Skepthical review: *Cross-Configuration Transfer Learning Framework for Robust Step Counting in Free-Living Conditions*

Summary

This paper evaluates how well a step-counting model trained in one wearable "configuration" generalizes to others without any adaptation. A LightGBM regressor is trained on high-fidelity hip-worn accelerometer data sampled at **100** Hz (source) using a leave-one-subject-out (LOSO) protocol on **39** participants. Using synchronized hip and wrist accelerometers available at **100** Hz and **25** Hz, the trained source model is applied zero-shot to three target configurations (Hip **25** Hz, Wrist **100** Hz, Wrist **25** Hz). Features are extracted from **2** s windows (with **1** s stride) and step counting is posed as window-level regression; participant-level total steps are reconstructed from window predictions and evaluated mainly via MAE/MAPE (plus $\mathbf{Std_AE}$), with non-parametric tests (Wilcoxon signed-rank; Mann–Whitney U / Kruskal–Wallis for demographics). The core finding is that zero-shot cross-configuration transfer degrades strongly—especially for wrist placement and lower sampling frequency—manifesting as systematic underestimation and higher inter-individual variability. The experimental question is important and the paired, multi-configuration dataset is a clear asset; however, several methodological details (especially window labeling/aggregation under overlap, resampling/FFT feature definitions, and model configuration) are currently under-specified, and the paper's framing as a "transfer learning framework" is not yet supported by implemented adaptation baselines or broader comparator models.

Strengths

- Well-motivated deployment problem: wearable heterogeneity in sensor placement (hip vs. wrist) and sampling rate (**100** vs. **25** Hz) is a real barrier (Sec. 1, Sec. 2.4).
- Valuable dataset design: synchronized hip and wrist recordings across two sampling rates with expert-annotated step ground truth (Sec. 2.1).
- Realistic evaluation protocol: LOSO-CV for the source configuration and participant-level totals aligned with practical usage (Sec. 2.3.1, Sec. 3.1).
- Systematic target shifts: evaluates location shift, sampling-rate shift, and their combination (Sec. 2.4, Sec. 3.2–3.3).
- Clear empirical result: naive zero-shot transfer fails substantially, particularly for Wrist **25** Hz, with errors dominated by underestimation and increased participant-to-participant variability (Sec. 3.2–3.5).
- Use of paired non-parametric testing to assess degradation across configurations is directionally appropriate given the within-subject multi-configuration setup (Sec. 2.6.1, Sec. 3.2).

- Figures effectively convey monotonic degradation trends across configurations and highlight variability, which is important for deployment risk assessment (Fig. 1–3).

Major issues

1. **Ambiguity/possible invalidity in participant-level step reconstruction under overlapping windows. The method uses 2 s windows with 1 s stride (50% overlap) and states that participant total steps are obtained by summing window-level predictions (Sec. 2.2.1–2.2.2, Sec. 2.4.2, Sec. 3.1).** With overlapping windows, naive summation can double-count contributions unless the labeling/aggregation is defined to be overlap-consistent (e.g., each true step assigned to exactly one window, or overlap-corrected weighting). As written, it is unclear whether each ground-truth step can appear in multiple window labels; if so, both training targets and reconstructed totals become scale-dependent on the overlap choice, and cross-configuration comparisons may be biased.

   Recommendation: In Sec. 2.2–2.4, explicitly define (i) how step timestamps are mapped to windows under overlap (including boundary steps), (ii) whether a step can contribute to multiple window labels, and (iii) the exact aggregation used to recover participant totals from overlapping predictions. Ensure the aggregation is mathematically consistent (e.g., use non-overlapping windows for total reconstruction; or assign each step to a unique window; or apply a principled weighting such that summing yields an unbiased total). Add a small schematic/pseudocode example (appendix is fine) and explicitly confirm that Hip 100 Hz LOSO totals are computed only from held-out predictions using the same policy.

2. **Framing as a "transfer learning framework" is not matched by the implemented methods: the paper evaluates only zero-shot transfer of a single LightGBM model trained on Hip 100 Hz (Sec. 1, Sec. 2.3–2.4, Sec. 4).** This is still a useful robustness evaluation, but current narrative/title may overstate methodological novelty and generality.

   Recommendation: Either (A) reframe throughout (title, abstract, Sec. 1, Sec. 4) as a systematic evaluation/benchmark of zero-shot cross-configuration robustness for step counting, or (B) add at least one concrete adaptation baseline consistent with "transfer learning" (e.g., simple target calibration, CORAL/feature alignment, per-configuration normalization, small labeled fine-tuning, or multi-configuration training) and compare against zero-shot transfer.

3. **Insufficient specification of window labels, regression target, and output post-processing. It is unclear whether zero-step windows are included, how negative predictions are handled (regression can output negatives), whether predictions are rounded/clipped before aggregation, and how boundary cases are treated (Sec. 2.2.1–2.2.2, Sec. 2.4.2, Sec. 3.1).** These choices directly affect systematic underestimation and the participant-level totals.

Recommendation: In Sec. 2.2–2.4, document: inclusion/exclusion of zero-step windows; whether targets are integers or real-valued; whether model outputs are constrained (clip to ≥ 0, rounding strategy); and how boundary steps are counted. Report an ablation or sensitivity check showing how these choices affect MAE/MAPE and bias (especially underestimation) across configurations.

4. **Missing key modeling/training details hinder reproducibility and interpretation of failure modes. LightGBM hyperparameters, objective, number of trees/iterations, regularization, early stopping, feature handling (scaling/normalization), random seeds, and any hyperparameter tuning protocol are not adequately reported (Sec. 2.2.2, Sec. 2.3.1, Sec. 3.1).** It is also ambiguous whether cross-configuration results use one final model trained on all Hip 100 Hz data or fold-specific LOSO models reused for target inference.

   Recommendation: Add a dedicated subsection (Sec. 2.3) listing all LightGBM settings (e.g., boosting type, objective, \text{learning_rate}, \text{n_estimators}, \text{num_leaves}, \text{max_depth}, \text{min_data_in_leaf}, subsampling/feature_fraction, L1/L2), tuning method (defaults vs. search; nested CV or not), early stopping, and random seed(s). Clearly state whether target inference uses a single final Hip 100 Hz model or the ensemble of LOSO-trained folds; ensure the policy is consistent and justified.

5. **Sampling-rate handling (25 Hz) and frequency-feature definitions are under-specified, threatening comparability across configurations. It is unclear whether 25 Hz is natively recorded or downsampled from 100 Hz, and if downsampled whether anti-alias filtering was applied (Sec. 2.1, Sec. 2.2.2).** FFT-based features (dominant frequency, spectral energy) depend on sampling rate, window length, FFT normalization, and one-sided vs two-sided spectra; without precise definitions, feature scale may change purely due to sampling-rate differences, confounding "transfer failure" conclusions (Sec. 2.2.2, Sec. 3.2–3.3).

   Recommendation: In Sec. 2.1 and Sec. 2.2.2, specify the acquisition path for 25 Hz (native vs resampled) and the exact resampling pipeline (filter type/order/cutoff). Precisely define FFT computation and spectral energy (normalization, one- vs two-sided spectrum, scaling with N). Consider normalizing spectral features to be sampling-rate comparable (e.g., power spectral density or per-Hz normalization) and report whether such normalization reduces Hip 25 Hz degradation.

6. **Lack of critical baselines limits interpretability of the observed degradation. Without (i) in-domain models trained/evaluated within each target configuration and (ii) a simple signal-processing/heuristic step counter, it is unclear whether target settings are intrinsically harder or whether the degradation is primarily cross-configuration mismatch (Sec. 2.3–2.4, Sec. 3.2–3.3, Sec. 4).** Considering only one model family (LightGBM) and only one source domain (Hip 100 Hz) further limits generality.

Recommendation: Add: (1) LOSO within-configuration baselines for Hip **25** Hz, Wrist **100** Hz, Wrist **25** Hz (same features/model) to quantify "best achievable" within each configuration; (2) at least one simple heuristic baseline (e.g., bandpass + peak detection on SVM) per configuration; and ideally (3) a second model class (e.g., random forest or linear model) or an alternate source configuration (e.g., Wrist **100** Hz) to test whether failure modes are model/source-specific. If infeasible, narrow claims in Sec. 4 to this specific LightGBM/Hip-**100** setup and elevate missing baselines as a primary limitation.

7. **Error analysis is not deep enough to support some of the broader interpretations (including demographic conclusions). The paper emphasizes MAE/MAPE/Std_AE and qualitative underestimation claims but does not report signed bias, predicted-vs-true correlation, or error dependence on total steps/activity composition (Sec. 3.3–3.5).** Demographic analysis (sex, coarse age bins) is underpowered at $n = 39$ and "non-significant" results are described in a way that may be read as "equitable/independent," which is stronger than supported (Sec. 2.6.2, Sec. 3.4, Sec. 4).

   Recommendation: In Sec. 3.2–3.5, add: mean signed error (bias) and its CI; predicted vs. true scatter plots with Pearson/Spearman and/or R^2 per configuration; error vs. true total steps plots; and (if available) bout-level or intensity-stratified analysis to localize failure (missed bouts vs within-bout undercounting). For demographics, report effect sizes (e.g., rank-biserial correlation, eta-squared) and confidence intervals, and rephrase conclusions as "no evidence detected" given limited power; optionally use regression/ANCOVA controlling for total steps or other covariates if available.

8. **Dataset/protocol description is insufficient to judge generalizability. Critical context is missing: recording duration (per participant, range), number of sessions, free-living vs scripted activities, activity mix, device make/model and dynamic range, placement/orientation details, non-wear handling, and annotation procedure quality (synchronization, number of annotators, reliability) (Sec. 2.1, Sec. 2.1.1).**

   Recommendation: Expand Sec. 2.1–2.1.1 with: device specs (make/model, range), placement/orientation (hip side, wrist side), duration and sessions statistics, activity contexts and mix, missing-data/non-wear criteria, and ground-truth annotation workflow (synchronization method, annotators, inter-rater agreement if available). In Sec. 4, explicitly scope generalization claims to similar populations/devices/contexts.

9. **Related work and positioning are not sufficiently focused on cross-placement/cross-device step counting and wearable domain adaptation; some citations appear tangential (Sec. 1, Sec. 4).** This weakens the "bigger picture" justification and makes it harder to see what is new beyond the empirical finding that naïve transfer fails.

Recommendation: Add a structured Related Work section (Sec. 1 or new Sec. 2.x) covering: (a) traditional pedometer/step-counting on hip vs wrist, (b) prior cross-placement generalization studies in step counting/activity recognition, and (c) wearable domain adaptation/transfer learning methods. Replace/relocate tangential references unless directly connected to the paper's methods or hypotheses. Clearly state the paper's novelty (dataset setup + evaluation protocol + quantified degradation) and limit claims accordingly.

Minor issues

1. Figures do not fully exploit the paired design (same participants across configurations) and omit key context (units, sample sizes). Figures **1** and **3** would be more informative if they displayed within-participant changes (paired slopes/differences), and statistical significance is not annotated directly. Figures **2–3** lack explicit units and/or aggregation clarity (per participant totals vs window-level) (Fig. 1–3, Sec. 3.2–3.5).

   Recommendation: Revise Fig. 1–3 to include paired plots (slope/spaghetti or paired-difference plots), annotate n per group/configuration, add units (steps, %, etc.), and clarify the aggregation level in captions. Optionally annotate corrected p-values/significance markers on the figures for standalone readability.

2. Statistical testing description/reporting is incomplete: unclear test-statistic definitions (e.g., Wilcoxon "Statistic" field), inconsistent p-value formatting (e.g., **0.0000**), and incomplete multiple-comparisons accounting (how many tests per family; why Bonferroni vs FDR) (Sec. 2.6.1–2.6.2, Tables 2–4).

   Recommendation: In Sec. 2.6 and table captions, define the reported statistic ($W/Z/etc.$), report p-values in scientific notation or thresholds (e.g., $p < 1e{-}4$), and explicitly state correction method, number of comparisons, and corrected α per test family. Ensure text claims match corrected p-values.

3. MAPE/APE edge cases are not specified: APE is undefined when $\mathbf{TrueSteps} = 0$, but handling is not described (Sec. 2.5).

   Recommendation: State how participants/segments with $\mathbf{TrueSteps} = 0$ are handled (excluded, epsilon, or alternative metric), and ensure the same rule is used consistently across configurations.

4. Feature set description lacks final dimensionality and interpretability hooks; this limits diagnosing why transfer fails (Sec. 2.2.2, Sec. 3.5).

   Recommendation: Report total feature dimensionality (per axis/SVM and overall). Add a brief LightGBM feature-importance summary (gain/split) and discuss whether frequency features or SVM dominate and how that relates to sampling-rate and placement sensitivity.

5. Text/table formatting issues reduce clarity: some table headers/captions appear corrupted/merged with narrative, and section heading/numbering styles are inconsistent (Sec. 3.2, Sec. 3.4, Sec. 2.6.2, Sec. 3.x).

   Recommendation: Proofread tables to ensure headers/captions are self-contained and not merged into text. Normalize section numbering/heading styles per venue guidelines.

6. Ethics, consent, and data/code availability are not clearly stated despite human wearable + demographic data (Sec. 2.1, Sec. 4).

   Recommendation: Add an ethics/IRB + informed consent statement and a brief data/code availability section (or justification if sharing is restricted), including privacy handling for demographic variables.

Very minor issues

1. Terminology ambiguity: "SVM" may be read as Support Vector Machine rather than Signal Vector Magnitude; SVM formula is typographically ambiguous without explicit time indexing (Sec. 2.2).

   Recommendation: Define SVM at first use as Signal Vector Magnitude and write it unambiguously, e.g., $SVM(t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$.

2. Citation and tone issues: inconsistent citation formatting (including malformed in-text citations) and occasional promotional language (Sec. 1, References).

   Recommendation: Standardize citations to venue style, fix malformed entries, and revise subjective phrasing to neutral scientific language.

3. General presentation polish: typos/fragmented sentences and figure accessibility concerns (font size, colorblind-safe palette, unclear legends) (Sec. 1, Sec. 3.5, Fig. 1–3).

   Recommendation: Proofread for typos and incomplete sentences; export higher-resolution/vector figures; use colorblind-safe palettes; and ensure legends/labels are consistent (e.g., "Hip **100** Hz", "Wrist **25** Hz").

Key statements and references

- ⚠ **Model performance during cross-configuration transfer was quantified using standard regression metrics, specifically Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), calculated for total step counts across participants, following established evaluation practices for machine learning systems in related work.**
- *Reference(s):* Cheng et al., 2022, Ferrer et al., 2024

- *Justification:* The papers support using standard regression metrics like MAE for evaluation (Cheng et al., 2022; Ferrer et al., 2024), and Ferrer et al. discuss best-practice evaluation design. However, neither paper mentions MAPE, cross-configuration transfer, or evaluation on total step counts across participants. Thus only the general use of MAE as a standard regression metric is supported, while the specific details are not.

- ✖ **The participant cohort and sensor configurations were designed in line with prior accelerometry studies: 39 participants wore tri-axial accelerometers simultaneously at the hip and wrist, with data collected at 100 Hz and 25 Hz to form four configurations (Hip 100Hz, Hip 25Hz, Wrist 100Hz, Wrist 25Hz), and ground truth step counts were obtained from expert-annotated video recordings.**

- *Reference(s):* Abadleh et al., 2018, Sun et al., 2024, Koffman et al., 2024
- *Justification:* Neither paper matches the described design. Abadleh et al., 2018 uses a smartphone accelerometer held in the hand for distance/step detection, with no 39-participant cohort, no hip/wrist placements, no 100 Hz vs 25 Hz configurations, and no expert-annotated video ground truth. Sun et al., 2024 uses a single ankle IMU sampled at 25 Hz in 48 valid subjects; step counting was preliminarily validated in 5 subjects against experimenter-observed counts, not video, and sensors were not placed simultaneously at hip and wrist nor sampled at 100 Hz.

- ⚠ **The exploratory data analysis computed Signal Vector Magnitude (SVM) as $\sqrt{(x^2 + y^2 + z^2)}$ and found that wrist-worn accelerometer data exhibited higher and more variable mean SVM (1.18 ± 0.45 g) than hip-worn data (1.05 ± 0.21 g), consistent with prior findings that wrist sensors capture more non-gait-related movements that can confound step detection.**

- *Reference(s):* Donckt et al., 2024, 2024a, Urbanek et al., 2016, Straczkiewicz et al., 2022
- *Justification:* Partially supported. Donckt et al., 2024, 2024a explicitly define the signal magnitude vector as $\sqrt{x^2 + y^2 + z^2}$. Both Urbanek et al., 2016 and Straczkiewicz et al., 2022 note that wrist-worn sensors capture additional arm movements and can reduce specificity/complicate walking or step-related detection (e.g., strong subharmonics and low specificity for household hand movements at the wrist). However, none of the attached papers report the stated mean SVM values (1.18 ± 0.45 g for wrist vs 1.05 ± 0.21 g for hip) or an EDA showing higher and more variable mean SVM at the wrist; Urbanek et al., 2016 instead reports VMC with hip slightly higher than wrist. Thus the formula and qualitative rationale are supported, but the quantitative wrist–hip SVM comparison is not.

- ✖ **To test whether cross-configuration performance degradation and demographic influences were statistically significant, the study employed non-parametric tests aligned with prior methodological work: Wilcoxon signed-rank tests compared absolute errors between configurations, while Mann-Whitney U and Kruskal-Wallis H tests assessed differences by sex and age group, using $p < 0.05$ (with multiple-comparison corrections such as Bonferroni and FDR) as the significance threshold.**
- *Reference(s):* Cheng et al., 2022, Ghosh et al. 2025, Sekkat et al., 2024
- *Justification:* The papers do not report using Wilcoxon signed-rank, Mann-Whitney U, or Kruskal-Wallis tests with multiple-comparison corrections. Sekkat et al., 2024 assess demographic effects mainly via logistic regression (Wald tests) and likelihood-ratio tests, with some complementary chi-squared and ANOVA analyses; Wilcoxon/Kruskal–Wallis are cited only as prior work. Ghosh et al., 2025 use Wasserstein and Kolmogorov–Smirnov tests (and DeLong for AUC), not the stated non-parametric tests or corrections. Cheng et al., 2022 is a survey and does not describe such a study design. Therefore the methodological statement is not supported by the attached papers.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains light mathematics: definitions of SVM, windowing relations between seconds and samples, feature descriptions (FFT-based energy and dominant frequency), error metrics (AE/APE/MAE/MAPE), and high-level descriptions of nonparametric tests. The central internal-consistency problem is the stated method for reconstructing total steps by summing predictions from overlapping windows, which is not algebraically consistent without correction.

Checked items

1. ✔ **Signal Vector Magnitude definition** (Sec. 2.1.1, p.3 (EDA))
    - **Claim:** SVM is computed as the magnitude of tri-axial acceleration: $\sqrt{x^2 + y^2 + z^2}$.
    - **Checks:** notation consistency, dimensional/unit sanity
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** x, y, z denote the three accelerometer axes for a given sample/time.
    - **Notes:** Mathematically correct as a vector magnitude; only minor typographic ambiguity due to missing parentheses/time index.
2. ✔ **Window size to sample-count conversion** (Sec. 2.2.1, p.3)

- **Claim:** A 2-second window corresponds to 200 samples at 100 Hz and 50 samples at 25 Hz; stride is 1 second (50% overlap).
- **Checks:** algebra/arithmetic consistency, definition consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Sampling rates are exactly 100 Hz and 25 Hz., Window duration is exactly 2.0 s and stride is 1.0 s.
- **Notes:** 2 s × 100 Hz = 200 samples; 2 s × 25 Hz = 50 samples; 1 s stride implies 50% overlap.

3. ✔ **Window label definition (steps per window)** (Sec. 2.2.2, p.3)

   - **Claim:** The target label for each 2-second window is the total number of steps occurring within that window's time interval.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Ground truth annotations can be mapped to each window interval.
   - **Notes:** The label definition is clear for a single window; however, it interacts critically with overlap during aggregation (see separate item).

4. ⚠ **FFT spectral energy feature definition** (Sec. 2.2.2, p.3)

   - **Claim:** Spectral energy is extracted as the sum of squared magnitudes of FFT frequency components.
   - **Checks:** definition completeness, scale/normalization consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** FFT is computed on each windowed signal per axis/SVM.
   - **Notes:** The paper omits normalization details (N-scaling; one-sided vs two-sided spectrum). Without this, the feature is mathematically under-specified and can change scale with window length/sampling rate even if the underlying signal is similar.

5. ✔ **Dominant frequency feature definition** (Sec. 2.2.2, p.3)

   - **Claim:** Dominant frequency is the frequency component with the highest magnitude in the FFT output.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** A frequency axis (bin-to-Hz mapping) is defined implicitly by sampling rate and window length.
   - **Notes:** Conceptually consistent; the paper does not specify handling of DC component or ties, but that is not an internal algebra error.

6. ✖ **Total step reconstruction from window predictions** (Sec. 2.4.2, p.4)

- **Claim:** Total predicted steps for a participant are obtained by summing predicted step counts across all **2**-second windows.
- **Checks:** algebra/logic consistency, sanity case (overlap counting)
- **Verdict:** FAIL; confidence: high; impact: critical
- **Assumptions/inputs:** Windows use **2** s duration with **1** s stride (**50%** overlap) as stated earlier., Window-level predictions approximate steps within each window interval.
- **Notes:** With **50%** overlap, each time segment (and steps within it) belongs to multiple windows. Therefore, summing 'steps in each window' generally overcounts relative to the true total. A simple limiting check: if a recording has constant cadence and the model predicts exactly the true step count in every window, summing over overlapped windows yields approximately a factor-of-2 inflation (edge effects aside). The paper provides no correction/derivation to justify equality with ground-truth totals.

7. ✔ **Absolute Error (AE) definition** (Sec. 2.5.1, p.4)

   - **Claim:** $AE = |\textbf{PredictedSteps} - \textbf{TrueSteps}|$ at the participant total-step level.
   - **Checks:** algebra correctness, definition consistency
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** **PredictedSteps** and **TrueSteps** refer to totals over the same observation period.
   - **Notes:** Standard and internally consistent, assuming totals are computed consistently (but the reconstruction method above threatens that assumption).

8. ⚠ **APE and MAPE definitions** (Sec. 2.5.1–2.5.2, p.4)

   - **Claim:** $APE = (AE/\textbf{TrueSteps}) \times 100$; $MAPE$ is the average APE across participants.
   - **Checks:** algebra correctness, domain constraints
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** $\textbf{TrueSteps} > 0$ for all participants/configurations.
   - **Notes:** Algebra is correct, but the paper does not state how it handles $\textbf{TrueSteps} = 0$ cases (division by zero). If such cases exist, $MAPE$ is undefined without a convention.

9. ✔ **Use of paired nonparametric test for degradation** (Sec. 2.6.1, p.4–5)

   - **Claim:** Wilcoxon signed-rank test compares participant-wise absolute errors between baseline and each target configuration.
   - **Checks:** assumption consistency
   - **Verdict:** PASS; confidence: medium; impact: minor

- **Assumptions/inputs:** Errors are paired by participant across configurations.
- **Notes:** The pairing logic is coherent with the described design (same participants measured under multiple configurations). The exact statistic reported later is not defined, but the test choice is mathematically consistent.

Limitations

- The audit is restricted to the provided PDF text/images; no code, supplementary materials, or appendices were available to clarify omitted mathematical steps (e.g., de-overlapping aggregation, FFT normalization).
- No equation numbering is present in the provided excerpt; locations are given by section and page.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

16 numeric checks were run: 15 PASS and 1 UNCERTAIN. Verified items include exact cohort count consistency, exact windowing/sample computations, correct overlap percentage, correct sampling-rate reduction factor, consistent ordering of MAE/MAPE across configurations, and several narrative percent/fold-change claims matching recomputation within stated tolerances.

Checked items

1. ✔ **C1_participant_sex_counts_sum** (p.2, §2.1 Data Collection and Participant Cohort)
    - **Claim:** Participant metadata includes sex (20 male, 19 female) for 39 participants.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** male + female == total_participants
2. ✔ **C2_participant_age_group_counts_sum** (p.2, §2.1 Data Collection and Participant Cohort)
    - **Claim:** Age range groups are reported as 15 aged 18-29, 14 aged 30-49, 10 aged 50+ (total 39 participants).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** age_18_29 + age_30_49 + age_50_plus == total_participants
3. ✔ **C3_window_samples_100Hz** (p.3, §2.2.1 Windowing)
    - **Claim:** A 2-second window at 100 Hz corresponds to 200 samples.
    - **Checks:** unit_consistent_recomputation

- **Verdict:** PASS
- **Notes:** window_duration × sampling_rate == reported_samples

4. ✔ **C4_window_samples_25Hz** (p.3, §2.2.1 Windowing)
   - **Claim:** A **2**-second window at **25** Hz corresponds to **50** samples.
   - **Checks:** unit_consistent_recomputation
   - **Verdict:** PASS
   - **Notes:** window_duration × sampling_rate == reported_samples
5. ✔ **C5_stride_overlap_percent** (p.3, §2.2.1 Windowing)
   - **Claim:** Using a **2**-second window with **1**-second stride corresponds to **50%** overlap.
   - **Checks:** ratio_to_percent
   - **Verdict:** PASS
   - **Notes:** overlap_percent = (\text{window_size} - \text{stride})/\text{window_size} \times 100 == reported_overlap
6. ✔ **C6_sampling_reduction_factor** (p.5, §3.2.1 Impact of sampling frequency (Hip_25Hz))
   - **Claim:** Reduction from **100** Hz to **25** Hz is described as a four-fold reduction in sampling frequency.
   - **Checks:** ratio_check
   - **Verdict:** PASS
   - **Notes:** high_rate / low_rate == reported_factor
7. ✔ **C7_SVM_mean_difference** (p.3, §2.1.1 Exploratory Data Analysis)
   - **Claim:** Mean SVM is 1.18 ± 0.45 g for wrist vs 1.05 ± 0.21 g for hip (wrist higher).
   - **Checks:** inequality_direction_and_difference
   - **Verdict:** PASS
   - **Notes:** Checked wrist_mean_svm > hip_mean_svm; computed difference provided.
8. ✔ **C8_Hip25_MAE_percent_increase** (p.5, §3.2.1 Impact of sampling frequency (Hip_25Hz) + Table 1)
   - **Claim:** MAE escalated by **182%** from **387.54** steps to **1093.61** steps.
   - **Checks:** percent_change
   - **Verdict:** PASS
   - **Notes:** ((\text{hip25_mae} - \text{baseline_mae})/\text{baseline_mae}) \times 100 \approx reported_increase_percent

9. ✔ **C9_Wrist100_MAE_percent_increase** (p.6, §3.2.2 Impact of sensor location (Wrist_100Hz) + Table 1)
    - **Claim:** For Wrist_100Hz, MAE **1731.98** steps represents a **347%** increase from the Hip_100Hz baseline (**387.54**).
    - **Checks:** percent_change
    - **Verdict:** PASS
    - **Notes:** ((\text{wrist100_mae} - \text{baseline_mae})/\text{baseline_mae}) \times 100 \approx reported_increase_percent
10. ✔ **C10_Wrist25_MAE_fold_increase** (p.6, §3.2.3 Combined impact (Wrist_25Hz) + Table 1)
    - **Claim:** Wrist_25Hz MAE **1978.11** steps is described as a more than five-fold increase compared to Hip_100Hz baseline MAE **387.54**.
    - **Checks:** fold_change_lower_bound
    - **Verdict:** PASS
    - **Notes:** (\text{wrist25_mae} / \text{baseline_mae}) > threshold_fold
11. ✔ **C11_Table1_MAPE_ordering** (p.5, Table 1)
    - **Claim:** MAPE increases monotonically across configurations: **12.88%** (Hip_100Hz) < **34.14%** (Hip_25Hz) < **56.98%** (Wrist_100Hz) < **66.91%** (Wrist_25Hz).
    - **Checks:** monotonic_ordering
    - **Verdict:** PASS
    - **Notes:** mape_hip100 < mape_hip25 < mape_wrist100 < mape_wrist25
12. ✔ **C12_Table1_MAE_ordering** (p.5, Table 1)
    - **Claim:** MAE increases monotonically across configurations: **387.54** (Hip_100Hz) < **1093.61** (Hip_25Hz) < **1731.98** (Wrist_100Hz) < **1978.11** (Wrist_25Hz).
    - **Checks:** monotonic_ordering
    - **Verdict:** PASS
    - **Notes:** mae_hip100 < mae_hip25 < mae_wrist100 < mae_wrist25
13. ⚠ **C13_interindividual_consistency_doubling_claim** (p.6, §3.3 Inter-individual consistency + Table 1)
    - **Claim:** $\text{Std}_\text{A}\text{E}$ more than doubled from **401.81** (Hip_100Hz) to **998.00** (Hip_25Hz).
    - **Checks:** fold_change_lower_bound
    - **Verdict:** UNCERTAIN
    - **Notes:** Insufficient inputs to compute fold change lower bound.

14. ✔ **C14_interindividual_consistency_quadrupled_claim_wrist100** (p.6, §3.3 Inter-individual consistency + Table 1)
    - **Claim:** $\mathbf{Std_AE}$ quadrupled to **1529.00** steps for Wrist_100Hz compared with **401.81** baseline.
    - **Checks:** approx_fold_change
    - **Verdict:** PASS
    - **Notes:** std_ae_wrist100 / std_ae_baseline ≈ claimed_fold
15. ✔ **C15_interindividual_consistency_quadrupled_claim_wrist25** (p.6, §3.3 Inter-individual consistency + Table 1)
    - **Claim:** $\mathbf{Std_AE}$ quadrupled to **1718.98** steps for Wrist_25Hz compared with **401.81** baseline.
    - **Checks:** approx_fold_change
    - **Verdict:** PASS
    - **Notes:** std_ae_wrist25 / std_ae_baseline ≈ claimed_fold
16. ✔ **C16_underestimation_percent_wrist25_total_steps** (p.7, §3.5 Analysis of error characteristics)
    - **Claim:** For Wrist_25Hz, predicted total steps across all participants is $\mathbf{37,636}$ vs ground truth $\mathbf{125,797}$; described as underestimation of over **70%**.
    - **Checks:** percent_difference
    - **Verdict:** PASS
    - **Notes:** ((\text{ground_truth_total} - \text{predicted_total})/\text{ground_truth_total}) \times 100 > claimed_underestimation_threshold

Limitations

- Audit is limited to the provided PDF text; no underlying datasets, per-participant errors, or per-window predictions are available for recomputation of reported metrics or statistical tests.
- Plot-based numerical verification is excluded (no pixel/value extraction from figures).
- Some narrative quantitative terms (e.g., 'quadrupled') are approximate; checks use reasonable numeric tolerances but cannot confirm authorial intent beyond arithmetic consistency.