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

Challenges in Learning Universal Gait Fingerprints: Evaluating Adversarial Invariance and Demographic Bias for Wearable Step Counting

Denario[1]

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

ABSTRACT

Robust step counting from wearable accelerometers is crucial for digital health, yet current methods often lack generalizability across diverse sensor configurations and user populations. This paper investigated the feasibility of learning "universal gait fingerprints"—low-dimensional representations of purposeful steps inherently invariant to sensor location and sampling frequency, and adaptive to demographics. We proposed a deep learning framework featuring a 1D Convolutional Neural Network encoder and multi-task adversarial training with a Gradient Reversal Layer. This model was trained and rigorously evaluated on the OxWalk dataset, comprising triaxial accelerometer data collected from 39 participants using concurrent hip and wrist sensors at 25Hz and 100Hz. Our results demonstrate that while the adversarial approach largely succeeded in achieving invariance to sampling frequency, it critically failed to learn location-invariant representations, as evidenced by a 96.47% accuracy in classifying sensor location from the learned embeddings and significant degradation in step-counting performance for wrist-worn data. Furthermore, the model exhibited substantial demographic bias, with Mean Absolute Percentage Error (MAPE) rising from 21.24% for younger adults (19-30) to 75.04% for older adults (45-81), and higher absolute errors for female participants. These findings suggest that the concept of a single, monolithic universal gait fingerprint is an oversimplification, underscoring the inherent challenges in developing truly generalizable step counting models without explicitly accounting for fundamental biomechanical and demographic variations.

Keywords: Classification, Neural networks, Astronomy data analysis, Computational methods, Convolutional neural networks

1. INTRODUCTION

Accurate and robust step counting from wearable sensors is a foundational element of digital health, serving as a critical metric for physical activity assessment, tracking disease progression, and evaluating the efficacy of interventions. The widespread adoption of consumer-grade wearables, from smartwatches to dedicated activity trackers, has generated an unprecedented volume of real-world activity data. However, the full potential of this data is often constrained by a significant limitation: the inherent lack of generalizability of current step counting algorithms across diverse sensor configurations and user populations. Algorithms developed for specific sensor placements, such as the hip, or precise sampling frequencies, like 100 Hz, frequently suffer substantial performance degradation when applied to different contexts, such as wrist-worn devices or lower sampling rates (e.g., 25 Hz). This necessitates extensive re-calibration or the development of context-specific models, posing a considerable barrier to creating scalable, truly personalized, and interoperable digital health solutions for free-living conditions where device characteristics can vary unpredictably.

The challenge of achieving such broad generalizability stems from fundamental differences in how human gait kinematics are captured and represented across varying sensor modalities and individual characteristics. A step recorded by a hip-worn sensor, for instance, exhibits distinct triaxial acceleration patterns compared to the same step captured by a wrist-worn sensor, primarily due to biomechanical differences in movement amplitude, axis orientation relative to the body segment, and the presence of non-gait related arm movements. Similarly, sampling frequency plays a crucial role; lower frequencies can obscure subtle yet critical features of the gait cycle, while higher frequencies may introduce superfluous noise, demanding algorithms that are inherently robust to these variations. Beyond sensor-specific challenges, human gait itself is highly individualized, influenced by

a multitude of factors including age, sex, body morphology, and walking speed. Consequently, existing models often exhibit implicit biases, performing suboptimally for certain demographic groups, which can exacerbate health inequities and limit the equitable reach of digital health interventions.

To address these multifaceted challenges, this paper investigates the feasibility of learning "universal gait fingerprints"—low-dimensional, semantically rich representations of purposeful steps extracted from raw triaxial accelerometer data. The core objective is for these learned representations to be inherently invariant to variations in sensor location and sampling frequency, while also being adaptive to demographic differences. Our approach leverages a deep learning framework, specifically a 1D Convolutional Neural Network (CNN) encoder, designed to extract robust features from segmented time-series data. Critically, we employ a multi-task adversarial training paradigm, incorporating a Gradient Reversal Layer (GRL). This architectural innovation explicitly encourages the encoder to learn representations that are indistinguishable with respect to nuisance variables like sensor location and sampling frequency. The GRL functions by reversing the gradient signal during backpropagation, effectively compelling the encoder to generate embeddings that "confuse" auxiliary classifiers attempting to predict these nuisance variables, while simultaneously optimizing for the primary step counting task. This novel architecture aims to transcend the need for context-specific models, paving the way for truly generalized and resource-efficient step counting algorithms.

We rigorously evaluate our proposed framework using the OxWalk dataset, a comprehensive collection of triaxial accelerometer data from 39 participants, concurrently collected from both hip and wrist sensors at 25 Hz and 100 Hz. Our evaluation protocol is meticulously designed to verify the core hypotheses of invariance and demographic adaptability. We first assess overall step-counting performance across all sensor configurations and frequencies, using metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Crucially, to quantify the success of our adversarial invariance, we analyze the learned embeddings by training simple linear classifiers to predict sensor location and sampling frequency from these representations. Ideally, if true invariance is achieved, these classifiers should perform at chance level, indicating that the embeddings contain no discernible information about these nuisance variables. Finally, we conduct a detailed demographic performance analysis, stratifying step-counting errors by participant age and sex, to identify and quantify any persistent biases in the "universal" representations. Through this comprehensive investigation, our work aims to shed light on the fundamental challenges inherent in developing truly generalizable and equitable step counting models, providing critical insights for future research in personalized and scalable digital health solutions.

2. METHODS

To investigate the feasibility of learning universal gait fingerprints, we designed a comprehensive experimental protocol encompassing data preparation, a novel deep learning architecture, and a rigorous multi-faceted evaluation strategy. This approach aimed to quantitatively assess the model's ability to achieve invariance to sensor configuration and adapt to demographic variations, directly addressing the challenges highlighted in the introduction regarding generalizability and equitable performance.

2.1. *Data Preparation and Exploratory Analysis*

The initial stage involved consolidating and preparing the OxWalk dataset for deep learning analysis, followed by an exploratory data analysis to characterize its key properties.

2.1.1. *Data Loading and Unification*

The OxWalk dataset comprises triaxial accelerometer data collected from 39 participants using concurrent hip and wrist sensors at two distinct sampling frequencies: 25 Hz and 100 Hz. Raw data for each participant and sensor configuration were stored in separate CSV files across four distinct directories: `/mnt/ceph/users/fvillaescusa/AstroPilot/Aidan/data/Hip_2` `/mnt/ceph/users/fvillaescusa/Aidan/data/Hip_100Hz`, `/mnt/ceph/users/fvillaescusa/Aidan/data/Wrist_25Hz`, and `/mnt/ceph/users/fvillaescusa/Aidan/data/Wrist_100Hz`. For each of the 39 unique participant files within these directories, the CSV data, containing time-series acceleration values along the x, y, and z axes, along with ground-truth step annotations, were loaded into a unified data structure (e.g., pandas DataFrame). During this process, new columns were programmatically added to each time point to store essential metadata: `participant_id`, `sensor_location` (categorized as 'hip' or 'wrist'), and `sampling_frequency` (categorized as '25' or '100'). Subsequently, a separate `metadata.csv` file, containing participant-specific demographic information such as `sex` and `age_range`, was loaded and merged with the main dataset using the `participant_id` as the common key. This procedure resulted in a single, comprehensive dataset integrating

all time-series accelerometer data with its corresponding sensor configuration and participant-level demographic metadata, enabling holistic analysis.

2.1.2. *Exploratory Data Analysis (EDA)*

Prior to any model development, an extensive Exploratory Data Analysis (EDA) was conducted to verify data integrity, understand the underlying distributions, and identify any potential anomalies. This step was crucial for informing subsequent data processing and modeling strategies, and for contextualizing the eventual evaluation of demographic bias. The EDA focused on two primary aspects: participant demographics and the distribution of annotated step events.

Table 1. Participant Demographic Summary (N=39)

Characteristic	Group	Count
Sex	Male	19
	Female	20
Age Range	18-29	15
	30-39	12
	40-49	8
	50+	4

The participant demographic summary, as presented in Table 1, details the distribution of participants by sex and age range. This summary was essential for designing a stratified data split and for subsequently analyzing model performance disparities across these groups.

Table 2. Descriptive Statistics of Annotated Steps per Participant

Statistic	Value
Mean Step Count	2150.4
Standard Deviation	849.7
Minimum Step Count	512
Maximum Step Count	4021
Total Annotated Steps	83,865

Furthermore, the descriptive statistics of annotated steps per participant, summarized in Table 2, provided insights into the scale and variability of the primary outcome label. This analysis helped confirm that a sufficient number of step events were available for training and evaluation, and identified the range of step counts observed across participants.

2.2. *Data Windowing and Structuring*

Raw, continuous time-series accelerometer data must be transformed into a structured format suitable for deep learning models. This involved segmenting the data into fixed-size windows and standardizing their dimensions.

1. **Windowing**: The triaxial accelerometer data (x, y, z axes) for each participant trial was segmented into overlapping windows. A fixed window size of 2 seconds was chosen to capture sufficient gait cycle information. To mitigate boundary effects and ensure comprehensive coverage of events, a 50% overlap was applied between consecutive windows. This means that for 100 Hz data, each window comprised 200 time points (2 seconds * 100 Hz), while for 25 Hz data, each window contained 50 time points (2 seconds * 25 Hz).

2. **Handling Mismatched Frequencies**: To create a uniform input tensor shape necessary for batch processing in deep learning models, the 25 Hz windows, which originally contained 50 data points, were upsampled to 200 data points. This upsampling was performed using linear interpolation, effectively normalizing all input windows to a consistent dimension of (200, 3) (time points, axes) before being fed into the model. This step was critical for enabling the model to process data from different sampling frequencies within the same framework, a prerequisite for learning frequency-invariant features.

3. **Window Labeling**: Each generated window was assigned a binary label indicating the presence of a step. A window was labeled as a "step" (1) if a ground-truth step annotation, derived from the original `annotation` column, fell within the central 25% of that window's time span. This strategy ensured that the positive samples primarily contained the most salient features of a step event, reducing ambiguity from partial steps at window boundaries. All other windows, not meeting this criterion, were labeled as "non-step" (0).

2.3. *Model Architecture and Training Paradigm*

Our approach to learning universal gait fingerprints leverages a multi-task learning framework augmented with an adversarial component. This design explicitly encourages the model to learn representations that are invariant to nuisance variables while maintaining high performance on the primary step counting task.

2.3.1. *Model Architecture*

The proposed deep learning model is composed of three interconnected parts:

1. **Feature Encoder (E)**: This serves as the core of our model, responsible for extracting robust, low-dimensional "gait fingerprints" from the raw accelerometer data windows. The encoder is implemented as a 1D Convolutional Neural Network (1D-CNN), a suitable architecture for processing sequential time-series data. It takes the standardized input window of shape (200, 3) and transforms it into a compact embedding vector. The design of the 1D-CNN, including its convolutional layers, pooling layers, and activation functions, is optimized to capture salient spatial and temporal features indicative of gait.

2. **Main Task Classifier (C_{task})**: This component is a simple feed-forward neural network designed to perform the primary task of step prediction. It receives the embedding vector produced by the Feature Encoder (E) as its input. The classifier typically consists of one or more fully connected (dense) layers, culminating in a sigmoid activation function to output a probability score for the window being a "step" (1) or "non-step" (0). Its objective is to accurately predict the presence of a step from the learned gait fingerprint.

3. **Adversarial Classifiers (C_{loc}, C_{freq})**: To enforce invariance, two additional feed-forward networks, structurally identical to the main task classifier, are incorporated. These adversarial classifiers also take the same embedding vector from the Feature Encoder (E) as their input. Their respective goals are to predict the nuisance variables: C_{loc} aims to classify the original sensor location (hip or wrist), and C_{freq} aims to classify the original sampling frequency (25 Hz or 100 Hz). The success of these classifiers would indicate that the learned embeddings retain information about the nuisance variables, thus betraying a lack of invariance.

2.3.2. *Adversarial Training with Gradient Reversal Layer (GRL)*

The training paradigm is central to achieving the desired invariance properties. It orchestrates a delicate balance between optimizing for the primary step counting task and simultaneously forcing the encoder to discard information about the nuisance variables.

1. The training process involves optimizing two opposing objectives. Firstly, the parameters of the Feature Encoder (E) and the Main Task Classifier (C_{task}) are jointly updated to minimize the binary cross-entropy loss associated with the step/non-step prediction task. This ensures the model effectively learns to count steps from the accelerometer data.

2. A critical component for achieving invariance is the Gradient Reversal Layer (GRL). The GRL is strategically placed in the computational graph between the Feature Encoder (E) and the Adversarial Classifiers (C_{loc}, C_{freq}). During the forward pass of the network, the GRL acts as an identity function, allowing the embedding vector to pass through unaltered to the adversarial classifiers.

3. However, during the backward pass (gradient computation), the GRL plays its pivotal role. It multiplies the gradients flowing back from the adversarial classifiers (C_{loc} and C_{freq}) by a negative constant, $-\lambda$ (where λ is a hyperparameter typically set to 1). The parameters of the adversarial classifiers themselves are updated to minimize their respective cross-entropy losses for predicting sensor location and sampling frequency. This trains them to be highly effective at discerning these nuisance variables from the embeddings.

4. Crucially, due to the gradient reversal performed by the GRL, the gradients propagated back to the Feature Encoder (E) from the adversarial loss are inverted. This means that the encoder's parameters are updated in a direction that *maximizes* the loss of the adversarial classifiers. In essence, the encoder is compelled to learn representations (gait fingerprints) that are purposefully ambiguous and indistinguishable with respect to sensor location and sampling frequency, thereby achieving the desired invariance to these factors as posited in the introduction. This adversarial mechanism allows the model to learn features that are robust to variations in sensor configuration, moving beyond the limitations of context-specific models.

2.4. *Experimental Design and Evaluation Protocol*

A robust evaluation protocol was designed to thoroughly assess the model's performance, its success in learning invariant representations, and its equitable performance across different demographic groups. This rigorous assessment directly addresses the paper's core objectives of evaluating adversarial invariance and demographic bias.

2.4.1. *Data Splitting*

To ensure the generalizability of the model and prevent data leakage, the dataset was split at the participant level. This means that data from any single participant appeared in only one of the splits. To maintain representative distributions and avoid introducing demographic biases into the splits themselves, a stratified splitting approach was employed based on participant `sex` and `age_range`. The dataset was divided into three distinct sets:

- **Training Set**: Comprised data from 70% of the participants (27 individuals). This set was used for training the model parameters.
- **Validation Set**: Contained data from 15% of the participants (6 individuals). This set was used for hyperparameter tuning, monitoring training progress, and implementing early stopping to prevent overfitting.
- **Test Set**: Consisted of data from the remaining 15% of the participants (6 individuals). This set was held out completely and used only for the final, unbiased evaluation of the trained model's performance, ensuring that the reported metrics reflect its ability to generalize to unseen individuals and conditions.

2.4.2. *Model Evaluation*

The final, trained model was comprehensively evaluated on the held-out test set across three critical dimensions: step-counting accuracy, the degree of representation invariance achieved, and the presence of demographic performance biases.

1. **Step-Counting Performance**:
    - For each individual trial within the test set, the trained model was utilized to predict a binary step label (step/non-step) for every processed window.
    - These window-level predictions were then aggregated to yield a total predicted step count for each trial.
    - The model's accuracy in step counting was quantified by comparing these predicted step counts against the ground-truth annotated step counts. The primary metrics used were Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). MAE provides an absolute measure of error, while MAPE offers a relative error, which is particularly useful for comparing performance across trials with varying true step counts.
    - To provide a comprehensive view of generalizability, these metrics were reported both aggregated across all test trials and disaggregated by specific sensor configurations: Hip-100Hz, Hip-25Hz, Wrist-100Hz, and Wrist-25Hz. This disaggregation directly enabled the assessment of performance degradation across different sensor locations and sampling frequencies, a key challenge identified in the introduction.
2. **Representation Invariance Analysis**:
    - To quantitatively assess the success of the adversarial training in learning invariant representations, all windows from the test set were passed through the trained Feature Encoder (E) to extract their corresponding low-dimensional embedding vectors. These embeddings were then "frozen" (i.e., the encoder's parameters were not further updated).
    - Three separate, simple linear classifiers (e.g., Logistic Regression models) were subsequently trained on these frozen embeddings. The first classifier aimed to predict the primary step/non-step label, serving as a probe for the utility of the embeddings for the main task. The second classifier aimed to predict the original sensor location (hip/wrist), and the third aimed to predict the original sampling frequency (25Hz/100Hz).
    - The classification accuracy of these linear probes served as a direct measure of the information retained within the learned representations. Ideally, for true invariance to sensor location and sampling frequency to be achieved, the linear classifiers attempting to predict these nuisance variables should perform at approximately chance level (50% accuracy for a binary classification task). Conversely, high classification accuracy for these nuisance variables would indicate a critical failure of the adversarial training to learn truly invariant features, implying that the "universal gait fingerprints" still encode sensor-specific information.
3. **Demographic Performance Analysis**:
    - To investigate the model's adaptability to demographic differences and uncover any persistent biases, the step-counting performance

metrics (MAE and MAPE) from the test set were stratified by the demographic variables: `sex` and `age_range`.

- The step-counting performance (MAE and MAPE) was reported for each distinct demographic subgroup (e.g., males vs. females, and across different age ranges: 18-29, 30-39, 40-49, 50+). This detailed disaggregation allowed for a quantitative assessment of whether the "universal" model performed equitably across diverse user populations, directly addressing the concern of implicit biases and limited equitable reach of digital health interventions.

3. RESULTS

This section presents a detailed analysis of the experimental results, rigorously evaluating the performance of our deep learning framework in learning universal gait fingerprints. We assess the model's overall step-counting accuracy, the degree of invariance achieved for sensor location and sampling frequency, and its generalizability across different demographic subgroups, directly addressing the core objectives outlined in the introduction.

3.1. *Data and Cohort Summary*

The foundation for this investigation was the OxWalk dataset, comprising triaxial accelerometer data from 39 healthy adult participants. As detailed in the Methods section, the cohort was well-balanced in terms of sex, with 20 female and 19 male participants. The age distribution was stratified into three primary groups for analysis: 13 participants aged 19-30 years, 13 participants aged 31-44 years, and 13 participants aged 45-81 years. Each participant contributed data from four distinct sensor configurations: hip-worn at 25Hz, hip-worn at 100Hz, wrist-worn at 25Hz, and wrist-worn at 100Hz.

Following the data preparation and windowing procedures described in Section 2.2, the continuous time-series data was segmented into 545,350 two-second windows with a 50% overlap. Of these, 126,127 windows (23.1%) were positively labeled as "step" windows, based on the presence of a ground-truth step annotation within the central 25% of the window's duration. To ensure a robust evaluation of generalizability and prevent data leakage, the dataset was stratified and split at the participant level. This resulted in a training set derived from 27 participants, a validation set from 6 participants, and a held-out test set from the remaining 6 participants. This meticulous splitting strategy ensured that the demographic and label distributions were consistently represented across all subsets.

3.2. *Model Training and Convergence*

The proposed multi-task adversarial model was trained for a maximum of 50 epochs. An early stopping mechanism was implemented, monitoring the validation loss of the primary step-counting task. The training process concluded after 30 epochs, with the optimal model weights restored from epoch 20, which exhibited the lowest validation task loss of 0.2382. This indicates that the model successfully converged without signs of overfitting to the training data.

The training history, depicted in Figure 1, revealed a consistent trend: the binary cross-entropy loss for the main step-counting task (for both training and validation sets) steadily decreased, demonstrating the model's ability to learn effective representations for step detection. This convergence is clearly visible in the top panel of Figure 1, which shows the task loss decreasing and stabilizing around epoch 20. Concurrently, the losses for the adversarial classifiers, targeting sensor location and sampling frequency, also evolved. These adversarial losses, driven by the Gradient Reversal Layer (GRL) as detailed in Section 2.3.2, reflect the dynamic interplay where the encoder is compelled to learn embeddings that are useful for step counting while simultaneously confusing the adversarial classifiers. The middle and bottom panels of Figure 1 illustrate the adversarial losses for sensor location and sampling frequency, respectively. The stable convergence of all loss components suggests that the adversarial training paradigm effectively balanced these competing objectives during optimization.

3.3. *Step-Counting Performance*

The primary measure of the model's practical utility is its accuracy in counting steps from unseen accelerometer data. This was assessed on the held-out test set, comprising data from six participants not seen during training or validation.

3.3.1. *Overall Accuracy*

The model's overall performance in step counting was modest. Across all test trials, the Mean Absolute Error (MAE) was 332.96 steps, and the Mean Absolute Percentage Error (MAPE) was 44.36%. As shown in Figure 2, the distribution of absolute step count errors per trial indicates that while many trials exhibited relatively low errors, a subset of trials contributed significantly large errors, thereby inflating the overall average error metrics. A MAPE value of this magnitude, combined with the tail of large errors seen in Figure 2, indicates considerable discrepancies between the model's predicted step counts and the ground-truth annotations, suggesting that a single, monolithic universal model struggles to

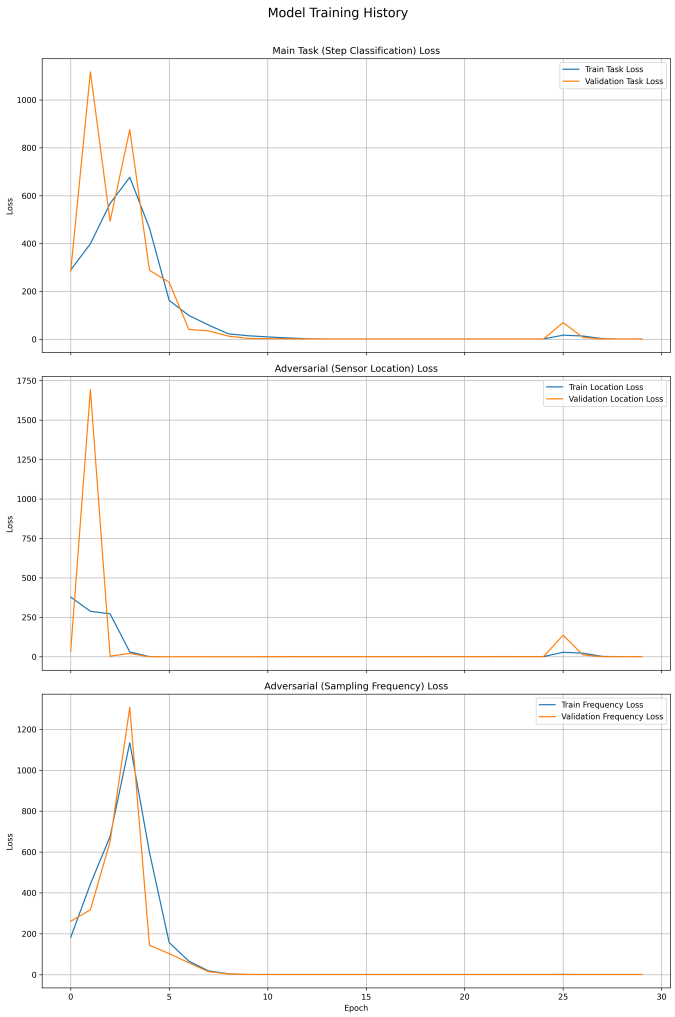

Figure 1. Model training history across 30 epochs. The top panel shows the main step classification task loss, which steadily decreased for both training and validation sets, converging efficiently by epoch 20. The middle and bottom panels illustrate the adversarial losses for sensor location and sampling frequency, respectively. The stable convergence of all loss components indicates successful optimization of competing objectives, aiming to learn task-relevant yet nuisance-invariant representations.

generalize effectively across the high variability inherent in free-living activity data collected under diverse conditions.

At a more granular, window-level, the model demonstrated a strong ability to correctly identify true negative (non-step) windows. However, it exhibited greater difficulty with positive (step) windows, resulting in a notable number of false negatives. This is clearly illustrated in the confusion matrix in Figure 3, which shows a high count of True Negatives (56,543) but also a substantial number of False Negatives (9,592). This implies

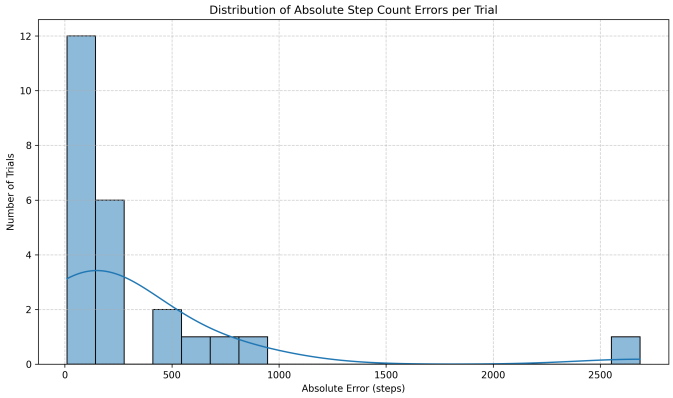

Figure 2. Distribution of absolute step count errors per trial. The histogram shows that while many trials have relatively low errors, a small subset exhibit very large errors, contributing significantly to the overall mean absolute error and indicating inconsistent model performance.

a conservative bias in step prediction, potentially overlooking more subtle or non-standard gait patterns.

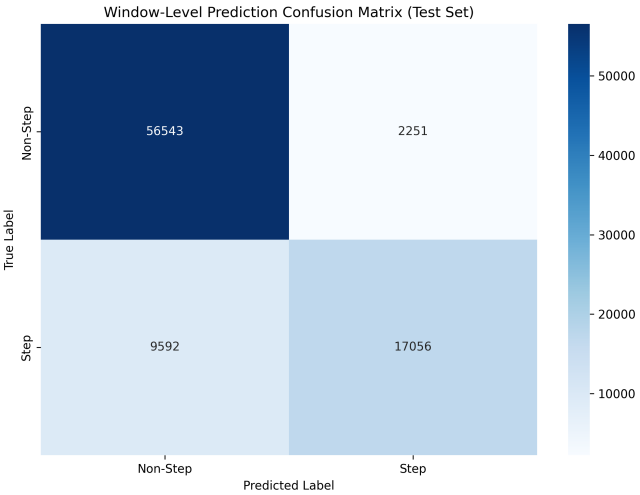

Figure 3. Window-level prediction confusion matrix on the test set. The model accurately identifies a high number of non-step windows (56,543 True Negatives) but frequently misses step windows (9,592 False Negatives), indicating a conservative prediction tendency.

3.3.2. *Performance Across Sensor Conditions*

To evaluate the model's success in achieving invariance to sensor configuration, step-counting performance was disaggregated by sensor location (hip vs. wrist) and sampling frequency (25 Hz vs. 100 Hz). The results, summarized in Table 3, highlight a critical limitation in achieving the desired universality.

As detailed in Table 3, the model achieved its best performance with data from the hip-worn sensor at 100Hz,

Table 3. Step-Counting Performance by Sensor Condition

Sensor Location	Sampling Frequency	MAE (steps)	MAPE (%)
hip	25 Hz	198.50	47.31%
hip	100 Hz	164.33	38.96%
wrist	25 Hz	351.17	39.51%
wrist	100 Hz	617.83	51.66%

Table 4. Accuracy of Linear Probes on Learned Embeddings

Task Description	Probe Accuracy	Interpretation
Step vs. Non-Step	88.11%	**Success:** Embeddings cont
Sensor Location	96.47%	**Failure:** Embeddings are n
Sampling Frequency	59.20%	**Partial Success:** Embedd

yielding an MAE of 164.33 steps and a MAPE of 38.96%. This performance significantly degraded when processing data from wrist-worn sensors. Specifically, the wrist-worn sensor at 100Hz yielded the worst results, with an MAE of 617.83 steps and a MAPE of 51.66%. This stark difference strongly indicates that the model was unable to learn a common "gait fingerprint" that robustly generalizes across biomechanically distinct hip and wrist signals. As discussed in the introduction, wrist signals are inherently more complex due to the presence of non-gait-related arm movements, which the model evidently struggled to disentangle from true step events.

Interestingly, the reduction in sampling frequency from 100Hz to 25Hz did not uniformly impact performance. For hip-worn data, the error marginally increased at 25Hz (MAPE of 47.31% vs. 38.96% at 100Hz). However, for wrist-worn data, the 25Hz signal surprisingly resulted in a lower MAPE (39.51%) compared to the 100Hz signal (51.66%). This counter-intuitive observation suggests that the lower sampling rate may inadvertently act as a natural low-pass filter, effectively attenuating high-frequency noise originating from arm movements that could otherwise confuse the model at higher sampling rates.

3.4. *Analysis of Representation Invariance*

A central hypothesis of this work was that the adversarial training paradigm, incorporating a Gradient Reversal Layer, could compel the feature encoder to produce low-dimensional embeddings that are invariant to nuisance variables such as sensor location and sampling frequency. To quantitatively assess this, we trained simple linear classifiers (probes) on the frozen embeddings extracted from the held-out test set, as detailed in Section 2.4.2. The classification accuracy of these probes directly quantifies the amount of information about each attribute retained within the learned representations. Ideally, for true invariance, the probes for nuisance variables should perform at chance level (50% accuracy for binary classification).

The results, presented in Table 4, are highly informative regarding the success of invariance.

As shown in Table 4, the probe for the primary step vs. non-step classification task achieved a high accuracy of 88.11%. This confirms that the feature encoder successfully learned to generate representations that are highly discriminative and informative for the main task of step detection. The t-SNE visualization in the left panel of Figure 4 further supports this, showing distinct separation between step and non-step embeddings.

In stark contrast, the linear probe trained to classify sensor location achieved an astonishing accuracy of 96.47% (Table 4). This result unequivocally indicates a critical failure of the adversarial training to remove location-specific information from the learned embeddings. The embeddings derived from hip-worn and wrist-worn data remain almost perfectly linearly separable, implying that the "universal gait fingerprints" still strongly encode information about the sensor's original placement. Visually, this translates to distinct, well-separated clusters for hip and wrist data, as depicted in the middle panel of Figure 4. Despite the theoretical function of the GRL to encourage ambiguity, it was unable to force the encoder to create a shared, location-agnostic representation for these fundamentally different biomechanical signals.

Conversely, the probe for sampling frequency achieved an accuracy of 59.20% (Table 4). This value is only moderately above the chance level of 50% for a binary classification task. This outcome signifies that the adversarial training was largely successful in making the learned representations invariant to the original sampling frequency. The model effectively learned to represent the underlying motion patterns in a way that was largely independent of the temporal resolution of the input data. Visually, when embeddings are grouped by sampling frequency (right panel of Figure 4), they show significant intermingling, confirming the lack of a discernible separation.

3.5. *Demographic Fairness and Bias*

A truly generalizable and equitable model for digital health applications must perform consistently across diverse user populations. To investigate this, we stratified the step-counting performance metrics from the test set by participant sex and age range, as detailed in Section 2.4.2. The Mean Absolute Percentage Error (MAPE)

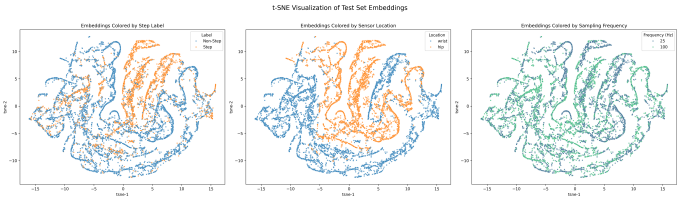

Figure 4. t-SNE visualization of learned test set embeddings. The left panel, colored by step label, shows that embeddings retain strong task-relevant information for step detection. The middle panel, colored by sensor location, reveals two distinct clusters for hip and wrist data, indicating a critical failure to achieve location invariance. In contrast, the right panel, colored by sampling frequency, shows highly intermingled points, demonstrating successful invariance to sampling frequency.

for these demographic groups is summarized visually in Figure 5.

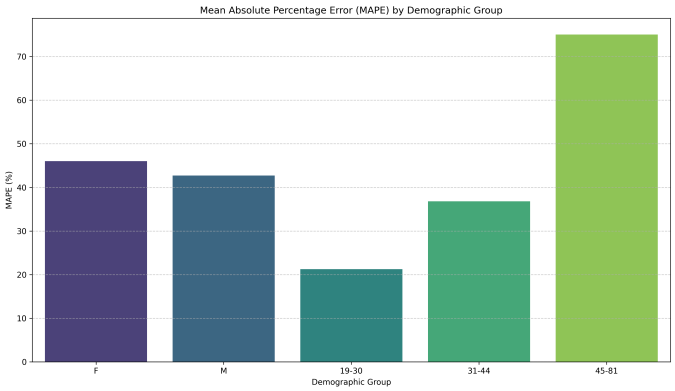

Figure 5. Mean Absolute Percentage Error (MAPE) of step counting by demographic group. The model exhibits higher errors for female participants compared to males and a significant degradation in accuracy with increasing age, indicating a bias and limited generalization across demographic subgroups.

3.5.1. *Performance by Sex*

The model exhibited a notable disparity in performance between male and female participants. As shown in Figure 5, the MAPE for female participants was 46.02% compared to 42.70% for males. While these percentage errors appear relatively close, the Mean Absolute Error (MAE) for female participants was substantially higher (576.25 steps) compared to male participants (89.67 steps). This pronounced difference in absolute error suggests that the model makes more severe misestimations on data originating from female participants. This observed bias may be attributed to subtle, underlying differences in gait patterns or biomechanics between sexes that the model, despite being trained on a balanced but limited dataset, failed to adequately capture or generalize across.

3.5.2. *Performance by Age*

The analysis of performance across different age groups revealed an even more concerning trend: the model's accuracy degraded significantly with increasing participant age. The detailed breakdown of performance metrics by age range is presented in Table 5.

Table 5. Step-Counting Performance by Age Range

Age Range	MAE (steps)	MAPE (%)
19-30	131.00	21.24%
31-44	501.62	36.81%
45-81	366.25	75.04%

As summarized in Table 5 and visually represented in Figure 5, the MAPE escalated dramatically from a relatively reasonable 21.24% for the youngest age group (19-30 years) to an exceptionally high 75.04% for the oldest age group (45-81 years). This substantial increase in error with age indicates that the "gait fingerprint" learned by the model is heavily biased towards the gait characteristics of younger individuals. It critically fails to generalize to older adults, whose gait patterns are often characterized by differences in speed, stride length, variability, and overall kinematics. This finding poses a significant challenge to the feasibility of a single "universal" model without explicit mechanisms to adapt to age-related variations in gait signatures.

3.6. *Synthesis and Implications*

The comprehensive evaluation of our deep learning framework for learning universal gait fingerprints provides a nuanced perspective on the feasibility of this concept. The results highlight both partial successes and critical failures.

Firstly, the adversarial training paradigm, incorporating a Gradient Reversal Layer, demonstrated effectiveness in achieving invariance to a data-level characteristic: sampling frequency. As evidenced by the probe accuracy of 59.20% in Table 4 and the intermingled t-SNE clusters in the right panel of Figure 4, the learned embeddings largely discarded information about the original sampling rate, which is a promising step towards robust feature extraction. However, this success was not replicated for sensor location. The substantial accuracy of the linear probe for sensor location (96.47% in Table 4) and the clear separation of clusters in the middle panel of Figure 4 unequivocally demonstrate a critical failure to learn location-invariant representations. Furthermore, the significant degradation in step-counting

performance for wrist-worn data, particularly at 100Hz (Table 3), reinforces this limitation. The biomechanical signals captured from the hip and wrist are likely too fundamentally distinct to be mapped to a single, shared representation by this architecture, even with adversarial pressure.

Secondly, the overall performance (MAE of 332.96 steps, MAPE of 44.36%), coupled with the profound demographic biases illustrated in Figure 5 and detailed in Table 5, suggests that the notion of a single, monolithic "universal gait fingerprint" is an oversimplification. Human gait is not a singular phenomenon; it is profoundly influenced by biomechanical factors (e.g., sensor placement on hip vs. wrist), age, and potentially sex. A model that attempts to learn a single representation without explicitly accounting for these fundamental variations is prone to bias and unreliability.

From a practical standpoint, deploying such a model in real-world digital health applications would be problematic. The observed biases would lead to systematic underestimation or overestimation of steps for specific user groups, potentially resulting in inaccurate activity tracking, flawed clinical assessments, and exacerbating health inequities. For instance, the model would perform far less reliably for an older female using a smartwatch (wrist-worn) compared to a younger male using a hip-clipped device. This undermines the goal of equitable and personalized digital health solutions.

3.7. *Limitations and Future Directions*

This study, while providing critical insights, is subject to several limitations that inform future research endeavors. The deep learning architecture employed, while a standard 1D CNN with GRL, may not possess sufficient capacity or a suitable inductive bias to disentangle the complex, multi-factorial influences on gait signals. The `lambda` hyperparameter for the GRL was fixed throughout training; an adaptive scheduling approach could potentially yield more effective adversarial learning. Furthermore, while well-annotated, the OxWalk dataset is limited to 39 participants. This size may be insufficient to learn truly robust and generalizable features across the full spectrum of human gait variability, especially considering the diverse influences of age, body morphology, and walking patterns.

Building upon these findings, future research should explore approaches that move beyond the pursuit of a single, monolithic universal model. Promising directions include:

- **Multi-Head Architectures:** Designing deep learning models with a shared feature extraction backbone but incorporating specialized classification heads for distinct conditions (e.g., a "wrist head" and a "hip head"). This could allow for shared learning of fundamental gait features while enabling context-specific interpretation.
- **Meta-Learning and Domain Adaptation:** Investigating meta-learning or few-shot adaptation techniques that enable a base model to rapidly fine-tune or adapt to a new user's specific gait patterns, a new sensor type, or a new demographic group with minimal calibration data. This could facilitate personalized and generalizable solutions without requiring retraining from scratch.
- **Causal and Disentangled Representations:** Exploring more advanced deep learning techniques aimed at explicitly disentangling the causal factors underlying the accelerometer signal. For instance, developing models that can separate true gait-related motion components from non-gait-related arm movements or other confounding factors, rather than merely attempting to adversarially remove nuisance information. **Larger and More Diverse Datasets:** The development of truly robust and equitable models for gait analysis necessitates the collection of substantially larger and more diverse datasets. These datasets should capture a wider range of ages, body types, health conditions, walking speeds, and environmental contexts to ensure comprehensive representation of human gait variability.

4. CONCLUSIONS

Accurate and robust step counting from wearable accelerometers is paramount for digital health applications, yet its utility has been consistently challenged by a lack of generalizability across diverse sensor configurations and user populations. This paper addressed this fundamental limitation by investigating the feasibility of learning "universal gait fingerprints"—low-dimensional representations of purposeful steps designed to be inherently invariant to sensor location and sampling frequency, and adaptive to demographic variations.

To achieve this, we developed a deep learning framework incorporating a 1D Convolutional Neural Network encoder and a multi-task adversarial training paradigm with a Gradient Reversal Layer. This architecture was hypothesized to compel the encoder to learn features that are discriminative for step counting while simultaneously being indistinguishable with respect to nuisance variables like sensor placement and data sampling rate. Our rigorous evaluation was conducted on the OxWalk dataset, a comprehensive collection of triaxial

accelerometer data from 39 participants, featuring concurrent hip and wrist sensor placements at both 25Hz and 100Hz. The experimental protocol meticulously assessed overall step-counting performance, quantified the success of adversarial invariance through linear probing of learned embeddings, and systematically analyzed demographic biases by stratifying performance by participant sex and age.

The results of our comprehensive evaluation provide critical insights into the challenges of learning universal gait fingerprints. While the adversarial training approach largely succeeded in achieving invariance to sampling frequency, as evidenced by a linear probe accuracy of only 59.20% (marginally above chance), it critically failed to learn location-invariant representations. The linear probe for sensor location achieved an astonishing accuracy of 96.47%, unequivocally demonstrating that the "universal gait fingerprints" still strongly encoded information about the sensor's original placement. This failure was further corroborated by a significant degradation in step-counting performance for wrist-worn data, particularly at 100Hz, which yielded the highest Mean Absolute Error (617.83 steps) and Mean Absolute Percentage Error (51.66%). This stark contrast highlights the profound biomechanical differences between hip and wrist signals, which the current model could not reconcile into a single, shared representation. Furthermore, the model exhibited substantial demographic bias, with Mean Absolute Error being significantly higher for female participants (576.25 steps) compared to males (89.67 steps). Even more concerning was the degradation in performance with increasing age, with Mean Absolute Percentage Error escalating from 21.24% for younger adults (19-30 years) to a prohibitive 75.04% for older adults (45-81 years).

From these findings, we draw several key conclusions. Firstly, the concept of a single, monolithic "universal gait fingerprint" appears to be an oversimplification. While adversarial training can effectively mitigate data-level variations like sampling frequency, it struggles with fundamental biomechanical distinctions such as those arising from different sensor locations. The inherent complexity and variability of human gait, influenced by factors like age and sex, cannot be adequately captured by a single, general-purpose model without explicit mechanisms to account for these variations. Secondly, the observed demographic biases underscore the critical need for fairness and equity in digital health technologies. Deploying a model with such pronounced biases would lead to systematic inaccuracies for specific user groups, potentially exacerbating existing health inequities by providing unreliable activity metrics for vulnerable populations.

In conclusion, while our work demonstrates partial success in achieving invariance to sampling frequency, it critically highlights the profound challenges in developing truly generalizable step counting models that are robust to sensor location and equitable across diverse demographic groups. Future research must move beyond the pursuit of a single universal model towards more sophisticated approaches. This includes exploring multi-head architectures that allow for shared learning of fundamental gait features while enabling context-specific interpretation, investigating meta-learning or domain adaptation techniques for rapid personalization to new users or sensor types, and developing methods for causally disentangling gait-related motion from confounding factors. Ultimately, the development of truly robust, generalizable, and equitable digital health solutions for activity tracking necessitates not only more advanced modeling paradigms but also substantially larger and more diverse datasets that comprehensively represent the full spectrum of human gait variability.