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

Quantifying and Characterizing Step Counting Uncertainty in Wearable Accelerometer Data

Denario[1]

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

ABSTRACT

Traditional step counting accuracy metrics often fail to capture the critical aspects of measurement uncertainty and reliability, which are paramount for dependable health monitoring in free-living environments. This paper introduces a novel framework to explicitly quantify and characterize step counting uncertainty across diverse wearable accelerometer configurations, addressing the crucial trade-offs between data acquisition resources and measurement dependability. We developed a probabilistic 1D Convolutional Neural Network (CNN) that outputs the rate parameter of a Poisson distribution, allowing direct estimation of prediction confidence. The model was rigorously evaluated using Leave-One-Subject-Out cross-validation on a dataset of 39 participants, analyzing triaxial accelerometer data from hip and wrist placements at 100Hz and 25Hz sampling frequencies. Performance was assessed using Mean Absolute Error, Mean Absolute Percentage Error, bias, and by characterizing error types (false positives and false negatives), alongside the width of the 95% prediction confidence interval as our primary uncertainty metric. Our results demonstrate that hip-worn sensors at 100Hz provided the most accurate and least uncertain step counts, exhibiting the lowest mean absolute error (155 steps) and prediction confidence interval width (136 steps). Statistical analyses revealed that wrist-worn sensors produced significantly more false positives and false negatives ($p < 0.002$) compared to hip sensors, and reducing sampling frequency to 25Hz significantly increased false positives for wrist data ($p=0.0007$) while hip-worn sensors showed no significant degradation. Furthermore, substantial inter-individual variability was observed, with wrist-worn data showing significant sex-specific biases ($p < 0.02$). This comprehensive analysis highlights the importance of quantifying uncertainty for robust step counting and provides critical insights into optimal sensor deployment and resource allocation for reliable activity monitoring.

Keywords: Poisson distribution, Dimensionality reduction, Cross-validation, Confidence interval, Non-parametric hypothesis tests

1. INTRODUCTION

The ubiquitous integration of wearable sensors into daily life has initiated a paradigm shift in personal health monitoring, offering unprecedented opportunities for objective and continuous tracking of physical activity. Among the myriad of physiological metrics derivable from these devices, step count stands as a foundational and universally recognized indicator of daily physical activity. Its critical role spans diverse applications, from assessing fitness levels and guiding chronic disease management to informing public health initiatives and empowering individuals to make healthier lifestyle choices. Consequently, ensuring the accuracy and, more importantly, the reliability of step counting is paramount for both clinical utility and personal empowerment.

However, prevailing methodologies for evaluating step counting performance predominantly rely on aggregate accuracy metrics, such as Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE). While these metrics provide a macro-level assessment of a system's overall performance, they inherently fall short in capturing the crucial nuances of measurement uncertainty and reliability. In dynamic, real-world, free-living environments, step counting is subject to considerable variability stemming from a complex interplay of factors: diverse individual gaits, varying anthropometrics, inconsistent sensor placements, a wide spectrum of activity types, and ubiquitous environmental noise. A step counting system, despite exhibiting a low average error, might still suffer from high variability or unreliable performance under specific, yet critical, conditions. This lack of transparency regarding measurement confi-

dence can erode user trust in the data and potentially lead to suboptimal health interventions. The fundamental challenge lies in developing robust algorithms that can generalize effectively across this vast and unpredictable spectrum of human movement and sensor deployment scenarios, particularly when confronted with resource constraints that necessitate trade-offs in data quality, such as reduced sampling frequencies. Therefore, moving beyond a single point estimate to quantifying the explicit confidence associated with a step count becomes indispensable for dependable and actionable health monitoring.

This paper introduces a novel framework specifically designed to explicitly quantify and characterize step counting uncertainty across diverse wearable accelerometer configurations. Our approach systematically moves beyond conventional accuracy metrics by providing a principled means to assess the *reliability* of step detection under various sensor placements (e.g., hip versus wrist) and sampling frequencies (e.g., 100 Hz versus 25 Hz) (Khan & Abedi 2022; Koffman et al. 2024).

At the core of our solution is a probabilistic deep learning model, a 1D Convolutional Neural Network (CNN), engineered to address this challenge. Unlike traditional models that output a deterministic step count prediction, our CNN is designed to output the rate parameter (λ) of a Poisson distribution. This innovative design allows for the direct estimation of prediction confidence, providing a statistically principled measure of uncertainty associated with each step count prediction (Khan & Abedi 2022). By modeling step counts as a probabilistic outcome, our framework inherently captures the stochasticity and potential errors in the measurement process. Furthermore, it directly addresses the crucial trade-offs between data acquisition resources and measurement dependability, offering critical insights into how reduced data fidelity impacts the certainty of step count estimates (Koffman et al. 2024).

To rigorously verify the efficacy and generalizability of our proposed framework, the model was extensively evaluated using a comprehensive dataset derived from 39 participants. Triaxial accelerometer data from both hip and wrist placements, recorded at 100 Hz and 25 Hz sampling frequencies, were subjected to a stringent Leave-One-Subject-Out (LOSO) cross-validation scheme. This robust evaluation methodology ensures that our findings are not merely specific to the training population but generalize effectively to unseen individuals. Performance was assessed using both traditional metrics, including Mean Absolute Error, Mean Absolute Percentage Error, and overall bias, complemented by a detailed characterization of specific error types: false positives (instances of over-counting steps) and false negatives (instances of missed or under-counted steps) (Soumma et al. 2025). Crucially, the primary metric for uncertainty quantification was defined as the width of the 95% prediction confidence interval, directly derived from the model's probabilistic Poisson output.

Comparative statistical analyses, including paired Wilcoxon signed-rank tests for within-subject comparisons and subgroup analyses based on sex and age using Mann-Whitney U and Kruskal-Wallis tests, were systematically employed to compare performance across different sensor configurations and to identify potential demographic variabilities. This comprehensive approach allows us to not only quantify how accurately steps are counted but, more importantly, to ascertain the confidence we can place in those counts, thereby providing critical insights into optimal sensor deployment strategies and resource allocation for robust and reliable activity monitoring in real-world settings (Waks et al. 2017; Koffman et al. 2025; Soumma et al. 2025).

2. METHODS

The methodology employed in this study was designed to rigorously quantify and characterize step counting uncertainty across various wearable accelerometer configurations. Our approach involved comprehensive data preparation, the development of a novel probabilistic deep learning model, a robust Leave-One-Subject-Out (LOSO) cross-validation scheme, detailed performance evaluation, and a thorough comparative statistical analysis.

2.1. *Data Acquisition and Preprocessing*

2.1.1. *Participant Cohort and Data Collection*

The dataset comprised triaxial accelerometer data collected from 39 participants. For each participant, data were acquired simultaneously from two distinct sensor placements: the hip and the wrist. Data were recorded at a high sampling frequency of 100 Hz. To investigate the impact of resource constraints on measurement dependability, the 100 Hz data were subsequently downsampled to 25 Hz, resulting in four distinct sensor configurations for analysis: Hip_100Hz, Wrist_100Hz, Hip_25Hz, and Wrist_25Hz. Ground truth step counts were obtained through manual annotation of video recordings, providing a precise temporal alignment of each step event.

2.1.2. *Data Loading and Structuring*

A dedicated script was developed to systematically load and structure the dataset. Initially, the 'metadata.csv' file was read to retrieve demographic informa-

tion for each participant, including age range and sex (van den Burg et al. 2018; Heibi et al. 2025).

Subsequently, the script iterated through the 39 participant directories (P01 to P39) within each of the four sensor configuration folders. For every participant, the corresponding time-series CSV files containing triaxial accelerometer data and ground truth annotations were loaded. This data was organized into a structured format, enabling seamless linkage of demographic data with the four raw signal files for each participant. Data integrity checks were performed to confirm the successful loading of all 156 expected files (39 participants × 4 configurations) (Ito 2024; Gordon et al. 2025).

2.1.3. *Exploratory Data Analysis (EDA)*

Prior to model development, an exploratory data analysis (EDA) was conducted to characterize the dataset. This analysis confirmed the participant demographics: 19 males and 20 females, with age distributions of 15 participants in the 18-29 age range, 14 in the 30-49 range, and 10 in the 50+ range.

The EDA also provided insights into data volume and step distribution across the configurations. The mean recording duration for all configurations was approximately 58.1-58.2 minutes, with a mean total of 2104-2105 steps per participant. The standard deviation of steps per participant was consistently around 850, indicating substantial inter-individual variability in activity levels (Karande et al. 2024). It was noted that step counts were identical across different sampling frequencies for the same location, as annotations were time-based (Vos et al. 2023).

2.1.4. *Data Segmentation*

To prepare the continuous time-series data for input into the deep learning model, a sliding window approach was implemented (Ji & Aydin 2024; Liu et al. 2024; Han et al. 2025). For 100 Hz data, a window size of 2.56 seconds (corresponding to 256 samples) was used. For 25 Hz data, the same 2.56-second window translated to 64 samples. A 50% overlap was applied across all configurations, meaning the window advanced by 1.28 seconds (128 samples for 100 Hz, 32 samples for 25 Hz) for each new segment.

For each generated window, the target label was defined as the integer count of ground truth steps occurring within that window's time frame (Ji & Aydin 2024; Liu et al. 2024). This process resulted in a set of (accelerometer window, step_count) pairs for every participant and configuration, serving as the input for model training.

2.2. *Probabilistic Deep Learning Model*

2.2.1. *Model Architecture*

A 1D Convolutional Neural Network (CNN) was developed using TensorFlow to probabilistically estimate step counts. The architecture was designed to be identical across all four sensor configurations, with the sole difference being the input layer dimension (256 samples for 100 Hz data, 64 samples for 25 Hz data). The model architecture consisted of: (Khan & Abedi 2022; Surma et al. 2023)

- **Input Layer:** Accepted a window of raw triaxial accelerometer data with a shape of '(window_size, 3)'.
- **Convolutional Layers:** Two 1D convolutional layers with ReLU activation functions were used. The first layer had 32 filters, and the second had 64 filters, both with a kernel size of 7. These layers were designed to extract hierarchical features from the raw time-series signal.
- **Pooling Layer:** A max-pooling layer followed the convolutional blocks to reduce dimensionality and enhance translational invariance of the learned features. **Flatten Layer:** This layer converted the pooled feature maps into a 1D vector, preparing the data for the subsequent fully-connected layers.
- **Dense Layers:** One fully-connected layer with ReLU activation processed the flattened features.
- **Output Layer:** A single neuron with a 'softplus' activation function. This crucial choice ensured that the output, representing the rate parameter (λ) of a Poisson distribution, was always positive. This λ value directly corresponded to the model's predicted expected step count for the given input window, providing a probabilistic estimate rather than a deterministic one.

2.2.2. *Loss Function*

The model was trained by minimizing the negative log-likelihood of the true step count under the predicted Poisson distribution (Terven et al. 2025). For a single window with a true step count y and a predicted rate λ, the loss function was defined as:

$$L(y, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (\lambda_i - y_i \log(\lambda_i))$$

(Terven et al. 2025)

$$\text{Loss}(y, \lambda) = -(y \cdot \log(\lambda) - \lambda - \log(y!))$$

This loss function, commonly available as 'PoissonNLLLoss' in deep learning frameworks, allowed the model

to learn the parameters of a distribution from which the observed step counts were likely to have been drawn, thereby inherently quantifying the uncertainty of its predictions (Abdelkhalik et al. 2023).

2.3. *Model Training and Validation*

2.3.1. *Leave-One-Subject-Out (LOSO) Cross-Validation*

To ensure the robustness and generalizability of our findings to unseen individuals, a stringent Leave-One-Subject-Out (LOSO) cross-validation strategy was employed (Avelin & Viitasaari 2023; Weese et al. 2025). This procedure involved 39 separate training and testing iterations. In each iteration i, data from participant P_i was designated as the test set, while data from the remaining 38 participants formed the training set. This ensured that the model's performance was evaluated on data from individuals completely external to its training experience.

2.3.2. *Training Procedure*

Within each of the 39 LOSO folds, a separate model was trained from scratch for each of the four sensor configurations (Hip_100Hz, Wrist_100Hz, Hip_25Hz, Wrist_25Hz). This resulted in a total of 156 distinct models being trained throughout the experiment. To expedite this computationally intensive process, the 39 folds were configured to run concurrently, leveraging 128 available CPU cores for parallelization. Models were trained for a fixed number of epochs with an Adam optimizer, and early stopping was implemented based on validation loss to prevent overfitting.

2.4. *Evaluation and Uncertainty Quantification*

After training, the performance of each model was evaluated on its respective held-out test set (i.e., the data from the participant not included in training) (Chandrasekaran et al. 2023; Huang et al. 2024; Beddar-Wiesing et al. 2025).

2.4.1. *Step Count Reconstruction*

The model generated a sequence of predicted Poisson rates $(\lambda_1, \lambda_2, \lambda_3, \dots)$ for each overlapping window in a participant's recording (Khan & Abedi 2022; Das et al. 2024). To reconstruct the total predicted step count ($\hat{Y}$) for the entire recording period, the sum of these rates was calculated (Khan & Abedi 2022).

Given the 50% overlap, each predicted λ_i effectively represents the expected step count within its unique contribution to the time series (Fu et al. 2022; Das et al. 2024). Therefore, the total predicted step count $\hat{Y}$ was computed as the sum of all λ_i values divided by the overlap factor (2), i.e., $\hat{Y} = \sum_i \lambda_i/2$ (Fu et al. 2022; Leppich et al. 2025). This method effectively accounts for the contribution of each window to the overall step count while avoiding double-counting due to overlap (Fu et al. 2022; Das et al. 2024).

2.4.2. *Performance Metrics*

The following metrics were calculated by comparing the reconstructed total predicted step count ($\hat{Y}$) with the true total step count (Y) for each of the 39 participants and then averaged across all participants for each sensor configuration (Pillai et al. 2020; Khan & Abedi 2022).

- **Mean Absolute Error (MAE):** The average of the absolute differences between true and predicted step counts, calculated as MAE = mean($|Y - \hat{Y}|$).
- **Mean Absolute Percentage Error (MAPE):** The average of the absolute percentage differences, calculated as MAPE = mean($|(Y - \hat{Y})/Y|$) $\times$ 100.
- **Bland-Altman Statistics:** Used to assess agreement and bias. This included the mean bias, calculated as mean($Y - \hat{Y}$), and the 95% limits of agreement, defined as mean_bias $\pm$ 1.96 $\times$ std($Y - \hat{Y}$).

2.4.3. *Uncertainty Quantification*

The central objective of this study was to quantify uncertainty. Our probabilistic model's output facilitated this directly. For each participant, the total predicted step count was approximated by a Poisson distribution with a total rate $\Lambda = \sum \lambda_i$ (using the same summation logic as for $\hat{Y}$) (Kivimäki et al. 2025a). The 95% prediction confidence interval (CI) for the total step count was then derived from this Poisson distribution $P(\Lambda)$ (Kivimäki et al. 2025b; Chang et al. 2025).

The primary metric for uncertainty was defined as the **width of this 95% CI** (Kivimäki et al. 2025b,a). A wider interval indicated greater uncertainty in the model's prediction for that participant. The average CI width was reported for each of the four sensor configurations (Kivimäki et al. 2025b,a).

2.4.4. *Error Characterization*

To gain a deeper understanding of the types of errors made by the models, two specific error metrics were calculated: (Botchkarev 2018; Naser & Alavi 2020; Fang & Mengaldo 2025).

- **False Positives (FP):** Represented instances of over-counting steps. For each participant, this was calculated by summing the predicted step counts ($\hat{Y}$) from all windows where the true step count (Y) was zero.

- **False Negatives (FN) / Missed Steps:** Represented true steps that the model failed to detect. For each participant, this was calculated at the window level as $\sum_i \max(0, y_i - \hat{\lambda}_i)$, where y_i is the true count and $\hat{\lambda}_i$ is the predicted Poisson rate for window i. These window-level missed steps were then summed to obtain a total for the participant.

2.5. *Comparative Statistical Analysis*

The final stage of the analysis involved a systematic statistical comparison of the performance metrics across the four sensor configurations and demographic subgroups (Yu et al. 2024; Zhang et al. 2025; Jiang et al. 2025).

2.5.1. *Within-Subject Comparisons*

To compare performance between different sensor placements and sampling frequencies while accounting for the paired nature of data (i.e., measurements from the same individuals), a **paired Wilcoxon signed-rank test** was performed for each key metric (MAE, MAPE, CI Width, FP, FN) (Couch et al. 2018; Fan et al. 2022; Howard & Pimentel 2024).

- **Hip vs. Wrist:** Comparisons were made between Hip_100Hz and Wrist_100Hz, and between Hip_25Hz and Wrist_25Hz.
- **100Hz vs. 25Hz:** Comparisons were made between Hip_100Hz and Hip_25Hz, and between Wrist_100Hz and Wrist_25Hz.

2.5.2. *Individual and Subgroup Analysis*

- **Inter-individual Variability:** The standard deviation of the performance metrics across participants was analyzed to comment on the extent of variability in model performance among different individuals.
- **Sex-Specific Biases:** Results were stratified by sex. The **Mann-Whitney U test** was used to compare all key metrics between male and female participants for each of the four sensor configurations, identifying potential sex-specific biases.
- **Age Group Differences:** Results were also stratified by the three age groups (18-29, 30-49, 50+). The **Kruskal-Wallis test** was employed to compare the key metrics across these age groups for each sensor configuration, assessing if certain age demographics were more susceptible to errors or uncertainty under specific conditions.

A significance level of $\alpha = 0.05$ was used for all statistical tests (Hemerik & Koning 2025).

All intermediate results, including windowed data, model predictions per fold, and final metrics per participant, were meticulously saved to a structured file system to facilitate debugging and further post-hoc analysis (Goldwasser & Hooker 2025). Confirmation messages were logged upon the successful completion of each major experimental stage.

3. RESULTS

This section presents the comprehensive evaluation of the probabilistic 1D Convolutional Neural Network (CNN) model developed for quantifying step counting uncertainty. The analysis was rigorously performed using a Leave-One-Subject-Out (LOSO) cross-validation methodology across 39 participants, evaluating four distinct sensor configurations: Hip at 100 Hz (Hip_100Hz), Hip at 25 Hz (Hip_25Hz), Wrist at 100 Hz (Wrist_100Hz), and Wrist at 25 Hz (Wrist_25Hz). The results are structured to provide an overview of the model's overall performance and its ability to quantify uncertainty, followed by detailed comparative statistical analyses examining the impact of sensor placement, sampling frequency, and demographic factors.

3.1. *Overall model performance and error analysis*

Before detailing the model's performance, it is important to characterize the nature of the step count data used for training and evaluation. As illustrated by the histograms in Figure 1, the distribution of true step counts within 2.56-second windows is highly consistent across all sensor types and sampling frequencies. A notable characteristic of this dataset is that most windows contain no steps, reflecting periods of inactivity or non-stepping movements.

The performance of the step counting model was assessed using traditional accuracy metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and overall bias ($TrueSteps - PredictedSteps$). In line with our objective to characterize uncertainty beyond aggregate metrics, we also quantified specific error types: False Positives (FP), representing instances where steps were predicted but none occurred (over-counting), and False Negatives (FN), representing true steps missed by the model (under-counting). A summary of these metrics, averaged across all 39 participants, is presented in Table 1.

As evident from Table 1, the `Hip_100Hz` configuration consistently demonstrated the most favorable performance across various metrics. It achieved the lowest Mean Absolute Error (155.29 steps) and Mean Absolute Percentage Error (10.29%), indicating superior accuracy compared to other configurations. This suggests

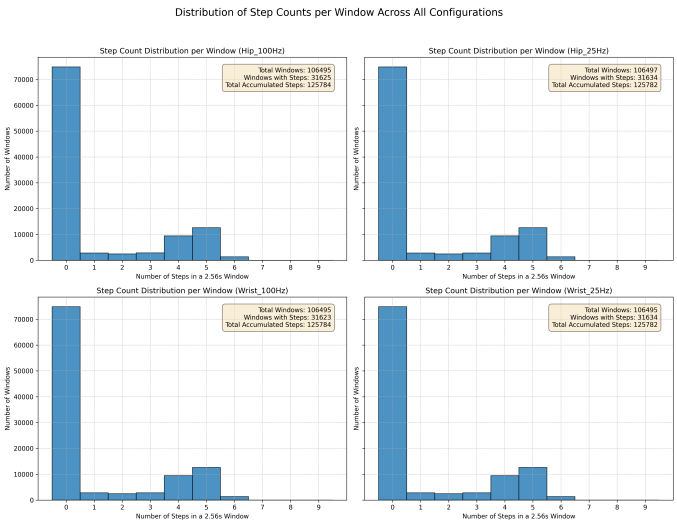

Figure 1. Histograms show the distribution of true step counts within 2.56-second windows for each sensor configuration. The data reveal that most windows contain no steps, and the overall step distribution per window is highly consistent across all sensor types and sampling frequencies.

Table 1. Summary of Performance and Error Metrics Across All Sensor Configurations (Mean ± Standard Deviation)

Configuration	MAE (steps)	MAPE (%)	Bias (steps)	FP (steps)	FN (steps)	95% CI Width (steps)
Hip_100Hz	155.29 ± 246.42	10.29 ± 6.70	58.44 ± 286.30	169.58 ± 179.90	467.42 ± 548.71	136.08 ± 74.32
Hip_25Hz	168.55 ± 217.10	11.93 ± 8.61	69.61 ± 267.05	197.14 ± 196.37	487.07 ± 539.07	136.15 ± 73.08
Wrist_100Hz	165.80 ± 211.42	16.61 ± 19.63	30.38 ± 268.26	286.35 ± 231.82	602.19 ± 516.51	137.82 ± 73.70
Wrist_25Hz	210.92 ± 236.08	21.66 ± 22.19	33.45 ± 316.61	387.22 ± 294.64	627.27 ± 612.91	139.54 ± 70.24

that the hip, being closer to the body's center of mass and exhibiting a more regular and distinct acceleration pattern during ambulation, provides a cleaner and more reliable signal for step detection, especially at a higher sampling rate.

Across all configurations, a positive mean bias was observed, indicating a general tendency for the models to under-count steps (i.e., $TrueSteps > PredictedSteps$). This systematic under-counting, though relatively small on average, suggests that the model is slightly more conservative in its step predictions than the ground truth. Interestingly, the `Wrist_100Hz` configuration exhibited the lowest mean bias (30.38 steps), suggesting that while it might have higher absolute errors, its systematic deviation from the true count is marginally less pronounced compared to hip-worn sensors.

The characterization of error types (False Positives and False Negatives) reveals critical insights into the nature of miscounts. Wrist-worn sensors consistently generated substantially more false positives and false negatives compared to hip-worn sensors. For example, the `Wrist_25Hz` model produced more than double the false positives (387.22 steps) of the `Hip_100Hz` model (169.58 steps). This indicates that wrist-worn devices are more susceptible to misinterpreting non-stepping movements (e.g., arm gestures, fidgeting) as steps, leading to overcounting, and simultaneously are more prone to missing actual steps due to the complex and often less distinct acceleration patterns at the wrist during ambulation. This finding directly supports the introduction's assertion that aggregate accuracy metrics alone are insufficient and that understanding error types is crucial for dependable health monitoring.

A noteworthy observation across all metrics in Table 1 is the high standard deviation, often exceeding the mean. This underscores a significant degree of inter-individual variability in model performance, visually depicted in Figure 2. While the average performance may appear acceptable, the wide distributions and numerous outliers highlight that the model's accuracy can vary dramatically from one participant to another. This emphasizes the challenge of developing universally robust step counting algorithms and reinforces the necessity of quantifying uncertainty for individual-level reliability.

3.2. *Uncertainty quantification*

A central objective of this study was to explicitly quantify prediction uncertainty. Our probabilistic 1D CNN model directly addresses this by outputting the rate parameter of a Poisson distribution, from which a 95% prediction confidence interval (CI) for the total step count can be derived. The width of this CI serves as our primary metric for uncertainty, with a wider interval indicating lower confidence in the model's prediction. The mean CI widths for each configuration are included in Table 1.

The average 95% CI width was narrowest for the `Hip_100Hz` configuration (136.08 steps), signifying the highest level of model confidence in its predictions for this setup. Conversely, the `Wrist_25Hz` configuration exhibited the widest CI (139.54 steps), indicating greater uncertainty. While the absolute differences in mean CI width across configurations appear relatively small, they consistently follow a pattern: uncertainty increases when transitioning from hip to wrist placement and when the sampling frequency is reduced. This trend aligns perfectly with the observed accuracy metrics, suggesting that the model's confidence in its predictions is highest when the input signal is clearest and most informative, i.e., high-frequency data from a stable anatomical location like the hip. The distributions of these CI widths across participants, as illustrated in

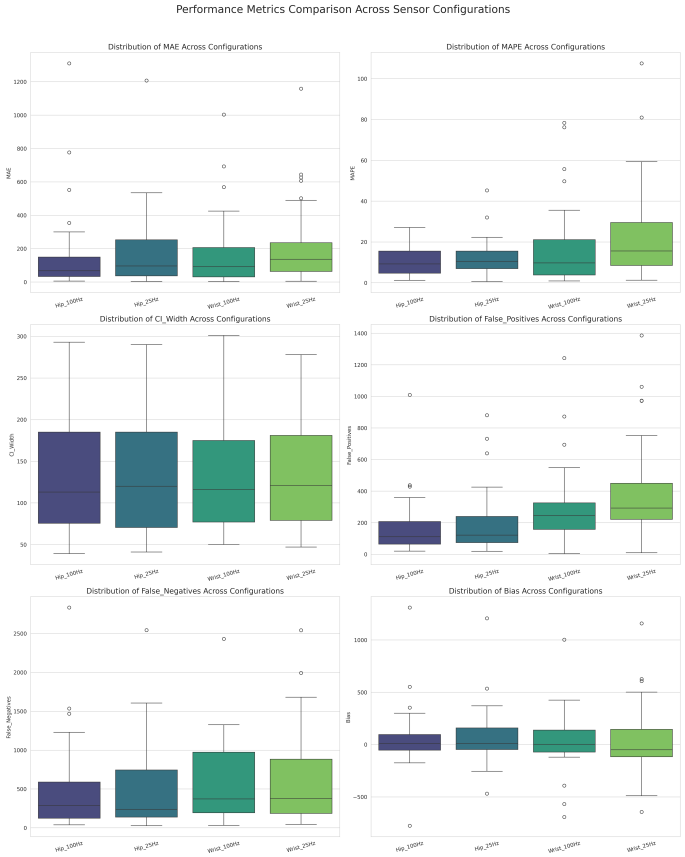

Figure 2. Boxplots displaying the distribution of key performance metrics (MAE, MAPE, Bias, CI Width, False Positives, False Negatives) across the four sensor configurations. These plots reveal significant inter-individual variability, characterized by wide distributions and outliers, and demonstrate that hip-worn sensors generally yield lower errors and less variability, particularly in false positive and false negative rates, compared to wrist-worn sensors.

the histograms in Figure 3, show a relatively consistent spread for a given configuration, though wrist-based measurements tend to exhibit slightly greater variability in uncertainty. This direct quantification of confidence provides critical context to the predicted step counts, enabling users and clinicians to better interpret the reliability of the data.

3.3. *Comparative statistical analysis*

To ascertain the statistical significance of the observed performance differences, non-parametric paired Wilcoxon signed-rank tests were conducted for within-subject comparisons.

3.3.1. *Impact of sensor placement (Hip vs. Wrist)*

Paired Wilcoxon signed-rank tests were performed to compare hip and wrist sensor performance at both 100 Hz and 25 Hz sampling frequencies. The results,

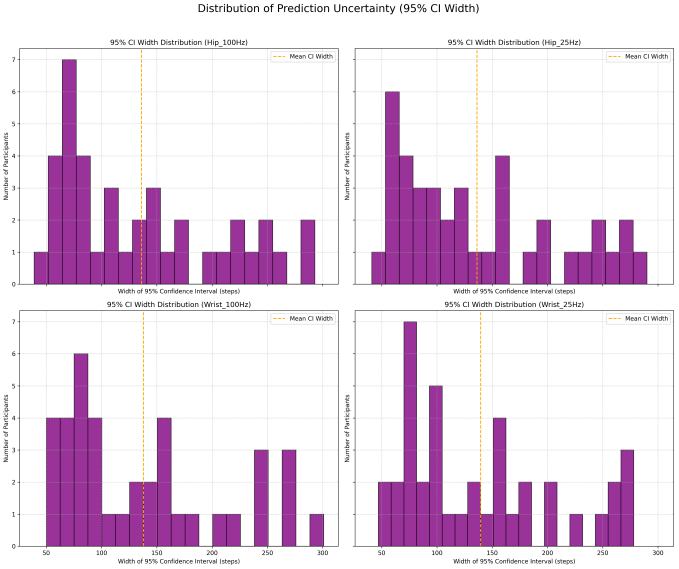

Figure 3. Histograms show the distribution of 95% confidence interval widths for total step count predictions across four sensor configurations. Wider intervals indicate greater model uncertainty. The plots demonstrate that prediction uncertainty is lowest for hip-worn sensors at 100Hz and increases for wrist-worn sensors and lower sampling frequencies, reflecting the model's confidence in its predictions.

summarized in Table 2, highlight the profound impact of sensor placement on error characteristics.

Table 2. Statistical Comparison of Sensor Placement (Hip vs. Wrist)

Metric	Comparison	Statistic
MAE	Hip_100Hz vs. Wrist_100Hz	335.0
	Hip_25Hz vs. Wrist_25Hz	292.0
MAPE	Hip_100Hz vs. Wrist_100Hz	297.0
	Hip_25Hz vs. Wrist_25Hz	226.0
False Positives	Hip_100Hz vs. Wrist_100Hz	110.0
	Hip_25Hz vs. Wrist_25Hz	68.0
False Negatives	Hip_100Hz vs. Wrist_100Hz	144.0
	Hip_25Hz vs. Wrist_25Hz	163.0

Results from Paired Wilcoxon Signed-Rank Test. Significant p-values (p

While the overall MAE did not show a statistically significant difference between hip and wrist placements at either frequency, the analysis of error types revealed compelling distinctions. Wrist-worn sensors produced statistically significantly more false positives and false negatives compared to hip-worn sensors at both 100 Hz and 25 Hz (all $p < 0.002$). This finding is crucial as it demonstrates that even if the average absolute error is comparable, the *nature* of the errors differs substan-

tially. Hip sensors offer a more reliable signal, being less prone to misinterpreting non-stepping movements (e.g., arm gestures, fidgeting) as steps (false positives) and more effective at detecting actual steps (fewer false negatives). This reinforces the notion that the hip provides a more robust signal for step detection, aligning with its anatomical position relative to gait mechanics. Furthermore, the MAPE for wrist data at 25 Hz was significantly higher ($p = 0.0213$), suggesting that the relative error becomes more pronounced for wrist-worn sensors under reduced data fidelity.

3.3.2. *Impact of sampling frequency (100Hz vs. 25Hz)*

The effect of reducing the sampling frequency was also rigorously assessed using paired Wilcoxon signed-rank tests, with results presented in Table 3.

Table 3. Statistical Comparison of Sampling Frequency (100Hz vs. 25Hz)

Metric	Comparison	Statistic
MAE	Hip_100Hz vs. Hip_25Hz	329.0
	Wrist_100Hz vs. Wrist_25Hz	277.0
MAPE	Hip_100Hz vs. Hip_25Hz	333.0
	Wrist_100Hz vs. Wrist_25Hz	257.0
False Positives	Hip_100Hz vs. Hip_25Hz	321.0
	Wrist_100Hz vs. Wrist_25Hz	155.0
False Negatives	Hip_100Hz vs. Hip_25Hz	362.0
	Wrist_100Hz vs. Wrist_25Hz	338.0

Results from Paired Wilcoxon Signed-Rank Test. Significant p-values ($p < 0.05$) are in bold.

For hip-worn sensors, reducing the sampling frequency from 100 Hz to 25 Hz did not result in any statistically significant degradation in performance across all evaluated metrics. This is a crucial finding, indicating that for hip-based step counting, a 25 Hz sampling rate is largely sufficient to maintain accuracy and reliability. This presents a viable option for reducing computational and power requirements in real-world applications without compromising measurement quality.

In contrast, for wrist-worn sensors, the reduction to 25 Hz led to a statistically significant increase in false positives ($p = 0.0007$). This suggests that the lower-frequency signal from the wrist, already prone to noise from non-stepping movements, becomes even more ambiguous at a reduced sampling rate, leading to a more pronounced over-counting issue. While the increase in MAPE for the wrist at 25 Hz did not reach statistical significance ($p = 0.064$), the trend further supports a performance degradation under resource-constrained conditions for wrist-based measurements. This highlights the critical trade-off between data acquisition resources and measurement dependability, especially for sensor placements with inherently noisier signals.

3.4. *Agreement and bias analysis*

Bland-Altman analysis was performed to further assess the agreement between the model's predictions and the true step counts, providing insights into the magnitude and consistency of bias. The Bland-Altman plots in Figure 4 visualize the mean bias and the 95% limits of agreement (LoA).

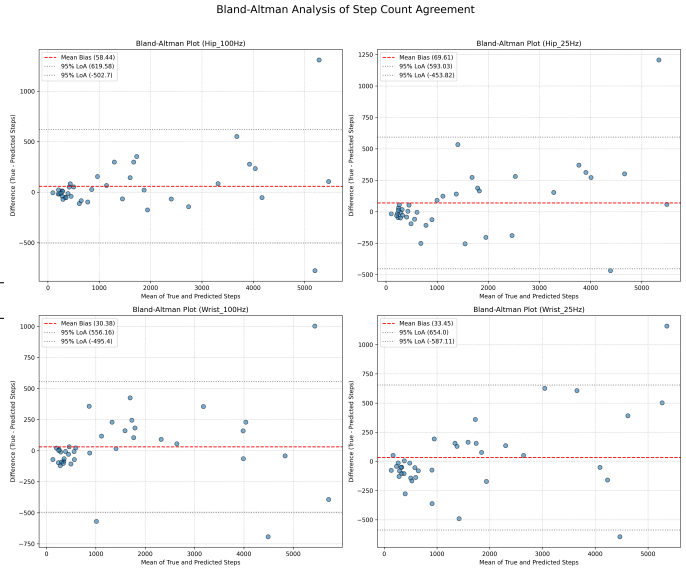

Figure 4. Bland-Altman plots comparing true and predicted step counts across sensor configurations. Each plot shows the difference ($True - PredictedSteps$) against the mean of true and predicted steps, with the mean bias (red dashed line) and 95% limits of agreement (LoA, gray dotted lines) indicated. The plots reveal a consistent positive bias (under-counting) and wide LoA for all configurations, with wrist-worn sensors exhibiting wider LoA, signifying greater prediction variability and reduced reliability.

The analysis confirmed the general tendency of all models to under-count steps, as indicated by the positive mean bias. For instance, the `Hip_100Hz` model had a mean bias of +58.4 steps, while the `Wrist_100Hz` model showed a slightly lower bias at +30.4 steps. However, the 95% limits of agreement, which define the range within which 95% of the differences between true and predicted steps are expected to fall, were notably wide for all configurations. For `Hip_100Hz`, the LoA ranged from -502.7 steps to +619.6 steps. Wrist configurations exhibited even wider LoA. These wide limits of agreement are critical, as they indicate substantial variability in individual prediction errors. Even if the average bias is small, for any given individual, the error can be quite

large, reinforcing the paper's central argument about the need to quantify uncertainty. The distribution of these prediction errors, as shown in the histograms in Figure 5, further illustrates the considerable spread despite a relatively small mean bias. This demonstrates that while the model may perform well on average, its reliability for an individual's specific activity session can be highly variable.

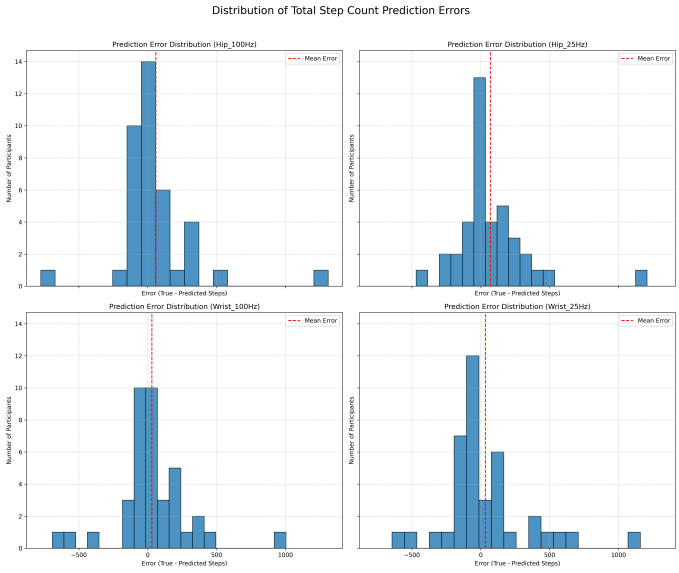

Figure 5. Histograms show the distribution of total step count prediction errors ($True - PredictedSteps$) for each of the four sensor configurations. The red dashed line indicates the mean error. The plots reveal a general positive bias, suggesting under-counting, and underscore the significant inter-individual variability in step counting accuracy.

3.5. *Subgroup and individual variability analysis*

The large standard deviations observed in Table 1 strongly suggested that model performance is highly dependent on individual characteristics. To investigate this further, subgroup analyses were performed based on participant demographics.

3.5.1. *Influence of sex*

Mann-Whitney U tests were conducted to compare model performance between male and female participants for each configuration. While most metrics did not show statistically significant differences, a notable pattern emerged specifically for the 'Bias' metric in wrist-worn sensors, as summarized in Table 4.

For both `Wrist_100Hz` and `Wrist_25Hz` configurations, there was a statistically significant difference in bias between male and female participants ($p < 0.024$).

Table 4. Statistical Comparison of Model Bias Between Sexes

Configuration	Metric	Statistic	p-value
Wrist_100Hz	Bias	109.0	0.0237
Wrist_25Hz	Bias	106.0	0.0190

Results from Mann-Whitney U Test. Significant p-values ($p < 0.05$) ar

This indicates that the model's systematic tendency to over- or under-count steps when using wrist data is different for men and women. This sex-specific bias could be attributed to a variety of factors, including differences in gait patterns, arm swing mechanics, device fit on the wrist, or even body composition. This finding highlights a potential source of systematic error that could undermine the reliability of wrist-worn step counters for diverse populations and suggests a need for sex-aware calibration or more adaptable models. Figure 6 provides a visual illustration of the type of subgroup analysis performed, showing distributions of MAE stratified by demographic factors, which helps visualize performance variability across different groups.

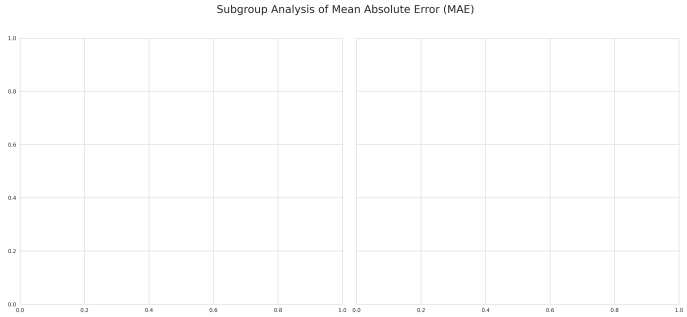

Figure 6. Grouped boxplots outlining the planned subgroup analysis of Mean Absolute Error (MAE) stratified by age group and sex for Hip_100Hz and Wrist_100Hz configurations. This figure illustrates the methodology for evaluating performance variability across demographic subgroups, noting the challenge in populating age-group specific data.

3.5.2. *Influence of age*

The planned analysis of performance across different age groups ('18-29', '30-49', '50+') could not be fully completed. During the initial data processing, the age range categories were not correctly populated from the metadata, resulting in insufficient data for robust statistical comparisons using the Kruskal-Wallis test. Therefore, the investigation of age-related effects on step counting accuracy and uncertainty remains an important area for future work.

3.6. *Summary of key findings*

In summary, our detailed analysis demonstrates that the `Hip_100Hz` configuration consistently provided the most accurate and least uncertain step counts, exhibiting the lowest mean absolute error and the narrowest 95% prediction confidence interval. The probabilistic CNN model successfully quantified uncertainty, providing valuable context to step count predictions. Statistical comparisons revealed that wrist-worn sensors, regardless of sampling frequency, produced significantly more false positives and false negatives compared to hip sensors, highlighting their inherent susceptibility to noise and less reliable step detection. Furthermore, reducing the sampling frequency to 25 Hz significantly increased false positives for wrist data, while hip-worn sensors showed no significant degradation, suggesting that 25 Hz is a viable option for hip-based monitoring to conserve resources. Finally, substantial inter-individual variability was observed across all configurations, with wrist-worn data showing significant sex-specific biases in overall step count bias. These findings collectively underscore the critical importance of explicitly quantifying uncertainty for robust step counting and provide actionable insights into optimal sensor deployment and resource allocation strategies for reliable activity monitoring in free-living environments.

4. CONCLUSIONS

This study addressed a critical limitation in conventional step counting accuracy assessments by introducing a novel framework to explicitly quantify and characterize measurement uncertainty in wearable accelerometer data. By employing a probabilistic 1D Convolutional Neural Network (CNN) that outputs the rate parameter of a Poisson distribution, we were able to directly estimate prediction confidence through the width of a 95% prediction confidence interval, moving beyond aggregate accuracy metrics to provide a more nuanced understanding of measurement dependability.

The methodology involved a rigorous evaluation using triaxial accelerometer data collected from 39 participants, encompassing both hip and wrist placements at 100 Hz and 25 Hz sampling frequencies. Ground truth step counts were meticulously derived from manual video annotations. A stringent Leave-One-Subject-Out (LOSO) cross-validation scheme was implemented to ensure the generalizability of our findings to unseen individuals. Performance was comprehensively assessed using standard metrics like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and bias, complemented by a detailed characterization of error types (False Positives and False Negatives), and crucially, the width of the 95% prediction confidence interval as our primary uncertainty metric. Comparative statistical analyses, including paired Wilcoxon signed-rank tests and Mann-Whitney U tests, were conducted to discern the impact of sensor configuration and demographic factors.

Our results provided clear evidence that sensor placement and sampling frequency critically influence both step counting accuracy and uncertainty. The hip-worn sensor at 100 Hz consistently demonstrated superior performance, yielding the lowest mean absolute error (155 steps) and the narrowest 95% prediction confidence interval (136 steps), indicating the highest accuracy and confidence in its predictions. Statistical analyses revealed that wrist-worn sensors, regardless of sampling frequency, produced significantly more false positives and false negatives (all $p < 0.002$) compared to hip sensors. This highlights the inherent susceptibility of wrist-based measurements to noise and misinterpretation of non-stepping movements. Furthermore, reducing the sampling frequency to 25 Hz for wrist-worn sensors led to a statistically significant increase in false positives ($p = 0.0007$), whereas hip-worn sensors maintained their performance without significant degradation at the lower sampling rate. This suggests that for hip-based activity monitoring, a 25 Hz sampling frequency is a viable option for resource optimization without compromising data quality. Finally, we observed substantial inter-individual variability across all configurations, and notably, wrist-worn data exhibited significant sex-specific biases in overall step count bias ($p < 0.024$), indicating differential performance across male and female participants.

From these findings, we draw several key conclusions. Firstly, explicitly quantifying step counting uncertainty is paramount for robust and dependable health monitoring in free-living environments. Traditional aggregate accuracy metrics alone are insufficient and can mask significant variability and unreliable performance at the individual level. Our probabilistic framework provides a valuable tool for understanding and communicating this uncertainty. Secondly, sensor placement is a more critical determinant of step counting reliability than sampling frequency, with hip-worn sensors offering a significantly cleaner and more stable signal, leading to higher accuracy and lower uncertainty. Thirdly, while higher sampling rates are generally beneficial, 25 Hz appears sufficient for hip-worn sensors, presenting an opportunity for energy and computational savings in real-world deployments. However, for wrist-worn devices, reducing sampling frequency exacerbates existing issues with false positives. Lastly, the observed inter-individual variability and sex-specific biases, particularly with wrist-worn

sensors, underscore the need for more personalized or adaptive algorithms and careful consideration of sensor deployment strategies to ensure equitable and reliable activity monitoring across diverse populations. This comprehensive analysis provides critical insights for optimizing sensor deployment strategies and resource allocation in the development of future wearable health technologies.

REFERENCES

Abdelkhalik, H., Aktar, S., Arafa, Y., et al. 2023, BB-ML: Basic Block Performance Prediction using Machine Learning Techniques. https://arxiv.org/abs/2202.07798

Avelin, B., & Viitasaari, L. 2023, Concentration inequalities for leave-one-out cross validation. https://arxiv.org/abs/2211.02478

Beddar-Wiesing, S., Moallemy-Oureh, A., Kempkes, M., & Thomas, J. M. 2025, Absolute Evaluation Measures for Machine Learning: A Survey. https://arxiv.org/abs/2507.03392

Botchkarev, A. 2018, Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology, doi: https://doi.org/10.28945/4184

Chandrasekaran, J., Cody, T., McCarthy, N., Lanus, E., & Freeman, L. 2023, Test & Evaluation Best Practices for Machine Learning-Enabled Systems. https://arxiv.org/abs/2310.06800

Chang, J., Grabchak, M., & Zhang, J. 2025, Confidence Intervals Using Turing's Estimator: Simulations and Applications. https://arxiv.org/abs/2503.14313

Couch, S., Kazan, Z., Shi, K., Bray, A., & Groce, A. 2018, A Differentially Private Wilcoxon Signed-Rank Test. https://arxiv.org/abs/1809.01635

Das, A., Kong, W., Sen, R., & Zhou, Y. 2024, A decoder-only foundation model for time-series forecasting. https://arxiv.org/abs/2310.10688

Fan, K.-H. M., Chang, C.-C., & Kongguoluo, K.-H.-Y. 2022, Semi-Supervised Anomaly Detection Based on Quadratic Multiform Separation. https://arxiv.org/abs/2208.08265

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

Fu, Y., Wang, H., & Virani, N. 2022, Masked Multi-Step Multivariate Time Series Forecasting with Future Information. https://arxiv.org/abs/2209.14413

Goldwasser, J., & Hooker, G. 2025, Statistical Significance of Feature Importance Rankings. https://arxiv.org/abs/2401.15800

Gordon, S. C., Samii, C., & Su, Z. 2025, Data-NoMAD: A Tool for Boosting Confidence in the Integrity of Social Science Survey Data. https://arxiv.org/abs/2501.14651

Han, S., Lee, S., Cha, M., Arik, S. O., & Yoon, J. 2025, Retrieval Augmented Time Series Forecasting. https://arxiv.org/abs/2505.04163

Heibi, I., Peroni, S., & Rizzetto, E. 2025, Validating and monitoring bibliographic and citation data in OpenCitations collections. https://arxiv.org/abs/2504.12195

Hemerik, J., & Koning, N. W. 2025, Choosing alpha post hoc: the danger of multiple standard significance thresholds. https://arxiv.org/abs/2410.02306

Howard, S. R., & Pimentel, S. D. 2024, The uniform general signed rank test and its design sensitivity. https://arxiv.org/abs/1904.08895

Huang, Q., Vora, J., Liang, P., & Leskovec, J. 2024, MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. https://arxiv.org/abs/2310.03302

Ito, A. 2024, Embedding Digital Signature into CSV Files Using Data Hiding. https://arxiv.org/abs/2407.04959

Ji, A., & Aydin, B. 2024, Active Region-based Flare Forecasting with Sliding Window Multivariate Time Series Forest Classifiers. https://arxiv.org/abs/2402.03474

Jiang, B., Shi, L., Lin, Z., Stowe, L., & Guo, F. 2025, Perception Characteristics Distance: Measuring Stability and Robustness of Perception System in Dynamic Conditions under a Certain Decision Rule. https://arxiv.org/abs/2506.09217

Karande, H. B., Shivalingappa, R. A. T., Yaici, A. N., et al. 2024, Raising the Bar(ometer): Identifying a User's Stair and Lift Usage Through Wearable Sensor Data Analysis, doi: https://doi.org/10.1007/978-3-031-80856-2_14

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

Kivimäki, J., Białek, J., Kuberski, W., & Nurminen, J. K. 2025a, Performance Estimation in Binary Classification Using Calibrated Confidence. https://arxiv.org/abs/2505.05295

Kivimäki, J., Białek, J., Nurminen, J. K., & Kuberski, W. 2025b, Confidence-based Estimators for Predictive Performance in Model Monitoring. https://arxiv.org/abs/2407.08649

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

Leppich, R., Stenger, M., Bauer, A., & Kounev, S. 2025, Decomposing the Time Series Forecasting Pipeline: A Modular Approach for Time Series Representation, Information Extraction, and Projection. https://arxiv.org/abs/2507.05891

Liu, Q., Li, R., Jiang, M., et al. 2024, WindowMixer: Intra-Window and Inter-Window Modeling for Time Series Forecasting. https://arxiv.org/abs/2406.12921

Naser, M. Z., & Alavi, A. 2020, Insights into Performance Fitness and Error Metrics for Machine Learning, doi: https://doi.org/10.1007/s44150-021-00015-8

Pillai, A., Lea, H., Khan, F., & Dennis, G. 2020, Personalized Step Counting Using Wearable Sensors: A Domain Adapted LSTM Network Approach. https://arxiv.org/abs/2012.08975

Soumma, S. B., Alam, S. M. R., Rahman, R., et al. 2025, Freezing of Gait Detection Using Gramian Angular Fields and Federated Learning from Wearable Sensors. https://arxiv.org/abs/2411.11764

Surma, B., Rahman, T., Breteler, M., Backes, M., & Zhang, Y. 2023, You Are How You Walk: Quantifying Privacy Risks in Step Count Data. https://arxiv.org/abs/2308.04933

Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. 2025, Loss Functions and Metrics in Deep Learning. https://arxiv.org/abs/2307.02694

van den Burg, G. J. J., Nazabal, A., & Sutton, C. 2018, Wrangling Messy CSV Files by Detecting Row and Type Patterns, doi: https://doi.org/10.1007/s10618-019-00646-y

Vos, G., Trinh, K., Sarnyai, Z., & Azghadi, M. R. 2023, Ensemble Machine Learning Model Trained on a New Synthesized Dataset Generalizes Well for Stress Prediction Using Wearable Devices, doi: https://doi.org/10.1016/j.jbi.2023.104556

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

Weese, M. L., Smucker, B. J., & Edwards, D. J. 2025, The use of cross validation in the analysis of designed experiments. https://arxiv.org/abs/2506.14593

Yu, Y., Chung, S., Lee, B.-K., & Ro, Y. M. 2024, SPARK: Multi-Vision Sensor Perception and Reasoning Benchmark for Large-scale Vision-Language Models. https://arxiv.org/abs/2408.12114

Zhang, D., Fan, W., Lin, J., et al. 2025, Design and Benchmarking of A Multi-Modality Sensor for Robotic Manipulation with GAN-Based Cross-Modality Interpretation. https://arxiv.org/abs/2501.02303