MSc Biomedical Engineering

Measuring fall risk of the elderly with IMU sensors by developing a Convolutional Neural Network

Carlijn Claassen

TU Delft

Measuring fall risk of the elderly with IMU sensors by developing a Convolutional Neural Network

By

Carlijn Claassen

in partial fulfilment of the requirements for the degree of

Master of Science
in Biomedical Engineering

at the Delft University of Technology,
to be defended publicly on Thursday August 25, 2022 at 10:00 AM.

Supervisor:	Dr. ir. E. van der Kruk	
Thesis committee:	Prof. dr. F.T.C. van der Helm,	TU Delft
	Prof. dr. ir. J. Harlaar,	TU Delft
	Drs. F. Lefeber,	Smart Floor

Measuring fall risk of the elderly with IMU sensors by developing a Convolutional neural Network

Department of Biomechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, TU Delft, The Netherlands,
August 18, 2022
C. M. Claassen

Abstract— The combination of the high number and the consequences of falls in older adults led to the development of fall risk assessments; non-sensor-based and sensor-based. Multiple studies used ML for older adults' fall risk prediction using raw IMU data. This study's objective was to develop a DL algorithm that predicts the fall risk of people living in a geriatric rehabilitation department using raw data collected from IMUs positioned at the ankles during the 10-m walk test. Raw IMU data of 97 participants were used. The participants were classified as low, increased or high fall risk based on the Performance Oriented Mobility Assessment (POMA). Accelerometer and gyroscope's resultant time-series sequences (n=1037) were used as input for the Convolutional Neural Network (CNN) that was optimised and trained with 80% and tested with 20% of the participants. The results were compared with the performance of an existing portable sensor-based fall risk assessment called the Smart Floor (SF). The macro F1 of the unweighted (40%) and weighted (41%) multiclass classification CNNs was lower than the macro F1 of the SF (49%). The binary classification CNN's macro F1 (56%) was slightly lower than the SF's performance. All CNNs were better at predicting high-risk sequences. All models had poor performance when all three POMA fall risk categories should have been predicted. Adjustments to the data collection and CNN optimisation methods should be performed to study the possibility of predicting fall risk using raw IMU data in geriatric rehabilitation centres.

***Index Terms*—** fall risk assessment, older adults, geriatric rehabilitation, inertial sensors, accelerometer, gyroscope, deep learning, convolutional neural network, gait analysis

I. INTRODUCTION

One-third of the people of 65-years of age and older experience a fall annually [1, 2]. A *fall* is defined as "an event which results in a person coming to rest inadvertently on the ground or floor or other low levels" [1]. The number of falls is expected to increase progressively in the upcoming years [2]. The elderly who have experienced a fall suffered among others, injuries, long-term institutional care [3], fear of falling [4], reduced activity [5, 6], and a lower quality of life [5]. Besides physical and mental health care consequences, the medical costs of falls in the Netherlands alone consisted of more than one billion euros in 2019 [2].

The timely detection of fall risk is beneficial to timely prescribe fall risk prevention programs like exercising, using assisting devices, wearing a protective garment or receiving increased surveillance. Therefore, over thirty fall risk assessments have been developed in the past decades [7]. Even though there is no uniformly used definition of fall risk, fall risk assessments can determine whether someone is prone to falling [7, 8]. Fall risk assessments can be validated using other fall risk assessments, retro- or prospective falls. Since the goal is to prevent falls, future falls should be the golden standard. However, a prospective study costs time and money, so most fall risk assessments are validated using retrospective falls or other assessments [7].

In a recent literature study, we categorised the fall risk assessments as non-sensor-based and sensor-based, focusing on clinical assessments or daily life activities (ADL) [7]. The sensors can be wearable (e.g., accelerometer, gyroscope, magnetometer) or portable (e.g., Microsoft Kinect) [7, 8].

Smart Floor is a company that developed a portable sensor-based fall risk assessment used in clinical assessments or ADL. They created a "Smart Floor" (SF) to predict fall risk in the elderly by analysing the gait pattern. The SF is a thin (<0.2 mm) sensor foil (SFfloor) placed underneath or on top of a floor (Figure 1) that collects data when combined with a *wearable* around the ankle (SFwearable), see Figure 2 [9]. The SFwearable consists of a Radio Frequency Identification (RFID) reader and an Inertial Measurement Unit (IMU).

The SF estimates the Performance-Oriented Mobility Assessment (POMA), a non-sensor-based clinical assessment that was first published in 1986 and has been used ever since to predict the fall risk of the elderly [10]. The company has chosen to use the POMA as a non-sensor-based indication of fall risk.

The POMA consists of 16 items focussing on balance and gait during physical tasks. All items are scored on an ordinary scale of 0 until 1 or 0 until 2. The balance part (POMA-B) consists of nine tasks, and participants can achieve a maximum of 16 points. The gait part (POMA-G) consists of seven items with a maximum of 12 points. The total score indicates a person's fall risk. A detailed overview of the items and the corresponding scoring system is displayed in Appendix-A. A total score between 19-24 shows an increased risk, and a score lower than 19 means a high risk of falling. If a participant has a score higher than 24, the participant has a low risk of falling. A score lower than 19 indicates a five times higher risk of falling [9, 11].

The RFID reader of the SFwearable collected the data used to determine four gait parameters: stride length, stride frequency, gait speed, and vertical foot acceleration [9]. With multiple logistic regression, the contribution of each parameter to the POMA score was determined using the regression coefficient leading to an SF fall risk score. The SF fall risk score can also be divided into low, increased, and high fall risk profiles. A score lower than 19 indicates a high fall risk, 19-24 an increased fall risk and a score higher than 24 a low fall risk. In a validation study, the SF is validated with 54 participants aged 55 years and older, temporarily living in a rehabilitation centre. In a rehabilitation centre, patients can recover from injury or a severe medical event until they return to their maximum functional potential. The SF correctly estimated the participants' three POMA fall risk categories with 54% [12].

Currently, the RFID data is used stand-alone to predict fall risk in care facilities, which requires the installation of the SFfloor. The floor installation is time-consuming and complicates fall risk measurements at a patient's home. Besides the RFID reader, the SFwearable also contains an IMU, which does not require interaction with the SFfloor. If the data collected by the IMU, embedded in the SFwearable, could be used to predict the fall risk, installation of the SFfloor is not required. This enables fall risk prediction at patients' homes.

Previous studies have shown that the prediction of fall risk in older adults using IMU data is possible with a machine learning (ML) or deep learning (DL) algorithm [7, 13, 14]. Smart Floor positions the SFwearable around the ankle; thus, the input will be the data collected from the IMU positioned at the ankle. The algorithm's output will be the fall risk classification based on the POMA score. Therefore, this project aims to predict fall risk of the elderly temporarily living in a geriatric rehabilitation centre using raw data collected from IMU sensors positioned at the ankles by developing a DL algorithm.

Figure 1: The sensor foil on top of the existing floor of the geriatric rehabilitation department (left) and the sensor foil covered with a loose lay floor (right).

Figure 2: Picture of SFwearable around the ankle.

A. Related work

Three studies were found that have used raw IMU data to predict fall risk of the elderly using DL. In the study of Nait Aicha et al. (2018), 296 participants wore an accelerometer with a sampling frequency of 100 Hz on the lower back for 1-week [13]. The participants were categorised as fallers and non-fallers based on prospective falls during six months of follow-up. The bouts of non-wearing, locomotion, sitting, standing, and lying were identified using the manufacturer's activity classification algorithm. The locomotion bouts were used in the study. They studied the performance of a Recurrent Neural Network model (RNN), specifically the Long-Short Term Memory (LSTM) model. They also studied the Convolutional Neural Network (CNN) performance and the performance of a CNN to which an LSTM layer was added, the so-called ConvLSTM. The LSTM and ConvLSTM models performed better than the CNN. Due to the long computation time of the LSTM model, only the ConvLSTM model was trained with a larger dataset. The addition of age and sex slightly improved the performance of the ConvLSTM. They also found that a stricter gait classification algorithm might improve the performance. The best-performing model had an area under the Receiver Operating Characteristic (ROC) curve of 0.75. The ROC is the visual representation of the performance of this classifier. The area under the ROC curve (AUC) is a numerical representation of the binary classifier.

Tunca, Salur, and Ersoy (2020) also used raw sensor data to predict the fall risk in older adults with an average age of 76 years [14]. They used a 3D accelerometer and 3D gyroscope that had a sampling frequency of 100 Hz. In this study, 76 participants with neurological disorders walked back and forth three times eight meters with the sensors on the dorsum of each foot. The fall risk classification was based on retrospective falls in the year prior to the assessment. Using a gait analysis system, they extracted the stride length, clearance, stance time, and swing time. These parameters were used to calculate more parameters like cycle, cadence, speed, and stance ratio. The parameters were used for multiple sample-to-label ML models (Support Vector Machine (SVM), Random Forest (RF), MultiLayer Perceptron (MLP)). The results were compared with sequence-to-label DL models (Hidden Markov Model (HMM), LSTM model). In the sequence-to-label models, LSTM and LSTM raw sequences were used. The stride length, clearance, stance time, and swing time parameters were calculated for ten strides in a window and arranged in chronological order in a sequence, leading to four-dimensional sequences of length ten. These sequences are the LSTM sequences. The LSTM raw sequences contained 10-stride windowed raw inertial data and were used to determine whether the model can implicitly learn the required features. The LSTM raw and LTSM sequences were tested with and without windowing and rotations as data augmentation techniques. They showed that the performance of the LSTM model with raw data input improved when windowing and rotations were used as data augmentation techniques. In contrast, the LSTM model had the best performance with only windowing. The LSTM in which the stride length, clearance, stance time, and swing time parameters were sequenced had the best performance of all models, with an AUC of 0.987. The raw LSTM had an AUC of 0.725.

In the study of Roshdibenam et al. (2021), 98 patients from an academic geriatrics clinic with a mean age of 75 years were classified as fallers and non-fallers using the geriatrician's fall risk assessment [15]. This consists of test scores from the Timed-up and Go (TUG) test, 30-sec stand and the 4-stage balance tests, measurement of orthostatic blood pressure, the clinician's observation of movement disorders, number of medications, number of diagnoses, age, gender, BMI, and the Staying Independent Brochure (SIB) score. The SIB is the subjects' report of risk factors, including subjects' history of falls. During the TUG, the participants wore three IMUs; two on the lace of the right and left shoe near the midfoot and one at the back of the neck on the subject's clothing. The IMUs contained a 3D accelerometer and 3D gyroscope that had a sampling frequency of 250 Hz. The three axes of the gyroscope and the accelerometer signals obtained during the TUG were used as input for developing a CNN and an SVM. The accelerometer and gyroscope data were kept separately, and the signals were segmented into three-second signal segments using a sliding window. The model's performance for each sensor and each sensor position was evaluated. The CNN with the gyroscope at the neck showed the best performance: an F1 score of 0.80 and an AUC of 0.75 for predicting the fall risk based on the geriatrician's assessment. However, they also studied the classification performance of prospective falls during six to twelve months of follow-up. The CNN showed a lower performance for predicting prospective falls. The neck's gyroscope still had the best performance, with an F1 score of 0.41 and an AUC of 0.56.

The studies of Nait Aicha et al. (2018) and Tunca, Salur, and Ersoy (2020) show the potential of using raw IMU data to predict fall risk in older adults. The patient characteristics can influence the performance of a model and sensor positions, which makes that the data obtained by the SF are not applicable in the models of the studies as mentioned earlier [16]. The patient characteristics in the study of Roshdibenam et al. (2021) are more comparable to patient characteristics of the SF's dataset. However, the sensor positions are different, and the CNN's performance could be improved.

B. Convolutional Neural Network

This study aims to predict fall risk using raw IMU data by developing a DL algorithm. The model should be able to process sensor data as input and give the fall risk profile based on the POMA as an output. Therefore, the model should be able to recognise fall risk-related patterns from the input data. A Convolutional Neural Network (CNN) is a type of ANN used for pattern recognition in, for example, image or language recognition. A CNN is a DL model, thus consisting of feature extraction and classification. The feature extraction part consists of convolution and pooling layers, and the classification part consists of a fully connected layer and the output. Because of the pattern recognition characteristics and the faster computation time compared to an LSTM [13], there is chosen for using a CNN.

Explaining CNNs can best be done by using image recognition as an example. An image consists of pixels that are grids of numbers that tell what the intensity of each pixel is. These grids of numbers can be analysed and manipulated to find patterns and characteristics. The filters in a convolutional layer of a CNN perform this manipulation. The filter is an $n \times m$-matrix that moves across the full image. A pooling layer can extract the filters' main features and pool together the essential characteristics.

A CNN always consists of a convolutional layer in which the input shape of the layers equals the shape of the input data. A pooling layer follows this layer. After these layers, additional layers can be added, consisting of iterating convolutional and pooling layers. The input of these layers is determined by the output of the layers before. The final layer is a fully connected layer of which the output shape is equal to the number of classes, in this case, three; low, increased, and high fall risk.

Building a CNN requires making decisions about, for example, the number of filters in each layer, the number of convolutional layers in the model, and the pooling method. Therefore, these parameters must be defined before training the model and are called hyperparameters. In a CNN, the hyperparameters can be set to different values. The best combination of the hyperparameters and the corresponding value should lead to the best performance of the model.

II. Methods

This section describes the methods for the data collection, preprocessing of the collected data, hyperparameter optimisation of the CNNs, and evaluation of the CNNs and SF. First, the definitions of the statistical values used during optimisation and evaluation will be outlined.

All methods and data analyses described were performed using Python programming language v3.9 (Python Software Foundation) [17].

A. Definitions statistical values

The macro averaged F1 (macro F1) was used as a performance value during CNN optimisation. The SF and CNN performances were evaluated using *One versus Rest* classification, which enables approaching a multiclass classification as a binary classification problem by studying every class separately. For example, when looking at high fall risk, the performance was determined by how well the high fall risk is predicted compared to the combination of increased and low fall risk.

The performances were analysed with the F1 score, macro F1, sensitivity, specificity, precision, a 3x3-confusion matrix and a ROC curve with the corresponding AUC. The ROC curve and the AUC were not obtained for the SF's performance because only the predicted labels were available, corresponding to one point on the ROC curve [18].

The F1 score gives the harmonic mean between the precision and sensitivity and was calculated with the following equation:

$$F1score = \frac{2 \times precision \times sensitivity}{precision + sensitivity}$$

In this function, the precision tells how many participants that were predicted with a particular fall risk have that fall risk and was calculated as follows:

$$precision = \frac{TP}{TP + FP}$$

Sensitivity tells how many participants with a particular fall risk were predicted with that fall risk and was calculated with:

$$sensitivity = \frac{TP}{TP + FN}$$

In these equations, the TP (true positive) is defined as the number of correctly classified participants for a specific category, e.g., high risk. When looking at high risk, the FN (false negative) indicates the number of participants that should have been a high risk but were predicted as increased or low risk. The FP (false positive) is the number of participants that should not have been classified as high risk but were classified as high risk.

The F1 score gives a value between 0 and 1, where a score of 1 indicates a well-performing model. The F1 score was determined for each fall risk category separately but can be combined into one value using the macro F1:

$$F1_{macro} = \frac{F1_{low} + F1_{increased} + F1_{high}}{3}$$

The specificity defines the probability that not having a specific fall risk gives the prediction of not having that fall risk. When looking at high fall risk, the specificity describes the probability that participants who did not have a high-risk classification also did not have a high-risk prediction. The specificity was calculated using the following equation:

$$specificity = \frac{TN}{TN + FP}$$

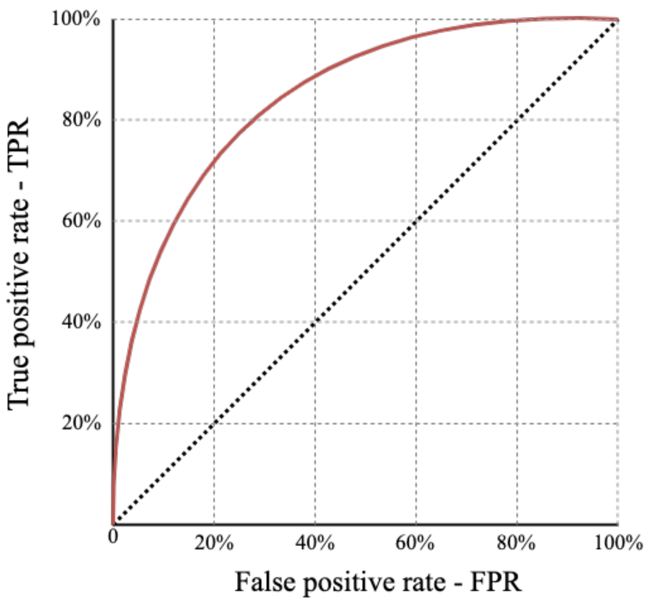


Figure 3: Example of a ROC curve based on the curve of [18]. The upper and left axes represent the truth. The diagonal line represents an AUC of 0.5.

The TN (true negative) indicates how many participants are correctly not classified with particular fall risk.

A 3x3-confusion matrix was used to visualise how many sequences were correctly classified as a low, increased or high fall risk. The true label described the categorisation based on the POMA and was displayed on the y-axis. On the x-axis, the predicted label was displayed. The predicted label is the categorisation made by the SF or CNN. The 3x3-confusion matrix shows the model's results on one threshold of the ROC curve. A threshold is a cutoff for the probability value.

The ROC curve visualises the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for every threshold. An example of a ROC curve is displayed in Figure 3. The TPR is equal to the sensitivity. The FPR is the percentage of participants with a specific fall risk that were mistakenly classified with another fall risk and was calculated with

$$FPR = \frac{FP}{TN + FP} = 1 - specificity$$

Generally, a higher TPR and a lower FPR indicate a better model's performance. The area under the ROC curve (AUC) can be used to describe the performance of the ROC curve in one number. In this case, the better the performance, the higher the AUC. The dotted line in Figure 3 describes an AUC of 0.5. A good classification model would have a ROC outlined in the top left corner and an AUC close to 1.

B. Data collection

In 2021 Smart Floor collected data from 101 older adults temporarily living in a geriatric rehabilitation department of a nursing home in Bergen op Zoom, Noord-Brabant, The Netherlands. Participants were included if they were 65 years or older and had a functional ambulation classification (FAC)

of 3 (walking independently with or without a walking aid). Most participants suffered multiple medical disorders like a combination of Chronic Obstructive Pulmonary disease (COPD), Diabetes Mellitus (DM), cardiovascular diseases or COVID-19.
The SF was installed on the floor of the common area of the department. Each participant performed fall risk assessments two days in a row. Each day, the participants performed the ten-meter walk test (10MWT), Timed-Up-and-Go test (TUG), and the POMA. A physical therapist administered the assessments. The participants wore the SFwearable around the left and right ankle during all tests.
Only the data obtained during the 10MWT was used in this study because not every participant's data for the TUG were available. During the 10MWT, the floor was marked with two lines separated ten-meter from each other. Each participant was asked to walk (if possible) three times from one line to the other at their own pace. The participant was allowed to rest for twenty seconds until two minutes between each trial. The help of a third party was not allowed, but the use of an assistive device was [19]. During each trial, the physical therapist counted to three. The participant started walking at three. A third observer started the sensor's recording when the participant started walking and stopped the recording when the participant passed the 10-meter line with one foot.
The POMA scores were used to classify the participants as high (0-18), increased (19-24) or low (25-28) fall risk.
The SFwearable was attached using velcro such that it fell over the Achilles tendon, see Figure 2. The SFwearable contained an RFID reader, accelerometer, gyroscope, magnetometer, WiFi transmitter, and battery. The study's objective was to use raw IMU data; thus, only data collected by the accelerometer, gyroscope, and magnetometer should have been used. However, the magnetometer collected unreliable results due to noise. Thus only the accelerometer and gyroscope data were used. The sampling frequency of the accelerometer and gyroscope was set to 52.3 Hz.
For every trial of the 10MWT, a data file with the collected data was generated per sensor. The data samples of the accelerometer and gyroscope for each timestamp were listed in that file. A maximum of six trials, thus twelve data files, were collected per person. A schematic overview of the collected data for one day is displayed in Figure 4.
The clinic's ethical committee approved the study. The assessments are part of regular care, and all participants participated voluntarily. Therefore, no METC approval was required. All participants signed informed consent. If a participant could not sign the consent, a family member signed it.

C. Preprocessing

Several preprocessing steps were applied to the raw IMU data to increase the model's performance and reduce the computation time. The sensor's recording was manually started when the physical therapist told the participant to start walking and ended when the participant walked ten meters. Therefore, each time series of the accelerometer and gyroscope was related to one trial of a 10MWT. This whole time series was used because the length of the time series could give information about the gait speed, which is correlated to prospective falls [20]. Every axis' time series collected per trial were combined in a data frame for each ankle separately, as is displayed in Figure 6.
Due to a technical defect, the sampling frequency varied, resulting in data samples with unequally distributed timestamps. A data sample is one data point in a time series; see Figure 6. Therefore, all timestamps were resampled to 19 ms, which equals the sampling frequency of 52.3 Hz. Therefore, all data samples were first upsampled to 1 ms timestamps using linear interpolation, followed by downsampling to 19 ms by taking the median. The data samples sometimes had identical timestamps but different accelerometer and gyroscope values. These timestamps were split into equally spaced timestamps between the first and the subsequent not identical timestamp, followed by 19 ms resampling. An example of the timestamps with corresponding data samples before and after splitting is displayed in Appendix-B.
The orientation of the IMU inside the SFwearable could be different in each SFwearable. The resultant of the accelerometer ($accel_{res}$) and gyroscope ($gyro_{res}$) were used to ensure that the different orientations did not affect the model. The resultants were calculated with the following equations:

$$accel_{res} = \sqrt{x_{accel}{}^2 + y_{accel}{}^2 + z_{accel}{}^2}$$

$$gyro_{res} = \sqrt{x_{gyro}{}^2 + y_{gyro}{}^2 + z_{gyro}{}^2}$$

In this report, the calculated $accel_{res}$ and $gyro_{res}$ will be referred to as a feature, and the combination of these two features is called a sequence. Thus, one sequence refers to one 10-m of the 10MWT and one ankle, see Figure 6.

Measurement disturbances were removed, followed by normalising and equalising the sequences to improve the computation time and performance.
The disturbances were removed by i) removing entire sequences and ii) removing parts of sequences.

Removing entire sequences
Some sequences had an extremely short total recording time. The average gait of a community-dwelling older adult is 0.6-1.45 with the desired speed of 1.2 m/s [21]. The mean gait speed in a geriatric rehabilitation setting is around 0.23 m/s, and using an assistive device reduces the gait speed. All sequences with a total time lower than 6.9, equal to walking 10 m with a gait speed of 1.45 m/s, were removed.

Removing parts of a sequence
When observing one sequence at a time, multiple outliers were observed. These outliers could be caused by environmental noise or disconnection in the wire. The outliers were removed using the difference between the minimum and maximum value, also called the range, in a specific window of the $accel_{res}$. This range was calculated with the following equation:

$$range = max_{values} - min_{values}$$

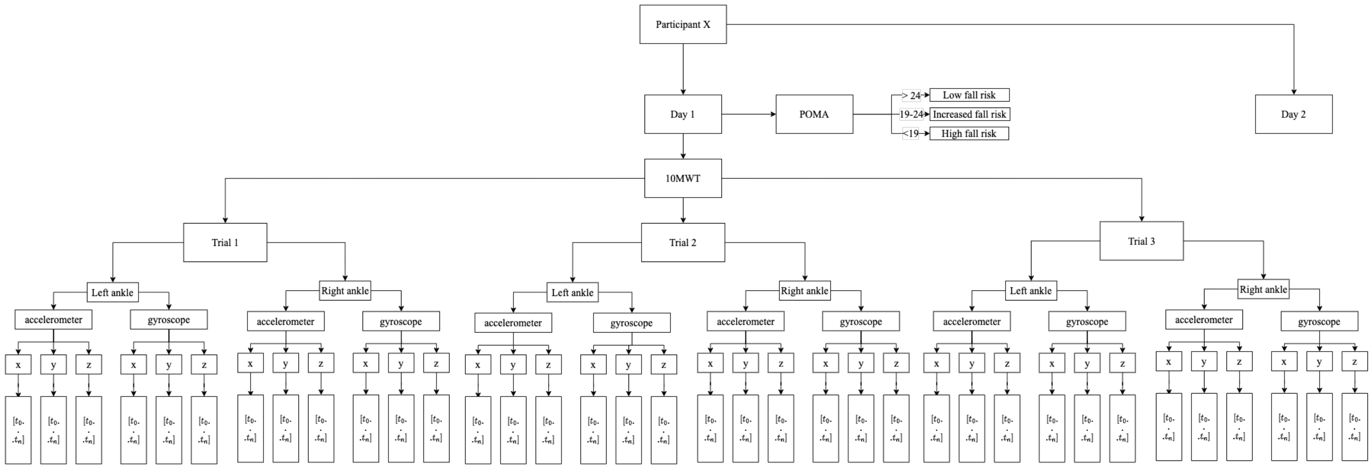


Figure 4: Example of the collected data on one day of a participant who performed three trials of the 10MWT resulting in twelve time-series.

In this equation, max_{values} describes the maximum value in the specific window and min_{values} describes the minimum value in the window.
If the range was higher than a specific value, all data samples inside the window were removed and replaced with *Not a Number* (NaN). The range was only calculated in $accel_{res}$, but the windows were removed in the whole sequence, thus in $accel_{res}$ and the corresponding $gyro_{res}$ data. A manual search found that a range of 30 and a window size of 20 data samples were the best cutoff values. The NaN values were replaced with the mean of each feature without outliers. The effect of removing parts of a sequence is displayed in Figure 5.

During normalisation, all $accel_{res}$ features from the left and right ankle, and all participants were combined to calculate the mean (μ) and standard deviation (σ). The overall mean and standard deviation were extracted from each data sample in a sequence using:

$$sample_{normalized} = \frac{x - \mu}{\sigma}$$

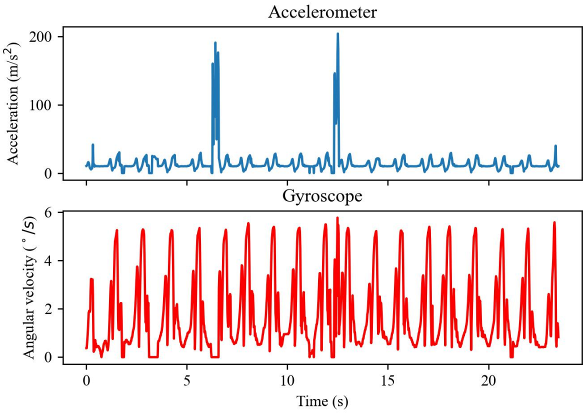
where x is the sample in a trial, μ is the mean of all trials combined, and σ is the standard deviation of all trials combined. The same was done for the $gyro_{res}$ features.

Equalisation of the sequences was necessary to create vectorised data required for DL libraries. When one participant walks faster, the performance time will be shorter, resulting in different sequence lengths. The sequences can be padded or truncated for equalisation. In this study, post padding was also used for equalisation, which was also done in the study of Roshdibenam et al. (2021) [15]. With post padding, all sequences were equalised to the most extended sequences' lengths by adding zeroes. The longest sequence had a length of n=3895.

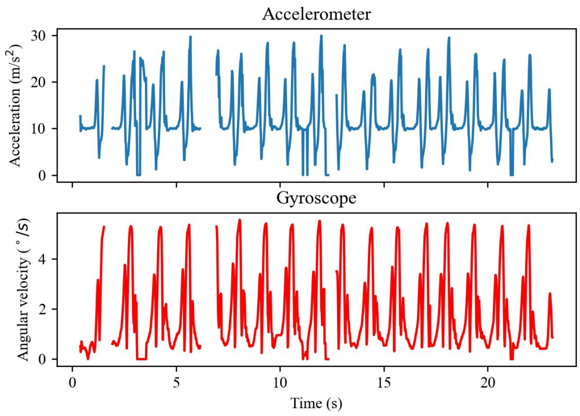
The CNN was given input X consisting of $accel_{res}$ and $gyro_{res}$ and the corresponding fall risk classification $y \in \{0, 1, 2\}$. The high fall risk category was labelled as 0, increased fall risk as 1, and low fall risk as 2.

D. Hyperparameter optimization

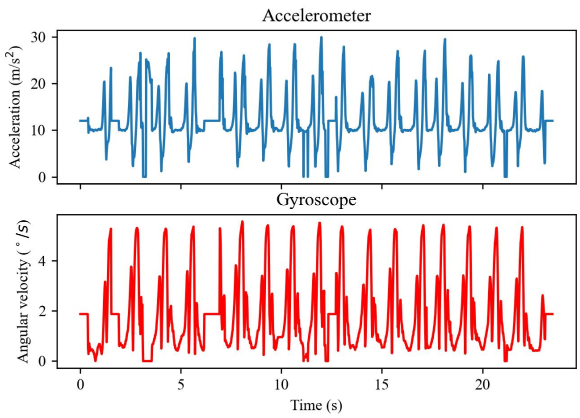
The model's architecture and hyperparameters can influence the model's performance. The architecture described in the study of Nait Aicha et al. (2018) had an iterating structure of convolutional layers and dropout layers with the addition of an LSTM layer in the ConvLSTM model [13]. The maximum number of filters in the CNN layers was 128 leading to a promising AUC of 0.75. The study of Rohsdibenam et al. (2021) had an architecture in which two convolutional layers were followed by one max-pooling layer, which showed a lower performance for predicting prospective falls [15]. Therefore, in the current study was also chosen for an iterating structure in which a max-pooling layer followed each convolutional layer. The convolutional layers contained a rectified linear unit (ReLU) activation function to introduce non-linearity. Besides, all convolutional layers except the first contained L1 and L2 regularisers, both equaling 0.005 to improve the model's fitting. After iterating a specified number of times, a dropout layer is added, followed by a fully connected layer with a ReLU activation. The final layer is the output layer with a softmax activation. A schematic overview of the model's architecture is displayed in Figure 7. After defining the architecture, the hyperparameters should be tuned. Manual search, random search, grid search, and Bayesian optimisation can be used to find the optimal hyperparameters. Manual searching is time-consuming and does not always give reliable results. Therefore, multiple automatic algorithms have been developed. With grid search, every possible combination of hyperparameters is tested to find the optimal combination, but this exhaustive searching method has a long computation time [22]. During a random search, one finds out which hyperparameters contribute to higher performance and which do not. Choosing only the hyperparameters that matter results in a more efficient search, but it is unreliable in more complex models. Another optimisation method is Bayesian optimisation. In Bayesian optimisation, the prior information is combined with sample information to obtain posterior information on the function distribution using the Bayesian formula. Using



(a) With outliers



(b) After outlier removal



(c) Nan replaced with mean

Figure 5: An example of a sequence of the resultant of the accelerometer and gyroscope with outliers (a), without outliers (b), and the Nan replaced with the mean (c).

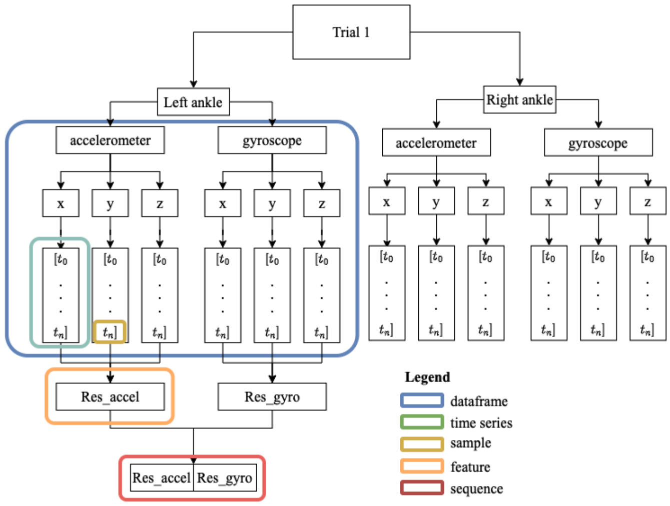


Figure 6: Schematic overview of the terms used in this report to describe different parts of the dataset.

this information, there can be determined where the function gets the optimal value [22]. In this study was chosen for Bayesian optimisation due to the computational cost of the grid search and the unreliability of random search.

Even with Bayesian optimisation, it would be time-consuming to study all possible hyperparameters. Therefore, a selection of hyperparameters was studied to optimise the model. These hyperparameters are the number of convolutional layers (1-7), the number of filters in these layers (multiples of 8, from 8-128), the kernel size (2-5), the dropout (0.0-0.5), and the number of filters in the fully connected layer (multiples of 16, from 8-256). The number of epochs (1-10) was also used as a variable to prevent the Bayesian search would choose a combination of parameters purely by the predefined number of epochs. However, the cross entropy loss function was plotted for the model with the best combination of hyperparameters to observe the fitting. This fitting was used to determine the number of epochs.

Train-test split

The dataset was split into a training and validation (80%) and a testing (20%) set at the participant level to ensure that all sequences of one participant were either in the training or testing set. This was done to ensure that the model learned fall risk classification characteristics instead of participant characteristics [13]. The distribution of the different fall risk profiles in both sets was comparable. For hyperparameter optimisation, the training and validation set was used. During optimisation, grouped stratified 5-fold cross-validation was applied. With grouped 5-fold cross-validation, the total training and validation dataset was divided into five groups, again at the participant level. At each fold, four groups were used as a training set, and one was used as a validation set. Every fold, another group of sequences was part of the training or validation set [14]. The average of every fold's results was the model's macro F1 score that was used for choosing the best hyperparameters.

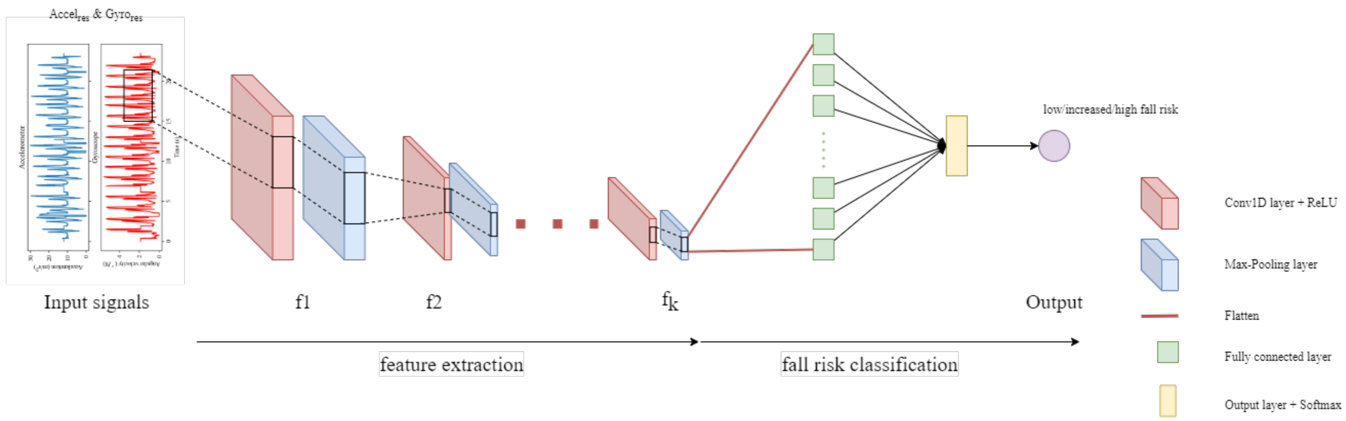


Figure 7: Schematic view of the CNN's architecture. Each layer is described by f_k in which k is the number of layers.

After optimisation, the model was trained using the combined training and validation data as training data. After training, the model was evaluated using the testing set that was not seen by the model before.

In the results of the unweighted multiclass classification, there was a discrepancy between the 3x3-confusion matrix and the ROC curve, which could have been caused by imbalanced data. Therefore, a weighted multiclass classification was also performed. The classes were weighted with

$$weight_{class_x} = \frac{n_{sequences}}{n_{classes} \times n_{sequences_x}}$$

in which $n_{sequences}$ is the number of sequences in the complete dataset, $n_{classes}$ is the number of classes, which is three (high, increased or low) in the multiclass classification problem, and $n_{sequences_x}$ is the number of sequences in class x.

Binary classification
When a physical therapist decides that everybody with an increased fall risk should follow a fall prevention program, it can be of clinical value that the model can differentiate between the combination of high and increased against low risk sequences. Therefore, binary classification was also studied in which the sequences with a high and increased fall risk were combined into one group and compared to the low fall risk sequences. Because of the unequal distribution of participants, the classes were weighted with

$$weight_{class_x} = \frac{n_{sequences}}{n_{classes} \times n_{sequences_x}}$$

in which $n_{sequences}$ is the number of sequences in the complete dataset, $n_{classes}$ is the number of classes, which is two (high or low) in the binary classification problem, and $n_{sequences_x}$ is the number of sequences in class x.

E. Smart Floor evaluation

The SF's estimates of the POMA score in this specific group of participants were determined using a regression algorithm based on data from patients in a nursing home, collected by Smart Floor in 2019 [9]. The SF scores were categorised into low, increased, and high fall risk. All values lower than 0 were classified into the high fall risk category, and all values higher than 28 were classified into the low fall risk category. The POMA scores lower than 0 and higher than 28 were excluded because these scores do not exist according to the POMA scoring system. The SF's performance was evaluated using the (macro) F1 score, sensitivity, specificity, precision, and the 3x3-matrix.

F. Model evaluation

The model was trained and tested five times, and the average of the results was used.

Multiclass classification
One versus Rest classification was used to extract the macro F1, sensitivity, specificity, precision, a 3x3-confusion matrix, and a ROC curve with corresponding AUC. The model's performance for each fall risk category was compared to the performance of the SF.

Binary classification
In binary classification *One versus Rest* classification was not required because the classes were already binary. The performance was studied using the ROC curve with the corresponding AUC, (macro) F1-score, sensitivity, specificity, precision, and a 2x2-confusion matrix.

III. Results

A. Data description

In total, 117 participants performed the assessments. 10MWT data were available from 101 participants. Because of a POMA score outside the range of 0-28, two participants were additionally excluded.
In total, 1100 sequences were collected, less than the expected twelve sequences per participant. This is because not every participant performed three trials per 10MWT, and the sensor did not always work as expected, leading to missing data.

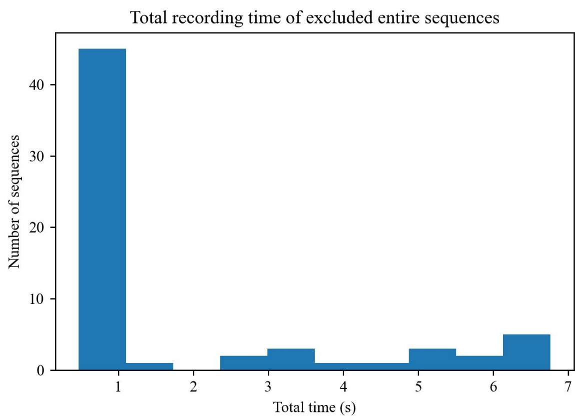


Figure 8: Histogram of sequences excluded based on the total recording time.

Table I: Number of participants, 10MWTs and trials for each fall risk profile based on POMA score after exclusion. Sixteen of the 97 participants had two different fall risk classifications, resulting in 113 classifications.

	Low	Increased	High	**Total**
# of participants	22	57	34	**113**
# of 10MWT	35	94	59	**188**
# of trials	188	518	331	**1037**

Because of a total time lower than 6.9 s, 63 sequences were removed. This resulted in excluding two participants, of whom no sequences were left. The completion time distribution of those sequences is displayed in Figure 8. In Figure 9 the flowchart of all excluded participants and sequences is displayed.

Of the resulting 97 participants, sixteen received two different POMA classifications of the physical therapist during the two assessments. In those cases, both classifications were taken into account. All sixteen participants received a higher fall risk on the first day than on the second. Eight participants received an increased risk on day one and low risk on day two, and eight received a high risk on day one and an increased risk on day two. The number of participants, 10MWTs, and sequences for each fall risk category are displayed in Table I.

The distribution of the sequences in each fall risk category for the training and validation and the test set are displayed in Table II.

Table II: Distribution of sequences for training/validation and testing set.

	Low risk n (%)	Increased risk n (%)	High risk n (%)	Total n
Training set	157 (18.5%)	417 (49.2%)	274 (32.3%)	848
Testing set	31 (16.4%)	101 (53.4%)	57 (30.2%)	189
Total	**188**	**518**	**331**	**1037**

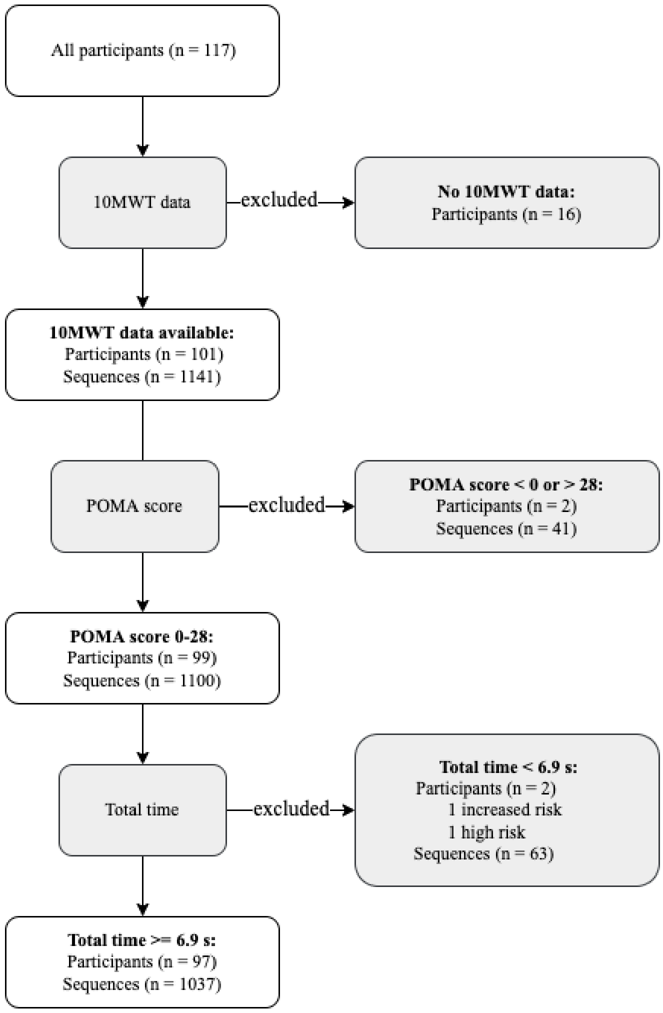


Figure 9: Flowchart of excluded participants and sequences.

B. *Smart Floor evaluation*

For 1065 sequences, the SF fall risk profile was determined. Of these values, 1027 had a POMA between 0 and 28; the others were excluded.

Of the 1027 sequences, 512 were correctly classified, as seen in the 3x3-confusion matrix displayed in Figure 10. The macro F1 score of the SF is 49%. Looking at the high fall risk, the SF has a specificity of 94% and sensitivity of 32% in combination with a precision of 70%, see Table III. The SF is thus better in predicting sequences with a low and increased fall risk than high fall risk. The increased fall risk versus the low and high fall risk has a specificity and sensitivity of respectively 51% and 54%. Thus, 54% of the sequences with an increased fall risk were correctly classified. The low fall risk was best classified with a sensitivity of 69%, and the precision is only 38%, meaning that only 38% of all sequences that were predicted with a low fall risk had a low fall risk.

C. *Model evaluation*

Multiclass classification

Unweighted

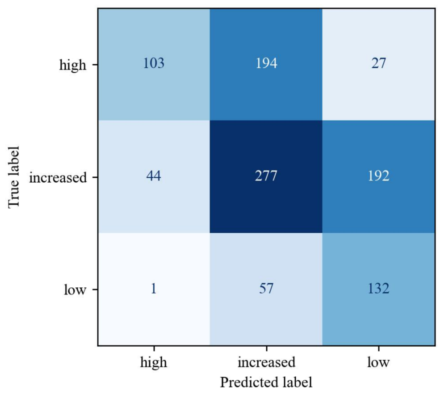


Figure 10: 3x3-confusion matrix of the Smart Floor.

Table III: Performance of the Smart Floor classification.

Fall risk	Sensitivity	Specificity	Precision	F1-score
High	32%	94%	70%	44%
Increased	54%	51%	52%	53%
Low	69%	74%	38%	49%
Macro avg	**52%**	**73%**	**53%**	**49%**

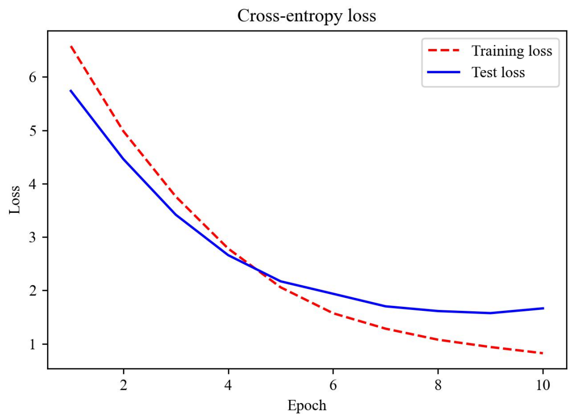
The Bayesian optimization determined the most optimal hyperparameters. The best combination of hyperparameters for the multiclass classification with unweighted and weighted classes is displayed in Table IV. The corresponding cross-entropy loss function of the unweighted multiclass is displayed in Figure 11a, the average cross-entropy of five iterations. The number of epochs was set to four.
The macro F1 of the unweighted multiclass classification model was 46%, lower than the SF classification. The sensitivity and precision of the increased and high fall risk are higher than the sensitivity of the SF. For the increased fall risk, the sensitivity is 85%, and the precision is 59%. The sensitivity of the high fall risk is 40% with a precision of 82%. Thus 40% of the high fall risk sequences were classified as a high fall risk, and 82% of all high-risk predictions were correctly classified. In Figure 12 can be seen that the low sensitivity of the high fall risk was mainly caused by high-risk sequences classified as increased risk.
The sensitivity and precision of the increased fall risk are 85% and 59%, respectively. Thus, 82% of the increased risk sequences were correctly classified. In the 3x3-matrix, Figure 12a, it can be seen that the increased sequences that were not predicted as increased were mainly predicted as low risk. The model is not a good classifier of sequences with low fall risk. Only 10% of the low fall risk sequences were correctly classified, and overall only 20% of the low-risk predictions were correct. The specificity of 92% shows that the model is better at predicting sequences that do not have a low risk than sequences that do.
In the 3x3-confusion matrix, the low fall risk sequences were mainly classified as increased risk, see Figure 12a. No sequences were classified as high risk. The wrong classification of low-risk sequences could be caused by the low number of total sequences with a low fall risk (n=31).
The ROC curves in Figure 12b show that the high fall risk sequences are best classified with an AUC of 0.80, while the increased fall risk sequences are least well classified with an AUC of 0.61. The imbalanced data could cause the discrepancy between the 3x3-confusion matrix and the ROC curve. Therefore, a weighted multiclass classification was also performed.

Weighted
In the weighted multiclass CNN, the low, increased, and high fall risk categories were weighted with 1.84, 0.67, and 1.04, respectively. The architecture of the best scoring weighted multiclass model is displayed in Table IV and the corresponding cross-entropy loss plot can be seen in Figure 11b. Based on this plot, the number of epochs was set to three.
The macro F1 score is 47%, which is only 1% higher than the unweighted multiclass classification. The sensitivity of the high and low fall risk increased with a maximum of 4% compared to the sensitivity of the unweighted model. The high fall risk has a sensitivity of 44% and precision of 81%, and the low fall risk category has a sensitivity of 13% and precision of 21%. Even though the slight increase, the model is not good at predicting high and low-risk sequences. The increased risk sequences are still best predicted.
In the 3x3-confusion matrix in Figure 13a, it can be seen that sequences with a low fall risk are mainly predicted as increased fall risk. The model barely misclassifies high-risk sequences as low-risk and vice versa.
The ROC curve in Figure 13b again shows that the increased risk (AUC=0.61) was less well predicted compared to the high (AUC=0.79) and low (AUC=0.66) risk. The fact that the values in Table VI show better results than the ROC could be caused, by the fact that those values are based on one point in the ROC curve, while the ROC curve gives a general view of the overall performance for different threshold.

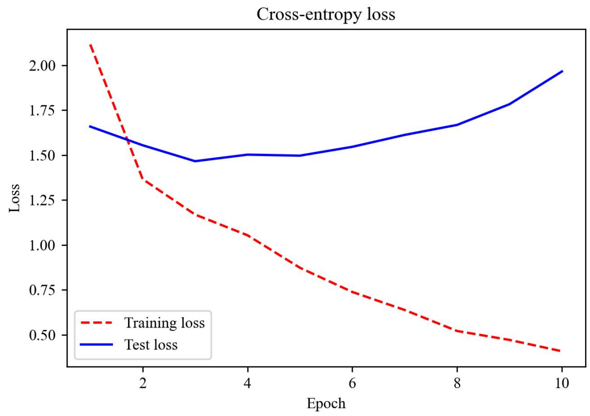
Binary classification
In the weighted binary CNN, the low and high risks were respectively weighted as 2.83 and 0.61. The CNN's architecture can be seen in Table VII. The number of epochs was set to five based on the cross-entropy loss function displayed in Figure 14.
The macro F1 score is 47%, the highest score compared to the SF and the multiclass classification models. This high score is caused by the high F1 score of the high-risk class, resulting from the higher sensitivity (46%) and precision (91%). Thus 46% of the high-risk sequences were correctly classified as high-risk sequences. Interestingly, the low fall risk sensitivity significantly increased to 77%, while the precision was comparable to the precisions of the other models (22%). This distribution is also visualised in the 2x2-confusion matrix of Figure 15a. Most low-risk sequences were predicted as low risk, while less than half of the high-risk sequences were

Table IV: CNN architecture for weighted and unweighted multiclass classification.

Class	(un)weighted	kernel size	# of filters layer 1	# of filters layer 2	# of filters layer 3	# of filters fully-connected layer	dropout	# of epochs
Multi	unweighted	3	40	48	96	224	0.5	3
Multi	weighted	2	8	64	-	240	0.2	1

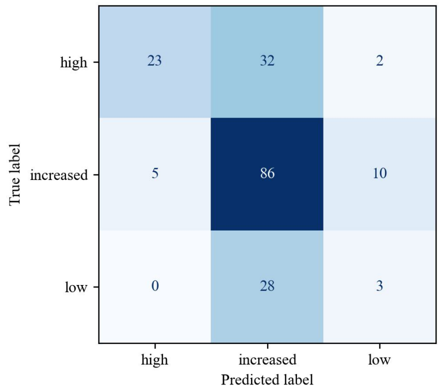


(a) Unweighted

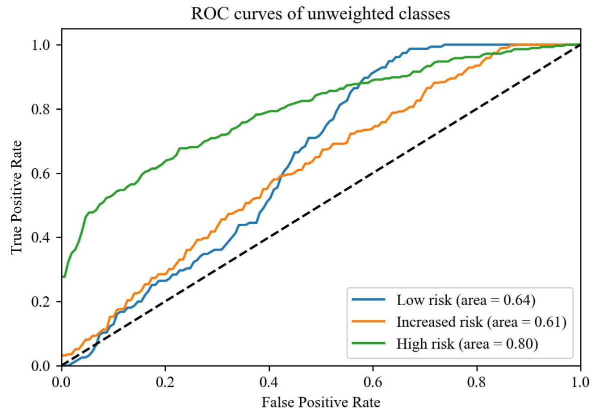


(b) Weighted

Figure 11: Cross-entropy loss functions of multiclass classification models with unweighted (a) and weighted (b) classes. Based on these graphs, the number of epochs of the unweighted model was set to four. The number of epochs of the weighted model was set to three.



(a) 3x3-confusion matrix



(b) ROC curve

Figure 12: The 3x3-confusion matrix (a) and ROC curve (b) of unweighted multiclass classification.

Table V: Performance of unweighted multiclass classification.

Fall risk	Sensitivity	Specificity	Precision	F1-score
High	40%	96%	82%	54%
Increased	85%	32%	59%	70%
Low	10%	92%	20%	13%
Macro avg	**45%**	**73%**	**54%**	**46%**

Table VI: Performance of weighted multiclass classification.

Fall risk	Sensitivity	Specificity	Precision	F1-score
High	44%	95%	81%	57%
Increased	81%	35%	59%	68%
Low	13%	91%	21%	16%
Macro avg	**46%**	**74%**	**54%**	**47%**

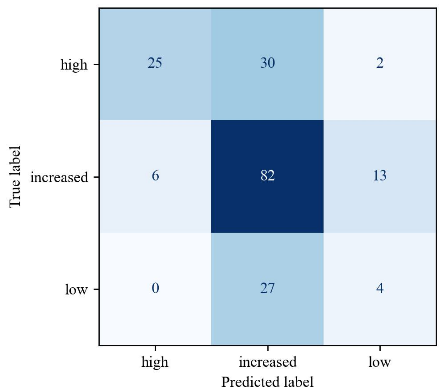
predicted as high risk. The ROC curve in Figure 15b gives the predictive capability of the binary model with the high class as a positive label. The model has an AUC of 0.64, which is lower than the high-risk AUCs of the weighted multiclass model.

IV. Discussion

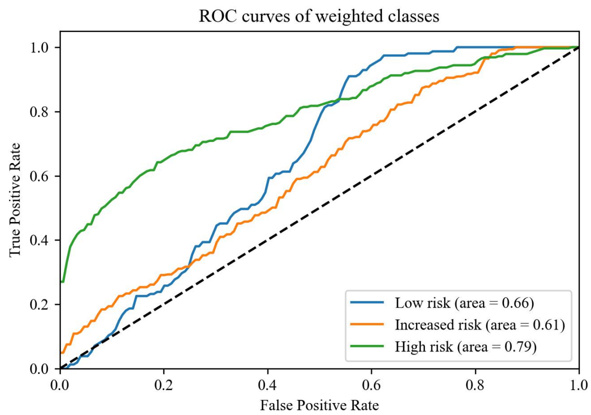
This study's objective was to predict the fall risk of people from a geriatric rehabilitation department using raw data collected with IMUs positioned at the ankles by developing a DL algorithm. A CNN was developed, trained and tested with raw IMU data collected during the 10MWT to predict the older adults' POMA fall risk category. The performance

Table VII: CNN architectures of the binary weighted classification model.

Class	(un)weighted	kernel size	# of filters layer 1	# of filters layer 2	# of filters layer 3	# of filters fully-connected layer	dropout	# of epochs
Binary	weighted	4	48	96	88	32	0.3	7



(a) 3x3-confusion matrix



(b) ROC curve

Figure 13: The 3x3-confusion matrix (a) and ROC curve (b) of weighted multiclass classification.

Table VIII: Performance of weighted binary classification.

Fall risk	Sensitivity	Specificity	Precision	F1-score
High	46%	77%	91%	61%
Low	77%	46%	22%	34%
Macro avg	**61%**	**61%**	**56%**	**47%**

of the SF for predicting the POMA fall risk categories of the same participants was also determined.

The developed CNNs, as well as the SF, had a poor classification performance. The SF had a macro F1 score of 49%, which is higher than the macro F1 score of the unweighted and the weighted multiclass classification CNNs, which had a macro F1 score of 40% and 41%, respectively. Even if the models had performances equal to the SF, the performance of the

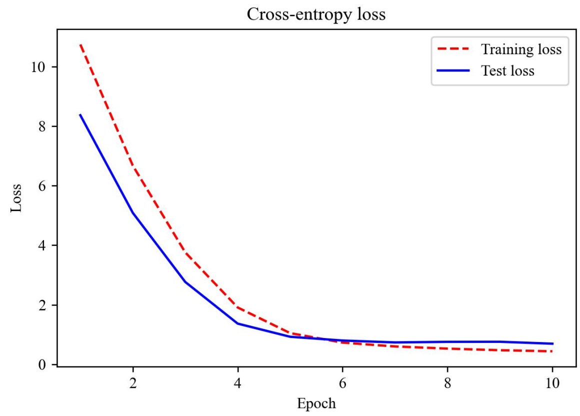


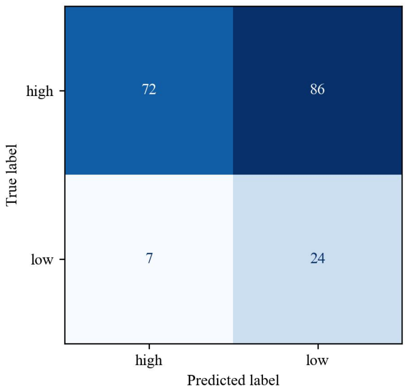
Figure 14: Cross-entropy loss functions of the binary classification model. Based on this plot, the number of epochs was set to 4.

models is low. The ROC curves of the multiclass unweighted and weighted showed high performance for the high fall risk sequences with an AUC of 0.80 and 0.79, respectively. The comparison with the increased and low risk combined will only be of clinical value if only the high fall risk class is essential to be determined.

The binary classification model had a macro F1 score of 56%, slightly higher than the SF performance. This model combined the high and increased models into one class. This model will only have a clinical value if the indication for a fall prevention program is equal in case of increased and high fall risk. Yet, 56% is still very low, meaning that it is not a reliable POMA fall risk classification model.

The low classification performance of the SF can be explained by the fact that the algorithm of the SF was only trained with data collected from participants living in a nursing home. This study tested the performance with data collected from participants temporarily living in a rehabilitation department.

The multiclass classification CNNs in this study had a comparable macro F1 to the low F1 score in the prospective gyroscope CNN model of Roshdibenam et al. (2021), which had an F1 score of 0.41 and an AUC of 0.56 [15]. In this study, all models' AUCs of all fall risk categories were higher. In contrast, the studies of Nait Aicha et al. (2018) and Tunca, Salur and Ersoy (2020) showed better results with an AUC of 0.75 and 0.82, respectively [13, 14]. Only the high fall risks of the multiclass unweighted and weighted classification models were close to these values. The difference in patient characteristics, sensor position, and categorisation method might explain the performance difference between the



(a) 2x2-confusion matrix

(b) ROC curve

Figure 15: The 2x2-confusion matrix (a) and ROC curve (b) of weighted binary classification.

current study's models and those found in other studies.
In the current study, the participants suffered multiple conditions, which might complicate fall risk prediction based on the gait analysis. For example, the difference in gait symmetry is associated with osteoarthritis and hip replacement, while reduced coordination is associated with rheumatoid arthritis and Parkinson's disease [23]. The participants in the study of Roshdibenam et al. (2021) also suffered multiple conditions [15], which could clarify the lower performance compared to the other studies with better performances. In those studies, the participants were healthy older adults and participants with neurological disorders [13, 14].
Positioning of the sensors in combination with the complexity of the task influences the performance of a model [7]. A more complex task is associated with better performance. In the current study, the 10MWT was performed, which only contained straight walking. The sensors were positioned on the ankles. In the better performing study of Nait Aicha et al. (2020), the participants wore the accelerometer on the lower back for a week during ADL [13]. ADL also consists of more complex tasks, which might explain the better performance of the model. In the study of Roshdibenam et al. (2021), the results of the different sensor positions were also evaluated [15]. The sensor on the neck outperformed the sensors on the left and right foot when worn during the TUG. However, the accelerometer and gyroscope data were separated. At the same time, the combination of both sensors might give more valuable information, which was done in the study of Tunca, Ersoy and Salur (2018) [14]. In that study, the participants performed straight walking with sensors on their feet.
The data collection methods of Tunca, Ersoy and Salur (2018) are comparable to the current study, except that the sensors were positioned at the ankle. However, the performance was worse than the performance in the study of Tunca, Ersoy and Salur (2018). Among others, the use of the resultants of the accelerometer and gyroscope could cause this. If the signals were out of phase, they were zeroed when the resultant was calculated. Additionally, the negative and positive orientations were removed by the calculations. As a result, there could have been a loss of characteristics that might be important for the fall risk classification.
The categorisation method in the current study was based on the POMA fall risk assessment. However, a well-performing model for classifying participants based on a fall risk assessment does not guarantee a good performance when predicting prospective falls [15]. The usefulness of the POMA for predicting prospective falls is limited [10]. Participants with a high POMA score have a chance of 50% for a prospective fall in six months, which is only 20% higher than the chance of falling without an assessment [7]. Furthermore, the POMA is better at predicting people who will not experience a fall than people who will [24]. The use of the POMA as fall risk assessment is thus questionable and might have affected the performance of the model.
This study has three main limitations. First, data resampling was required due to the timestamp's varying sampling frequencies and associated distributions. The samples with identical timestamps were split into equally spaced timestamps, resulting in exceptionally high sampling frequencies.
Second, Bayesian optimisation with 60 iterations and a selection of hyperparameters was used for the hyperparameter optimisation. This method was chosen because of the available time and materials. The low number of iterations in combination with the Bayesian optimisation might have resulted in missing valuable combinations of hyperparameters.
Third, the left and right sensor sequences were separately used as input for the CNNs. Therefore, the model could not have detected asymmetry in the gait pattern, while this is a fall risk characteristic in, for example, stroke patients [25].

A. *Recommendations*

In the future, multiple adjustments could be made to improve the model's performance resulting in a better fall risk prediction.
When the IMUs are positioned the same in every measurement, all three axes of sensors can be used individually. This enables the identification of characteristics in every direction and guarantees that the signals are not zeroes when they are

out of phase.

Combining the sequences of the sensors on the left and right ankles makes it possible to let the model identify asymmetry characteristics, a meaningful fall risk characteristic [25].

Furthermore, only the time domain sequences were used to develop a CNN, while adding the signals in the frequency domain to the input signals could be promising. Participants with a history of falling could have a lower amplitude and slope of the dominant frequency in the vertical axis and a higher amplitude and slope in the mediolateral axis compared to people without a history of falling [26].

It will be of added value to use more iterations or a Grid search for the optimisation if the circumstances allow it. This reduces or removes the chance that essential combinations are missed.

Collecting data during more complex tasks or ADL might give a better classification performance [7]. When the sensors are worn during ADL, this gives better insight into the movement of people in their living environment while performing daily tasks. Nait Aicha et al. (2018) also studied sensor data collected during ADL, which showed promising results [13].

For predicting prospective falls, the positioning of the sensors might also be reconsidered. Wearing a 3D accelerometer on the lower back showed the capability to predict prospective falling after a 25m walk test, while a 3D accelerometer on the shanks did not [27]. When ADL is used as an assessment, the sensors should be positioned on the lower back [13].

Finally, predicting prospective falls would have higher clinical relevance than predicting the POMA fall risk classification, of which the classification performance is questionable. The model could then be used to predict whether someone is prone to falling in the coming semester or not. This would make it easier to decide whether a fall risk prevention program and what type of program should be started.

V. Conclusion

The multiclass CNNs had a lower performance than the SF and binary classification. All CNNs were better at predicting high-risk sequences. However, the models had a poor performance when all three POMA fall risk categories should have been predicted. Adjustments to the data collection and CNN optimisation methods should be performed to study the possibility of predicting fall risk using raw IMU data in geriatric rehabilitation centres.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, dr. ir. E. van der Kruk for her support, helpful contact and critical view during my thesis. Furthermore, I would like to offer my thanks to Drs. F. Lefeber and the other employees of Smart Floor for the assitance and sharing the data. I would also express my gratitude to prof. dr. F.C.T. van der Helm for the insights he gave me and the interesting conversations we had. Also, I am thankful for the great explanations of Dr. D.M.J. Tax. Finally, I must thank my family and friends for their moral support during the project.

References

[1] World Health Organization, “Falls,” 04 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls

[2] VeiligheidNL, “Feiten cijfers — Valpreventie — VeiligheidNL,” 09 2020. [Online]. Available: https://www.veiligheid.nl/valpreventie/feiten-cijfers

[3] T. Masud and R. O. Morris, “Epidemiology of falls,” *Age and ageing*, vol. 30, no. suppl_4, pp. 3–7, 2001.

[4] B. J. Vellas, S. J. Wayne, L. J. Romero, R. N. Baumgartner, and P. J. Garry, “Fear of falling and restriction of mobility in elderly fallers,” *Age and ageing*, vol. 26, no. 3, pp. 189–193, 1997.

[5] J. Howland, E. W. Peterson, W. C. Levin, L. Fried, D. Pordon, and S. Bak, “Fear of falling among the community-dwelling elderly,” *Journal of aging and health*, vol. 5, no. 2, pp. 229–243, 1993.

[6] M. E. Tinetti and C. S. Williams, “The effect of falls and fall injuries on functioning in community-dwelling older persons,” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 53, no. 2, pp. M112–M119, 1998.

[7] C. M. Claassen, “Predictability of fall risk assessments in community-dwelling older adults: a scoping review,” Master’s thesis, Department of Biomechanical Engineering, Delft University of Technology, 2022.

[8] G. Zhao, L. Chen, and H. Ning, “Sensor-based fall risk assessment: A survey,” in *Healthcare*, vol. 9, no. 11. Multidisciplinary Digital Publishing Institute, 2021, p. 1448.

[9] K. Drost, F. Lefeber, and R. Dierkx, “Een innovatief meetinstrument voor valrisico bij ouderen,” Tech. Rep., 12 2020. [Online]. Available: https://www.zorginnovatie.nl/sites/default/files/test-reports/2012-04%20NTGF_Artikel%20SmartFloor.pdf

[10] M. J. Faber, R. J. Bosscher, and P. C. van Wieringen, “Clinimetric properties of the performance-oriented mobility assessment,” *Physical therapy*, vol. 86, no. 7, pp. 944–954, 2006.

[11] Meetinstrumenten in de zorg, “Performance-Oriented Mobility Assessment (volgens Tinetti) –,” 12 2020. [Online]. Available: https://meetinstrumentenzorg.nl/instrumenten/performance-oriented-mobility-assement-volgens-tinetti/

[12] K. Vriends-Evers, “De construct validiteit van de smart floor als meetinstrument voor de geriatrische revalidatie: een cross-sectioneel onderzoek,” Master’s thesis, Geriatric Physical Therapy, Avans+, 2021.

[13] A. Nait Aicha, G. Englebienne, K. S. Van Schooten, M. Pijnappels, and B. Kröse, “Deep learning to predict falls in older adults based on daily-life trunk accelerometry,” *Sensors*, vol. 18, no. 5, p. 1654, 2018.

[14] C. Tunca, G. Salur, and C. Ersoy, “Deep learning for fall risk assessment with inertial sensors: Utilizing domain knowledge in spatio-temporal gait parameters,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 1994–2005, 2019.

[15] V. Roshdibenam, G. J. Jogerst, N. R. Butler, and S. Baek, “Machine learning prediction of fall risk in older adults using timed up and go test kinematics,” *Sensors*, vol. 21, no. 10, p. 3481, 2021.

[16] J. Bezold, J. Krell-Roesch, T. Eckert, D. Jekauc, and A. Woll, “Sensor-based fall risk assessment in older adults with or without cognitive impairment: a systematic review,” *European review of aging and physical activity*, vol. 18, no. 1, pp. 1–14, 2021.

[17] Python Software Foundation, “Python.” [Online]. Available: https://www.python.org

[18] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, “A Survey on Graphical Methods for Classification Predictive Performance Evaluation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, pp. 1601–1618, 2011.

[19] Meetinstrumenten in de zorg, "10 Meter Walk Test / Tien meter looptest –," 11 2020. [Online]. Available: https://meetinstrumentenzorg.nl/instrumenten/10-meter-walk-test-tien-meter-looptest/

[20] L. J. Viccaro, S. Perera, and S. A. Studenski, "Is timed up and go better than gait speed in predicting health, function, and falls in older adults?" *Journal of the American Geriatrics Society*, vol. 59, no. 5, pp. 887–892, 2011.

[21] V. Hornyak, J. M. VanSwearingen, and J. S. Brach, "Measurement of gait speed," *Topics in geriatric rehabilitation*, vol. 28, no. 1, pp. 27–32, 2012.

[22] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[23] U. Lindemann, "Spatiotemporal gait analysis of older persons in clinical practice and research," *Zeitschrift für Gerontologie und Geriatrie*, vol. 53, no. 2, pp. 171–178, 2020.

[24] P. R. Trueblood, N. Hodson-Chennault, A. McCubbin, and D. Youngclarke, "Performance and impairment-based assessments among community dwelling elderly: sensitivity and specificity," *Journal of geriatric physical therapy*, vol. 24, no. 1, pp. 2–6, 2001.

[25] T.-S. Wei, P.-T. Liu, L.-W. Chang, and S.-Y. Liu, "Gait asymmetry, ankle spasticity, and depression as independent predictors of falls in ambulatory stroke patients," *PloS one*, vol. 12, no. 5, p. e0177136, 2017.

[26] A. Weiss, M. Brozgol, M. Dorfman, T. Herman, S. Shema, N. Giladi, and J. M. Hausdorff, "Does the Evaluation of Gait Quality During Daily Life Provide Insight Into Fall Risk? A Novel Approach Using 3-Day Accelerometer Recordings," *Neurorehabilitation and Neural Repair*, vol. 27, no. 8, pp. 742–752, 2013.

[27] L. Bizovska, Z. Svoboda, M. Janura, M. C. Bisi, and N. Vuillerme, "Local dynamic stability during gait for predicting falls in elderly people: A one-year prospective study," *PloS one*, vol. 13, no. 5, p. e0197091, 2018.

VI. Appendix

A. POMA

The POMA measurement tool that can be used to score each item of the POMA and calculate the total score.

Tinetti Performance Oriented Mobility Assessment (POMA)	Date
Balance Tests: Subject is seated on hard, armless chair	
SITTING BALANCE Leans or slides in chair =0, Steady, safe =1	
ARISES Unable without help =0; Able, uses arms =1, Able without using arms = 2	
ATTEMPTS TO RISE: Unable w/o help=0; Able, requires > 1 attempt =1; Able in 1 attempt =2	
IMMEDIATE STANDING BALANCE (first 5 seconds) Unsteady (sway/stagger/feet move)=0; Steady, w/ support =1;Steady w/o support =2	
STANDING BALANCE Unsteady =0; Steady, stance > 4 inch BOS & requires support =1; Narrow stance, w/o support =2	
STERNAL NUDGE (feet close together) Begins to fall =0; Staggers, grabs, catches self =1; Steady =2	
EYES CLOSED (feet close together) Unsteady =0; Steady =1	
TURNING 360 DEGREES Discontinuous steps =0; Continuous steps =1	
TURNING 360 DEGREES Unsteady (staggers, grabs) =0;Steady =1	
SITTING DOWN Unsafe (misjudges distance, falls) =0;Uses arms, or not a smooth motion =1; Safe, smooth motion =2	
BALANCE SCORE TOTAL	/16
Gait Tests: Subject stands with examiner, walks down hallway or across room, first at "usual" pace, then back at "rapid, but safe" pace (using usual walking aids)	
GAIT INITATION (immediate after told "go) Any hesitancy, multiple attempts to start =0; No hesitancy =1	
STEP LENGTH R swing foot passes L stance leg =1; L swing foot passes R =1	
FOOT CLEARANCE R foot completely clears floor =1; L foot completely clears floor =1	
STEP SYMMETRY R and L step length unequal =0; R and L step length equal=1	
STEP CONTINUITY Stop/discontinuity between steps =0; Steps appear continuous =1	
PATH (excursion) Marked deviation =0; Mild/moderate deviation or use of aid =1; Straight without device=2	
TRUNK Marked sway or uses device =0; No sway but knee or trunk flexion or spread arms while walking =1; None of the above deviations=2	
BASE OF SUPPORT Heels apart =0; Heels close while walking =1	
GAIT SCORE TOTAL	/12
ASSISTIVE DEVICE	
TOTAL SCORE (BALANCE + GAIT)	/28
FALL RISK **(minimal >23, Mod. 19-23, High < 19)**	
Therapist initials	

Figure 16: POMA score sheet [11]

B. Splitting timestamps with resampling

Table IX: An example of a dataframe with identical timestamps.

human_timestamp	timestamp	accel_x	accel_y	accel_z	gyro_x	gyro_y	gyro_z
2021-03-23 08:57:15.036	1616489835036	-5.6	-15.9	-4.1	1.6	0.2	-1.2
2021-03-23 08:57:15.037	1616489835037	-4.9	-17.5	-3.2	2.3	0.0	-1.9
2021-03-23 08:57:15.038	1616489835038	-2.0	-18.8	-1.3	3.4	-0.0	-3.1
2021-03-23 08:57:15.038	1616489835038	-3.6	-18.9	-1.9	3.0	-0.1	-2.6
2021-03-23 08:57:15.039	1616489835039	-0.6	-15.5	-1.2	3.7	0.5	-3.7
2021-03-23 08:57:15.039	1616489835039	-0.8	-17.5	-1.2	3.6	0.2	-3.5
2021-03-23 08:57:15.040	1616489835040	-0.7	-13.3	-0.6	3.6	0.4	-3.8
2021-03-23 08:57:15.041	1616489835041	-0.8	-10.9	-0.1	3.5	0.3	-3.6
2021-03-23 08:57:15.041	1616489835041	-1.0	-8.8	0.3	3.4	0.1	-3.4
2021-03-23 08:57:15.042	1616489835042	-1.1	-7.3	0.3	3.2	-0.1	-3.2
2021-03-23 08:57:15.042	1616489835042	-1.1	-6.2	0.1	2.9	-0.3	-3.0

Table X: Dataframe after linear interpolation to create equally spaced timestamps from identical timestamps.

human_timestamp	timestamp	accel_x	accel_y	accel_z	gyro_x	gyro_y	gyro_z
2021-03-23 08:57:15.036000000	1.616490e+12	-5.6	-15.9	-4.1	1.6	0.2	-1.2
2021-03-23 08:57:15.036999936	1.616490e+12	-4.9	-17.5	-3.2	2.3	0.0	-1.9
2021-03-23 08:57:15.038000128	1.616490e+12	-2.0	-18.8	-1.3	3.4	-0.0	-3.1
2021-03-23 08:57:15.038500096	1.616490e+12	-3.6	-18.9	-1.9	3.0	-0.1	-2.6
2021-03-23 08:57:15.039000064	1.616490e+12	-0.6	-15.5	-1.2	3.7	0.5	-3.7
2021-03-23 08:57:15.039500032	1.616490e+12	-0.8	-17.5	-1.2	3.6	0.2	-3.5
2021-03-23 08:57:15.040000000	1.616490e+12	-0.7	-13.3	-0.6	3.6	0.4	-3.8
2021-03-23 08:57:15.040999936	1.616490e+12	-0.8	-10.9	-0.1	3.5	0.3	-3.6
2021-03-23 08:57:15.041499904	1.616490e+12	-1.0	-8.8	0.3	3.4	0.1	-3.4
2021-03-23 08:57:15.041999872	1.616490e+12	-1.1	-7.3	0.3	3.2	-0.1	-3.2
2021-03-23 08:57:15.042500096	1.616490e+12	-1.1	-6.2	0.1	2.9	-0.3	-3.0

Table XI: Dataframe after resampling with 19ms.

human_timestamp	accel_x	accel_y	accel_z	gyro_x	gyro_y	gyro_z
2021-03-23 08:57:15.020	-4.25	-18.15	-2.55	2.65	0.00	-2.25
2021-03-23 08:57:15.039	-0.70	-11.10	1.00	-0.90	0.30	0.80
2021-03-23 08:57:15.058	-0.15	-9.80	2.20	-0.40	0.00	0.50
2021-03-23 08:57:15.077	-0.85	-11.70	-0.10	0.75	0.05	-1.00
2021-03-23 08:57:15.096	-0.40	-10.10	2.00	-0.50	0.00	0.70
2021-03-23 08:57:15.115	-0.25	-9.85	-0.00	-0.00	-0.10	0.25
2021-03-23 08:57:15.134	-0.70	-10.50	1.30	-0.90	0.20	0.80
2021-03-23 08:57:15.153	0.10	-9.70	2.55	-0.40	-0.10	0.55

C. Rewritten literature review

Predictability of fall risk assessments in community-dwelling older adults: a scoping review

August 11, 2022
N.F.J.Waterval, C. M. Claassen, F.C.T. van der Helm, E. van der Kruk

Abstract— **Background: Fall risk increases with age, and in adults over 65 years old one-third experiences a fall annually. Due to the aging population, the number of falls and related medical costs will progressively increase. Prediction of whom will fall in the future may help to timely intervene and reduce the number of falls. Therefore, the aim of this scoping review is to determine the predictive value of fall risk assessments in community-dwelling older adults using prospective studies.**
Methods: 37 studies were included that evaluated clinical assessments (questionnaires, physical assessments, or a combination), sensor-based clinical assessments, or sensor-based daily life assessments using prospective study designs. The posttest probability of falling or not-falling was calculated.
Results: In general, fallers were better classified than non-fallers. Questionnaires had a lower predictive capability compared to the other assessment types. Contrary to conclusions drawn in reviews that include retrospective studies, the predictive value of physical tests evaluated in prospective studies varies largely, with only smaller sampled studies showing good predictive capabilities. Sensor-based fall risk assessments performed best. However, predictive value of sensor-based assessments has only been evaluated in relatively small samples. The performance of sensor-based improves with task complexity; sensor-data of straight walking was insufficient to predict future fallers.
Conclusion: Fall risk prediction using sensor-data collected during clinical test or daily living seems to outperform conventional tests, but their validity needs to be confirmed by large prospective studies.

Index Terms— Fall risk assessment, prediction, community-dwelling, older adults, clinical, sensor-based, validity

I. Introduction

One-third of people above 65-years of age experience a fall annually, and in more than 50 percent of the cases medical assistance is needed [1–3].Besides direct injuries, elderly who experience a fall often suffer from fear of falling [4], reduced activity [5, 6], and a lower quality of life [5]. Due to the rapidly aging population, fall prevalence and associated medical cost will progressively increase in the next decades.

Fall risk factors can be divided into extrinsic and intrinsic risk factors [7–9]. Extrinsic risk factors are external to the individual and are also called environmental factors. These factors include poor lighting, unsafe stairs, slippery floors, a loose carpet, or unsafe footwear [7, 10]. Intrinsic factors can be divided into age-related physiological changes, pathological predisposing factors[9], and drugs[11]. To maintain balance, multiple physiological systems need to work in synergy, such as the sensory system, central nervous system and motor system [12]. Most falls in institutionalized elderly are the consequence of age-related reduction of physiological capacity [11]. The sensory system loses sensitivity due to a loss of sensors, while response to stimuli is delayed and less effective due to a reduction in white and grey matter in the central nervous system and reduced nerve conduction velocity [12]add citation: van der Kruk, E., Silverman, A. K., Koizia, L., Reilly, P., Fertleman, M., Bull, A. M. (2021). Age-related compensation: Neuromusculoskeletal capacity, reserve movement objectives. Journal of Biomechanics, 122, 110385.. Intramuscular changes cause reduced muscle force and speed of force transmission [12]

Several fall risk assessments have been developed which aim to identify people at risk of falling to enable timely prescription of fall prevention programs and assistive devices [13]. Clinical fall risk assessments have been used for decades, and include questionnaires (e.g., Falls Efficacy Scale-international (FES-i)[14], Activity specific Balance Confidence (ABC-scale)[15]) or physical tests (e.g., Tinetti Performance Oriented Mobility Assessment (POMA) [16], Timed Up and Go test (TUG)[17]). Sensor-based fall risk assessments can provide a more objective and less time-consuming approach [18, 19]. These assessments can be sensor-based clinical assessments, where a sensor is used during the performance of a clinical fall risk assessment[20, 21], or sensor-based activities of daily life (ADL) assessments, where sensors are used at home during daily life activities[22, 23]. Either wearable (e.g., inertial measurement unit (IMU)) or portable sensors (e.g., Microsoft Kinect, pressure sensor) have been used in this approach [19].

The validity of these fall risk assessments have been described in various recent reviews, focusing on sensor-based [24, 25] or clinical fall risk assessments [26]. However, these reviews included retrospective studies, while to determine the validity and diagnostic value of fall risk assessments, prospective falls are considered the gold standard [27]. Consequently, a comprehensive overview of the predictive capability of both sensor-based and clinical fall risk assessment based on prospective falls was lacking. Furthermore, validity of sensor-based clinical and ADL assessments depends on "sensor location, sensor attachment and the type of assessment chosen for the recording of sensor data" [25], yet this was not considered in previous reviews. The primary aim of this review is therefore to re-examine the validity of fall risk assessments to predict future falls in community-dwelling older adults when only studies with a prospective study design are included.

II. Methods

A. Search strategy

In December 2021 the databases of Scopus, PubMed, IEEE Xplore, and Web of Science were searched using the following search string: (fall risk predict* OR fall risk assess* OR fall risk classif* OR fall risk measur*) AND ("older adult" OR aged OR elder* OR senior* OR geriatric) in the title, abstract or keywords and in all fields was searched for (accura* OR sensitiv* OR specific*). The search strings for the different databases are described in the supplementary material -A. Results were exported to EndNote (EndNote X9.3.3, Philadelphia, USA) for further analysis.

Table I: 2x2-matrix

	(Recurrent) fallers	Non-fallers	Total
Positive assessment	TP	FP	Participants with positive assessment
Negative assessment	FN	TN	Participants with negative assessment
Total	Fallers	Non-fallers	Total participants

TP: true positive, FN: falls negative, TN: true negative, FP: falls positive

B. Selection criteria

Duplicates, conference proceedings, books, and serials were removed from the results. The search included 50 high-quality review articles on fall prediction published between 19... and 2020. Therefore, the selection of high-quality studies was made by searching the references from these recently published reviews using the following inclusion criteria: i) inclusion of community-dwelling people with an age ≥60, or the mean age minus the standard deviation is ≥60, ii) use of prospective methods for the categorization of fallers/non-fallers or low/high fall risk to avoid recall bias [28, 29], iii) data are prescribed (or could be determined) on the number of fallers who were positive or negative for the fall risk assessment or summary statistics (sensitivity, specificity, or area under the curve (AUC)) and iv) written in English or Dutch. Studies were excluded if i) assistance by another person was allowed during the assessment, or ii) the focus was only on the detection of (near) falls instead of predicting falls.

C. Data extraction

Articles were categorized based on the type of fall risk assessments: clinical (without sensors) or sensor-based (during clinical or ADL assessments). The following data were extracted: number of participants, number of fallers and fall criteria, follow-up time, percentage female, mean age (±SD), sensitivity, specificity, and positive and negative likelihood ratio.
For clinical assessments, studies that reported a predictive value in multivariate analysis and that reported validity measures were included. The specific assessment and cut-off score were extracted.
For the sensor-based assessments, sensor-type, sensor location, type of assessment, cut-off score, and validity measures were extracted.

D. Analysis

The precision of fall risk assessments has been quantified with various measures, such as sensitivity and specificity, number of true positives, the positive and negative likelihood ratios (LR) and posttest probability (PoTP) Table II. To compare results, we calculated the PoTP for each study. The PoTP defines how much fall risk has shifted compared to pretest probability (PrTP) [30]. For example, falls have a prevalence of 30% in the population of older adults, so the chance of falling is 30%. If a fall risk assessment has a +PoTP of 60%, then a person with a positive assessment has a 60% chance of falling, while rhe PrTP was 30%. On the other hand, if the -PoTP is 20%, then the chance someone with a negative assessment will fall reduces from 30% to 20%. Ideally, an assessment would have a high +PoTP and a low -PoTP, which indicates that people with a positive assessment experience a fall, and with a negative assessment do not fall.
The PrTP in this study is set at 30%, based on the fall risk of older adults in the general population[1, 2, 30]. The PoTP is calculated based on the available measures [30].

III. Results

A. Study selection

The combination of the results from IEEE Xplore (n=91), PubMed (n=152), Scopus (n=1168), and Web of Science (n=256) resulted in 1667 articles. After removing 437 duplicates, 1230 articles were left, of which 1127 were journal articles. After scanning the titles and abstracts, 50 reviews were included for full-text screening, of which 178 articles were extracted. After the full-text screening of these articles, 37 prospective studies were included. In Figure 1 the flow chart of the study selection has been displayed.

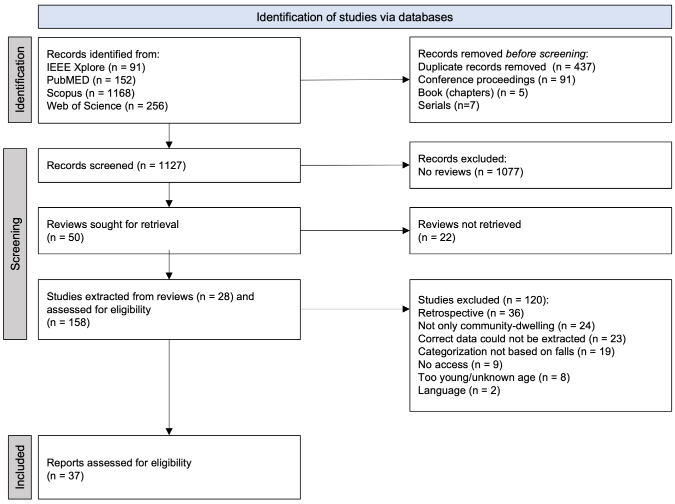


Figure 1: Flowchart of study selection.

Thirty articles focused on clinical assessments, based on questionnaires or physical tests, while two focused on sensor-based ADL assessments, and five on sensor-based clinical assessments.

B. Clinical assessments without sensors

Thirty articles focused on clinical assessments without sensors, published between 2000 and 2018. Study characteristics are displayed in Table III. In total, the studies included 12406 participants of whom 3084 were classified as fallers based on prospective data. The follow-up period varied from six months to three years [31–40]. The most used fall criterion was "at least one fall during follow-up"[31–33, 35–39, 41–53].
The 30 studies used a total of 39 different clinical assessments without sensors (explained in Appendix-B). These assessments can be classified as questionnaires, questionnaires combined with physical performance, and only physical performance. The performance of each of the assessments is displayed in Table IV.

Questionnaires

The predictive value of questionnaires were evaluated by a total of five studies evaluating four questionnaires; the *Fall-risk screening test* [40], *Fall Risk for Older People in the Community (FROP-Com)*[50, 51], *Geriatric Depression Scale (GDS)* [44], and the combination of the *history of falls* and *independent bathing*[54].
The best performance for classification of fallers (+PoTP) based on questionnaires was the *Fall-risk screening test* (+PoTP:52%) [40]. This test was evaluated in one study (n=1285) with a 3-year follow-up. It demonstrated classification of non-fallers (-PoTP) of 20% [40]. The *FROP-Com* was evaluated in two studies with a 12 month follow-up and had a slightly lower +PoTP, but a slightly better -PoTP: (+PoTP:46%,-PoTP:17%) (n=344) [50] and (+PoTP:44%,-PoTP:18%)(n=192) [51]. The +PoTP of *GDS* was slighlty better than *FROP-Com* (48%), yet the -PoTP was worse (24%)(n=260) [44]. The combination of the *history of falls* and *independent bathing* was studied by one study for one fall in 12-months and classified non-fallers (-PoTP:13%) better than fallers (+PoTP:39%) (n=192) [54].

Physical performance

Assessments that use a stand-alone physical performance test are the

Table II: Explanation of validity terms.

Term	Definition	Calculation
True positive (TP)	Correct prediction of fallers	*sensitivity*fallers*
True negative (TN)	Correct prediction of non-fallers	*specificity*(non-fallers)*
False positive (FP)	Non-fallers that were incorrectly predicted as fallers	non-fallers$-TN$
False negative (FN)	Fallers that were incorrectly predicted as non-fallers	fallers$-TP$
Sensitivity	Percentage of fallers correctly identified	$\frac{TP}{TP+FN}$
Specificity	Percentage of non-fallers correctly identified	$\frac{TN}{TN+FP}$
Positive likelihood ratio (+LR)	Probability that a positive assessment will be expected in a person with a high fall risk divided by a person with a low fall risk	$\frac{sensitivity}{(1-specificity)}$
Negative likelihood ratio (-LR)	Probability that a negative assessment will be expected in a person with a low fall risk divided by a person with a high fall risk	$\frac{(1-\text{sensitivity}}{specificity}$
Pretest probability (PrTP)	Prevalence of falls in the population	30%
Pretest odds (PrTO)	The odds that the patient has a disorder before the assessment has been performed	$\frac{\text{PrTP}}{1-\text{PrTP}}$
Posttest odds (PoTO)	The odds that the patient has a disorder after the assessment results are known	*+PoTP=PrTO*(+LR)* or *-PoTP=PrTO*(-LR)*
Positive posttest probability (+PoTP)	Chance that a person with a positive assessment will fall	$\frac{\text{(+PoTO)}}{1+\text{(+PoTO)}}$
Negative posttest probability (-PoTP)	Chance that a person with a negative assessment will fall	$\frac{\text{-PoTO}}{1+\text{(-PoTO)}}$

alternate step test[55], *Adjusted maximum step length*[46], *Berg Balance Scale (BBS)[34, 47, 53], Dynamic Gait Index (DGI)*[39], *Functional Gait Assessment (FGA)*[39], *Five-times sit-to-stand (FTSS)*[56, 57], *Zur balance scale*[53], *Getting up from lying on the floor*[58], *One leg balance (OLB)*[57], *Tinetti Performance-Oriented Mobility Assessment (POMA)*[20, 38, 49, 52], *Risk assessment*[57], *stair ascent*[59], *Test battery*[45], *Timed gait*[52, 59], *TUG*[31, 33–39, 56], and *Walking While Talking (WWT)*[52].

The *TUG* had the highest predictive value, but the +PoTP varied from 31% to 91% and the -PoTP from 7% to 29% within the nine articles, all but one with a follow-up of 6 months Table III. The best scores (+PoTP:91%, -PoTP:7%) are from a small study (n=35), where the average of three *TUG* tests was used [39]. The next best results for the *TUG* had +PoTPs of 55% and 48% and -PoTPs of 25% and 15% using larger sample sizes (n=259 and n=60, respectively) [31, 33]. The studies with the largest sample size for the TUG evaluation (n=621 and n=868)($\geq$ 2 falls) had an even lower +PoTP of 42% and 37%, and -PoTP of 29% and 28%[34, 57].

The second best predictor of fall risk in fallers with a physical test only was the *WWT* (complex:+PoTP:79%, -PoTP:22%; simple:+PoTP:65%, -PoTP:21%)[52]. This test was however evaluated in a single prospective study with a relatively small sample size (n=59) and results should therefore be interpreted with caution. These same limitations (n$\leq$ 94) hold for the studies on *Zur balance scale* (+PoTP:74%, -PoTP:16%)[53], *FGA* (FGA)(+PoTP: 71%, -PoTP:0.00%)[39], the *DGI* (+PoTP: 64%, -PoTP: 0.00%)[39], the *classification tree* (+PoTP:53%, -PoTP:15%)[36], and the *test battery* (+PoTP:27%, -PoTP:33%)[45].

The POMA was verified in four studies, but in distinct forms: the complete short assessment (full POMA) [20, 38], the balance part of the short assessment only (14-item balance assessments)[20, 52], and the balance part of the long assessment only (9-task balance part)[49]. The different forms has comparable posttest probabilities, respectively: The full POMA (n=180 +PoTP:50%, -PoTP:27% [38], n=131 +PoTP:63%, -PoTP:15% [20]), 14-item balance assessments (n=225 +PoTP:38%,-PoTP:20% [52], n=131 +PoTP:65%, -PoTP:20% [20]) and the nine task balance part (n=59 +PoTP:48%,-PoTP:18%)[49].

The *risk model for recurrent falls, performance-based FRAT* ($\geq$6)(+PoTP:62%, -PoTP:23%) (n=362)[55], and *Getting up from lying on the floor* (+PoTP:55%, -PoTP:25%)(n=307) [58] demonstrated good predictive value in large cohort studies. Tiedemann et al. (2010) compared several cut-off scores for the performance-based FRAT (0-1,2-3,4-5, $\geq$ 6) and reported the best posttest probabilities to be $\geq$ 6 [55] .

Timed gait was used in different tasks. In the study of Verghese et al. (2002) (n=59), the duration of a participant to walk 6 meters, turn, and return at a normal walking speed was measured[52], while Tiedemann et al. (2008) (n=347) timed a 6 meter straight walk at a normal pace[59]. The timed gait with turn showed a slightly better predictive value (+PoTP:52%, -PoTP:24%)[52] than without the turn (+PoTP:40%, -PoTP:24%)[59].

All other physical performance based assessments had a +PoTP of less than 50% Table III.

Questionnaires combined with physical performance

Nine different assessment tools combining questionnaires and physical tests were evaluated: the *Downton Fall Risk Index (DFRI)*[35], *Risk model for recurrent falls*[60], *classification tree for recurrent falls*[34], *mJH-FRAT*[43], the *AGS/BGS/AAOS Fall screening algorithm*[48], *performance-based FRAT* (based on the five domains *PPA* and questions on fall history and medication use)[55],the *clinical screening tool* (questions and one leg balance test (OLB))[41], and the combination of the *Physical Performance Assessment (PPA)* and *FES-I*[42]. All combinations of questionnaires and physical assessments contained *history of falls* as one of the questions, except in the combination of the *PPA* and *FES-I*[42].

The best faller identification procedure was the *clinical screening tool* combined with the *OLB* with a +PoTP of 70% and -PoTOP of 26% (n=1759) [41]. This is over 30% higher and 3% lower +PoTP and -PoTP respectively compared to the OLB alone [41]. The best combination to classify non-fallers was the *DFRI*, evaluated by one study with a follow-up of 12 months (n=116) who found a -PoTP:0%, yet also a very low +PoTP of only 30% [35]. Three combinations resulted in +PoTP larger than 60%: The *Risk model for recurrent fallers* had a +PoTP of 66% and a -PoTP of 17% (n=287) (fall criterion: $\geq$ 2 falls follow-up 6 months)[60], the *classification tree for recurrent falls* had the same +PoTP (66%), but a higher -PoTP (23%) (n=868) (fall criterion: $\geq$ 2 falls follow-up 9 months)[34], the *performance-based FRAT* had a +PoTP (62%) and comparable -PoTP of 23% (n=362) (fall criterion: $\geq$ 2 falls follow-up 12 months)[55]. The mJH-FRAT (+PoTP: 59%, -PoTP: 14%)[43], the AGS/BGS/AAO Fall screening algorithm (+PoTP: 52%, -PoTP:21%)[48], and the combination of the *PPA* and the *FES-I* (n=600) (+PoTP: 37%, -PoTP: 23%) (fall criterion: $\geq$ 2 falls follow-up 12 months) each had a +PoTP below

60%.

C. Sensor-based assessments

In total, seven sensor-based assessments articles were included, of which five evaluated the performance of sensor-based clinical assessments, and two the performance of sensor-based ADL assessments. The articles were written between 2013 and 2018, and 1306 participants were included, of whom 695 were classified as fallers, see Table V. The fall criterion of at least one fall during follow-up was mostly used[21, 22, 61–64]. Three articles had a follow-up time of 6 months[62–64], three articles had a follow-up time of 12 months [20, 22, 61], and one article had a follow-up time of 24 months[21].
The seven studies used Inertial Measurement Units (IMUs) (accelerometer, gyroscope) placed on a variety of locations on the body Figure 2; Two studies combined the accelerometer with an insole pressure sensor under the plantar foot[63, 64].
The performance of each assessment is displayed in Table VI.

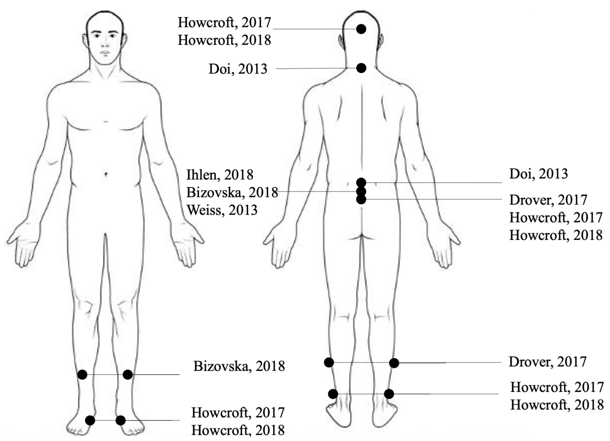


Figure 2: Locations of all wearable sensors of the included sensor-based assessments.

Sensor-based clinical assessments

The performance of identification of fallers using sensor-based clinical assessment is dependent on the task, the sensor location, feature extraction, and the classification method.

The tasks used in the prospective sensor-based studies were standardized walking tests [20, 61–64], of which four only used data from straight walking, and one straight walking and turns during a 6-minute walk test (6MWT)[62], see Table VI. The latter, using accelerometers on the shank, analysed straight walking and turns separately and combined. The data during a turn (+PoTP:59%, -PoTP: 17%) and turn and straight walking (+PoTP:58%, -PoTP: 18%) had a higher predictive value than straight walking alone (+PoTP:33%, -PoTP:26%) [62].

Howcroft, Koftmann, and Lemaire (2017 and 2018) compared sensor locations for optimal classification of prospective falls. They analyzed single-task (ST) and dual-task (DT) walking with sensors on the head, pelvis, ankles, and with a pressure insole [63, 64]. Based on their prospective study result they conclude that in DT the pelvis accelerometer had the best single-sensor predictive capability, while in ST the head location performs better. Overall, their conclusion was that the multi-location sensors outperformed the single-sensor approach. Bizovska et al. (2018) used sensors on the trunk and shanks and found that the sensors on the shanks did not contribute to a distinction between fallers and non-fallers.

Although the studies used similar sensors, feature extraction and classification methods differed. Drover et al. used the maximum, mean, and SD of acceleration in all directions of the three axes, acceleration frequency, and ratio of even/odd harmonics from sensors on the shank [62]. They found that the best overall machine learning method was a random forest classifier and five turn-based features selected with select-five best method from cross-validation.Doi et al. found that the harmonic rate in the vertical direction of the sensor on the upper trunk was the discriminate factor for a 15m straight walk (+PoTP:65%, -PoTP:14%) [61]. In the study of Bizovska et al. (2018), [20], the trunk medial-lateral (ML) acceleration in short term (slopes of mean log divergence curve between 0 and 0.5 stride) and Lyapunov exponents (stLE) had the best predictive power during a 25m straight walk (+PoTP:60%, -PoTP: 19%). Note that the non-sensor based POMA balance score (+PoTP:65%, -PoTP:20%) and the POMA total score (+PoTP:63%, -PoTP:15%) each had a better posttest probability value compared to the Trunk stLE ML alone. Howcroft, Koftmann, and Lemaire (2017 and 2018) found the sensors on the head and right shank to be discriminative for the ST-walking assessment using a vector machine (SVM) (+PoTP:33%, -PoTP:28%)[63, 64]. For the DT walking assessment, the sensor on the head, pelvis, and left shank showed discriminative power (+PoTP: 34%, -PoTP:27%) as determined using a neural network. In the study of 2018, the insole pressure and sensor on the left shank made the distinction between faller and non-faller by using a support vector machine: +PoTP: 65%, -PoTP:20%[64].

Sensor-based ADL assessments

Two studies used sensors in daily living; one for three days[23] and one for a week[22]. Both studies used a fall criterion of 2 falls in 6 months and accelerometer data on the lower back (lumbar spine). Weiss et al. combined the 3-day measurements with the Dynamic Gait Index (DGI) in 71 participants of whom 12 were fallers. Fallers and non-fallers could be classified by the total activity duration, DGI, and the anterior-posterior acceleration range and width, extracted from the frequency in the power spectral density[23]. The combination of DGI and 3-days ADL had a sublime result of +PoTP of 100% and a -PoTP of 10%).[23].
Ihlen et al. used the phase-dependent generalized multiscale entropy (PGME) to define time series irregularities in 303 participants. They investigated the high-frequency intra-step modulation of trunk acceleration signals of walking[22] and found a +PoTP of 74% and a -PoTP of 14% [22]. Combining the fall history, conventional gait, and demographic variables with the sensor resulted in a worse +PoTP (64%) and better -PoTP (8%).

IV. Discussion

The aim of this review was to provide an overview of the predictive value of current fall assessment tools in community-dwelling older adults. Assessment tools were classified as either questionnaires, physical performance, a combination of questionnaires and physical performance and sensor based assessment. For all classification groups, in general, fallers are better classified than non-fallers. Questionnaires have a lower predictive value compared to the other groups, while sensor-based assessment seem to perform best, but have mainly been studied in relatively small samples.
As expected, questionnaires have a lower predictive capability than the other assessment types, no study found a +PoTP value above 52%. The predictive value was improved by combining questionnaires with physical performance tests and include *fall history* in the questionnaires. All combined assessments analyzed for this review used *fall history* except for the *PPA* in combination with the *FES-I*. The performance of the *PPA* increased with addition of the *history of falls* and *medication use* as was seen in the *performance-based FRAT*. However, taking fall history into account has its limitations as it can not be used to predict elderly at risk of becoming fallers. Therefore, although easy to administer, the use of questionnaires to predict future falls is not sufficient.

Table III: Study characteristics of clinical assessments.

Author, year	Total (n)	Female (%)	Mean age (SD)	Fallers	Fall criteria	Follow-up time (months)	Tool
Alexandre, 2012	60	51.7	66	SF: 19	>=1	6, self report of a fall, reported after 3, 6 and 12 months	TUG
Bergland, 2005	307	100	80.8 (range 75-93)	IF: 61	Severe IF	12, fall calendar, report every 3 months	Getting up from lying on the floor
Bongue, 2011	1759	51	70.7 (4.6)	SF: 563	>=1	12, monthy collected by phone	Clinical screening tool
Buatois, 2008	1958	50	70 (4)	SF: 239 MF: 183	>=2	18-36, questionnaire	FTSS
Buatois, 2010	619	50	70.1 (4.4)	MF: 55	>=2	25, questionnaire	FTSS
							TUG
							OLB
							Risk assessment
Coll-Planas, 2006	192	82.3	82.2	SF: 116	>=1	12, self-report	History of falls & independent bathing/showering
Delbaere, 2010	500	54	77.9 (4.6)	IF/MF: 166	>=1 IF, >=2 falls	12, fall diary, monthly report	PPA
							PPA followed by FES-I
Gerdhem, 2005	984	100	75 (range: 75.01-75.99)	MF: 232	>=1	12, questionnaire after 12 months	mJH-FRAT
Hnidzo, 2013	107	34.6	79.8	SF: 33, IF: 7	>=1	6, daily fall calendar, report each month	mJH-FRAT
					IFs		
Kojima, 2015	259	63.3	72.6 (5.9)	SF: 38, MF: 21	>=1	6, fall diaries, monthly report by mail	TUG
Kwan, 2012	260	30.5	74.9 (6.4)	SF: 51, MF: 35	>=1	24, montlhy telephone contact	GDS
					>=2		
Laessoe, 2007	94	74	73.7 (2.9)	SF: 14	>=1	12, fall diary, report by phone every six and twelve months	Test battery
Leclerc, 2009	868	77.2	F: 79.5 (6.6) NF: 79.0 (6.9)	MF: 99	>=2	6, calendar with montlhy report by telephone	Classification tree for risk of recurrent falling
							BBS
							TUG
Lindemann, 2008	56	57	67.7 (6.0)	SF: 30	>=1	12, daily fall calendar, every 2 months called send calendars after 12 months	Step length assessment
							AmeanVL
							AmaxVL
							AmeanVL + history of falls (12 months)
							AmaxVL + history of falls (12 months)
Moller, 2012	153	67	81.5 (6.3)	SF: 38 of whom IF:68.4%	>=1	6, question about falls in past 3 months	TUG
						12	TUG
						6	DFRI
						12	DFRI
Muhaidat, 2014	62	NF:65.3 F:69.2	NF: 75 (11.5) F: 82 (12)	SF: 13	>=1	6, fall diary, report each months by post	Classification tree
Muir, 2008	187	35	79.47 (5.83)	SF: 80 of whom MF: 33, IF: 55	>=1	12, daily registration in fall calendar monthly report by mail	BBS
					>=2		
					IFs		
Muir, 2010	117	NR	79.7 (5.3)	SF: 26, MF: 26, IF: 36	>=1	12, daily falls calendar with montlhy submission	AGS/BGS/AAOS Fall screening algorithm
					IF		
Raiche, 2000	225	NR	80.0 (4.4)	SF: 53	>=1	12, calendar to record the date of falls, monthly report by phone	14-item POMA
Russell, 2008	344	69.2	75.9 (8.5)	SF: 64, MF: 100	>=1	12, falls diary, reported every 2 months by phone	FROP-com
Russell, 2009	344	69.2	75.9		>=1 >=2	12, fall diaries	FROP-Com
	263						
Samah, 2018	305	43.9	67.67 (5.5)	SF: 40	>=1	6, falls diary, monthly follow-up by phone call	TUG
Stalenhoef, 2002	287	19.5	Female 78.5 (5.2) Male 77.2 (4.9)	SF: 49, MF:46	>=2	9, 6-weekly report by telephone	Risk Model for recurrent falls
Tiedemann, 2008	362	NR	80.4 (4.5)	SF: 99. MF: 80	>= 2	12, montlhy fall calendars	Alternate step test
							FTSS
							Timed gait (6 m)
							Stair ascent
Tiedemann, 2010	362	65	80.25 (4.5)	MF: 80	>=2	12, monthly fall calendars	Performance-based FRAT
Tromp, 2001	1285	51	75.2 (6.5)	MF: 146	>=2	36, weekly report of falls on fall calendar, mail calendar every 3 months	Fall-risk screening test
Trueblood, 2001	180	79.4	77.9 (7.26)	SF: 30, of whom IF: 16	>=1	6, report via phone at 4 and 6 months	POMA
							TUG
							Reaction time etc.*
Verghese, 2002	59	57.6	79.6 (6.3)	SF: 13	>=1	12, phone interview at 6 and 12 months	POMA, balance
							Timed Gait
							WWT-simple
							WWT-complex
Wrisley, 2010	35	51.4	72.9 (7.8)	SF: 6	>=1	6, postage-paid fall calendar postcards report monthly	DGI
							FGA
							TUG
Zur, 2016	76	79	83 (5)	SF: 8, MF: 5	>=1	18, from medical records	BBS
							Zur balance scale

Abbreviations: AGS/BGS/AAOS: American Geriatrics Society/British Geriatrics Society/American College of Orthopaedic Surgeons, AmaxVL: adjusted maximum valid step length, AmeanVL: adjusted mean valid step length, BBS: Berg Balance Scale, DFRI: Downton Fall Risk Index, F: faller, FES-I: fall efficacy scale international, FGA: functional gait assessment, FROP-Com: fall risk of older people in the community, FTSS: Five-times sit-to-stand, GDS: geriatric depression scale, IF: injurious falls/fallers, mJH-FRAT MF: multi-falls/fallers, NF: Non-faller, OLB: one-leg balance test, POMA: Tinetti Performance Oriented Mobility Assessment, PPA: physical performance assessment, SF: single faller, TUG: Timed Up and Go test, WWT: walking while talking.
*Age, fal history, reaction time, movement velocity, standing on a firm plate with eyes open

Table IV: Performance of clinical assessments.

Author, year	Fall criteria	Tool	Cut off	TP	FP	FN	TN	Sn (95% CI)	Sp (95% CI)	LR+	LR-	+PoTP	-PoTP
Alexandre, 2012	>=1	TUG	>=12,47 s	14	14	5	27	0.74 (0.49-0.91)	0.66 (0.49-0.80)	2,16	0,40	0,48	0,15
Bergland, 2005	Severe IF	Getting up from lying on the floor	No	20	11	41	83	0.33 (0.21-0.46)	0.88 (0.80-0.94)	2,80	0,76	0,55	0,25
Bongue, 2011	>=1	Clinical screening tool	>7/12	108	42	455	1154	0.19 (0.16-0.23)	0.97 (0.95-0.97)	5,49	0,84	0,70	0,26
Buatois, 2008	>=2	FTSS	>15 s	101	621	82	1154	0.55 (0.48-0.63)	0.65 (0.63-0.67)	1,57	0,69	0,40	0,23
Buatois, 2010	>=2	FTSS	>15 s	26	195	29	369	0.47 (0.34-0.61)	0.65 (0.61-0.69)	1,37	0,81	0,37	0,26
		TUG	>12 s	6	36	49	528	0.11 (0.04-0.22)	0.94 (0.91-0.95)	1,71	0,95	0,42	0,29
		OLB	<5 s	5	38	50	526	0.09 (0.03-0.20)	0.93 (0.91-0.95)	1,35	0,97	0,37	0,29
		Risk assessment	>=5	41	226	14	338	0.75 (0.61-0.85)	0.60 (0.56-0.64)	1,86	0,42	0,44	0,15
Coll-Planas, 2006	IF	History of falls & independent bathing/showering	>=1 (12 months), no	99	44	17	32	0.85 (0.78-0.91)	0.42 (0.31-0.54)	1,47	0,35	0,39	0,13
Delbaere, 2010	>=1 IF or >=2 falls	PPA	>=0.6	116	186	50	148	0.70 (0.62-0.77)	0.44 (0.39-0.50)	1,25	0,68	0,35	0,23
		PPA followed by FES-I	PPA <=0.6, FES>=23 or PPA >=0.6 and FES>=20	103	153	63	181	0.62 (0.54-0.69)	0.54 (0.49-0.60)	1,35	0,70	0,37	0,23
Gerdhem, 2005	>=1	mJH-FRAT	NR	162	158	70	594	0.70 (0.63-0.76)	0.79 (0.76-0.82)	3,33	0,38	0,59	0,14
Hnidzo, 2013	>=1	mJH-FRAT	6	4	65	36	2	0.10 (0.08-0.25)	0.03 (0.04-0.10)	0,10	30.00	0,04	0,93
			12	37	41	3	26	0.93 (0.80-0.98)	0.39 (0.27-0.52)	1,52	0,18	0,39	0,07
			13	34	34	6	33	0.86 (0.70-0.94)	0.49 (0.37-0.62)	1,70	0,28	0,42	0,11
			14	29	32	11	35	0.72 (0.56-0.85)	0.52 (0.40-0.65)	1,51	0,53	0,39	0,18
	IF		14	1	53	6	47	0.10 (0.00-0.58)	0.47 (0.37-0.57)	0,19	1,90	0,08	0,45
			17	1	34	6	66	0.10 (0.00-0.58)	0.66 (0.56-0.75)	0,29	1,37	0,11	0,37
Kojima, 2015	>=1	TUG	>=12.6 s	18	21	41	179	0.31 (0.19-0.44)	0.90 (0.84-0.93)	2,91	0,78	0,55	0,25
Kwan, 2012	>=1	GDS	>=6	28	28	58	146	0.33 (0.23-0.44)	0.84 (0.78-0.89)	2,00	0,80	0,46	0,26
	>=2			14	42	21	183	0.40 (0.24-0.58)	0.81 (0.76-0.86)	2,14	0,74	0,48	0,24
Laessoe, 2007	>=1	Test battery	6.9	7	46	7	34	0.50 (0.23-0.77)	0.43 (0.32-0.54)	0,88	1,16	0,27	0,33
Leclerc, 2009	>=2	Classification tree for risk of recurrent falling	See Appendix-B	36	63	63	706	0.36 (0.27-0.47)	0.92 (0.90-0.94)	4,44	0,69	0,66	0,23
		BBS	<=30	19	69	80	700	0.19 (0.12-0.28)	0.91 (0.89-0.93)	2,11	0,89	0,48	0,28
		TUG	>=30 s	25	138	74	631	0.25 (0.17-0.35)	0.82 (0.79-0.85)	1,39	0,91	0,37	0,28
Lindemann, 2008	>=1	Step length assessment											
		AmeanVSL	<64% of body height	23	10	7	16	0.77 (0.58-0.90)	0.62 (0.41-0.80)	2,00	0,40	0,46	0,15
		AmaxVSL	<66% of body height	21	8	9	18	0.70 (0.51-0.85)	0.69 (0.48-0.86)	2,30	0,40	0,50	0,15
		AmeanVSL + history of falls (12 months)	64% of body height, >=1	28	12	2	14	0.93 (0.78-0.99)	0.54 (0.33-0.73)	2,00	0,10	0,46	0,04
		AmaxVSL + history of falls (12 months)	66% of body height, >=1	27	11	3	15	0.90 (0.73-0.98)	0.58 (0.37-0.77)	2,10	0,20	0,47	0,08
Möller, 2012	>=1	TUG	>= 12 s	28	79	10	36	0.73 (0.57-0.87)	0.32 (0.23-0.41)	1,07	0,85	0,31	0,27
		TUG	>= 12 s	14	27	4	16	0.78 (0.52-0.94)	0.37 (0.23-0.53)	1,24	0,60	0,35	0,20
		DFRI	>= 3	24	66	7	19	0.77 (0.59-0.9)	0.22 (0.14-0.33)	1,01	0,00	0,30	0,00
		DFRI	>= 3	23	57	6	18	0.79 (0.60-0.92)	0.24 (0.15-0.35)	0,86	0,03	0,27	0,01
Muhaidat, 2014	>=1	Classification tree	NR	9	13	4	36	0.69 (0.39-0.91)	0.73 (0.59-0.85)	2,61	0,42	0,53	0,15
Muir, 2008	>=1	BBS	>=50	14	43	19	111	0.42 (0.25-0.61)	0.72 (0.64-0.79)	1,52	0,80	0,39	0,26
			<=45	20	14	60	93	0.25 (0.16-0.36)	0.87 (0.79-0.93)	1,92	0,86	0,45	0,27
			<=54	49	73	31	34	0.61 (0.50-0.72)	0.53 (0.23-0.41)	1,30	0,74	0,36	0,24
	>=2		<=45	14	43	19	111	0.42 (0.25-0.61)	0.87 (0.64-0.79)	3,23	0,67	0,58	0,22
			<=53	22	88	11	66	0.69 (0.48-0.82)	0.57 (0.35-0.51)	1,60	0,54	0,41	0,19
	IF		<=45	16	41	39	91	0.29 (0.18-0.43)	0.86 (0.60-0.77)	2,07	0,83	0,47	0,26
			<=54	30	92	25	40	0.62 (0.41-0.68)	0.51 (0.23-0.39)	1,27	0,75	0,35	0,24
Muir, 2010	>=1	AGS/BGS/AAOS Fall screening algorithm	NR	13	16	13	75	0.50 (0.30-0.70)	0.82 (0.73-0.90)	2,71	0,61	0,54	0,21
	IF			20	18	16	63	0.56 (0.38-0.72)	0.78 (0.67-0.86)	2,50	0,57	0,52	0,20
Raiche, 2000	>=1	14-item POMA	<= 36	37	83	16	89	0.70 (0.56-0.82)	0.52 (0.44-0.59)	1,46	0,58	0,38	0,20
Russell, 2008	>=1	FROP-com	>=20	108	27	8	49	0.66 (0.87-0.97)	0.64 (0.53-0.75)	1,85	0,53	0,44	0,18
			>=18	117	79	47	101	0.71 (0.64-0.78)	0.56 (0.49-0.63)	1,62	0,51	0,41	0,18
Russell, 2009	>=1	FROP-Com	>=4	110	60	54	120	0.67 (0.59-0.74)	0.67 (0.59-0.74)	2,02	0,49	0,46	0,17
	>=2		>=4	70	100	30	80	0.70 (0.60-0.79)	0.44 (0.37-0.52)	1,26	0,68	0,35	0,22
		FROP-Com, NFPP	>=4	80	96	42	45	0.66 (0.56-0.74)	0.32 (0.24-0.40)	0,96	1,08	0,29	0,32
Samah, 2018	>=1	TUG	>=8	68	72	13	150	0.84 (0.74-0.91)	0.32 (0.61-0.74)	1,24	0,49	0,35	0,17
Stalenhoef, 2002	>=2	Risk Model for recurrent falls	>=0.3	27	31	19	210	0.59 (0.43-0.73)	0.87 (0.82-0.91)	4,54	0,47	0,66	0,17
Tiedemann, 2008	>= 2	Alternate step test	>=10 s	55	124	25	158	0.69 (0.57-0.79)	0.56 (0.50-0.62)	1,90	0,50	0,45	0,18
		FTSS	>=12 s	48	127	25	155	0.60 (0.54-0.76)	0.55 (0.49-0.61)	1,33	0,73	0,36	0,24
		Timed gait (6 m)	>=6 s	40	90	25	192	0.50 (0.49-0.73)	0.68 (0.62-0.73)	1,56	0,74	0,40	0,24
		Stair ascent	>=5 s	43	118	37	164	0.54 (0.42-0.65)	0.58 (0.52-0.64)	1,29	0,79	0,36	0,25
Tiedemann, 2010	>=2	Performance-based FRAT	0-1	3	49	77	233	0.04 (0.08-0.11)	0.83 (0.78-0.87)	0,22	1,16	0,08	0,33
			2-3	16	121	64	161	0.20 (0.12-0.30)	0.57 (0.51-0.63)	0,47	1,40	0,17	0,38
			4-5	32	85	48	197	0.40 (0.29-0.52)	0.70 (0.64-0.75)	1,33	0,86	0,36	0,27
			>=6	29	27	51	255	0.36 (0.26-0.48)	0.90 (0.86-0.94)	3,79	0,71	0,62	0,23
Tromp, 2001	>=2	Fall-risk screening test	>=7	79	239	67	900	0.54 (0.46-0.62)	0.79 (0.77-0.81)	2,57	0,58	0,52	0,20
Trueblood, 2001	>=1	POMA	8	6	14	24	137	0.21 (0.07-0.39)	0.91 (0.85-0.95)	2,33	0,87	0,50	0,27
		TUG	12	3	8	27	143	0.10 (0.02-0.27)	0.95 (0.90-0.98)	2,00	0,95	0,46	0,29
		Reaction time etc.*	NR	20	46	10	114	0.67 (0.47-0.83)	0.71 (0.64-0.78)	2,32	0,47	0,50	0,17
Verghese, 2002	>=1	POMA, balance	<=10	8	14	5	32	0.65 (0.32-0.86)	0.70 (0.54-0.82)	2,13	0,50	0,48	0,18
		Timed Gait	>= 12 s	5	7	8	39	0.38 (0.14-0.68)	0.85 (0.71-0.94)	2,51	0,73	0,52	0,24
		WWT-simple	>= 20 s	6	5	7	41	0.46 (0.19-0.75)	0.89 (0.76-0.96)	4,35	0,60	0,65	0,21
		WWT-complex	>= 26 s	8	2	8	44	0.39 (0.25-0.75)	0.96 (0.85-0.99)	8,75	0,64	0,79	0,22
Wrisley, 2010	>=1	DGI	<=20/24	6	7	0	22	0.10 (0.54-1.00)	0.76 (0.56-0.90)	4,14	0,00	0,64	0,00
		FGA	<=20/30	6	5	0	24	0.10 (0.54-1.00)	0.83 (0.64-0.94)	5,80	0,00	0,71	0,00
		TUG	>12.3 s	5	1	1	28	0.83 (0.36-1.00)	0.97 (0.82-1.00)	24.17	0,17	0,91	0,07
Zur, 2016	>=1	BBS	NR	5	6	3	62	0.66 (0.24-0.91)	0.91 (0.82-0.67)	7,33	0,37	0,76	0,14
		Zur balance scale	NR	5	6	3	62	0.60 (0.24-0.91)	0.91 (0.82-0.67)	6,67	0,44	0,74	0,16

AGS/BGS/AAOS: American Geriatrics Society/British Geriatrics Society/American College of Orthopaedic Surgeons, AmaxVL: adjusted maximum valid step length, AmeanVL: adjusted mean valid step length, BBS: Berg Balance Scale, DFRI: Downton Fall Risk Index, F: faller, FES-I: fall efficacy scale international, FGA: functional gait assessment, FROP-Com: fall risk of older people in the community, FTSS: Five-times sit-to-stand, GDS: geriatric depression scale, IF: injurious falls/fallers, mJH-FRAT, MF: multi-falls/fallers, NF: Non-faller, NFPP: no fall prevention program, OLB: one-leg balance test, POMA: Tinetti Performance Oriented Mobility Assessment, PPA: physical performance assessment, SF: single faller, TUG: Timed Up and Go test, WWT: walking while talking.

*Age, fal history, reaction time, movement velocity, standing on a firm plate with eyes open

Table V: Overview study characteristics sensor-based assessments

Author	Type of sensor	Sensor location	Total (n)	Fallers	Female (%)	Mean age (SD)	Fall criteria	Follow-up time (months)
Bizovska, 2018	3 3D accelerometers	Trunk (near L5) Both shanks (15 cm above malleolus)	131	SF: 35, MF: 15	NR	NF: 70.5 (6.4) MF: 71.2 (5.3)	>=2	12, every 14 days called to report
Doi, 2013	2 3D accelerometers	Upper trunk (C7) Lower trunk (L3)	73	SF: 16	78.1	80.3	>=1	12, self reporting weekly collection
Drover, 2017	3 accelerometers	Lower back Left and right lateral shank	71 578	SF: 28		74.15 (7,0)	>=1	6, fall occurrence survey
Howcroft, 2017	1 accelerometer, 1 pressure sensor	Acc: posterior head posterior pelvis lateral shank just above the ankle. Pressure insole: plantar H-RS H-P-LS	19	SF: 7	58,7	75.2 (6.6)	>=1	6, fall calendar report monthly
Howcroft, 2018	1 accelerometer, 1 pressure sensor	Acc: posterior head, posterior pelvis Lateral shank just above the ankle Pressure insole: plantar	19	SF: 7	58.7	75.2 (6.6)	>=1	6, fall calendar, report monthly
Ihlen, 2018	1 3D accelerometer	Lower back	303	SF: 58, MF:46	SF: 51, MF: 48.8	SF:76 (6.8) MF: 75.9 (6.7)	>=1	12, montlhy phone calls
							>=2	6
Weiss, 2013	1 3D accelerometer, 1 3D gyroscope	Lower back	71	MF: 12	65	78.36 (4.71)	>=2	6

6MWT:6-minute walk test, Accel: accelerometer, ADL:activities of daily living, DGI: dynamic gait index, DT:dual-task, INF: infinity, MF: multi-faller, PGME:phase-dependent generalized multiscale entropy, SF: single faller, ST: single-task, stLE ML:short term Lyapunox exponents, TUG: Timed-Up and Go.

Table VI: Results of sensor-based assessments.

Author	Sensor worn assessment	Analyzed	Fall criteria	TP	FP	FN	TN	Sn (95% CI)	Sp (95% CI)	+LR	-LR	+PoTP	-PoTP
Bizovska, 2018	25 m walking	POMA balance score	>=2	7	9	8	72	0.47 (0.21-0.73)	0.89 (0.80-0.95)	4,27	0,60	0,65	0,20
		POMA total score		10	14	5	67	0.67 (0.38-0.88)	0.83 (0.73-0.90)	3,94	0,40	0,63	0,15
		Trunk stLE ML		8	12	7	69	0.53 (0.27-0.79)	0.85 (0.76-0.92)	3,53	0,55	0,60	0,19
		POMA balance score, trunk stLE ML		11	21	4	60	0.57 (0.45-0.92)	0.72 (0.63-0.83)	2,04	0,60	0,47	0,20
		POMA total score, trunk stLE ML		13	23	2	58	0.87 (0.60-0.98)	0.72 (0.61-0.82)	3,11	0,18	0,57	0,07
		POMA balance score & POMA total score & Trunk stLE ML		12	23	3	58	0.8 (0.52-0.96)	0.72 (0.61-0.81)	2,86	0,28	0,55	0,11
Doi, 2013	15 m walking	10 m	>=1	11	9	5	48	0.69 (0.41-0.89)	0.84 (72.13-92.52)	4,35	0,37	0,65	0,14
Drover, 2017	6MWT	Straight and turn-walking	>=1	16	8	12	35	0.58 (0.37-0.76)	0.81 (0.67-0.92)	3,04	0,52	0,57	0,18
		Straight		17	23	11	20	0.62 (0.41-0.79)	0.46 (0.31-0.62)	1,15	0,83	0,33	0,26
		Turn		17	8	11	35	0.61 (0.41-0.79)	0.82 (0.67-0.92)	3,36	0,48	0,59	0,17
Howcroft, 2017	7.62 m walking (ST, DT) & 6MWT (ST)	ST	>=1	3	5	4	7	0.43 (0.1-0.82)	0.62 (0.28-0.85)	1,12	0,92	0,33	0,28
		DT-walking		3	4	4	8	0.43 (0.1-0.82)	0.65 (0.35-0.9)	1,23	0,88	0,34	0,27
Howcroft, 2018	ST	ST-walking	>=1	4	4	3	8	0.86 (0.67-0.96)	1.00 (0.92-1.00)	INF	0,14	1,00	0,06
Ihlen, 2018	1-week ADL	PGME	>=1	41	40	17	159	0.71 (0.70-0.72)	0.80 (0.79-0.81)	3,55	0,36	0,60	0,13
		Conventional gait & demographic variables		35	48	23	151	0.61 (0.60-0.62)	0.76 (0.75-0.77)	2,54	0,51	0,52	0,18
		Fall history		27	66	31	133	0.47 (0.46-0.48)	0.67 (0.65-0.68)	1,42	0,79	0,38	0,25
		All combined		44	40	15	159	0.75 (0.73-0.76)	0.80 (0.79-0.81)	3,75	0,31	0,62	0,12
		PGME	>=2	31	20	27	179	0.67 (0.66-0.69)	0.69 (0.68-0.70)	6,70	0,37	0,74	0,14
		Conventional gait & demographic variables		36	48	22	151	0.79 (0.77-0.80)	0.76 (0.75-0.77)	3,29	0,28	0,59	0,11
		Fall history		25	70	33	129	0.55 (0.53-0.56)	0.65 (0.64-0.67)	1,57	0,69	0,40	0,23
		All combined		38	40	20	159	0.83 (0.82-0.84)	0.8 (0.82-0.84)	4,15	0,21	0,64	0,08
Weiss, 2013	3-days ADL	Dynamic gait index (without sensors)	>=2	4	1	8	58	0.33 (0.09-0.65)	0.98 (0.91-1.00)	15,14	0,68	0,87	0,23
		DGI (without sensors) & 3-days ADL		9	0	3	59	0.75 (0.43-0.95)	1.00 (0.94-0.10)	INF	0,25	1,00	0,10

6MWT:6-minute walk test, ADL:activities of daily living, DGI: dynamic gait index, DT:dual-task, INF: infinity, MF: multi-faller, PGME:phase-dependent generalized multiscale entropy, SF: single faller, ST: single-task, stLE ML:short term Lyapunox exponents, TUG: Timed-Up and Go.

Popular physical fall assessment tools such as TUG, BBS, and FTSS have been evaluated by multiple studies, but do not show high predictive value for prospective falls. All studies using the TUG, except for one, showed a +PoTP of 55% or lower, indicating a poor fall prediction capability. The one study with a higher performance (+PoTP of 91%) only included 6 fallers and is therefore not representative [39]. Although the BBS showed promising outcomes in a small sampled study (n=76)[53], in two larger studies (n=187 n=868), the BBS showed a +PoTP between 36 and 48% and a -PoTP between 24 and 28% for varying cut-off scores [34][47], while the FTSS scored even worse. These results are in contrast with the review of Lusardi et al. (2017), who concluded that the BBS, TUG, and FTSS are currently the most evidence-supported functional measures to determine individual fall risk. However, this review included retrospective studies[30], for which much better associations between TUG, BBS and FTSS and falls have been reported, as demonstrated by Beauchet et al. (2011)[27]. Based on the prospective studies it must be concluded that the predictive value of TUG, BBS, and FTSS is poor, and should not be used as primary tool to assess fall risk. Furthermore, this discrepancy between prospective and retrospective studies indicates the importance of using prospective study designs to evaluate fall risk assessment tools.

Besides these common tools, *risk model for recurrent falls*, *performance-based FRAT*, and *Getting up from lying on the floor* are physical assessments that showed promising predictive values

(+PoTP> 55%) in moderate sample sizes (N between 145 - 303). However, for the *performance-based FRAT* the predictive value depends largely on the cut-off score used [55]. In the future,research should aim to specify the best cut-off scores, and their validity and reliability in different settings (e.g. environment, age of population) before using them to predict fall risk in daily practice.

Sensor-based assessments demonstrate promising predictive value, although results depend on the task complexity, location of the sensor, and feature extraction and have only been evaluated in small samples ($N < 131$) with consequently a low number of fallers [25]. Straight walking does not yield enough discrimative information to predict fallers, as with sensor data from the +PoTP was below 40%. More complex tasks, like including a turn in walking, or dual-task walking, showed better predictability, which is in line with conclusions of Bayot et al. (2020) [65]. Furthermore, adding complexity to the task seems better than adding sensors, as indicated by the finding that adding sensors to a double task (talking) while straight walking does not improve the predictive value beyond adding turns to the double task [63][52]. Focusing on ADL, which generally are complex tasks, rather than sensor-based clinical tests therefore seems promising. The sensor-based ADL assessments show promising results (+PoTP $>$ 74%), but require further large-scale prospective studies.

Limitations & future work

The strength of this review is that only studies with a prospective design were included to determine the predictive value. As shown for the TUG, BBS and FTSS including retrospective studies may lead to an overestimation of the predictive value of the different tools. A disadvantage of our scoping review is that only studies mentioned in one of the 28 high-quality and recently published reviews were included. Additionally, the method to detect falls differed from a calendar to record falls [49] or contact by phone every 12 days to few months [22]. Differences in accuracy of the fall detection method might affected the outcomes, and limit the direct comparison between studies [30].As our review suggest that sensor-based fall-risk assessment potentially outperform traditional assessments, large, high quality prospective research in this field is needed. As no studies after 2020 were included in our selected reviews, we searched from that date on and found that since then only two sensor-based prospective fall prediction studies have been published, both with a relatively small sample size ($n < 74$) Future research should aim to determine the best location of the sensors, task to perform with the sensors and whether sensors worn in ADL can predict falls. This could not only make fall prediction better, but potentially less time-consuming as no physical test or questionnaires need to be performed.

V. Conclusion

Traditional questionnaires and physical test to assess fall risk, have limited predictive value when evaluated in prospective study designs. Use of sensors during assessment or ADL might improve fall risk prediction, although this warrants large prospective studies to determine the validity of the sensor location and task complexity.

References

[1] World Health Organization, "Falls," 04 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls

[2] VeiligheidNL, "Feiten cijfers — Valpreventie — VeiligheidNL," 09 2020. [Online]. Available: https://www.veiligheid.nl/valpreventie/feiten-cijfers

[3] C. Stam and B. Blatter, "Letsels 2020: Kerncijfers LIS," Tech. Rep. 2, 11 2021.

[4] B. J. Vellas, S. J. Wayne, L. J. Romero, R. N. Baumgartner, and P. J. Garry, "Fear of falling and restriction of mobility in elderly fallers," *Age and ageing*, vol. 26, no. 3, pp. 189–193, 1997.

[5] J. Howland, E. W. Peterson, W. C. Levin, L. Fried, D. Pordon, and S. Bak, "Fear of falling among the community-dwelling elderly," *Journal of aging and health*, vol. 5, no. 2, pp. 229–243, 1993.

[6] M. E. Tinetti and C. S. Williams, "The effect of falls and fall injuries on functioning in community-dwelling older persons," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 53, no. 2, pp. M112–M119, 1998.

[7] K. A. Faulkner, J. A. Cauley, S. A. Studenski, D. P. Landsittel, S. R. Cummings, K. E. Ensrud, M. Donaldson, and M. Nevitt, "Lifestyle predicts falls independent of physical risk factors," *Osteoporosis international*, vol. 20, no. 12, pp. 2025–2034, 2009.

[8] G. F. Fuller, "Falls in the elderly," *American family physician*, vol. 61, no. 7, p. 2159, 2000.

[9] P. Pasquetti, L. Apicella, and G. Mangone, "Pathogenesis and treatment of falls in elderly," *Clinical cases in mineral and bone metabolism*, vol. 11, no. 3, p. 222, 2014.

[10] C. Boelens, E. E. Hekman, and G. J. Verkerke, "Risk factors for falls of older citizens," *Technology and Health care*, vol. 21, no. 5, pp. 521–533, 2013.

[11] A. Bueno-Cavanillas, F. Padilla-Ruiz, J. J. Jiménez-Moléon, C. Peinado-Alonso, and R. Gálvez-Vargas, "Risk factors in falls among the elderly according to extrinsic and intrinsic precipitating causes," *European journal of epidemiology*, vol. 16, no. 9, pp. 849–859, 2000.

[12] J. H. Van Dieën and M. Pijnappels, "Balance control in older adults," in *Locomotion and Posture in Older Adults*. Springer, 2017, pp. 237–262.

[13] L. Clemson, R. G. Cumming, H. Kendig, M. Swann, R. Heard, and K. Taylor, "The effectiveness of a community-based program for reducing the incidence of falls in the elderly: A randomized trial," *Journal of the American Geriatrics Society*, vol. 52, no. 9, pp. 1487–1494, 2004.

[14] L. Yardley, N. Beyer, K. Hauer, G. Kempen, C. Piot-Ziegler, and C. Todd, "Development and initial validation of the falls efficacy scale-international (fes-i)," *Age and ageing*, vol. 34, no. 6, pp. 614–619, 2005.

[15] L. E. Powell and A. M. Myers, "The activities-specific balance confidence (abc) scale," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 50, no. 1, pp. M28–M34, 1995.

[16] M. E. Tinetti, T. F. Williams, and R. Mayewski, "Fall risk index for elderly patients based on number of chronic disabilities," *The American journal of medicine*, vol. 80, no. 3, pp. 429–434, 1986.

[17] D. Podsiadlo and S. Richardson, "The timed "up & go": a test of basic functional mobility for frail elderly persons," *Journal of the American geriatrics Society*, vol. 39, no. 2, pp. 142–148, 1991.

[18] M. Marschollek, A. Rehwald, K. Wolf, M. Gietzelt, G. Nemitz, H. M. Zu Schwabedissen, and R. Haux, "Sensor-based fall risk assessment–an expert 'to go'," *Methods of information in medicine*, vol. 50, no. 05, pp. 420–426, 2011.

[19] G. Zhao, L. Chen, and H. Ning, "Sensor-based fall risk assessment: A survey," in *Healthcare*, vol. 9, no. 11. Multidisciplinary Digital Publishing Institute, 2021, p. 1448.

[20] L. Bizovska, Z. Svoboda, M. Janura, M. C. Bisi, and N. Vuillerme, "Local dynamic stability during gait for predicting falls in elderly people: A one-year prospective study," *PloS one*, vol. 13, no. 5, p. e0197091, 2018.

[21] B. R. Greene, D. McGrath, and B. Caulfield, "A comparison of cross-sectional and prospective algorithms for falls risk assessment," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE,

2014, pp. 4527–4530.

[22] E. A. Ihlen, K. S. Van Schooten, S. M. Bruijn, J. H. Van Dieen, B. Vereijken, J. L. Helbostad, and M. Pijnappels, "Improved prediction of falls in community-dwelling older adults through phase-dependent entropy of daily-life walking," *Frontiers in aging neuroscience*, vol. 10, p. 44, 2018.

[23] A. Weiss, M. Brozgol, M. Dorfman, T. Herman, S. Shema, N. Giladi, and J. M. Hausdorff, "Does the evaluation of gait quality during daily life provide insight into fall risk? a novel approach using 3-day accelerometer recordings," *Neurorehabilitation and neural repair*, vol. 27, no. 8, pp. 742–752, 2013.

[24] A. Kristoffersson, J. Du, and M. Ehn, "Performance and characteristics of wearable sensor systems discriminating and classifying older adults according to fall risk: a systematic review," *Sensors*, vol. 21, no. 17, p. 5863, 2021.

[25] J. Bezold, J. Krell-Roesch, T. Eckert, D. Jekauc, and A. Woll, "Sensor-based fall risk assessment in older adults with or without cognitive impairment: a systematic review," *European review of aging and physical activity*, vol. 18, no. 1, pp. 1–14, 2021.

[26] V. Strini, R. Schiavolin, and A. Prendin, "Fall risk assessment scales: A systematic literature review," *Nursing Reports*, vol. 11, no. 2, pp. 430–443, 2021.

[27] O. Beauchet, B. Fantino, G. Allali, S. Muir, M. Montero-Odasso, and C. Annweiler, "Timed up and go test and risk of falls in older adults: a systematic review," *The journal of nutrition, health & aging*, vol. 15, no. 10, pp. 933–938, 2011.

[28] S. R. Cummings, M. C. Nevitt, and S. Kidd, "Forgetting falls: the limited accuracy of recall of falls in the elderly," *Journal of the American Geriatrics Society*, vol. 36, no. 7, pp. 613–616, 1988.

[29] D. A. Ganz, T. Higashi, and L. Z. Rubenstein, "Monitoring falls in cohort studies of community-dwelling older people: effect of the recall interval," *Journal of the American Geriatrics Society*, vol. 53, no. 12, pp. 2190–2194, 2005.

[30] M. M. Lusardi, S. Fritz, A. Middleton, L. Allison, M. Wingood, E. Phillips, M. Criss, S. Verma, J. Osborne, and K. K. Chui, "Determining risk of falls in community dwelling older adults: a systematic review and meta-analysis using posttest probability," *Journal of geriatric physical therapy (2001)*, vol. 40, no. 1, p. 1, 2017.

[31] T. S. Alexandre, D. M. Meira, N. C. Rico, and S. K. Mizuta, "Accuracy of timed up and go test for screening risk of falls among community-dwelling elderly," *Brazilian Journal of Physical Therapy*, vol. 16, pp. 381–388, 2012.

[32] S. Hnizdo, R. A. Archuleta, B. Taylor, and S. C. Kim, "Validity and reliability of the modified john hopkins fall risk assessment tool for elderly patients in home health care," *Geriatric nursing*, vol. 34, no. 5, pp. 423–427, 2013.

[33] G. Kojima, T. Masud, D. Kendrick, R. Morris, S. Gawler, J. Treml, and S. Iliffe, "Does the timed up and go test predict future falls among british community-dwelling older people? prospective cohort study nested within a randomised controlled trial," *BMC geriatrics*, vol. 15, no. 1, pp. 1–7, 2015.

[34] B. S. Leclerc, C. Bégin, É. Cadieux, L. Goulet, J.-F. Allaire, J. Meloche, N. Leduc, and M.-J. Kergoat, "A classification and regression tree for predicting recurrent falling among community-dwelling seniors using home-care services," *Canadian journal of public health*, vol. 100, no. 4, pp. 263–267, 2009.

[35] U. Olsson Möller, J. Kristensson, P. Midlöv, C. Ekdahl, and U. Jakobsson, "Predictive validity and cut-off scores in four diagnostic tests for falls–a study in frail older people at home," *Physical & Occupational Therapy in Geriatrics*, vol. 30, no. 3, pp. 189–201, 2012.

[36] J. Muhaidat, A. Kerr, J. J. Evans, M. Pilling, and D. A. Skelton, "Validity of simple gait-related dual-task tests in predicting falls in community-dwelling older adults," *Archives of physical medicine and rehabilitation*, vol. 95, no. 1, pp. 58–64, 2014.

[37] Z. A. Samah, D. K. A. Singh, R. R. Murukesu, S. Shahar, N. Nordin, M. A. Omar, and A. Chin, "Discriminative and predictive ability of physical performance measures in identifying fall risk among older adults," *Sains Malaysiana*, vol. 47, no. 11, pp. 2769–2776, 2018.

[38] P. R. Trueblood, N. Hodson-Chennault, A. McCubbin, and D. Youngclarke, "Performance and impairment-based assessments among community dwelling elderly: sensitivity and specificity," *Journal of geriatric physical therapy*, vol. 24, no. 1, pp. 2–6, 2001.

[39] D. M. Wrisley and N. A. Kumar, "Functional gait assessment: concurrent, discriminative, and predictive validity in community-dwelling older adults," *Physical therapy*, vol. 90, no. 5, pp. 761–773, 2010.

[40] A. Tromp, S. Pluijm, J. Smit, D. Deeg, L. Bouter, and P. Lips, "Fall-risk screening test: a prospective study on predictors for falls in community-dwelling elderly," *Journal of clinical epidemiology*, vol. 54, no. 8, pp. 837–844, 2001.

[41] B. Bongue, C. Dupré, O. Beauchet, A. Rossat, B. Fantino, and A. Colvez, "A screening tool with five risk factors was developed for fall-risk prediction in community-dwelling elderly," *Journal of clinical epidemiology*, vol. 64, no. 10, pp. 1152–1160, 2011.

[42] K. Delbaere, J. C. Close, H. Brodaty, P. Sachdev, and S. R. Lord, "Determinants of disparities between perceived and physiological risk of falling among elderly people: cohort study," *Bmj*, vol. 341, 2010.

[43] P. Gerdhem, K. A. Ringsberg, K. Åkesson, and K. J. Obrant, "Clinical history and biologic age predicted falls better than objective functional tests," *Journal of clinical epidemiology*, vol. 58, no. 3, pp. 226–232, 2005.

[44] M. M.-S. Kwan, S.-I. Lin, J. C. Close, and S. R. Lord, "Depressive symptoms in addition to visual impairment, reduced strength and poor balance predict falls in older taiwanese people," *Age and ageing*, vol. 41, no. 5, pp. 606–612, 2012.

[45] U. Laessoe, H. C. Hoeck, O. Simonsen, T. Sinkjaer, and M. Voigt, "Fall risk in an active elderly population–can it be assessed?" *Journal of negative results in biomedicine*, vol. 6, no. 1, pp. 1–11, 2007.

[46] U. Lindemann, L. Lundin-Olsson, K. Hauer, M. Wengert, C. Becker, and K. Pfeiffer, "Maximum step length as a potential screening tool for falls in non-disabled older adults living in the community," *Aging clinical and experimental research*, vol. 20, no. 5, pp. 394–399, 2008.

[47] S. W. Muir, K. Berg, B. Chesworth, and M. Speechley, "Use of the berg balance scale for predicting multiple falls in community-dwelling elderly people: a prospective study," *Physical therapy*, vol. 88, no. 4, pp. 449–459, 2008.

[48] S. W. Muir, K. Berg, B. Chesworth, N. Klar, and M. Speechley, "Application of a fall screening algorithm stratified fall risk but missed preventive opportunities in community-dwelling older adults: a prospective study," *Journal of geriatric physical therapy*, vol. 33, no. 4, pp. 165–172, 2010.

[49] M. Raîche, R. Hébert, F. Prince, and H. Corriveau, "Screening older adults at risk of falling with the tinetti balance scale," *The Lancet*, vol. 356, no. 9234, pp. 1001–1002, 2000.

[50] M. A. Russell, K. D. Hill, I. Blackberry, L. M. Day, and S. C. Dharmage, "The reliability and predictive accuracy of the falls risk for older people in the community assessment (frop-com) tool," *Age and ageing*, vol. 37, no. 6, pp. 634–639, 2008.

[51] M. A. Russell, K. D. Hill, L. M. Day, I. Blackberry, L. C. Gurrin, and S. C. Dharmage, "Development of the falls risk for older people in the community (frop-com) screening tool," *Age and ageing*, vol. 38, no. 1, pp. 40–46, 2009.

[52] J. Verghese, H. Buschke, L. Viola, M. Katz, C. Hall, G. Kuslansky, and R. Lipton, "Validity of divided attention tasks in predicting falls in older individuals: a preliminary study," *Journal of the American Geriatrics Society*, vol. 50, no. 9, pp. 1572–1576, 2002.

[53] O. Zur, T. Shaki, and E. Carmeli, "Concurrent validity and reliability of a new balance scale used in older adults," in *Respiratory Medicine and Science*. Springer, 2015, pp. 63–70.

[54] L. Coll-Planas, M. Kron, S. Sander, U. Rißmann, C. Becker, and T. Nikolaus, "Accidental falls among community-dwelling older adults," *Zeitschrift für Gerontologie und Geriatrie*, vol. 39, no. 4, pp. 277–282, 2006.

[55] A. Tiedemann, S. R. Lord, and C. Sherrington, "The development and validation of a brief performance-based fall risk assessment tool for use in primary care," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 65, no. 8, pp. 896–903, 2010.

[56] S. Buatois, D. Miljkovic, P. Manckoundia, R. Gueguen, P. Miget, G. Vançon, P. Perrin, and A. Benetos, "Five times sit to stand test is a predictor of recurrent falls in healthy community-living subjects aged 65 and older," *Journal of the American Geriatrics Society*, vol. 56, no. 8, pp. 1575–1577, 2008.

[57] S. Buatois, C. Perret-Guillaume, R. Gueguen, P. Miget, G. Vançon, P. Perrin, and A. Benetos, "A simple clinical scale to stratify risk of recurrent falls in community-dwelling adults aged 65 years and older," *Physical therapy*, vol. 90, no. 4, pp. 550–560, 2010.

[58] A. Bergland and K. Laake, "Concurrent and predictive validity of "getting up from lying on the floor"," *Aging clinical and experimental research*, vol. 17, no. 3, pp. 181–185, 2005.

[59] A. Tiedemann, H. Shimada, C. Sherrington, S. Murray, and S. Lord, "The comparative ability of eight functional mobility tests for predicting falls in community-dwelling older people," *Age and ageing*, vol. 37, no. 4, pp. 430–435, 2008.

[60] P. Stalenhoef, J. Diederiks, J. Knottnerus, A. Kester, and H. Crebolder, "A risk model for the prediction of recurrent falls in community-dwelling elderly: a prospective cohort study," *Journal of clinical epidemiology*, vol. 55, no. 11, pp. 1088–1094, 2002.

[61] T. Doi, S. Hirata, R. Ono, K. Tsutsumimoto, S. Misu, and H. Ando, "The harmonic ratio of trunk acceleration predicts falling among older people: results of a 1-year prospective study," *Journal of neuroengineering and rehabilitation*, vol. 10, no. 1, pp. 1–6, 2013.

[62] D. Drover, J. Howcroft, J. Kofman, and E. D. Lemaire, "Faller classification in older adults using wearable sensors based on turn and straight-walking accelerometer-based features," *Sensors*, vol. 17, no. 6, p. 1321, 2017.

[63] J. Howcroft, J. Kofman, and E. D. Lemaire, "Prospective fall-risk prediction models for older adults based on wearable sensors," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 10, pp. 1812–1820, 2017.

[64] J. Howcroft, E. D. Lemaire, and J. Kofman, "Prospective elderly fall prediction by older-adult fall-risk modeling with feature selection," *Biomedical Signal Processing and Control*, vol. 43, pp. 320–328, 2018.

[65] M. Bayot, K. Dujardin, L. Dissaux, C. Tard, L. Defebvre, C. T. Bonnet, E. Allart, G. Allali, and A. Delval, "Can dual-task paradigms predict falls better than single task?–a systematic literature review," *Neurophysiologie Clinique*, vol. 50, no. 6, pp. 401–440, 2020.

VI. Appendix

A. Complete search strategy

Scopus
TITLE-ABS-KEY((”fall* risk assess*” OR ”fall* risk predict*” OR ”fall* risk classif*” OR ”fall* risk measur*”) AND (”older adult” OR elder* OR aged OR geriatr* OR senior*)) AND (ALL(accura* OR sensitiv* OR specific*))

PubMed
(”fall risk assess*”[Title/Abstract] OR ”fall risk predict*”[Title/Abstract] OR ”fall risk classif*”[Title/Abstract] OR ”fall risk measur*”[Title/Abstract]) AND (”older adult”[Title/Abstract] OR ”elder*”[Title/Abstract] OR ”aged”[Title/Abstract] OR ”senior*”[Title/Abstract] OR ”geriatric*”[Title/Abstract]) AND (”accura*”[All Fields] OR ”specifi*”[All Fields] OR ”sensitiv*”[All Fields])

IEEE Xplore
((”All Metadata”:”fall risk assess*” OR ”All Metadata”:”fall risk predict*” OR ”All Metadata”:”fall risk classification” OR ”All Metadata”:”fall risk measur*”) AND (”All Metadata”:”older adult” OR ”All Metadata”:elder* OR ”All Metadata”:aged OR ”All Metadata”:geriatr* OR ”All Metadata”:senior) AND (”Full Text Metadata”:accura* OR ”Full Text Metadata”:specificity OR ”Full Text Metadata”:sensitiv*))

Web of Science
(TS=(”fall risk assess*”) OR TS=(”fall risk predict*”) OR TS=(”fall risk classif*”) OR TS=(”fall risk measur*”)) AND (TS=(”older adult”) OR TS=(elder*) OR TS=(aged) OR TS=(geriatric*) OR TS=(senior*)) AND (ALL=(specific*) OR ALL=(sensitiv*) OR ALL=(accura*))

B. Explanation of different assessments

Adjusted maximum step length[46]
Ruler tape was attached to the floor in front of the participant. The participant is instructed to step out maximally with the preferred leg while maintaining the stance leg in the initial position. When returning to the initial position, the stepping foot can step back in several small steps. To find out the preferred leg, the participant can make two sub-maximal steps with each leg. After that, the test is conducted five times with the preferred leg. The arms are allowed to move freely but not touch or grab anything.
The distance between the toes of both feet is the maximum step length. The step length was adjusted for body height.

American Geriatrics Society/British Geriatrics Society/American College of Orthopaedic Surgeons (AGS/BGS/AAOS) Guidelines for Fall Prevention screening algorithm[48]
According to the AGS/BGS/AAOS Guidelines for Fall Prevention screening algorithm, people with a history of one fall should receive an assessment for balance and gait impairment. If the participant has balance/gait impairment, fall evaluation is required. If the participant has a history of multiple falls, fall evaluation is needed, and no gait and balance assessment are necessary. People without a history of falling should be reviewed yearly. The algorithm is displayed in Figure 3

Alternate step test (AST)[55]
In this test, the participant places the whole foot onto a step (18 cm high and 40 cm deep). The participant should alternate between the right and left foot for eight repetitions. The test should be completed as fast as possible, and the time to complete the task is measured. This test requires speed, strength, and balance.

Berg balance scale (BBS)[34, 47, 53]
The BBS is a balance scale that contains 14 items focussing on

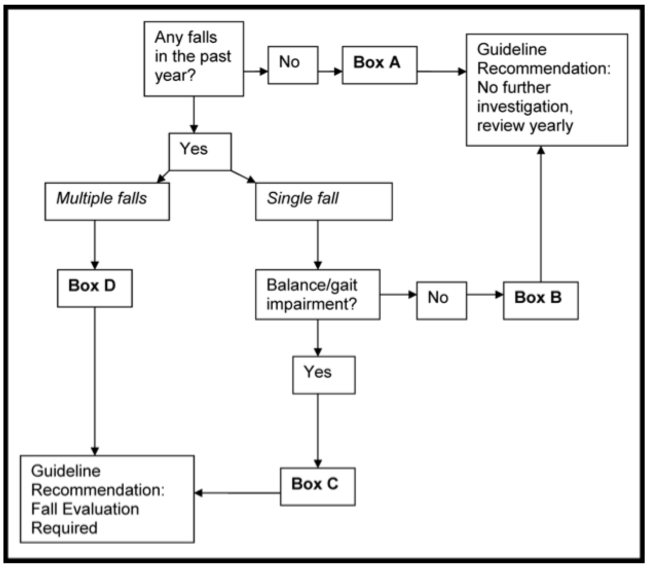


Figure 3: AGS/BGS/AAOS fall screening algorithm according to Muir et al. (2010a)[48].

sitting and standing[39]. Every item should be scored on an ordinal scale from 0-4[47]. Zero indicates an inability to perform the task, and four means that the task is performed independently. Examples of the tasks are standing from a chair and standing with feet together[39].

Classification tree[36]
The classification tree is a combination of the following nine tasks: straight walking and visuospatial clock task, walking with turns and naming animals, walking with turns and counting backward in 3s, avoiding stationary obstacles and naming animals, avoiding a moving obstacle, and carrying a cup, TUG and having a cup, stair descent and naming animals, WWT-complex, a combination of straight walking, visuospatial clock task, and holding a cup (TT test)[36]. For each test, the time is measured. In addition, the secondary task performance was recorded and assessed as total answers/second (performance speed) and errors/total answers (accuracy).

Classification tree for risk of recurrent falling[34]
The classification tree for risk of recurrent falling is based on the history of falls, BBS, type of housing, and alcohol use. The amount of falls in the past three months ($\geq$2) leads to the other parts of the flow chart: BBS score (cut-off: 30), type of housing (private residential/other types of housing), and alcohol in the past six months (yes/no). The whole flow chart is displayed in Figure 4.

Clinical screening tool[41]
The clinical screening tool consists of 6 items of categorical variables: sex, living alone (in couple, family, alone), Osteoarthritis, history of falls (0, 1, 2, 3 and more), psychoactive drug use, and balance impairments[41]. Balance impairment is measured by the change of position of the arms during the first 5 seconds when performing the OLB test.

Downton Fall Risk Index (DFRI)[35]
Eleven fall risk items: history of falls in preceding 12 months (reported by the patient), use of tranquilizers or sedatives, diuretics, antihypertensive drugs, antiparkinsonian drugs, antidepressants (determined by drug prescriptions), visual impairment (moderate to severe impairment or blindness, and needing glasses daily), hearing impairment (moderate to severe impairment or deafness), limb impairment (extremity paresis or muscle weakness), cognitive

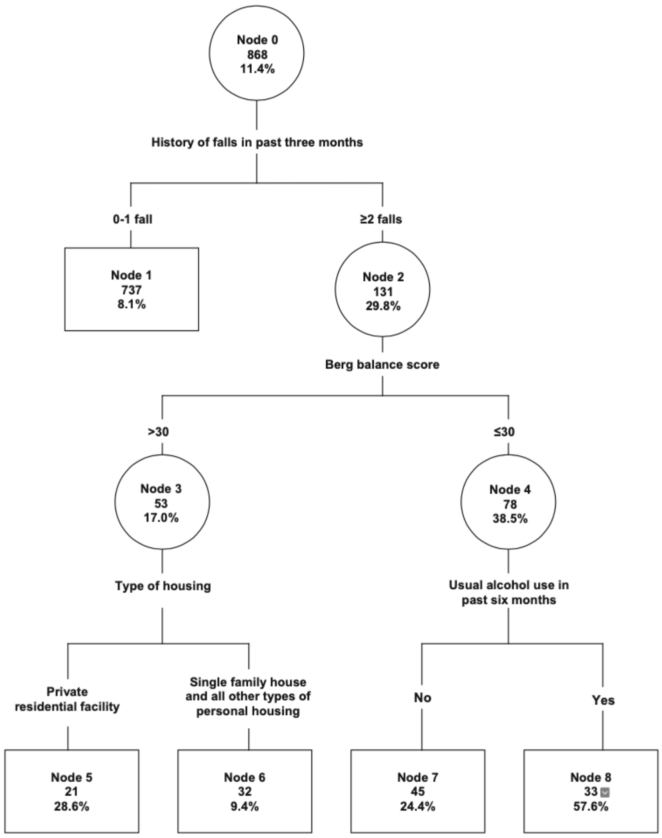


Figure 4: Classification tree for predicting the risk of recurrent falls at six months follow-up[34].In each node the following information is displayed from top to bottom: node number, number of subjects in that node, incidence of recurrent faller in that node.

impairment (MMSE<=23), and walking ability[35]. A participant can obtain a score between 0-11, where a score ≥ 3 indicates a high risk of falls.

Dynamic Gait Index (DGI)[21, 39]
For the DGI, the participant performs the following tasks: walking at normal, fast, and slow speeds, walking with horizontal and vertical head turns, walking over and around obstacles, and ascending and descending stairs[39]. All items are rated on a 4-level ordinal scale. A lower score indicates greater impairments. A score of =¡19 indicates an increased risk of falls.

Fall risk screening test[40]
The fall risk screening test for recurrent fallers consists of fall history, urinary incontinence, visual impairment, and use of benzodiazepines[40]. The score of each item is the regression coefficient multiplied by five and rounded off to the nearest integer. The sum of the score of each item determines the end score (max. = 15), which is related to the fall risk.

Fall Risk for Older People in the Community (FROP-Com)[50, 50]
The FROP-Com consists of 26 questions, covering 13 risk factors: history of falls, medication, medical condition, feet and footwear, cognitive status, continence, nutritional status, environment, functional behavior, function, balance, gait/physical activity[50]. The questions should be answered on a dichotomous or ordinal scale from 0-3. The overall score is the sum of the score of each question resulting in a maximum score of 60. A higher score indicates a greater risk of falling.

Fall Efficacy Scale International (FES-I)[42]
The participant is asked about the concern about falling during ADL[42]. A higher score indicates a greater concern about falling.

Functional Gait Assessment (FGA)[39]
The FGA is a modification of the DGI to improve the reliability of the DGI and reduce the ceiling effect seen with the DGI in patients with vestibular disorders[39]. Participants are asked to walk at regular, fast, and slow speeds, with vertical and horizontal head, turns, with eyes closed, over obstacles, in tandem, backward, and ascending and descending stairs. All items are scored on a 4-level ordinal scale. The total score ranges from 0-30, with lower scores indicating more significant impairments.

Five-times sit-to-stand (FTSS)[55, 56]
Participants are asked to rise from a chair five times as quickly as possible with their arms folded[55, 56]. The chair had a height of 43 cm in the study of Tiedemann et al. (2010) and a height of 45 cm in Buatois et al. (2008).

Geriatric depression scale (GDS)[44]
The participant is scored on questions of the 15 items GDS (GDS-15). A score of ≥ 6 is suggestive of depression[44].

Getting up from lying on the floor[58]
Participants lay down on the floor in a supine position with their hands on the floor and their head on a pillow[58]. The participant was asked to get up from lying on the floor without aid at their own time. The outcome was whether the subject managed (1) or not (0).

Modified John Hopkins Fall Risk Assessment Tool (mJH-FRAT)[35]
The JH-FRAT is initially developed for assessing fall risk in acute care. In the mJH-FRAT, seven areas are evaluated: patient age, prior fall history, elimination, medications, use of patient care equipment, mobility, and cognition[35]. The total score categorized patients into three risk groups: low risk (0-6), moderate risk (7-13), high risk (14-35).

One leg balance test (OLB)[56]
The participant is instructed to remain upright on one leg without support for at least 5 seconds. The participant failed if he/she could not keep standing.

Performance-based FRAT[55]
The tests are based on the five major physiological domains of the Physiological Profile Assessment (PPA). The domains are vision, lower limb sensation, lower limb strength, reaction time, and standing balance[55]. For vision, a low (10%) contrast visual acuity measured at a distance of 3 m was used. Lower limb sensation was tested by using a single Semmes-Weinstein-type pressure monofilament on the ankle. For lower limb strength, reaction time, and standing balance, the sit-to-stand, alternate step test, and near tandem stand test were used. In addition, the fall history (at least one fall in previous year) and use of medication (four of more, psychotropic) were asked.

Tinetti Performance Oriented Mobility Assessment (POMA)[20, 38, 49, 52]
The long version of the POMA consists of 24 items, while the frequently used short version consists of 16 items. Both POMAs consist of a balance and a gait part using walking, sitting, and standing positions. The long version consists of 14 items for balance (max score 24) and ten items for gait (max score 16)[49]. Raiche et al. (2000) and Bizovska et al. (2018) studied the balance part of the long version of the POMA[20, 49]. Trueblood et al. (2001) studied the full short version, containing balance (max score=16) and gait (max score=12) part[38]. Verghese et al. (2002) studied the nine task balance part of the short version[52].

Physiological Performance Assessment[42]
The five major physiological domains of the Physiological Profile Assessment (PPA) are visual contrast sensitivity (Melbourne edge test), proprioception (measured with a lower limb-matching task), quadriceps strength (measured isometrically in the dominant leg while the participant is seated with the hip and knee flexed at 90°, simple reaction time (measured with light as stimulus and a higher press as a response), and postural sway (measured with a sway meter recording displacements of the body at the level of the pelvis while the participants stand on a foam rubber mat with eyes open)[42].

Risk assessment[57]
The risk assessment of Buatois et al. (2010) consists of the FTSS (cut-off>15 s), TUG (cut-off >12 s), and the OLB (cut-off ¡5 s)[57]. This assessment results in a score from 0 to 16. A higher score gives a higher fall risk.

Risk model for recurrent fallers[60]
Risk model for the prediction of recurrent falls ($\geq$2) containing: gender, $\geq$2 fall in the previous year, depression (SCL90$\geq$22), hand dynamometry (man$\leq$22 kg, woman $\leq$12 kg), abnormal postural sway[60]. The regression coefficient is multiplied by five and rounded to the nearest integer. The sum of the scores can determine the risk of falling.

Six-minute walk test(6MWT)[62, 63]
During the six-minute walk test, the participant walks along a hallway making straight left and right turns around two cones spaced 100 ft (30.34 m) apart[62]. In clinical assessments, the distance that was covered is measured.

Stair ascent[59]
In the stair ascent test of Tiedemann et al. (2008), the participant had to ascent eight steps (15cm high, 27.5 cm deep)[59]. The time was measured from the moment the patient raised the foot of the floor to climb the first step until both feet were placed on the eighth step.

Test battery[45]
The test battery consists of 9 items: standing balance (0-6 performance scale), stepping ability (time required), general function (time required), reaction time (averaged time to step), general leg strength (time required), dual-task (speed reduction, gait variability (autocorrelation), gait cadence (steps per second), vision (acuity/contrast/field (0-7))[45].

Timed gait[59]
During the 6-meter walk test of Tiedemann et al. (2008), the participant was asked to walk 6 meters at their speed[59]. The timed gait test of Verghese et al. was performed over two times 6 meters, tus included a turn[52].

Timed Up and Go (TUG)[21, 31, 33–39, 56]
The TUG is a functional mobility test where the participant should rise from a chair, walk 3 meters, turn, walk back, and sit down again[33, 35]. The time a participant needs to perform the test is measured. A longer time indicates worse balance and mobility performance[33].

Walking while talking test (WWT)[52]
Participants walk 20 feet, turn and return (40 feet in total) while reciting letters of the alphabet for the simple version [52]. During the complex WWT, participants recite alternate alphabet letters (a, c, e. . .). The duration of the walking is timed.

Zur balance scale[53]
In this test, a piece of Styrofoam (30 kg/m^3) is covered tightly with a piece of fabric[53]. The participant performs the Romberg stance and Tandem stance with the eyes opened and closed and vertical and horizontal head movements. These tests are once performed on the floor and once on the Styrofoam plate. The score given per item is the time maintaining the balance and the number of head movements with a maximum of 10 s. The end score is the number of head movements multiplied by two and the total time divided by 2.