.003
Total PA (MET-min/week)	1874.3 ± 1789.1	1389.7 ± 1409.7	2182.6 ± 1932.4	< 0.001
Education (levels)	2.1 ± 1.1	1.7 ± 1.1	2.3 ± 1.1	< 0.001
MMSE score	26.5 ± 2.9	25.7 ± 3.3	26.9 ± 2.6	< 0.001

Table 1. Demographic characteristics of participants. *Mean ± SD* mean and standard deviation, *PA* physical activity, *METs* metabolic equivalents, *MMSE* mini-mental state examination; the level of education was defined as a categorical variable [elementary school education or less (education level 1), middle school education (education level 2), high school education (education level 3), college degree (education level 4), or higher (education level 5)]; boldface denotes a significant difference between high- and low-risk of falls; significant difference: $p < 0.05$.

$$\hat{y}_j = \sum_{k=1}^K f_k(x_i), \quad (7)$$

where $f_k \in \{f(x) = \omega_q\}$ ($q: \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T$) is the regression tree's space, and q denotes the independent structure of each tree with T leaves. Each f_k corresponds to an independent tree structure q and leaf weights ω . To learn the set of functions, the following regularized objective is minimized.

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (8)$$

where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2$, l denotes the model loss function, and Ω denotes the regularized term. The dataset was split into a training set (70%) and a testing set (30%). Ten-fold cross-validation with a random split was used for all the processes.

The model was measured using the prediction performance of the model by computing C-statistics (i.e., the area under the receiver operating characteristic [ROC] curve), prospective prediction results, and decision curves. The accuracy of each identified parameter was estimated as follows:

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{False Negative} + \text{False Positive} + \text{True Negative})} \quad (9)$$

The sensitivity, measuring how accurately the high-risk fall group is identified, and the specificity, measuring how accurately the low-risk fall group is identified, were calculated as

$$\text{Sensitivity} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \quad (10)$$

$$\text{Specificity} = \frac{\text{True Negative}}{(\text{True Negative} + \text{False Positive})} \quad (11)$$

Using the area under the ROC curves, we evaluated the accuracy of the gait variables in predicting the risk of falls in older adults. All models were adjusted by age, sex, BMI, level of education, and PA levels as covariates. Level of education was defined as a categorical variable (elementary school education or less, middle school education, high school education, college degree, or higher). All analyses were performed with R statistical software (version 3.6.1, RStudio). The level of statistical significance was set at 0.05.

Results

Table 1 shows the demographic and cognitive characteristics of participants. Compared to the participant in the low-risk fall group, participants with a high risk of falls were relatively older, with higher BMI, lower PA levels and education levels, and poorer cognition. From Table 2, 23 of 27 gait variables were significantly impaired in the high-risk of the fall group. The gait variables presented highly correlated characteristics, which is essential, as these gait variables are not independent regarding their correlation.

Selected optimal features of gait by XGBoost. The XGBoost algorithm was used to extract the optimal features affecting the risk of falls from a total of 34 features. The classification model considered high- and low-risk groups according to fall risk levels. Figure 2 shows the ROC curves; the corresponding values of area under the curve (AUC) for each speed are presented, and the accuracy of each classification model was approximately 68%, 70%, and 67% in the slower-walking, preferred-walking, and faster-walking speed models, respectively. Moreover, the sensitivities were approximately 43%, 53%, and 51% in the slower-walking, preferred-walk-

Variables	All participants (n = 746)	High-risk (n = 290)	Low-risk (n = 456)	p-value
Slower speed				
Walking speed (m/s)	0.9 ± 0.1	0.8 ± 0.1	0.9 ± 0.1	<0.001
Stride length (m)	1.1 ± 0.1	1.1 ± 0.1	1.2 ± 0.1	<0.001
Cadence (beats/min)	97.8 ± 10.9	97.9 ± 11.1	97.7 ± 10.7	0.799
Stance phase (%)	59.2 ± 1.6	59.5 ± 1.7	59.0 ± 1.6	<0.001
Stride time (s)	1.2 ± 0.1	1.2 ± 0.1	1.2 ± 0.1	0.895
CV of stride length (%)	2.7 ± 1.3	3.0 ± 1.4	2.5 ± 1.2	<0.001
CV of stance phase (%)	4.4 ± 2.2	5.0 ± 2.5	4.2 ± 2.0	<0.001
CV of stride time (%)	2.7 ± 1.3	3.0 ± 1.4	2.5 ± 1.1	<0.001
Gait asymmetry (%)	2.6 ± 2.4	2.8 ± 2.6	2.5 ± 2.3	0.123
Preferred speed				
Walking speed (m/s)	1.2 ± 0.2	1.1 ± 0.2	1.2 ± 0.2	<0.001
Stride length (m)	1.2 ± 0.2	1.1 ± 0.2	1.3 ± 0.1	<0.001
Cadence (beats/min)	115.9 ± 10.2	114.8 ± 10.3	116.7 ± 10.1	0.016
Stance phase (%)	57.5 ± 1.7	58.0 ± 1.8	57.2 ± 1.6	<0.001
Stride time (s)	1.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.1	0.014
CV of stride length (%)	2.0 ± 1.1	2.2 ± 1.3	1.9 ± 0.9	<0.001
CV of stance phase (%)	3.0 ± 1.7	3.3 ± 2.0	2.8 ± 1.4	0.001
CV of stride time (%)	2.0 ± 1.1	2.2 ± 1.3	1.9 ± 0.9	<0.001
Gait asymmetry (%)	2.0 ± 2.0	2.1 ± 2.1	2.0 ± 2.0	0.752
Faster speed				
Walking speed (m/s)	1.5 ± 0.2	1.4 ± 0.2	1.5 ± 0.2	<0.001
Stride length (m)	1.4 ± 0.2	1.3 ± 0.2	1.4 ± 0.2	<0.001
Cadence (beats/min)	130.3 ± 11.7	127.8 ± 12.1	131.8 ± 11.1	<0.001
Stance phase (%)	55.7 ± 1.8	56.3 ± 1.9	55.3 ± 1.7	<0.001
Stride time (s)	0.9 ± 0.1	0.9 ± 0.1	0.9 ± 0.1	<0.001
CV of stride length (%)	1.9 ± 1.0	2.0 ± 1.3	1.7 ± 0.8	0.001
CV of stance phase (%)	2.4 ± 1.3	2.6 ± 1.4	2.2 ± 1.1	<0.001
CV of stride time (%)	1.9 ± 1.0	2.0 ± 1.3	1.7 ± 0.8	<0.001
Gait asymmetry (%)	2.1 ± 2.0	2.3 ± 2.2	1.9 ± 1.8	0.015

Table 2. Gait characteristics of participants. Mean ± SD mean and standard deviation, CV coefficient of variance, a significant difference between high- and low-risk of falls: $p < 0.05$.

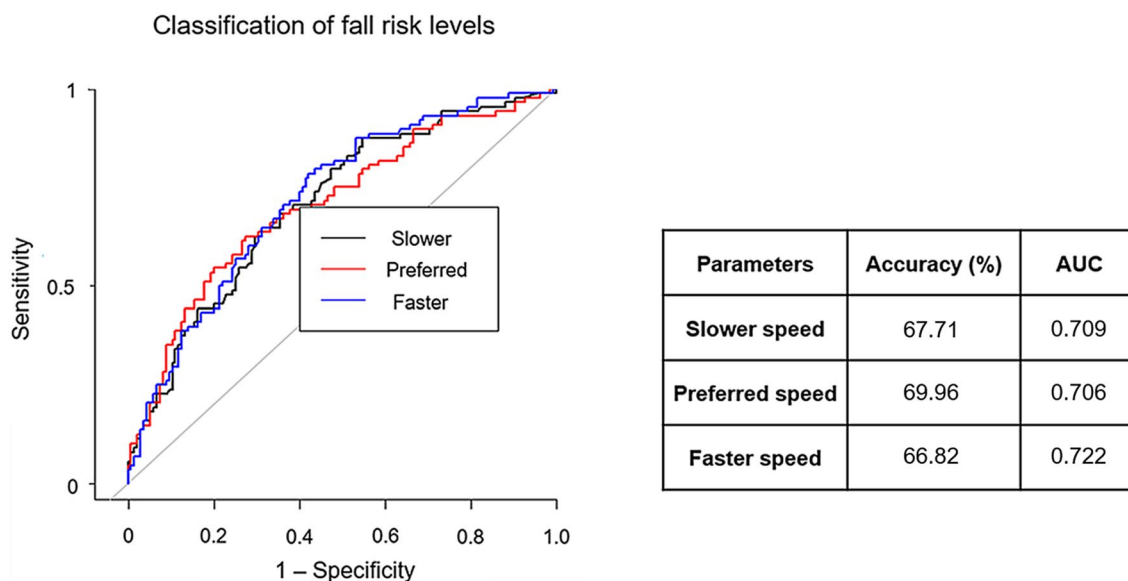


Figure 2. Prediction ability of the XGBoost models for three different walking speeds. ROC curves show the superiority in classifying high vs low level of fall risk. The corresponding values of AUC for each speed are presented.

Models	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PLR (95% CI)	NLR (95% CI)
Slower speed	0.71 (0.64–0.78)	0.43 (0.33–0.54)	0.84 (0.76–0.89)	0.63 (0.50–0.75)	0.69 (0.62–0.76)	2.65 (1.69–4.16)	0.68 (0.56–0.83)
Preferred speed	0.71 (0.64–0.78)	0.53 (0.42–0.64)	0.81 (0.73–0.87)	0.64 (0.52–0.75)	0.73 (0.65–0.80)	2.77 (1.87–4.12)	0.58 (0.45–0.73)
Faster speed	0.72 (0.66–0.79)	0.51 (0.40–0.62)	0.77 (0.69–0.84)	0.59 (0.47–0.70)	0.71 (0.63–0.78)	2.23 (1.54–3.23)	0.63 (0.53–0.83)

Table 3. Prediction results of the three different walking speed models of the XGBoost in the risk of falls. *AUC* area under the curve, *NLR* negative likelihood ratio, *NPV* negative predictive value, *PLR* positive likelihood ratio, *PPV* positive predictive value, *CI* confidence interval.

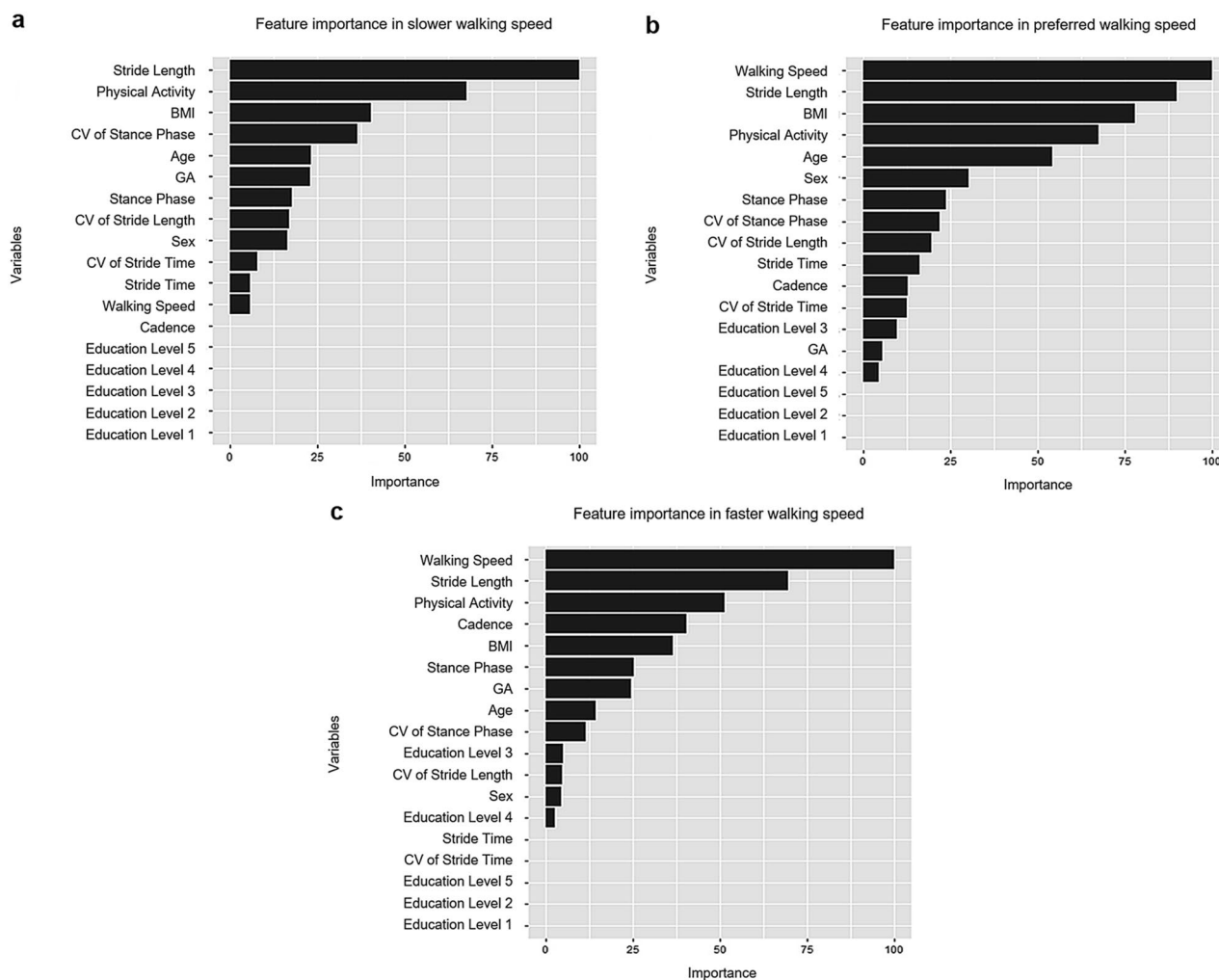


Figure 3. Important features for each walking speeds using the XGBoost models. (a) Results of feature selection for the slower-walking, (b) preferred-walking, and (c) faster-walking speed models.

ing, and faster-walking speed models, respectively. The specificities were approximately 84%, 81%, and 77% in the slower walking, preferred walking, and faster-walking speed models, respectively (Table 3).

In the study, the feature importance was calculated through XGBoost to determine the features having an optimal effect when determining the risk of falls. The feature importance is the score result indicating how each variable contributes to the model accuracy when creating the XGBoost model. Figure 3 shows the result of deriving the importance of the main features among all the explanatory variables. As shown in Fig. 3, the most important features are stride length (slower speed) and walking speed (preferred and faster speed) for determining in which the fall into a high- or low-risk groups. Additional important features in the slower-walking speed model among the top 10 were CV of stance phase, GA, stance phase, CV of stride length, CV of stride time, and non-spatiotemporal parameters such as PA level, BMI, age, and gender. Additional features in the preferred-walking speed model included stride length, stance phase, CV of stance phase and stride length, and stride time with

non-spatiotemporal parameters such as BMI, PA level, age, and gender among the top 10. Additional features in the faster-walking speed model among the top 10 included the stride length, cadence, stance phase, GA, and CV of stance phase with non-spatiotemporal parameters such as PA level, BMI, and age. Overall, the stride length and stance phase were the common features among the top 10 in all walking speeds models. The variability domain appeared to be an important factor in the risk of falls in older adults when the walking speed was slow. Moreover, as the walking speed increased, the pace and rhythm domains appeared to be important factors.

Discussion

The analysis of 746 data from older adults was performed using the ML algorithm XGBoost, an approach that allows the identification of optimal features of gait to predict the risk of falls. The XGBoost model achieved high predictive performance using only gait variables. The developed model also achieved acceptable sensitivity and specificity for predicting the risk of falls. To our best knowledge, this is the first study that has applied the ML approach using the XGBoost to identify the predicting gait features for the risk of fall analysis in older adults. The main findings of this study can be summarized as follows: (1) Stride length, walking speed, and stance phase of gait features were identified using XGBoost; these features accurately classified the fall risk levels. (2) The most relevant features were preferred- and faster-walking speed to determine in which group, high- or low-risk, falls can be classified. (3) The XGBoost algorithm could be a useful tool to identify the predicting gait features of the risk of falls in older adults. These findings are discussed in detail below.

Nine gait variables in each walking speed were used as input features to identify gait variables for predicting the risk of falls. In our model, the stride length at slower-walking and the walking speed variable at preferred-walking and faster-walking speeds were the most important features to predict the risk of falls. Moreover, the stance phase is also the common variable among the top 10 for all walking speeds models. In gait assessment, gait variables such as the walking speed have been associated with a high risk of falls^{6,28}. A previous study found that a decline in walking speed is one of the early markers of falls²⁹. The age-related gait characteristics change in older adults with slow walking speed and a shorter stride (or step) length, could lead to an increase in the stance phase. Thus, walking speed could not always be considered as an independent variable to predict the falls. Simultaneously, the stride length and stance phase were also important features to predict the risk of falls, as shown in our models; therefore, these gait variables should be considered together. The slow walking speed with a shorter stride length may contribute to a longer stance phase in response to the insufficient generating capacity of lower extremity torque. This could result from the force-length relationship, owing to lower strength in older adults, because walking speed is modulated using propulsive force generation during the stance phase of walking^{18,30}. This gait pattern may produce dynamic instability, which could lead to an increase in the risk of falls⁷. Moreover, a longer stance phase disrupted the gait harmony^{31,32} (golden ratio between the stance and the swing phase) caused by an impairment in the reciprocal circuits between the cerebellum and the basal ganglia. This can be involved in the regulation of gait, because the overlapped area cooperates to modulate the motor and cognitive functions during walking in the older adults^{33,34}. This is supported by our previous studies where a longer stance phase, owing to the slow walking speed with a shorter stride length, as well as decreased muscular strength, was strongly associated with the lower global cognitive functions in older adults^{18,30}. The results also showed that global cognitive function in a high-risk group indicated lower cognitive functions than a low-risk group. Moreover, our findings are similar to a previous study where the association between an increase in gait variability and an increase in fall risks in older adults was analyzed³⁵. Our result showed that the variability domain appeared to be an important factor in the high risk of falls when the walking speed was slow. Gait variability is increased in response to the stride-to-stride fluctuations to generate force using muscle in the aging process with the partial summation of overlapping twitches due to impaired cognitive functions during modulation in slow walking¹⁸.

Our results showed that one of the most important features was preferred and faster walking speed to determine in which group, high- or low-risk, the fall can be classified. Namely, an increase in walking speed may increase the risk of falls rather than the slow walking speed. Walking can be defined as a process of continuous loss and recovery of balance that initiates as the center of mass (COM) moving forward, translating the body system mechanically, and recovering a dynamic balance by moving another foot forward to avoid falls³⁶. The COM motion in the mediolateral direction could decrease, whereas in the vertical direction could increase as the walking speed increases following a sinusoidal pattern³⁷. Thus, altering the COM motion due to the increase in walking speed may contribute to the decline of dynamic stability during walking³⁸. Furthermore, dynamic instability due to impaired postural regulation during walking in older adults increases the potential risk of falls because the postural regulation may be integrated through the descending commands for movement being transmitted to brainstem, which is involved in postural control, providing a way to adjust the magnitude and timing of postural changes during stance phase³⁹. On increasing the walking speed, this impaired postural regulation could not dissipate the momentum generated with a fast walking speed despite the momentum control of COM being essential to maintain the dynamic stability³⁸.

Selection of the ML techniques to predict the factor affecting the risk of falls was based on ML framework. Different ML models such as XGBoost, logistic regression, classification and regression tree, random forest, and deep learning were employed (Tables S1–S3). However, in this study, better fall status prediction results were obtained using the XGBoost. Based on our three different walking speed models, we suggest that the XGBoost algorithm could be a useful tool to identify the predicting gait features of the risk of falls in older adults. In the models, the results can classify the high-risk group from the low-risk group with an overall accuracy ranging from 67 to 70% with the sensitivity ranging from 43 to 53% and specificity ranging between 77 and 84%. A previous study reported that the XGBoost algorithm showed high predictive classification accuracy on falls, which is similar to our models⁴⁰. In addition, the preferred walking speed model had better classification ability to predict the risk of falls among three different walking speed models. The selected optimal features of gait obtained by

XGBoost are similar to numerous previous studies regarding the features predicting the risk of falls^{5–8}. Therefore, these findings pave the way for a better understanding of the utility of ML-XGBoost algorithm to help informed prediction of potential falls.

Our study presented several potential limitations. First, we were unable to consider the fall efficacy scale, assessing the fear of falls. We evaluated the fear of falls using only the question that 'Are you afraid of falls? However, we assumed that our fall-levels definition could be properly classified as the risk of falls, even though only one question was asked to assess fear of falls. Second, our datasets have an imbalance between sex and age. Older adults are reported as a risk factor of falls^{2,41}. The classification results should improve with a more homogeneous dataset. Lastly, one may assume that the relatively higher distribution datasets of high-risk groups might affect the predictability (lower sensitivities). To improve the classification performance, a comparison of the ML models' performances should be conducted. Moreover, ML techniques with higher predictability and a filtering technique for human motion should be developed. A method considering the further expanding of the number of samples or collecting various samples and additional variables contributing to improving the predictability can be added to the classification model. These additional possible ways could improve the XGBoost model classification or even other transparent models. However, we concluded that the three XGBoost approaches consistently showed outstanding predictability. Further studies should evaluate the findings on a much larger dataset in realistic environmental conditions.

Conclusions

In this study, the ML-XGBoost approach was used to identify the most important features for predicting the risk of falls in older adults. The XGBoost algorithm showed the highest classification accuracy of 70% and selected the optimal features such as stride length at slower-walking and the walking speed variable at preferred-walking and faster-walking speeds. Moreover, the stance phase in all walking speeds was also selected as the optimal feature for precisely fall risk levels classification in older adults. Additionally, the results showed that the increase in walking speed should increase fall risks. These gait features should be considered for predicting the risk of falls in older adults. The fall risk assessment by the ML approaches with inertial measurement unit sensors improved the classification of the individuals with a high risk of falling. Our results are useful for the foundation for future works on fall prevention. Moreover, our ML approaches could inspire the fall risk potential assessment through the gait analysis of older adults.

Received: 25 August 2020; Accepted: 21 April 2021

Published online: 09 June 2021

References

- Mancini, M. *et al.* Continuous monitoring of turning mobility and its association to falls and cognitive function: A pilot study. *J. Gerontol. A* **71**, 1102–1108 (2016).
- Fuller, G. F. Falls in the elderly. *Am. Fam. Physician* **61**, 2159 (2000).
- Muir, S. W., Gopaul, K. & Montero Odasso, M. M. The role of cognitive impairment in fall risk among older adults: A systematic review and meta-analysis. *Age Ageing* **41**, 299–308 (2012).
- Robinovitch, S. N. *et al.* Video capture of the circumstances of falls in elderly people residing in long-term care: An observational study. *Lancet* **381**, 47–54 (2013).
- Gulley, E., Ayers, E. & Verghese, J. A comparison of turn and straight walking phases as predictors of incident falls. *Gait Posture* **79**, 239–243 (2020).
- Verghese, J., Holtzer, R., Lipton, R. B. & Wang, C. Quantitative gait markers and incident fall risk in older adults. *J. Gerontol. A* **64**, 896–901 (2009).
- Aboutorabi, A., Arazpour, M., Bahramzadeh, M., Hutchins, S. W. & Fadayevatan, R. The effect of aging on gait parameters in able-bodied older subjects: A literature review. *Aging Clin. Exp. Res.* **28**, 393–405 (2016).
- Herssens, N. *et al.* Do spatiotemporal parameters and gait variability differ across the lifespan of healthy adults? A systematic review. *Gait Posture* **64**, 181–190 (2018).
- Van Kan, G. A., Houles, M. & Vellas, B. Identifying sarcopenia. *Curr. Opin. Clin. Nutr. Metab. Care* **15**, 436–441 (2012).
- Wahid, F., Begg, R. K., Hass, C. J., Halgamuge, S. & Ackland, D. C. Classification of Parkinson's disease gait using spatial-temporal gait features. *IEEE J. Biomed. Health Inform.* **19**, 1794–1802 (2015).
- Davis, K. D. *et al.* Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. *Nat. Rev. Neurol.* **13**, 624–638 (2017).
- Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining* 785–794 (2016).
- Dey, A. Machine learning algorithms: A review. *Int. J. Comput. Sci. Inform. Tech.* **7**, 1174–1179 (2016).
- Lai, D. T., Begg, R. & Palaniswami, M. Svm models for diagnosing balance problems using statistical features of the mtc signal. *Int. J. Comput. Intell. Appl.* **7**, 317–331 (2008).
- Kabeshova, A. *et al.* Artificial neural network and falls in community-dwellers: A new approach to identify the risk of recurrent falling?. *J. Am. Med. Dir. Assoc.* **16**, 277–281 (2015).
- Qiu, H., Rehman, R. Z. U., Yu, X. & Xiong, S. Application of wearable inertial sensors and a new test battery for distinguishing retrospective fallers from non-fallers among community-dwelling older people. *Sci. Rep.* **8**, 1–10 (2018).
- Lee, M., Youm, C., Noh, B. & Park, H. Gait characteristics based on shoe-type inertial measurement units in healthy young adults during treadmill walking. *Sensors* **20**, 2095 (2020).
- Noh, B., Youm, C., Lee, M. & Park, H. Age-specific differences in gait domains and global cognitive function in older women: Gait characteristics based on gait speed modification. *PeerJ* **8**, e8820. <https://doi.org/10.7717/peerj.8820> (2020).
- Lee, M., Youm, C., Jeon, J., Cheon, S. M. & Park, H. Validity of shoe-type inertial measurement units for Parkinson's disease patients during treadmill walking. *J. Neuroeng. Rehabil.* **15**, 38 (2018).
- Oyeyemi, A. L., Umar, M., Oguiche, F., Aliyu, S. U. & Oyeyemi, A. Y. Accelerometer-determined physical activity and its comparison with the International Physical Activity Questionnaire in a sample of Nigerian adults. *PLoS ONE* **9**, e87233. <https://doi.org/10.1371/journal.pone.0087233> (2014).

21. Nakakubo, S. *et al.* Association of walk ratio during normal gait speed and fall in community-dwelling elderly people. *Gait Posture* **66**, 151–154 (2018).
22. Silva, J. *et al.* Comparing machine learning approaches for fall risk assessment. *Bio-inspired Syst. Signal Process.* **5**, 223–230 (2017).
23. Bergen, G., Stevens, M. R. & Burns, E. R. Falls and fall injuries among adults aged ≥ 65 years: United States, 2014. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 993–998 (2016).
24. Chung, M. J. & Wang, M. J. J. The change of gait parameters during walking at different percentage of preferred walking speed for healthy adults aged 20–60 years. *Gait Posture* **31**, 131–135 (2010).
25. Kim, Y. K., Joo, J. Y., Jeong, S. H., Jeon, J. H. & Jung, D. Y. Effects of walking speed and age on the directional stride regularity and gait variability in treadmill walking. *J. Mech. Sci. Technol.* **30**, 2899–2906 (2016).
26. Winter, M.V. Normal gait. *Gait analysis: an introduction. Gait analysis (Fourth Edition)* 47–100 (Butterworth-Heinemann).
27. Plotnik, M., Giladi, N. & Hausdorff, J. M. A new measure for quantifying the bilateral coordination of human gait: Effects of aging and Parkinson's disease. *Exp. Brain Res.* **181**, 561–570 (2007).
28. Bohannon, R. W., Andrews, A. W. & Thomas, M. W. Walking speed: Reference values and correlates for older adults. *J. Orthop. Sports Phys. Ther.* **24**, 86–90 (1996).
29. Kyrødalen, I. L., Thingstad, P., Sandvik, L. & Ormstad, H. Associations between gait speed and well-known fall risk factors among community-dwelling older adults. *Physiother. Res. Int.* **24**, e1743. <https://doi.org/10.1002/pri.1743> (2019).
30. Noh, B., Youm, C., Lee, M. & Park, H. Associating gait phase and physical fitness with global cognitive function in the aged. *Int. J. Environ. Res. Public Health* **17**, 4786. <https://doi.org/10.3390/ijerph17134786> (2020).
31. Iosa, M. *et al.* The golden ratio of gait harmony: Repetitive proportions of repetitive gait phases. *Biomed. Res. Int.* **2013**, 918642. <https://doi.org/10.1155/2013/918642> (2013).
32. Serrao, M. *et al.* Harmony as a convergence attractor that minimizes the energy expenditure and variability in physiological gait and the loss of harmony in cerebellar ataxia. *Clin. Biomech.* **48**, 15–23 (2017).
33. Holtzer, R., Epstein, N., Mahoney, J. R., Izzetoglu, M. & Blumen, H. M. Neuroimaging of mobility in aging: A targeted review. *J. Gerontol. A.* **69**, 1375–1388 (2014).
34. Kikkert, L. H. J., Vuillerme, N., van Campen, J. P., Hortobágyi, T. & Lamoth, C. J. Walking ability to predict future cognitive decline in old adults: A scoping review. *Ageing Res. Rev.* **27**, 1–14 (2016).
35. Hausdorff, J. M., Rios, D. A. & Edelberg, H. K. Gait variability and fall risk in community-living older adults: A 1-year prospective study. *Arch. Phys. Med. Rehabil.* **82**, 1050–1056 (2001).
36. Tesio, L. & Rota, V. The motion of body center of mass during walking: A review oriented to clinical applications. *Front. Neurol.* **10**, 999. <https://doi.org/10.3389/fneur.2019.00999> (2019).
37. Orendurff, M. S. *et al.* The effect of walking speed on center of mass displacement. *J. Rehabil. Res. Dev.* **41**, 829–834 (2004).
38. Meyer, G. & Ayalon, M. Biomechanical aspects of dynamic stability. *Eur. Rev. Aging Phys. Act.* **3**, 29–33 (2006).
39. Drew, T., Prentice, S. & Schepens, B. Cortical and brainstem control of locomotion. *Prog. Brain Res.* **143**, 251–261 (2004).
40. Gao, C. *et al.* Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Sci. Rep.* **8**, 1–21 (2018).
41. Callis, N. Falls prevention: Identification of predictive fall risk factors. *Appl. Nurs. Res.* **29**, 53–58 (2016).

Acknowledgements

This work was supported by the Sports Promotion Fund of Seoul Olympic Sports Promotion Foundation from the Ministry of Culture, Sports and Tourism (Grant Number B0080605000494). The authors would like to thank the biomechanics laboratory staff at Dong-A University for their assistance with data collection.

Author contributions

B.N., C.Y., E.G., H.J., and O.K. designed the research. B.N., C.Y., E.G., M.L., and H.P performed material preparation, data collection, and analysis. B.N. and C.Y. wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-91797-w>.

Correspondence and requests for materials should be addressed to C.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021