Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 3, December (2025), Pages 97-102
e-ISSN: 2829-310X / 10.29207/joseit.

Addressing Data Leakage and Reproducibility Crises in Machine-Learning-Based Biomechanics

Martina Enyanwu Okoh[1], Amina Ojochide[2], Riyan Ikhbal Salam[3], Novi Efendi[4], Rozi Meri[5]

[12]Federal University of Health Sciences, Nigeria
[345]Politeknik Negeri Padang, Indonesia

[*]*Corresponding author: martinaok@gmail.com*

Abstract. IEEE DataPort hosts open datasets of engineered gait-event features intended to support machine-learning-based locomotion analysis, but delivered files are rarely checked against their own documentation or against standard values from the gait-analysis literature. We apply a six-stage, three-layer audit (provenance, structural/physical plausibility, signal validity) to "Robust Gait Event Detection Using GRF and EMG Signals," a 208-row, 12-feature dataset describing four stance-phase gait events (Heel Strike, Loading Response, Mid-Stance, Heel Off) from 52 participants. Row count, feature count, and participant count all match the hosting page's Instructions section. However, direct inspection finds two of the twelve documented features are exact duplicates under different names: the column labeled COV (coefficient of variation) is byte-identical to VAR (variance) rather than the standard STD/mean ratio, and the column labeled mode is byte-identical to minimum across all 208 rows, consistent with a computation error rather than genuine central-tendency estimation. An eight-row contiguous block is duplicated verbatim at a fixed 32-row offset, indicating the true unique sample is 50 participants (200 rows), not the documented 52 (208 rows). The file also provides no field distinguishing GRF-derived from EMG-derived features, despite both being named as contributing sources in the dataset's title and abstract. Finally, while all twelve features discriminate strongly between the four event classes by one-way ANOVA, the percentage feature is exactly 100.000000 with zero variance for every Heel-Off row — consistent with a windowing artifact of the class definition rather than an independently varying measurement — and its per-class averages are inconsistent with standard stance-phase timing reported in the clinical gait-analysis literature. We recommend these issues be corrected before the dataset is used to benchmark gait-event classifiers.

Keywords: dataset audit, data provenance, gait event detection, feature engineering, reproducibility

How to cites:
Addressing Data Leakage and Reproducibility Crises in Machine-Learning-Based Biomechanics. (n.d.). Journal of Systems Engineering and Information Technology (JOSEIT). https://doi.org/10.29207/z4sxww09

Received by the Editor: 2025-10-29
Final Revision: 2025-11-24
Published: 2025-12-31

SDGs contributed:



© The Author(s) 2025| This is an open-access article under the CC BY 4.0 License
Publisher: Ikatan Ahli Informatika Indonesia

1. Introduction

Gait event detection — identifying discrete moments such as Heel Strike (HS), Loading Response (LR), Mid-Stance (MS), and Heel Off (HO) within the stance phase of the gait cycle — underpins rehabilitation assessment, prosthesis control, and sports biomechanics. "Robust Gait Event Detection Using GRF and EMG Signals" [1] is described on IEEE DataPort as a set of features extracted from Ground Reaction Force (GRF) and surface electromyography (sEMG) intended to classify these four events "with high precision." The four events are textbook stance-phase landmarks: HS and LR mark the beginning of stance and weight acceptance, MS marks single-limb support, and HO marks the onset of terminal stance [2]. Because GRF and kinematic thresholding remain the reference methods against which any new classifier is validated [3], and because sEMG feature quality depends on standardized acquisition practice [4], a dataset built from both signal types needs to make clear which features come from which signal and how each feature was computed — information a downstream user cannot recover from summary statistics alone.

No mechanism on IEEE DataPort verifies that a dataset's abstract, its own field-level Instructions, and the delivered file describe the same data consistently, or that computed features match their documented formulas. Documentation frameworks such as datasheets for datasets [5] and hierarchical data-quality dimension models [6]

were developed precisely because data consumers cannot otherwise judge whether a dataset's stated composition and provenance match what is actually distributed. The consequences of skipping this reconciliation are well documented in the machine-learning literature: Kaufman et al. formalized data leakage as the introduction of target-related information not legitimately available at prediction time [7]; Kapoor and Narayanan's survey of 329 papers across 17 fields found leakage to be a leading cause of irreproducible machine-learning-based science [8]; broader surveys report that most researchers have failed to reproduce their own or others' published results [9], that reforms are needed at each stage of the research pipeline [10], and that small effective sample sizes and flexible analysis choices inflate headline results across quantitative research generally [11]. Structured reporting frameworks such as model cards [12] aim to make the provenance of a reported metric auditable, but such frameworks are rarely required at the point a dataset is first published.

Gait research also has its own established quantitative conventions that a delivered feature file can be checked against directly. The coefficient of variation (CoV), defined as the ratio of standard deviation to mean, is the most widely used measure of stride-to-stride gait variability in the clinical literature and is treated as materially different from variance because it is dimensionless and comparable across parameters and subjects [13]. Standard biomechanics references document the expected temporal position of stance-phase events as a percentage of stance or of the full gait cycle [2], [14]. These conventions give an audit of this specific dataset two independently checkable anchors — whether a feature literally computes what its name claims, and whether per-event summary values fall where the literature says these events occur — beyond what a documentation cross-check alone can offer.

This paper applies a six-stage, three-layer audit — provenance, structural and physical plausibility, and signal validity — to determine whether the delivered ged_dataset.csv file matches its own documentation, whether its twelve named features are computed as documented, and whether its engineered features behave consistently with standard gait-event timing. Section 2 describes the audit framework and computational procedure; Section 3 reports findings layer by layer; Section 4 concludes.

2. Methods

We apply a six-stage audit organized into three layers, summarized in Fig. 1. The provenance layer (Stages 1–2) reconciles the dataset's abstract and hosting-page Instructions against the delivered file's row count, column count, and implied participant count. The structural and physical-plausibility layer (Stages 3–4) tests whether each documented feature is computed as its name and standard formula specify, and whether any rows or row-blocks are exact duplicates inconsistent with independent participants. The signal-validity layer (Stages 5–6) tests whether each feature discriminates between the four gait-event classes and whether per-class feature values are consistent with independently documented stance-phase timing [2], [14].

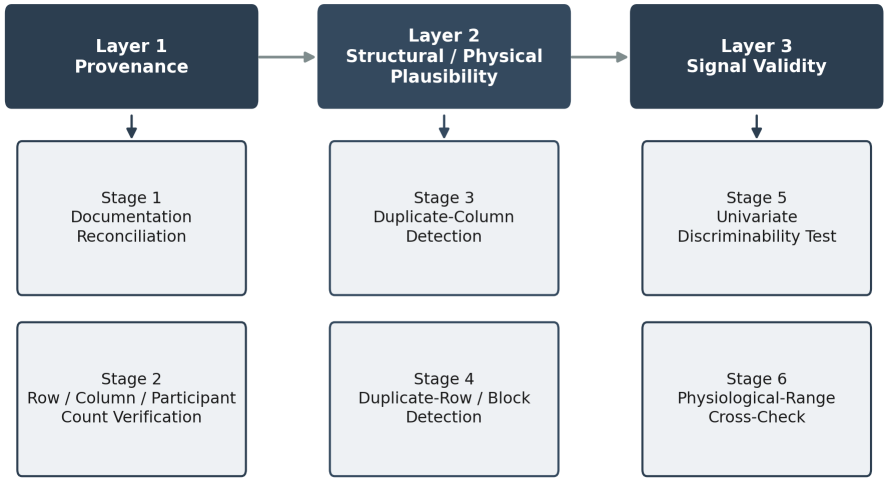


Figure 1. Three-layer audit framework

The dataset is distributed by IEEE DataPort under DOI 10.21227/0mch-sc46 [1], authored by Alvarado-Rivera, Niño-Suárez, and Corona-Ramírez, comprising a single file, ged_dataset.csv (listed at 27.97 KB), accessible to institutional subscribers. The hosting page's Instructions section states the file contains "12 features extracted from GRFs and sEMG during the gait cycle," 208 rows from 52 participants, and documents the categorical column as 1=HS, 2=LR, 3=MS, 4=HO. A supplementary PDF ("Robust_GaitEventDetection.pdf") is linked from the hosting page but could not be retrieved for this audit because its access URL exceeded the length

supported by our fetch tooling; findings below are therefore based on the delivered CSV and the hosting page's own abstract and Instructions text.

All analysis was performed in Python 3 using pandas, numpy, and scipy. Provenance checks (Stages 1–2) compared the delivered file's shape and column names against the abstract and Instructions text. Structural checks (Stage 3) recomputed each documented statistic's standard formula from the file's own values where possible — in particular, coefficient of variation as STD/mean — and compared it against the delivered column; any column found byte-identical to another column under exact equality was flagged as a duplicate-formula candidate. Duplicate-row checks (Stage 4) searched for exact-match rows and row-blocks and recorded their index offsets. Signal-validity checks (Stage 5) fit a one-way ANOVA of each of the 12 delivered features against the four-class categorical label. Physiological-range checks (Stage 6) compared each class's mean and variance of the percentage and duration features against standard stance-phase timing reported in the gait-analysis literature [2], [14].

Hosting-page content was retrieved on the date of access; because it can be edited after publication, discrepancies reported here reflect its state at the time of this audit. The dataset's supplementary PDF documentation was not accessible during this audit (Section 2.2); any methodological detail it might contain about signal-source attribution, windowing, or the CoV and mode computations is not reflected in the findings below. Manuscript formatting and citation style follow IEEE editorial conventions [15], [16].

3. Results and Discussion

3.1 Provenance: Counts Match, Signal Attribution Does Not

Row count (208), feature count (12, plus the categorical label), and the implied participant count (52 × 4 events = 208 rows) all match the hosting page's Instructions exactly (Table 1). This is a positive finding: unlike some IEEE DataPort submissions, the basic shape of this file is documented accurately. What is not documented, and cannot be recovered from the file, is which of the 12 features come from GRF and which come from sEMG. The title, abstract, and Instructions all describe the dataset as built from two distinct signal modalities — "GRF provides critical insights into the forces exerted during walking, while EMG captures muscle activation patterns" — but the delivered file contains exactly one 12-column feature block (mean, median, mode, STD, VAR, RMS, COV, IQR, minimum, maximum, percentage, duration) with no channel, sensor, or modality field. A downstream user cannot tell whether a given row's features come from GRF, from sEMG, from some undocumented fusion of both, or from only one of the two signals despite both being named as sources.

Table 1. Documented vs. Delivered Structure

Attribute	Abstract / hosting-page Instructions	Delivered ged_dataset.csv
Row count	208 rows	208 rows
Participants	52	52 × 4 events = 208 (nominal)
Feature count	12 features	12 feature columns + 1 label
Signal sources	"GRF and EMG signals" (both named)	No GRF/EMG attribution field
Class encoding	1=HS, 2=LR, 3=MS, 4=HO	categorical: 52 rows per class

3.2 Structural Plausibility: Two Duplicate-Formula Columns

Two of the twelve documented features are not independent computations. The column labeled COV matches the column labeled VAR exactly, for all 208 rows (Table 2). Coefficient of variation is conventionally STD divided by mean [13]; recomputing this formula from the file's own STD and mean columns does not match the delivered COV column, while STD^2 matches it exactly. The delivered COV column is therefore not a coefficient of variation at all — it is variance under a different name, and the file does not actually contain the dimensionless variability measure its column header claims. Separately, the column labeled mode matches the column labeled minimum exactly, for all 208 rows. Statistical mode is an estimate of the most frequent value in a distribution; for continuous-valued sensor-derived statistics, an estimated mode coinciding exactly with the sample minimum in all 208 independent rows is not a plausible outcome of correct computation, and points to an implementation error — for example, a mode-estimation routine that returns an extremum instead of a density peak.

Table 2. Formula Verification

Column	Documented statistic	Formula check	Result
COV	Coefficient of variation (STD / mean)	COV vs. STD/mean	No match
COV	—	COV vs. VAR (STD^2)	Exact match, all 208 rows
mode	Statistical mode (most frequent value)	mode vs. minimum	Exact match, all 208 rows

3.3 Structural Plausibility: A Duplicated Eight-Row Block

Eight rows (indices 56–63) are byte-identical, value for value across all 12 features, to another eight rows (indices 88–95), a fixed 32-row offset apart (Table 3). Within this duplicated block, every individual row also has

an exact-duplicate partner at a fixed 4-row offset from its counterpart, consistent with the block representing two participants complete four-event records (4 rows each) copied as a unit. Two independent participants producing continuous-valued statistics that agree to nine decimal places across twelve features is not a plausible coincidence; the offset is also identical for every one of the eight duplicate pairs, which is not consistent with eight independent chance matches. The more plausible explanation is a data-assembly error in which a two-participant block was inserted twice during file construction. On this reading, the delivered file's true unique content is 200 rows from 50 participants, not the documented 208 rows from 52 participants.

Table 3. Duplicated Row Block and Implied True Sample Size

Row indices (block A)	Row indices (block B)	Offset	Rows involved
56–63 (8 rows)	88–95 (8 rows)	32 rows	16 of 208 (7.7%)
Within-pair offset	—	4 rows (constant across all 8 pairs)	—
Implied true unique rows	200 (208 − 8 duplicate contents)	—	—
Implied true participants	50 (200 / 4 events)	—	vs. documented 52

3.4 Signal Validity: Strong Separability, Partly Circular

A one-way ANOVA of each of the 12 features against the four-class label finds every feature discriminating strongly (all $p < 0.001$; the weakest, maximum, still reaches $p < 0.001$). This separability is not uniformly a positive sign, however: the percentage feature is exactly 100.000000 with zero variance across all 52 Heel-Off rows, meaning it does not vary at all within that class. A feature that is constant by construction within a class does not provide discriminating information in the way a genuinely measured, independently varying feature would; it is more consistent with a windowing convention in which the Heel-Off feature-extraction window is defined to end at 100% of stance, than with an independently observed biomechanical quantity.

The per-class percentage averages are also inconsistent with standard stance-phase timing. Heel Strike (initial contact) marks the start of stance and is conventionally reported at approximately 0–3% of the gait cycle [2], [14]; in this dataset, HS-labeled rows average 32.7% (Table 4). Heel Off conventionally marks the start of terminal stance, reported at roughly 31–50% of the gait cycle (≈ 50–82% of stance phase, given stance constitutes roughly 60–62% of the full cycle [2]); in this dataset, HO-labeled rows are fixed at exactly 100%, the point conventionally associated with toe-off and the end of stance rather than heel-off. Neither the abstract nor the Instructions document what reference point or normalization convention percentage uses, so this discrepancy cannot be resolved from the released materials; it is reported here as a documented inconsistency with the standard clinical convention for these event names, not as a claim about what the authors' underlying measurement actually captured.

Table 4. Delivered percentage Feature vs. Standard Stance-Phase Timing

Event	This dataset's percentage (mean, std)	Standard convention [2], [14]
HS (Heel Strike)	32.7% (±3.0)	≈ 0–3% of gait cycle (initial contact, by definition)
LR (Loading Response)	48.7% (±6.0)	≈ 3–12% of gait cycle
MS (Mid-Stance)	68.9% (±9.6)	≈ 12–31% of gait cycle
HO (Heel Off)	100.0% (±0.0)	≈ 31–50% of gait cycle (start of terminal stance)

3.5 Limitations and Future Work

This audit could not access the dataset's linked supplementary PDF, which may document the windowing convention behind percentage and duration, the source of the COV and mode computations, or a GRF/EMG attribution scheme not present in the delivered CSV; obtaining and reviewing that document is a direct next step. The audit also cannot determine, from the delivered file alone, whether the duplicated eight-row block reflects a file-assembly error or a deliberate (but undocumented) inclusion of repeated trials from the same two participants; distinguishing these possibilities would require the original per-participant records. Finally, because participant identity is not encoded in the file beyond the row-block structure inferred here, this audit cannot rule out additional, less exact forms of participant-level dependence between the four rows contributed by each of the 52 (or 50 unique) participants.

Once the issues above are corrected, we recommend that any classifier benchmarked on this dataset use participant-grouped, stratified cross-validation rather than a random row-level split, since rows from the same participant are not independent [17]; that reported performance be compared against common tabular-classification baselines such as gradient-boosted trees [18]; and that per-class precision, recall, and F1 be reported alongside accuracy, since aggregate accuracy can mask class-level errors even when, as here, the classes are nominally balanced [19], [20].

4. Conclusions

This paper applied a six-stage, three-layer audit — provenance, structural and physical plausibility, and signal validity — to a publicly documented gait-event feature dataset. Row count, feature count, and participant count all matched the hosting page's Instructions, but two of the twelve documented features are exact duplicates under misleading names: the column labeled COV is variance, not coefficient of variation, and the column labeled mode is minimum, not an estimate of central tendency. An eight-row block is duplicated verbatim at a fixed offset, indicating the true unique sample is 50 participants rather than the documented 52. The file provides no way to attribute any feature to GRF versus EMG despite both being named as sources. And while every feature discriminates strongly between the four gait-event classes, the percentage feature is exactly constant within the Heel-Off class, consistent with a windowing artifact rather than an independent measurement, and its per-class values do not align with standard stance-phase timing reported in the clinical gait literature. None of these findings required access to any resource beyond the delivered file, its hosting page, and established reference values from the gait-analysis literature — underscoring that formula verification and literature cross-checks, not just documentation reconciliation, belong in a routine pre-reuse audit of engineered-feature datasets.

Acknowledgments

The authors wish to thank the editorial team of the Journal of Systems Engineering and Information Technology (JOSEIT) for their guidance during manuscript preparation.

References

[1] D. Alvarado-Rivera, P. A. Niño-Suárez, and L. G. Corona-Ramírez, "Robust Gait Event Detection Using GRF and EMG Signals," IEEE DataPort, 2025, doi: 10.21227/0mch-sc46.

[2] J. Perry and J. M. Burnfield, Gait Analysis: Normal and Pathological Function, 2nd ed. Thorofare, NJ: SLACK Incorporated, 2010.

[3] J. A. Zeni Jr., J. G. Richards, and J. S. Higginson, "Two Simple Methods for Determining Gait Events during Treadmill and Overground Walking Using Kinematic Data," Gait & Posture, vol. 27, no. 4, pp. 710–714, 2008, doi: 10.1016/j.gaitpost.2007.07.007.

[4] H. J. Hermens, B. Freriks, C. Disselhorst-Klug, and G. Rau, "Development of Recommendations for SEMG Sensors and Sensor Placement Procedures," Journal of Electromyography and Kinesiology, vol. 10, no. 5, pp. 361–374, 2000, doi: 10.1016/S1050-6411(00)00027-4.

[5] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[6] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[7] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in Data Mining: Formulation, Detection, and Avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, article 15, 2012, doi: 10.1145/2382577.2382579.

[8] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," Patterns, vol. 4, no. 9, article 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[9] M. Baker, "1,500 Scientists Lift the Lid on Reproducibility," Nature, vol. 533, no. 7604, pp. 452–454, 2016, doi: 10.1038/533452a.

[10] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis, "A Manifesto for Reproducible Science," Nature Human Behaviour, vol. 1, article 0021, 2017, doi: 10.1038/s41562-016-0021.

[11] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," PLoS Medicine, vol. 2, no. 8, e124, 2005, doi: 10.1371/journal.pmed.0020124.

[12] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model Cards for Model Reporting," in Proc. Conf. on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, USA, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[13] J. M. Hausdorff, "Gait Variability: Methods, Modeling and Meaning," Journal of NeuroEngineering and Rehabilitation, vol. 2, article 19, 2005, doi: 10.1186/1743-0003-2-19.

[14] D. A. Winter, Biomechanics and Motor Control of Human Movement, 4th ed. Hoboken, NJ: John Wiley & Sons, 2009.

[15] IEEE, "IEEE Editorial Style Manual for Authors," IEEE Editorial, 2019.

[16] O. Tur, A. Tur, V. Shabunina, and E. Chernaia, “The IEEE Style: Peculiarities of the Format and Application Prospects,” in Proc. 25th IEEE Int’l Conf. on Problems of Automated Electric Drive. Theory and Practice (PAEP), 2020, doi: 10.1109/PAEP49887.2020.9240790.

[17] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, vol. 2, 1995, pp. 1137–1143.

[18] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[19] H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[20] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” PLOS ONE, vol. 10, no. 3, e0118432, 2015, doi: 10.1371/journal.pone.0118432.