Skepthical review: *3.3. Robustness to sensor location (hip vs. wrist)*

Summary

This manuscript evaluates how robust nine simple accelerometer-derived time- and frequency-domain features are for discriminating "Step" vs "Non-Step" **2 s** windows (**50%** overlap) across two sensor locations (hip, wrist) and two sampling frequencies (**100 Hz**, **25 Hz**) (Sec. 2.1–2.4). Triaxial acceleration is converted to ENMO, windowed, labeled as "Step" if at least one annotated step event falls inside the window, and reduced to scalar features including magnitude/variability summaries (mean, SD/variance, IQR), peak-based metrics, and spectral summaries (dominant frequency, spectral energy in **0.5–3 Hz**, spectral entropy) (Sec. 2.2–2.3). Each feature is assessed as a univariate classifier via ROC-AUC for each condition (Sec. 2.4; Fig. 1; Table 1). The main empirical pattern is that variability/magnitude features (especially SD/variance, IQR, and band-limited spectral energy) yield high AUCs, remain comparatively strong at **25 Hz**, and generally perform better at the hip than at the wrist, while simple peak-based features degrade markedly at the wrist and some spectral features behave inconsistently (Sec. 3.2–3.4, 3.6). The paper is clearly organized and the pipeline is simple and interpretable, but key methodological details are currently under-specified (data provenance, step annotation/synchronization, preprocessing and exact feature definitions). In addition, the evaluation pools hundreds of thousands of overlapping windows without participant-wise evaluation or uncertainty quantification (Sec. 2.4.1, 3.1–3.4), which limits claims about robustness and generalization. A planned demographic subgroup analysis failed due to a metadata merge error (Sec. 2.4.3, 3.5), and an internal inconsistency in reported step-event totals (Sec. 2.1 vs Sec. 3.1) needs correction.

Strengths

- Addresses a practically important "bigger picture" question for wearable deployment: which simple features remain reliable under sensor placement (hip vs wrist) and reduced sampling (**100 Hz** vs **25 Hz**) (Sec. 1).
- Simultaneous hip and wrist recordings within the same **39** participants enable clean within-subject comparisons across locations and sampling rates, reducing confounding (Sec. 2.1).
- Simple, transparent pipeline (ENMO → fixed windows → predefined feature set → AUC) supports interpretability and makes trends easy to communicate (Sec. 2.2–2.4).
- Results yield actionable guidance: variability/magnitude features and **0.5–3 Hz** spectral energy are consistently strong, whereas naive peak counting is unreliable at the wrist (Sec. 3.2–3.4, 3.6; Fig. 1).
- Limitations are at least acknowledged (notably the failed demographic merge), providing a clear set of next steps (Sec. 3.5, 4).

Major issues

1. **Data provenance, protocol, and ground-truth step annotation/synchronization are under-specified, limiting reproducibility and external validity (Sec. 2.1, Sec. 2.2.2, Sec. 3.1).** It is unclear what study the 39 participants come from, what activities/environments were included (e.g., treadmill vs free-living; walking only vs mixed tasks), how step events were obtained (manual video, foot sensors, device algorithm), annotation resolution/quality control, and how hip and wrist streams (and 100 vs 25 Hz streams) were time-aligned and drift-corrected. These details are critical because labels are derived from step timestamps.

   Recommendation: Expand Sec. 2.1 and Sec. 2.2.2 to document: (i) dataset/study origin (new vs reused; cite source if reused), setting and activity protocol, recording durations, inclusion/exclusion; (ii) step annotation method, definition of a "step event," temporal resolution, QC/validation; (iii) synchronization/alignment procedure across sensors and sampling rates (timestamping, drift handling). Explicitly describe how ambiguous segments (turns, transitions, stairs, shuffling, arm-only motion) were handled, and discuss likely label noise implications in Sec. 3.6/4.

2. **Unit of analysis and dependence: AUC is computed once per feature/condition on pooled, overlapping windows across all participants (Sec. 2.4.1, Sec. 3.1–3.4).** With 50% overlap and strong within-subject temporal correlation, treating ∼hundreds of thousands of windows as independent can inflate effective sample size, obscure between-participant heterogeneity, and overstate "robustness"/generalization to unseen individuals.

   Recommendation: Revise Sec. 2.4.1 to adopt a participant-wise evaluation: report per-participant AUCs (or leave-one-subject-out / participant-wise k-fold CV) and summarize distributions (mean±SD or median[IQR]) for each feature and condition in Sec. 3.2–3.4. If you keep a pooled AUC, also report participant-level results as the primary robustness evidence and use pooled results only as a descriptive complement.

3. **No uncertainty quantification or formal comparisons for AUC differences (Sec. 2.4.1–2.4.2, Sec. 3.2–3.4, Sec. 3.6).** Claims about "minimal degradation," "robustness," or location/sampling effects rely on point estimates without confidence intervals/tests, making it unclear which differences are statistically/practically meaningful.

   Recommendation: Add uncertainty estimates using a clustering-respecting method: e.g., bootstrap resampling by participant (not by window) to obtain 95% CIs for AUCs and key contrasts (hip vs wrist; 100 vs 25; top features vs others). If using participant-level AUCs, perform paired comparisons across participants (with multiple-comparison control where appropriate). Update Sec. 3.6 and Sec. 4 to ground robustness statements in these intervals.

4. **Task/label definition creates a mismatch with "step counting" claims: windows are labeled "Step" if ≥ 1 step occurs in a 2 s window (Sec. 2.2.2), which conflates step presence with step density and "walking-like" movement.** Many features (SD/IQR/spectral energy) will scale with the number of steps and intensity within the window; windows with 1 step vs 4 steps are treated identically, and mixed-activity windows are possible. This limits how directly conclusions translate to event-level step detection/counting (Sec. 1, Sec. 3.6, Sec. 4).

   Recommendation: Clarify throughout (Abstract, Sec. 1, Sec. 3.6, Sec. 4) that the evaluated task is window-level discrimination of step-containing windows (a proxy for walking/movement bouts), not end-to-end step counting. In Sec. 2.2.2 and Sec. 3.1, report the distribution of steps-per-positive-window and discuss mixed-window prevalence and boundary effects. If feasible, add a small sensitivity analysis: alternative labeling rules (e.g., ≥ 2 steps; or majority-of-samples walking), and/or a regression to predict step count per window, to test whether the main feature rankings and "25 Hz sufficiency" claims hold.

5. **Feature definitions and signal-processing details are incomplete, and some are internally inconsistent across sampling rates (Sec. 2.2.1–2.3.2).** In particular: (i) ENMO is written with gravity subtraction inside the square root (dimensionally inconsistent) (Intro p.2; Sec. 2.2.1 p.2–3); (ii) it is unclear whether ENMO is truncated at 0 (common ENMO definition); (iii) peak detection is under-specified (min distance, edge handling, filtering/smoothing) and is likely sampling-rate sensitive; (iv) FFT features lack details (demeaning/DC handling, window function, scaling/normalization, bin selection for 0.5–3 Hz at 25 vs 100 Hz); (v) spectral energy wording conflicts ("power spectrum" vs "squared magnitudes," potentially implying $|\mathbf{FFT}|^4$) (Sec. 2.3.2).

   Recommendation: Fix and standardize definitions in Sec. 2.2.1–2.3.2: (a) correct ENMO with explicit parentheses and units, e.g., $\mathbf{ENMO} = \mathbf{max}(0, |\mathbf{a}| - 1g)$ if truncation is intended; use the same expression everywhere; (b) specify whether 25 Hz is native or downsampled, and if downsampled, detail anti-alias filtering/decimation (Sec. 2.2.2); (c) fully specify peak detection algorithm and parameters (library/function, prominence, min peak distance in seconds, preprocessing, boundary treatment) (Sec. 2.3.1); (d) fully specify FFT pipeline (demean/detrend, DC exclusion, window type, zero padding, PSD normalization) and give explicit formulas for dominant frequency, spectral energy (as a band power/integral), and spectral entropy (including probability normalization and log base) (Sec. 2.3.2). Ensure frequency-band implementation is comparable across 25/100 Hz (bin mapping).

6. **The demographic subgroup analysis failed due to a metadata merge/parsing error, but the failure mode and data-integrity implications are not sufficiently diagnosed (Sec. 2.4.3, Sec. 3.1, Sec. 3.5).** Without clear verification,

readers cannot be sure only demographics were impacted (and not participant IDs, file-to-subject mapping, or label alignment).

Recommendation: In Sec. 2.4.3 and Sec. 3.5, provide a concrete diagnosis (which keys failed, how IDs differed, what rows were missing/duplicated). Explicitly document integrity checks that confirm the main feature–label dataset is correct (participant-level counts, file hashes/checksums, sanity checks on time ranges). If any participants/sessions were excluded, list them. If feasible, fix the merge and include at least a minimal subgroup summary; otherwise, strengthen generalization caveats in Sec. 4.

7. **Internal inconsistency in reported step-event totals: $48,721$ (Methods Sec. 2.1) vs $62,904$ (Results Sec. 3.1).** This undermines trust in dataset accounting and downstream results.

   Recommendation: Reconcile these numbers by explicitly stating scopes/filters for each count (e.g., before/after exclusions, per-condition vs combined, annotation revisions). Correct the manuscript so the same quantity is reported consistently, and ensure any derived percentages in Sec. 3.1 match the corrected totals.

8. **Conclusions about robustness and the sufficiency of 25 Hz are broader/stronger than warranted given the described cohort (39 healthy adults), unclear activity context, missing subgroup analyses, and the current pooled-window evaluation (Sec. 3.6, Sec. 4).**

   Recommendation: In Sec. 3.6 and Sec. 4, restrict claims to the study conditions once fully specified (protocol, population, annotation method). Rephrase statements implying broad deployment across "diverse wearable applications" to note that validation is still needed in older adults, clinical/pathological gait, and multi-day free-living settings. Tie any "**25 Hz** is sufficient" claim to specific features and to participant-level uncertainty estimates (after addressing the evaluation issues above).

Minor issues

1. Windowing/labeling choices are not justified or stress-tested: **2 s** windows with **50%** overlap (Sec. 2.2.2) may smooth short bouts and create mixed-content windows, affecting AUC and feature rankings; overlap further increases dependence.

   Recommendation: In Sec. 2.2.2, justify **2 s** relative to plausible cadence ranges and intended deployment. If feasible, add a brief sensitivity analysis (e.g., **1 s** and/or **4 s** windows; different overlaps) and summarize whether the main rankings and hip/wrist and **100/25** conclusions remain stable (Sec. 3.6).

2. AUC direction handling is unclear for features expected to be lower during walking (e.g., spectral entropy as described) (Sec. 2.3.2, Sec. 2.4.1). The text's interpretation of AUC as $P(\mathrm{feature} | \mathrm{step} > \mathrm{feature} | \mathrm{non-step})$ can conflict with "lower is more step-like."

Recommendation: In Sec. 2.4.1, state explicitly how direction is handled (e.g., compute ROC with the appropriate inequality per feature; or report $\mathbf{max(AUC, 1-AUC)}$; or negate features like entropy). Ensure plots/tables use a consistent convention.

3. Comparability across sampling rates is not explicitly validated for sampling-sensitive features (peak_count, spectral energy) (Sec. 2.3, Sec. 3.4). Without careful normalization, observed "**25 Hz** sufficiency" may reflect definition/implementation artifacts rather than true robustness.

   Recommendation: Add a short paragraph in Sec. 2.3.2 and/or Sec. 3.4 explaining why each frequency-domain measure is sampling-rate invariant under your implementation (or how it is normalized to approximate a band-power integral). If **25 Hz** is downsampled from **100 Hz**, consider reporting a controlled downsampling experiment to isolate sampling effects from device/hardware differences.

4. Figure 1 (and Table 1) currently emphasize point estimates and are hard to use for "robustness deltas" (Sec. 3.2–3.4). The grouped bars are dense, and there are no uncertainty indicators.

   Recommendation: Revise Fig. 1 to show uncertainty (participant-level points or CIs) and to better emphasize within-feature deltas across conditions (e.g., dot-and-line per feature; small multiples). Add reference lines at $\mathbf{AUC = 0.5}$ and optionally other benchmarks; use colorblind-safe encoding; ensure vector export and readable labels. Consider adding an explicit $\mathbf{\Delta AUC}$ panel (hip→wrist, $\mathbf{100 \rightarrow 25}$).

5. Related work and feature-set justification are thin and sometimes cite tangential domains rather than gait/step-counting literature (Sec. 1, Sec. 2.3, References). Novelty (systematic per-feature robustness across placement and sampling) is present but not crisply positioned.

   Recommendation: Add a focused related-work paragraph in Sec. 1 on prior findings about hip vs wrist and sampling-rate reduction for step counting, and clearly state the paper's unique contribution. In Sec. 2.3, justify why these nine features are a representative low-cost set and briefly note omitted but relevant feature families (e.g., autocorrelation/periodicity measures). Tighten citations toward directly relevant accelerometry/gait sources.

6. Class imbalance and operational metrics: AUC is prevalence-insensitive, but deployment decisions often need operating-point metrics; imbalance may differ by condition (Sec. 2.4.1, Sec. 3.1).

   Recommendation: Report class prevalence per condition (Hip/**100**, Hip/**25**, Wrist/**100** , Wrist/**25**) in Sec. 3.1. Consider adding PR-AUC or sensitivity/specificity at fixed FPR (or fixed sensitivity) for the top features, to complement AUC and support "practical" conclusions (Sec. 3.6).

7. Ethics/data governance are not mentioned for human-participant data (Sec. 2.1).

Recommendation: Add a brief ethics statement (IRB/ethics approval, consent, and data-sharing constraints) in Sec. 2 (new subsection if appropriate) following venue norms.

8. Placeholder/incorrect author–affiliation line and potentially irrelevant references reduce professionalism and may violate venue requirements (Title page; References).

   Recommendation: Replace the placeholder affiliation text with correct author/affiliation information (or proper anonymization for review). Audit the bibliography to remove/replace unrelated references and ensure each citation supports the corresponding claim.

Very minor issues

1. Typographical/style inconsistencies (units "100Hz" vs "100~Hz", "hipworn", inconsistent feature naming/capitalization, quote styles for class labels, reference formatting) (Sec. 2.2.2, Sec. 3.x, Sec. 4, References).

   Recommendation: Proofread and standardize units, hyphenation, feature names/abbreviations (define once and reuse consistently), quote style for "Step/Non-Step," and reference formatting per the target venue.

2. Section heading formatting inconsistency (e.g., stray leading "#" in a heading) (Sec. 3.3).

   Recommendation: Normalize heading formatting/numbering to match the manuscript's overall style and the target venue template.

3. Dominant frequency definition may inadvertently select DC if ENMO is not mean-centered or if DC is not excluded (Sec. 2.3.2).

   Recommendation: State whether ENMO windows are demeaned and whether the DC bin is excluded; define the frequency search range used for "dominant frequency."

Key statements and references

- ⚠ **The ENMO (Euclidean Norm Minus One) signal, defined as ENMO** $= \sqrt{a_x^2 + a_y^2 + a_z^2} - 1g$**, is used as an orientation-invariant measure of motion intensity for wearable accelerometer data, effectively removing the constant gravitational component and centering the signal around zero during stillness, and has been adopted in prior accelerometry studies as a basis for subsequent feature engineering and analysis.**
- *Reference(s):* Suibkitwanchai et al., **2020**, Peng and Dinçer, **2024**, Acar-Denizli and Delicado, **2024**
- *Justification:* Suibkitwanchai et al., **2020** state they compute the Euclidean norm from triaxial acceleration and then remove Earth's gravity by using ENMO (Euclidean norm minus one; citing van Hees, **2013**), and they base all subsequent analyses (IS,

IV, DFA, PoV) on ENMO. Peng and Dinçer, 2024 describe enmo as the Euclidean norm of accelerometer signals with negative values rounded to zero and note it is a commonly computed feature; they also use it as an input for event detection. Acar-Denizli and Delicado, 2024 cite multiple studies that analyze data in ENMO units (e.g., Wrobel et al., 2021; Wu et al., 2019), indicating adoption in prior work. However, none of the papers explicitly describe ENMO as an "orientation-invariant" measure or explicitly state that it "centers the signal around zero during stillness," though these properties are implied by the norm and gravity subtraction. Hence, the claim is only partially supported.

- **✖ Ground-truth step annotations from prior work are used to label 2-second ENMO windows as "Step" if they contain one or more annotated step events and "Non-Step" otherwise, enabling supervised evaluation of step vs. non-step discrimination performance of accelerometer-derived features.**
- *Reference(s):* Waks et al., 2017
- *Justification:* Waks et al., 2017 derives ground-truth gait events from ankle gyroscopes (per prior methods) to evaluate step counts and gait phase timing, using peak detection on the acceleration norm and sensor fusion. The study does not use ENMO, does not segment data into 2-second windows, and does not label windows as Step/Non-Step for supervised feature discrimination.
- **✖ The selection of nine accelerometer-derived features—including time-domain measures (mean, standard deviation, variance, interquartile range, peak count, mean peak prominence) and frequency-domain measures (dominant frequency, spectral energy 0.5–3.0 Hz, spectral entropy)—is motivated by prior literature demonstrating their utility for characterizing motion and gait-related activity from wearable sensors.**
- *Reference(s):* Sigcha et al., 2024, Mridula et al., 2023, Togootogtokh and Klasen, 2021
- *Justification:* Neither Mridula et al., 2023 nor Togootogtokh and Klasen, 2021 discuss accelerometer-derived features or gait/motion analysis from wearable sensors. Mridula et al., 2023 focuses on EEG and eye-movement features (FFT power spectrum, differential entropy) for emotion recognition, and Togootogtokh and Klasen, 2021 focuses on speech features (melspectrogram) for emotion recognition. The specific accelerometer features listed and their motivation from prior literature are not addressed.
- **⚠ The Area Under the Receiver Operating Characteristic Curve (AUC) is employed as the primary scalar metric to quantify the discriminative power of each individual feature for distinguishing "Step" from "Non-Step" windows, following established methodology for evaluating binary classifier performance and interpreting discrimination ability in risk prediction and signal classification contexts.**

- *Reference(s):* Stern, 2021, Fewell, 2024
- *Justification:* General methodology is supported but the specific application is not shown. Stern, 2021 describes AUC as a measure of separation/discrimination and its use in evaluating binary risk prediction models. Fewell, 2024 discusses ROC curves in signal vs background classification and states that area under the curve is a commonly used, satisfactory metric. However, neither paper mentions using AUC as the primary scalar metric for each individual feature or the specific 'Step' vs 'Non-Step' window task, so the claim is only partially supported.

Mathematical consistency audit

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

Maths relevance: light

The paper is primarily methodological/empirical with limited explicit mathematics. The main explicit formula is the ENMO transformation; other 'math' consists of standard descriptive-statistics feature definitions (mean/SD/variance/IQR/peaks) and high-level descriptions of FFT-derived features and AUC. There are no multi-step derivations to audit; the audit therefore focuses on dimensional consistency, unambiguous formulas, and definition consistency between sections.

Checked items

1. ✖ **ENMO definition (dimensional consistency and parentheses)** (Introduction (p.2) and Sec. 2.2.1 (p.2–3))
    - **Claim:** ENMO is computed from triaxial acceleration as $\mathbf{ENMO} = \sqrt{a_x^2 + a_y^2 + a_z^2 - 1g}$, providing an orientation-invariant motion intensity centered around zero.
    - **Checks:** dimensional/units consistency, algebra/parentheses ambiguity, sanity case (stillness)
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** a_x, a_y, a_z denote acceleration components with consistent units, $1g$ denotes gravitational acceleration
    - **Notes:** As written, $a_x^2 + a_y^2 + a_z^2$ has units of acceleration^2, but $1g$ has units of acceleration, so the subtraction under the square root is ill-typed. It also risks a negative radicand. To match the stated intent ('minus one g' after taking the Euclidean norm), the formula must be written unambiguously as $\mathbf{ENMO} = \sqrt{a_x^2 + a_y^2 + a_z^2} - 1g$ (optionally with truncation if intended).
2. ✔ **Window sample count at 100 Hz** (Sec. 2.2.2, p.3)
    - **Claim:** A 2-second window corresponds to 200 samples at 100 Hz.
    - **Checks:** arithmetic consistency

- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Sampling frequency is exactly 100 samples/second, Window duration is exactly 2 seconds
- **Notes:** $2\ \text{s} \times 100\ \text{Hz} = 200$ samples.

3. ✔ **Window sample count at 25 Hz** (Sec. 2.2.2, p.3)
   - **Claim:** A 2-second window corresponds to 50 samples at 25 Hz.
   - **Checks:** arithmetic consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Sampling frequency is exactly 25 samples/second, Window duration is exactly 2 seconds
   - **Notes:** $2\ \text{s} \times 25\ \text{Hz} = 50$ samples.
4. ✔ **50% overlap implies 1-second stride** (Sec. 2.2.2, p.3)
   - **Claim:** Using 50% overlap with a 2-second window means each window advances by 1 second.
   - **Checks:** definition consistency, arithmetic consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Overlap is defined as fraction of window length shared between consecutive windows
   - **Notes:** Stride $= (1 - 0.5) \times 2\ \text{s} = 1\ \text{s}$.
5. ✔ **Variance definition equals SD squared** (Sec. 2.3.1, p.3)
   - **Claim:** Signal variance is defined as the square of the standard deviation.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Standard deviation is the (nonnegative) square root of variance
   - **Notes:** This is mathematically consistent as a definitional relationship (assuming population vs sample convention is consistently applied; the paper does not specify which).
6. ✔ **IQR definition** (Sec. 2.3.1, p.3)
   - **Claim:** IQR is defined as the 75th percentile minus the 25th percentile of ENMO within the window.
   - **Checks:** definition consistency, sanity bounds
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Percentiles are computed on the set of window samples
   - **Notes:** IQR ≥ 0 by definition; consistent with use as dispersion.
7. ⚠ **Spectral energy definition vs 'power spectrum' wording** (Sec. 2.3.2, p.3)

- **Claim:** After FFT and obtaining a power spectrum, spectral energy (**0.5–3.0 Hz**) is computed as the sum of squared magnitudes of FFT components in that band.
- **Checks:** notation/definition consistency, dimensional consistency
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** Power spectrum may mean $|X[k]|^2$, FFT magnitude means $|X[k]|$
- **Notes:** The text simultaneously references a 'power spectrum' and then 'squared magnitudes of FFT components'. If power spectrum is already $|X|^2$, then summing squared magnitudes would imply $|X|^4$. Likely the intended computation is $\sum |X|^2$ over the band, but this is not stated unambiguously.

8. ⚠ **Spectral entropy definition missing** (Sec. 2.3.2, p.3)
    - **Claim:** Spectral entropy measures flatness/uniformity of the power spectrum; lower indicates concentrated energy.
    - **Checks:** missing definition/formula check
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** Entropy requires a normalized distribution over frequency bins
    - **Notes:** No explicit entropy formula is provided (normalization, log base, frequency range, handling of zeros), preventing verification and comparability across conditions.

9. ⚠ **AUC interpretation vs feature directionality** (Sec. 2.4.1, p.4; plus feature descriptions in Sec. 2.3.2, p.3)
    - **Claim:** AUC is the probability a random Step window is ranked higher (has a higher feature value) than a random Non-Step window.
    - **Checks:** definition consistency across sections, sign/direction sanity check
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** AUC computed from raw feature values with Step as positive class, Some features may be negatively associated with Step (e.g., entropy lower for steps)
    - **Notes:** The paper states lower spectral entropy corresponds to rhythmic walking, which would imply Step windows may have *lower* values than Non-Step. Under the stated AUC-as-'higher-is-more-positive' interpretation, such a feature would yield **AUC < 0.5** unless the direction is inverted. The paper does not specify direction handling.

10. ⚠ **Dominant frequency definition and DC handling** (Sec. 2.3.2, p.3)
    - **Claim:** Dominant frequency is the frequency with the highest magnitude in the power spectrum.
    - **Checks:** definition completeness / edge cases

- **Verdict:** UNCERTAIN; confidence: medium; impact: minor
- **Assumptions/inputs:** FFT is applied to windowed ENMO data
- **Notes:** If ENMO has a nonzero mean in a window, the **0 Hz** (DC) bin can dominate unless excluded or the signal is demeaned. The paper does not specify whether DC is excluded or whether demeaning/windowing is applied.

Limitations

- Audit was performed on the provided parsed text of the **9**-page PDF; equation numbering, tables, and any mathematical details embedded solely in figures were not fully available for verification.
- The paper contains few explicit formulas and no multi-step derivations; several signal-processing quantities (FFT scaling, power spectrum definition, entropy formula) are described only qualitatively, limiting the ability to verify mathematical equivalence across conditions without inventing missing steps.

Numerical results audit

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

Out of **22** numeric checks, **21** passed and **1** failed. The only failure is a cross-section inconsistency in total annotated step events between Methods and Results. All other checked arithmetic relationships (percentages, window counts, sampling/window math, and reported AUC drops/differences) match exactly or within stated tolerances.

Checked items

1. ✔ **C1** (p.2, Methods 2.1 Study Participants and Data Collection)
    - **Claim:** Cohort consisted of **19** males (**48.7%**) and **20** females (**51.3%**) out of **39** participants.
    - **Checks:** percent_of_total_and_sum_to_100
    - **Verdict:** PASS
    - **Notes:** Percent checks and sums (counts and percents) all consistent within tolerance.
2. ✔ **C2** (p.2, Methods 2.1 Study Participants and Data Collection)
    - **Claim:** Age distribution: **25** individuals (**64.1%**) aged **18–39**; **14** individuals (**35.9%**) aged **40–65** (total **39**).
    - **Checks:** percent_of_total_and_sum_to_100
    - **Verdict:** PASS
    - **Notes:** Percent checks and sums (counts and percents) all consistent within tolerance.
3. ✔ **C3** (p.4, Results 3.1 Data preparation and cohort summary)

- **Claim:** 156 data files correspond to 39 participants across four experimental conditions.
- **Checks:** product_equals_total
- **Verdict:** PASS
- **Notes:** $39 \times 4 = 156$ exactly.

4. ✔ **C4** (p.4, Results 3.1 Data preparation and cohort summary)
   - **Claim:** Total windows $545,350$ equals Step windows $157,710$ plus Non-Step windows $387,640$.
   - **Checks:** parts_sum_to_total
   - **Verdict:** PASS
   - **Notes:** $157,710 + 387,640 = 545,350$ exactly.
5. ✖ **C5** (p.2, Methods 2.1 vs p.4, Results 3.1)
   - **Claim:** Total annotated step events reported as $48,721$ (Methods) vs $62,904$ (Results).
   - **Checks:** cross_section_numeric_consistency
   - **Verdict:** FAIL
   - **Notes:** Mismatch detected: Results minus Methods $= 14,183$; equality would be expected if same scope/definition.
6. ✔ **C6** (p.3, Methods 2.2.2 Sliding Window Segmentation and Labeling)
   - **Claim:** 2-second window corresponds to 200 samples at 100 Hz and 50 samples at 25 Hz.
   - **Checks:** unit_rate_times_duration
   - **Verdict:** PASS
   - **Notes:** $100 \times 2 = 200$ and $25 \times 2 = 50$ exactly.
7. ✔ **C7** (p.3, Methods 2.2.2)
   - **Claim:** 50% overlap implies 1-second stride for a 2-second window.
   - **Checks:** overlap_stride_consistency
   - **Verdict:** PASS
   - **Notes:** Stride computed as $2 \times (1 - 0.5) = 1$ exactly.
8. ✔ **C8** (p.6, Results 3.4 (hip sampling frequency robustness))
   - **Claim:** Hip: std_dev and variance AUC dropped by 0.0036 (from 0.9839 at Hip/100 Hz to 0.9803 at Hip/25 Hz).
   - **Checks:** difference_equals_reported_drop
   - **Verdict:** PASS
   - **Notes:** $0.9839 - 0.9803$ matches reported drop within tolerance.
9. ✔ **C9** (p.5, Results 3.3 (location robustness at 100 Hz))

- **Claim:** At **100 Hz**: std_dev and variance AUC decreased by **0.0704** (from **0.9839** at Hip/**100 Hz** to **0.9135** at Wrist/**100 Hz**).
- **Checks:** difference_equals_reported_drop
- **Verdict:** PASS
- **Notes:** **0.9839 − 0.9135** matches reported drop within tolerance.

10. ✔ **C10** (p.5, Results 3.3 (location robustness at **100 Hz**))
    - **Claim:** At **100 Hz**: spectral_energy drop of **0.0509** (from **0.9839** to **0.9330**).
    - **Checks:** difference_equals_reported_drop
    - **Verdict:** PASS
    - **Notes:** **0.9839 − 0.9330** matches reported drop within tolerance.
11. ✔ **C11** (p.5, Results 3.3 (IQR location robustness at **100 Hz**))
    - **Claim:** IQR AUC decreased by **0.0395** at **100 Hz** (**0.9817** at Hip/**100 Hz** to **0.9422** at Wrist/**100 Hz**).
    - **Checks:** difference_equals_reported_drop
    - **Verdict:** PASS
    - **Notes:** **0.9817 − 0.9422** matches reported drop within tolerance.
12. ✔ **C12** (p.5, Results 3.3 (IQR location robustness at **25 Hz**))
    - **Claim:** IQR AUC decreased by **0.0357** at **25 Hz** (**0.9771** at Hip/**25 Hz** to **0.9414** at Wrist/**25 Hz**).
    - **Checks:** difference_equals_reported_drop
    - **Verdict:** PASS
    - **Notes:** **0.9771 − 0.9414** matches reported drop within tolerance.
13. ✔ **C13** (p.5, Results 3.3 (peak_count location degradation at **100 Hz**))
    - **Claim:** peak_count AUC plummeted by **0.2809** (from **0.9598** at Hip/**100 Hz** to **0.6789** at Wrist/**100 Hz**).
    - **Checks:** difference_equals_reported_drop
    - **Verdict:** PASS
    - **Notes:** **0.9598 − 0.6789** matches reported drop within tolerance.
14. ✔ **C14** (p.6, Results 3.4 (wrist sampling frequency robustness: peak_count increase))
    - **Claim:** peak_count at wrist increased in AUC from **0.6789** (**100 Hz**) to **0.7658** (**25 Hz**).
    - **Checks:** direction_and_difference
    - **Verdict:** PASS
    - **Notes:** Direction check passed (**0.7658 > 0.6789**); computed increase = **0.0869**.

15. ✔ **C15** (p.6, Results 3.4 (hip sampling frequency robustness: dominant_freq increase))
    - **Claim:** dominant_freq at hip increased from **0.6500** (**100 Hz**) to **0.7317** (**25 Hz**).
    - **Checks:** direction_and_difference
    - **Verdict:** PASS
    - **Notes:** Direction check passed (**0.7317** > **0.6500**); computed increase = 0.0817.
16. ✔ **C16** (p.5, Results 3.2 (dominant_freq examples))
    - **Claim:** dominant_freq yielded AUCs **0.5265** at Wrist/**100 Hz** and **0.7317** at Hip/**25 Hz** (examples).
    - **Checks:** range_and_consistency_with_random_chance_reference
    - **Verdict:** PASS
    - **Notes:** Both AUCs within $[0, 1]$; wrist/**100 Hz** above **0.5** by **0.0265**.
17. ✔ **C17** (p.5, Results 3.2 (spectral_entropy range statement))
    - **Claim:** spectral_entropy AUCs ranged from **0.5136** to **0.8474**.
    - **Checks:** range_ordering_and_bounds
    - **Verdict:** PASS
    - **Notes:** Ordering and bounds check passed: $0 \leq 0.5136 \leq 0.8474 \leq 1$.
18. ✔ **C18** (p.6, Results 3.4 (spectral_entropy frequency reduction, hip))
    - **Claim:** spectral_entropy AUC at hip dropped from **0.7210** (**100 Hz**) to **0.5136** (**25 Hz**).
    - **Checks:** difference_and_direction
    - **Verdict:** PASS
    - **Notes:** Direction check passed (**0.7210** > **0.5136**); computed drop = **0.2074**.
19. ✔ **C19** (p.6, Results 3.4 (spectral_entropy frequency reduction, wrist))
    - **Claim:** spectral_entropy AUC at wrist dropped from **0.8474** (**100 Hz**) to **0.7164** (**25 Hz**).
    - **Checks:** difference_and_direction
    - **Verdict:** PASS
    - **Notes:** Direction check passed (**0.8474** > **0.7164**); computed drop = **0.1310**.
20. ✔ **C20** (p.6, Results 3.4 (mean feature hip drop))
    - **Claim:** Mean AUC at hip dropped from **0.8488** (**100 Hz**) to **0.7293** (**25 Hz**).
    - **Checks:** difference_and_direction
    - **Verdict:** PASS
    - **Notes:** Direction check passed (**0.8488** > **0.7293**); computed drop = **0.1195**.

21. ✔ **C21** (p.6, Results 3.4 (wrist sampling frequency robustness: iqr difference))
    - **Claim:** IQR at wrist had difference of 0.0008 between 100 **Hz** and 25 **Hz** (AUCs 0.9422 and 0.9414).
    - **Checks:** difference_equals_reported
    - **Verdict:** PASS
    - **Notes:** $|0.9422 - 0.9414|$ matches reported 0.0008 within tolerance.
22. ✔ **C22** (p.5-7, Narrative claims about thresholds)
    - **Claim:** Claim: variability/energy features achieved high AUCs (all > 0.91) across all conditions; and later: IQR (AUC > 0.94) and spectral_energy (AUC > 0.92) for wrist data.
    - **Checks:** threshold_claim_vs_listed_values
    - **Verdict:** PASS
    - **Notes:** For explicitly listed wrist values: IQR 0.9422 and 0.9414 exceed 0.94; spectral_energy 0.9330 exceeds 0.92.

Limitations

- Only the provided PDF text was used; Table 1 and Table 2 numeric contents were not available in the parsed text, preventing verification of claims that require the full table.
- No values were extracted from plots/figures (no pixel/bar reading), so any claim relying solely on Figure 1 without explicit numbers in the text is not directly checkable.
- Many aggregate quantities (e.g., total windows from recording durations) cannot be recomputed without per-participant durations and preprocessing rules not explicitly specified.
- The detected mismatch in annotated step-event totals ($48,721$ vs $62,904$) cannot be resolved to a single correct value without additional clarification about definitions/scope.