

Smartphone-Based Sit-to-Stand Power Assessment for Frailty Risk Screening

Ozair Ismail (06966695)

Abstract

Frailty affects ~10% of adults over 65 and is the strongest predictor of falls and loss of independence. The 30-second chair stand test is clinically validated but only captures repetition count, missing movement quality indicators that better predict adverse outcomes. We develop a two-stage system: (1) an ML-based sit-to-stand event detector trained on smartphone inertial data, and (2) a rule-based quality assessment layer extracting six clinically meaningful features mapped to Fried frailty dimensions. We compare a threshold baseline, Logistic Regression, and Random Forest using 30-fold LOSO-CV on UCI HAPT. Random Forest achieves the best internal performance (event $F1 = 0.755$, rep count $MAE = 0.60$), but completely fails on external validation with elderly subjects from SisFall (0% recall), while Logistic Regression generalizes well (92.6% recall). Gyroscope features contribute 54% of feature importance. The quality assessment layer demonstrates extraction of per-rep power, acceleration, angular velocity, and fatigue indicators that capture pre-frailty risk invisible to repetition counting alone.

1. Introduction

Frailty is a clinical syndrome of reduced physiological reserve that dramatically increases vulnerability to adverse health outcomes. The 30-second chair stand test (30s CST) is widely used for lower-limb functional assessment, but standard administration only counts repetitions with a stopwatch. Research shows that sensor-derived kinematic parameters differentiate frailty levels even when repetition counts are identical [1]. This motivates automating and enriching the test using ubiquitous smartphone sensors.

The input to our system is raw triaxial accelerometer and gyroscope data (6 channels at 50Hz) from a waist-mounted smartphone. We use a Random Forest, Logistic Regression, and a threshold baseline to output a binary prediction per 2.56-second window: sit-to-stand (1) or other activity (0). Consecutive positive windows are clustered into discrete rep events. A rule-based quality assessment layer then extracts six per-rep features — peak dynamic acceleration, relative muscle power, time-per-rep, peak gyroscope magnitude, coefficient of variation, and fatigue slope — each mapped to a Fried frailty dimension and compared against published reference ranges. Output is educational/screening, not diagnostic.

2. Related Work

Clinical IMU studies. Millor et al. [1] used a single lumbar IMU during the 30s CST and found that kinematic parameters (particularly angular velocity peaks) differentiated three frailty levels where rep count could not ($p < 0.001$). Van Lummel et al. [2] demonstrated that instrumented STS durations were more strongly associated with health status than manually recorded durations. These validate sensor-derived features but use dedicated research-grade hardware.

Smartphone-based approaches. Galán-Mercant & Cuesta-Vargas [3] used an iPhone 4 at the waist and found frail elderly produced peak vertical acceleration of $\sim 2.7 \text{ m/s}^2$ vs. $\sim 8.5 \text{ m/s}^2$ non-frail — the only smartphone study with frailty-specific thresholds (dataset not public). Sher et al. [4] achieved 99% cycle detection using rule-based signal processing on 660 cycles.

Power-based assessment. Alcázar et al. [5] validated a relative STS power equation with age.sex normative cut-off points. Park et al. [6] identified optimal sensor-derived features for frailty phenotypes using 5 wearable sensors (AUC 0.86). Our work approximates Park et al.'s multi-sensor framework with a single smartphone.

3. Dataset

UCI HAPT [7] (training): Raw accelerometer and gyroscope from 30 participants (ages 19–48), Samsung Galaxy S II at waist, 50Hz. 12 activity classes, framed as binary: sit-to-stand (ID 8) vs. everything else. 62 sit-to-stand segments averaging 2.59s. Signals windowed into 2.56s frames (128 samples) with 50% overlap. Windows labeled by majority vote (50% purity threshold — lowered from 80% after analysis showed 16% of sit-to-stand segments too short for the stricter threshold). Eight channels (3 accel axes, 3 gyro axes, accel magnitude, gyro magnitude) \times 6 statistics = **48 features**. Final: 17,453 windows, 126 positive (0.72%).

SisFall [8] (external validation): 15 elderly subjects (ages 60–75), 149 sit-to-stand trials (D07 slow, D08 fast). Resampled 200Hz \rightarrow 50Hz via resample_poly with anti-aliasing. Same feature pipeline.

4. Methods

We compare three approaches: (1) **Threshold Baseline** — predicts sit-to-stand if $\text{accel_mag_max} > t_1$ AND $\text{accel_mag_range} > t_2$ (grid-searched, 2 features, no learning). (2) **Logistic Regression** — $P(y=1|x) = \sigma(w \cdot x + b)$, 48 features, `class_weight='balanced'`, StandardScaler. (3) **Random Forest** — 100 trees, bootstrap sampling, random feature subsets, `class_weight='balanced'`. Evaluated via 30-fold LOSO-CV; metrics: precision, recall, F1, PR-AUC for sit-to-stand. Event-level: cluster consecutive positives, match to ground truth ($\pm 1.0\text{s}$ tolerance), report rep count MAE.

Quality Assessment (rule-based, not ML): For each detected rep, extract features from raw signal within rep boundaries after gravity removal (4th-order Butterworth high-pass, 0.3Hz). Six indicators: (1) peak dynamic acceleration magnitude (weakness), (2) relative muscle power via adapted Alcázar equation: $\text{Power} = [0.9 \times 9.81 \times (h \times 0.5 - c)] / (t \times 0.5)$ W/kg (weakness), (3) time-per-rep (slowness), (4) peak gyroscope magnitude (slowness), (5) CV across reps (exhaustion), (6) fatigue slope via linear regression on per-rep metrics (exhaustion). Each maps to a Fried frailty dimension with published reference comparisons.

5. Experiments, Results, and Discussion

5.1 Experiment 1: Internal Validation (LOSO-CV)

Table 1: Window-level results (mean \pm std, 30 folds)

Model	Precision	Recall	F1	PR-AUC
Threshold Baseline	0.011 ± 0.002	0.933 ± 0.135	0.022 ± 0.004	—
Logistic Regression	0.186 ± 0.070	0.933 ± 0.117	0.305 ± 0.096	0.675 ± 0.148
Random Forest	0.778 ± 0.377	0.454 ± 0.283	0.554 ± 0.301	0.742 ± 0.212

RF achieves the best F1 and PR-AUC. The pooled confusion matrix (Figure 1) shows 58 TP, only 5 FP (all SITTING windows at the transition boundary), 68 FN, and 17,322 TN.

Post-processing interaction: Standard filtering (smoothing, min 2-window duration) destroyed RF's sparse correct predictions. With minimal post-processing (single positive window = event):

Table 2: Event-level results (minimal post-processing)

Model	Rep MAE	Event Prec	Event Rec	Event F1
Threshold Baseline	14.73	0.101	0.823	0.180
Logistic Regression	3.23	0.151	0.387	0.217
Random Forest	0.60	0.909	0.645	0.755

RF achieves MAE 0.60 reps (20/30 exact), 91% event precision. Feature importance (Figure 2): gx_energy (gyro x-axis) ranks #1; gyroscope features hold 5 of top 10 positions (54% of total Gini importance), aligning with Millor et al.'s finding that angular velocity peaks are the strongest frailty differentiator [1].

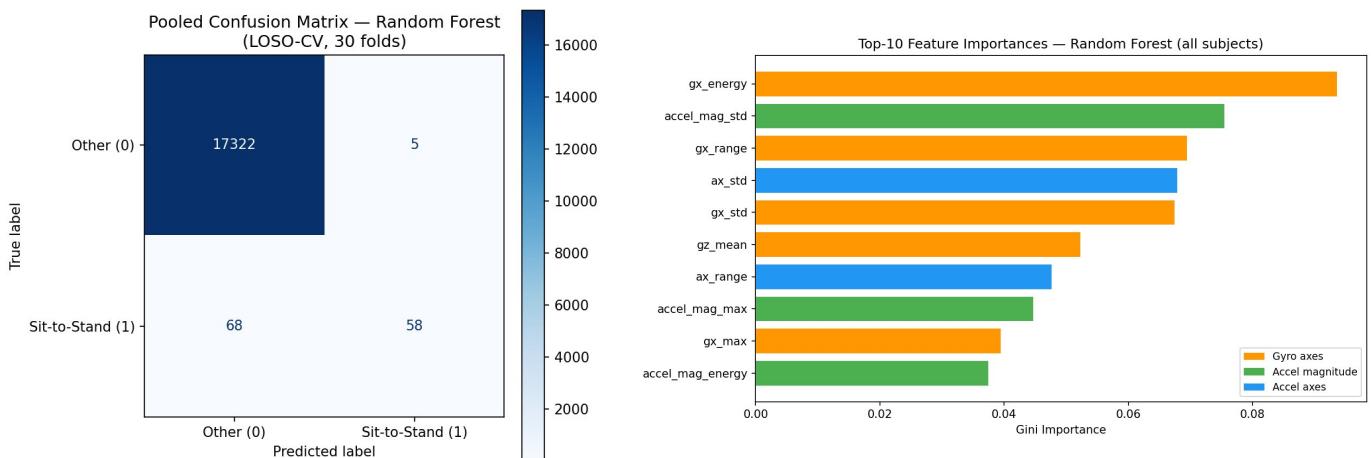


Figure 1: Pooled confusion matrix (RF, LOSO-CV). Only 5 FPs, all sitting windows.

Figure 2: Top-10 feature importances (RF). Gyroscope features dominate (orange).

5.2 Experiment 2: External Validation (SisFall Elderly)

Table 3: External validation — event recall on 149 elderly trials

Model	Features	Overall	D07 (slow)	D08 (fast)
Threshold Baseline	48	0.960	0.919	1.000
Logistic Regression	48	0.926	0.892	0.960
Log. Reg. (accel-only)	24	0.148	0.095	0.200
Random Forest	48	0.000	0.000	0.000

RF completely fails — zero detections across 1,192 windows. Its nonlinear boundaries overfit to UCI HAPT-specific patterns. LR generalizes well (92.6%), capturing more universal biomechanical signatures. This is the bias-variance tradeoff in practice: the strongest internal model generalizes worst. D08 (fast) detected more reliably than D07 (slow), confirming faster movements produce more salient signals.

5.3 Experiment 3: Feature Ablation

Removing gyroscope cuts RF event F1 from 0.755 to 0.341 ($\Delta 0.41$). LR accel-only external recall drops from 0.926 to 0.148 ($\Delta 0.78$). Gyroscope captures rotational dynamics critical for cross-population generalization.

5.4 Quality Assessment Layer

All six indicators produce physiologically plausible values across 62 UCI HAPT reps: peak dynamic acceleration 2.07–9.06 m/s^2 (mean 4.54), time-per-rep 1.48–3.66s, peak gyro 1.01–4.92 rad/s, power 1.88–4.65 W/kg. Correlation between time and acceleration: $r = -0.40$ ($p = 0.029$), confirming expected biomechanics. Against published thresholds: all subjects intermediate for acceleration (between Galán-Mercant’s frail 2.7 and non-frail 8.5 m/s^2), 7/30 flagged for exhaustion ($\text{CV} > 0.30$). Subject 17’s acceleration (2.80 m/s^2) is near the frail threshold despite being a young adult, illustrating individual variation invisible to rep count.

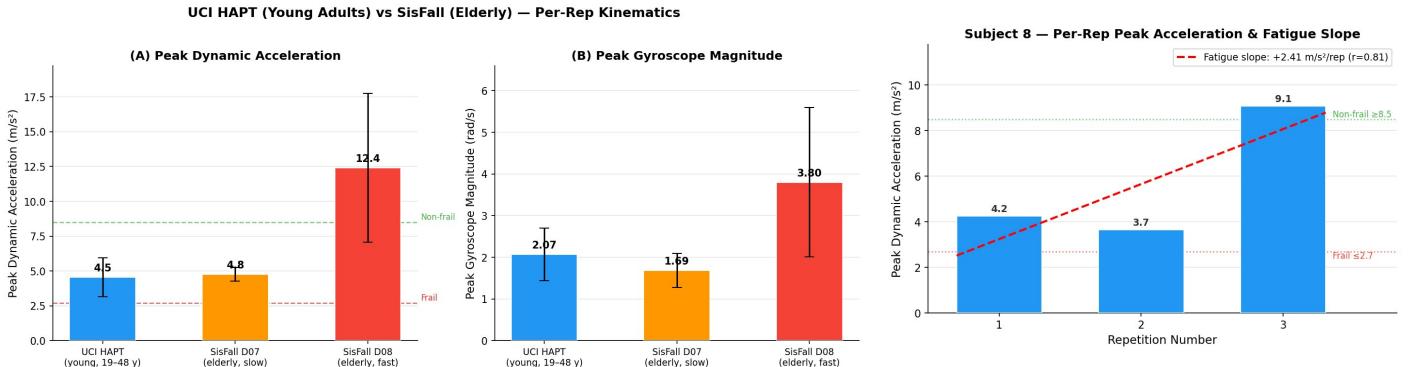


Figure 3: Young (UCI HAPT) vs elderly (SisFall) per-rep kinematics. D07 slow \approx young natural pace; D08 fast exceeds both.

Figure 4: Per-rep fatigue slope example (Subject 8). Dashed lines show frail/non-frail thresholds.

Cross-dataset comparison (Figure 3) revealed that movement speed matters more than age for peak metrics: elderly D07 slow (4.76 m/s^2) \approx UCI HAPT young (4.54 m/s^2), while elderly D08 fast (12.41 m/s^2) far exceeds both. Fatigue and variability indicators tracking performance across reps (Figure 4) are likely more sensitive to age-related differences than single-rep peaks.

Limitations: UCI HAPT has only 2–3 reps/subject (not a 30s CST), so CV and fatigue slope are illustrative. Published thresholds derive from different protocols, sensors, and populations. Acceleration magnitude is a proxy for vertical acceleration. Alcázar equation adapted from 5-rep form. No direct Fried phenotype validation.

6. Conclusion and Future Work

We demonstrated a two-stage smartphone-based system for sit-to-stand assessment. RF achieved the best internal event detection ($\text{F1} = 0.755$, $\text{MAE} = 0.60$) but completely failed on external elderly data, while LR generalized well (92.6% recall). Gyroscope features are critical (54% importance, $+0.41$ event F1). The quality assessment layer extracts six clinically meaningful indicators capturing pre-frailty risk invisible to standard rep counting. Future work: (1) train on a 30s CST dataset with elderly participants and Fried labels, (2) explore domain adaptation for RF cross-population generalization, (3) extend to

continuous daily monitoring, (4) clinical validation against clinician-administered assessments.

Contributions

Solo project. All work performed by Ozair Ismail.

References

- [1] Millor, N., et al. (2013). An evaluation of the 30-s chair stand test in older adults. *J. NeuroEng. Rehab.*, 10, 86.
- [2] Van Lummel, R. C., et al. (2016). The instrumented sit-to-stand test has greater clinical relevance. *PLoS ONE*, 11(7), e0157968.
- [3] Galán-Mercant, A. & Cuesta-Vargas, A. I. (2014). Mobile Romberg test assessment. *BMC Research Notes*, 7, 640.
- [4] Sher, T., et al. (2025). Waist-mounted smartphone 30s CST cycle detection (rule-based, 99% on 660 cycles).
- [5] Alcázar, J., et al. (2021). Relative sit-to-stand power: aging trajectories and normative data. *J. Cachexia, Sarcopenia and Muscle*, 12(4), 1013–1028.
- [6] Park, C., et al. (2021). Optimal sensor-based frailty phenotype assessment. *IEEE J. Biomed. Health Inform.*, 25(8), 3057–3067.
- [7] Reyes-Ortiz, J. L., et al. (2015). Transition-aware human activity recognition. *Neurocomputing*, 171, 754–767.
- [8] Sucerquia, A., et al. (2017). SisFall: A fall and movement dataset. *Sensors*, 17(1), 198.