Multimodal Machine Learning Approach for Timed-Up-and-Go Test Score Prediction from Wearable Gait Data

Nabid Faiem[1(✉)], Tunc Asuroglu[1,2], Koray Acici[3], Antti Kallonen[1], and Mark van Gils[1]

[1] Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland
nabid.faiem@tuni.fi
[2] VTT Technical Research Centre of Finland, 33101 Tampere, Finland
[3] Department of Artificial Intelligence and Data Engineering, Ankara University, 06830 Ankara, Turkey

Abstract. Falls are a major cause of injury and disability in older adults and people with Parkinson's disease (PD). The Timed-Up-and-Go (TUG) test assesses mobility and falls risk but requires supervised clinical testing, limiting its regular use. We investigated whether TUG time can be estimated from two-minute walking data captured using in-shoe vertical ground reaction force (GRF) sensors, which are easy to use during daily life and provide detailed gait information. While prior studies have estimated TUG from wearable sensors, how different input modalities such as gait representations and demographics contribute individually and in combination has received little attention.

Using the PhysioNet GRF dataset, we analyzed 100 participants (62 PD, 38 controls) with demographics (age, sex, height, and weight) and 16-channel GRF recordings. A Perceiver-based multimodal framework was trained to estimate TUG time from raw GRF timeseries, GRF-derived summary features, and demographics in unimodal and multimodal configurations and benchmarked against conventional feature-based regressors trained on extracted features. Models were evaluated using ten-fold subject-wise cross-validation with mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (CC).

Combining raw timeseries and features (MAE 0.80 $\pm$ 0.34 s, RMSE 1.27 $\pm$ 0.85 s, CC 0.93 $\pm$ 0.07) achieved best performance, outperforming unimodal models, conventional regressors, and prior GRF-based TUG results. PD participants showed higher prediction errors (MAE 0.97 s) than controls (MAE 0.41 s), with accuracy decreasing in older age groups. Adding demographics provided no consistent benefit.

These findings suggest that steady-walking vertical GRF can approximate TUG time and could support continuous mobility assessment in PD.

Keywords: Parkinson's disease · Fall-risk assessment · Timed-Up-and-Go (TUG) · Gait analysis · GRF sensors · In-shoe wearable sensors · Perceiver architecture · Multimodal deep learning

© The Author(s) 2026
M. Särestöniemi et al. (Eds.): NCDHWS 2026, CCIS 3010, pp. 31–50, 2026.
https://doi.org/10.1007/978-3-032-28819-6_3

1 Introduction

Functional mobility is essential for independent living and safe performance of daily activities. Impaired mobility not only complicates health-related issues but also diminishes quality of life. Poor balance, gait impairment, and reduced functional mobility significantly increase the risk of falls among older adults [1] and in patients with neurological disorders, such as Parkinson's disease (PD) [2]. The risk further increases in advanced PD stages, where fall-related injuries may cause severe disabilities or even death. Epidemiological studies suggest that roughly one-third of individuals aged 65 years or older fall at least once per year [3], and fall-related mortality has risen in many countries over recent decades. The Global Burden of Disease study reported an age-standardized mortality rate of 9.2 per 100,000 in 2017 [1]. Deaths due to falls have also steadily increased across Europe since 1990 [1]. These events impose not only a personal burden but also a substantial economic cost. In the USA alone, healthcare spending attributes to non-fatal falls in adults aged 65 years and older were estimated at US$ 80 billion in 2020 [4], and average treatment costs per fall-related hospitalization can exceed US$ 18.000 [5]. Falls also impose major costs on European health systems. The cost of treating fall-related injuries in the EU is estimated to be about 25 billion Euros per year [6].

The Timed-Up-and-Go (TUG) test is one of the most widely used tools for assessing mobility and fall risk. Podsiadlo and Richardson formalized the timed version of "Get-Up-and-Go" [7] in 1991 and showed that the time taken to rise from a chair, walk 3 m, turn, walk back and sit down is a reliable and valid measure of functional mobility that correlates strongly with balance, gait speed and activities of daily living [8]. Longer completion times are associated with greater fall risk. In community-dwelling older adults, a cut-off of 13.5 s is often cited, with high sensitivity and specificity for fall prediction [9], although alternative thresholds between approximately 8 to 16 s have been reported [10]. For PD patients specifically, a threshold of 11.5 s has been associated with increased fall risk and disease severity [2]. Despite this utility, TUG assessment has limitations. It requires a supervised and task-specific protocol. The notion of a "comfortable pace" is subjective and may vary between raters and patients. TUG time exhibits test-retest variability. Clinical physiotherapy guidance in PD report minimal detectable change (MDC) values of approximately 3.5–11 s [11]. Most critically, TUG time compresses complex movement into a single scaler outcome, providing no direct insight into which specific gait components or transitions are most impaired.

Quantitative gait analysis offers a more detailed assessment of mobility by capturing spatiotemporal dynamics and biomechanical patterns relevant to fall risk. Wearable sensors such as inertial measurement units (IMUs), accelerometers, and ground reaction force (GRF) sensors have enabled continuous and non-invasive gait monitoring [12]. GRF sensors are particularly informative because they reflect how forces are distributed across the feet and thus indirectly, joint mechanics and muscle activity. Previous studies have shown that gait signals can be used to estimate PD symptoms' severity [13–15] and support fall risk prediction through spatiotemporal gait features [14]. However, many approaches rely solely on sensor data or extracted features, overlooking clinically relevant factors such as demographic characteristics (e.g., age, sex, height, weight) which may provide complementary predictive value.

In our previous work [13], we demonstrated the feasibility of predicting Unified Parkinson's Disease Rating Scale (UPDRS) scores from raw multichannel GRF time-series signals combined with expert-derived features using the Perceiver architecture from Google DeepMind [16]. While promising, both modalities originated from the same data sources (in-shoe GRF sensors), and the outcome focused on overall motor severity rather than functional mobility and fall risk. Moreover, the impact of including additional modalities beyond gait-derived signals, such as demographics, on prediction performance was not systemically examined. This highlights a key gap in the present study context, specifically the absence of a unified multimodal framework for TUG prediction that integrates GRF recordings and demographic covariates on the same dataset.

In this study, we extend our previous GRF-based work in two directions. First, we target TUG completion time as a proxy for functional mobility and fall risk, and we investigate whether it can be estimated from two-minute walking recording without administering the TUG test itself. Second, we explicitly consider demographics as an additional modality and systemically compare unimodal and multimodal configurations to quantify the impact of combining demographics with GRF-derived features and GRF timeseries signals.

The major contributions of this paper can be summarized as follows:

1. Evaluate the feasibility of estimating TUG completion time from vertical GRF-based gait signals acquired during two-minute walking tests in people with PD and healthy controls and examine how prediction performance varies across diagnostic status and age groups.
2. Compare unimodal and multimodal using raw GRF timeseries signals, GRF-derived summary features, and demographics covariates to quantify the contribution of each input modality and their combinations.
3. Benchmark a Perceiver-based multimodal architecture, using both unimodal and multimodal input configurations, against standard machine-learning (ML) regressors (KNN, SVR, Decision Tree and Random Forest) trained on the same hand-crafted feature set.

The rest of the paper is organized as follows. Section 2 discusses related work on TUG assessment, wearable-based gait analysis and multimodal modelling in PD. Section 3 describes the dataset, preprocessing and component of the proposed framework. Section 4 presents the results of the performance evaluation of the model configurations. Section 5 discusses the findings, strengths and limitations, challenges, and implications for future research. Section 6 concludes the paper.

2 Related Work

The TUG test is a standard clinical tool for assessing basic functional mobility and fall risk, particularly in older adults and people with neurological conditions. Podsiadlo and Richardson reported that TUG test is reliable (inter-rater and intra-rater) and shows a good correlation with gait speed, balance and activities of daily life in frail older adults [8]. Subsequent studies and reviews have shown its prognostic value for falls [9, 10].

A major limitation of conventional TUG is that it compresses complex movement into a single scalar outcome. This has motivated the development of instrumented TUG (iTUG) protocols using wearable sensors, which can quantify spatiotemporal gait characteristics and postural transition beyond raw completion time [17–20]. Recent reviews conclude that iTUG provides more sensitive fall-risk and mobility metrics than conventional TUG [21].

Beyond iTUG, several studies have investigated whether TUG time can be estimated from gait outside formal TUG trials. Choi et al. used IMU-based 3D accelerometers on the pelvis and both feet during outdoor walking in 37 older adults, extracting 120 step-based gait features [22]. Elastic-net regression was applied for feature reduction, followed by ridge regression to predict TUG time. Across repeated train-test splits, the model achieved small mean absolute prediction errors (MAPE) of approximately 0.865 s for the foot configuration, 0.918 s for the combined foot-pelvis configuration, and 0.921 s for the pelvis configuration. This study therefore illustrates how feature-based regression can leverage wearable gait data collected in real-world community settings to approximate conventional TUG assessments [22].

GRF and planer-pressure sensors are less commonly used for TUG prediction than IMUs, but they are important in PD gait research. Systematic reviews of wearable-based digital biomarkers [23] highlight the PhysioNet vertical GRF database [24], which uses in-shoe vertical GRF sensors in people with PD and healthy controls, and show that GRF-derived spatiotemporal and kinetic features support accurate PD diagnosis and severity assessment. Complementary to these GRF-based studies, Hao et al. proposed a sensor-agnostic ML pipeline that predicts TUG time from day-to-day, unscripted walking rather than controlled TUG tests [25]. They used 303 participants in three cohorts (healthy older adults, PD patients and individuals with mild cognitive impairment or dementia) and three wearable setups (lower-back accelerometer, wrist-accelerometer and in-sole GRF sensors). A random forest model evaluated with leave-one-subject-out validation achieved errors (MAE of 1.7 ± 1.7 s), within the minimal detectable change for most predictions, supporting the feasibility of using heterogeneous wearable signals collected in free-living conditions as a passive proxy for TUG-related mobility biomarker.

Mostafavi et al. [26] applied a Gramian Angular Field (GSF) and convolutional neural network (CNN) pipeline to the PhysioNet vertical GRF dataset [24]. Raw gait signals from eight in-shoe sensors under each foot (16 channels) were segmented into overlapping 10–60 s windows, and each window was converted into a 2D GAF image that encodes temporal correlations. These images were then used as input to CNNs for PD versus healthy control classification and for regression of continuous clinical scores (Hoehn and Yahr, UPDRS, UPDRS motor and TUG time). Predictions from multiple windows were averaged to obtain a single TUG estimate per participant. With this framework, the authors achieved 98.6% accuracy for the classification task and an R^2 of 0.80 for TUG regression with a MAE of 1.32 s and RMSE of 1.75 s (TUG range 7.27–36.34 s) [26]. This work demonstrated that transformed GRF signals combined with CNNs can support continuous TUG prediction.

Our previous work on the same dataset [24] proposed a Perceiver-based multimodal framework to estimate UPDRS scores from raw 16-channel vertical GRF signals and GRF-derived time- and frequency- domain features [13]. Using gait data from 93 PD

patients and 73 controls, the multimodal Perceiver achieved MAE of 2.23 ± 1.31, an RMSE of 5.75 ± 4.16 and correlation coefficient of 0.93 ± 0.08, for UPDRS prediction, outperforming earlier GRF-based approaches in terms of MAE and correlation [13]. This supports the idea that combining raw sensor signals with engineered features in a single architecture can improve PD severity estimation. These findings are consistent with observations that multimodal architectures can improve biomedical prediction. Acosta et al. reviewed multimodal biomedical AI and concluded that integrating heterogeneous data sources such as image, electronic health records and wearable data often yields more accurate and robust models than single-modality design [27]. Recent PD-focused multimodal frameworks, such as MultiParkNet, which fuses speech, drawing, neuroimaging and cardiovascular signals [28], and the model of Li et al., which combines structural MRI, wearable accelerometry, genetic risk scores and lifestyle neurodegenerative disease prediction [29], also report clear gains over unimodal baselines.

Taken together, these studies suggest that TUG can be approximated from wearable gait signals using both classical and deep learning methods. However, most existing work either focuses on IMUs and free-living accelerometry, or, in the case of GRF-based approaches, relies on image-like transformations and CNNs. The relative contribution of raw GRF timeseries, GRF-derived features and demographic factors within a unified multimodal architecture remains largely unexplored, and direct comparisons of unimodal versus multimodal models on a common GRF-based dataset are uncommon. In this context, we investigate how different GRF representations and demographics relate to TUG prediction performance and how multimodal fusion behaves compared with unimodal designs in PD and health controls.

3 Materials and Methods

This study proposes a multimodal machine learning framework based on the Perceiver architecture to estimate TUG scores in individuals with PD and healthy controls. The pipeline consists of data curation, signal preprocessing and feature extraction, model architecture and evaluation protocol.

3.1 Dataset Description

The publicly available Parkinson's gait dataset from PhysioNet [24] was used which includes recording from three independent research groups [30–32]. This dataset includes gait recordings from 93 patients with idiopathic PD and 73 healthy subjects.

Researchers measured gait signals using eight GRF sensors positioned under each foot. These sensors measured the force signal in Newtons at a sampling rate of 100 Hz. Recordings were acquired under two walking conditions:

- **Single-task gait:** walking at a self-selected comfortable pace.
- **Dual-task gait:** walking while performing a concurrent cognitive task (serial subtraction by sevens from a pre-defined number).

Both single-task and dual-task gait recordings were used when available: otherwise, only the available recording was included. In addition to raw gait signals, demographic

metadata were available: sex, age, height, weight, UPDRS score, and TUG completion time. Figure 1 presents the distribution of TUG scores across the cohort and their variation with age. For the present study, we included only participants with complete demographics profiles (age, sex, height and weight) and valid TUG scores. Recordings shorter than two minutes were excluded to maintain a consistent minimum signal duration. After applying these criteria, 100 participants (62 PD, 38 controls) were retained for analysis. Figure 2 illustrates typical GRF sensors outputs over a single gait cycle, highlighting stance and swing phases.

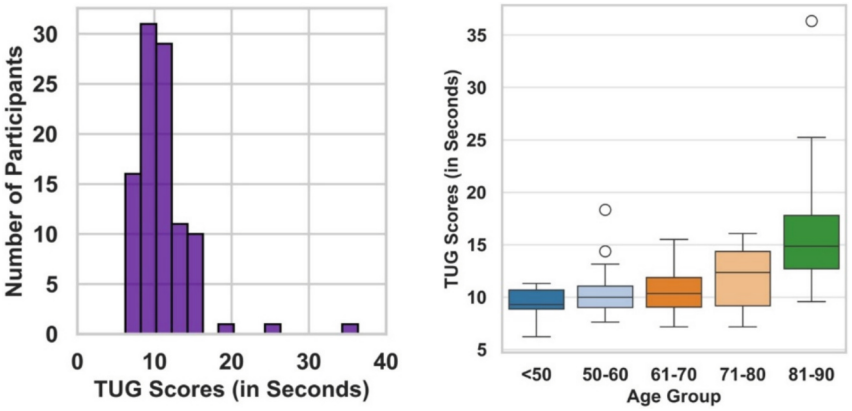


Fig. 1. (Left) Histogram of TUG scores distribution. Most participants completed the test within 5–16 s. (Right) Box plot of TUG scores distribution among participants across different age groups. The data show a gradual increase in median TUG scores with age.

3.2 Pre-Processing and Feature Extraction

To minimize the end-of-trial artifacts and to obtain fixed length for the Perceiver implementation, we extracted 10,025 samples segments from each GRF recording, corresponding to approximately 100 s of continuous walking at 100 Hz. Recordings that did not contain at least this duration of usable data were discarded. This procedure ensured that all measurements had the same number of time points and allowed the GRF signals to reshape into matrices size of 10025×16 (samples $\times$ sensors) for model input.

Signal denoising was performed using a median filter of length three to suppress spikes while preserving gait dynamics. Feature extraction was then conducted to complement raw gait signals with interpretable descriptors. Both time- and frequency-domain features were derived for each GRF sensor over the 100-s segment and are presented in Table 1. These features, validated in prior PD studies [13–15], capture the variability, and spectral properties of the GRF signals.

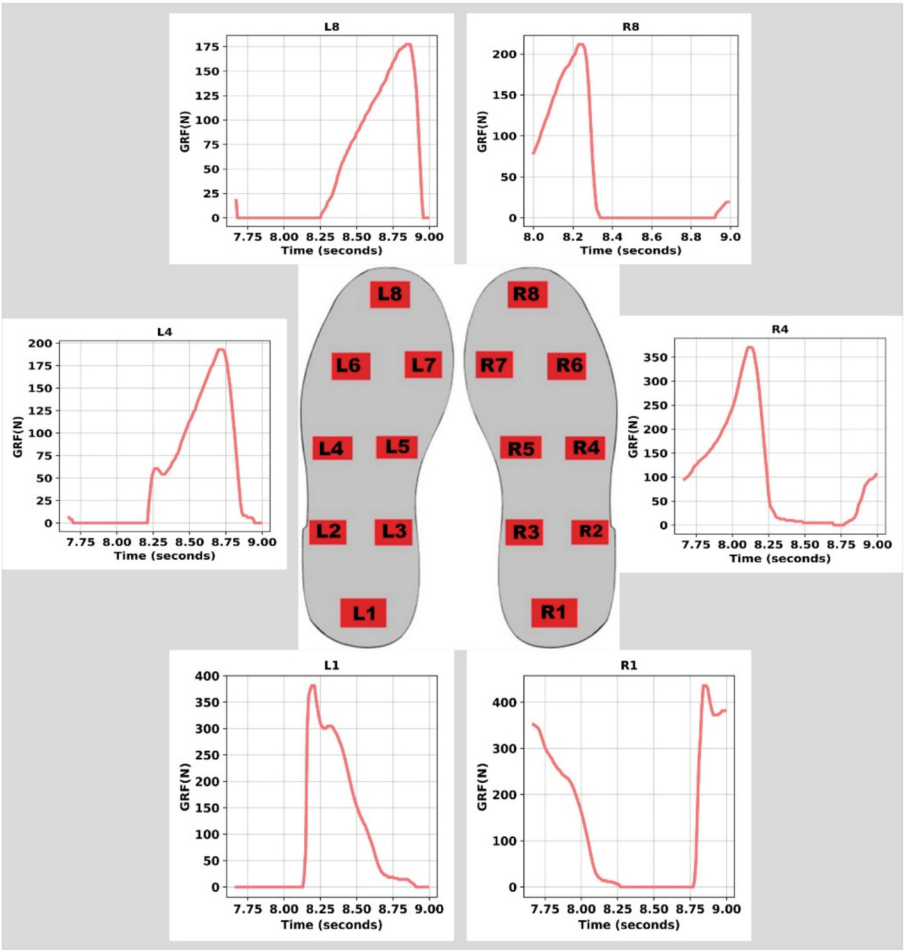


Fig. 2. This figure illustrates the force data measured by GRF sensors placed on the participant's left (L) and right (R) feet, numbered from 1 to 8 to indicate their positions on the feet. Timeseries plots depict the force exerted during a gait cycle. Signal variations from these sensors provide insights into the distinct gait characteristics, reflecting dynamic changes across the stance and swing phases of the gait cycle.

All features were z-score standardized according to Eq. 1 using the mean and standard deviation estimated from the full cohort.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

In this equation, $\boldsymbol{x}$ denotes the feature value, $\boldsymbol{\mu}$ the cohort mean and $\boldsymbol{\sigma}$ the standard deviation.

Table 1. Time and Frequency based features extracted from each sensor of all participants

Feature Domain	Features calculated from each GRF sensor
Time	mean, harmonic mean, median, range, interquartile range (IQR), mean absolute deviation, maximum amplitude and minimum/maximum spread, skewness, kurtosis, root mean square (RMS), energy, power, and entropy
Frequency	mean, minimum, maximum, normalized, energy, power, and phase

In addition to gait signals, we used four demographic variables as a separate modality: age, sex, height and weight. Sex encoded as a binary categorical variable (male = 1, female = 2). After preprocessing, demographic attributes were reshaped to for each subject were represented as a 4 x 1 vector.

In this study, ***raw timeseries*** denote the original 16-channel GRF signals, ***summary features*** refer to handcrafted features computed from the GRF segment, and ***demographics*** represent subject-level covariates.

3.3 Perceiver Architecture

The Perceiver architecture was selected for its ability to process multimodal data efficiently without introducing domain-specific biases [16]. A learnable latent array interacts with the input through cross-attention to form a fixed-size latent bottleneck, and a latent Transformer applies self-attention only over latent indices to refine this representation. Cross-attention scales approximately as $O(MN)$, where M is the number of input length and N is the number of latent vectors. Latent self-attention scales as $O(LN2)$, where L is the number of latent self-attention layers. The overall complexity is $O(MN + LN2)$, where $O(\cdot)$ denotes asymptotic (Big-O) complexity.

In Perceiver networks, each modality is represented as an array of input vectors with positional information added to preserve temporal structure for one-dimensional inputs and spatial structure for two-dimensional inputs. In multimodal settings, a learned modality-specific embedding is concatenated to each input vector prior to attention. All input vectors from a given modality shares the same embedding. This embedding identifies the modality and can be sized to align channel dimensions across modalities. This input formulation preserves a modality-agnostic nature of the architecture and supports learned fusion across heterogenous inputs.

Despite its strengths, the design has a clear trade-off. Computation and memory still include a quadratic term in the latent vectors N and using a large N or many latent self-attention layers L can become computationally expensive. The latent bottleneck may miss fine details unless inputs are re-queried. Adding additional cross-attention layers allows the latent array to re-query the inputs and can improve accuracy, but it increases the number of input-linear operations and therefore raises computing and latency (Fig. 3).

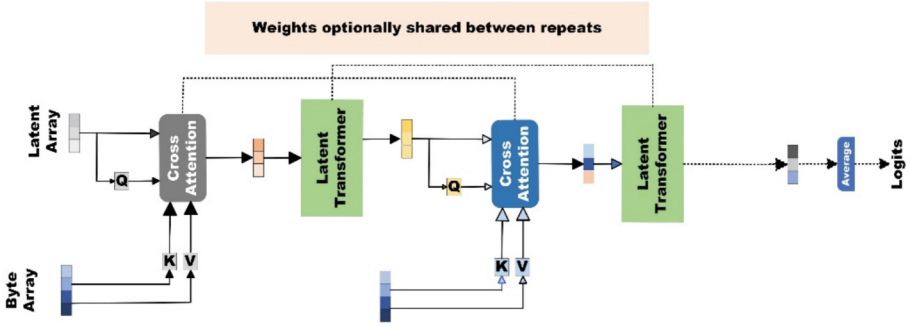


Fig. 3. The Perceiver architecture uses an iterative approach that alternates between cross-attention layers and several latent self-attention blocks. By alternating between these two types of blocks, the model can receptively attend to the input data and find patterns between the input and output layers more effectively. This mechanism allows the model to reduce the dimensionality of the input data while preserving the most critical information. The weight sharing strategy of architecture can also improve predictive performance.

3.4 Proposed Framework

The multimodal Perceiver was employed to estimate TUG scores as a continuous outcome. The open-source PyTorch implementation [33] was used in its standard form, configuring only the input modalities and a single-unit regression output. Following preprocessing, GRF timeseries were reshaped to 10025×16 (samples $\times$ sensors), extracted features to 368×1 and demographic attributes to 4×1. Each modality was flattened into a one-dimensional vector to conform to the PyTorch Multimodal Perceiver implementation [33]. Thus, the model operated on three modalities: raw GRF timeseries, GRF-derived features and demographics.

Inside a Perceiver block, positional encodings and modality identifiers are concatenated to each input vector, producing modality-aware arrays as described in Sect. 3.3. These inputs are jointly processed in a single forward pass. A learnable latent array queries the full set of input vectors via cross-attention, forming a fixed-size latent bottleneck. This latent representation is refined through several latent-only self-attention blocks. When enabled, re-entrant cross-attention allows the latent re-query the inputs, acquiring additional detail without reprocessing the entire sequence.

After refinement, latent states are averaged across latent indices to produce a global summary vector. A lightweight linear projection head maps this summary to the predicted TUG score. Because the model operates on pooled latent representations, it returns a single prediction per sample in each forward pass.

3.5 Experimental Setup

The cohort included an unequal number of PD and control participants (62 PD vs 38 controls). Although our objective was regression, this group imbalance could bias training if the TUG time distribution differed between groups. We, therefore, used PyTorch's weighted random sampler to balance PD and control representation in mini batches, so that both groups contributed more evenly to the optimization.

Hyperparameters for the Perceiver model were selected through iterative experimentation. Initial values were chosen based on prior experience and adjusted incrementally to minimize prediction errors. This process led to the selection of a network depth of 4, with 6 layers each for cross-attention and latent self-attention. Weight sharing was enabled between the cross-attention and the latent blocks.

Training was conducted with a mini-batch size of six due to GPU memory constraints. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4}, and a combined regression loss defined as the sum of MAE and MSE. Validation loss and evaluation metrics were computed on the entire held-out validation fold without rebalancing to simulate the real-world data distribution.

The framework was developed in Python using the PyTorch library and executed within a JupyterLab environment. All experiments were run on a workstation equipped with an AMD Ryzen 5 5600X 6-Core Processor, 16 GB RAM, and a 12 GB NVIDIA GeForce RTX 3060 GPU. CUDA support was utilized to accelerate training. Typical training sessions for processing either multimodal data or timeseries input configurations required approximately 50 h. Once trained, the model requires only a single forward pass and can provide predictions quickly on a standard laptop CPU.

3.6 Evaluation

Model performance was assessed by comparing predicted TUG scores against ground truth values using three matrices: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation Coefficient (CC). These metrics quantify the average deviation, squared error and linear relationship between predicted and actual scores, respectively.

$$MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \tag{2}$$

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \tag{3}$$

Here $\boldsymbol{n}$ denotes the number of samples, $\boldsymbol{p}$ the predicted value, and $\boldsymbol{a}$ the actual value.

To obtain robust estimation of generalization performance and to avoid information leakage between training and validation sets, we employed ten-fold cross validation (CV) with subject-wise splits. The 100 participants were randomly divided into ten disjoint folds at the subject level. In each iteration, nine folds ($\approx 90\%$ of subjects) were used for training and one-fold ($\approx 10\%$ of subjects) for validation. All recordings (both walking conditions) from a given subject were kept within the same fold, to prevent any information leakage during training and validations sets. Pearson CC was computed between predicted and actual TUG scores across all subjects in the validation fold for each CV run. This process was then repeated ten times, and the final performance was reported as the mean and standard deviation (SD) of MAE, RMSE, and CC across all folds.

The best-performing model was identified based on the lowest mean and SD in MAE and RMSE, and the highest mean with lowest SD in CC. Comparative analysis was

conducted between multimodal and unimodal models to determine the most effective configuration. Scatter plots of actual versus predicted TUG scores were used to visually illustrate model performance.

In addition to overall performance, subgroup analyses were performed by diagnostic group (PD vs healthy controls) and age range using subject-level out-of-fold predictions from the ten-fold subject-wise CV. For participant with multiple measurements, observed and predicted TUG scores were average within each subject before subgroup analysis. Subjects were then grouped into predefined age categories for the age-based analysis. We also examined model performance near the clinically relevant 11.5 s TUG threshold in participants with PD. Specifically, we assessed whether subjects with observed TUG values within $\pm$ 1.0 s and $\pm$ 1.5 s of this cut-off were placed on the correct side of the threshold based on predicted TUG values.

For benchmarking, conventional regressors including KNN, SVR, Decision Tree and Random Forest were implemented using scikit-learn library. Hyperparameters for these benchmark regressors were selected through iterative experimentation, starting from standard defaults and adjusting key settings incrementally to minimize prediction error. These models were trained and evaluated under the same ten-fold CV protocol as the proposed framework. Since these algorithms cannot directly handle raw timeseries signals, they were applied only to the handcrafted extracted features.

4 Results

Table 2 summarizes the accuracy across input configurations. Among unimodal models, the GRF-derived summary features model achieved low prediction error (MAE of 0.82 $\pm$ 0.42 s, RMSE of 1.38 $\pm$ 1.06 s and CC of 0.92 $\pm$ 0.07). The raw GRF timeseries model performed similarly (MAE of 0.87 $\pm$ 0.31 s, RMSE of 1.34 $\pm$ 0.80 s and CC of 0.94 $\pm$ 0.05). In contrast, the demographics-only model produced substantially higher errors and a lower correlation (MAE: 1.80 $\pm$ 0.63 s, RMSE: 2.63 $\pm$ 1.63 s, and CC: 0.55 $\pm$ 0.37), indicating that gait information captured by the GRF signals contributes most to prediction accuracy.

The Multimodal (Summary features + raw GRF timeseries signal) configuration yielded the best overall performance (MAE 0.80 $\pm$ 0.34 s, RMSE 1.27 $\pm$ 0.85 s and CC 0.93 $\pm$ 0.07). In contrast, the three-modality setting (Summary features + raw GRF timeseries + demographics) produced similar but slightly inferior results (MAE of 0.87 $\pm$ 0.32, RMSE of 1.29 $\pm$ 0.73, and CC of 0.94 $\pm$ 0.05). The modestly lower CC for the summary features and raw timeseries configuration is largely driven by a single outlier participant with a time of 36.34 s, for whom all models under-predict. Overall, these findings indicate that adding demographic data does not systematically improve prediction accuracy, whereas fusing raw GRF timeseries signal with handcrafted features offers a consistent, though modest advantage over unimodal inputs.

Figure 4 visualizes the relationship between actual and predicted TUG scores across the three main configurations (Feature, Timeseries and Multimodal data). Predictions cluster closely around the reference line ($y = x$), indicating strong agreement across all models.

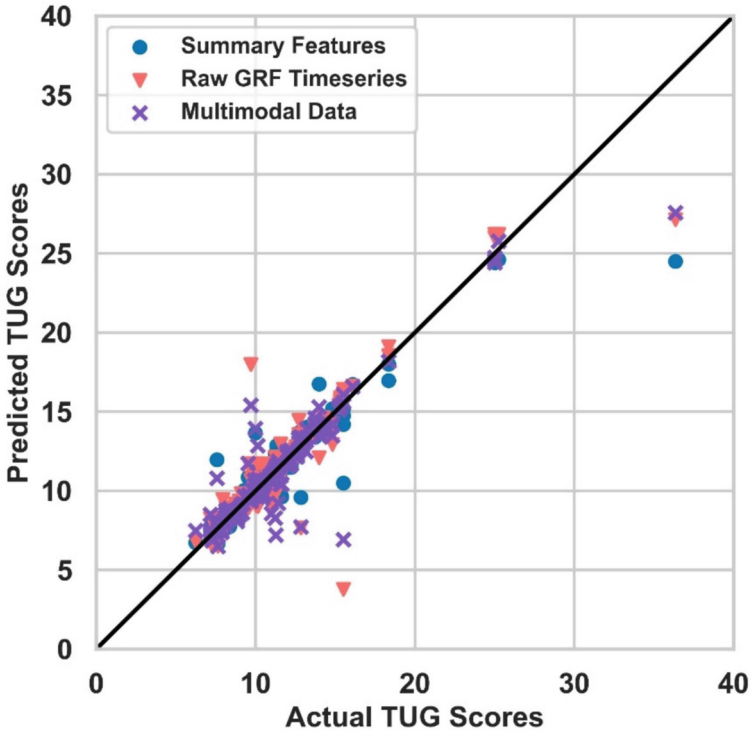


Fig. 4. Scatter plot showing the predicted versus the actual TUG scores for three main Perceiver configurations: GRF-derived summary features (blue circles), raw GRF timeseries signals (pink down-pointing triangle) and Multimodal Data (purple cross). The black dashed line represents the ideal prediction (y = x). Here "Multimodal Data" refers to the combination of GRF-derived summary features and raw GRF timeseries signals.

Table 2. Performance comparison across input modalities

Input Modality	MAE Mean ± SD	RMSE Mean ± SD	CC Mean ± SD
Demography	1.80 ± 0.63	2.63 ± 1.63	0.55 ± 0.37
Features	0.82 ± 0.42	1.38 ± 1.06	0.92 ± 0.07
Timeseries	0.87 ± 0.31	1.34 ± 0.80	0.94 ± 0.05
Multimodal Data (Features + Timeseries)	0.80 ± 0.34	1.27 ± 0.85	0.93 ± 0.07
Multimodal Data (Features + Timeseries + Demographics)	0.87 ± 0.32	1.29 ± 0.73	0.94 ± 0.05

For the best performing model (Multimodal: Summary features + Raw timeseries signal), errors were lower in healthy controls (MAE 0.41 s, RMSE 0.57 s and CC 0.93)

than in participants with PD (MAE 0.97 s, RMSE 2.15 s and CC 0.86), consistence with the broader range and higher variability of TUG times in PD (Table 3).

Table 3. Diagnostic-group performance of the best performed model Multimodal Data (GRF summary features + raw GRF timeseries). N denotes unique participants in each group

Diagnostic-group	N	MAE(s)	RMSE (s)	CC
Control	38	0.41	0.57	0.93
PD	62	0.97	2.15	0.86

Errors were lowest in the 50–70 years ranges (MAE $\approx$ 0.53–0.67 s) and increased in older participants, with the largest error in the 81–90 age group (MAE 1.88 s and RMSE 4.59 s), reflecting largest variability in gait and mobility. The $<$ 50 years bin contains only two participants and should be interpreted cautiously (Table 4).

Table 4. Age-stratified performance of the best performed model Multimodal Data (GRF summary features + raw GRF timeseries). N denotes unique participants in each group

Age bins	N	MAE (s)	RMSE (s)
$<$ 50 years	2	0.64	0.70
50–60 years	34	0.53	0.76
61–70 years	42	0.67	1.24
71–80 years	14	0.95	1.79
81–90 years	8	1.88	4.59

Using the PD-specific cutoff, 15 of 19 subjects (78.9%) within $\pm$ 1.0 s and 26 out of 32 subjects (81.3%) within $\pm$ 1.5 s of the threshold was placed on the correct side of the cutoff based on predicted TUG values.

Table 5 summarizes representative studies that predict continuous TUG time from wearable or clinical data. Direct numerical comparison is difficult because cohorts, sensing modalities and protocols differ, but several patterns emerge. Choi et al. IMU-based accelerometers on the pelvis and feet during outdoor walking in 37 older adults and obtained MAPE of approximately 0.87–0.92 s, depending on sensor configuration [22].

The sensor-agnostic random forest pipeline [25], trained on heterogenous free-living accelerometer and GRF data from 303 participants, reported a substantially larger MAE of 1.7 $\pm$ 1.7 s. On the same PhysioNet GRF dataset Mostafavi et al. [26] applied a GAF + CNN model and achieved MAE 1.32 s and RMSE of 1.75 s for TUG regression. Other heterogenous wearable or clinical-variable models typically report MAE $\geq$ 1.7 $\pm$ 1.7 s. In

this context, our Multimodal Data (Summary features + Raw timeseries) configuration attains a lower MAE (0.80 ± 0.34 s) and comparable or higher correlations (CC 0.93 ± 0.07), with an RMSE (1.27 ± 0.85 s) below the GRF-based GAF + CNN results and broadly in line with IMU-based regression. These findings suggest that vertical GRF, when modelled with multimodal Perceiver architecture, is comparable with more commonly used IMU setups for TUG predictions.

Table 5. Comparison with previous studies on TUG score estimation

Study	Modality	Method	Reported Metrics
Choi et al. [22]	IMU-based 3D accelerometers on pelvis and both feet during outdoor walking	Elastic-net feature selection + ridge regression	MAPE ≈ 0.865 s (feet), 0.918 s (feet + pelvis), 0.921 s (pelvis only) RMSE/CC not reported
Hao et al. [25]	Lower-back and wrist accelerometers plus in-shoe vertical GRF (16 sensors)	Random Forest regression	MAE: 1.7 ± 1.7 s other metrics not reported
Mostafavi et al. [26]	In-shoe vertical GRF (16 sensors)	GAF transformation + CNN	MAE: 1.32 s; RMSE: 1.75 s; R^2: 0.80
Multimodal Data	**In-shoe vertical GRF (16 sensors)**	**Perceiver-based multimodal regression**	**MAE: 0.80 ± 0.34; RMSE: 1.27 ± 0.85; CC: 0.93 ± 0.07**

The Perceiver-based framework consistently outperforms traditional machine learning models. Conventional regressors applied to extracted features showed substantially higher errors rates and weaker correlations compared to the Perceiver-based approach as shown in Table 6. For example, KNN produced an MAE of 2.01 ± 0.62 s, RMSE of 2.82 ± 1.47 s and CC of 0.51 ± 0.21, while Decision Tree performed worst overall (MAE: 2.65 ± 1.29 s, RMSE: 3.51 ± 1.22 s, and CC: 0.28 ± 0.29).

Table 6. Performance of Conventional regressors

Input Modality	Hyperparameters	MAE Mean ± SD	RMSE Mean ± SD	CC Mean ± SD
Features		0.82 ± 0.42	1.38 ± 1.06	0.92 ± 0.07
Multimodal Data		0.80 ± 0.34	1.27 ± 0.85	0.93 ± 0.07
KNN	n_neighbors = 23	2.01 ± 0.62	2.82 ± 1.47	0.51 ± 0.21
SVR	C = 10.0, gamma =" scale", epsilon = 0.1	2.04 ± 0.66	2.80 ± 1.43	0.47 ± 0.29
Decision Tree	min_samples_leaf = 3	2.65 ± 0.69	3.51 ± 1.22	0.28 ± 0.29

(*continued*)

Table 6. (*continued*)

Input Modality	Hyperparameters	MAE Mean ± SD	RMSE Mean ± SD	CC Mean ± SD
Random Forest	n_estimators = 500, min_samples_leaf = 2	1.94 ± 0.64	2.62 ± 1.29	0.51 ± 0.28

5 Discussion

This work explored whether TUG completion time can be estimated from vertical GRF signals recorded during two-minute walk in people with PD and healthy controls. Overall, the results suggest that steady-walking GRF contains substantial information related to functional mobility. Both the raw GRF timeseries signal and GRF summary features-based Perceiver models achieved high predictive accuracy, and the combined GRF configurations (Raw timeseries signals + summary features) produced the best average performance. The Perceiver-based models also outperformed conventional feature-based regressors trained on the same feature-set, which supports the value of learning directly from long raw signals and fusing them with interpretable gait descriptors within a single architecture.

Figure 4 provides a quantitative view of these results, showing that the predictions from all Perceiver configurations cluster closely around the reference line, indicating strong agreement with the measured TUG times for most participants. The largest deviations occur at the extremes of the score distribution, where the slowest participant is consistently underpredicted. Compared with the unimodal models, the Multimodal Data configuration (Raw timeseries signals and summary features) shows prediction slightly closer to the reference line, consistent with the performance metrics reported in Table 2.

The comparison between unimodal and multimodal inputs provides insight into multimodal learning in this context. The raw GRF timeseries signal and GRF summary features models explained most of the variability in TUG time, and combining these two GRF representations produced only a modest improvement but consistent reduction in MAE and RMSE. Although the improvement was modest, it may still be practically relevant, as the multimodal model provided a more reliable estimate of TUG time than either modality alone. This suggests that TUG-related mobility information can be captured from structured walking recordings. However, further work is needed to determine its utility for continuous monitoring in real-world settings. In contrast, adding age, sex, height and weight did not provide consistent benefit and slightly reduced performance. This matches the weak performance of the demographics model and suggests that, once GRF inputs are incorporated, basic demographics add little extra information. These findings should not be interpreted as evidence that demographics are unimportant for mobility in general, but rather that in this dataset they contribute little beyond what is already captured by GRF-based gait dynamics. This is consistent with multimodal learning theory, where fusion is most useful when modalities provide complementary information rather than redundant signals [34, 35].

Subgroup analyses showed higher accuracy in controls than in participants with PD. Prediction error increased error in older age groups, consistent with higher variability in TUG performance in these subgroups. In participants with PD, the model also showed reasonable performance near the clinically relevant 11.5 s cutoff, although some near-threshold cases were placed on the wrong side of the threshold. Therefore, small prediction errors near the cutoff may lead to incorrect fall-risk classification in PD, and predictions in this range should be interpreted carefully.

The Perceiver architecture is well suited for long GRF sequences and for combining multiple modalities in a single model. However, its fusion behavior can influence the contribution of low-dimensional inputs when a high-dimensional modality is much more informative. The model compresses the inputs into a fixed-size bottleneck using cross-attention and refines this representation with latent self-attention blocks [16]. When modalities differ strongly in dimensionality and predictive strength, training can prioritize the modality that reduces the loss fastest. This can lead to modality dominance (optimization imbalance), where gradients are driven mainly by the informative modality unless specific architectural or training strategies are used [36]. In our setting, GRF inputs likely dominate learning, so demographics may be down weighted and contribute little to the final prediction.

Computational efficiency should also be considered when interpreting these results, Although the dataset size was moderate, the multimodal and timeseries-based models included long GRF sequences, which substantially increased computational cost. Therefore, training required approximately 50 h.

Model behavior at the extremes of the TUG distribution also reflects the dataset composition. The cohort mainly includes people with mild-to-moderate PD and healthy controls, and very long TUG times are rare. As a result, severe cases are under-represented during training, prediction errors tend to increase for these rare targets. This corresponds to imbalanced regression, where extreme target values are scarce and models trained with standard losses become calibrated to the dense middle of the distribution [37]. In addition, the analyzed recordings capture level-ground walking and do not include key TUG components such as sit-to-stand, turning, or stand-to-sit phases. The model should therefore be viewed as learning a GRF-based proxy for TUG as a global mobility marker rather than a full surrogate for all TUG phases. Although the model achieved sub-second MAE which is below reported MDC ranges in PD [11], this indicates potential feasibility for longitudinal monitoring. However, the clinical impact of prediction error depends on where the true TUG value lies, as errors near commonly used cut-offs may change risk categorization. For this reason, predictions for individuals with very poor mobility and predictions near the decision threshold should be interpreted with caution.

Several methodological points also affect how these results should be interpreted. MAE and RMSE were selected as primary evaluation metrics because they express prediction error directly in seconds, making them more interpretable for TUG estimation. In contrast, explained-variance metrics such as R^2 are dimensionless and do not reflect the absolute magnitude of error on the original scale. Both the Perceiver models and the conventional regressors were evaluated without systematic hyperparameters optimization. Therefore, the observed performance differences should be interpreted with these methodological considerations in mind. Although both single-task and dual-task

gait recordings were included, they were not analyzed separately because both recording conditions were not available for all subjects. Therefore, the reported performance reflects overall model accuracy across available recordings and does not permit a direct comparison between single-task and dual-task conditions. Finally, because the PhysioNet dataset aggregates recordings from multiple studies, cross-cohort validation (e.g., train on one or two cohorts and test on the third) would provide a stronger test of generalizability.

Overall, these findings support steady-walking vertical GRF as a practical tool for functional mobility in this dataset, while highlighting the need for stronger validation and robustness testing before clinical use.

6 Conclusion

This study evaluated a Perceiver-based multimodal framework to estimate TUG time from in-shoe vertical GRF recordings collected during straight, steady walking in people with PD and healthy controls from a publicly available dataset. Models trained on raw GRF timeseries and GRF-derived features achieved sub-second error, and combining both representations within a single architecture outperformed conventional feature-based regressors and prior studies. In this cohort, basic demographic variables showed limited predictive value and did not consistently improve performance when added to GRF inputs. Subgroup analyses indicated higher accuracy in controls than in PD, and errors increased with age.

Future work should prioritize external validation on independent cohorts with broader disease stages to assess generalizations and clinical robustness. Performance should also be reported stratified by TUG severity bands, age, and PD or control status to identify whether accuracy degrades for slow walkers or differs across clinical groups. In addition, model performance should be evaluated separately for single-task and dual-task gait recordings to determine whether cognitive-motor interference influences TUG prediction in PD. A more systematic hyperparameter tuning strategy should be adopted to improve the rigor and reproducibility of the experiments. If non-gait modalities are intended to contribute reliably, modality fusion should be designed explicitly (e.g., modality-specific encoders, conditional fusion, or training strategies that mitigate modality dominance). Finally, longitudinal data and uncertainty-aware prediction could support patient-specific monitoring by indicating low-confidence estimates, particularly for severe cases that are under-represented during training. In addition to TUG estimation, future work should evaluate the model for fall prediction and its potential use as a biomarker for rehabilitation and treatment outcomes.

References

1. Haagsma, J.A., Olij, B.F., Majdan, M., et al.: Falls in older aged adults in 22 European countries: incidence, mortality and burden of disease from 1990 to 2017. Inj. Prev. **26**, i67–i74 (2020). https://doi.org/10.1136/injuryprev-2019-043347
2. Nocera, J.R., Stegemöller, E.L., Malaty, I.A., et al.: Using the timed up & go test in a clinical setting to predict falling in Parkinson's disease. Arch. Phys. Med. Rehabil. **94**, 1300–1305 (2013). https://doi.org/10.1016/j.apmr.2013.02.020

3. World Health Organization: Step Safely: Strategies for Preventing and Managing Falls Across the Life-Course. https://www.who.int/publications/i/item/978924002191-4. Accessed 07 Feb 2024
4. Haddad, Y.K., Miller, G.F., Kakara, R., et al.: Healthcare spending for non-fatal falls among older adults, USA. Inj. Prev. **30**, 272–276 (2024). https://doi.org/10.1136/ip-2023-045023
5. National Council of Aging, Get the Facts on Falls Prevention, https://www.ncoa.org/article/get-the-facts-on-falls-prevention/. Accessed 10 Dec 2025
6. Almada, M., Brochado, P., Portela, D., et al.: Prevalence of fall and associated factors among community-dwelling European older adults: a cross-sectional study. J. Frailty Aging, **10**, 10–16 (2021). https://doi.org/10.14283/jfa.2020.44
7. Mathias, S., Nayak, U.S., Isaacs, B.: Balance in elderly patients: the "get-up and go" test. Arch. Phys. Med. Rehabil. **67**, 387–389 (1986)
8. Podsiadlo, D., Richardson, S.: The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J. Am. Geriatr. Soc. **39**, 142–148 (1991). https://doi.org/10.1111/j.1532-5415.1991.tb01616.x
9. Shumway-Cook, A., Brauer, S., Woollacott, M.: Predicting the probability for falls in community-dwelling older adults using the timed up & go test. Phys. Ther. **80**, 896–903 (2000). https://doi.org/10.1093/ptj/80.9.896
10. Barry, E., Galvin, R., Keogh, C., et al.: 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. **14**, 14 (2014). https://doi.org/10.1186/1471-2318-14-14
11. Keus, S.H.J., Munneke, M., Graziano, M., et al.: European Physiotherapy Guideline for Parkinson's Disease. KNGF/ParkinsonNet, The Netherlands (2014)
12. Kristoffersson, A., Lindén, M.: A systematic review of wearable sensors for monitoring physical activity. Sensors **22**, 573 (2022). https://doi.org/10.3390/s22020573
13. Faiem, N., Asuroglu, T., Acici, K., et al.: Assessment of Parkinson's disease severity using gait data: a deep learning-based multimodal approach. In: Särestöniemi, M., et al (eds.) Digital Health and Wireless Solutions, pp. 29–48. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-59091-7_3
14. Aşuroğlu, T., Oğul, H.: A deep learning approach for parkinson's disease severity assessment. Health Technol. **12**, 943–953 (2022). https://doi.org/10.1007/s12553-022-00698-z
15. Aşuroğlu, T., Açıcı, K., Berke Erdaş, Ç., et al.: Parkinson's disease monitoring from gait analysis via foot-worn sensors. Biocybern. Biomed. Eng. **38**, 760–772 (2018). https://doi.org/10.1016/j.bbe.2018.06.002
16. Jaegle, A., Gimeno, F., Brock, A., et al.: Perceiver: General Perception with Iterative Attention, http://arxiv.org/abs/2103.03206 (2021)
17. Zampieri, C., Salarian, A., Carlson-Kuhta, P., et al.: The instrumented timed up and go test: potential outcome measure for disease modifying therapies in Parkinson's disease. J. Neurol. Neurosurg. Psychiatry **81**, 171–176 (2010). https://doi.org/10.1136/jnnp.2009.173740
18. Greene, B.R., O'Donovan, A., Romero-Ortuno, R., et al.: Quantitative falls risk assessment using the timed up and go test. IEEE Trans. Biomed. Eng. **57**(12), 2918–2926 (2010). https://doi.org/10.1109/TBME.2010.2083659
19. Palmerini, L., Mellone, S., Rocchi, L., et al.: Dimensionality reduction for the quantitative evaluation of a smartphone-based Timed Up and Go test. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 7179–7182 (2011). https://doi.org/10.1109/IEMBS.2011.6091814
20. Vervoort, D., Vuillerme, N., Kosse, N., et al.: Multivariate analyses and classification of inertial sensor data to identify aging effects on the timed-up-and-go test. PLoS ONE **11**, e0155984 (2016). https://doi.org/10.1371/journal.pone.0155984

21. Ortega-Bastidas, P., Gómez, B., Aqueveque, P., et al.: Instrumented timed up and go test (iTUG)—more than assessing time to predict falls: a systematic review. Sensors. **23**, 3426 (2023). https://doi.org/10.3390/s23073426
22. Choi, J., Parker, S.M., Knarr, B.A., et al.: Wearable sensor-based prediction model of timed up and go test in older adults. Sensors. **21**, 6831 (2021). https://doi.org/10.3390/s21206831
23. Polvorinos-Fernández, C., Sigcha, L., Borzì, L., et al.: Evaluating motor symptoms in parkinson's disease through wearable sensors: a systematic review of digital biomarkers. Appl. Sci. **14**, 10189 (2024). https://doi.org/10.3390/app142210189
24. Hausdorff, J.M.: Gait in Parkinson's Disease. https://physionet.org/content/gaitpdb/. (2008). https://doi.org/10.13026/C24H3N
25. Hao, T., Yamada, Y., Rogers, J.L., et al.: An automated digital biomarker of mobility. In: 2023 IEEE International Conference on Digital Health (ICDH), pp. 100–108. IEEE, Chicago, IL, USA (2023). https://doi.org/10.1109/ICDH60066.2023.00022
26. Mostafavi, S.M., Soumma, S.B., Peterson, D., et al.: Detection and severity assessment of parkinson's disease through analyzing wearable sensor data using Gramian angular fields and deep convolutional neural networks. Sensors. **25**, 3421 (2025). https://doi.org/10.3390/s25113421
27. Acosta, J.N., Falcone, G.J., Rajpurkar, P., et al.: Multimodal biomedical AI. Nat. Med. **28**, 1773–1784 (2022). https://doi.org/10.1038/s41591-022-01981-2
28. Sar, A., Puri, P.S., Naz, H., et al.: Multi-modal deep learning framework for early detection of Parkinson's disease using neurological and physiological data for high-fidelity diagnosis. Sci. Rep. **15**, 34835 (2025). https://doi.org/10.1038/s41598-025-21407-6
29. Li, A., Lian, J., Vardhanabhuti, V.: Multi-modal machine learning approach for early detection of neurodegenerative diseases leveraging brain MRI and wearable sensor data. PLOS Digit Health. **4**, e0000795 (2025). https://doi.org/10.1371/journal.pdig.0000795
30. Yogev, G., Giladi, N., Peretz, C., et al.: Dual tasking, gait rhythmicity, and Parkinson's disease: which aspects of gait are attention demanding? Eur. J. Neurosci. **22**, 1248–1256 (2005). https://doi.org/10.1111/j.1460-9568.2005.04298.x
31. Hausdorff, J.M., Lowenthal, J., Herman, T., et al.: Rhythmic auditory stimulation modulates gait variability in Parkinson's disease. Eur. J. Neurosci. **26**, 2369–2375 (2007). https://doi.org/10.1111/j.1460-9568.2007.05810.x
32. Frenkel-Toledo, S., Giladi, N., Peretz, C., et al.: Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson's disease. Mov. Disord. **20**, 1109–1114 (2005). https://doi.org/10.1002/mds.20507
33. Campagne, F.: Adapting Perceiver for learning with text modalities (2021). https://github.com/fac2003/perceiver-multi-modality-pytorch
34. Baltrusaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607
35. Li, S., Tang, H.: Multimodal Alignment and Fusion: A Survey, https://arxiv.org/abs/2411.17040 (2024). https://doi.org/10.48550/ARXIV.2411.17040
36. Peng, X., Wei, Y., Deng, A., et al.: Balanced Multimodal Learning via On-the-fly Gradient Modulation. https://arxiv.org/abs/2203.15332 (2022). https://doi.org/10.48550/ARXIV.2203.15332
37. Ribeiro, R.P., Moniz, N.: Imbalanced regression and extreme value prediction. Mach. Learn. **109**, 1803–1835 (2020). https://doi.org/10.1007/s10994-020-05900-9

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

