Healthcare and Rehabilitation 2 (2026) 100069

Contents lists available at ScienceDirect

Healthcare and Rehabilitation

journal homepage: www.keaipublishing.com/en/journals/hcr/

Full Length Article

Comparative study of demographic information, clinical scales and questionnaires, and mobility tests for fall risk assessment in older adults

Peng Wu[a], Jianlei Fang[a], Ziyun Ding[b], Zeyang Guan[a], Jiachen Wang[a], Yikai He[c], Yihao Zhang[a], Huanghe Zhang[a,d,*]

[a] Center for Robotics, School of Control Science and Engineering, Shandong University, Jinan 250061, China
[b] School of Engineering, University of Birmingham, Birmingham B15 2TT, United Kingdom
[c] School of Qilu Transportation, Shandong University, Jinan 250002, China
[d] International Joint Research Center for Perception and Control of Intelligent Rehabilitation Systems of Sichuan Province, Chengdu 610106, China

ARTICLE INFO

Keywords:
Fall risk assessment
Demographic information
Clinical scales and questionnaires
Mobility tests and contexts
Machine learning

ABSTRACT

Background: Falls are the leading cause of injury and mortality in older adults; however, the relative contributions of different fall risk factor domains remain unclear. A past fall history strongly predicts future falls, making fall history classification critical for prospective risk assessment.
Objective: This study compared three domains for classifying fall history status in older adults as the basis for fall risk assessment.
Study design: Cross-sectional observational study.
Methods: We analyzed the G-STRIDE dataset (163 older adults; mean age [standard deviation] = 82.6 [6.2] years; 72.4% female; 52.8% fallers). Three domains were examined: demographic information (DGI), clinical scales and questionnaires (CSQ), and mobility tests and contexts (MTC). Four classifiers (logistic regression, support vector machine, random forest, and artificial neural network) were evaluated using 10-fold cross-validation, leave-one-out, and hold-out validation. Bootstrap 95% confidence intervals (CIs) and paired t-tests were used for area under the receiver operating characteristic curve (AUC) comparisons.
Results: MTC alone achieved AUC = 0.89 (95% CI: 0.83–0.94), significantly outperforming DGI (AUC = 0.76, $P < 0.001$). DGI plus MTC showed a marginal advantage over DGI plus CSQ ($P = 0.064$). The evolutionary optimization identified a seven-variable subset dominated by mobility measures that matched the full-model performance (AUC = 0.90). A multi-method feature importance analysis identified the examination location, frailty index, and short Falls Efficacy Scale-International as the top predictors. The external validation of GAIT2CARE (N = 127) achieved an AUC of 0.802 for DGI plus MTC.
Conclusions: Objective mobility tests combined with demographic data provided efficient fall risk assessment without extensive questionnaire-based assessments, supporting streamlined clinical screening.

1. Introduction

7w? > Falls represent a major public health challenge in aging societies worldwide and are a leading cause of injury-related morbidity and mortality among older adults. Each year, approximately one in three community-dwelling adults aged 65 years and older experiences at least one fall, with substantially higher rates in institutional settings such as nursing homes and assisted living facilities.[1,2] The consequences of falls extend beyond physical injury and include reduced independence, fear of falling, activity restriction, higher healthcare costs, and premature institutionalization. As the global population continues to age, the burden of fall-related injuries on healthcare systems is projected to increase significantly, emphasizing the urgent need for effective fall risk screening and prevention strategies.[3,4] Importantly,

Abbreviations: ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; BMI, body mass index; CI, confidence interval; CSQ, clinical scales and questionnaires; DGI, demographic information; FES-I, Falls Efficacy Scale-International; GELU, Gaussian error linear unit; LR, logistic regression; LOO, leave-one-out; MDI, mean decrease in impurity; MTC, Mobility Tests and Contexts; RFs, random forests; ROC, receiver operating characteristic; SPPB, Short Physical Performance Battery; SVM, support vector machine; TUG, Timed Up and Go

* Correspondence to: Center for Robotics, School of Control Science and Engineering, Shandong University, Jinan, Shandong Province 250061, China.
E-mail address: zhanghuanghe@sdu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.hcr.2026.100069
Received 19 December 2025; Received in revised form 11 March 2026; Accepted 15 March 2026

extensive epidemiological evidence supports that a history of falls is one of the strongest predictors of future fall events in older adults, with systematic reviews reporting that prior fallers face a two- to three-fold increase in prospective fall risk compared with nonfallers.[1,5,6] This relationship between past and future falls underpins the clinical rationale for fall history classification; accurately identifying individuals with a positive fall history serves as a direct proxy for prospective fall risk assessment, enabling timely intervention before recurrent falls occur.

Traditionally, fall risk assessment relies on fall risk factors that encompass patient-specific factors that can be evaluated via clinical examinations and structured assessment protocols. Fall risk factors can be broadly categorized into three domains: demographic information, clinical assessments, and mobility tests. Understanding the relative contribution of each domain to fall risk prediction is essential for designing efficient screening protocols that balance the assessment burden with predictive accuracy. Demographic characteristics have long been recognized as fundamental indicators of vulnerability to falls. Age is perhaps the most robust predictor, with fall risk increasing substantially after 65 years of age and accelerating further beyond 75 and 80 years of age, due to progressive declines in physiological reserves, sensory function, and neuromuscular control.[1,2] Gender differences are also well documented; women exhibit higher fall rates than men, which is attributable to factors including lower bone mineral density, greater prevalence of osteoporosis, differences in muscle mass distribution, and potentially higher engagement in household activities that involve fall hazards. Body mass index (BMI) and anthropometric measures influence fall risk through multiple pathways, as underweight individuals may lack the muscular strength to recover balance, whereas obesity impairs mobility and shifts the center of mass, both of which increase fall susceptibility. Living environment, particularly residence in a nursing home versus community dwelling, reflects not only frailty and functional dependence but also exposure to different environmental hazards and care structures. However, demographic information captures only broad risk categories and provides limited insight into the specific functional impairments and physiological vulnerabilities that directly precipitate falls. Therefore, screening based solely on demographic profiles lacks the precision required for targeted intervention planning and may miss high-risk individuals whose functional status is disproportionately poor relative to their demographic profile.

Clinical assessment instruments offer a more comprehensive evaluation of fall risk by quantifying functional capacity, frailty, and psychological factors. The Short Physical Performance Battery (SPPB) combines balance tests, gait speed assessment, and chair stand performance to provide an integrated measure of lower extremity function, which correlates strongly with mobility disability and fall risk.[7,8] The Fried frailty phenotype operationalizes frailty using five criteria, namely unintentional weight loss, self-reported exhaustion, low physical activity, slow gait speed, and weak grip strength, enabling clinicians to identify individuals with compromised physiological reserves who are at an elevated risk of adverse outcomes, including falls.[9] Fear of falling, commonly assessed using the Falls Efficacy Scale-International (FES-I), captures the psychological dimension of fall risk: individuals with a high fear of falling often restrict their activities, leading to deconditioning, muscle weakness, and ultimately increased fall susceptibility.[1] These clinical instruments provide actionable information that can guide multifactorial interventions targeting specific deficits. The primary advantage of such assessments lies in their ability to reveal modifiable risk factors and stratify patients according to their intervention needs. However, clinical assessment has notable limitations. Many rely on subjective self-reports or clinician assessment, introducing potential bias and variability. Questionnaire-based scales can be time-consuming to administer and require trained personnel to accurately score and interpret results. Furthermore, the incremental predictive value of extensive questionnaire batteries over simpler objective tests has not been systematically quantified, raising questions about the optimal balance between assessment comprehensiveness and clinical feasibility in routine screening.[10,11]

Mobility tests provide objective, performance-based measurements of functional capacity that directly reflect the neuromuscular and biomechanical systems responsible for postural stability and safe ambulation. The Timed Up and Go (TUG) test, in which participants rise from a chair, walk 3 m, turn, return, and sit down, assesses multiple domains, including lower limb strength, dynamic balance, gait speed, and functional mobility.[7,12] However, systematic reviews and meta-analyses have questioned the ability of the TUG to predict falls in isolation, suggesting that it should not be used as a standalone screening tool in community-dwelling older adults.[13] Similarly, the 4-metre walk test yields gait speed, a powerful predictor of adverse health outcomes and a sensitive marker of functional decline. Studies comparing the TUG and gait speed have found that both measures predicted geriatric outcomes with similar accuracy, although gait speed might be simpler to implement.[14] These tests are quick, require minimal equipment, and can be performed in diverse clinical settings by nursing staff or trained assistants without specialized biomechanical expertise.[12] Their objective nature reduces inter-rater variability and enhances reproducibility across sites and populations. Moreover, mobility tests capture real-time functional performance, rather than historical or perceived ability, thus providing a snapshot of an individual's current physiological capacity. Recent advances in wearable sensors have further supported the value of objective gait assessment for fall risk evaluation, particularly through measures related to walking performance and velocity.[15–19] These developments reinforce the clinical relevance of objective mobility testing, but they do not remove the need to determine which assessment domains provide the best balance between predictive value and assessment burden. Despite these advantages, mobility tests alone do not reveal the underlying causes of impaired performance because poor TUG or gait speed scores may result from muscle weakness, joint pain, neurological deficits, fear of falling, and environmental factors. Interpreting the clinical significance of test scores and determining precise fall risk thresholds without integrating additional contextual information is challenging. Furthermore, many clinicians lack the analytical tools to translate continuous test metrics into actionable risk categories, suggesting that machine learning approaches may be necessary to fully exploit the predictive potential of objective mobility data.[20–22]

The relative contributions of demographic information, clinical assessments, and objective mobility tests to fall risk prediction remain unclear. Although each domain has been studied individually, systematic comparisons to isolate their incremental predictive value are scarce. Most previous studies have aggregated diverse clinical variables into composite models without dissecting domain-specific contributions, leaving clinicians uncertain as to which assessments offer the best return on investment in terms of time, training, and resources.[10,11] This gap is particularly important in resource-limited settings, where comprehensive assessments may be impractical and efficient screening protocols are needed. The integration of autonomous mobile robots with wearable sensors has been promising for engaging older adults in walking exercises while collecting accurate gait data, suggesting new possibilities for continuous fall risk monitoring outside traditional clinical environments.[23] Recent studies have demonstrated that combining wearable inertial sensors with clinical tests, such as the TUG and 6-minute walking tests, could improve the accuracy of fall risk prediction in nursing home residents.[24] Moreover, the interplay between simple demographic screening, moderately intensive questionnaire-based evaluations, and objective testing has not been explored in a unified analytical framework that considers different machine learning algorithms and validation strategies. Machine learning approaches, including XGBoost, random forests (RFs), and neural networks, have been promising for fall risk prediction, with some studies achieving classification accuracies exceeding 80% when combining multiple data sources.[25,26] Addressing this gap requires a structured multilevel analysis that systematically varies the assessment effort while quantifying the predictive performance across clinically meaningful feature sets and robust evaluation schemes.

Table 1
Baseline demographic and anthropometric characteristics of the G-STRIDE study population by fall history status.

Characteristic	Overall (N = 163)	Fallers (n = 86)	Non-fallers (n = 77)
Age (years), mean ± SD	82.6 ± 6.2	84.2 ± 5.5	80.9 ± 6.5
Female, n (%)	118 (72.4)	67 (77.9)	51 (66.2)
Weight (kg), mean ± SD	64.3 ± 13.1	63.1 ± 13.4	65.6 ± 12.7
Height (m), mean ± SD	1.57 ± 0.10	1.52 ± 0.08	1.62 ± 0.10
BMI (kg/m^2), mean ± SD	26.2 ± 5.0	27.2 ± 5.6	25.0 ± 4.1
Nursing home resident, n (%)	53 (32.5)	31 (36.0)	22 (28.6)

Values are presented as mean ± standard deviation or n (%).
This table is descriptive; no between-group hypothesis tests are reported.
Abbreviations: BMI, body mass index; SD, standard deviation.

Beyond comparing predictive accuracy alone, the clinical contribution of the present study lies in quantifying the trade-off between assessment burden and discriminatory performance across progressively more complex feature sets. This comparative framework allows a workflow-oriented interpretation: basic demographic information may support initial low-burden screening, objective mobility tests may serve as the main second-level assessment because they provide strong discrimination with limited administration time, and more comprehensive questionnaire-based assessments may be reserved for cases in which additional clinical characterization is needed. In this way, the study aims not only to compare models, but also to inform how fall risk screening protocols may be streamlined in rehabilitation and geriatric care settings.

The present study addresses these limitations via a comprehensive comparative analysis of fall risk factors for fall history classification in older adults. We structured the clinical variables into three distinct domains and constructed six feature sets corresponding to escalating levels of assessment effort, ranging from minimal demographic profiles to comprehensive multidomain evaluations. We trained and evaluated four standard machine learning classifiers, namely logistic regression (LR), support vector machine (SVM), RF, and artificial neural network (ANN), using three internal validation strategies to quantify the influence of the modeling and evaluation choices. Additionally, we performed an exhaustive search of low-dimensional SVM models to identify compact feature combinations that approached the performance of full clinical models. We further employed an evolutionary optimization procedure for automated feature subset selection to discover parsimonious variable combinations that retained high predictive accuracy while reducing the clinical assessment burden.

We selected three domains—demographic information (DGI), clinical scales and questionnaires (CSQ), and mobility tests and contexts (MTC)—because they represent the core components of routine clinical fall risk assessments, which could be obtained without specialized instrumentation. Other potentially relevant factors, such as cognitive function and medication use, were not available in the dataset and were acknowledged as limitations.

Based on prior literature and clinical reasoning, we formulated three hypotheses:

(1) **H1:** MTC provide a stronger discriminative capacity for fall history classification than CSQ, given that mobility tests directly measure functional capacity and are supported by substantial evidence as objective fall risk indicators.[7,12]
(2) **H2:** Adding DGI to mobility-based models improves classification performance because demographic factors such as age, sex, and body composition capture complementary risk dimensions that are not reflected in test performance alone.[1]
(3) **H3:** Combining all three domains—DGI, MTC, and CSQ—will yield the best overall classification performance, because each domain captures different dimensions of fall risk.

Through this systematic approach, we aim to clarify how different fall risk factor domains contribute to fall history discrimination, guide practical protocol design by quantifying the cost-benefit tradeoffs inherent in multilevel assessments, and identify efficient screening strategies suitable for deployment in diverse clinical and community settings.

2. Methods

2.1. Datasets

We analyzed the G-STRIDE dataset, a publicly available cohort of older adults recruited from two outpatient clinics in public hospitals and three public nursing homes in Spain.[27,28] The dataset contains comprehensive clinical, functional, and gait assessments as well as fall history status in the year before the recorded visit. A total of 163 participants were included in the analysis (mean age [standard deviation] = 82.6 [6.2] years; range: 70–98 years; 118 females [72.4%]; 86 fallers [52.8%]).

Participants were labeled as fallers if they self-reported at least one fall during the previous 12 months, and as nonfallers otherwise. A fall history was ascertained through structured interviews conducted by trained clinicians during the assessment visit. Recent work on this dataset has demonstrated that machine learning models trained on wearable-derived gait features can effectively distinguish fallers from nonfallers, thereby providing a foundation for the present comparative analysis of fall risk factor domains.[29] For the present study, we used only clinical and questionnaire variables, focusing exclusively on fall risk factors that could be assessed without specialized instrumentation.

Table 1 presents the baseline characteristics of the study population stratified by fall history status.

2.2. Fall risk factor domains and feature sets

We considered three clinically meaningful domains of fall risk factors, representing patient-specific factors that could be assessed through routine clinical evaluations. Table 2 summarizes the number of variables and assessment burden for each domain.

DGI represents basic demographic and anthropometric information. It includes six variables: age, sex, residence in a nursing home, height,

Table 2
Number of variables included in each fall-risk-factor domain and in the combined feature set.

Domain	Variables, n
Demographic Information	6
Clinical Scales and Questionnaires	12
Mobility Tests and Contexts	5
All combined	23

weight, and BMI. These variables are usually available from medical records or brief interviews and form the minimum profile required for low-cost screening.

CSQ provides subjective or clinician-rated information obtained using standardized instruments. It includes 12 variables: the individual components of the Fried frailty phenotype (unintentional weight loss, exhaustion, low activity level, slow gait, and low grip strength), manual grip strength (in kg), frailty index, SPPB subscores (balance, gait speed, and chair stand), total score, and short FES-I score.[1,7] The term "low activity level" refers to participants who report engaging in physical activity less than once per week, based on structured interview responses. These assessments require questionnaires and scoring procedures but do not require specialized hardware.

MTC represents objective performance-based assessments of mobility and balance along with contextual testing conditions. This domain contains five variables: the 4-metre walk test (measured as time in seconds and derived gait speed in m/s), TUG test duration in seconds, examination location (categorized as consultation room, nursing home, or familiar environment, such as the participant's home), and type of walking surface (categorized as a plane or varied surface, including carpets, tiles, or uneven flooring). These categories were predefined in the original dataset based on clinical documentation. The tests are simple, fast, and widely recommended in clinical guidelines for mobility assessment.[12]

From these domains, we defined six feature sets:

(1) DGI only (6 variables).
(2) CSQ only (12 variables).
(3) MTC only (5 variables).
(4) DGI combined with CSQ (18 variables).
(5) DGI combined with MTC (11 variables).
(6) All variables: DGI combined with CSQ and MTC (23 variables).

This design enables a direct comparison among pure demographic screening, moderately intensive questionnaire-based assessments, and objective testing. It also supports a cost-benefit interpretation of the amount of predictive gain obtained when moving from one assessment level to another.

2.3. Preprocessing

Starting with the original spreadsheet, we implemented a reproducible preprocessing pipeline in Python. We first retained only variables that belonged to the aforementioned three fall risk factor domains and removed identifiers, administrative fields, and all external sensor metrics.

The missing data was minimal for most variables. The manual grip strength had three missing values (1.8%), the manual force clinical assessment had three missing values (1.8%), and the short FES-I score had two missing values (1.2%). The TUG test duration had 16 missing values (9.8%), which were imputed using linear regression, as described below. All other variables had complete data.

For categorical items with values "yes," "no," and "incapable," we mapped "yes" and "incapable" to one and "no" to zero. This coding reflects the fact that the inability to perform a test (incapable) typically indicates severe impairment, similar to a positive finding for fall risk factors such as slow gait or weak grip strength.[9] Other non-numeric entries representing missing information were converted to missing values.

The duration of the TUG test was sometimes missing for participants who only completed the 4-metre walk. We fitted a simple linear regression between the TUG time and 4-metre walk time for the 147 participants who completed both tests. The resulting regression equation was as follows: TUG time $= 2.30 \times$ 4-metre walk time $+ 3.08$, with coefficient of determination $R^2 = 0.70$ (Pearson $r = 0.84$, $P < 0.001$). This strong linear relationship supported the validity of the imputation approach. Fig. 1 shows this relationship, displaying the observed data points together with the regression line and imputed values.

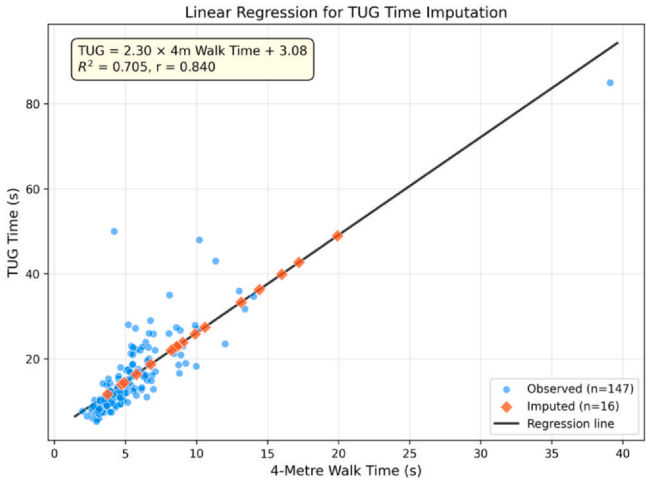

Fig. 1. Linear regression between TUG time and 4-metre walk time.
Blue circles represent the 147 observed data points, and the solid line shows the fitted regression. Orange diamonds indicate the 16 imputed TUG values. The annotation reports the regression equation and goodness-of-fit statistics. Abbreviation: TUG , Timed Up and Go.

For continuous variables, missing values were imputed using the median of the training subset for each validation fold. For binary variables, missing values were imputed with the most frequent category (mode) in the training data. We selected mode imputation for binary variables because it preserved the original distribution and avoided introducing artificial intermediate values. However, we acknowledge that this approach may slightly underestimate variance and affect class balance if missing data are not randomly distributed.

For models requiring standardized inputs (LR, SVM, and ANN), z-score normalization was applied using the mean and standard deviation computed from the training partition. The RF was operated at the original scale. All preprocessing steps, including imputation and standardization, were performed separately within each outer training partition so that information from the test data was never used for computing the transformation parameters. This procedure prevents information leakage and yields more realistic performance estimates.[10,30,31]

2.4. Machine learning models

We evaluated four commonly used binary classifiers spanning a range of model complexities and interpretabilities. These classifiers were selected based on their widespread use in clinical prediction modelling, complementary strengths, and representation in the fall risk literature.[10,20,21] LR provides interpretable coefficients and serves as the baseline linear model. An SVM with a radial basis function kernel can capture nonlinear relationships. RF offers robustness against overfitting and provides feature importance measures. An ANN is a flexible nonlinear approach that can model complex interactions.

LR was implemented with an ℓ_2 penalty and regularization parameter $C = 1$. We used the L-BFGS optimizer with a maximum of 1000 iterations and applied standardization, as described above. The regularization parameter was set to the default value, commonly used in clinical applications.

SVM employed a radial basis function kernel with regularization parameter $C = 1$ and gamma equal to the reciprocal of the product of the number of features and their variance (the "scale" option in scikit-learn). Probability estimation was enabled, such that posterior probabilities were available for receiver operating characteristic (ROC) analysis.[20] These default hyperparameters were retained to ensure a fair comparison across feature sets without introducing additional optimization bias.

RF was implemented as an ensemble of 100 decision trees with bootstrap sampling and the Gini impurity criterion. RF was trained on the original unstandardized features. The number of trees was set to 100, as this typically provided a stable performance while remaining computationally efficient.

ANN is a fully connected feedforward network with three hidden layers of 128 neurons each. The hidden layers use Gaussian error linear unit (GELU) activations, followed by dropout (0.3) for regularization, and the output layer uses a sigmoid activation. We adopted this architecture as a stable, general-purpose nonlinear baseline that can capture interactions among demographic, questionnaire, and mobility variables while retaining a consistent model structure across feature sets. At the same time, given the modest sample size and tabular clinical data, we treat ANN as a comparative benchmark rather than a preferred deployment model, and its results are interpreted with appropriate caution regarding potential overparameterization and overfitting.

We acknowledge that the hyperparameters are fixed rather than tuned through cross-validation, which may not yield optimal performance for each model. However, this approach ensures a fair comparison across feature sets and avoids the potential overfitting that may result from an extensive hyperparameter search on a small dataset.

For all models, the decision threshold for computing classification accuracy was set to 0.4. This value was chosen to provide a slightly more sensitive operating point for faller identification in this imbalanced clinical dataset, where the cost of missing at-risk individuals may be greater than the cost of additional false positives. We therefore report AUC as the primary threshold-independent discrimination metric, while accuracy at the 0.4 threshold is presented as a secondary summary of one clinically reasonable operating point rather than a universally optimal decision rule.

2.5. Validation schemes

Given the limited sample size, we adopted three internal validation schemes.

Ten-fold cross-validation used stratified splits with similar faller proportions across folds. For each fold, a model was trained on nine folds and evaluated on the remaining fold. The performance was aggregated by concatenating the predictions from all folds. This validation approach is consistent with recent work demonstrating that transductive learning models can substantially reduce measurement errors in the ambulatory gait analysis of elderly residents in assisted living facilities.[32]

Leave-one-out (LOO) cross validation considered each participant once as a test case while the remaining participants formed the training set. This scheme maximizes data use but can be computationally demanding.

Single hold-out evaluation used a single stratified split with 80% of the data for training and 20% for testing. The models were trained only once on the training subset, and their performance on the hold out test set provided a single estimate. This strategy is common in clinical machine learning studies but can yield unstable estimates in small datasets.[10,11]

For each model and feature set, we computed the ROC curves and corresponding area under the curve, as well as the classification accuracy. For the full feature set, we summarized the accuracy distributions across repeated experiments using box plots.

2.6. Optimization-based feature subset selection

In addition to the exhaustive enumeration of one- and two-variable models, we sought to identify the optimal feature subsets of arbitrary size that balanced predictive performance against assessment complexity. We formulated this task as a combinatorial optimization problem and employed an evolutionary search strategy to efficiently explore the space of possible variable combinations.

Specifically, we represent each candidate subset as a binary indicator vector of length P, where P denotes the total number of fall risk factors in the full pool. A value of one at position i indicates the inclusion of the corresponding variable, whereas zero indicates its exclusion. The quality of each subset S is quantified using a fitness function defined as

$$\mathrm{F}(S) = \mathrm{AUC}_{\mathrm{LR},\mathit{five}-\mathrm{fold}}(S) - \lambda\frac{|S|}{P},$$

where $\mathrm{AUC}_{(\mathrm{LR,\ five\text{-}fold})}(S)$ represents the mean area under the ROC curve obtained via stratified 5-fold cross validation using LR on the selected features, $|S|$ denotes the number of included variables, and $\lambda = 0.01$ serves as a sparsity penalty that discourages excessively large subsets. This formulation encourages the discovery of parsimonious feature combinations that can achieve high discrimination while remaining clinically practical.

The search proceeds through the iterative refinement of a population of candidate solutions. We initialized a population of 40 individuals with random binary vectors and evolved them over 40 generations. In each generation, individuals were selected for reproduction based on tournament selection with a tournament size of three. Offspring were generated through a single-point crossover with a probability of 0.8, and random bit flip mutations were applied with a probability of 0.05 per gene. To prevent degenerate solutions with no selected features, all zero chromosomes were repaired by activating a randomly chosen variable. Throughout the search, we tracked the best individual encountered with respect to fitness, and recorded its constituent variables, mean area under the curve, and subset cardinality.

Upon convergence, we extracted the optimal subset and re-evaluated it using all four classifiers under the three validation schemes described previously. This post hoc evaluation confirmed whether the selected variables were generalized across different modelling choices and provided a fair comparison against the full feature set and exhaustively enumerated low-dimensional models.

2.7. Statistical analysis

All analyses were performed in Python 3.13 using scikit-learn 1.8.0, SciPy 1.17.0 and PyTorch 2.10.0. Continuous variables were expressed as mean ± standard deviation and categorical variables as counts (percentages). Classification performance is reported as accuracy and area under the receiver operating characteristic curve (AUC) with bootstrap 95% confidence intervals (CI). Pairwise comparisons of AUC across feature sets were conducted using paired t tests on fold-level AUC values from ten-fold cross-validation and the DeLong test on pooled out-of-fold predicted probabilities. All tests were two-sided with a significance level of $\alpha = 0.05$.

3. Results

3.1. Performance across feature sets in cross-validation

We evaluated the six feature sets and four classifiers using 10-fold cross-validation. Bootstrap 95% CIs were computed with 1000 resamples.

DGI alone already provides predictive power above chance, with LR achieving an AUC of 0.76 (95% CI: 0.68–0.84). CSQ alone performs at an intermediate level, with LR reaching an AUC of 0.82 (95% CI: 0.75–0.88). MTC yields the highest performance among single domains, with LR achieving an AUC of 0.89 (95% CI: 0.83–0.94).

Combining DGI with CSQ improves over either domain alone (AUC = 0.84, 95% CI: 0.77–0.90 for LR) but does not surpass DGI combined with MTC (AUC = 0.89, 95% CI: 0.83–0.94). Adding CSQ on top of DGI and MTC produces similar performance (AUC = 0.87, 95% CI: 0.81–0.92 for LR). These trends indicate that objective mobility measures extract most of the discriminative information, while

Table 3
Pairwise comparisons of AUC across feature sets for logistic regression under 10-fold cross-validation.

Comparison	t-statistic	*P* value	*Z* (DeLong)	*P* (DeLong)
MTC vs. DGI	4.85	< 0.001	1.624	0.104
MTC vs. CSQ	1.95	0.083	1.269	0.205
DGI + MTC vs. DGI	5.73	< 0.001	1.866	0.062
DGI + MTC vs. CSQ	1.63	0.138	1.183	0.237
DGI + MTC vs. DGI + CSQ	2.11	0.064	1.008	0.313
DGI + MTC vs. All variables	0.90	0.394	0.297	0.766
All variables vs. DGI	3.54	0.006	1.569	0.117
CSQ vs. DGI	1.12	0.291	0.438	0.662

P values are from paired t tests on fold-level AUC values from 10-fold cross-validation. Z (DeLong) and *P* (DeLong) are from the DeLong test for comparing correlated receiver operating characteristic curves on pooled out-of-fold predictions.
Abbreviations: AUC, area under the receiver operating characteristic curve; DGI, demographic information; MTC, mobility tests and contexts; CSQ, clinical scales and questionnaires.

demographic variables contribute additional but modest gains. Questionnaire-based assessments provide limited further improvement when objective tests are already available.

To assess the statistical significance of these performance differences, we conducted paired *t*-tests on the fold-level AUC values from the 10-fold cross-validation procedure. Table 3 summarizes the key comparisons using logistic regression (LR), and Table 4 summarizes the accuracy and AUC values across feature sets and models.

MTC significantly outperformed DGI alone ($P < 0.001$). DGI plus MTC showed a marginal advantage over DGI plus CSQ ($P = 0.064$), while All Variables significantly outperformed DGI alone ($P = 0.006$). The differences between MTC alone and CSQ alone, and between DGI plus MTC and All Variables were not statistically significant, indicating that clinical scales provided limited incremental value when mobility tests were already included.

Table 4
Classification accuracy and AUC for each feature set and model under 10-fold cross-validation.

Feature set	Model	Accuracy	AUC [95% confidence interval]
DGI	LR	0.69	0.76 [0.68–0.84]
DGI	SVM	0.63	0.70 [0.61–0.78]
DGI	RF	0.64	0.73 [0.65–0.81]
DGI	ANN	0.69	0.72 [0.63–0.80]
CSQ	LR	0.75	0.82 [0.75–0.88]
CSQ	SVM	0.75	0.80 [0.72–0.86]
CSQ	RF	0.79	0.83 [0.76–0.89]
CSQ	ANN	0.76	0.82 [0.75–0.88]
MTC	LR	0.81	0.89 [0.83–0.94]
MTC	SVM	0.77	0.85 [0.78–0.91]
MTC	RF	0.79	0.86 [0.80–0.92]
MTC	ANN	0.80	0.87 [0.81–0.93]
DGI + CSQ	LR	0.75	0.84 [0.77–0.90]
DGI + CSQ	SVM	0.74	0.82 [0.75–0.88]
DGI + CSQ	RF	0.75	0.83 [0.76–0.89]
DGI + CSQ	ANN	0.79	0.83 [0.76–0.89]
DGI + MTC	LR	0.80	0.89 [0.83–0.94]
DGI + MTC	SVM	0.79	0.88 [0.82–0.93]
DGI + MTC	RF	0.75	0.85 [0.78–0.91]
DGI + MTC	ANN	0.83	0.88 [0.82–0.93]
All variables	LR	0.76	0.87 [0.81–0.92]
All variables	SVM	0.80	0.88 [0.82–0.93]
All variables	RF	0.79	0.87 [0.81–0.92]
All variables	ANN	0.80	0.85 [0.78–0.91]

Values in the AUC column are reported as AUC [95% confidence interval].
Abbreviations: ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; CSQ, clinical scales and questionnaires; DGI, demographic information; LR, logistic regression; MTC, mobility tests and contexts; RF, random forest; SVM, support vector machine.

3.2. Receiver operating characteristic curves for the six feature sets

Fig. 2 shows characteristic curves for the six feature sets. Each panel shows the LR, SVM, RF, and ANN models evaluated under LOO, 10-fold, and single hold-out validations. Across all feature sets, the three validation schemes produced consistent trends, but the single hold-out results exhibited wider variability, especially for ANNs. For the MTC and DGI plus MTC sets, the curves are consistently shifted towards the upper left corner, confirming the strong discriminative capacity of these domains.

3.3. Detailed performance of the full clinical feature set

Fig. 3 shows the full clinical feature set that combines all three domains. Fig. 3(a) shows the ROC curves for the four models and three validation schemes. Fig. 3(b) shows box plots of the accuracy across repeated runs on the full feature set. The cross-validation schemes yielded relatively tight accuracy distributions, with median values of approximately 80% for all models. Single hold-out results are more dispersed and occasionally underestimate the achievable performance, particularly for the ANN. These observations highlight the importance of robust resampling-based validation using small clinical datasets.[10,11]

3.4. Optimal feature subsets identified by the selection procedure

The optimization-based feature selection procedure converged to a compact subset of seven variables from the full clinical pool. Table 5 presents the composition of this optimal subset organized by domain.

The selected subset emphasized the mobility test variables, with four of the seven features belonging to this domain. Gait speed derived from the 4-metre walk and TUG test duration appeared as expected, given their established predictive value. Contextual factors, including examination location and walking surface type, also contribute, likely reflecting the environmental influences on test performance. From the CSQ, the low activity level item from the Fried frailty phenotype and chair stand component of the SPPB were retained; the latter captured lower extremity strength and power, whereas most other questionnaire-based items were excluded. Height was the sole demographic variable selected, potentially serving as a proxy for stride length and overall body size.

Cross-validated performance for the optimal seven-feature subset remained strong across all four classifiers: LR achieved an AUC of 0.90 (95% CI: 0.84–0.95) under 10-fold cross-validation, while RF, SVM, and ANN yielded AUC values of 0.85 (95% CI: 0.78–0.91), 0.87 (95% CI: 0.81–0.93), and 0.89 (95% CI: 0.83–0.94), respectively. These results demonstrate that the selected seven-variable subset retains nearly all discriminative information present in the complete clinical pool while substantially reducing assessment burden.

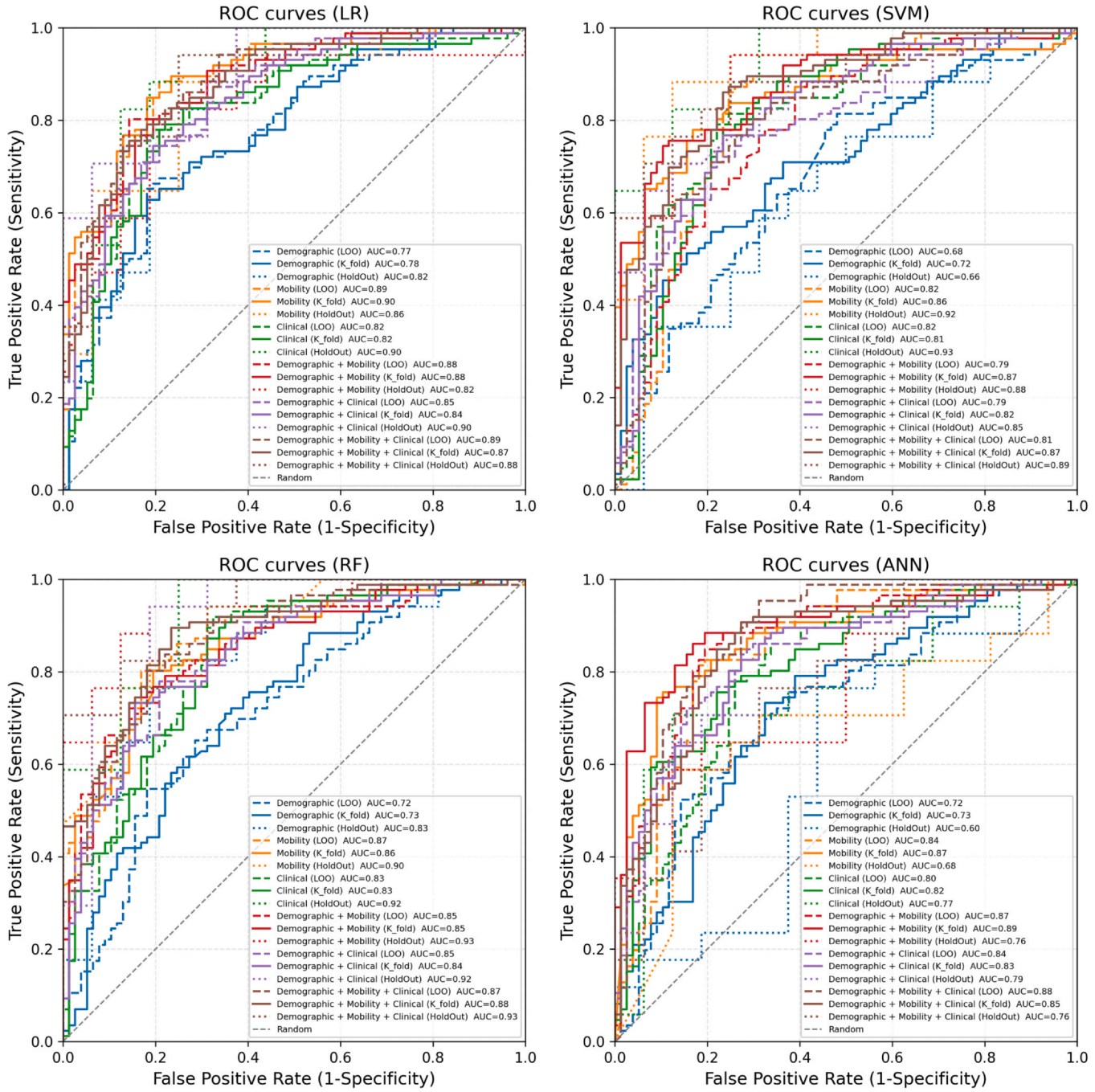

Fig. 2. ROC curves for the four classifiers. (a) LR. (b) SVM. (c) RF. (d) ANN. Each panel shows six feature sets (DGI, CSQ, MTC, DGI + CSQ, DGI + MTC, and All) under LOO, 10-fold cross-validation, and hold-out validation. Abbreviations: ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; LOO, leave-one-out; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine.

Fig. 4 shows the ROC curves for the optimal subset for all three validation schemes. The curves are clustered in the upper left region of the plot, confirming strong discrimination. The performance remains stable across the LOO, 10-fold, and hold-out evaluations, suggesting that the selected features generalize well and are not artifacts of a particular data partition.

For logistic regression (LR) under 10-fold cross-validation, the optimal subset achieved an AUC of 0.90 (95% CI: 0.84–0.95), compared with 0.87 (95% CI: 0.81–0.92) for the full feature set and 0.82 (95% CI: 0.75–0.88) for CSQ alone. This improvement with fewer variables indicates that removing redundant or noisy features can enhance model performance while simplifying clinical implementation.

3.5. Feature importance analysis

To ensure the robustness of the feature importance findings, we employed four complementary methods: the RF mean decrease in impurity (MDI), LR absolute coefficients, model-agnostic permutation importance, and gradient boosting feature importance. All methods were applied to the full feature set.

Fig. 5 presents a heat map comparing the normalized importance across all four methods for the top 15 features.

Table 6 reports the LR coefficients for the most important features, indicating the direction of each feature effect on fall history classification.

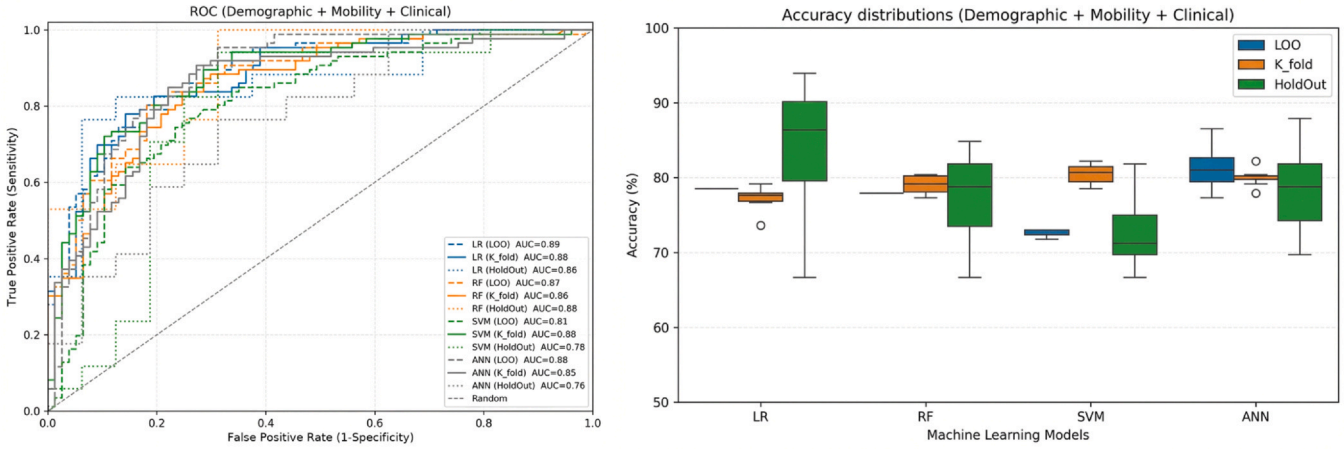

Fig. 3. Detailed performance of the four models on the union of DGI, CSQ, and MTC. (a) ROC curves for the full feature set. (b) Accuracy distributions over repeated runs for the full feature set.

Table 5
Variables included in the optimal seven-feature subset identified by evolutionary optimization.

Domain	Selected variables
DGI	Height (m)
CSQ	Low activity level; chair-stand test (SPPB score)
MTC	Gait speed from 4-metre walk (m/s); TUG duration (s); examination location; walking surface type

Abbreviations: CSQ, clinical scales and questionnaires; DGI, demographic information; MTC, mobility tests and contexts; SPPB, Short Physical Performance Battery; TUG, Timed Up and Go.

Across all four methods, the examination location, frailty index, and short FES-I consistently emerged among the top-ranked predictors. The mean rank across the methods (Table 7) shows that the examination location ranks first, followed by the frailty index, short FES-I, test surface type, and height.

This multi-method analysis confirms that mobility-related variables and clinical frailty indicators dominate feature importance across all analytical approaches. The convergence across the four independent methods strengthens the reliability of these findings compared to relying on a single importance metric.

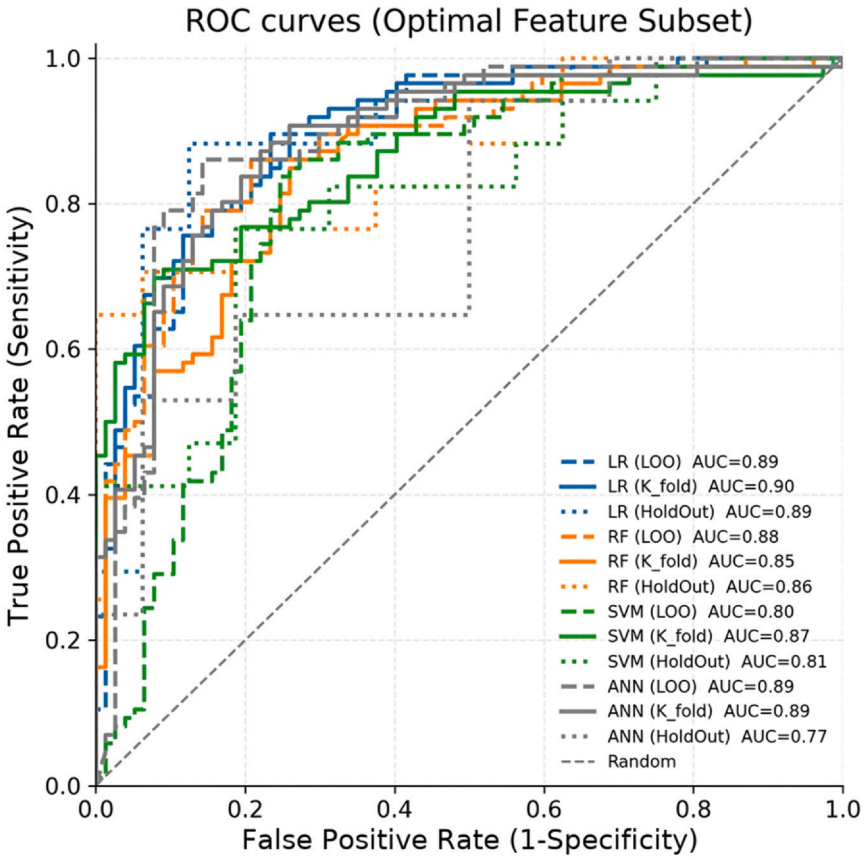

Fig. 4. ROC curves for the optimal feature subset. The figure shows LR, SVM, RF, and ANN under LOO, 10-fold cross-validation, and hold-out validation.

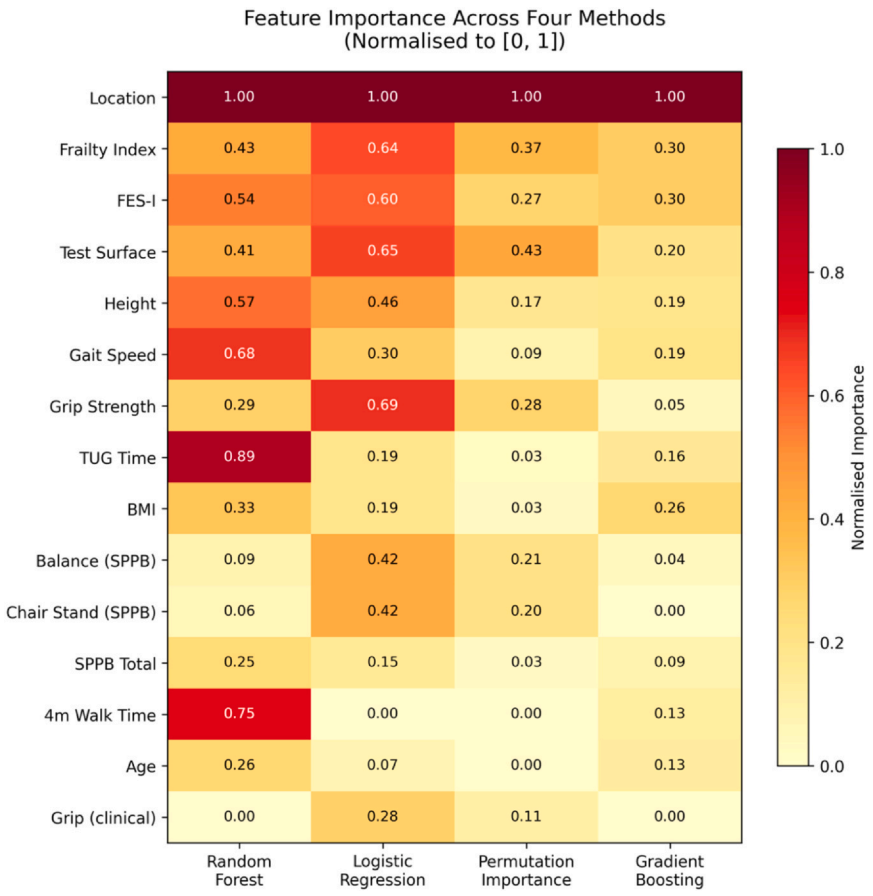

Fig. 5. Feature importance heatmap across four methods: RF (MDI), LR (absolute coefficients), Permutation Importance, and Gradient Boosting. Values are normalised to [0, 1] for each method. Darker colours indicate higher importance. Abbreviations: BMI, body mass index; FES-I, Falls Efficacy Scale-International; LR, logistic regression; MDI, mean decrease in impurity; RF, random forest; SPPB, Short Physical Performance Battery; TUG, Timed Up and Go.

Table 6
Logistic regression coefficients for selected predictors and their direction of association with fall history classification.

Feature	Coefficient	Direction of effect
Examination location	-1.20	-
Frailty index	+ 0.79	+
Short FES-I score	+ 0.74	+
Grip strength (kg)	+ 0.84	+
Test surface type	-0.80	-
Height (m)	-0.57	-
Gait speed (m/s)	-0.40	-

Positive coefficients indicate a positive association with fall history classification; negative coefficients indicate an inverse association. Abbreviation: FES-I, falls efficacy scale-international.

4. Discussion

The present study builds on the extensive literature on fall risk assessment while addressing a specific gap regarding the relative contributions of different fall risk factor domains. Our results are consistent with prior studies. The AUC values achieved in this study for mobility-based models are comparable to those reported by Noh *et al.* using XGBoost on gait-derived features,[25] and Gonzalez-Castro *et al.* across various machine learning approaches for fall risk classification.[22] The slightly higher performance in our study may reflect the clinical population, which includes both outpatient clinic and nursing home participants, with pronounced functional variability.

Many previous studies have examined machine learning models for fall history classification using various clinical variables, clinical scales, or combinations of both.[10,20,21,33] Recent studies using the G-STRIDE system have demonstrated its effectiveness in assessing fall risk and

Table 7
Top 10 predictors ranked by mean importance across four feature-importance methods.

Rank	Feature	RF	LR	Perm.	Gradient Boosting	Mean rank
1	Examination location	0.121	1.201	0.081	0.313	1.0
2	Frailty index	0.057	0.785	0.031	0.095	4.2
3	Short FES-I	0.069	0.742	0.022	0.094	4.8
4	Test surface type	0.056	0.797	0.035	0.062	4.8
5	Height	0.073	0.574	0.015	0.059	6.5
6	Gait speed (m/s)	0.086	0.407	0.007	0.061	7.2
7	Grip strength	0.041	0.842	0.023	0.014	7.8
8	TUG time (s)	0.109	0.264	0.003	0.049	8.5
9	BMI (kg/m^2)	0.046	0.259	0.003	0.083	10.2
10	Balance test (SPPB)	0.019	0.536	0.017	0.012	10.8

Values are normalized importance scores. Abbreviations: RF, random forest; LR, logistic regression; Perm., permutation importance; BMI, body mass index; FES-I, falls efficacy scale-international; SPPB, Short Physical Performance Battery; TUG, Timed Up and Go.

frailty in older adults through comprehensive functional analysis.[28,34] Subsequent research explored the stratification of older adults by frailty status and falls and showed that functional parameters could effectively classify participants according to risk.[35] Furthermore, machine learning comparisons have shown that integrating multiple data types improved the fall history classification accuracy, with models trained on combined datasets achieving superior performance.[11] However, few studies have systematically separated demographic data, questionnaire-based assessments, and objective mobility tests, and even fewer have examined how different fall risk factor domains and validation schemes jointly influenced model performance. By structuring the variables into three domains and six feature sets, and analyzing four models under three validation strategies, we provide a more complete picture of how multilevel clinical assessments contribute to fall history classification.

Using DGI alone, we observed a moderate yet clinically meaningful predictive capacity. Data on age, sex, residence in a nursing home, height, weight, and BMI were readily available and obtained without additional testing. Although these variables do not capture detailed functional status, they reflect the overall vulnerability and living context, which explains their ability to distinguish fallers from nonfallers to some extent. Among the three single-domain feature sets, the MTC showed the strongest performance, followed by the CSQ and DGI. This ordering supported our first hypothesis (H1) that mobility tests would provide a stronger discriminative capacity than clinical scales, with statistical testing confirming that MTC significantly outperformed DGI ($P < 0.001$), although the difference between MTC and CSQ did not reach significance ($P = 0.083$).

When demographic data were combined with each of the two clinical domains, different patterns emerged. DGI plus CSQ is more complex to administer because it requires detailed questionnaires and clinician ratings; however, it does not clearly surpass DGI plus MTC in this dataset. DGI plus MTC achieved strong discrimination, supporting our second hypothesis (H2) that adding demographics improved performance, as Demo + Mobility significantly outperformed Demographics alone ($P < 0.001$), with a marginal advantage over Demo + Clinical ($P = 0.064$). Therefore, MTC appear to offer a favorable balance between predictive power and assessment burden. Objective tests such as the 4-metre walk and TUG are quick, inexpensive, and can be performed by trained staff without subjective interpretation, making them particularly attractive for routine screening.[12]

The combination of all three domains does not yield notable gains over DGI plus MTC despite the increased complexity. The full model achieved a performance similar to that of the simpler combination with overlapping CIs and no significant differences ($P = 0.394$). This observation provides a more nuanced interpretation of our third hypothesis (H3). Although the full model performed well, it did not significantly outperform the simpler DGI + MTC combination, suggesting that CSQ added limited incremental value once objective mobility testing was already available in this cohort.

The limited incremental value of CSQ may be explained by several factors. First, many questionnaire-based items capture the constructs (frailty and fear of falling) that are already reflected in objective mobility performance. Second, self-reported measures may introduce recall bias and subjective variability, which reduce the discriminative power. Third, the SPPB subscores included in the CSQ domain partially overlapped with the mobility tests, creating redundancy. Finally, in this clinical population, objective functional impairment may be more pronounced and easier to detect through mobility testing than questionnaires.

From a practical viewpoint, these findings support a tiered screening strategy. Basic demographic information may serve as a low-burden first step for broad triage, whereas the addition of mobility testing appears particularly useful when a more discriminative yet still feasible evaluation is needed. Comprehensive questionnaire-based assessment may remain valuable in selected cases, such as when psychosocial contributors or multidimensional geriatric concerns require closer characterization, but the present results suggest that it may not always be necessary for initial screening. In busy rehabilitation settings, such a staged workflow could reduce assessment burden while preserving clinically meaningful discrimination.

Regarding the machine learning models, all four classifiers performed competitively. LR, RF, and SVM have already achieved high areas under the curve and accuracy on the best feature sets. The ANN achieved the highest cross-validated performance, especially on DGI plus MTC; however, it also showed greater variability under the single-hold-out evaluation. In applications in which interpretability and stability are essential, simpler models may be preferred, particularly when they can be combined with feature importance analyses or other explainable artificial intelligence methods.[36] Nevertheless, the neural network provides a useful upper bound on achievable performance with the given clinical information.

An exhaustive search for low-dimensional SVM models offers an additional perspective. Two variable models that combine gait speed or TUG indices with fear of falling or related questionnaire scores achieve a performance comparable to that of the full clinical models. This finding reinforces the central role of mobility and balance in fall risk and highlights the potential of simple models for rapid screening. Such compact predictors may be valuable in primary care, community programmes, or telehealth scenarios, where only a few measurements can be collected.

The optimization-based feature selection procedure provides further insight into the variables that contribute the most to fall risk discrimination. The optimal subset of seven variables was dominated by mobility test measures, with gait speed and TUG test duration retained alongside the contextual factors. These findings align with the clinical intuition and existing guidelines that emphasize mobility assessment for fall risk screening.[7,12] Notably, most CSQ items were excluded from the optimal subset, and only the low-activity level component of the Fried frailty phenotype and chair stand test from the SPPB were selected. This suggests that although questionnaire-based scales capture important dimensions, such as frailty and fear of falling, their incremental predictive value beyond objective mobility tests is limited in this cohort. The inclusion of contextual factors such as examination location and walking surface type in the optimal subset was noteworthy. These variables may reflect environmental influences on test performance or serve as proxies for frailty in participant and care settings. Their selection indicated that even simple contextual information can enhance the prediction accuracy.

Multi-method feature importance analysis further validated these conclusions. Examination location emerged as the dominant predictor across multiple methods, with the highest mean importance rank (1.0), followed by the frailty index (4.2), and short FES-I (4.8). Gait speed, although ranked sixth overall (mean rank 7.2), remains consistent with its well-established role as a sensitive marker of the overall health status of older adults. The hierarchical ordering of feature importance, with examination location, frailty index, and short FES-I at the top, mirrored the composition of the optimization-derived optimal subset, with mobility-related and contextual variables occupying the top positions. This convergence across different analytical approaches strengthens confidence in the robustness of our findings and their clinical relevance. Recent studies have further demonstrated that instrumented footwear combined with machine learning models can effectively predict the fall risk in institutionalized older adults by leveraging stride-to-stride gait data, thus providing additional support for the importance of objective gait assessment in fall risk stratification.[37]

The observation that the optimal subset achieved equal or better discrimination than the full clinical set despite containing fewer than half of the variables has important practical implications. Clinicians can potentially streamline fall risk assessments by focusing on a core set of mobility tests supplemented with minimal demographic and frailty information. This parsimonious approach reduces assessment time,

minimizes patient burden, and facilitates implementation in resource-constrained settings. The consistency of the results across different classifiers and validation schemes strengthens confidence in the robustness of the selected features.

From a clinical perspective, these findings support a tiered screening strategy for fall-risk assessment in older adults. In routine rehabilitation or community screening, basic demographic information and brief mobility tests may serve as a practical first-line assessment because they require less time, less patient effort, and fewer questionnaire-based responses, while still preserving substantial discriminatory value. More comprehensive clinical scales and questionnaires may then be reserved for cases in which the initial screen is inconclusive, for patients with complex functional or psychosocial presentations, or when a broader multidimensional rehabilitation plan is being formulated. In this way, the present results suggest that simplified assessments may improve clinical efficiency without necessarily compromising the ability to identify individuals at elevated fall risk.

5. Limitations and future work

Several limitations should be acknowledged. The sample size was modest (N = 163), participants were recruited within one country, and fall history was based on retrospective self-report, which may limit generalizability and introduce recall bias. In addition, some clinically relevant fall-risk factors, including medication use, cognitive status, visual function, and comorbidity burden, were not available. Although we included both internal validation and external testing on the GAIT2CARE dataset, broader validation in more diverse populations is still needed because both datasets originated from the same healthcare system. Accordingly, the present results should be interpreted as comparative performance estimates rather than clinically deployable prediction models. In addition, contextual variables such as examination location and walking surface type may partly encode environment- or site-specific information, so some contextual bias cannot be excluded. Calibration assessment, decision-curve analysis, and formal threshold optimization were beyond the scope of this comparative study and should be prioritized in future external validation work. Nevertheless, the consistency of results across models and validation settings supports the robustness of the main comparative findings and the practical value of simplified mobility-based screening approaches.

6. Conclusion

We presented a comprehensive analysis of fall risk assessment in older adults using fall risk factors from the publicly available G-STRIDE dataset (N = 163; mean age 82.6 years; 52.8% fallers). By organizing the variables into DGI, CSQ, and MTC domains, we examined six feature sets, four machine learning models, and three validation schemes. By exploring low-dimensional SVM models and employing an optimization-based feature selection procedure, we obtained a nuanced view of how different fall risk factor domains contributed to classification performance. Objective MTC emerged as the most informative domain (AUC = 0.89, 95% CI: 0.83–0.94), significantly outperforming demographic features alone ($P < 0.001$), and when combined with basic demographic data, they already provided high discrimination between fallers and nonfallers. CSQ contributed additional but modest improvements, with overlapping CIs and no statistically significant incremental gain ($P = 0.394$). The optimization-based selection procedure identified a parsimonious subset of seven variables dominated by mobility measures that matched the discrimination of the complete clinical pool. A multi-method feature importance analysis using four complementary approaches (RF, LR, permutation importance, and gradient boosting) consistently identified examination location, frailty index, and short FES-I as the dominant predictors. External validation on the independent GAIT2CARE dataset (N = 127) demonstrated good generalizability, with DGI plus MTC features achieving an AUC of 0.802 (95% CI: 0.675–0.907). These findings support the development of practical fall risk assessment tools that rely primarily on objective mobility assessments combined with basic demographic information and provide a foundation for future work that integrates multiple intrinsic and extrinsic risk factors and validates the findings on external cohorts.

CRediT authorship contribution statement

Peng Wu: Conceptualisation, Methodology, Software, Formal analysis, Writing - Original Draft; **Jianlei Fang:** Software, Validation, Data Curation; **Ziyun Ding:** Conceptualisation, Formal analysis, Investigation, Project administration, Validation; **Zeyang Guan:** Methodology, Software, Validation; **Jiachen Wang:** Investigation, Resources; **Yikai He:** Investigation, Resources; **Yihao Zhang:** Software, Visualisation; **Huanghe Zhang:** Conceptualisation, Supervision, Writing - Review & Editing, Project administration, Funding acquisition. All authors reviewed and approved the final manuscript.

Ethical approval

This study used the publicly available G-STRIDE dataset, which was collected with appropriate ethics approval from the participating institutions in Spain. The original data collection was approved by the relevant ethics committees, and all participants provided written informed consent. As this secondary analysis used de-identified publicly available data, no additional ethics approval was required.

Funding

This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62403281, in part by the Taishan Scholars Project (Young Expert Program) under Grant NO. tsqn202408040, in part by the Shandong Excellent Young Scientists Fund Program (Overseas) under Grant 2024HWYQ-019, and in part by the Open Project Fund of International Joint Research Center for Perception and Control of Intelligent Rehabilitation Systems of Sichuan Province under Grant No. 25-H-01.

Data availability

The G-STRIDE dataset is publicly available on Zenodo: https://doi.org/10.5281/zenodo.8003441.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used generative AI tools including GPT-5.2 solely for language polishing and format modification. The AI tools were not involved in data analysis, result interpretation or conclusion formulation. All AI-generated content was rigorously reviewed and revised by the authors, who take full responsibility for the content of the published article.

Acknowledgements

The authors thank the G-STRIDE research team for making their dataset publicly available.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.hcr.2026.100069.

References

1. Montero-Odasso M, van der Velde N, Martin FC, et al. World guidelines for falls prevention and management for older adults: a global initiative. *Age Ageing.* 2022;51(9):afac205. https://doi.org/10.1093/ageing/afac205
2. Salari N, Darvishi N, Ahmadipanah M, Shohaimi S, Mohammadi M. Global prevalence of falls in the older adults: a comprehensive systematic review and meta-analysis. *J Orthop Surg Res.* 2022;17(1):334. https://doi.org/10.1186/s13018-022-03222-1
3. Chen Y, Dai F, Huang S, et al. Global, regional, and national burden of falls among older adults: findings from the Global Burden of Disease Study 2021 and Projections to 2040. *npj Aging.* 2025;11(1):85. https://doi.org/10.1038/s41514-025-00275-4
4. Centers for Disease Control and Prevention. Older adult falls data. Accessed January 15, 2025. ⟨https://www.cdc.gov/falls/⟩.
5. Deandrea S, Lucenteforte E, Bravi F, Foschi R, La Vecchia C, Negri E. Risk factors for falls in community-dwelling older people: a systematic review and meta-analysis. *Epidemiology.* 2010;21(5):658–668. https://doi.org/10.1097/EDE.0b013e3181e89905
6. Tromp AM, Pluijm SME, Smit JH, Deeg DJH, Bouter LM, Lips P. Fall-risk screening test: a prospective study on predictors for falls in community-dwelling elderly. *J Clin Epidemiol.* 2001;54(8):837–844. https://doi.org/10.1016/S0895-4356(01)00349-3
7. Beck Jepsen D, Robinson K, Ogliari G, et al. Predicting falls in older adults: an umbrella review of instruments assessing gait, balance, and functional mobility. *BMC Geriatr.* 2022;22(1):615. https://doi.org/10.1186/s12877-022-03271-5
8. Lauretani F, Ticinesi A, Gionti L, et al. Short-Physical Performance Battery (SPPB) score is associated with falls in older outpatients. *Aging Clin Exp Res.* 2019;31(10):1435–1442. https://doi.org/10.1007/s40520-018-1082-y
9. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci.* 2001;56(3):M146–M157. https://doi.org/10.1093/gerona/56.3.M146
10. Leghissa M, Carrera Á, Iglesias CA. Machine learning approaches for frailty detection, prediction and classification in elderly people: a systematic review. *Int J Med Inf.* 2023;178:105172. https://doi.org/10.1016/j.ijmedinf.2023.105172
11. González-Castro A, Benítez-Andrades JA, González-González R, Prada-García C, Leirós-Rodríguez R. Predicting fall risk in older adults: a machine learning comparison of accelerometric and non-accelerometric factors. *Digit Health.* 2025;11:20552076251331752. https://doi.org/10.1177/20552076251331752
12. Jia S, Si Y, Guo C, et al. The prediction model of fall risk for the elderly based on gait analysis. *BMC Public Health.* 2024;24(1):2206. https://doi.org/10.1186/s12889-024-19760-8
13. Barry E, Galvin R, Keogh C, Horgan F, Fahey T. Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. *BMC Geriatr.* 2014;14:14. https://doi.org/10.1186/1471-2318-14-14
14. Herman T, Giladi N, Hausdorff JM. Properties of the 'timed up and go' test: more than meets the eye. *Gerontology.* 2011;57(3):203–210. https://doi.org/10.1159/000314963
15. Zhang H, Guo Y, Zanotto D. Accurate ambulatory gait analysis in walking and running using machine learning models. *IEEE Trans Neural Syst Rehabil Eng.* 2020;28(1):191–202. https://doi.org/10.1109/TNSRE.2019.2958679
16. Zhang H, Wu C, Huang Y, et al. Two-dimensional deep convolutional neural networks for estimating stride length and velocity in institutionalized older adults. *IEEE Sens J.* 2024;24(17):28267–28275. https://doi.org/10.1109/JSEN.2024.3408900
17. Wang J, Guan Z, Liang T, et al. Multi-task learning for gait phase and gait cycle percentage prediction with wearable sensors in frail older adults. *IEEE J Biomed Health Inf.* 2025. https://doi.org/10.1109/JBHI.2025.3643724
18. Montesinos L, Castaldo R, Pecchia L. Wearable inertial sensors for fall risk assessment and prediction in older adults: a systematic review and meta-analysis. *IEEE Trans Neural Syst Rehabil Eng.* 2018;26(3):573–582. https://doi.org/10.1109/TNSRE.2017.2771383
19. Chen M, Wang H, Yu L, et al. A systematic review of wearable sensor-based technologies for fall risk assessment in older adults. *Sensors (Basel).* 2022;22(18):6752. https://doi.org/10.3390/s22186752
20. Yu X, Cai Y, Yang R, Ma F, Kim W. Revisiting sensor-based intelligent fall risk assessment for older people: a systematic review. *Eng Appl Artif Intell.* 2025;144:110176. https://doi.org/10.1016/j.engappai.2025.110176
21. Lockhart TE, Soangra R, Yoon H, et al. Prediction of fall risk among community-dwelling older adults using a wearable system. *Sci Rep.* 2021;11(1):20976. https://doi.org/10.1038/s41598-021-00458-5
22. González-Castro A, Leirós-Rodríguez R, Prada-García C, Benítez-Andrades JA. The applications of artificial intelligence for assessing fall risk: systematic review. *J Med Internet Res.* 2024;26:e54934. https://doi.org/10.2196/54934
23. Zhang H., Chen Z., Zanotto D., Guo Y. Robot-assisted and wearable sensor-mediated autonomous gait analysis. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA). 2020:6795-6802. https://doi.org/10.1109/ICRA40945.2020.9196757.
24. Buisseret F, Catinus L, Grenard R, et al. Timed up and go and six-minute walking tests with wearable inertial sensor: one step further for the prediction of the risk of fall in elderly nursing home people. *Sensors (Basel).* 2020;20(11):3207. https://doi.org/10.3390/s20113207
25. Noh B, Youm C, Goh E, et al. XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes. *Sci Rep.* 2021;11(1):12183. https://doi.org/10.1038/s41598-021-91797-w
26. Speiser JL, Callahan KE, Houston DK, et al. Machine learning in aging: an example of developing prediction models for serious fall injury in older adults. *J Gerontol A Biol Sci Med Sci.* 2021;76(4):647–654. https://doi.org/10.1093/gerona/glaa138
27. García-de-Villa S, Neira GG, Álvarez MN, et al. A database with frailty, functional and inertial gait metrics for the research of fall causes in older adults. *Sci Data.* 2023;10(1):566. https://doi.org/10.1038/s41597-023-02428-0
28. Neira-Álvarez M, Rodríguez-Sánchez C, Huertas-Hoyas E, et al. Predictors of fall risk in older adults using the G-STRIDE inertial sensor: an observational multicenter case-control study. *BMC Geriatr.* 2023;23(1):737. https://doi.org/10.1186/s12877-023-04379-y
29. Fang J., Guan Z., Wang J., et al. Machine learning models for fall risk assessment using wearable-derived gait features on the GSTRIDE dataset. In: Proceedings of the International Conference on Deep Learning and Computer Vision (DLCV). 2025. In press.
30. Guan Z, Cai J, Wang J, et al. Accuracy and precision of wearable-derived gait parameters: how these affect the performance of models for fall prediction in the elderly. *IEEE Trans Neural Syst Rehabil Eng.* 2025;33:4255–4266. https://doi.org/10.1109/TNSRE.2025.3623129
31. Cai J, Guan Z, Wang J, et al. Impact of gait parameters and their variability on fall risk assessment accuracy using wearable sensor. *IEEE Trans Neural Syst Rehabil Eng.* 2025;33:1996–2003. https://doi.org/10.1109/TNSRE.2025.3572109
32. Zhang H, Duong TTH, Rao AK, et al. Transductive learning models for accurate ambulatory gait analysis in elderly residents of assisted living facilities. *IEEE Trans Neural Syst Rehabil Eng.* 2022;30:124–134. https://doi.org/10.1109/TNSRE.2022.3143094
33. Mohan D, Chong PHJ, Gutierrez J. A novel cooperative AI-based fall risk prediction model for older adults. *Sensors (Basel).* 2025;25(13):3991. https://doi.org/10.3390/s25133991
34. Álvarez MN, Ruiz ARJ, Neira GG, et al. Assessing falls in the elderly population using G-STRIDE foot-mounted inertial sensor. *Sci Rep.* 2023;13(1):9208. https://doi.org/10.1038/s41598-023-36241-x
35. Neira-Álvarez M, Huertas-Hoyas E, Novak R, et al. Stratification of older adults according to frailty status and falls using gait parameters explored using an inertial system. *Appl Sci (Basel).* 2024;14(15):6704. https://doi.org/10.3390/app14156704
36. Tang YT, Romero-Ortuno R. Using explainable artificial intelligence for the prediction of falls in older adults. *Algorithms.* 2022;15(10):353. https://doi.org/10.3390/a15100353
37. Zhang H, Wu C, Huang Y, Song R, Zanotto D, Agrawal SK. Fall risk prediction using instrumented footwear in institutionalized older adults. *IEEE Trans Neural Syst Rehabil Eng.* 2024;32:4260–4269. https://doi.org/10.1109/TNSRE.2024.3510300