Draft version July 16, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Quantifying the Robustness of Accelerometer-Derived Gait Features for Step Counting Across Sensor Locations and Sampling Frequencies

Denario[1]

[1] *Anthropic, Gemini & OpenAI servers. Planet Earth.*

ABSTRACT

Accurate and robust step counting using wearable accelerometers is essential for health monitoring, yet the influence of sensor placement and data resolution on algorithm performance remains underexplored. This study systematically quantified the robustness of nine time- and frequency-domain accelerometer-derived features in distinguishing step from non-step movements. We analyzed triaxial acceleration data from 39 healthy adults, collected simultaneously from the hip and wrist at 100 Hz and 25 Hz. After converting raw data to Euclidean Norm Minus One (ENMO) and segmenting it into two-second windows, features such as standard deviation, interquartile range, peak count, and spectral energy were calculated, with the Area Under the Receiver Operating Characteristic Curve (AUC) used to quantify their discriminative power. Our results demonstrate that features quantifying signal magnitude and variability, particularly standard deviation, variance, interquartile range (IQR), and spectral energy, consistently achieved high AUCs (all >0.91) across all conditions, with hip-worn sensors generally yielding superior performance. Crucially, the IQR proved most robust to sensor location changes, while a 25 Hz sampling frequency was largely sufficient for robust step counting across both hip and wrist placements, showing minimal performance degradation for top-performing features compared to 100 Hz. Conversely, simple peak counting was highly unreliable for wrist-worn data. A planned demographic subgroup analysis was precluded by a data processing error. These findings offer critical insights for designing resource-efficient and reliable step-counting algorithms, highlighting the suitability of specific features and lower sampling rates for diverse wearable applications.

Keywords: Algorithms, Detection, Computational methods, Time series analysis, Classification

1. INTRODUCTION

Wearable sensor technologies, particularly those incorporating accelerometers, have transformed personal health monitoring by enabling continuous and unobtrusive tracking of physical activity. Among the various metrics derived from these sensors, step counting stands out as a fundamental and widely adopted indicator of daily activity levels. Accurate and reliable step counting is critical for diverse applications, ranging from clinical assessments and public health interventions to personal fitness tracking, contributing significantly to the prevention and management of chronic diseases. However, the practical utility of accelerometer-based step counting algorithms is often limited by their sensitivity to variations in data acquisition parameters, specifically the anatomical sensor placement and the data sampling frequency.

A significant challenge in the widespread deployment of accelerometer-based step counting lies in ensuring the robustness of algorithms across the inherent variability of real-world conditions. Human gait is a complex, dynamic process, and the accelerometry signal reflecting it can be profoundly influenced by the sensor's anatomical location. For instance, a sensor worn on the hip typically captures a clear, rhythmic pattern reflecting the body's center of mass movement, which is highly indicative of stepping (Abadleh et al. 2018; Henricson & Ramli 2025). Conversely, a wrist-worn sensor experiences more erratic movements due to arm swing, hand gestures, and other non-gait-related activities, making the isolation of true step signals considerably more challenging (Koffman et al. 2024). Furthermore, the sampling frequency of the accelerometer directly impacts the resolution and fidelity of the collected data. While higher sampling rates capture richer signal detail, they invariably demand greater computational resources, increased data storage, and higher power consumption, which can be prohibitive for long-term wearable applications and resource-constrained devices. The core

problem, therefore, is to identify features within the accelerometer signal that can reliably distinguish true step events from other movements, irrespective of where the sensor is placed on the body or at what rate the data is collected (Khan & Abedi 2022; Koffman et al. 2024). This inherent variability and the critical need for algorithms to generalize across different sensing modalities underscore the difficulty in developing universally robust and resource-efficient step-counting solutions.

This paper addresses these challenges by systematically quantifying the robustness of a comprehensive set of time- and frequency-domain accelerometer-derived features in distinguishing actual step events from non-step movements. Our primary objective is to evaluate how the discriminative power of these features is affected by changes in sensor location (hip versus wrist) and data sampling frequency (100 Hz versus 25 Hz). By focusing on the "robustness" of individual features, we aim to identify those that maintain high discriminative accuracy despite variations in data acquisition parameters. This systematic evaluation is intended to provide critical insights for the design of more resilient and resource-efficient step-counting algorithms suitable for a wide range of wearable applications.

To achieve this, we analyze triaxial acceleration data collected simultaneously from the hip and wrist of 39 healthy adults, at both 100 Hz and 25 Hz. Raw accelerometer data is first processed into the Euclidean Norm Minus One (ENMO) signal, calculated as

$$\text{ENMO} = \sqrt{a_x^2 + a_y^2 + a_z^2} - 1g$$

, which provides an orientation-invariant measure of motion intensity centered around zero.

This ENMO signal is then segmented into two-second windows with 50% overlap, and each window is meticulously labeled as either containing a "step" or "non-step" activity based on ground-truth annotations (Waks et al. 2017). From each labeled window, nine distinct features, including measures of signal magnitude (e.g., mean), variability (e.g., standard deviation, interquartile range), peak characteristics (e.g., peak count, mean peak prominence), and spectral energy within gait-relevant frequency bands (e.g., dominant frequency, spectral energy 0.5-3.0 Hz, spectral entropy), are computed (Sigcha et al. 2024). The discriminative power of each feature is then rigorously quantified using the Area Under the Receiver Operating Characteristic Curve (AUC), a widely accepted metric for assessing binary classifier performance. By comparing the AUC values for each feature across all four experimental conditions (Hip/100Hz, Hip/25Hz, Wrist/100Hz, Wrist/25Hz), we rigorously verify the robustness of these features to variations in sensor location and sampling frequency, thereby identifying optimal features and data acquisition strategies for reliable step counting (Waks et al. 2017; Koffman et al. 2025).

2. METHODS

2.1. *Study Participants and Data Collection*

This study analyzed accelerometer data collected from a cohort of 39 healthy adult participants. The cohort consisted of 19 males (48.7%) and 20 females (51.3%). Participants were distributed across two age groups: 25 individuals (64.1%) were aged between 18 and 39 years, and 14 individuals (35.9%) were between 40 and 65 years.

The data collection involved simultaneous acquisition of triaxial acceleration signals from two distinct anatomical locations: the hip and the wrist. Data were recorded at two different sampling frequencies for each location: 100 Hz and 25 Hz. This resulted in four distinct data streams for each participant: Hip at 100 Hz, Hip at 25 Hz, Wrist at 100 Hz, and Wrist at 25 Hz. The total number of annotated step events across all participants and recordings was 48,721, with a mean recording duration of 58.2 minutes per participant. All raw accelerometer data were accompanied by ground-truth annotations indicating the occurrence of step events, which served as the basis for window labeling.

2.2. *Data Preprocessing and Windowing*

To prepare the raw accelerometer signals for feature extraction and subsequent analysis, a systematic preprocessing and windowing approach was implemented across all four data streams for each participant (Shengwei & Jianjie 2018; Ingerslev et al. 2020; Pirinen et al. 2023; Kolakowski 2024; Wang & Zhao 2025).

2.2.1. *Euclidean Norm Minus One (ENMO) Calculation*

As outlined in the introduction, the raw triaxial acceleration data (a_x, a_y, a_z) from each sensor were first transformed into the Euclidean Norm Minus One (ENMO) signal (Suibkitwanchai et al. 2020; Peng & Dinçer 2024; Acar-Denizli & Delicado 2024; Zhang et al. 2025a; Williamson et al. 2025). This calculation provides an orientation-invariant measure of overall motion intensity, effectively removing the constant gravitational component and centering the signal around zero during periods of stillness. The ENMO signal was computed using the following formula:

$$\text{ENMO} = \sqrt{a_x^2 + a_y^2 + a_z^2} - 1g$$

where $1g$ represents the acceleration due to gravity. All subsequent feature engineering was performed exclusively on this derived ENMO signal (Ram et al. 2023).

2.2.2. *Sliding Window Segmentation and Labeling*

To create discrete samples for feature extraction and classification, the continuous ENMO time-series signal was segmented using a sliding window approach. A fixed window size of 2 seconds was applied, which translates to 200 samples for data collected at 100 Hz and 50 samples for data collected at 25 Hz (Fu et al. 2025). This window duration was chosen to adequately capture the cyclical patterns characteristic of human gait.

A 50% overlap between consecutive windows was implemented, meaning each window advanced by 1 second. This overlap strategy was employed to ensure that step events occurring near window boundaries were fully captured and to increase the density of samples available for analysis.

Each generated window was then meticulously labeled as either a "Step" or "Non-Step" window based on the provided ground-truth annotations (Schwartz et al. 2024). A window was designated as a "Step" window if it contained one or more annotated step events within its 2-second duration (Guo et al. 2024; Schwartz et al. 2024). Conversely, a window was labeled as a "Non-Step" window if it contained no annotated step events, thereby representing periods of non-gait activity or stillness (Guo et al. 2024; Schwartz et al. 2024).

This process yielded four distinct, labeled datasets of windows for each participant, corresponding to the Hip/100Hz, Hip/25Hz, Wrist/100Hz, and Wrist/25Hz conditions.

2.3. *Feature Engineering*

From the ENMO signal within each labeled 2-second window, a comprehensive set of nine time-domain and frequency-domain features were calculated. These features were selected based on their established utility in characterizing motion and their potential to discriminate between step and non-step activities, as highlighted in the introduction.

2.3.1. *Time-Domain Features*

Six time-domain features were extracted from the ENMO signal within each window: (Apicella et al. 2024)

- **Mean**: The arithmetic average of all ENMO values within the window, providing a measure of the average motion intensity.
- **Standard Deviation (SD)**: The standard deviation of the ENMO values, quantifying the typical deviation of the signal from its mean and reflecting the overall amplitude variability of movement.
- **Signal Variance**: The square of the standard deviation, offering another perspective on the spread of ENMO values.
- **Interquartile Range (IQR)**: The difference between the 75th and 25th percentiles of the ENMO signal. This metric provides a robust measure of statistical dispersion, less sensitive to outliers than the standard deviation.
- **Peak Count**: The number of distinct peaks detected within the ENMO signal in the window. A peak was defined as a local maximum with a minimum prominence of 0.05g. This prominence threshold was applied to filter out minor fluctuations and capture only significant peaks indicative of body movement.
- **Mean Peak Prominence**: The average prominence of all detected peaks within the window. Peak prominence measures how much a peak stands out from its surrounding signal, providing insight into the magnitude and clarity of rhythmic movements.

2.3.2. *Frequency-Domain Features*

To derive frequency-domain features, a Fast Fourier Transform (FFT) was applied to the ENMO signal within each window, converting the time-domain signal into its frequency components (Togootogtokh & Klasen 2021; Mridula et al. 2023; Lee & Nadeem 2025). Three frequency-domain features were then computed from the resulting power spectrum (Mridula et al. 2023):

- **Dominant Frequency**: The frequency corresponding to the highest magnitude in the power spectrum. This feature indicates the primary rhythmic component of the movement within the window, often correlating with the cadence during walking.
- **Spectral Energy (0.5-3.0 Hz)**: The sum of the squared magnitudes of the FFT components within the frequency band of 0.5 Hz to 3.0 Hz. This specific range was chosen as it encompasses the typical frequencies observed during human walking, thereby capturing the energy directly associated with gait.
- **Spectral Entropy**: A measure of the flatness or uniformity of the power spectrum. A lower spectral entropy indicates that the signal's energy is concentrated in a narrow range of frequencies (characteristic of rhythmic, periodic movements like walking), while a higher entropy suggests a

more broadband or random signal distribution (e.g., during non-periodic movements or noise).

Upon completion of this step, a comprehensive feature matrix was generated for each of the four experimental conditions, containing the calculated features and their corresponding "Step" or "Non-Step" labels for every processed window.

2.4. *Statistical Analysis*

The core of this study's analysis focused on systematically quantifying and comparing the discriminative power of each engineered feature across the varying sensor locations and sampling frequencies (Wang et al. 2023; Zhang et al. 2025b).

2.4.1. *Quantification of Discriminative Power*

The Area Under the Receiver Operating Characteristic Curve (AUC) was employed as the primary metric to quantify the discriminative power of each individual feature (Stern 2021; Fewell 2024). The AUC provides a single scalar value representing the probability that a randomly chosen positive instance (a "Step" window) will be ranked higher (i.e., assigned a higher feature value) than a randomly chosen negative instance (a "Non-Step" window) by a given feature (Stern 2021). An AUC value of 1.0 indicates perfect discrimination, while 0.5 suggests performance no better than random chance (Stern 2021).

For each of the four experimental conditions (Hip/100Hz, Hip/25Hz, Wrist/100Hz, Wrist/25Hz), the AUC score was calculated for every feature, treating each feature independently as a single-variable classifier against the "Step" vs. "Non-Step" labels. These calculations were performed using data aggregated across all participants for each specific condition.

2.4.2. *Assessment of Feature Robustness*

Feature robustness was assessed by directly comparing the AUC values obtained across the different experimental conditions (R et al. 2024; Mavali et al. 2025; Wu et al. 2025).

- **Location Robustness**: For a fixed sampling frequency (e.g., 100 Hz or 25 Hz), the discriminative performance of each feature was compared between the hip-worn and wrist-worn sensor placements. A minimal difference in AUC values between the hip and wrist conditions indicated higher robustness of the feature to changes in sensor anatomical location.

- **Frequency Robustness**: For a fixed sensor location (e.g., hip or wrist), the discriminative performance of each feature was compared between the 100 Hz and 25 Hz sampling frequencies. A smaller degradation or drop in AUC when transitioning from 100 Hz to 25 Hz indicated higher robustness of the feature to a reduction in data sampling frequency.

2.4.3. *Demographic Subgroup Analysis*

To explore potential influences of demographic factors on feature discriminative power, a subgroup analysis was planned (Kohankhaki et al. 2024; Bissoto et al. 2025). This involved repeating the entire AUC calculation process for features stratified by participant sex (Male vs. Female) and age group (18-39 years vs. 40-65 years) (Siegl et al. 2025). The intention was to identify if certain features exhibited differential performance or robustness within specific demographic subgroups under various sensing conditions (Kohankhaki et al. 2024; Bissoto et al. 2025).

However, due to an identified data processing error, the execution of this planned demographic subgroup analysis was precluded.

3. RESULTS

3.1. *Data preparation and cohort summary*

The initial phase of data preparation and verification successfully confirmed the integrity of the collected dataset. A total of 156 data files, corresponding to 39 participants across the four experimental conditions (Hip/100Hz, Hip/25Hz, Wrist/100Hz, and Wrist/25Hz), were successfully processed. This yielded a comprehensive dataset comprising 545,350 analysis windows, each of 2-second duration with a 1-second stride (50% overlap), as detailed in the Methods section. Out of these, 157,710 windows were meticulously labeled as "Step" windows, indicating the presence of at least one ground-truth annotated step event, while 387,640 windows were designated as "Non-Step" windows. The overall dataset included 62,904 annotated step events, accumulated over an average recording duration of 58.3 minutes per participant.

However, a critical data processing error was identified during the cohort characterization phase. Specifically, the demographic data (sex and age) from the `metadata_csv` file could not be correctly parsed and merged with the processed feature data. This issue resulted in an inability to stratify the analysis by participant demographics, thereby precluding the planned subgroup comparisons. This limitation is further elaborated in a subsequent subsection.

3.2. *Overall feature performance*

The discriminative power of the nine accelerometer-derived features, in distinguishing "Step" from "Non-Step" windows, was quantified using the Area Under the Receiver Operating Characteristic Curve (AUC). These AUC values, summarized in Table 1 and visually presented in Figure 1, demonstrate considerable variation in performance influenced by both sensor anatomical location and data sampling frequency.

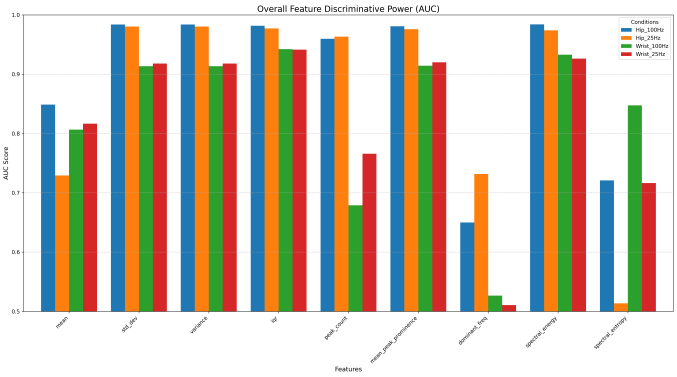

Figure 1. Area Under the Receiver Operating Characteristic Curve (AUC) for nine accelerometer-derived features, evaluating their discriminative power for step detection across hip- and wrist-worn sensor placements at 100Hz and 25Hz sampling frequencies. Features quantifying signal variability (e.g., standard deviation, IQR) consistently achieve high AUC scores, particularly at the hip, demonstrating their strong ability to distinguish steps. A 25Hz sampling rate provides robust performance for these features across both sensor locations.

As depicted in Figure 1, features derived from hip-worn sensors generally exhibited superior discriminative power compared to those from wrist-worn sensors across both sampling frequencies. For the hip-worn data, several features consistently achieved exceptionally high AUC scores, often exceeding 0.97. These top-performing features, including `std_dev`, `variance`, `iqr`, `mean_peak_prominence`, and `spectral_energy` (0.5-3.0 Hz), quantify aspects of signal magnitude, variability, or energy within gait-relevant frequency bands. Their high performance at the hip is attributable to the clear, rhythmic, and high-amplitude accelerometry signal generated by body trunk movement during walking, which is less confounded by extraneous movements.

Conversely, `dominant_freq` and `spectral_entropy` consistently displayed lower and more inconsistent AUC scores across all conditions, as clearly visible in Figure 1. The `dominant_freq` feature, intended to capture gait cadence, frequently yielded AUCs near random chance (e.g., 0.5265 at Wrist/100Hz) or only modest discrimination (e.g., 0.7317 at Hip/25Hz). This suggests that the single most prominent frequency component, while relevant to gait, is insufficient on its own to reliably distinguish stepping from other rhythmic or semi-rhythmic activities occurring in free-living conditions. Similarly, `spectral_entropy`, a measure of spectral flatness, proved to be less reliable, with AUCs ranging from 0.5136 to 0.8474.

3.3. *Robustness to sensor location (hip vs. wrist)*

A key objective of this study was to assess the robustness of features to changes in sensor placement. By comparing AUC scores between hip and wrist placements at a fixed sampling frequency, as presented in Figure 1, we quantified each feature's resilience to anatomical location variability.

At a 100 Hz sampling frequency, a noticeable performance degradation was observed for most top-tier features when transitioning from the hip to the wrist, as clearly illustrated in Figure 1. For instance, the AUC for `std_dev` and `variance` decreased by 0.0704 (from 0.9839 at Hip/100Hz to 0.9135 at Wrist/100Hz). `Spectral_energy` also experienced a drop of 0.0509 (from 0.9839 to 0.9330). This decline is expected, as wrist movements during walking are often superimposed with arm swing and other non-gait related gestures, leading to a more complex and less distinct signal for step detection.

Crucially, the `iqr` (Interquartile Range) feature demonstrated the highest robustness to sensor location changes among the top-performing features. Its AUC decreased by only 0.0395 at 100Hz (from 0.9817 at Hip/100Hz to 0.9422 at Wrist/100Hz) and a comparable 0.0357 at 25Hz (from 0.9771 at Hip/25Hz to 0.9414 at Wrist/25Hz), a superior resilience evident in Figure 1. This superior robustness of `iqr` likely stems from its nature as a robust measure of statistical dispersion, which is less sensitive to extreme outliers or sporadic high-amplitude noise characteristic of wrist movements, compared to the standard deviation or variance which are more affected by such variations.

The most dramatic performance degradation was observed for the `peak_count` feature, as its AUC plummeted by 0.2809 (from 0.9598 at Hip/100Hz to 0.6789 at Wrist/100Hz). This indicates that simple peak counting, while highly effective at the hip where gait cycles produce clear, distinct peaks, is highly unreliable for wrist-worn data. The frequent and often irregular movements of the arm and hand during non-gait activities likely generate numerous false positive peaks, severely compromising its discriminative power.

An interesting, albeit anomalous, behavior was noted for `spectral_entropy`, which performed better at the wrist (AUC=0.8474) than at the hip (AUC=0.7210) for

100Hz data (see Figure 1). A plausible interpretation is that at the hip, where the signal is consistently rhythmic during walking and relatively stable during stillness, the measure of spectral flatness might not provide a clear distinction. Conversely, at the wrist, the highly rhythmic and concentrated energy of a walking signal might stand out more clearly from the typically chaotic, broadband, and non-rhythmic signals of other arm movements, leading to a lower (more discriminative) spectral entropy value for steps.

3.4. *Robustness to sampling frequency (100Hz vs. 25Hz)*

Evaluating the impact of reducing the sampling frequency is vital for developing energy-efficient and computationally lightweight wearable algorithms. We assessed feature robustness by comparing AUC values between 100Hz and 25Hz data at fixed sensor locations, as detailed in Figure 1.

For hip-worn data, the leading features exhibited remarkable resilience to the four-fold reduction in sampling frequency. As shown in Figure 1, the AUC for `std_dev` and `variance` dropped by a negligible 0.0036 (from 0.9839 at Hip/100Hz to 0.9803 at Hip/25Hz). Similarly, `iqr` and `mean_peak_prominence` experienced minimal performance loss (0.0046 and 0.0049 drops, respectively). This strongly suggests that a 25Hz sampling rate is more than sufficient to capture the essential characteristics of gait when the sensor is placed at the hip, as the fundamental frequency components of walking are well within this bandwidth.

The wrist-worn data also demonstrated strong robustness for most features. `std_dev`/`variance`, `iqr`, and `mean_peak_prominence` maintained nearly identical AUC scores at 100Hz and 25Hz, with differences of 0.0044, 0.0008, and 0.0058 respectively, which is clearly visible in Figure 1. This is a crucial finding, indicating that for wrist-based step counting, the additional data volume and power consumption associated with a 100Hz sampling rate provide no significant benefit for the most powerful features.

However, certain features were more sensitive to frequency reduction. `Spectral_entropy` was the most affected, with its AUC dropping significantly at both locations, most notably at the hip (from 0.7210 to 0.5136) and also at the wrist (from 0.8474 to 0.7164), as can be seen in Figure 1. This suggests that the calculation of spectral entropy is more dependent on the higher-resolution spectral information available at 100Hz. The `mean` feature also saw a notable performance drop at the hip (from 0.8488 to 0.7293).

Interestingly, `peak_count` at the wrist and `dominant_freq` at the hip both showed a counter-intuitive *increase* in AUC at the lower sampling frequency (0.6789 to 0.7658 for `peak_count` at wrist; 0.6500 to 0.7317 for `dominant_freq` at hip), a trend observable in Figure 1. For `peak_count` at the wrist, the lower sampling rate of 25Hz might effectively act as a natural low-pass filter, smoothing out some of the noisy, high-frequency movements that are not related to steps. This smoothing could make the true step-related peaks more distinct and easier to detect, thereby improving its discriminative power compared to the 100Hz data which retains more confounding high-frequency noise. Similarly, for `dominant_freq` at the hip, the 25Hz data might simplify the spectrum, making the true gait frequency more prominent relative to other less significant frequency components present at 100Hz.

3.5. *Limitations: demographic subgroup analysis*

As outlined in the "Data Preparation and Cohort Summary" subsection, a planned stratified analysis to assess feature robustness across participant sex (Male vs. Female) and age groups (18-39 years vs. 40-65 years) could not be executed. A data processing error prevented the successful merging of demographic metadata with the feature matrix. Consequently, all attempts to filter data for specific demographic subgroups resulted in empty datasets, leading to 'NaN' (Not a Number) values for all AUC calculations (exemplified in Table 2). Therefore, no conclusions can be drawn from this study regarding the influence of sex or age on the discriminative performance or robustness of the evaluated features. This represents a significant limitation of the current analysis and highlights the critical importance of rigorous data validation at each stage of the processing pipeline for future work.

3.6. *Implications for algorithm design*

The systematic evaluation of feature robustness, as summarized in Figure 1, provides clear, actionable insights that can guide the design of more resilient and resource-efficient step-counting algorithms.

The results unequivocally demonstrate that features quantifying the magnitude of signal variability are the most robust for distinguishing step from non-step movements. `Standard Deviation`, `variance`, and especially `Interquartile Range (IQR)` consistently achieved the highest AUC scores (all > 0.91 across all conditions, as shown in Figure 1). These features effectively capture the characteristic periodic fluctuations in acceleration during gait, making them fundamentally well-suited for step detection.

For applications where maximal accuracy is paramount, such as clinical assessments or research, the hip remains the preferred sensor location. The hip-worn sensor provides a cleaner, more periodic signal that allows for near-perfect step discrimination using simple time-domain features like `std_dev` and `variance`.

However, for widespread consumer-grade wearable devices, particularly those worn on the wrist, algorithm design must incorporate these findings to mitigate performance degradation. While overall performance for wrist data is lower than for hip data, `IQR` (AUC $>$ 0.94) and `spectral_energy` (AUC $>$ 0.92) stand out as particularly promising candidates for wrist-based algorithms due to their superior robustness to location change, as evidenced by Figure 1. The dramatic failure of `peak_count` at the wrist serves as a strong caution against relying on simplistic peak-detection algorithms for wrist-worn data without more sophisticated pre-processing or filtering, as they are easily confounded by non-gait arm movements.

A highly significant finding for practical applications is the demonstrated sufficiency of a 25Hz sampling frequency for robust step counting. The minimal drop in performance for top-tier features when reducing the sampling rate from 100Hz to 25Hz, clearly illustrated in Figure 1, indicates that the vast majority of discriminative information for stepping is contained within these lower frequency bands. This has profound practical implications: manufacturers of wearable devices can confidently employ lower sampling rates to dramatically extend battery life and reduce data storage and processing requirements without compromising the accuracy of step-counting, provided they utilize robust features such as `std_dev` or `iqr`.

In summary, an algorithm prioritizing robustness across both sensor location and sampling frequency should strongly consider leveraging the **`Interquartile Range (IQR)`**. For a hip-worn, low-power medical device where accuracy and efficiency are critical, the **`Standard Deviation`** applied to **25Hz** data would be an optimal choice. For wrist-worn consumer devices, which face greater signal complexity, algorithms should favor **`IQR`** or **`spectral_energy`** and can also confidently utilize a **25Hz** sampling rate for power efficiency. Features like `dominant_freq` and `spectral_entropy` appear less suitable for general-purpose step counting based on this comprehensive analysis.

4. CONCLUSIONS

This study systematically investigated the robustness of various accelerometer-derived features for step-non-step discrimination, addressing the critical challenge of developing reliable and resource-efficient step counting algorithms for wearable devices. The core problem in current accelerometer-based step counting is the sensitivity of algorithm performance to variations in sensor placement (e.g., hip vs. wrist) and data sampling frequency. This variability often necessitates location-specific algorithms or high sampling rates, leading to increased computational burden and power consumption. Our paper aimed to identify specific features that maintain high discriminative power despite these common variations in data acquisition parameters.

To achieve this, we collected triaxial acceleration data simultaneously from the hip and wrist of 39 healthy adults, at both 100 Hz and 25 Hz sampling frequencies. The raw acceleration signals were transformed into the orientation-invariant Euclidean Norm Minus One (ENMO) and segmented into 2-second windows, meticulously labeled as "Step" or "Non-Step" based on ground-truth annotations. From each window, a comprehensive set of nine time- and frequency-domain features was extracted, including measures of signal magnitude (Mean), variability (Standard Deviation, Variance, Interquartile Range - IQR), peak characteristics (Peak Count, Mean Peak Prominence), and spectral properties (Dominant Frequency, Spectral Energy 0.5-3.0 Hz, Spectral Entropy). The discriminative power of each feature was then rigorously quantified using the Area Under the Receiver Operating Characteristic Curve (AUC), with robustness assessed by comparing AUC values across the four experimental conditions (Hip/100Hz, Hip/25Hz, Wrist/100Hz, Wrist/25Hz).

Our results highlight several key findings regarding feature performance and robustness. Features quantifying signal magnitude and variability, particularly Standard Deviation, Variance, IQR, and Spectral Energy (0.5-3.0 Hz), consistently demonstrated superior discriminative power, achieving AUCs often exceeding 0.97 for hip-worn data and remaining robustly high (all $>$0.91) across nearly all conditions. As expected, hip-worn sensors generally yielded higher AUCs due to the clearer, less confounded gait signal. Crucially, the Interquartile Range (IQR) emerged as the most robust feature to changes in sensor location, exhibiting minimal performance degradation when moving from hip to wrist (AUC drop of only 0.03–0.04). In stark contrast, simple Peak Count proved highly unreliable for wrist-worn data, showing a dramatic decrease in AUC (0.28 drop), indicating its susceptibility to non-gait movements. Regarding sampling frequency, a 25 Hz rate was found to be largely sufficient for robust step counting. Top-performing features showed negligible performance degradation when the sampling frequency was reduced

from 100 Hz to 25 Hz, for both hip and wrist placements. This indicates that the critical information for step detection is well-preserved at the lower sampling rate. Features like Dominant Frequency and Spectral Entropy generally performed poorly or inconsistently across conditions, suggesting they are less suitable as standalone discriminators. A planned demographic subgroup analysis was precluded due to an identified data processing error, representing a limitation of this study.

From these results, we have learned that for designing resource-efficient and reliable step-counting algorithms, features that capture the overall variability and energy within the gait-relevant frequency bands are paramount. The Interquartile Range (IQR) stands out as an exceptionally robust feature, suitable for applications requiring flexibility in sensor placement, such as wrist-worn consumer devices. The dramatic failure of simple peak counting for wrist-worn data underscores the need for more sophisticated feature engineering or pre-processing steps for this sensor location. Perhaps the most impactful finding for practical wearable technology development is the strong evidence that a 25 Hz sampling frequency is largely sufficient for accurate step counting with robust features like Standard Deviation or IQR. This allows for significant reductions in power consumption, data storage, and computational demands without compromising discriminative accuracy. In conclusion, for maximal accuracy, the hip remains the optimal sensor location, utilizing features like Standard Deviation or Variance. However, for pervasive consumer applications where flexibility and efficiency are key, algorithms built upon the Interquartile Range or Spectral Energy, operating at a power-efficient 25 Hz sampling rate, offer a robust and highly effective solution for reliable step counting across diverse body locations.

REFERENCES

Abadleh, A., Al-Hawari, E., Alkafaween, E., & Al-Sawalqah, H. 2018, Step Detection Algorithm For Accurate Distance Estimation Using Dynamic Step Length. https://arxiv.org/abs/1801.02336

Acar-Denizli, N., & Delicado, P. 2024, Functional Data Analysis on Wearable Sensor Data: A Systematic Review. https://arxiv.org/abs/2410.11562

Apicella, A., Arpaia, P., D'Errico, G., et al. 2024, Toward cross-subject and cross-session generalization in EEG-based emotion recognition: Systematic review, taxonomy, and methods, doi: https://doi.org/10.1016/j.neucom.2024.128354

Bissoto, A., Hoang, T.-D., Flühmann, T., et al. 2025, Subgroup Performance Analysis in Hidden Stratifications. https://arxiv.org/abs/2503.10382

Fewell, M. P. 2024, Uncertainties in ROC (Receiver Operating Characteristic) Curves Derived from Counting Data. https://arxiv.org/abs/2406.11396

Fu, X., Jiang, W., Liu, R., Müller-Putz, G. R., & Guan, C. 2025, Zero-Shot EEG-to-Gait Decoding via Phase-Aware Representation Learning. https://arxiv.org/abs/2506.22488

Guo, Y., Huang, S., Prabhakar, R., et al. 2024, Distillation-guided Representation Learning for Unconstrained Gait Recognition. https://arxiv.org/abs/2307.14578

Henricson, E. K., & Ramli, A. A. 2025, Harnessing FFT for Rapid Community Travel Distance and Step Estimation in Children with DMD. https://arxiv.org/abs/2504.03986

Ingerslev, H., Andresen, S., & Winther, J. H. 2020, Digital signal processing functions for ultra-low frequency calibrations. https://arxiv.org/abs/2005.09070

Khan, S. S., & Abedi, A. 2022, Step Counting with Attention-based LSTM. https://arxiv.org/abs/2211.13114

Koffman, L., Crainiceanu, C., & III, J. M. 2024, Comparing Step Counting Algorithms for High-Resolution Wrist Accelerometry Data in NHANES 2011-2014, doi: https://doi.org/10.1249/MSS.0000000000003616

Koffman, L., III, J. M., & Crainiceanu, C. 2025, Walking Fingerprinting Using Wrist Accelerometry During Activities of Daily Living in NHANES. https://arxiv.org/abs/2506.17160

Kohankhaki, F., Raza, S., Bamgbose, O., Pandya, D., & Dolatabadi, E. 2024, Practical Guide for Causal Pathways and Sub-group Disparity Analysis. https://arxiv.org/abs/2407.02702

Kolakowski, M. 2024, Utilizing acceleration measurements to improve TDOA based localization, doi: https://doi.org/10.1109/SPS.2017.8053694

Lee, H., & Nadeem, M. 2025, Toward Efficient Speech Emotion Recognition via Spectral Learning and Attention. https://arxiv.org/abs/2507.03251

Mavali, S., Ricker, J., Pape, D., Fischer, A., & Schönherr, L. 2025, Adversarial Robustness of AI-Generated Image Detectors in the Real World. https://arxiv.org/abs/2410.01574

Mridula, D. T., Ferdaus, A. A., & Pias, T. S. 2023, Exploring Emotions in EEG: Deep Learning Approach with Feature Fusion. https://arxiv.org/abs/2311.10155

Peng, C., & Dinçer, T. 2024, Event Detection via Probability Density Function Regression. https://arxiv.org/abs/2408.12792

Pirinen, P., Klein, P., Lahme, S. Z., et al. 2023, Learning digital signal processing using an interactive Jupyter notebook and smartphone accelerometer data. https://arxiv.org/abs/2306.08436

R, A. P., Bhattacharyya, A., Vaughan, J., & Nair, V. N. 2024, Assessing Robustness of Machine Learning Models using Covariate Perturbations. https://arxiv.org/abs/2408.01300

Ram, A., S., S. S. V., Keshari, S., & Jiang, Z. 2023, Annotating sleep states in children from wrist-worn accelerometer data using Machine Learning. https://arxiv.org/abs/2312.07561

Schwartz, D., Quinn, L., Fritz, N. E., et al. 2024, Detecting Daily Living Gait Amid Huntington's Disease Chorea using a Foundation Deep Learning Model. https://arxiv.org/abs/2412.11286

Shengwei, M., & Jianjie, L. 2018, Design of a PCIe Interface Card Control Software Based on WDF. https://arxiv.org/abs/1803.09052

Siegl, T., Coşkun, K., Hiller, B., et al. 2025, SubROC: AUC-Based Discovery of Exceptional Subgroup Performance for Binary Classifiers. https://arxiv.org/abs/2505.11283

Sigcha, L., Borzì, L., Pavón, I., et al. 2024, Improvement of Performance in Freezing of Gait detection in Parkinsons Disease using Transformer networks and a single waist worn triaxial accelerometer, doi: https://doi.org/10.1016/j.engappai.2022.105482

Stern, R. H. 2021, Interpretation of the Area Under the ROC Curve for Risk Prediction Models. https://arxiv.org/abs/2102.11053

Suibkitwanchai, K., Sykulski, A. M., Algorta, G. P., Waller, D., & Walshe, C. 2020, Nonparametric Time Series Summary Statistics for High-Frequency Accelerometry Data from Individuals with Advanced Dementia, doi: https://doi.org/10.1371/journal.pone.0239368

Togootogtokh, E., & Klasen, C. 2021, DeepEMO: Deep Learning for Speech Emotion Recognition. https://arxiv.org/abs/2109.04081

Waks, Z., Mazeh, I., Admati, C., et al. 2017, Wrist Sensor Fusion Enables Robust Gait Quantification Across Walking Scenarios. https://arxiv.org/abs/1711.06974

Wang, X., Mishra, V. K., & Kuo, C. C. J. 2023, Enhancing Edge Intelligence with Highly Discriminant LNT Features. https://arxiv.org/abs/2312.14968

Wang, Y., & Zhao, Y. 2025, HEROS-GAN: Honed-Energy Regularized and Optimal Supervised GAN for Enhancing Accuracy and Range of Low-Cost Accelerometers. https://arxiv.org/abs/2502.18064

Williamson, J. R., Alini, A., Telfer, B. A., Potter, A. W., & Friedl, K. E. 2025, Estimating Visceral Adiposity from Wrist-Worn Accelerometry. https://arxiv.org/abs/2506.09167

Wu, M., Lin, L., Zhang, W., et al. 2025, Preserving AUC Fairness in Learning with Noisy Protected Groups. https://arxiv.org/abs/2505.18532

Zhang, A. H., He-Mo, A., Yin, R. F., et al. 2025a, Mamba-based Deep Learning Approaches for Sleep Staging on a Wireless Multimodal Wearable System without Electroencephalography. https://arxiv.org/abs/2412.15947

Zhang, Y., Ayush, K., Qiao, S., et al. 2025b, SensorLM: Learning the Language of Wearable Sensors. https://arxiv.org/abs/2506.09108