BMEG 400K Grand Challenge Final Report

Ellie McGregor

1. How does your system make a decision regarding falls, near-falls, and activities of daily living?

The system I developed for the Grand Challenge classifies falls, near-falls, and activities of daily living (ADLs) using six wearable sensors: a three-lead electrocardiogram (ECG), a galvanic skin sensor (GSS), and three inertial measurement units (IMUs) on the back and thighs. This system uses a two-stage Random Forest classification and weighted scoring pipeline. First, it determines whether a signal window reflects an ADL or a potential fall. Then, if it's a non-ADL event, it distinguishes between a fall and a near-fall. I designed this method to make separating ADLs from fall-like motions easier and more reliable, enhancing both performance and interpretability.

Before classification, I filtered each signal based on physiological properties. ECG signals underwent a 0.5 to 20 Hz Butterworth bandpass filter to eliminate drift and muscle noise, followed by a 60 Hz notch filter for powerline interference. GSS data used a 10 Hz low-pass filter due to minimal high-frequency activity. IMU signals were bandpass filtered from 0.3 to 50 Hz to retain relevant motion while removing slow drift and high-frequency jitter. All were 4th-order Butterworth filters for their smooth frequency response, guided by standard practice [1-3], class assignments (notably Assignments 2 and 3), and exploratory signal plotting.

After filtering, all data were divided into 1.5-second segments with a 0.5-second overlap. The analysis of labeled segment durations (see Appendix 2) indicated that this setup effectively captured several complete fall or near-fall events without dividing them across segments. Additionally, it permitted the collection of segments featuring transitions and mid-event measurements for longer episodes, which is vital for real-time analysis. The overlap improves the clarity of event boundaries and helps ensure that brief transitions are not overlooked. I assigned labels based on a rule: if 50% or more overlapped with a labeled fall, it was marked as "F"; if it was near-fall, labeled "NF"; otherwise, it received "ADL."

I converted each window into a structured set of features to train the model and make predictions. I used a mix of classical statistical descriptors and sensor-specific metrics. Classical features included time-domain stats like mean, standard deviation, min, max, and kurtosis and frequency features such as peak frequencies, power ratios, centroid frequency, and total and band-specific power. These features were taken from Assignments 2 to 4 and served as a consistent foundation across sensors.

I incorporated sensor-specific features based on what each modality reveals during an event. For ECG, I included heart rate variability, the HF/LF ratio, instantaneous heart rate, and area under the curve. These features captured stress and cardiac response, showing significant shifts during falls, particularly with heart rate spikes and HRV drops [4]. For GSS, I extracted the slope of the signal, total range, and time to peak, reflecting sympathetic nervous system activation, as literature shows time to peak is shorter during falls [5]. IMU features included signal magnitude area (SMA), jerk, tilt angle, and axis correlations, reflecting full-body coordination, posture changes, and movement intensity [6]. Activity and trial were excluded as features or factors since the device wouldn't know what a person is doing in a real-world scenario.

A key design decision was to merge data from all three IMUs into a single dataset during training instead of treating each sensor independently. This enabled the model to learn broader patterns across different body positions and utilize useful features even when one sensor was noisy or degraded. Since IMU placement introduces variability in real-world applications, combining them during training provides a more flexible foundation for future work. I also chose not to normalize GSS features. I observed significant

variation in baseline GSS values across participants (Figure 1), and normalization wouldn't effectively address this variability.

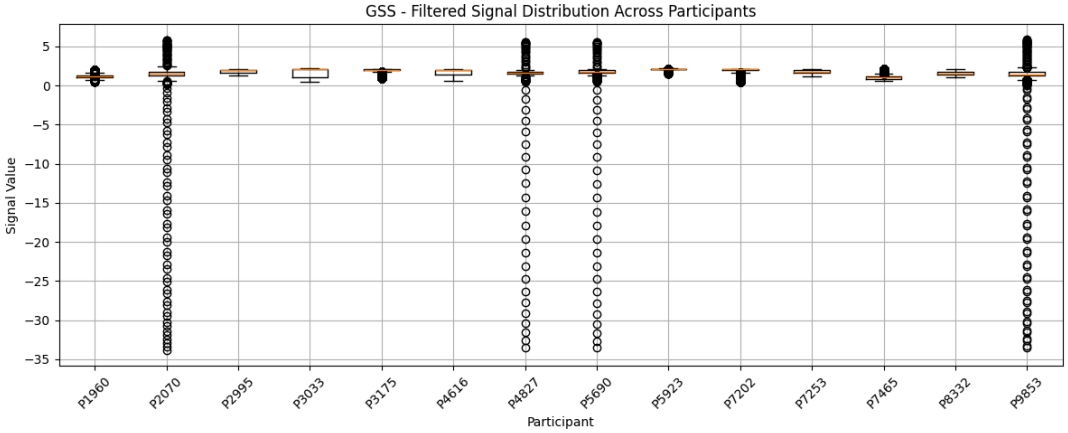


Figure 1. Filtered GSS signal distribution across participants. The boxplots highlight large baseline variability in GSS values, emphasizing why normalization was not applied.

I used Random Forests as the classification algorithm for each sensor-specific model because they perform well with tabular data, can manage a large number of features without overfitting, and do not assume a specific data distribution [7]. They are robust to noise, quick to train, and relatively easy to interpret. The system operates in two stages. In Stage 1, separate Random Forests for ECG, GSS, and IMU each predict whether a window is ADL or non-ADL. Their output probabilities are combined using a weighted average to make a final decision between ADL and non-ADL. In Stage 2, new models are trained for fall versus near-fall classification, again for each sensor, and their predictions are combined in the same manner. For both stages, I use weights of 70% IMU, 20% ECG, and 10% GSS. These were chosen based on validation with participant P3033, where IMUs were the most consistent, and ECG and GSS contributed to refining edge cases (see Q3). Overall, this structure is robust to IMU placement variability, participant differences, and enhances specificity by focusing each stage on a narrower classification task.

2. How did you train your system to detect falls, near-falls, and ADLs?

The model described in Question 1 was trained using a two-stage Random Forest architecture, with separate models for each sensor (ECG, GSS, and IMU) and classification stage (ADL vs non-ADL, then fall vs near-fall). For training, I excluded participant P3033 and used them solely for validation. This decision was based on signal distributions explored during preprocessing, where P3033 had relatively average values across sensors, making them a strong candidate to assess generalization (Appendix 4). The remaining participants provided a diverse training set in terms of movement styles and physiological responses.

Class imbalance was a key concern. ADL windows heavily outweighed fall and near-fall events (Appendix 3). To mitigate this, I used SMOTE to oversample the minority classes during training [8]. In addition, I randomly downsampled ADL windows in some cases to further balance the data. These steps helped improve model sensitivity and reduce bias toward the dominant class, especially in Stage 1.

Each Random Forest model was trained using 300 estimators, a maximum depth of 50, and a minimum leaf size of 2. These values were chosen to provide a balance between model complexity and training time, and to ensure that the trees had enough depth to separate subtle physiological patterns without overfitting. I chose Random Forests because they are non-parametric, handle mixed-scale features well, and perform internal feature selection, making them a good match for the broad feature set described in

Question 1 [7]. While I considered applying ANOVA-based feature filtering to reduce dimensionality, early tests showed limited improvement, and the built-in feature importance scoring in Random Forests was sufficient.

Stage 2 followed the same training structure but was trained only on windows that were not classified as ADLs. This allowed the fall vs. near-fall models to focus specifically on meaningful events without being distracted by ADL variation. By separating these tasks, the model increased specificity and reduced false positives in the final decision.

Once sensor-specific models were trained, I validated the system using participant P3033. Final predictions in both stages were generated using a weighted ensemble of the sensor models, with weights of 0.7 for IMU, 0.2 for ECG, and 0.1 for GSS. These values were selected based on P3033 performance and reflected the relative consistency of each modality. IMU models provided the most robust classification performance, while ECG and GSS added valuable context in ambiguous or borderline windows.

Overall, the training process was designed to align with the data organization provided and our goals while maximizing generalization to previously unseen participants. The combination of SMOTE, validation, and ensemble scoring helped create a flexible and interpretable classification pipeline.

3. What are your current results and what do they mean?

During initial exploratory analysis, I found that fall and near-fall durations had a median of approximately 2.4 seconds and 1.5 seconds, respectively. This confirmed that the 1.5-second windows I used, with a 0.5-second overlap, were well-suited to capture complete events and transitional motion. While I intended to include further analysis of raw signal ranges and event-specific feature distributions across sensors, this proved too computationally expensive. Instead, I focused on systematically evaluating model iterations, feature choices, and performance across both stages of the classification pipeline.

Early Models

The first round of models was trained using a basic Random Forest setup with 100 trees and no class balancing. Each sensor (ECG, GSS, IMU) was trained independently to classify windows as ADL or non-ADL. These early models showed high overall accuracy, but failed to reliably detect fall and near-fall windows, reflected in poor recall and F1-scores for the non-ADL class. The confusion matrices showed that many fall-related events were misclassified as ADLs (Table 1).

Table 1. Early Stage 1 Classification Results for Each Sensor

Sensor	Accuracy	Precision (ADL)	Recall (ADL)	F1 (ADL)	Precision (Non-ADL)	Recall (Non-ADL)	F1 (Non-ADL)
ECG	83.5%	0.84	0.99	0.91	0.70	0.08	0.15
GSS	82.9%	0.83	0.99	0.91	0.56	0.06	0.11
IMU	87.2%	0.88	0.98	0.93	0.79	0.32	0.46

These results revealed a strong bias toward the dominant class (ADL), with precision and recall dropping off sharply for fall-related events. To fix this, I made several improvements.

Final Model

In the final pipeline, I applied three key strategies to improve detection:

- I used SMOTE during training to oversample fall and near-fall events.
- I re-tuned the classification threshold for each model using the validation F1-score on P3033.

- I expanded the Random Forests to 300 trees, used class weighting (falls/near-falls = 2x ADLs), and set min_samples_leaf=2.

Table 2. Final Stage 1 Ensemble Performance (ADL vs. Non-ADL) on P3033

Sensor	Accuracy	Precision (ADL)	Recall (ADL)	F1 (ADL)	Precision (Non-ADL)	Recall (Non-ADL)	F1 (Non-ADL)
ECG	62.9%	0.87	0.65	0.74	0.25	0.54	0.34
GSS	19.7%	0.81	0.03	0.06	0.18	0.97	0.30
IMU	81.8%	0.92	0.86	0.89	0.49	0.64	0.56
Weighted Ensemble	83.3%	0.96	0.84	0.89	0.52	0.82	0.64

The metrics in Table 2 demonstrate strong sensitivity to fall-related events. The ensemble significantly improved recall compared to individual sensors and reduced missed fall-related windows from 937 to 180.

The confusion matrix below (Table 3) shows raw counts using the true label distribution from P3033 (4717 ADL, 1017 non-ADL).

Table 3. Stage 1 Confusion Matrix (Ensemble Model on P3033)

	Predicted ADL	Predicted Non-ADL
True ADL	3939	778
True Non-ADL	180	836

Stage 2 was evaluated only on the 836 windows classified as non-ADL by Stage 1, using a weighted ensemble across ECG, GSS, and IMU. However, due to an oversight in the validation script, predictions were filtered to only include windows classified as non-ADL by all three sensor models, rather than using the intended weighted score. This reduced the validation set from 836 to 202 windows, unintentionally biasing performance metrics upward.

Table 4. Stage 2 Confusion Matrix (Filtered Validation Subset)

	Predicted NF	Predicted F
True NF	39	20
True F	0	143

Table 5. Stage 2 Filtered Metrics (Fall vs. Near-Fall)

Accuracy	Precision (F)	Recall (F)	F1 (F)	Precision (NF)	Recall (NF)	F1 (NF)
90.1%	87.7%	100%	93.5%	100%	66.1%	79.6%

While Stage 2 results (Tables 4 and 5) appear strong, they are not accurate reflections of real-world performance. The filtering bug caused an overly narrow validation set, meaning that the model was evaluated only on the easiest non-ADL windows. This will be corrected in the final GitHub submission.

Still, the Stage 1 ensemble reliably identifies fall-related events (recall = 82.3%) with high ADL precision (95.6%). Although the system currently overpredicts falls compared to near-falls, this bias is acceptable in real-world scenarios where the primary goal is to detect risky events. In future work, more

emphasis can be placed on refining near-fall classification. Still, this two-stage system already provides a practical and interpretable fall detection pipeline that can be deployed on wearable sensors.

4. What would you improve for the future?

The two-stage architecture utilized in this project—initially separating ADLs from fall-related activities, followed by distinguishing between near-falls and falls—was one of the most successful elements of the pipeline. It diminished false positives resulting from ADLs being misclassified as events and enabled more focused modelling in the second stage. The incorporation of a weighted ensemble of ECG, GSS, and IMU predictions enhanced robustness by leveraging the unique strengths of each sensor, even when one modality experienced noise. Furthermore, the 0.5-second window overlap proved beneficial in improving temporal resolution and capturing transitions between activities.

One improvement area is the system's generalizability to varied sensor placements. The test dataset includes inconsistently labeled or placed sensors (e.g., “sternum” instead of “back”). To address this, I treated all IMU positions as interchangeable during training. This approach worked reasonably well, but future models could benefit from orientation-invariant features or improved sensor fusion logic that considers spatial positioning. Another improvement is to reduce reliance on GSS and ECG. Although these signals provide useful physiological context, they were less predictive and more participant-dependent than expected. Practically, IMUs are easier to wear and maintain. A simplified IMU-only version of the model may be more suitable for real-world deployment, especially where comfort and compliance are crucial, such as with older populations [9].

An important correction for future iterations is addressing a filtering error in the Stage 2 validation logic. Due to a mistake in the code, Stage 2 predictions were made solely on windows classified as non-ADL by all three individual sensors, rather than using the combined ensemble prediction. This oversight excluded many windows that should have been passed forward and likely inflated the reported Stage 2 metrics. Although the model architecture remains sound, this issue undermines the reliability of the Stage 2 results. The corrected logic will be included in the final GitHub submission.

Another key direction is to continue expanding the dataset by including more participants with diverse movement styles and sensor configurations. An increase in training data would enhance model generalization and facilitate the development of personalized thresholds or calibration logic. This is particularly significant for ECG and GSS signals, which exhibit high variability among participants. Future efforts could also investigate adaptive baselines or real-time recalibration mechanisms for these modalities to strengthen the system's robustness.

While Random Forests were chosen for their speed, interpretability, and compatibility with tabular data, they are limited in capturing time-based patterns and inter-feature interactions [7]. Exploring more complex machine learning approaches that better account for temporal dynamics may lead to stronger performance. These models could help the classifier understand how signal patterns evolve across adjacent windows, improving detection of ambiguous or short-duration events.

Overall, further investigation into participant variability and sensor placement differences will be essential to improving the model’s robustness. Exploring alternative machine learning approaches that can better capture temporal and inter-sensor dynamics may also enhance performance. Additionally, incorporating participant feedback on preferred sensor locations will be necessary for designing a real-time system that is comfortable, wearable, and reliable. Together, these improvements would support the development of a fall and near-fall detection tool that is practical for deployment in older adult populations.

References

[1] Analog Devices. *AD8232: Heart Rate Monitor Front End Datasheet*, 2021. [Online]. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/ad8232.pdf

[2] Tobii, "Galvanic skin response (GSR): Overview," Tobii Pro Insight, 2022. [Online]. Available: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/galvanic-skin-response/

[3] Reddit, "Best practices for IMU sensor filtering," *r/DSP*, 2022. [Online]. Available: https://www.reddit.com/r/DSP/comments/u5b2k9/imu_signal_filtering_best_practices/

[4] M. Shaikh, P. Tiwari, and A. Goel, "Analysis of heart rate variability in fall detection using wearable sensors," *Biomed. Signal Process. Control.*, vol. 68, p. 102605, 2021.

[5] H. F. Posada-Quintero et al., "Time-varying analysis of electrodermal activity during exercise," *PLoS One*, vol. 11, no. 12, p. e0167377, 2016.

[6] D. M. Karantonis et al., "Implementation of a real-time human movement classifier using a triaxial accelerometer," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 156–167, 2006.

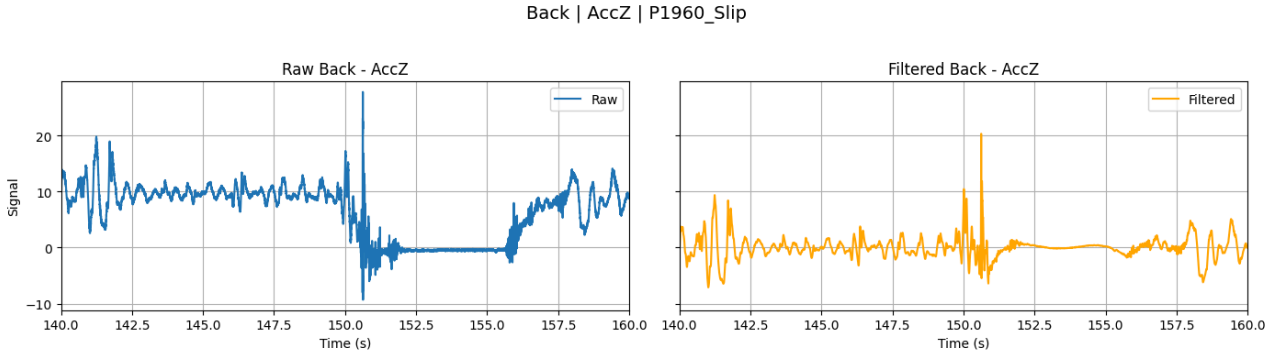
[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

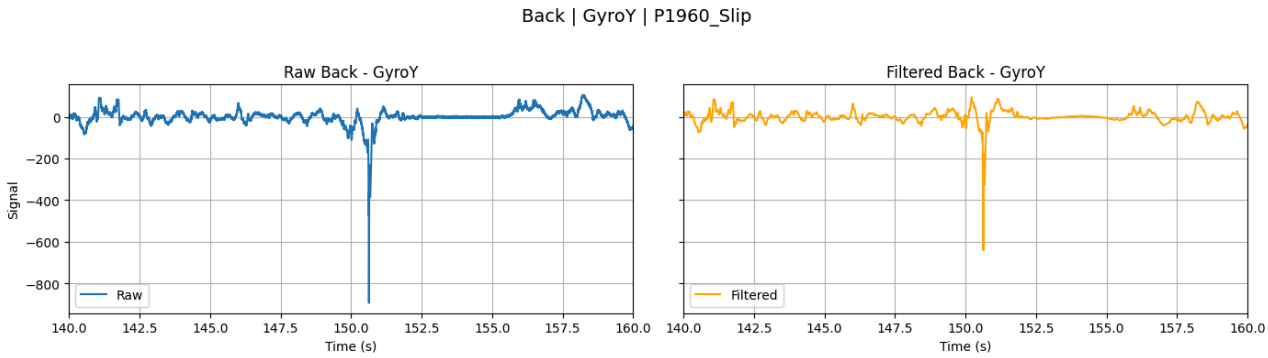
[8] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

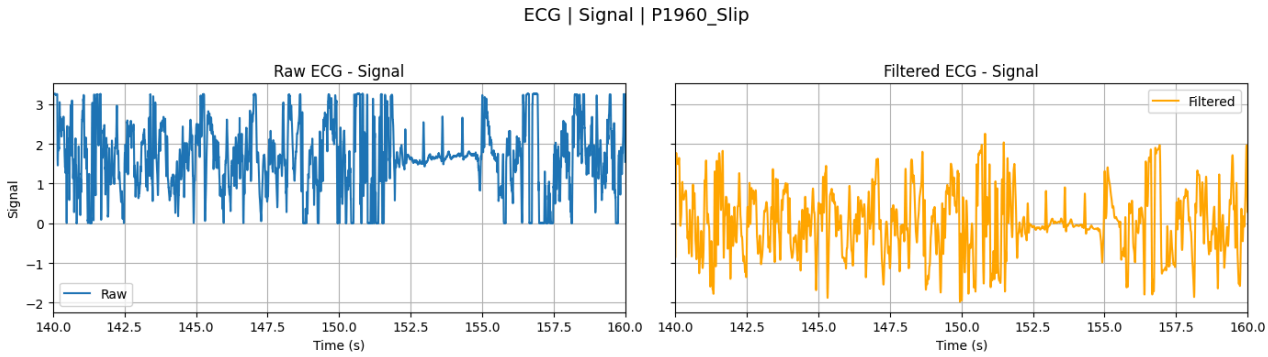
[9] M. Smuck et al., "The emerging clinical role of wearables: factors for successful implementation," *NPJ Digit. Med.*, vol. 4, no. 1, p. 45, 2021.

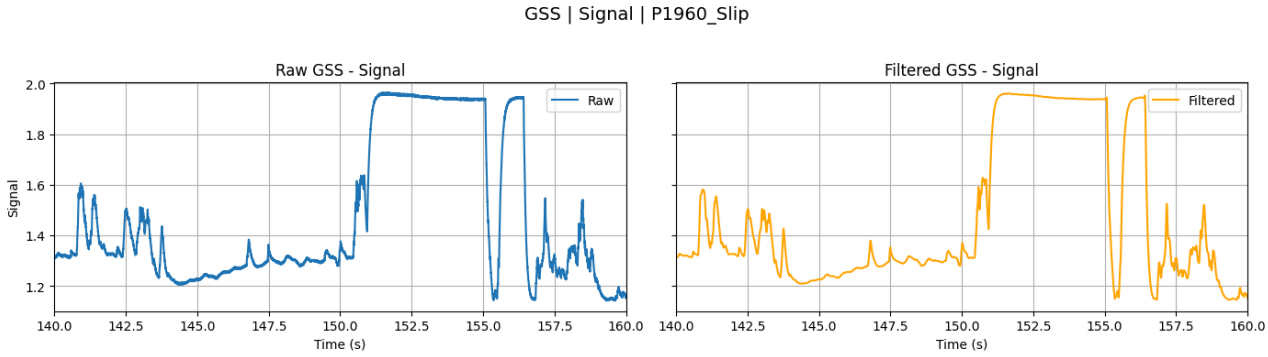
Appendix

Appendix 1 - Raw vs. Filtered Signals for P1960_Slip_Back



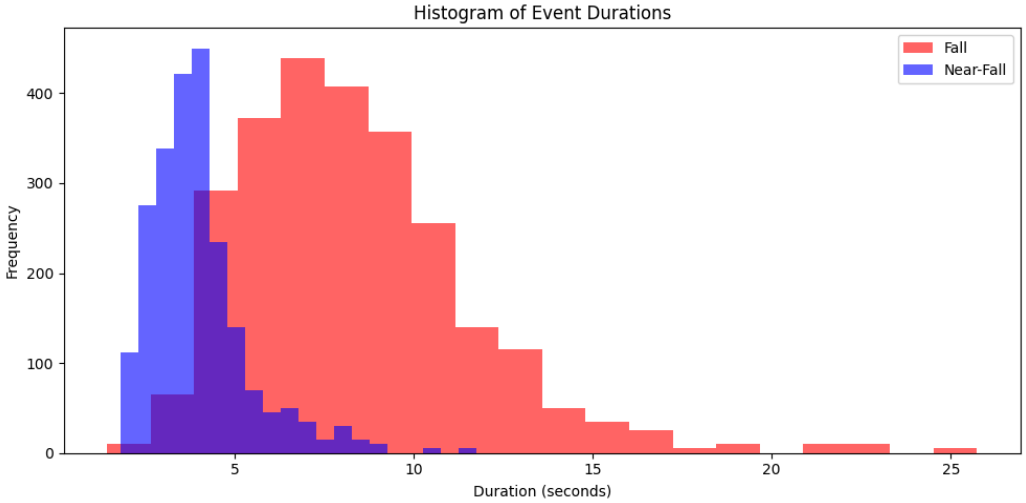


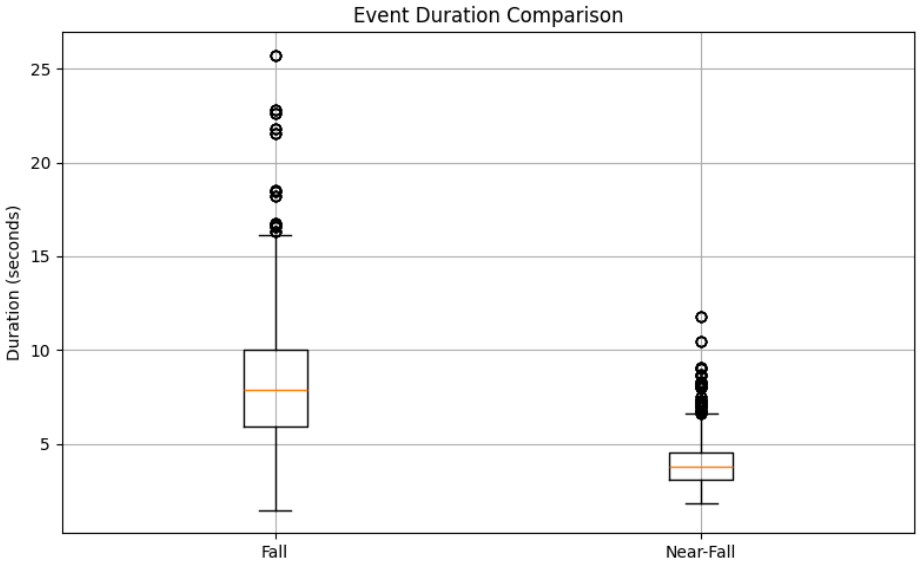




Appendix 2 - Timing and Distribution of Fall and Near Fall Events

Summary Stat	Falls	Near Falls
Count	2602	2252
Mean	8.32s	3.99s
Median	7.85s	3.77s
Minimum	1.44s	1.81s
Maximum	25.73s	11.76s
Standard Deviation	3.24s	1.38s
IQR	4.09s	1.40s





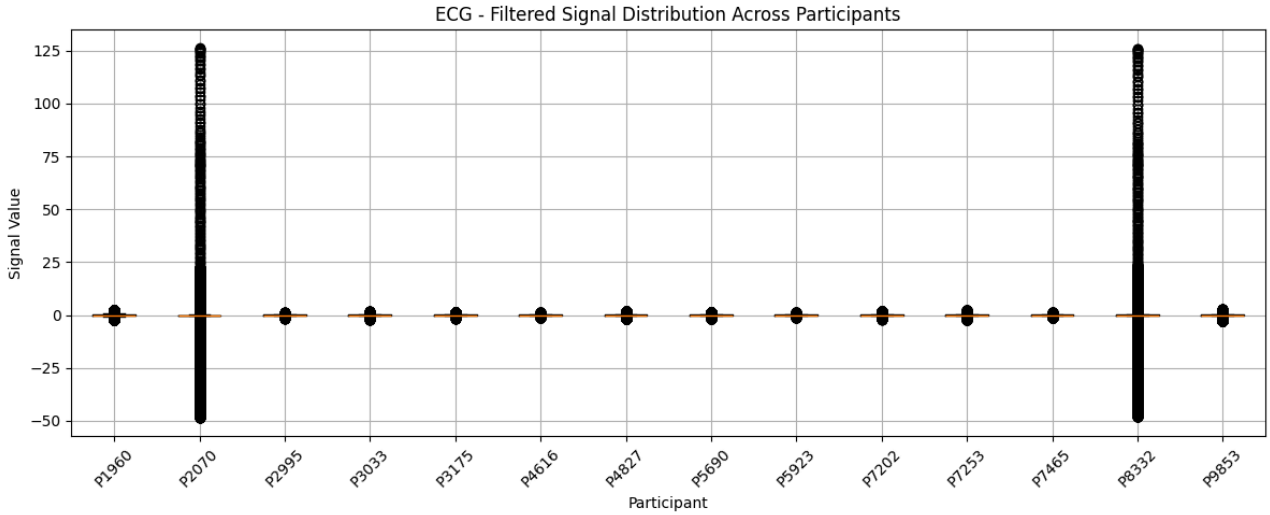
Appendix 3 - Windowed Data File Counts

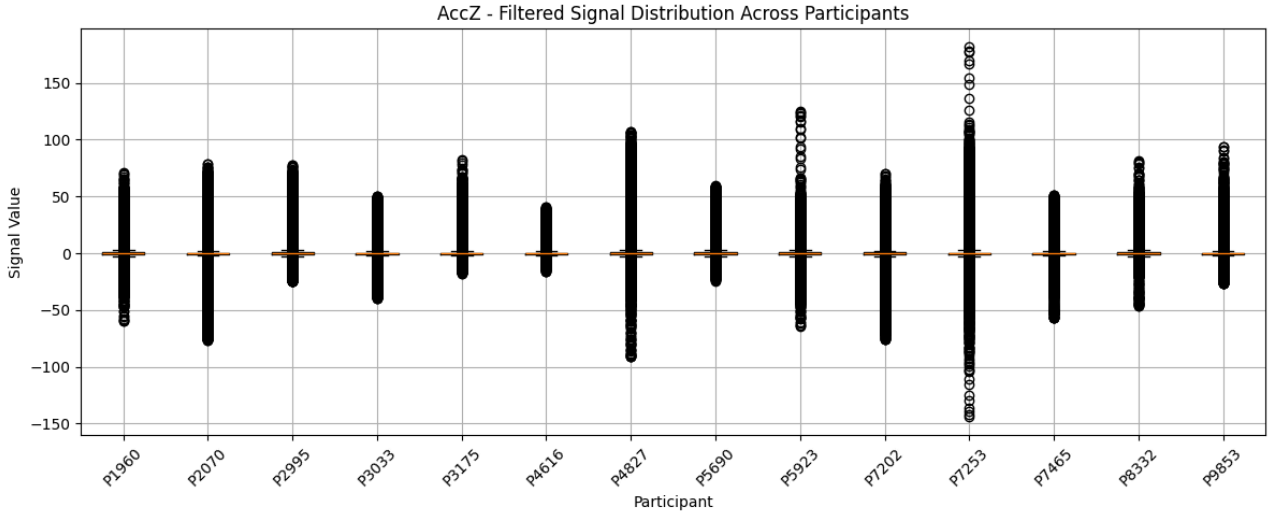
ADL: 302123

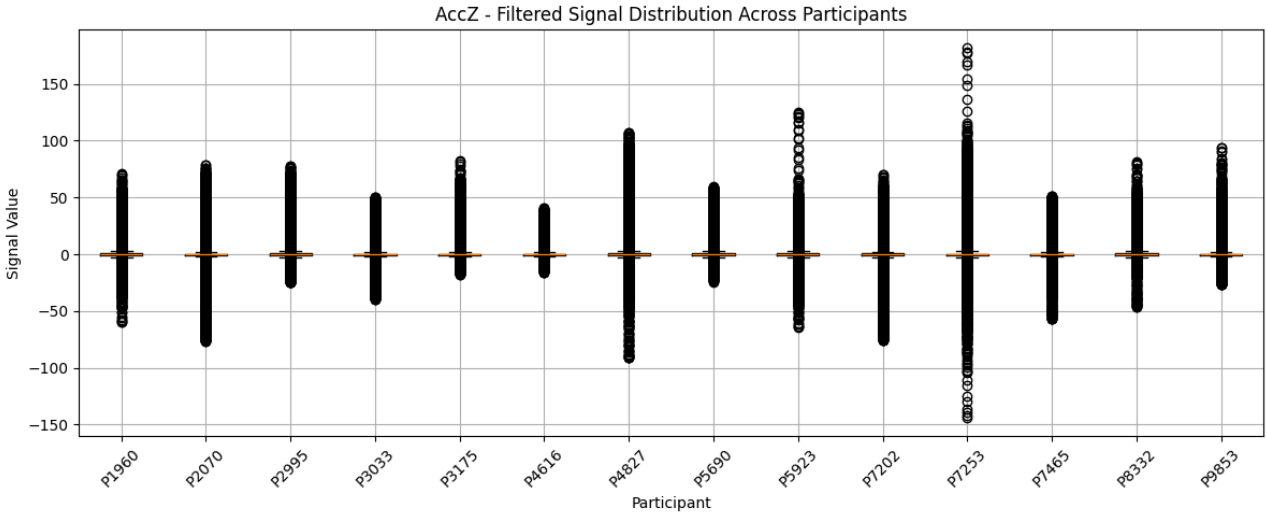
NF: 17926

F: 43162

Appendix 4 - Signal Value Variability Across Participants






AccZ - Filtered Signal Distribution Across Participants
Signal Value
150
100
50
0
−50
−100
−150
P1960
P2070
P2995
P3033
P3175
P4616
P4827
P5690
P5923
P7202
P7253
P7465
P8332
P9853
Participant