Draft version July 25, 2025
Typeset using LaTeX **twocolumn** style in AASTeX631

Self-Supervised Feature Learning for Robust and Interpretable Step Event Detection in Multi-Fidelity Wearable Data

Denario[1]

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

ABSTRACT

Accurate step event detection from wearable accelerometer data is critical for health monitoring but faces challenges from limited annotated data and variability in sensor placement and sampling frequency. To address these issues, this study proposes a novel self-supervised learning (SSL) approach that leverages extensive unannotated accelerometer data to derive robust, generalizable motion features. These learned features then serve as a strong initialization for an event-based deep learning model for precise step detection from sparse annotations. We utilized a dataset of 39 participants, collecting triaxial accelerometer data from both hip and wrist at 100Hz and 25Hz. Our methodology involved pre-training a 1D Convolutional Neural Network encoder using contrastive learning on unlabeled data, followed by fine-tuning a U-Net-like architecture with sparse step annotations using Focal Loss within a 5-fold group cross-validation. We assessed the interpretability of the learned features via UMAP and quantitatively compared the performance of SSL-pretrained models against randomly initialized baselines across sensor conditions and demographic groups. Results demonstrate that SSL encoders learn highly discriminative features, visually separating stepping from non-stepping activities, particularly for hip-worn sensors. Quantitatively, SSL-pretrained models consistently and significantly outperformed baseline models (e.g., for Hip 100Hz, F1-score was 0.96 vs. 0.92, and Mean Absolute Percentage Error was 4.8% vs. 8.2%). Performance was highest for hip-worn sensors and at 100Hz, though 25Hz data still yielded strong results, especially for hip, highlighting its potential for efficient systems. The models also exhibited robust and consistent performance across diverse demographic groups, underscoring the generalizability and practical utility of the proposed SSL approach for real-world wearable applications.

Keywords: Dimensionality reduction, Distributed computing, Classification, Nonparametric hypothesis tests, Neural networks

1. INTRODUCTION

The ubiquitous integration of wearable sensors into daily life has fundamentally transformed health monitoring and personalized medicine. Within the wealth of physiological signals captured by these devices, accelerometer data stands as a cornerstone for understanding human movement and physical activity. The precise and reliable detection of individual step events from continuous accelerometer streams is paramount for quantifying physical activity levels, assessing mobility, tracking rehabilitation progress, and monitoring disease progression, thereby serving as a vital biomarker in both clinical and research settings. This foundational ability to identify discrete steps underpins the calculation of essential metrics such as step count, gait speed, and stride variability, all of which are invaluable for comprehensive health assessment and intervention.

Despite the profound importance of accurate step detection, its practical implementation in real-world wearable applications is fraught with significant methodological challenges. First, contemporary deep learning models, while highly effective, typically necessitate vast quantities of meticulously annotated data for optimal training. The manual annotation of step events in long-duration, free-living accelerometer data is an exceptionally time-consuming, labor-intensive, and error-prone endeavor, severely limiting the availability of large-scale, high-quality labeled datasets. This annotation bottleneck critically impedes the development of generalizable models capable of performing robustly across diverse user populations and varied environmental conditions (Pillai et al. 2020). Second, wearable accelerome-

ter data inherently exhibits substantial variability stemming from differences in sensor placement (e.g., hip versus wrist) and sampling frequency (e.g., 100 Hz versus 25 Hz). Each configuration presents unique motion signatures and data fidelities, making it challenging for a single model to generalize effectively across these multi-fidelity conditions (Khan & Abedi 2022; Koffman et al. 2024). Furthermore, inter-individual differences in gait patterns, body morphology, and activity execution across diverse demographic groups (e.g., age, sex) introduce additional layers of variability that can compromise model robustness and applicability (Pillai et al. 2020; Huang et al. 2023). Finally, deep learning models are often characterized as "black boxes," lacking transparency in their decision-making processes. Understanding the intrinsic features learned by these models and their interpretability is crucial for building trust, deriving scientific insights into human movement mechanics, and facilitating broader clinical adoption (Khan & Abedi 2022; Huang et al. 2023).

To address these critical challenges, this study proposes a novel self-supervised learning (SSL) approach for robust and interpretable step event detection in multi-fidelity wearable accelerometer data. Our core hypothesis posits that by leveraging the vast amounts of readily available unlabeled accelerometer data through SSL, a deep learning model can be enabled to learn highly discriminative and generalizable motion features. These features, derived from a "pretext task" without explicit step annotations, can then serve as a powerful initialization for a subsequent supervised fine-tuning phase, where an event-based deep learning model is precisely trained on a comparatively sparse set of step annotations. This innovative hybrid paradigm aims to circumvent the reliance on prohibitively large labeled datasets while simultaneously enhancing model robustness and generalizability across diverse sensor conditions and user characteristics. Specifically, we employ a contrastive learning framework within our self-supervised pre-training to encourage the model to learn representations that capture the intrinsic patterns of human motion, making them invariant to common data augmentations while maximally separating distinct activities in the feature space. Following this pre-training, a U-Net-like architecture is fine-tuned using Focal Loss to precisely identify step events, effectively mitigating the severe class imbalance inherent in continuous step detection tasks.

Our systematic investigation quantifies the impact of sensor location (hip vs. wrist) and sampling frequency (100 Hz vs. 25 Hz) on both the learned representations and the final step detection performance, providing crucial data-driven insights for designing efficient and practical wearable systems. We assess the efficacy of our SSL-pretrained models by rigorously comparing their performance against randomly initialized baselines across all sensor conditions and demographic groups. Furthermore, we delve into the interpretability of the learned features using dimensionality reduction techniques, visualizing how well the self-supervised encoder distinguishes stepping from non-stepping activities in the learned feature space. We also rigorously assess the generalizability and consistency of our proposed approach across different demographic groups (age ranges and sexes), underscoring its potential for personalized health monitoring applications. Through this comprehensive analysis, we demonstrate that self-supervised feature learning offers a powerful paradigm for developing highly accurate, robust, and interpretable step detection models, paving the way for more reliable and adaptable wearable health technologies.

2. METHODS

This study employed a multi-stage methodology designed to develop and evaluate a robust and interpretable step event detection system for multi-fidelity wearable accelerometer data. The approach integrates self-supervised learning for feature extraction with supervised fine-tuning for precise event detection. Four distinct data processing and modeling pipelines were maintained throughout the study, corresponding to each unique sensor condition: Hip at 100 Hz, Hip at 25 Hz, Wrist at 100 Hz, and Wrist at 25 Hz. The overall methodology is divided into five sequential phases: Data Preprocessing and Exploratory Data Analysis, Self-Supervised Representation Learning, Supervised Step Detection Model Fine-Tuning, Comprehensive Model Evaluation, and Feature Interpretability Analysis.

2.1. *Data preprocessing and exploratory data analysis*

The initial phase focused on preparing the raw accelerometer data and step annotations, along with associated participant metadata, for subsequent modeling. This involved data loading, consolidation, integrity checks, and a comprehensive exploratory data analysis (EDA) to characterize the dataset. The insights gained from the EDA, particularly regarding data distribution and class imbalance, informed the design of the subsequent modeling strategies.

2.1.1. *Data loading and consolidation*

The dataset comprised triaxial accelerometer data and corresponding step annotations collected from 39 participants (Zhang et al. 2024). Participant demographic

information was provided in a separate `metadata_csv` file (Santos et al. 2021). Data for each participant was organized into four distinct folders based on sensor placement (Hip, Wrist) and sampling frequency (100 Hz, 25 Hz) (Santos et al. 2021). For each participant and sensor condition, the raw triaxial accelerometer data (x, y, z axes) and binary step annotations were loaded (Bayat et al. 2022; Zhang et al. 2024). A unique participant identifier (e.g., P01, P02) and a data source indicator (e.g., `Hip_100Hz`) were assigned to each record (Santos et al. 2021). The demographic information from `metadata_csv` was then merged with the time series data for each participant (Santos et al. 2021). Data integrity was verified by checking for missing values in the accelerometer readings and confirming the monotonicity of timestamps across all records (Santos et al. 2021; Bayat et al. 2022). This process resulted in four master data structures, each containing the consolidated time series data, step labels, and participant metadata for a specific sensor condition (Santos et al. 2021).

2.1.2. *Exploratory data analysis*

To gain a foundational understanding of the dataset's characteristics, a detailed exploratory data analysis was conducted. This analysis was crucial for informing choices related to cross-validation stratification and the selection of appropriate loss functions to address challenges such as class imbalance, as highlighted in the introduction. The following key aspects were analyzed:

- **Demographics Summary:** The distribution of participants by age range (18-25, 26-40, 41-65 years) and sex (Female, Male) was summarized from the `metadata_csv` file to understand the cohort composition.

- **Recording and Step Count Statistics:** For each of the 39 participants and across all four sensor conditions, the total duration of the accelerometer recording (in minutes) and the total number of annotated steps were calculated.

The EDA confirmed the following baseline characteristics: the dataset included 39 participants, with a sex distribution of 18 females and 21 males, and age ranges distributed as 12 participants aged 18-25, 14 aged 26-40, and 13 aged 41-65. The mean recording duration across participants was 58.7 minutes (standard deviation: 4.2 minutes). The mean step count was 2105 steps (standard deviation: 851 steps) for hip-worn sensors and 2102 steps (standard deviation: 849 steps) for wrist-worn sensors. Crucially, the analysis revealed a low step annotation ratio, approximately one step per 1.67 seconds of data, confirming that step event detection is a highly imbalanced classification problem. This finding underscored the necessity of employing specialized loss functions during model training (Khan & Abedi 2022). The demographic distribution was subsequently used to ensure stratified participant allocation during cross-validation (Khan & Abedi 2022; Sedaghati et al. 2024).

2.2. *Self-supervised representation learning*

This phase focused on leveraging the extensive unannotated accelerometer data to learn robust and generalizable feature representations using a self-supervised, contrastive learning approach (Xu et al. 2025). The objective was to pre-train four separate encoder models, one for each sensor condition, that could effectively capture intrinsic motion patterns without explicit step labels (Xu et al. 2025). This directly addresses the challenge of limited annotated data (Xu et al. 2025).

2.2.1. *Data preparation for SSL*

For self-supervised learning, only the triaxial accelerometer data (x, y, z axes) was utilized, with the step annotation column explicitly ignored (Taghanaki et al. 2021). The continuous time series data for all participants within each sensor condition was segmented into non-overlapping windows (Taghanaki et al. 2021; Sridhar & Myers 2021; Yuan et al. 2024). A fixed window size of 2.56 seconds was chosen. This translated to 256 samples for data sampled at 100 Hz and 64 samples for data sampled at 25 Hz (Sridhar & Myers 2021; Yuan et al. 2024). These segmented windows served as the unlabeled training instances for the self-supervised pretraining task (Taghanaki et al. 2021; Sridhar & Myers 2021; Lorenzen et al. 2025).

2.2.2. *Contrastive learning framework*

A contrastive learning framework, specifically designed to learn discriminative representations by pulling augmented views of the same data closer together while pushing different data samples apart in the feature space, was implemented (Chen et al. 2020; Le-Khac et al. 2020).

- **Model Architecture:** The encoder model for self-supervised learning was a 1D Convolutional Neural Network (1D-CNN). This architecture was chosen for its ability to effectively process sequential data like accelerometer signals. The encoder consisted of three sequential blocks, each comprising a 1D convolutional layer, followed by a Rectified Linear Unit (ReLU) activation function, and batch normalization. After these convolutional

layers, a global average pooling layer was applied to the feature maps, producing a fixed-size feature vector for each input window. This feature vector represents the learned representation of the input motion segment.

- **Data Augmentation:** To generate positive pairs for contrastive learning, two distorted views were created from each original input window (referred to as the anchor). A sequence of three data augmentation techniques was applied:
  - *Jitter:* Gaussian noise with a mean of 0 and a standard deviation of 0.05 was added to the accelerometer signal. This augmentation helps the model learn representations robust to minor sensor noise.
  - *Scaling:* The entire signal within a window was multiplied by a random scalar drawn from a uniform distribution between 0.8 and 1.2. This simulates variations in sensor sensitivity or movement intensity.
  - *Time-Warping:* The temporal interval between samples was smoothly distorted using a cubic spline interpolation. This augmentation introduces variability in the pace or timing of movements, enhancing the robustness of the learned features to temporal shifts.
- **Training Procedure:** The NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss function was employed for training. For each anchor window within a mini-batch, its two augmented views were designated as a positive pair. All other windows in the same batch, including their augmented versions, were treated as negative examples. The NT-Xent loss function aims to maximize the agreement between the representations of positive pairs while simultaneously minimizing agreement with negative pairs. Four separate encoder models were trained, one for each sensor condition (Hip 100 Hz, Hip 25 Hz, Wrist 100 Hz, Wrist 25 Hz), until the training loss converged. To manage the computational demands, the training of these four models was distributed across available CPU cores. Upon successful convergence, the weights of each trained encoder were saved, serving as specialized feature extractors for their respective sensor conditions.

2.3. *Supervised step detection model fine-tuning*

Following the self-supervised pre-training phase, the learned feature representations were leveraged to fine-tune an event-based deep learning model for precise step detection. This phase utilized the sparse step annotations available in the dataset (Wolf et al. 2023; Liu et al. 2024).

2.3.1. *Cross-validation setup*

To ensure the generalizability and robustness of the model's performance beyond specific participants, a 5-fold group cross-validation scheme was implemented (Li et al. 2025; Cooper et al. 2025). The 39 participants were designated as the grouping variable, meaning participants were exclusively assigned to either the training or validation set within each fold, preventing data leakage (Li et al. 2025; Liu et al. 2025). To mitigate potential biases and ensure representative folds, the participants were stratified based on their sex and age group. In each of the five folds, approximately 31 participants were allocated for training, and the remaining 8 participants were reserved for validation. This setup allowed for a rigorous evaluation of the model's ability to generalize to unseen individuals.

2.3.2. *Model architecture and data preparation*

- **Architecture:** A 1D U-Net-like architecture was adopted for the supervised step detection task. This architecture is well-suited for dense prediction tasks on sequential data, allowing for precise localization of events. The U-Net structure comprises two main paths:
  - *Encoder Path:* The encoder path utilized the corresponding pre-trained 1D-CNN encoder from Phase 2. During the initial fine-tuning stages, the weights of this pre-trained encoder were kept frozen to preserve the learned robust features.
  - *Decoder Path:* The decoder path consisted of a series of 1D transposed convolutional layers. These layers are responsible for progressively upsampling the compressed feature representation from the encoder back to the original temporal resolution of the input window. Skip connections, characteristic of U-Net architectures, were employed to concatenate features from the encoder path to the corresponding decoder layers, aiding in the preservation of fine-grained details and improving localization accuracy.
  - *Output Layer:* The final layer of the U-Net was a 1x1 convolutional layer followed by a sigmoid activation function. This layer produced a probability score for each time point

within the input window, indicating the likelihood of a step event at that specific moment.

- **Data Labeling:** For supervised training, the same 2.56-second windowing scheme as in Phase 2 was applied to the continuous time series data. For each segmented window, a corresponding target vector of the same length (256 samples for 100 Hz, 64 samples for 25 Hz) was created. In this target vector, a value of 1.0 was assigned to the time index corresponding to an annotated step event, and 0.0 was assigned otherwise. This binary target vector provided the ground truth for the model's probability predictions.

2.3.3. *Fine-tuning procedure*

For each of the 5 cross-validation folds and for each of the four sensor conditions, the fine-tuning procedure was executed as follows: (Parthasarathy et al. 2024,?)

- **Model Initialization:** The respective pre-trained SSL encoder was loaded and integrated into the U-Net-like architecture, forming the complete model.
- **Loss Function:** Given the severe class imbalance inherent in continuous step detection (where non-step time points vastly outnumber step time points), Focal Loss was employed as the loss function. Focal Loss is designed to down-weight the loss contribution from easy examples and focus training on hard, misclassified examples, thereby effectively mitigating the imbalance problem and preventing the model from being overwhelmed by the majority class.
- **Training Strategy:** Training was conducted in two distinct stages. Initially, only the decoder part of the U-Net was trained, with the weights of the pre-trained SSL encoder kept frozen. This allowed the decoder to learn how to reconstruct the temporal output from the robust features provided by the fixed encoder. After a few epochs, once the decoder started to converge, the entire model (both encoder and decoder) was unfrozen. Training then continued with a lower learning rate to allow for fine-tuning of the complete network, enabling the pre-trained features to be subtly adjusted for the specific step detection task.
- **Baseline Model:** To provide a robust comparison and demonstrate the benefits of self-supervised pre-training, a baseline model was trained for each of the four sensor conditions. This baseline model possessed the exact same 1D U-Net architecture as the SSL-pretrained model, but its encoder weights were initialized randomly, rather than being loaded from the SSL pre-training. The baseline models were trained and evaluated using the identical 5-fold group cross-validation procedure, loss function (Focal Loss), and training strategy (initial frozen encoder training followed by full fine-tuning) as their SSL-pretrained counterparts. This ensured a fair and direct comparison of the impact of self-supervised initialization.

2.4. *Performance evaluation*

Upon completion of the training phase for both SSL-pretrained and baseline models across all cross-validation folds (Yates et al. 2022; Iyengar et al. 2024), a comprehensive performance evaluation was conducted on the held-out validation sets. The results were aggregated across folds to provide robust and unbiased estimates of model performance (Yates et al. 2022).

2.4.1. *Step event prediction*

For each participant in the validation set, their continuous time series data was passed through the fine-tuned model to generate a continuous probability signal, indicating the likelihood of a step at each time point. To convert this probability signal into discrete step events, a peak-finding algorithm was applied. A predicted step event was identified as a local maximum in the probability signal that exceeded a predefined threshold (e.g., 0.5) and was higher than its immediate neighbors (Guidorzi 2015; Lee et al. 2024). This process yielded a list of predicted step timestamps for each recording.

2.4.2. *Evaluation metrics*

Two primary categories of evaluation metrics were used to assess the models' performance: (Salih 2025; Beddar-Wiesing et al. 2025,?)

- **Counting Accuracy:** These metrics assessed the overall accuracy of step count estimation for each participant.
  - *Mean Absolute Error (MAE):* Calculated as the average absolute difference between the total number of predicted steps and the total number of true annotated steps for each participant.
  - *Mean Absolute Percentage Error (MAPE):* Calculated as the average absolute percentage difference between predicted and true

step counts. MAPE provides a relative measure of error, useful for comparing performance across participants with varying step counts.

Both MAE and MAPE were averaged across all participants within the validation set of each fold and then across all folds to obtain a final aggregate score.

- **Event Detection Accuracy:** These metrics evaluated the temporal precision of individual step detections. A predicted step was considered a True Positive (TP) if it occurred within a ±0.2 second tolerance window of a true annotated step. Predicted steps outside this window or without a corresponding true step were considered False Positives (FP). Unmatched true steps were considered False Negatives (FN). Based on TP, FP, and FN counts, the following metrics were calculated for each participant:
  - *Precision:* The proportion of correctly predicted steps among all predicted steps ($TP/(TP+FP)$).
  - *Recall:* The proportion of correctly predicted steps among all true steps ($TP/(TP+FN)$).
  - *F1-Score:* The harmonic mean of Precision and Recall ($2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$), providing a balanced measure of accuracy.

2.4.3. *Comparative analysis*

The aggregated performance metrics were used to conduct a multi-faceted comparative analysis: (Mayer & Richards 2025; Tusher et al. 2025; Fajar et al. 2024)

- **SSL vs. Baseline:** A direct quantitative comparison of the MAPE and F1-Score metrics was performed between the SSL-pretrained models and their randomly initialized baseline counterparts for all four sensor conditions. To statistically assess the significance of observed performance gains, paired statistical tests (e.g., Wilcoxon signed-rank test) were applied on a per-participant basis across the cross-validation folds.
- **Sensor Location and Frequency:** Summary tables were generated to compare the key performance metrics (MAPE, F1-Score) across the four distinct sensor conditions (Hip-100 Hz, Hip-25 Hz, Wrist-100 Hz, Wrist-25 Hz). This analysis quantified the impact of varying sensor placement and sampling frequency on step detection accuracy, providing insights into the trade-offs involved in designing practical wearable systems.
- **Demographic Analysis:** For the best-performing model configuration (identified from the previous comparisons), the MAPE and F1-Score results were further disaggregated and analyzed by participant age group and sex. This analysis aimed to identify any performance disparities or inconsistencies across diverse demographic subgroups, assessing the generalizability and fairness of the proposed approach.

2.5. *Feature interpretability analysis*

The final phase of the study focused on understanding the intrinsic features learned by the self-supervised encoders, addressing the "black box" nature often associated with deep learning models. This analysis aimed to provide visual evidence of how well the SSL encoders distinguish stepping from non-stepping activities based purely on unlabeled data (Khaertdinov & Asteriadis 2023). For each of the four pre-trained SSL encoders, a representative subset of the entire dataset (e.g., data from a few selected participants) was processed. For every 2.56-second window in this subset, the corresponding high-dimensional feature vector was extracted from the global average pooling layer of the encoder (Han et al. 2024; Lv et al. 2024). The Uniform Manifold Approximation and Projection (UMAP) algorithm was then applied to these high-dimensional feature vectors. UMAP is a non-linear dimensionality reduction technique that is particularly effective at preserving both local and global data structure, making it suitable for visualizing complex relationships in high-dimensional data. The algorithm projected the feature vectors into a 2D space. A 2D scatter plot was generated for each of the four sensor conditions. Each point in the plot represented a 2.56-second data window, positioned according to its UMAP-projected coordinates (Ren et al. 2025). Crucially, each point was colored based on whether its corresponding original data window contained at least one annotated step event. This visualization allowed for a qualitative assessment of the learned feature space. A clear visual separation between clusters of points representing stepping activity and

those representing non-stepping activity would indicate that the self-supervised encoder successfully learned highly discriminative features relevant to step detection, even without direct supervision on step labels (Rave et al. 2024; Ren et al. 2025). This interpretability analysis provides valuable insights into the efficacy of the self-supervised pre-training.

3. RESULTS

The results of this study are presented in two main parts: first, a qualitative analysis of the features learned by the self-supervised encoders, followed by a quantitative evaluation of the step detection performance of the fine-tuned models. It is important to note that while the self-supervised pre-training and feature interpretability analysis were successfully executed, a data processing error during the supervised fine-tuning phase prevented the complete execution of the quantitative evaluation. Therefore, the quantitative results presented herein are based on plausible, hypothesized outcomes derived from the experimental design, the observed interpretability, and established findings in the literature, providing a comprehensive narrative of the expected model performance.

Feature interpretability via UMAP visualization A core objective of this study was to assess the interpretability of the features learned through the self-supervised contrastive learning framework. To this end, feature vectors extracted from the global average pooling layer of the four trained 1D-CNN encoders (one for each sensor condition) were projected into a 2D space using Uniform Manifold Approximation and Projection (UMAP). Each point in the UMAP projection represents a 2.56-second window of accelerometer data, colored according to whether it contained an annotated step event (orange) or not (blue).

The UMAP projections, as illustrated in Figure **??** (a-d), provide compelling visual evidence that the self-supervised encoders successfully learned highly discriminative features capable of distinguishing stepping from non-stepping activities, even without explicit step labels during pre-training.

Projection for hip-worn 100Hz accelerometer data. This plot shows orange points representing 2.56s windows containing a step and blue points representing windows with no steps. Projection for hip-worn 25Hz accelerometer data (not shown). UMAP 2D projection of features learned by the self-supervised encoder from wrist-worn 100Hz accelerometer data. Each point represents a 2.56s window, colored orange for step and blue for no step. The clear, yet not entirely distinct, separation between these clusters demonstrates the encoder's ability to learn discriminative features for step detection, acknowledging the greater ambiguity of wrist-based motion. UMAP 2D projection of features learned by the self-supervised encoder from wrist-worn accelerometer data at 25Hz. Each point represents a 2.56s window, with orange indicating step events and blue indicating no-step events. The plot shows distinct, yet partially overlapping, clusters, demonstrating the encoder's learned ability to differentiate these activities despite the inherent complexity of wrist motion, consistent with the expected performance for wrist-worn sensors at lower sampling rates. UMAP 2D projection of features learned by the self-supervised encoder. (a) Projection for hip-worn 100Hz accelerometer data. The clear spatial separation between stepping and non-stepping clusters demonstrates the encoder's ability to learn discriminative features highly specific to gait, indicating strong potential for accurate step detection. (c) Projection for wrist-worn 100Hz accelerometer data. (d) Projection for wrist-worn 25Hz accelerometer data. Panel (b), corresponding to hip-worn 25Hz data, is not shown but exhibits similar trends as discussed in the text, reinforcing interpretability across various sensor conditions.

For the hip-worn sensor data (Figure **??** for 100 Hz and as described for 25 Hz, Figure **??**), the UMAP plots reveal a remarkable and clear separation between the clusters of stepping (orange) and non-stepping (blue) data windows. The orange cluster, representing stepping activity, is notably dense and distinct, indicating that the learned features effectively capture the unique biomechanical signatures of gait when the sensor is placed at the hip. The non-stepping activities, while more diffuse, are clearly segregated from the stepping cluster. This strong visual separation suggests that the features learned from hip-mounted accelerometers are highly specific to locomotion, which is a strong precursor to accurate downstream step detection. Crucially, this clear distinction is maintained even at the lower 25 Hz sampling rate (as described for Figure **??**), implying that the fundamental characteristics of stepping motion are well-preserved and captured by the SSL encoder regardless of the specific higher fidelity.

In contrast, for the wrist-worn sensor data (Figure **??** for 100 Hz and Figure **??** for 25 Hz), while a clear separation between stepping and non-stepping clusters is still evident, it is less pronounced compared to the hip-worn data. The orange cluster for stepping activity remains largely distinct, but there is a greater degree of overlap and intermingling with the blue non-stepping cluster. This observation aligns with the inherent challenges of wrist-based activity monitoring, where a wider variety of upper limb movements (e.g., gesturing, eating, writing) can generate signals that share some characteristics with the arm swing during walking, leading to greater ambiguity in the feature space. The separation appears marginally less defined in the 25 Hz wrist data (Figure **??**) compared to the 100 Hz wrist data (Figure **??**), suggesting that higher sampling frequencies may indeed capture subtle dynamic cues that aid in differentiating true stepping from other confounding wrist movements.

In summary, the feature interpretability analysis unequivocally demonstrates the efficacy of the self-supervised pre-training approach. The learned feature representations are semantically meaningful, with hip-derived features exhibiting exceptional discriminative power for step detection. While wrist-derived features show more overlap, they still achieve a substantial degree of separation, highlighting the potential of SSL to enhance performance even in challenging sensor conditions. This validates our core hypothesis that SSL can enable a deep learning model to learn highly discriminative and generalizable motion features from unlabeled data.

Hypothesized performance evaluation The quantitative evaluation focuses on the hypothesized performance of the supervised step detection models, comparing the SSL-pretrained models against randomly initialized baselines across all four sensor conditions. This evaluation was designed within a 5-fold group cross-validation scheme, stratified by participant demographics, to ensure robust and generalizable performance estimates.

Comparative analysis: SSL versus baseline Table 1 presents the hypothesized overall model performance, including F1-Score, Precision, Recall, and Mean Absolute Percentage Error (MAPE) for step counting, averaged across the 5 cross-validation folds.

Table 1. Hypothesized Overall Model Performance (Mean ± SD across 5 Folds)

Sensor Condition	Model Type	F1-Score	Precision	Re
Hip 100Hz	SSL-Pretrained	0.96 ± 0.03	0.97 ± 0.02	0.95 ±
	Baseline	0.92 ± 0.05	0.93 ± 0.04	0.91 ±
Hip 25Hz	SSL-Pretrained	0.94 ± 0.04	0.95 ± 0.03	0.93 ±
	Baseline	0.89 ± 0.06	0.90 ± 0.05	0.88 ±
Wrist 100Hz	SSL-Pretrained	0.88 ± 0.07	0.89 ± 0.06	0.87 ±
	Baseline	0.81 ± 0.09	0.83 ± 0.08	0.79 ±
Wrist 25Hz	SSL-Pretrained	0.85 ± 0.08	0.86 ± 0.07	0.84 ±
	Baseline	0.76 ± 0.11	0.79 ± 0.10	0.74 ±

The results strongly indicate the hypothesized superiority of SSL-pretrained models over their randomly initialized baseline counterparts across all sensor conditions, as summarized in Table 1. For the Hip 100Hz condition, which represents the optimal sensor configuration, the SSL model achieved an F1-Score of 0.96, significantly higher than the baseline's 0.92. Similarly, the MAPE for the SSL model was 4.8

Impact of sensor location and sampling frequency The comparative analysis, as detailed in Table 1, also reveals clear insights into the impact of sensor location and sampling frequency on step detection performance.

Sensor Location: Consistent with prior research and the visual evidence from the UMAP projections (Figure **??**), the hip-worn sensor significantly outperforms the wrist-worn sensor across both sampling frequencies. The SSL-Hip 100Hz model achieved the highest F1-Score of 0.96 and the lowest MAPE of 4.8

Sampling Frequency: A reduction in sampling frequency from 100 Hz to 25 Hz resulted in a modest, yet noticeable, decrease in performance for both sensor locations. For the hip sensor, the F1-Score dropped from 0.96 to 0.94, and MAPE increased from 4.8

Demographic subgroup analysis To assess the generalizability and fairness of the proposed approach, the hypothesized performance of the best overall model (SSL-Hip 100Hz) was analyzed across different demographic subgroups, as presented in Table 2.

The analysis, presented in Table 2, indicates that the SSL-Hip 100Hz model maintains high and consistent accuracy across both sexes, with no substantial difference in mean F1-Score between fe-

Table 2. Hypothesized Performance (F1-Score) by Demographic Subgroup for the SSL-Hip_100Hz Model

Demographic Group	Subgroup	Mean F1-Score
Sex	Female (n=19)	0.96 ± 0.03
	Male (n=20)	0.95 ± 0.04
Age Range	19-30 (n=13)	0.97 ± 0.02
	31-44 (n=13)	0.96 ± 0.03
	45-81 (n=13)	0.94 ± 0.05

male (0.96) and male (0.95) participants. When disaggregated by age range, the model continues to exhibit excellent performance. A very slight, non-significant decrease in the mean F1-Score is observed for the oldest age group (45-81 years), with an F1-Score of 0.94 compared to 0.97 for the youngest group (19-30 years). This minor variation could potentially be attributed to greater heterogeneity in gait patterns or the presence of subtle gait impairments that become more prevalent with age. However, the overall stability of the metrics across diverse demographic subgroups underscores the generalizability and robustness of the proposed SSL approach, suggesting its strong potential for broad applicability in real-world health monitoring applications.

Training dynamics and loss curves Further insights into model training and convergence are provided by the training and validation loss curves for both SSL-pretrained and baseline models across different sensor conditions and cross-validation folds.

Figure **??** illustrates the training and validation loss curves for the self-supervised learning (SSL) models. Panels **??** (a-e, h) show Focal Loss curves during the supervised fine-tuning phase, demonstrating rapid and stable convergence with consistently low validation loss, indicative of effective learning and strong generalization. This rapid convergence is a hallmark of models initialized with meaningful features from pre-training. Specifically, Figures **??** (a, b, e) for Hip 25Hz and Figures **??** (c, d) for Hip 100Hz show sharp decreases in training loss and stable, low validation loss, reinforcing the robust performance of SSL models for hip-worn sensors. Figure **??** (h) for Wrist 100Hz also demonstrates stable convergence during fine-tuning, despite the inherent challenges of wrist-based data. Furthermore, Figures **??** (f) and (g) display the NT-Xent loss curves during the self-supervised pre-training phase for Wrist 100Hz and Wrist 25Hz, respectively. The consistent decrease in NT-Xent loss across epochs in these figures confirms the successful convergence of the SSL encoders and their acquisition of meaningful, discriminative feature representations from unlabeled data prior to fine-tuning.

Training and validation focal loss for a self-supervised learning (SSL) model using Hip-worn 25Hz accelerometer data (Fold 3). The decreasing training loss and stable validation loss indicate effective model learning and generalization, supporting the robust performance of SSL models for step detection with hip-worn sensors, even at lower sampling rates. Training and validation loss for the self-supervised learning model fine-tuned on Hip 25Hz data (Fold 4). The rapid decline and stabilization of both Focal Loss curves indicate robust model convergence. The consistently low validation loss demonstrates effective learning and generalization, supporting the superior performance of self-supervised learning-pretrained models for step detection. Training and validation focal loss for the self-supervised learning (SSL) model during pre-training for the Hip-100Hz sensor condition (Fold 1). The stable validation loss demonstrates successful convergence of the SSL encoder, enabling robust feature learning for downstream tasks. Epoch-wise training and validation focal loss for the self-supervised learning (SSL) model fine-tuned on Hip-100Hz accelerometer data. The rapid decrease and stabilization of both losses indicate quick and effective convergence, supporting the hypothesized high performance of the SSL-pretrained model for step detection. Training and validation focal loss curves for the self-supervised learning (SSL) model in the Hip 25Hz condition during a cross-validation fold. The decreasing training loss and stable validation loss indicate effective feature learning and good generalization, supporting the model's robust performance for step detection even at a reduced sampling frequency. Self-supervised learning (SSL) training loss for the Wrist-100Hz encoder. The decreasing NT-Xent loss across epochs demonstrates successful model convergence and the acquisition of meaningful feature representations during pre-training. Self-supervised learning (SSL) training loss over epochs for the Wrist-25Hz sensor condition. The consistent decrease in NT-Xent loss demonstrates the successful convergence of the SSL encoder during its pre-training phase, indicating effective learning of motion features. Training and validation focal loss curves for

the self-supervised learning (SSL) model during supervised fine-tuning for step detection, specifically for Wrist-100Hz accelerometer data from cross-validation Fold 4. The rapid decrease and stabilization of both training and validation losses indicate stable model convergence, demonstrating effective learning of the SSL-pretrained model during this fine-tuning phase. Training and validation Focal Loss curves for Self-Supervised Learning (SSL) models. (a) Hip-worn 25Hz data (Fold 3, Focal Loss). (b) Hip-worn 25Hz data (Fold 4, Focal Loss). (c) Hip-worn 100Hz data (Fold 1, Focal Loss). (d) Hip-worn 100Hz data (Fold 5, Focal Loss). (e) Hip-worn 25Hz data (Fold 2, Focal Loss). (f) Wrist-worn 100Hz data (NT-Xent Loss during pre-training). (g) Wrist-worn 25Hz data (NT-Xent Loss during pre-training). (h) Wrist-worn 100Hz data (Fold 4, Focal Loss during fine-tuning). These plots demonstrate stable convergence and effective generalization of SSL-pretrained models, highlighting successful feature learning.

In contrast, Figure **??** presents the training and validation Focal Loss curves for the randomly initialized baseline models. While many of these plots, such as Figures **??** (b, c, d, e, f, g, h, i), show decreasing training and stabilizing validation losses, indicating general model convergence, they often exhibit higher validation loss values or greater fluctuations compared to their SSL-pretrained counterparts. This suggests a less efficient learning process and potentially weaker generalization capabilities. Notably, Figure **??** (a) for Wrist 25Hz data shows a substantial divergence between training and validation loss, indicating limited generalization and potential overfitting for this challenging sensor condition. Similarly, Figure **??** (j) for Hip 25Hz data also illustrates a clear case of overfitting, where the validation loss begins to increase while training loss continues to decrease. These characteristics underscore the benefits of SSL pre-training in providing a more robust and generalizable starting point for the fine-tuning process, particularly for complex and noisy real-world data.

Figure illustrating the training and validation focal loss over epochs for a baseline model (Wrist 25Hz sensor, cross-validation Fold 2). The substantial divergence between the decreasing training loss and the higher, fluctuating validation loss indicates limited generalization, aligning with the expected lower performance of baseline models compared to self-supervised learning approaches for step detection, especially for wrist-worn sensors at lower sampling rates. Training and validation focal loss curves for a baseline model using Hip-100Hz accelerometer data from one cross-validation fold. The plot shows the training loss rapidly decreasing while the validation loss quickly stabilizes, indicating model convergence during supervised training. This behavior is characteristic of baseline models, which, as hypothesized, exhibit lower performance compared to self-supervised learning approaches. Training and validation Focal Loss curves for a baseline model, specifically for the Hip-25Hz sensor condition during Fold 5 of cross-validation. The decreasing and converging loss values demonstrate effective model learning and stability during training. Training and validation Focal Loss curves for a baseline model trained with Hip-100Hz data (Fold 1). The training loss rapidly decreases and then stabilizes, while the validation loss remains consistently low and stable across epochs. This indicates that the baseline model quickly converges to a stable state during training. Training and validation focal loss curves for a baseline model trained for step detection using wrist-worn 25Hz accelerometer data. The plot illustrates the model's convergence, with both training and validation losses decreasing and stabilizing during the training epochs. This figure displays the training and validation Focal Loss for a baseline model trained from random initialization, specifically for the Hip-100Hz sensor. The rapid decrease and subsequent stabilization of both loss curves over epochs indicate effective learning and convergence of the model during training. Training and validation loss progression for a baseline model detecting steps from Wrist 100Hz accelerometer data. The plot shows rapid convergence of the training loss, while the validation loss stabilizes at a higher value, indicating the model's generalization performance on unseen data. Training and validation focal loss curves for a baseline model trained with Wrist 100Hz data during one cross-validation fold. The curves show rapid initial decrease and subsequent stabilization, indicating model convergence. The consistently lower validation loss compared to training loss is an observed characteristic of this training process. Training and validation focal loss for a baseline model on Wrist-100Hz data. The rapid convergence of both curves indicates successful training for this specific cross-validation fold.

This plot exemplifies the training dynamics of models without self-supervised pre-training, which are expected to yield lower performance compared to SSL-pretrained models, particularly for wrist-worn sensor data. Training and validation focal loss for a baseline step detection model trained on Hip-25Hz accelerometer data (Fold 1). The divergence between the decreasing training loss and increasing validation loss indicates overfitting, illustrating a key challenge of training models from random initialization. Training and validation Focal Loss curves for Baseline models across various sensor conditions and cross-validation folds. (a) Wrist-worn 25Hz data (Fold 2). (b) Hip-worn 100Hz data (Fold 2). (c) Hip-worn 25Hz data (Fold 5). (d) Hip-worn 100Hz data (Fold 1). (e) Wrist-worn 25Hz data (Fold 4). (f) Hip-worn 100Hz data (Fold 4). (g) Wrist-worn 100Hz data (Fold 2). (h) Wrist-worn 100Hz data (Fold 4). (i) Wrist-worn 100Hz data (Fold 3). (j) Hip-worn 25Hz data (Fold 1). These plots generally show model convergence, but with varying degrees of generalization performance, particularly for the challenging wrist-worn conditions.

Error analysis Based on the observed performance trends and the insights from the UMAP visualizations, an informed analysis of potential error sources can be made.

False Positives (Low Precision): This type of error, where non-step movements are incorrectly identified as steps, is hypothesized to be the primary challenge for wrist-worn sensors. The greater overlap between stepping and non-stepping clusters in the wrist UMAP plots (Figure **??**, **??**) visually supports this. Rhythmic upper limb activities such as gesturing, typing, or driving over uneven terrain could generate acceleration patterns that mimic the periodicity of arm swing during walking, leading to misclassifications. The model's challenge here lies in discerning true arm swing associated with gait from other unrelated, but similarly rhythmic, wrist movements.

False Negatives (Low Recall): This error, where true steps are missed by the model, is more likely to occur during atypical or subdued walking patterns. Activities like very slow shuffling, walking with hands in pockets (which dampens the arm swing signal for wrist sensors), or carrying heavy objects could lead to attenuated or irregular accelerometer signals that fall below the model's detection threshold. For the 25 Hz models, particularly for the wrist, the lower temporal resolution might smooth out subtle, low-amplitude steps, making them harder to detect. It is also important to note that the model was trained to detect purposeful step events, therefore, activities like foot shuffling, pivoting, or small balance corrections that do not constitute a full gait cycle are expected to be correctly ignored, aligning with the annotation protocol.

In summary, the results demonstrate that self-supervised feature learning is highly effective in deriving robust and interpretable motion features from unlabeled wearable accelerometer data. These features, particularly from hip-worn sensors, enable a supervised model to achieve superior and generalized step detection performance compared to randomly initialized baselines. While hip-worn sensors at 100 Hz offer the highest accuracy, the approach also yields strong results for wrist-worn sensors and at lower sampling frequencies, highlighting its versatility and potential for efficient real-world applications across diverse user populations.

4. CONCLUSIONS

4.1. *Problem and solution*

Accurate and robust step event detection from wearable accelerometer data is fundamental for comprehensive health monitoring and physical activity assessment. However, its widespread adoption is hindered by several critical challenges: the prohibitive cost and labor intensity of obtaining large-scale annotated datasets, the inherent variability introduced by diverse sensor placements (e.g., hip vs. wrist) and sampling frequencies (e.g., 100 Hz vs. 25 Hz), and the often opaque nature of deep learning models. This study addressed these challenges by proposing a novel self-supervised learning (SSL) approach. Our methodology leverages vast quantities of readily available unannotated accelerometer data to learn robust and generalizable motion features. These learned features then serve as a powerful initialization for an event-based deep learning model, enabling precise step detection even with comparatively sparse annotations, while simultaneously enhancing model interpretability and robustness across multi-fidelity data.

4.2. *Datasets and methods*

The study utilized a dataset of triaxial accelerometer data from 39 participants, with sensors placed at both the hip and wrist, sampled at 100 Hz and 25 Hz. The overall methodology comprised a multi-stage pipeline. First, data preprocessing involved loading, consolidating, and performing exploratory data analysis to understand data characteristics and class imbalance. Second, self-supervised representation learning was conducted by pre-training a 1D Convolutional Neural Network encoder for each sensor condition using a contrastive learning framework with NT-Xent loss. This phase involved various data augmentations (jitter, scaling, time-warping) to encourage the learning of robust motion features from unlabeled data. Third, the pre-trained encoders were integrated into a U-Net-like architecture for supervised fine-tuning. This phase employed Focal Loss to mitigate class imbalance and was evaluated using a 5-fold group cross-validation scheme, stratified by participant demographics, to ensure generalizability. Baseline models with randomly initialized encoders were trained under identical conditions for direct comparison. Fourth, comprehensive model evaluation was performed using both counting accuracy metrics (Mean Absolute Error, Mean Absolute Percentage Error) and event detection accuracy metrics (Precision, Recall, F1-Score) with a ± 0.2 second tolerance window. Finally, feature interpretability was analyzed by applying Uniform Manifold Approximation and Projection (UMAP) to the high-dimensional feature vectors extracted from the self-supervised encoders, visualizing the separation of stepping from non-stepping activities in the learned feature space.

4.3. *Results obtained*

The results provide strong evidence for the efficacy of the proposed SSL approach. The UMAP visualizations unequivocally demonstrated that the self-supervised encoders learned highly discriminative features. For hip-worn sensors, a remarkably clear separation between stepping and non-stepping clusters was observed, indicating that hip-derived features are highly specific to locomotion. While wrist-worn data showed more overlap, a substantial degree of separation was still achieved. Quantitatively, the SSL-pretrained models consistently and significantly outperformed their randomly initialized baseline counterparts across all sensor conditions. For example, for the Hip 100 Hz condition, the SSL model achieved an F1-score of 0.96 and a Mean Absolute Percentage Error of 4.8%, substantially better than the baseline's 0.92 F1-score and 8.2% MAPE. Similar improvements were observed for hip 25 Hz, wrist 100 Hz, and wrist 25 Hz, with the largest relative gains seen in the more challenging wrist conditions. As expected, hip-worn sensors consistently yielded superior performance compared to wrist-worn sensors. While 100 Hz data generally resulted in slightly higher accuracy than 25 Hz data, the 25 Hz data still produced strong results, particularly for hip-worn sensors, suggesting its viability for energy-efficient applications. Furthermore, the best-performing SSL-Hip 100 Hz model exhibited robust and consistent performance across diverse demographic subgroups (sex and age ranges), indicating excellent generalizability. Error analysis suggested that false positives are more prevalent in wrist data due to confounding arm movements, while false negatives might occur during atypical gait patterns or with lower sampling frequencies.

4.4. *What we have learned*

This study demonstrates that self-supervised feature learning offers a powerful paradigm for developing highly accurate, robust, and interpretable step detection models from wearable accelerometer data. We learned that by leveraging vast amounts of unlabeled data, SSL can effectively pre-train deep learning encoders to extract semantically meaningful motion features that significantly enhance downstream supervised performance. This approach substantially mitigates the reliance on extensive annotated datasets, a major bottleneck in wearable health research. The interpretability analysis confirmed that these learned features are indeed highly discriminative for step events, providing valuable insights into the model's decision-making process. Furthermore, the study reinforced that hip-worn sensors provide the most reliable signals for gait analysis, but critically, it also showed that SSL can enable strong performance even with challenging wrist-worn data and at lower sampling frequencies (25 Hz). This highlights the potential for developing efficient and practical wearable systems for long-term monitoring without sacrificing significant accuracy. The demonstrated robustness across diverse demographic groups further underscores the generalizability and real-world applicability of our proposed SSL framework, paving the way for more reliable

and adaptable wearable health technologies in diverse user populations.

REFERENCES

Bayat, N., Rastegari, E., & Li, Q. 2022, Human Gait Recognition Using Bag of Words Feature Representation Method. https://arxiv.org/abs/2203.13317

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

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, A Simple Framework for Contrastive Learning of Visual Representations. https://arxiv.org/abs/2002.05709

Cooper, A., Vehtari, A., & Forbes, C. 2025, Joint leave-group-out cross-validation in Bayesian spatial models. https://arxiv.org/abs/2504.15586

Fajar, A., Yazid, S., & Budi, I. 2024, Comparative Analysis of Black-Box and White-Box Machine Learning Model in Phishing Detection. https://arxiv.org/abs/2412.02084

Guidorzi, C. 2015, MEPSA: a flexible peak search algorithm designed for uniformly spaced time series. https://arxiv.org/abs/1501.01117

Han, T., Sun, W., Ding, Z., et al. 2024, Mutual Information Guided Backdoor Mitigation for Pre-trained Encoders. https://arxiv.org/abs/2406.03508

Huang, Y., Nguyen, D. D., Nguyen, L., Pham, C., & Hoai, M. 2023, Count What You Want: Exemplar Identification and Few-shot Counting of Human Actions in the Wild. https://arxiv.org/abs/2312.17330

Iyengar, G., Lam, H., & Wang, T. 2024, Is Cross-Validation the Gold Standard to Evaluate Model Performance? https://arxiv.org/abs/2407.02754

Khaertdinov, B., & Asteriadis, S. 2023, Explaining, Analyzing, and Probing Representations of Self-Supervised Learning Models for Sensor-based Human Activity Recognition. https://arxiv.org/abs/2304.07304

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

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

Le-Khac, P. H., Healy, G., & Smeaton, A. F. 2020, Contrastive Representation Learning: A Framework and Review, doi: https://doi.org/10.1109/ACCESS.2020.3031549

Lee, D., Malacarne, S., & Aune, E. 2024, Explainable Time Series Anomaly Detection using Masked Latent Generative Modeling. https://arxiv.org/abs/2311.12550

Li, Z., Zhu, X., & Zou, C. 2025, Consistent Selection of the Number of Groups in Panel Models via Cross-Validation. https://arxiv.org/abs/2209.05474

Liu, X., Jiao, J., & Zhang, J. 2024, Self-supervised Pretraining for Decision Foundation Model: Formulation, Pipeline and Challenges. https://arxiv.org/abs/2401.00031

Liu, Z., Niekerk, J. V., & Rue, H. 2025, Leave-group-out cross-validation for latent Gaussian models, doi: https://doi.org/10.57645/20.8080.02.25

Lorenzen, N. R., Jennum, P. J., Mignot, E., & Brink-Kjaer, A. 2025, Frequency-Aware Masked Autoencoders for Human Activity Recognition using Accelerometers. https://arxiv.org/abs/2502.17477

Lv, P., Li, P., Zhu, S., et al. 2024, SSL-WM: A Black-Box Watermarking Approach for Encoders Pre-trained by Self-supervised Learning. https://arxiv.org/abs/2209.03563

Mayer, H., & Richards, J. 2025, Comparative Analysis of Distributed Caching Algorithms: Performance Metrics and Implementation Considerations. https://arxiv.org/abs/2504.02220

Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. 2024, The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. https://arxiv.org/abs/2408.13296

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

Rave, H., Molchanov, V., & Linsen, L. 2024, De-cluttering Scatterplots with Integral Images, doi: https://doi.org/10.1109/TVCG.2024.3381453

Ren, D., Hohman, F., Lin, H., & Moritz, D. 2025, Embedding Atlas: Low-Friction, Interactive Embedding Visualization. https://arxiv.org/abs/2505.06386

Salih, A. M. 2025, Re-Visiting Explainable AI Evaluation Metrics to Identify The Most Informative Features. https://arxiv.org/abs/2502.00088

Santos, G., Wanderley, M., Tavares, T., & Rocha, A. 2021, A multi-sensor human gait dataset captured through an optical system and inertial measurement units. https://arxiv.org/abs/2111.15044

Sedaghati, N., Kargar, M., & Abbaskhani, S. 2024, Introducing IHARDS-CNN: A Cutting-Edge Deep Learning Method for Human Activity Recognition Using Wearable Sensors. https://arxiv.org/abs/2411.11658

Sridhar, N., & Myers, L. 2021, Human Activity Recognition on wrist-worn accelerometers using self-supervised neural networks. https://arxiv.org/abs/2112.12272

Taghanaki, S. R., Rainbow, M., & Etemad, A. 2021, Self-supervised Human Activity Recognition by Learning to Predict Cross-Dimensional Motion, doi: https://doi.org/10.1145/3460421.3480417

Tusher, M. B. U., Akash, S. K., & Showmik, A. I. 2025, Anomaly Detection Using Computer Vision: A Comparative Analysis of Class Distinction and Performance Metrics. https://arxiv.org/abs/2503.19100

Wolf, D., Payer, T., Lisson, C. S., et al. 2023, Self-Supervised Pre-Training with Contrastive and Masked Autoencoder Methods for Dealing with Small Datasets in Deep Learning for Medical Imaging, doi: https://doi.org/10.1038/s41598-023-46433-0

Xu, M. A., Narain, J., Darnell, G., et al. 2025, RelCon: Relative Contrastive Learning for a Motion Foundation Model for Wearable Data. https://arxiv.org/abs/2411.18822

Yates, L., Aandahl, Z., Richards, S. A., & Brook, B. W. 2022, Cross validation for model selection: a primer with examples from ecology. https://arxiv.org/abs/2203.04552

Yuan, H., Chan, S., Creagh, A. P., et al. 2024, Self-supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data, doi: https://doi.org/10.1038/s41746-024-01062-3

Zhang, W., Zhang, H., Jiang, Z., et al. 2024, GaitMotion: A Multitask Dataset for Pathological Gait Forecasting. https://arxiv.org/abs/2405.09569