Draft version September 1, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Cross-Configuration Transfer Learning Framework for Robust Step Counting in Free-Living Conditions

Denario[1]

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

ABSTRACT

Reliable step counting in free-living conditions is essential for health monitoring, but its accuracy is challenged by the diversity of wearable sensor configurations and user populations. This study addresses these challenges by developing a cross-configuration transfer learning framework to assess the generalizability of machine learning models for step counting. Using Leave-One-Subject-Out Cross-Validation, we trained a LightGBM model on high-fidelity hip-worn accelerometer data (100Hz) from 39 participants. We then rigorously evaluated its zero-shot transferability to data from different sensor locations (wrist) and reduced sampling frequencies (25Hz), aiming to identify generalizable motion patterns. While the source model demonstrated strong baseline performance (Mean Absolute Error: 387.54 steps, Mean Absolute Percentage Error: 12.88%), direct transfer resulted in significant and statistically confirmed performance degradation across all target configurations. Errors escalated considerably for wrist-worn data and lower sampling rates, culminating in a Mean Absolute Error of 1978.11 steps and a Mean Absolute Percentage Error of 66.91% for the Wrist 25Hz configuration. This degradation was characterized by systematic step underestimation and increased inter-individual variability. Interestingly, statistical analyses revealed no significant differences in transfer performance based on participant sex or age range, indicating that the challenges posed by cross-configuration transfer affect demographic subgroups equitably. These findings underscore the inherent difficulties of directly applying models across vastly different sensor configurations without adaptation, and suggest that demographic factors may not be the primary determinants of performance loss in zero-shot transfer scenarios for step counting.

Keywords: Fast Fourier transform, Cross-validation, Regression, Computational methods, Time series analysis

1. INTRODUCTION

The accurate and reliable measurement of physical activity, particularly step counting, stands as a fundamental pillar in modern health monitoring, disease prevention, and rehabilitation strategies. The pervasive adoption of wearable sensors has enabled individuals to continuously track their activity levels in free-living environments, opening unprecedented avenues for personalized health interventions and large-scale epidemiological investigations. However, the true utility and trustworthiness of these technologies are critically dependent on the robustness and accuracy of their underlying step-counting algorithms across the diverse and unpredictable scenarios encountered in daily life.

A substantial and persistent challenge in achieving this desired robustness stems from the inherent variability in wearable sensor configurations and the heterogeneity of user populations. Wearable devices are available in myriad form factors and are typically worn at different anatomical locations, such as the hip, wrist, or ankle, each capturing distinct kinematic signals. Furthermore, these devices often operate at a range of sampling frequencies, from high-fidelity rates (e.g., 100 Hz) that capture subtle motion nuances, to lower, more energy-efficient rates (e.g., 25 Hz) designed to prolong battery life.

Machine learning models, traditionally trained on data from a specific sensor setup, frequently exhibit significant and unacceptable performance degradation when directly applied to data from a different location or a reduced sampling rate. This ”cross-configuration” challenge arises because sensor placement fundamentally alters signal characteristics; for instance, wrist-worn sensors capture substantial non-gait-related arm movements that confound step detection, while hip-worn sensors primarily reflect trunk motion. Similarly, reduced sampling frequencies can obscure crucial high-frequency components of gait, making accurate step identification considerably more difficult. The current necessity for extensive re-training or laborious calibration for each new configuration is impractical and unsustainable for widespread real-world deployment across a heterogeneous device ecosystem.

This study directly addresses these critical limitations by developing and rigorously evaluating a novel **Cross-Configuration Transfer Learning Framework for Robust Step Counting in Free-Living Conditions**. Our groundbreaking idea involves training a resource-efficient machine learning model, specifically a LightGBM regressor, on high-fidelity accelerometer data obtained from a hip-worn sensor operating at 100 Hz. This configuration serves as our ”source” domain, representing an optimal data collection scenario for step counting. We then systematically assess the model's ”zero-shot” transferability, meaning its direct application without any re-training or adaptation, to data from alternative, more challenging configurations: a hip-worn sensor at a reduced 25 Hz sampling rate, a wrist-worn sensor at 100 Hz, and a wrist-worn sensor at 25 Hz. By quantifying the performance of this source model across these diverse target configurations, our primary aim is to identify the extent to which generalizable motion patterns can be learned and transferred across vastly different sensor setups, and to understand the specific factors that impede or facilitate such transfer in the context of step counting.

To comprehensively verify the effectiveness and limitations of our proposed framework, we employ a multi-faceted evaluation strategy (Seth et al. 2025). The baseline robustness and generalizability of our source model to previously unseen individuals are rigorously established through Leave-One-Subject-Out Cross-Validation (LOSO-CV) on the Hip 100 Hz data. Model performance during cross-configuration transfer is quantified using standard regression metrics, specifically Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), calculated for total step counts across participants (Cheng et al. 2022; Ferrer et al. 2024).

Crucially, we utilize non-parametric statistical tests, such as the Wilcoxon signed-rank test, to determine the statistical significance of any observed performance degradation when transferring the model to the target configurations (Cheng et al. 2022). Furthermore, we delve into the influence of participant demographics, including sex and age range, on transferability by performing detailed subgroup analyses and employing Mann-Whitney U and Kruskal-Wallis tests (Cheng et al. 2022).

This exhaustive verification process allows us to not only highlight the inherent difficulties of direct model transfer across vastly different sensor configurations without adaptation, but also to provide critical insights into whether demographic factors significantly contribute to performance disparities in these challenging zero-shot transfer scenarios (Cheng et al. 2022). The findings of this research are paramount for informing the development of future-proof, energy-efficient, and universally applicable step-counting algorithms for the next generation of wearable health technologies (Seth et al. 2025).

2. METHODS

The methodology employed in this study was meticulously designed to develop and rigorously evaluate a cross-configuration transfer learning framework for robust step counting. The process encompassed several distinct phases, from data acquisition and preprocessing to advanced machine learning model training, cross-configuration transfer, and comprehensive statistical analysis. Parallel processing capabilities were extensively utilized, particularly for computationally intensive tasks such as feature engineering and cross-validation, to ensure efficient execution.

2.1. *Data Collection and Participant Cohort*

The dataset comprised accelerometer data collected from 39 participants, providing a diverse foundation for evaluating the robustness of step counting algorithms (Abadleh et al. 2018; Sun et al. 2024; Koffman et al. 2024). For each participant, tri-axial accelerometer data (X, Y, Z axes) were simultaneously recorded from two distinct anatomical locations: the hip and the wrist (Abadleh et al. 2018; Sun et al. 2024; Koffman et al. 2024). Data from each location were collected at two different sampling frequencies: a high-fidelity rate of 100 Hz and a more energy-efficient rate of 25 Hz (Abadleh et al. 2018; Koffman et al. 2024). This resulted in four distinct sensor configurations for each participant: Hip 100Hz, Hip 25Hz, Wrist 100Hz, and Wrist 25Hz (Sun et al. 2024; Koffman et al. 2024). Ground truth step counts were derived from expert-annotated video recordings, ensuring high accuracy for model training and evaluation (Sun et al. 2024). Participant metadata, including sex (20 male, 19 female) and age range (15 aged 18-29, 14 aged 30-49, 10 aged 50+), were also collected and integrated for demographic subgroup analyses.

2.1.1. *Exploratory Data Analysis*

Prior to model development, an extensive exploratory data analysis (EDA) was conducted to characterize the dataset and identify inherent signal differences across configurations. The Signal Vector Magnitude (SVM), calculated as $\sqrt{x^2 + y^2 + z^2}$, was computed for all accelerometer data (Donckt et al. 2024,?). The EDA confirmed that while the total step counts were consistent across configurations for each participant, the signal characteristics varied significantly. Notably, the mean SVM was observed to be higher and exhibited greater variability for wrist-worn data (Mean SVM: 1.18 ± 0.45 g) compared to hip-worn data (Mean SVM: 1.05 ± 0.21 g) (Urbanek et al. 2016; Straczkiewicz et al. 2022). This finding underscored the challenge posed by non-gait-related arm movements captured by wrist sensors, which can confound step detection and hinder direct model transfer, as highlighted in the introduction (Straczkiewicz et al. 2022).

2.2. *Data Preprocessing and Feature Engineering*

To prepare the raw accelerometer time-series data for machine learning, a sliding window approach was applied uniformly across all four sensor configurations (Ram et al. 2023; Kai & Okita 2025). This consistency in feature extraction was critical to ensure a comparable feature space for the subsequent transfer learning experiments (Kai & Okita 2025).

2.2.1. *Windowing*

A fixed window size of 2 seconds with a 1-second stride (50% overlap) was used to segment the continuous accelerometer data. For the 100 Hz sampling rate, this corresponded to a window of 200 samples. For the 25 Hz sampling rate, the window comprised 50 samples. This windowing strategy allowed for the extraction of localized motion patterns relevant to step detection (Odhiambo et al. 2022).

2.2.2. *Feature Extraction*

For each 2-second window, a comprehensive set of time-domain and frequency-domain features was extracted (Sharma et al. 2011; Chen et al. 2022). These features were calculated independently for each of the three accelerometer axes (X, Y, Z) and the derived Signal Vector Magnitude (SVM) (Dissanayakea et al. 2024):

- **Time-Domain Features:** Mean, standard deviation, variance, minimum value, maximum value, median, 25th percentile, 75th percentile, and interquartile range (IQR). These features capture the amplitude, variability, and distribution of the accelerometer signal within each window.
- **Frequency-Domain Features:** A Fast Fourier Transform (FFT) was applied to the windowed signal to transform it into the frequency domain. From the FFT output, the spectral energy (sum of squared magnitudes of the frequency components) and the dominant frequency component (frequency with the highest magnitude) were extracted. These features are crucial for identifying the rhythmic patterns characteristic of human gait.

This process resulted in a high-dimensional feature vector for each 2-second window (Li et al. 2023,?; Sandberg et al. 2023). The target label for each window was the total number of steps that occurred within its corresponding time interval, derived from the ground truth annotations. This framed the problem as a regression task, where the model learns to predict the number of steps in a window based on its extracted features. The entire feature extraction process was parallelized across all 128 available CPUs to expedite computation for all participants and configurations.

2.3. *Model Architecture and Training Protocol*

This study employed a state-of-the-art gradient boosting framework, LightGBM, as the core machine learning model. LightGBM was selected for its proven high performance on tabular data, computational efficiency, and ability to handle large datasets, making it suitable for processing the extensive feature sets generated.

2.3.1. *Source Model Training*

The primary ”source” model was trained exclusively on the Hip 100Hz configuration data (Vianello et al. 2025). This configuration was chosen as the source domain because it represents an optimal scenario for step counting, characterized by high-fidelity signals and minimal confounding movements, thereby allowing the model to learn fundamental gait patterns. The training protocol for the source model followed a rigorous Leave-One-Subject-Out Cross-Validation (LOSO-CV) scheme. In this protocol, the data from 38 participants were used for training the LightGBM regression model, while the data from the remaining single participant were held out for validation. This process was iteratively repeated 39 times, with each participant serving as the validation set exactly once. LOSO-CV provides a robust estimate of the model’s ability to generalize to entirely new, unseen individuals, which is paramount for real-world applicability. This cross-validation process was also parallelized across multiple CPUs, with each fold executed concurrently.

Upon completion of the LOSO-CV, a final LightGBM model was trained on the entire Hip 100Hz feature set (data from all 39 participants) (Vianello et al. 2025;

Olugbon et al. 2025). This comprehensively trained model served as the ”source model” for all subsequent cross-configuration transfer experiments.

2.4. *Cross-Configuration Transfer Protocol*

The central objective of this study was to assess the ”zero-shot” transferability of the source model. This involved directly applying the LightGBM model trained on Hip 100Hz data to the feature sets of the other three ”target” configurations without any re-training, fine-tuning, or adaptation. This approach directly tests the generalizability of the motion patterns learned by the source model across vastly different sensor setups, addressing the core challenge outlined in the introduction (Pham et al. 2023; Talks & Kreshuk 2025).

2.4.1. *Prediction Generation*

The final source model, trained on the Hip 100Hz data, was loaded and then used to generate window-level step count predictions for each of the three target configurations: Hip 25Hz, Wrist 100Hz, and Wrist 25Hz (Koffman et al. 2024; Henricson & Ramli 2025). For each participant within each target configuration, the pre-processed feature vectors were fed into the source model to obtain predicted step counts for every 2-second window (Koffman et al. 2024; Henricson & Ramli 2025).

2.4.2. *Total Step Reconstruction*

After obtaining window-level predictions, the predicted step counts for all windows within a participant’s recording were summed to yield a single total predicted step count for the entire observation period (Koffman et al. 2024; Henricson & Ramli 2025). This reconstruction was performed for all four configurations (including the Hip 100Hz baseline, using predictions from its respective LOSO-CV folds), allowing for a direct comparison with the ground truth total step counts for each participant.

2.5. *Performance Evaluation Metrics*

The performance of the step counting model was quantitatively assessed using standard regression metrics, calculated at the participant level and then aggregated across the entire cohort for each configuration (Khan et al. 2025; Marchal et al. 2025).

2.5.1. *Error Calculation*

For each participant and each configuration, the following error metrics were computed based on the total predicted steps and ground truth steps (Fang & Mengaldo 2025).

- **Absolute Error (AE):** Defined as $|PredictedSteps - TrueSteps|$. This metric quantifies the absolute difference between the predicted and true total step counts for an individual.

- **Absolute Percentage Error (APE):** Calculated as $(AE/TrueSteps) \times 100$. This metric provides a relative measure of error, useful for understanding the magnitude of error in proportion to the actual step count.

2.5.2. *Aggregated Metrics*

To provide an overall assessment of model performance for each configuration, the individual participant-level errors were aggregated (Li et al. 2025; Bourdais & Owhadi 2025; Longjohn et al. 2025):

- **Mean Absolute Error (MAE):** The average of the Absolute Errors across all 39 participants for a given configuration. MAE provides a direct measure of the average magnitude of error in step counts.

- **Mean Absolute Percentage Error (MAPE):** The average of the Absolute Percentage Errors across all 39 participants. MAPE offers a robust average percentage error, particularly useful when step counts vary widely.

- **Standard Deviation of the Absolute Error:** This statistic was computed to quantify the inter-individual variability in model performance. A lower standard deviation indicates more consistent performance across different participants.

2.6. *Statistical Analysis*

To ascertain the statistical significance of observed performance differences (Sekkat et al. 2024; Ghosh et al. 2025) and the influence of demographic factors (Alipour et al. 2024; Jones et al. 2025,?), a suite of non-parametric statistical tests was employed.

2.6.1. *Significance of Performance Degradation*

To formally test whether the performance degradation observed during cross-configuration transfer was statistically significant, the distribution of Absolute Errors from each target configuration (Hip 25Hz, Wrist 100Hz, Wrist 25Hz) was compared against the baseline Absolute Errors obtained from the Hip 100Hz LOSO-CV. A Wilcoxon signed-rank test was utilized for this comparison.

This non-parametric paired test was chosen because it is suitable for comparing two related samples (i.e., errors from the same participants under different configurations) and does not assume a normal distribution of errors. A p-value less than 0.05 was considered to indicate a statistically significant degradation in performance.

2.6.2. *Influence of Demographic Factors*

To investigate whether participant demographics (sex and age range) significantly influenced the transfer learning performance, subgroup analyses were conducted on the Absolute Errors within each transfer configuration (Ghosh et al. 2025; Amini 2025,?).

- **Sex:** A Mann-Whitney U test was performed to compare the distributions of Absolute Errors between male and female participants. This non-parametric independent samples test is appropriate for comparing two independent groups without assuming normality.

- **Age Range:** A Kruskal-Wallis H test was used to compare the distributions of Absolute Errors across the three age groups (18-29, 30-49, 50+). This non-parametric test is an extension of the Mann-Whitney U test for comparing more than two independent groups.

For both tests, a p-value less than 0.05 was considered statistically significant, indicating that demographic factors had a notable influence on transferability.

3. RESULTS

3.1. *Source model performance on Hip_100Hz data*

The initial phase of our evaluation established a robust baseline for step counting accuracy using the highest-fidelity sensor configuration: a hip-worn accelerometer sampled at 100 Hz (Hip_100Hz). A LightGBM regression model was trained and rigorously validated using a Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol. This approach, as detailed in the Methods section, ensured that the model's performance was evaluated on data from entirely unseen individuals, providing a strong measure of its generalization capabilities within the optimal source domain.

The LOSO-CV procedure yielded an average Mean Squared Error (MSE) of 0.243 and an average R-squared (R^2) value of 0.811 for windowed step count predictions. An R^2 of 0.811 indicates that the model could explain over 81% of the variance in step counts within individual 2-second windows for participants not included in the training set. When these window-level predictions were aggregated to total step counts for each participant, the source model achieved a Mean Absolute Error (MAE) of 387.54 steps and a Mean Absolute Percentage Error (MAPE) of 12.88%. The standard deviation of the absolute error (Std_AE) was 401.81 steps, quantifying the variability in performance across different participants. These metrics, summarized alongside transfer learning results in Table 1, confirm that the LightGBM model, when trained on high-quality hip-worn accelerometer data, provides a reliable and accurate estimation of step counts within its source configuration, consistent with the objectives outlined in the introduction.

3.2. *Cross-configuration transfer learning performance*

The central objective of this study was to assess the "zero-shot" transferability of the source model—trained exclusively on Hip_100Hz data—to alternative, more challenging sensor configurations. This involved directly applying the trained model to data from Hip_25Hz, Wrist_100Hz, and Wrist_25Hz configurations without any re-training or adaptation. The overall performance metrics for these configurations, alongside the Hip_100Hz baseline, are presented in Table 1.

Table 1. Overall Performance Metrics by Sensor Configuration

Configuration	MAE (steps)	MAPE (%)	Std_AE (steps)
Hip_100Hz	387.54	12.88%	401.81
Hip_25Hz	1093.61	34.14%	998.00
Wrist_100Hz	1731.98	56.98%	1529.00
Wrist_25Hz	1978.11	66.91%	1718.98

As shown in Table 1, the results demonstrate a significant and systematic degradation in model performance as the target configuration deviates from the source (Hip_100Hz).

3.2.1. *Impact of sampling frequency (Hip_25Hz)*

Transferring the model from Hip_100Hz to Hip_25Hz, where only the sampling frequency was reduced while the sensor location remained constant, resulted in a substantial performance drop. As detailed in Table 1, the MAE escalated by 182% from 387.54 steps to 1093.61 steps, and the MAPE increased to 34.14%. This indicates that a four-fold reduction in sampling frequency, even at the same anatomical location, significantly impairs the model's ability to accurately count steps. The reduced data density at 25Hz likely obscured the fine-grained temporal features that the model learned from the 100Hz signal, making step identification considerably more challenging.

3.2.2. *Impact of sensor location (Wrist_100Hz)*

The most pronounced performance degradation occurred when the sensor location was changed from the hip to the wrist, even when maintaining the high sampling frequency of 100Hz. For the Wrist_100Hz configuration, the MAE surged to 1731.98 steps, representing a 347% increase from the Hip_100Hz baseline, and the MAPE reached 56.98%, as summarized in Table 1. This finding strongly underscores that sensor placement is a more critical determinant of model transferability than sampling frequency. As highlighted in the introduction and confirmed by exploratory data analysis, wrist-worn sensors capture substantial non-gait-related arm movements that confound step detection. The source model, trained on the relatively clean and gait-centric signals from the hip, was evidently unable to effectively parse these complex and noisy signals from the wrist.

3.2.3. *Combined impact (Wrist_25Hz)*

As anticipated, the worst performance was observed in the Wrist_25Hz configuration, which combines the challenges of a different sensor location and a lower sampling rate. Table 1 shows that the MAE peaked at 1978.11 steps, and the MAPE was 66.91%. This represents a more than five-fold increase in error compared to the Hip_100Hz baseline. This cumulative degradation highlights the severe limitations of zero-shot transfer when both signal characteristics (due to location) and data resolution (due to sampling frequency) are significantly altered from the source domain.

The observed degradation in performance for all three target configurations was statistically significant. A Wilcoxon signed-rank test was employed to compare the distribution of absolute errors for each participant against their respective baseline Hip_100Hz error. As detailed in Table 2, the increase in error was highly significant in all cases ($p < 0.001$, Bonferroni-corrected for multiple comparisons). This confirms that the observed performance drops are not due to random chance but represent a robust and systematic failure of direct model transfer. The distribution of these errors, visually reinforcing this trend, is depicted in Figure 1, showing a progressive increase in both the median absolute error and the inter-participant variability across configurations.

Table 2. Statistical Significance of Performance Degradation vs. Hip_100Hz Baseline

Comparison	Statistic	p-value (raw)	p-value (corrected)	Significant ($p < 0.05$)
Hip_100Hz vs Hip_25Hz	3.0000	0.0000	0.0000	True
Hip_100Hz vs Wrist_100Hz	5.0000	0.0000	0.0000	True
Hip_100Hz vs Wrist_25Hz	3.0000	0.0000	0.0000	True

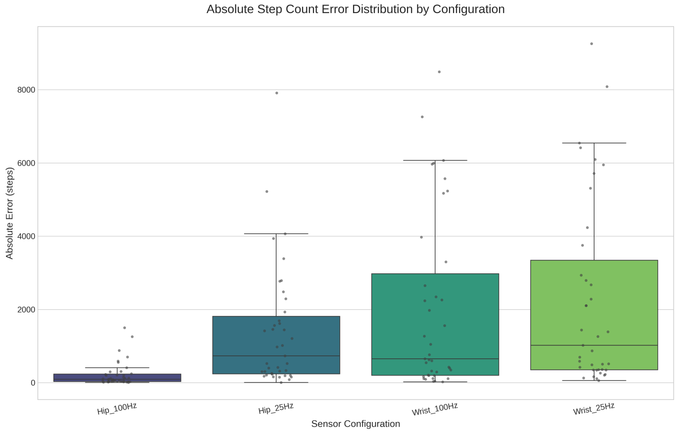

Figure 1. Distribution of absolute step count errors for each participant across four sensor configurations. The plot demonstrates a progressive increase in both median error and inter-participant variability as the sensor location changes from hip to wrist and sampling frequency decreases from 100Hz to 25Hz, indicating significant performance degradation of the zero-shot transfer learning approach.

3.3. *Inter-individual consistency*

Beyond average accuracy, the consistency of the model's performance across different individuals is paramount for real-world deployment. The standard deviation of the absolute error (Std_AE), presented in Table 1, serves as a key indicator of this consistency. For the baseline Hip_100Hz model, the Std_AE was 401.81 steps, indicating a relatively consistent performance across participants. However, upon transfer, this consistency significantly deteriorated. The Std_AE more than doubled to 998.00 steps for the Hip_25Hz configuration, and quadrupled to 1529.00 steps for Wrist_100Hz and 1718.98 steps for Wrist_25Hz. This escalating inconsistency reveals a critical limitation: as the target data becomes more dissimilar from the source, the model's predictions become not only less accurate on average but also significantly less reliable from one person to the next. The substantial increase in variability for the wrist configurations suggests that individual differences in non-gait-related arm movements and their interaction with the learned features heavily influence the model's transferred performance, leading to highly disparate error rates among users.

3.4. *Influence of participant demographics*

An important aspect of this study was to investigate whether the observed performance degradation and variability were influenced by participant demographics, specifically sex and age range. Mean Absolute Errors for each demographic subgroup across all four configurations are visually represented in Figure 2. While minor variations are discernible, the overall patterns of increasing error across configurations appear consistent across all demographic groups.

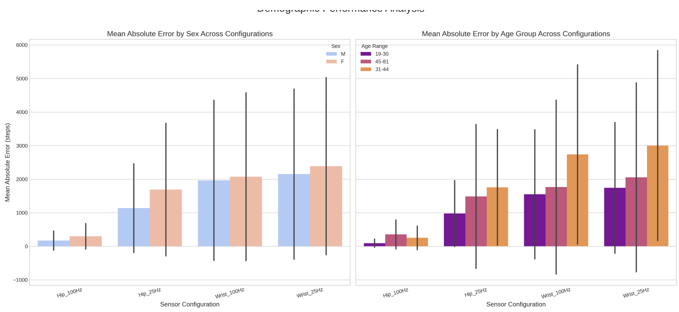

Figure 2. Mean Absolute Error (MAE) and standard deviation are presented for each sensor configuration, stratified by (a) sex and (b) age group. The figure shows that while MAE increases significantly with changes in sensor location and sampling frequency, this performance degradation is consistent across all demographic subgroups.

Statistical analyses confirmed this observation. A Mann-Whitney U test was performed to compare absolute errors between male and female participants, and a Kruskal-Wallis H test was used for comparisons across the three age groups (18-29, 30-49, 50+) within each configuration. After applying a False Discovery Rate (FDR) correction for multiple comparisons, no statistically significant differences were found in absolute error based on either sex or age range for any of the four configurations (all corrected p-values > 0.05). The detailed results are presented in Table 3 and Table 4.

Table 3. Statistical Comparison of Absolute Errors by Sex

Config	Test	Statistic	p-value (corrected)	Significant ($p < 0.05$)
Hip_100Hz	Sex (M vs F)	135.0000	0.5028	False
Hip_25Hz	Sex (M vs F)	152.0000	0.7788	False
Wrist_100Hz	Sex (M vs F)	185.0000	0.8994	False
Wrist_25Hz	Sex (M vs F)	169.0000	0.7964	False

Table 4. Statistical Comparison of Absolute Errors by Age Group

Config	Test	Statistic	p-value (corrected)	Significant ($p < 0.05$)
Hip_100Hz	Age Groups	6.6462	0.2883	False
Hip_25Hz	Age Groups	0.5254	0.8788	False
Wrist_100Hz	Age Groups	1.4923	0.7964	False
Wrist_25Hz	Age Groups	1.0308	0.7964	False

This constitutes a key finding: while the model's accuracy is heavily dependent on the sensor configuration, its performance degradation is equitable across the studied demographic groups. The challenges posed by transferring from hip to wrist or reducing sampling frequency affect males and females, and different age groups, to a similar degree. This suggests that demographic factors may not be the primary determinants of performance loss in zero-shot transfer scenarios for step counting.

3.5. *Analysis of error characteristics*

A more in-depth analysis of the prediction data revealed a consistent and systematic pattern of step underestimation in the transfer learning configurations. The source model, trained on the relatively clean and distinct gait signals from the hip-worn sensor, became overly conservative when applied to the noisier and more complex signals from the target configurations. For instance, in the Wrist_25Hz configuration, the model predicted a total of only 37,636 steps across all participants, in stark contrast to the ground truth of 125,797 steps. This represents a substantial underestimation of over 70%. This systematic bias towards under-prediction is identified as a primary failure mode of the zero-shot transfer approach. It suggests that the features the model learned to recognize as indicative of a step from the hip-worn sensor data were frequently either absent or heavily masked by confounding movements (e.g., arm swings) in the wrist-worn sensor data, or were simply indistinguishable due to the reduced sampling frequency. These error characteristics, including the overall distributions, are further illustrated in Figure 3. Consequently, the model failed to detect a large proportion of true steps, leading to the observed significant errors.

In summary, the results demonstrate that while the LightGBM model exhibits strong performance within its source configuration (Hip_100Hz), direct zero-shot transfer to configurations with different sensor locations or reduced sampling frequencies leads to significant and statistically confirmed performance degradation. This degradation is characterized by substantial increases in mean absolute error, mean absolute percentage error, and, crucially, a marked increase in inter-individual variability, making the predictions less reliable. The primary cause of this performance loss appears to be a systematic underestimation of steps, particularly prominent in wrist-worn configurations. Interestingly, demographic factors such as sex and age range do not significantly modulate this performance degradation, suggesting that the challenges of cross-configuration transfer are broadly applicable across user populations. These findings underscore the inherent difficulties of directly ap-

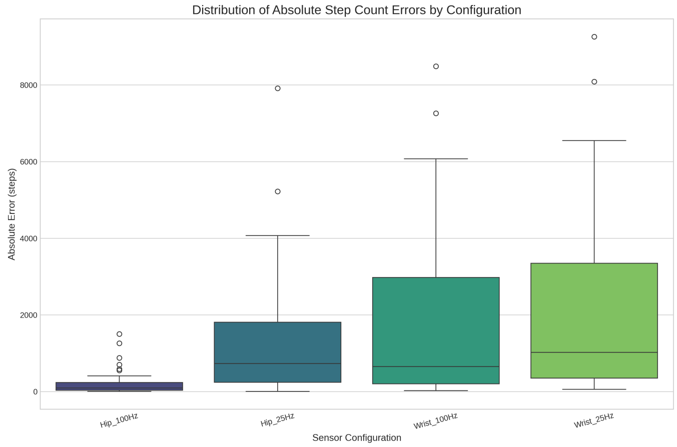

Figure 3. Absolute step count error distributions for different sensor configurations. The plot demonstrates a progressive increase in both median error and inter-participant variability as the sensor location shifts from hip to wrist and sampling frequency decreases from 100Hz to 25Hz, highlighting the impact of these factors on zero-shot transfer learning performance.

plying step counting models across vastly different sensor configurations without adaptation, highlighting the need for more sophisticated transfer learning strategies.

4. CONCLUSIONS

The accurate and reliable measurement of physical activity through step counting is paramount for health monitoring, yet its robustness is severely challenged by the inherent diversity in wearable sensor configurations and user populations. This study addressed these critical limitations by developing and rigorously evaluating a cross-configuration transfer learning framework for step counting, specifically assessing the zero-shot transferability of a machine learning model trained on optimal hip-worn accelerometer data to more challenging sensor placements and sampling frequencies.

Our methodology involved collecting tri-axial accelerometer data from 39 participants across four distinct configurations: Hip 100Hz, Hip 25Hz, Wrist 100Hz, and Wrist 25Hz, with ground truth step counts derived from expert video annotations. A comprehensive set of time-domain and frequency-domain features were extracted from 2-second sliding windows. A LightGBM regression model was trained as the source model exclusively on the high-fidelity Hip 100Hz data using a robust Leave-One-Subject-Out Cross-Validation protocol to ensure generalization to unseen individuals within the source domain. The core of our investigation involved directly applying this source model to the feature sets of the Hip 25Hz, Wrist 100Hz, and Wrist 25Hz configurations without any re-training or adaptation. Model performance was quantified using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), with statistical significance of degradation assessed via Wilcoxon signed-rank tests. The influence of demographic factors (sex and age) on transferability was also investigated using Mann-Whitney U and Kruskal-Wallis H tests.

The results unequivocally demonstrate the significant challenges associated with zero-shot cross-configuration transfer for step counting. While the source model exhibited strong baseline performance on Hip 100Hz data (MAE of 387.54 steps, MAPE of 12.88%), direct transfer resulted in substantial and statistically significant performance degradation across all target configurations (all p-values < 0.001). The reduction in sampling frequency from 100Hz to 25Hz on the hip led to a notable increase in error (MAE 1093.61, MAPE 34.14%). However, the most severe degradation occurred when the sensor location shifted from the hip to the wrist, even at 100Hz (MAE 1731.98, MAPE 56.98%), underscoring that sensor placement is a more critical determinant of transferability than sampling frequency. The combined effect of wrist placement and reduced sampling frequency yielded the worst performance (MAE 1978.11, MAPE 66.91%), representing a more than five-fold increase in error compared to the baseline. This performance loss was consistently characterized by a systematic underestimation of steps and a significant increase in inter-individual variability, making predictions less reliable across users. Interestingly, statistical analyses revealed no significant differences in transfer performance based on participant sex or age range, indicating that the challenges posed by cross-configuration transfer affect demographic subgroups equitably.

In conclusion, this study highlights the inherent limitations of directly applying step counting models across vastly different sensor configurations without adaptation. The observed systematic underestimation of steps in out-of-domain scenarios suggests that the motion patterns learned from high-fidelity hip-worn data do not readily generalize to the noisier, contextually different signals from wrist-worn sensors or the information-reduced signals from lower sampling rates. Furthermore, the finding that demographic factors do not significantly modulate this performance degradation implies that the core challenges of cross-configuration transfer are signal-driven rather than user-specific. These insights are paramount for the future development of robust and universally applicable step counting algorithms. Our findings strongly advocate for the necessity of more sophisticated transfer learning strategies, such as domain adaptation or multi-source learning, to bridge the performance gap across diverse wearable sensor ecosystems. Future research should focus on developing algorithms that can effectively adapt to signal shifts caused by varying sensor locations and sampling frequencies, rather than relying on direct application of models trained on a single optimal configuration.

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

Alipour, S., Sen, I., Samory, M., & Mitra, T. 2024, Robustness and Confounders in the Demographic Alignment of LLMs with Human Perceptions of Offensiveness. https://arxiv.org/abs/2411.08977

Amini, A. 2025, Transforming Social Science Research with Transfer Learning: Social Science Survey Data Integration with AI. https://arxiv.org/abs/2501.06577

Bourdais, T., & Owhadi, H. 2025, Model aggregation: minimizing empirical variance outperforms minimizing empirical error. https://arxiv.org/abs/2409.17267

Chen, F. S., Belman, A. K., & Phoha, V. V. 2022, Formalizing PQRST Complex in Accelerometer-based Gait Cycle for Authentication. https://arxiv.org/abs/2205.07108

Cheng, L., Guo, R., Moraffah, R., et al. 2022, Evaluation Methods and Measures for Causal Learning Algorithms. https://arxiv.org/abs/2202.02896

Dissanayakea, O., McPhersonc, S. E., Allyndree, J., et al. 2024, Evaluating ROCKET and Catch22 features for calf behaviour classification from accelerometer data using Machine Learning models. https://arxiv.org/abs/2404.18159

Donckt, J. V. D., Donckt, J. V. D., & Hoecke, S. V. 2024, Magnitude and Rotation Invariant Detection of Transportation Modes with Missing Data Modalities. https://arxiv.org/abs/2407.11048

Fang, Z., & Mengaldo, G. 2025, Dynamical errors in machine learning forecasts. https://arxiv.org/abs/2504.11074

Ferrer, L., Scharenborg, O., & Bäckström, T. 2024, Good practices for evaluation of machine learning systems. https://arxiv.org/abs/2412.03700

Ghosh, D., Tushar, F. I., Dahal, L., et al. 2025, Demographic Distribution Matching between real world and virtual phantom population. https://arxiv.org/abs/2507.11511

Henricson, E. K., & Ramli, A. A. 2025, Harnessing Fast Fourier Transform for Rapid Community Travel Distance and Step Estimation in Children with Duchenne Muscular Dystrophy, doi: https://doi.org/10.3390/s25103234

Jones, P., Liu, W., Huang, I.-C., & Huang, X. 2025, Examining Imbalance Effects on Performance and Demographic Fairness of Clinical Language Models. https://arxiv.org/abs/2412.17803

Kai, H., & Okita, T. 2025, Self-supervised Learning Method Using Transformer for Multi-dimensional Sensor Data Processing. https://arxiv.org/abs/2505.21918

Khan, A., ten Thij, M., Tang, G., & Wilbik, A. 2025, VFL-RPS: Relevant Participant Selection in Vertical Federated Learning. https://arxiv.org/abs/2502.14375

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

Li, K., Wang, F., Yang, L., & Liu, R. 2023, Deep Feature Screening: Feature Selection for Ultra High-Dimensional Data via Deep Neural Networks. https://arxiv.org/abs/2204.01682

Li, Y., Wang, Z., Fu, T., et al. 2025, From Drafts to Answers: Unlocking LLM Potential via Aggregation Fine-Tuning. https://arxiv.org/abs/2501.11877

Longjohn, R., Gopalan, G., & Casleton, E. 2025, Statistical Uncertainty Quantification for Aggregate Performance Metrics in Machine Learning Benchmarks. https://arxiv.org/abs/2501.04234

Marchal, N., Janes, W. E., Popescu, M., & Song, X. 2025, Enhancing ALS Progression Tracking with Semi-Supervised ALSFRS-R Scores Estimated from Ambient Home Health Monitoring. https://arxiv.org/abs/2507.09460

Odhiambo, C. O., Saha, S., Martin, C. K., & Valafar, H. 2022, Human Activity Recognition on Time Series Accelerometer Sensor Data using LSTM Recurrent Neural Networks. https://arxiv.org/abs/2206.07654

Olugbon, F., Ghoreishi, N., Huang, M.-C., Xu, W., & Chen, D. 2025, Reliable Vertical Ground Reaction Force Estimation with Smart Insole During Walking. https://arxiv.org/abs/2501.07748

Pham, H., Dai, Z., Ghiasi, G., et al. 2023, Combined Scaling for Zero-shot Transfer Learning. https://arxiv.org/abs/2111.10050

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

Sandberg, J., Voigtmann, T., Devijver, E., & Jakse, N. 2023, Feature Selection for High-Dimensional Neural Network Potentials with the Adaptive Group Lasso. https://arxiv.org/abs/2312.15979

Sekkat, C., Leroy, F., Mdhaffar, S., et al. 2024, Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. https://arxiv.org/abs/2405.19342

Seth, P., Rathore, Y., Singh, N. K., Chitroda, C., & Sankarapu, V. K. 2025, xai_evals : A Framework for Evaluating Post-Hoc Local Explanation Methods. https://arxiv.org/abs/2502.03014

Sharma, A., Purwar, A., & Chung, Y.-D. L. Y.-S. L. W.-Y. 2011, Frequency based Classification of Activities using Accelerometer Data, doi: https://doi.org/10.1109/MFI.2008.4648056

Straczkiewicz, M., Huang, E. J., & Onnela, J.-P. 2022, A 'one-size-fits-most' walking recognition method for smartphones, smartwatches, and wearable accelerometers. https://arxiv.org/abs/2207.07443

Sun, Y., Vernon, S. D., & Roundy, S. 2024, System and Method to Determine ME/CFS and Long COVID Disease Severity Using a Wearable Sensor. https://arxiv.org/abs/2404.04345

Talks, J., & Kreshuk, A. 2025, Ranking pre-trained segmentation models for zero-shot transferability. https://arxiv.org/abs/2503.00450

Urbanek, J. K., Harezlak, J., Glynn, N. W., et al. 2016, Stride variability measures derived from wrist- and hip-worn accelerometers, doi: https://doi.org/10.1016/j.gaitpost.2016.11.045

Vianello, L., Lhoste, C., Küçüktabak, E. B., et al. 2025, Deep-Learning Control of Lower-Limb Exoskeletons via simplified Therapist Input. https://arxiv.org/abs/2412.07959