Skepthical review: *Quantifying and Characterizing Step Counting Uncertainty in Wearable Accelerometer Data*

Summary

The paper presents a probabilistic step-counting framework for triaxial accelerometer data that aims to report not only point estimates but also uncertainty. A 1D CNN predicts a Poisson rate (Sec. 2.2–2.2.2) for step counts in short overlapping windows (2.56 s, 50% overlap), and window-level outputs are aggregated to participant-level step totals with a Poisson-based 95% interval (Sec. 2.4.1–2.4.3). The method is evaluated via leave-one-subject-out (LOSO) cross-validation on 39 participants across hip vs wrist placement and 100 Hz vs 25 Hz sampling, reporting MAE/MAPE/bias as well as FP/FN-style measures, interval width, and Bland–Altman analysis, plus non-parametric comparisons and a sex subgroup analysis (Sec. 3.1–3.5).

The core contribution—making uncertainty explicit and using it to compare sensing configurations—is timely and practically relevant. However, the current manuscript has several technical and reporting issues that undermine trust in the uncertainty claims and some conclusions: (i) the aggregation/interval construction is ambiguous and internally inconsistent under overlapping windows; (ii) interval calibration (coverage) is not evaluated; (iii) FP/FN definitions are hard to interpret under overlap and continuous-rate outputs; (iv) Table 1 appears corrupted/inconsistent; and (v) important methodological details, baseline comparisons, and statistical reporting (effect sizes/multiple comparisons) are missing or incomplete. Addressing these would substantially strengthen both scientific validity and practical usefulness.

Strengths

- Clear motivation for uncertainty-aware step counting and for comparing sensor placement and sampling-frequency trade-offs (Introduction, Sec. 1; Discussion/Conclusion, Sec. 4).
- Simple, interpretable probabilistic modeling choice: predicting a positive Poisson rate via softplus and training with a Poisson negative log-likelihood (Sec. 2.2.1–2.2.2).
- Reasonable and relevant experimental axes (hip vs wrist; 100 Hz vs 25 Hz) evaluated under LOSO cross-validation on 39 participants (Sec. 2.1–2.3).
- Evaluation goes beyond a single accuracy metric, including bias and Bland–Altman analysis, and attempts to quantify uncertainty via interval widths (Sec. 2.4; Sec. 3.2–3.4).
- The narrative generally emphasizes inter-individual variability and the risk of over-relying on aggregate MAE/MAPE, which is an important practical point for wearable monitoring (Sec. 3.4; Sec. 4).

Major issues

1. **Uncertainty/interval construction is ambiguous and internally inconsistent under overlap, and calibration is not assessed (Sec. 2.4.1–2.4.3; Results Sec. 3.2).** The text states an overlap-corrected total prediction $\hat{Y} = (\sum \lambda_i)/2$ (Sec. 2.4.1) but then constructs the Poisson total rate as $\Lambda = \sum \lambda_i$ (Sec. 2.4.3), which would center the interval on a different mean than the reported point estimate. Additionally, 50% overlapping windows violate the independence assumptions typically used to justify summing Poisson variables, and the reported mean 95% interval widths (Table 1 / Sec. 3.2) appear surprisingly narrow given the scale of total steps, raising concern that the implemented formula may not match the written description (or that a heuristic is being used without validation). Finally, the manuscript calls these "confidence intervals" in places, but what is computed appears to be a predictive interval conditional on $\boldsymbol{\Lambda}$.

   Recommendation: Make the uncertainty pipeline mathematically and implementationally explicit and consistent: (i) define precisely what λ_i represents (expected steps per full window vs per non-overlapping segment) and how overlap is handled; (ii) ensure the same overlap correction is applied to both the point estimate and the Poisson mean used for intervals (e.g., if $\hat{Y} = \sum \lambda_i/2$ then use $\Lambda_{\text{total}} = \sum \lambda_i/2$ for the Poisson-based predictive interval), and state how boundary windows are treated; (iii) explicitly acknowledge dependence induced by overlap and label the Poisson-sum interval as an approximation unless you provide a derivation under stated assumptions; (iv) rename to "(95%) prediction interval" unless parameter uncertainty is being estimated; and (v) add an uncertainty calibration analysis in Sec. 3.2/3.4: empirical coverage of nominal 50/80/95% intervals by configuration, plus a check that interval width correlates with absolute error. If coverage is off, consider either non-overlapping windows for interval construction, a dispersion/variance inflation factor, or a post-hoc calibration scaling of $\boldsymbol{\Lambda}$, reporting before/after coverage and width.

2. **FP/FN definitions are difficult to interpret and likely inconsistent with overlap-corrected aggregation and continuous Poisson-rate outputs (Sec. 2.4.4; referenced in Sec. 3.1–3.3).** FP is described as "sum of predicted steps over windows with zero true steps," and FN as $\sum \mathbf{max}(0, y_i - \hat{\lambda}_i)$. With overlapping windows, the same underlying time/steps contribute to multiple windows, so window-level sums can double-count. Moreover, using $\hat{\lambda}_i$ (a real-valued mean) yields fractional "FP/FN," which are not event-level false detections/misses in the conventional sense, making cross-configuration comparisons and interpretation (especially wrist vs hip) potentially misleading.

   Recommendation: Rework Sec. 2.4.4 to provide physically interpretable and overlap-consistent error decompositions. Options: (a) compute FP/FN on a non-overlapping timeline (e.g., per-sample, per-second, or per non-overlapping window) by reconstructing predicted counts onto unique time bins, then aggregating; (b) if you keep window-

level measures, explicitly correct for overlap (and justify the correction) and rename them to avoid event-detection connotations (e.g., "overcount mass in zero-step windows" and "undercount mass" / "count shortfall"). In either case, give explicit formulas with window indices ($\hat{\lambda}_i$, y_i) and state whether you interpret these as expected counts. Recompute and update Sec. 3.1–3.3 and Table 1 accordingly, and add a brief note explaining how these quantities relate to bias and absolute error to prevent apparent contradictions (e.g., similar MAE but different over/under counting patterns).

3. **Table 1 (core quantitative results) is corrupted/incomplete and internally inconsistent with surrounding text (Sec. 3.1).** The Hip_100Hz row contains narrative text in a numeric cell; FN/CI-width entries appear missing or misaligned for hip conditions; and at least one FP value conflicts with the text below the table (Hip_25Hz vs Hip_100Hz FP attribution). This prevents reliable verification of key claims (best configuration, CI width comparisons, FP/FN differences).

   Recommendation: Rebuild Table 1 directly from the stored per-fold/per-subject logs, ensuring each configuration (Hip/Wrist × 100/25 Hz) has complete, correctly aligned mean ± SD entries for MAE, MAPE, bias, FP, FN, and interval width. Remove prose from table cells and place clarifications in the caption or main text. Then audit Sec. 3.1–3.2 and Sec. 4 to ensure every numeric claim (e.g., FP values, CI widths, "more than double") matches the corrected table.

4. **Observation model choice (Poisson) is under-justified given overlap, zero-inflation, and likely overdispersion, which directly affects uncertainty claims (Sec. 2.2.2; hinted in Sec. 3.1–3.2).** Step counts can be overdispersed relative to Poisson (variance > mean) due to heterogeneous activities, cadence variability, and label noise; and overlapping windows induce strong temporal dependence. Using a Poisson model may understate predictive variance and give misleadingly narrow intervals, especially if the interval is interpreted as "model confidence."

   Recommendation: In Sec. 2.2.2 and Sec. 4 (Limitations), add a concrete diagnostic: report mean/variance of y_i per configuration (and proportion of $y_i = 0$) to assess overdispersion/zero inflation. If overdispersion is present, add at least a sensitivity analysis: Negative Binomial likelihood, quasi-Poisson (variance inflation factor estimated from residuals), or a calibrated dispersion term used when constructing predictive intervals. Explicitly discuss how overlap-induced dependence affects the generative interpretation, and frame the Poisson approach as a pragmatic approximation unless a more principled non-overlapping or point-process formulation is adopted.

5. **Missing baseline comparisons make it hard to quantify the incremental benefit of the proposed probabilistic CNN (Sec. 2–3).** The experiments compare only variants of the same Poisson-CNN across placements/frequencies. Without (i) a deterministic deep regressor baseline and (ii) a simple classical/heuristic step

counter, the reader cannot tell whether the main gain is accuracy, uncertainty quantification, or simply model capacity—and whether uncertainty adds value beyond, e.g., residual-based intervals.

Recommendation: Add at least two baselines under the same LOSO protocol and preprocessing: (i) a deterministic CNN (or LSTM/TCN) trained with MAE/MSE, with uncertainty via residual quantiles or conformal prediction; and (ii) a standard threshold/peak-based step counter (hip/wrist tuned identically across folds if possible). Report the same metrics (MAE/MAPE/bias and your revised FP/FN-style measures) plus Bland–Altman. For uncertainty, compare empirical coverage/width to your Poisson-based intervals. Summarize in Sec. 3.1–3.4 and update Sec. 4 to state clearly what is improved by the probabilistic formulation.

6. **Methodological details are insufficient for reproducibility and for evaluating overfitting/leakage risk under LOSO (Sec. 2.1–2.3).** Key missing items include: full CNN layer-by-layer specification; preprocessing (normalization, gravity removal, filtering); downsampling/anti-aliasing and time alignment of annotations at 25 Hz; training hyperparameters (batch size, optimizer settings, learning rate schedule, epochs); early stopping criterion/patience; random seeds; and the exact within-fold train/validation split (and confirmation that the held-out subject is not used for validation).

   Recommendation: Expand Sec. 2.1–2.3 with an implementation-ready description: (i) device/hardware details and accelerometer range; (ii) filtering and downsampling procedure (explicitly state any anti-alias filtering), and how step annotations are mapped to samples/windows after downsampling; (iii) preprocessing per axis (standardization, gravity removal, coordinate handling); (iv) exact CNN architecture (filters, kernel sizes, strides/padding, pooling, dense units, dropout/BN); (v) training details (optimizer, LR, batch size, epochs, early stopping metric/patience); (vi) within-fold validation scheme (window-level split across training subjects vs subject-level split), guaranteeing no leakage from the test subject; and (vii) code/data availability statement (Sec. 2.5 or Sec. 4) plus seed/control for reproducibility.

7. **Statistical testing/reporting is incomplete given the number of comparisons and the manuscript's reliance on p-values for conclusions (Sec. 2.5; Sec. 3.3; Abstract/Conclusion).** Tables 2–3 reportedly include test statistics but omit p-values/directions; effect sizes and confidence intervals are not provided; and no multiple-comparison correction is described despite many hypothesis tests across multiple metrics/configurations/subgroups. Additionally, the Abstract/Introduction/Conclusion cite specific p-values and mention age-related Kruskal–Wallis tests even though the age analysis is not completed (Sec. 3.5.2).

   Recommendation: Revise Sec. 2.5 and Sec. 3.3–3.5 to (i) report effect sizes (e.g., rank-biserial correlation or Cliff's delta) and median paired differences with confidence intervals for key comparisons; (ii) add p-values and the direction of effects to Tables 2–

3; and (iii) apply and disclose a multiple-testing correction (e.g., Holm) within coherent families (by metric or by research question), or clearly state that results are exploratory with unadjusted p-values and temper claims accordingly. Ensure every p-value mentioned in the Abstract/Introduction/Conclusion is traceable to a specific test in Sec. 3, and remove/relocate any age-related inferential claims unless the analysis is actually performed.

8. **Age-related subgroup analysis is inconsistent/unfinished but still appears in the methods narrative and figures, risking reader confusion and overclaiming (Sec. 2.1.3; 2.5.2; 3.5.2; Fig. 6).** The manuscript describes age stratification and cites a figure outlining age analyses, yet later states age metadata were incorrect and analysis could not be completed.

   Recommendation: Choose one of two clean resolutions: (a) repair the age metadata pipeline and rerun the planned age-stratified analyses (including explicit group sizes and tests), updating Fig. 6 with real results; or (b) remove age-analysis claims from the Abstract/Introduction/Conclusion and reposition Fig. 6 as a schematic in Methods/Supplement clearly labeled as "planned future analysis." In Sec. 3.5.2, describe the metadata issue precisely and confirm which covariates (e.g., sex) are unaffected.

9. **Scope/generalizability claims are stronger than the data support (Introduction; Sec. 4).** The dataset is moderate (39 participants) with recordings on the order of ∼58 minutes and may not reflect true multi-day free-living variability, device heterogeneity, or broader population diversity. Given the approximations in overlap + Poisson uncertainty, claims that the framework provides robust free-living guidance should be tempered.

   Recommendation: Add a dedicated Limitations paragraph (Sec. 4) that explicitly covers dataset size, recording duration, controlled vs truly free-living conditions, single-device/single-cohort constraints, and the modeling approximations (Poisson, overlap dependence). Rephrase any strong prescriptive statements ("critical insights," "25 Hz largely sufficient") to be conditional on this dataset and evaluation setting, and highlight needed external validation.

Minor issues

1. Loss function is presented in inconsistent forms and it is unclear which was implemented (Sec. 2.2.2). One equation omits the $\log(y!)$ term and another includes it; the averaging level (per window vs per batch) is not clearly stated.

   Recommendation: Present one canonical Poisson NLL matching the code (e.g., framework PoissonNLLLoss settings). State explicitly whether $\log(y!)$ is included/ignored, and clarify the reduction (mean over batch) used during training.

2. Terminology: "confidence interval" vs "prediction interval" is potentially misleading throughout (Sec. 2.4.3; Sec. 3.2; Sec. 4).

Recommendation: Use "prediction interval" (conditional on Λ and model outputs) unless you incorporate parameter uncertainty (e.g., Bayesian weights/ensembles). If you keep "confidence interval," define precisely what parameter it is intended to cover and how.

3. Windowing design choice (2.56 s, 50% overlap) is not justified beyond convention (Sec. 2.1.4). Given that overlap is central to aggregation/uncertainty issues, the choice deserves a brief rationale or sensitivity check.

   Recommendation: Provide a short justification linked to step cadence ranges and CNN receptive field needs, and/or add a small sensitivity analysis (e.g., 1.28 s vs 2.56 s; overlap 0% vs 50%) focusing on both MAE and interval calibration/width.

4. Relationship between interval width and actual error is discussed mostly qualitatively (Sec. 3.2; Sec. 3.4). Without quantitative linkage, it is unclear whether CI width is a useful confidence indicator at the participant level.

   Recommendation: Report median (IQR) interval widths by configuration and test hip vs wrist and 100 vs 25 Hz differences (non-parametric). Compute correlation (Spearman) between interval width and absolute error across participants; optionally show a scatter plot with regression/LOESS in Supplement.

5. Exploratory data analysis/activity context is thin (Sec. 2.1.3; Sec. 3.1). The manuscript notes many zero-step windows and similar distributions but provides little quantitative context (e.g., proportion of zero windows; activity protocol).

   Recommendation: Add concise EDA: percent of windows with $\boldsymbol{y} = \mathbf{0}$, mean/median $\boldsymbol{y}$ per window, and total steps per recording/configuration. Briefly describe the recording protocol (activities included) and note whether activity labels exist; this will contextualize wrist FP behavior.

6. Bland–Altman analysis presentation may pool repeated measures or omit key diagnostics (Sec. 3.4; Fig. 4/5). It is unclear whether repeated-measures BA considerations are needed, and proportional bias is not assessed.

   Recommendation: Clarify the unit of analysis (per participant totals) and confirm independence assumptions. Add proportional-bias checks (regression of differences on means) and, if applicable, report confidence intervals for bias and limits of agreement.

7. Figures 1/3/4/5 use inconsistent binning/axis scaling/normalization, reducing cross-panel comparability (Sec. 3; Figs. 1, 3–5).

   Recommendation: Standardize axis limits and bin widths across comparable panels; use density/proportion normalization for histograms; annotate sample sizes and key summary stats (median/IQR or mean/SD). For discrete counts (Fig. 1), consider bar-style histograms.

8. Ethics and human-subjects data provenance are not stated (Sec. 2.1.1).

Recommendation: Add an ethics/provenance statement: dataset source, IRB/ethics approval (or reference to the original approval if using a public dataset), consent, and anonymization.

9. Code/data availability is not clearly stated despite mention of logging intermediate results (Sec. 2.5.2; Sec. 4).

   Recommendation: Include a clear availability statement: links (if public), what will be released (code, configs, trained weights), and any access restrictions.

10. Related work on uncertainty estimation in time-series/HAR is underdeveloped relative to the paper's positioning (Introduction, Sec. 1).

    Recommendation: Add a short paragraph situating this approach vs alternatives (MC dropout, deep ensembles, Bayesian NNs, conformal prediction for counts), emphasizing trade-offs in calibration, complexity, and interpretability.

Very minor issues

1. Notation for FP uses $\mathring{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ (participant totals) where window-level quantities ($\boldsymbol{y_i}$, $\boldsymbol{\lambda}_i$) are intended (Sec. 2.4.4).

   Recommendation: Rewrite FP/FN definitions fully in window-indexed notation and define all symbols at first use.

2. MAPE is undefined when the true total $\boldsymbol{Y} = \mathbf{0}$ (Sec. 2.4.2).

   Recommendation: State that all evaluated recordings have $\boldsymbol{Y} > \mathbf{0}$, or define a convention (e.g., exclude, add epsilon, or use SMAPE).

3. Figure 6 appears placeholder-like (missing labels/legend/sample sizes) and is referenced as if results exist (Sec. 3.5; Fig. 6).

   Recommendation: Either replace with completed subgroup results or explicitly label as schematic and move out of Results; ensure labels/legend/sample sizes are present.

4. Reference formatting and relevance appear inconsistent; some citations seem tangential and styles are mixed (References; in-text citations in Sec. 1–2).

   Recommendation: Standardize to venue format, ensure every citation is necessary and directly supports a claim, and ensure consistent numbering/mapping between in-text and bibliography.

5. Author/affiliation line is non-standard for scientific submission (front matter).

   Recommendation: Replace with standard affiliations or anonymize appropriately for double-blind review.

6. Scattered typographical/formatting issues (broken words, inconsistent "100Hz/100 Hz", "wristworn/wrist-worn", stray heading characters) (Sec. 1–4).

Recommendation: Proofread and normalize units, hyphenation, heading styles, and math/p-value formatting throughout.

Key statements and references

- **✖ The proposed framework extends prior work on step counting under varying sensor placements and sampling frequencies by explicitly quantifying uncertainty for configurations such as hip versus wrist and 100 Hz versus 25 Hz, building on earlier analyses of high-resolution wrist accelerometry step counting and algorithm comparisons in large cohorts.**
- *Reference(s):* Khan and Abedi, 2022, Koffman et al., 2024
- *Justification:* Khan and Abedi, 2022 propose an attention-based LSTM for step counting and evaluate it across datasets with different sensor placements (ankle, hip, wrist) and sampling rates (15, 25, 100 Hz; with 100 Hz downsampled), reporting performance metrics and some standard deviations (e.g., Tables II–IV). However, they do not present an explicit uncertainty quantification framework for configurations like hip vs. wrist or 100 Hz vs. 25 Hz, nor do they build on or reference earlier analyses of high-resolution wrist accelerometry and algorithm comparisons in large cohorts. The paper focuses on accuracy metrics rather than uncertainty estimation and uses relatively small datasets.

- **✖ The probabilistic 1D CNN is designed to output the Poisson rate parameter λ for each window, following prior work on step counting with deep models, and uses the sum of predicted rates divided by the overlap factor (2) to reconstruct total step counts from overlapping windows, as previously proposed for time-series forecasting with overlapping segments.**
- *Reference(s):* Khan and Abedi, 2022, Fu et al., 2022, Leppich et al., 2025
- *Justification:* None of the attached papers describe a probabilistic 1D CNN that outputs a Poisson rate λ per window or reconstructs counts by summing predicted rates and dividing by an overlap factor. Khan and Abedi, 2022 formulate step counting as signal-level LSTM regression without windowing and do not use Poisson modeling; their related CNN/LSTM works are step/non-step classifiers, not Poisson outputs. Fu et al., 2022 (MMMF) and Leppich et al., 2025 (REP-Net) are general time-series forecasting frameworks and do not mention Poisson outputs or overlap-based aggregation.

- **⚠ The training objective minimizes the negative log-likelihood of the observed step counts under a Poisson model, using the standard PoissonNLLLoss formulation $L(y, \lambda) = -(y \cdot \log(\lambda) - \lambda - \log(y!))$, consistent with prior work on loss functions for deep learning and Poisson regression–style modeling.**
- *Reference(s):* Terven et al., 2025, Abdelkhalik et al., 2023

- *Justification:* Abdelkhalik et al., 2023 state they train a Poisson Neural Network using PyTorch's PoissonNLLLoss and give the loss explicitly as $\mathbf{loss(input, target)} = \mathbf{input} - \mathbf{target} \cdot \mathbf{log(input + eps)}$, i.e., the Poisson negative log-likelihood without the $\mathbf{log}(\boldsymbol{y}!)$ constant. Terven et al., 2025 likewise derives the Poisson loss as $\boldsymbol{\lambda} - \boldsymbol{y} \cdot \mathbf{log}(\boldsymbol{\lambda})$, noting that constant terms like $\mathbf{log}(\boldsymbol{y}!)$ are omitted. Thus, the use of a Poisson NLL objective is supported, but the specific full formula including $-\mathbf{log}(\boldsymbol{y}!)$ is not what is used; the constant term is omitted in both descriptions.

- **✖ Uncertainty quantification in this framework follows recent confidence-estimation methods by approximating each participant's total predicted step count as Poisson with rate $\Lambda = \sum \lambda_i$ and defining the primary uncertainty metric as the width of the 95% prediction confidence interval derived from this distribution.**

- *Reference(s):* Kivimäki et al., 2025a, Kivimäki et al., 2025b, Chang et al., 2025
- *Justification:* None of the attached papers model totals as a Poisson($\Lambda = \sum \lambda_i$) process or define uncertainty as the width of a 95% prediction CI for such a Poisson. Kivimäki et al., 2025a and Kivimäki et al., 2025b quantify uncertainty using distributions based on the Poisson binomial for counts of correct predictions and derive confidence intervals from those distributions, not a Poisson with summed rates. Chang et al., 2025 studies CIs for Turing's estimator (with normal/Poisson asymptotics) but does not discuss participant step counts or summing Poisson rates, nor make CI width the primary uncertainty metric.

- **✖ Within-subject comparisons of sensor configurations employ non-parametric tests grounded in recent methodological work: paired Wilcoxon signed-rank tests for hip versus wrist and 100 Hz versus 25 Hz comparisons, Mann-Whitney U tests for sex-stratified analyses, and Kruskal-Wallis tests for age-group comparisons, all conducted under an $\alpha = 0.05$ significance threshold to control Type I error as discussed in recent statistical literature on multiple significance thresholds.**

- *Reference(s):* Couch et al., 2018, Howard and Pimentel, 2024, Hemerik and Koning, 2025
- *Justification:* None of the cited papers describe within-subject sensor comparisons (hip vs wrist, 100 Hz vs 25 Hz) or the stated analysis plan using Mann–Whitney U or Kruskal–Wallis tests. Couch et al., 2018 and Howard and Pimentel, 2024 focus on Wilcoxon-type signed-rank methodology (with $\boldsymbol{\alpha = 0.05}$ in simulations) but do not cover the other tests or sensor context. Hemerik and Koning, 2025 discuss the dangers of multiple significance thresholds and Type I error, not the specific analyses claimed. Thus the 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: substantial

The paper's central analytic machinery is a probabilistic step-count model using a Poisson likelihood (via a CNN that outputs λ), reconstruction of participant-level step totals from overlapping windows, and uncertainty quantification via a Poisson-derived 95% interval width. Most metric definitions (MAE/MAPE/bias/Bland–Altman) are standard. The main internal inconsistency is that overlap correction is applied to the point estimate $\hat{Y}$ (divide by 2) but not to the Poisson rate Λ used for confidence intervals, even though the CI width is a primary contribution. Additional ambiguity arises from using Poisson additivity under overlapping (dependent) windows and from inconsistent window-vs-total notation in FP/FN definitions.

Checked items

1. ✔ **Window length conversion (100 Hz)** (Sec. 2.1.4, p.3)
    - **Claim:** A 2.56-second window corresponds to 256 samples at 100 Hz.
    - **Checks:** algebra, dimensional/units
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sampling frequency is exactly 100 samples/second.
    - **Notes:** 2.56 s × 100 Hz = 256 samples.
2. ✔ **Window length conversion (25 Hz)** (Sec. 2.1.4, p.3)
    - **Claim:** A 2.56-second window corresponds to 64 samples at 25 Hz.
    - **Checks:** algebra, dimensional/units
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sampling frequency is exactly 25 samples/second.
    - **Notes:** 2.56 s × 25 Hz = 64 samples.
3. ✔ **50% overlap stride conversion** (Sec. 2.1.4, p.3)
    - **Claim:** A 50% overlap implies a 1.28-second stride, i.e., 128 samples at 100 Hz and 32 samples at 25 Hz.
    - **Checks:** algebra, dimensional/units
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Stride is half the window length.
    - **Notes:** Half of 2.56 s is 1.28 s; 1.28×100=128 and 1.28×25=32.
4. ✔ **Poisson output positivity via softplus** (Sec. 2.2.1, p.3)
    - **Claim:** The model outputs a Poisson rate parameter λ guaranteed positive via softplus.
    - **Checks:** definition consistency

- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** $\text{Softplus}(x) = \log(1 + e^x) > 0$.
- **Notes:** Ensures $\lambda > 0$, matching Poisson requirements.

5. ⚠ **Poisson negative log-likelihood expression** (Sec. 2.2.2, p.3)

   - **Claim:** The loss corresponds to the negative log-likelihood under a Poisson model.
   - **Checks:** algebra, definition consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** Poisson pmf with mean λ., Optimization may drop constants independent of λ.
   - **Notes:** Two forms are shown: (i) $(\lambda_i - y_i \log \lambda_i)$ averaged over i, and (ii) $-(y \log \lambda - \lambda - \log(y!)) = \lambda - y \log \lambda + \log(y!)$. These are consistent up to the additive constant $\log(y!)$ and the averaging convention, but the text says "for a single window" while displaying a dataset average $1/n \sum_i$. Clarify per-sample vs average and note explicitly that $\log(y!)$ can be omitted.

6. ⚠ **Overlap-corrected total step reconstruction** (Sec. 2.4.1, p.4)

   - **Claim:** Total predicted steps for a participant are reconstructed as $\hat{Y} = (\sum_i \lambda_i)/2$ to avoid double counting due to 50% overlap.
   - **Checks:** derivation logic, sanity/limiting cases
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Each step event is contained in exactly two windows (except boundaries)., λ_i corresponds to the expected count for the full window.
   - **Notes:** Dividing by 2 is a plausible heuristic if every underlying event is counted twice due to 50% overlap, but this requires assumptions about boundary handling and about how window-level expectations map to unique time support. The paper states "unique contribution" but does not define it mathematically.

7. ✖ **Poisson total-rate used for CI (scaling mismatch)** (Sec. 2.4.3, p.4)

   - **Claim:** The total predicted step count distribution is approximated as Poisson with total rate $\Lambda = \sum_i \lambda_i$, using the same summation logic as for $\hat{Y}$.
   - **Checks:** definition consistency, algebra
   - **Verdict:** FAIL; confidence: high; impact: critical
   - **Assumptions/inputs:** $\hat{Y}$ aggregation is intended to match the mean of the uncertainty distribution.
   - **Notes:** $\hat{Y}$ is explicitly defined as $(\sum_i \lambda_i)/2$ in Sec. 2.4.1, but Λ is defined as $\sum_i \lambda_i$ in Sec. 2.4.3 while claiming the same logic. This makes the CI correspond to a mean that is (approximately) twice the reported point estimate $\hat{Y}$

(or conversely makes $\hat{Y}$ inconsistent with the Poisson mean). Since CI width is a primary uncertainty metric, this inconsistency is central.

8. ⚠ **Poisson additivity under overlapping windows** (Sec. 2.4.3, p.4)
    - **Claim:** A Poisson CI for the total is derived from a Poisson distribution with a total rate obtained by summing window rates.
    - **Checks:** derivation logic, assumption checking
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Window-level counts are independent Poisson random variables or represent disjoint contributions whose sum is Poisson.
    - **Notes:** If windows overlap, the associated counts are not disjoint and generally not independent, so the sum-of-Poissons rationale requires additional modeling (e.g., mapping each λ_i to a non-overlapping sub-interval). The paper does not provide this derivation; it states an approximation.
9. ✔ **MAE definition** (Sec. 2.4.2, p.4)
    - **Claim:** MAE = mean($|Y - \hat{Y}|$) at participant-total level.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Y and $\hat{Y}$ are totals per participant.
    - **Notes:** Standard definition; consistent with later bias/BA usage.
10. ✔ **MAPE definition** (Sec. 2.4.2, p.4)
    - **Claim:** MAPE = mean($|(Y - \hat{Y})/Y|$) $\times$ 100.
    - **Checks:** definition consistency, domain constraints
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $Y \neq 0$ for all participants included in MAPE computation.
    - **Notes:** Mathematically standard but undefined at $Y = 0$; the paper does not state a convention. If all participants have positive totals, it is fine.
11. ✔ **Bias and Bland–Altman limits of agreement** (Sec. 2.4.2, p.4; Sec. 3.4, p.8)
    - **Claim:** Bias is mean($Y - \hat{Y}$) and 95% LoA are $\mathbf{mean}_{\text{bias}} \pm \mathbf{1.96} \cdot \mathbf{std}(Y - \hat{Y})$. Positive bias indicates under-counting ($Y > \hat{Y}$).
    - **Checks:** definition consistency, sign consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Differences are computed as true minus predicted consistently.
    - **Notes:** Sign convention is consistent with the interpretation in Results (positive mean bias → undercount).
12. ✖ **False positives (FP) definition uses inconsistent symbols** (Sec. 2.4.4, p.4)

- **Claim:** FP is computed by summing predicted step counts from windows where the true step count is zero.
- **Checks:** notation/definition consistency
- **Verdict:** FAIL; confidence: high; impact: moderate
- **Assumptions/inputs:** Window-level true labels are y_i; window-level predictions are λ_i or a derived count.
- **Notes:** The text says: sum predicted step counts ($\hat{Y}$) from all windows where the true step count (Y) was zero. Earlier, Y and $\hat{Y}$ denote participant totals, while window-level variables are y_i and λ_i. As written, FP is not well-defined and should be expressed with window indices (e.g., $\sum_i \lambda_i \cdot \mathbf{1}y_i = 0$ or similar), also considering overlap correction.

13. ⚠ **False negatives (FN) / missed steps formula** (Sec. 2.4.4, p.5)
    - **Claim:** FN is computed as $\sum_i \mathbf{max}(0, y_i - \lambda_i)$ aggregated over windows.
    - **Checks:** dimensional/units, definition clarity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** λ_i is interpreted as predicted expected count for window i.
    - **Notes:** Expression is algebraically well-formed, but it yields fractional values because λ_i is a mean, not an integer count. This can be valid as an 'expected missed steps' surrogate, but the interpretation is not explicitly stated. Also, overlap may cause double-counting of missed steps unless corrected.
14. ⚠ **Uncertainty metric defined as 95% CI width** (Sec. 2.4.3, p.4; Sec. 3.2, pp.6–7)
    - **Claim:** Uncertainty is quantified by the width of a 95% CI derived from the Poisson distribution for the total.
    - **Checks:** definition completeness
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor
    - **Assumptions/inputs:** A specific CI construction/quantile rule is used consistently.
    - **Notes:** The paper does not specify how the 95% Poisson interval is computed (e.g., exact quantiles, central vs equal-tailed, continuity corrections). This does not create an algebraic contradiction, but it prevents verification of the stated CI derivation details.

Limitations

- The audit is restricted to the mathematics explicitly stated in the provided PDF text; several steps (notably the derivation/justification of overlap correction and the Poisson aggregation used for uncertainty) are asserted rather than derived, limiting verifiability.

- Figures are descriptive; no additional hidden equations were available beyond the text excerpts, so any mathematical definitions embedded only in figure images may not have been audited.

Numerical results audit

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

Executed checks (n=16) all passed. Verified exact integer arithmetic for participant demographics, expected file/model counts, LOSO split counts, and segmentation sample/overlap computations. Verified Table 1 narrative consistency for bias sign interpretation, lowest MAE, narrowest/widest CI widths, and a "more than double" false-positive comparison. Verified Bland–Altman LoA bounds for Hip_100Hz by recomputing mean_bias ± 1.96·SD within the stated rounding tolerance.

Checked items

1. ✔ **C01_participant_sex_counts_sum** (p.3, Sec. 2.1.3 (EDA demographics))
    - **Claim:** Participant demographics: 19 males and 20 females (total 39 participants).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Computed sum(parts)=39.0 vs reported total=39.0.
2. ✔ **C02_age_group_counts_sum** (p.3, Sec. 2.1.3 (EDA demographics))
    - **Claim:** Age distributions: 15 participants in 18-29, 14 in 30-49, and 10 in 50+ (total 39).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Computed sum(parts)=39.0 vs reported total=39.0.
3. ✔ **C03_expected_file_count** (p.3, Sec. 2.1.2 (Data integrity checks))
    - **Claim:** 156 expected files (39 participants × 4 configurations).
    - **Checks:** product_equals_total
    - **Verdict:** PASS
    - **Notes:** Computed product=156.0 vs reported total=156.0.
4. ✔ **C04_window_samples_100hz** (p.3, Sec. 2.1.4 (Data segmentation))
    - **Claim:** For 100 Hz data, a window size of 2.56 seconds corresponds to 256 samples.
    - **Checks:** unit_consistent_recomputation
    - **Verdict:** PASS

- **Notes:** Computed samples=256.0 from sampling_rate*duration.

5. ✔ **C05_window_samples_25hz** (p.3, Sec. 2.1.4 (Data segmentation))
   - **Claim:** For 25 Hz data, a 2.56-second window translates to 64 samples.
   - **Checks:** unit_consistent_recomputation
   - **Verdict:** PASS
   - **Notes:** Computed samples=64.0 from sampling_rate*duration.
6. ✔ **C06_overlap_step_seconds** (p.3, Sec. 2.1.4 (Data segmentation))
   - **Claim:** A 50% overlap means the window advanced by 1.28 seconds for each new segment (half of 2.56 seconds).
   - **Checks:** fraction_of_value
   - **Verdict:** PASS
   - **Notes:** Computed advance=1.28 from base*(1-overlap_fraction).
7. ✔ **C07_overlap_step_samples_100hz** (p.3, Sec. 2.1.4 (Data segmentation))
   - **Claim:** At 100 Hz, 50% overlap implies advancing by 128 samples (half of 256).
   - **Checks:** fraction_of_value
   - **Verdict:** PASS
   - **Notes:** Computed advance=128.0 from base*(1-overlap_fraction).
8. ✔ **C08_overlap_step_samples_25hz** (p.3, Sec. 2.1.4 (Data segmentation))
   - **Claim:** At 25 Hz, 50% overlap implies advancing by 32 samples (half of 64).
   - **Checks:** fraction_of_value
   - **Verdict:** PASS
   - **Notes:** Computed advance=32.0 from base*(1-overlap_fraction).
9. ✔ **C09_total_models_trained** (p.4, Sec. 2.3.2 (Training Procedure))
   - **Claim:** 39 LOSO folds × 4 configurations = 156 distinct models trained.
   - **Checks:** product_equals_total
   - **Verdict:** PASS
   - **Notes:** Computed product=156.0 vs reported total=156.0.
10. ✔ **C10_training_split_counts** (p.4, Sec. 2.3.1 (LOSO Cross-Validation))
    - **Claim:** Each LOSO iteration uses 1 participant for test and the remaining 38 for training (total 39).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Computed sum(parts)=39.0 vs reported total=39.0.

11. ✔ **C11_table1_bias_sign_matches_under_counting_claim** (p.6, Sec. 3.1 + Table 1 (Bias definition and narrative))
    - **Claim:** Bias is defined as (TrueSteps − PredictedSteps), and the text claims a positive mean bias indicates under-counting; Table 1 reports positive mean bias for all configurations.
    - **Checks:** sign_consistency_check
    - **Verdict:** PASS
    - **Notes:** Checked all listed means are strictly > 0.
12. ✔ **C12_table1_claim_lowest_mae_is_hip_100hz** (p.5-6, Sec. 3.1 narrative + Table 1)
    - **Claim:** Hip_100Hz achieved the lowest Mean Absolute Error (155.29 steps).
    - **Checks:** min_among_list
    - **Verdict:** PASS
    - **Notes:** Verified claimed configuration matches the minimum value.
13. ✔ **C13_table1_claim_narrowest_ci_width_is_hip_100hz** (p.6, Sec. 3.2 narrative + Table 1)
    - **Claim:** Average 95% CI width was narrowest for Hip_100Hz (136.08 steps).
    - **Checks:** min_among_list
    - **Verdict:** PASS
    - **Notes:** Verified claimed configuration matches the minimum value.
14. ✔ **C14_table1_claim_widest_ci_width_is_wrist_25hz** (p.6, Sec. 3.2 narrative + Table 1)
    - **Claim:** Wrist_25Hz exhibited the widest CI (139.54 steps).
    - **Checks:** max_among_list
    - **Verdict:** PASS
    - **Notes:** Verified claimed configuration matches the maximum value.
15. ✔ **C15_false_positive_double_claim** (p.6, Sec. 3.1 narrative + Table 1)
    - **Claim:** Wrist_25Hz produced more than double the false positives (387.22 steps) of Hip_100Hz (169.58 steps).
    - **Checks:** inequality_ratio_check
    - **Verdict:** PASS
    - **Notes:** Checked $387.22 > 2.0 \times 169.58$ $(= 339.16)$; ratio=2.2834060620356174.
16. ✔ **C16_bland_altman_loa_from_bias_sd_hip100** (p.8, Sec. 3.4 narrative + Table 1 (Hip_100Hz bias ± SD) + stated LoA)

- **Claim:** For Hip_100Hz, mean bias is +58.4 steps and LoA ranged from -502.7 to +619.6 steps (LoA = mean_bias ± 1.96×SD). Table 1 reports bias 58.44 ± 286.30.
- **Checks:** derived_limits_of_agreement
- **Verdict:** PASS
- **Notes:** Computed LoA as mean_bias ± multiplier×SD and compared to claimed bounds.

Limitations

- Only parsed text from the PDF was available; no underlying datasets or per-subject metric vectors are included, preventing recomputation of means/SDs and statistical tests.
- Figures are referenced but numeric extraction from plot graphics/pixels is out of scope; only tabulated/narrative numbers were used.
- Some narrative numbers are rounded (e.g., Bland–Altman LoA); checks must allow small rounding tolerances.