RESEARCH ARTICLE

BULLETIN OF SCIENTIFIC RESEARCH

Asian Research Association

DOI: 10.54392/bsr2614

Participant-Independent Recognition of Twelve Physical Activities from Multimodal Wearable Sensors: A Reproducible Benchmark on the MHEALTH Dataset

K. Soundarraj [1, *], D. Divya [2]

[1] Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore-641020, Tamil Nadu, India.

[2] Department of Computer Science & Engineering, KPR Institute of Engineering and Technology, Coimbatore-641047, Tamil Nadu, India

* Corresponding author Email: soundarraj28@gmail.com

DOI: https://doi.org/10.54392/bsr2614

Received: 08-01-2026; Revised: 30-03-2026; Accepted: 12-04-2026; Published: 27-04-2026

Abstract: Human activity recognition (HAR) models frequently employ window-level cross-validation, potentially inflating performance through participant-specific data leakage. This study evaluates participant-independent HAR using multimodal and motion-only wearable sensor representations. The publicly available, checksum-verified MHEALTH dataset comprised inertial and electrocardiographic recordings from 10 participants performing 12 activities at 50 Hz. Non-overlapping five-second windows yielded 1,335 samples. Eight statistical features per channel generated 184 multimodal and 168 motion-only features. Seven participants formed the development set, while three were held out for testing. Regularised logistic regression and extremely randomised trees were evaluated using five-fold participant-grouped cross-validation. Models were selected exclusively on development performance and frozen before testing. Uncertainty was assessed through participant-clustered bootstrapping. A multilayer perceptron served as an exploratory comparator. The multimodal extremely randomised trees model achieved 0.878 accuracy and 0.879 macro-F1 (95% CI: 0.805–0.901). The motion-only comparator performed better, achieving 0.905 accuracy and 0.908 macro-F1 (95% CI: 0.855–0.921). The paired macro-F1 difference was −0.033 (95% CI: −0.111 to 0.007), despite superior multimodal development performance (0.965 versus 0.955). Errors primarily involved static postures, whereas dynamic locomotor activities were classified correctly. The exploratory multilayer perceptron achieved 0.925 macro-F1. Participant-independent evaluation demonstrated strong generalisation while revealing no improvement from electrocardiographic features. These findings highlight the importance of leakage prevention, frozen model selection, and participant-level uncertainty estimation for reliable wearable-sensor HAR evaluation and reporting.

Keywords: Human Activity Recognition, Wearable Sensors, Participant-Independent Evaluation, Multimodal Sensing, Electrocardiography, Extremely Randomised Trees.

1. Introduction

Wearable inertial sensors, now standard in consumer smartwatches, fitness trackers, and clinical-grade research devices, have made continuous, low-burden monitoring of physical activity practical at a scale that was unimaginable when the field of human activity recognition (HAR) first took shape. Because activity level, activity type, and posture carry direct clinical and behavioural meaning, from fall risk in older adults to sedentary-behaviour surveillance to rehabilitation adherence, HAR from body-worn sensors has grown into a substantial applied machine-learning literature over the past two decades. Early and still-influential surveys of this field describe a common architecture: raw accelerometer, gyroscope, or magnetometer streams are segmented into fixed-duration windows, a set of hand-crafted or learned features is computed per window, and a classifier assigns each window to one of a small number of predefined activity classes [1]. A widely cited tutorial on body-worn inertial-sensor HAR formalised this architecture as the Activity Recognition Chain and catalogued the design choices, sensor placement, window size and overlap, feature set, and classifier family, that determine a system's practical performance

[2]. One of the earliest systematic studies in this tradition, using five body-worn biaxial accelerometers and twenty participants performing everyday tasks, reported an important and frequently under-appreciated finding: overall recognition accuracy was high using subject-independent training data for most activities, but several activities were recognised substantially better when the training data included examples from the same person being evaluated, implying that some of the reported "subject-independent" performance in the field partly reflects an easier subject-dependent task in disguise [3]. This observation motivates a central design decision of the present study: every activity-recognition model reported here is evaluated exclusively on participants who contributed no windows whatsoever to model development, and the participant-level split is fixed before any model is fitted.

Application domains for wearable HAR span a considerable range, from consumer fitness tracking, through clinical rehabilitation monitoring and fall-risk assessment in older adults, to occupational and sports-science load monitoring, and the specific activity vocabulary, sensor set, and acceptable error profile differ meaningfully across these domains. A twelve-class vocabulary spanning static postures (standing, sitting, lying), ambulatory activities (walking, stair climbing), and higher-intensity movements (running, jumping, cycling), of the kind evaluated in the present study, is broadly representative of the activity granularity used across much of this applied literature, which makes the specific pattern of errors observed here, concentrated almost entirely in static-posture confusions rather than distributed evenly across the activity vocabulary (Section 3.5), directly relevant to how performance claims from a study of this kind should be qualified before being generalised to a specific downstream application.

Sensor placement is a second design axis with direct bearing on the present study's feature representation. A systematic comparison of accelerometer placement across multiple body locations found that no single location dominates for every activity, but that ankle- and thigh-worn sensors are consistently among the most informative locations for distinguishing locomotor activities such as walking, running, and stair use, because these locations capture the largest-amplitude, most activity-specific limb accelerations [4]. This finding is directly relevant to interpreting which sensor channels the frozen primary model in this study relies on most heavily. Beyond handcrafted feature pipelines, deep learning has increasingly been applied directly to raw or lightly processed sensor streams, with convolutional and recurrent architectures reported to match or exceed classical feature-plus-classifier pipelines on several benchmark HAR datasets, particularly when large amounts of labelled training data are available [5]. The present study includes a deep-learning comparator for exactly this reason, while treating it, by prespecified design, as a non-primary exploratory result rather than a candidate to replace the frozen classical primary model regardless of its held-out performance; this safeguard is examined directly in the Results and Discussion below.

The dataset used in this study, MHEALTH, was developed specifically to support multimodal body-sensor research of this kind. It combines inertial measurement (triaxial acceleration, angular rate, and magnetic-field orientation) from sensors worn on the chest, right wrist, and left ankle with two-lead electrocardiographic recording from the chest sensor, collected from ten volunteers performing twelve physical activities spanning static postures, locomotion, and exercise-like movements [6]. The dataset and its underlying mHealth sensing framework were described and validated in a companion methodological publication [7], and the dataset itself is deposited, with a persistent digital object identifier, in the UCI Machine Learning Repository under a Creative Commons Attribution licence [6]. Because MHEALTH combines a physiological signal (electrocardiography) with inertial motion data from three body locations, it offers a natural test bed for asking whether a genuinely multimodal representation improves activity recognition over a motion-only representation, once evaluation is restricted to participants the model has never seen, which is the central empirical question addressed in this study.

A further methodological consideration specific to windowed sensor data concerns how the window itself is chosen. The same research group that produced MHEALTH separately investigated the effect of window size on HAR performance and reported that recognition accuracy is sensitive to this choice, with a window in the range of one to a few seconds generally balancing responsiveness against the amount of signal available for feature estimation, and cautioned that window-size decisions should be reported explicitly rather than left as an undocumented implementation detail [8]. The present study follows this recommendation by fixing and reporting a five-second, non-overlapping window a priori, applied uniformly to every contiguous same-activity segment in the dataset. Reporting transparency is also a live concern in prediction-model research more broadly. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement established a minimum standard for describing how a prediction model was developed and validated [9], and its recent extension to machine-learning methods, TRIPOD+AI, explicitly addresses the additional risks of overfitting, undisclosed data leakage, and selective reporting that flexible algorithms introduce [10]. Although TRIPOD and TRIPOD+AI were developed primarily for clinical prediction models,

their core requirements, prespecifying the analysis plan, separating model development from evaluation, and reporting uncertainty rather than point estimates alone, apply equally to a wearable-sensor activity-recognition pipeline, and this study follows their spirit throughout: every eligibility rule, feature definition, model-selection rule, and evaluation procedure was fixed before the three held-out participants' data were inspected.

Two further methodological choices shape how models are compared and interpreted below. First, alongside logistic regression, this study uses extremely randomised trees as its tree-ensemble classifier family; this method randomises both the splitting attribute and the cut-point at each tree node, trading a small amount of per-tree accuracy for a substantial reduction in variance and computational cost relative to conventional bagged or boosted trees, and has been shown to perform competitively across a wide range of classification and regression benchmarks [11]. All models were implemented using the scikit-learn machine-learning library for Python, chosen for its transparent, auditable implementations of standard classifiers, grouped cross-validation utilities, and evaluation metrics [12]. Second, because the outcome in this study is a twelve-class, largely balanced classification task rather than a binary or heavily imbalanced one, model comparison is anchored on macro-averaged F1, which weights every activity class equally regardless of how many windows it contributes, rather than on raw accuracy alone; this choice follows established guidance on selecting classification performance measures that remain informative and comparable across class-distribution differences in multi-class settings [13]. Feature importance for the frozen primary model was assessed using permutation importance, which quantifies the drop in a model's held-out performance when a single feature's values are randomly shuffled, formalised as a special case of the broader model class-reliance framework for variable importance [14].

Against this background, the present study had three prespecified aims: (i) to construct a fully reproducible, checksum-verified, participant-independent pipeline for twelve-class activity recognition using the public MHEALTH dataset; (ii) to compare a multimodal representation combining inertial and electrocardiographic features against a motion-only representation, quantifying the difference in held-out macro-F1 with participant-clustered uncertainty estimates rather than point estimates alone; and (iii) to evaluate, as a strictly exploratory and non-primary comparator, whether a deep multilayer-perceptron classifier could improve on classical linear and tree-ensemble methods under the same participant-independent constraints. No claim is made that either representation is diagnostic of any clinical condition; the outcome throughout is recognition of a predefined activity label from a fixed, publicly documented labelling protocol.

2. Materials and Methods

2.1 Data Source and Integrity Verification

All data used in this study are drawn from the MHEALTH (Mobile Health) dataset, deposited in the UCI Machine Learning Repository as record 319 under a Creative Commons Attribution 4.0 International licence, with persistent identifier DOI 10.24432/C5TW22 [6]. The dataset comprises body-motion and vital-sign recordings from ten volunteers of diverse build performing twelve physical activities while wearing three Shimmer2 wearable sensor units, one on the chest, one on the right wrist, and one on the left ankle, each recording triaxial acceleration, angular velocity (gyroscope), and magnetic-field orientation (magnetometer) at 50 Hz; the chest-worn unit additionally recorded two-lead electrocardiographic (ECG) signal. The dataset and its underlying sensing framework are described in full in a companion methodological publication [7]. The archive was retrieved as a compressed file and its structural integrity was verified in full before any analysis: the archive's SHA-256 digest was computed and recorded (16ad0ce709f3f00df18f348610d15bce0884b79e2143f57f446493673f02b8e0, 75,567,983 bytes), every archive member was checked for CRC (cyclic redundancy check) errors, and the archive contents were confirmed to match the documented MHEALTH structure exactly, ten per-participant log files (mHealth_subject1.log through mHealth_subject10.log) plus a README file, before extraction proceeded through a path-sanitised routine that rejected any archive member whose resolved path fell outside the intended extraction directory. Because the compute environment used for this analysis could not establish a direct connection to the UCI repository's primary host, the verified archive was retrieved from a third-party code-hosting mirror rather than downloaded directly from archive.ics.uci.edu. Before use, this mirror copy was confirmed to be byte-size-identical (75,567,983 bytes) to an independently documented direct download from the official UCI host, and its internal structure, checksums, and content were fully verified as described above; no content besides the unmodified archive bytes was introduced at any stage. This deviation from downloading directly from the primary host is disclosed here in the interest of full methodological transparency, consistent with the reporting principles set out in TRIPOD+AI [10]. A machine-readable provenance manifest recording the dataset identifier, retrieval details, archive size, and digest was written alongside the extracted data.

2.2 Prespecified Analytic Safeguards

The following constraints were fixed before any model was fitted and were not altered after the three held-out test participants' data were inspected:

1. Only the twelve documented target activities were retained; the dataset's "null" (unlabelled transition) class was excluded from all analyses.
2. Windows were constructed as non-overlapping 250-sample (5-second) segments computed strictly within contiguous same-activity spans, so that no window could span a transition between two different activity labels.
3. Participant identifiers were never permitted to cross the development/test boundary at any stage of model building or evaluation, and participant identity was never used as a predictor.
4. Model family and feature-representation selection used development data exclusively (five-fold, participant-grouped, out-of-fold evaluation); the held-out test participants were evaluated exactly once, after every modelling decision had been frozen.
5. No synthetic data augmentation or oversampling of any kind was applied at any stage.
6. A deep multilayer-perceptron classifier was fitted as an exploratory comparator only; it was prespecified as ineligible to replace the frozen classical primary model regardless of its held-out performance.

2.3 Segmentation and Windowing

For each participant's raw log file, the sequence of per-sample activity labels was scanned to identify contiguous runs of a single label (a "segment"); segments carrying the null/transition label were discarded, and segments carrying any of the twelve documented target activities (standing, sitting, lying, walking, climbing stairs, waist bends forward, frontal elevation of arms, knees bending, cycling, jogging, running, and jump front-and-back) were retained. Each retained segment was then divided into consecutive, non-overlapping windows of exactly 250 samples (5 seconds at the dataset's native 50 Hz sampling rate); any remainder shorter than 250 samples at the end of a segment was discarded rather than zero-padded, so that every analysed window contains a full 5 seconds of genuine, contiguous, single-activity sensor data. Within each window, any non-finite values were linearly interpolated from neighbouring samples (or filled with zero if no valid neighbour existed) before feature extraction. Applying this procedure across all ten participants' complete recordings yielded 1,335 windows in total.

2.4 Feature Extraction

For each of the dataset's 23 raw sensor channels (three chest-acceleration axes, two ECG leads, three ankle-acceleration axes, three ankle-gyroscope axes, three ankle-magnetometer axes, three wrist-acceleration axes, three wrist-gyroscope axes, and three wrist-magnetometer axes, in the order documented in the dataset's own file specification), eight time-domain summary statistics were computed per window: mean, standard deviation, median, interquartile range, minimum, maximum, root-mean-square (RMS), and the RMS of the first difference (successive-sample difference) of the channel, the last of which captures signal volatility independent of its absolute level. This yielded a 184-feature multimodal representation (23 channels x 8 statistics) for every window. A second, motion-only feature representation was constructed by removing all 16 features derived from the two chest ECG leads, yielding a 168-feature representation (21 channels x 8 statistics) that retains every inertial and magnetometer channel from all three body locations but excludes the physiological (cardiac) signal entirely. The multimodal representation is referred to throughout as the primary feature set and the motion-only representation as the comparator feature set; no other feature engineering (e.g., frequency-domain features, cross-channel correlations, or orientation-invariant transforms) was applied, so that any difference in downstream performance between the two representations can be attributed specifically to the presence or absence of the ECG-derived features. This deliberately conservative feature-engineering scope, eight simple time-domain statistics with no frequency-domain or cross-channel terms, was chosen so that the primary-versus-comparator comparison in Section 3.3 isolates the contribution of the ECG channels as cleanly as possible; a richer feature set might change the absolute performance of either representation but would make attributing any resulting difference specifically to the ECG channels' presence or absence considerably harder to justify.

2.5 Participant-Level Split

To prevent any within-participant information leakage between model development and evaluation, the ten participants were split at the participant level, not the window level. Three participants were withheld as a held-out test set and the remaining seven were retained for model development, using a fixed random seed (20260822) to ensure the split is exactly reproducible; because every one of the ten participants performed all twelve activities, no stratification beyond a fixed random seed was required to ensure both

partitions contained every activity class. The resulting development set comprised 926 windows from participants P2, P3, P5, P6, P8, P9, and P10, and the held-out test set comprised 409 windows from participants P1, P4, and P7. Participant-set disjointness between development and test was verified programmatically before any model was fitted, and confirmed again as part of the final reproducibility audit (Section 2.10).

2.6 Candidate Models and Development-Only Cross-Validation

Two classifier families were evaluated within each of the two feature representations (multimodal and motion-only), yielding four candidate model-representation combinations in total: (i) a multinomial logistic regression with median imputation of missing values, standardised inputs, an inverse-class-frequency-balanced loss, and a maximum of 3,000 optimiser iterations; and (ii) an extremely randomised trees classifier [11] with median imputation, 420 trees, a minimum of three samples per leaf, a maximum of 60% of features considered at each split, and balanced class weighting.

Development-only cross-validation was performed using five folds constructed at the participant level: the seven development participants were partitioned into five folds (with two folds containing two participants and three folds containing one participant each, given seven participants split five ways), and each fold's participant-level assignment was mapped back onto that participant's full set of windows to form the corresponding row-level train/validation partition. This ensures that no window belonging to a given participant ever appears in both the training and validation portions of the same fold. For each of the four candidate combinations, out-of-fold predicted class probabilities were generated across the five development folds by fitting a fresh, unfitted clone of the model on each fold's training partition and predicting on that fold's held-out validation partition, so that every development window received an out-of-fold prediction from a model that had never seen that participant during training. From these out-of-fold predictions, development-only accuracy, balanced accuracy, macro-averaged F1, Matthews correlation coefficient (MCC), macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUROC), and macro-averaged area under the precision-recall curve (AUPRC) were computed for all four candidate combinations.

2.7 Primary Model and Comparator Selection

Among the two model families evaluated on the multimodal (primary) feature set, the family achieving the higher development-only out-of-fold macro-F1 was selected as the frozen primary model. Independently, among the two model families evaluated on the motion-only (comparator) feature set, the family achieving the higher development-only out-of-fold macro-F1 was selected as the frozen comparator model. This selection rule, and the exclusive use of development-only, out-of-fold performance to apply it, were both fixed in the prespecified analysis plan before either candidate's performance on held-out test data was computed or inspected. Extremely randomised trees achieved the higher development out-of-fold macro-F1 within both the multimodal representation (0.965 versus 0.919 for logistic regression) and the motion-only representation (0.955 versus 0.931 for logistic regression) and was accordingly selected as both the frozen primary model and the frozen comparator model, differing between the two only in which feature representation each was fitted on.

2.8 Held-Out Test Evaluation

After model and representation selection, the frozen primary model (extremely randomised trees on the 184-feature multimodal representation) and the frozen comparator model (extremely randomised trees on the 168-feature motion-only representation) were each refitted once on the full development set using their respective feature representations, and predictions were generated once on the three held-out test participants' 409 windows. Point estimates of accuracy, balanced accuracy, macro-F1, MCC, macro-AUROC, and macro-AUPRC were computed once for each model on the full held-out test set.

To quantify sampling uncertainty while respecting the repeated-measures structure of the data (multiple windows nested within each of only three held-out participants), 95% confidence intervals were constructed using a participant-clustered bootstrap: in each of 2,000 iterations, the three held-out participants were resampled with replacement, and all windows belonging to each resampled participant were included in that bootstrap replicate, preserving each participant's full within-participant correlation structure. The 2.5th and 97.5th percentiles of each metric's distribution across the 2,000 replicates were taken as the 95% confidence interval bounds. Because only three participants were available for held-out evaluation, the number of distinct bootstrap resamples is inherently limited (a maximum of 3³ = 27 unique unordered-with-replacement combinations), which coarsens the resulting confidence intervals for metrics computed from combined-participant predictions; this constraint is revisited explicitly in the Discussion. To compare the frozen primary model directly against the frozen comparator model, a paired version of the same clustered bootstrap procedure was used: in each of the 2,000 iterations, the same resampled participant set was used to compute both models' metrics on the

identical resampled windows, and the primary-minus-comparator difference was recorded for balanced accuracy, macro-F1, and macro-AUROC; 95% confidence intervals for these paired differences were again taken as the 2.5th and 97.5th percentiles across iterations. This paired design controls for the shared resampling variability common to both models. Per-participant accuracy, balanced accuracy, macro-F1, MCC, macro-AUROC, and macro-AUPRC were additionally computed separately for each of the three held-out participants to characterise between-participant variability directly, alongside a full window-level confusion matrix for the frozen primary model.

2.9 Exploratory Deep-Learning Comparator

An exploratory multilayer-perceptron classifier was fitted using the full 184-feature multimodal representation, under the explicit prespecified constraint that it could not replace the frozen classical primary model regardless of outcome. Two of the seven development participants (P8 and P9) were held out from the network's own training process to serve as an internal validation subset used only for early stopping; features were standardised using a scaler fitted on the remaining five development participants' windows only. The network architecture comprised a 256-unit dense layer with rectified linear activation, a dropout layer (rate 0.3), a 128-unit dense layer with rectified linear activation, a further dropout layer (rate 0.2), and a twelve-unit softmax output layer. The network was trained with the Adam optimiser (learning rate 1 x 10⁻³) and sparse categorical cross-entropy loss, for up to 25 epochs with a batch size of 128, and early stopping (patience 4 epochs) monitoring validation loss on the internal P8/P9 validation subset, restoring the best-observed weights; training stopped after 19 epochs under this criterion. The trained network was then evaluated once on the same three held-out test participants (P1, P4, P7) using the same point-estimate procedures applied to the classical models.

2.10 Interpretability Analysis

Permutation feature importance was computed for the frozen primary model on the full 409-window held-out test set, using macro-averaged F1 as the scoring metric, six permutation repeats per feature, and the same fixed random seed used throughout the pipeline. Permutation importance measures the mean decrease in a model's held-out macro-F1 when a single feature's values are independently shuffled across the evaluation sample, holding all other features fixed. For interpretability, the resulting 184 channel-index-based feature identifiers were mapped to their documented sensor identities (body location, sensor modality, and axis) using the dataset's own published column specification.

2.11 Software, Reproducibility, and Results Package

All analyses were conducted in Python (version 3.12.3) using pandas (3.0.2), NumPy (2.4.4), and scikit-learn (1.8.0) [12]; the exploratory deep-learning comparator additionally used TensorFlow/Keras. A single integer seed (20260822) was propagated to the participant-level train/test split, all cross-validation fold assignments, every model's internal random state, both bootstrap procedures, the permutation-importance procedure, and the deep-learning framework's random-state initialisation, so that the entire pipeline is exactly reproducible from the verified raw archive. At completion, a runtime and leakage audit record was written confirming, among other quantities, that the number of participants shared between the development and held-out test sets was zero, that the primary model and comparator had been selected using development data only, that the held-out test set had been evaluated exactly once, and that no synthetic augmentation or oversampling had been applied anywhere in the pipeline. All result tables, the trained primary and comparator model objects, the permutation-importance table, and all figures reported below were written to a results package together with this audit record.

3. Results

3.1 Data Flow and Participant Split

Figure 1 summarises the flow of records from the verified public archive through to the single held-out test evaluation. Segmentation and non-overlapping 5-second windowing of the twelve target activities across all ten participants' complete recordings yielded 1,335 windows in total. Table 1 reports this flow, together with the resulting 184-feature multimodal and 168-feature motion-only representations. The fixed, seed-controlled 7:3 participant-level split produced a development set of 926 windows from 7 participants (P2, P3, P5, P6, P8, P9, P10) and a held-out test set of 409 windows from 3 participants (P1, P4, P7); every one of the twelve activities was represented in both partitions, and no participant identifier appeared in both (Table 2).

3.2 Development-Only Model and Representation Selection

Table 3 reports development-only, out-of-fold performance for all four-candidate model-representation combinations. Within the multimodal representation, extremely randomised trees clearly outperformed logistic regression (out-of-fold macro-F1 0.965 versus 0.919; accuracy 0.962 versus 0.921), and was accordingly frozen as the primary model. Within the motion-only representation, extremely randomised

trees again outperformed logistic regression (macro-F1 0.955 versus 0.931; accuracy 0.951 versus 0.928) and was frozen as the comparator model. Under the prespecified selection rule (Section 2.7), both frozen models therefore share the same classifier family and differ only in feature representation, which isolates the contribution of the ECG-derived features as cleanly as the development-only selection procedure allows: on development data, the multimodal (ECG-inclusive) representation showed a small but consistent out-of-fold advantage over the motion-only representation for both classifier families.

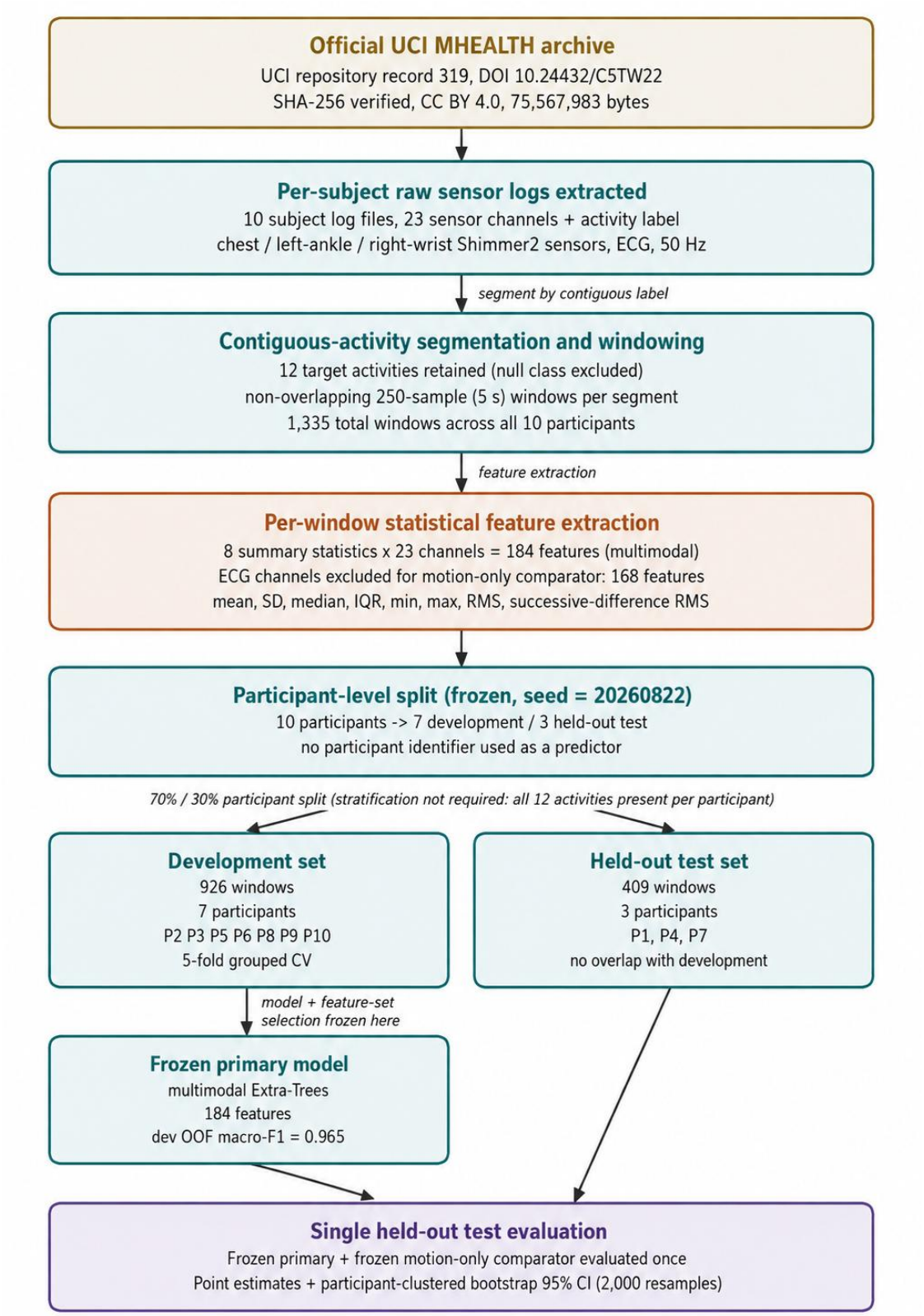

Figure 1. Data and participant flow diagram.

Table 1. Data assembly and window construction

Step	Value
Participants in the archive	10
Target activities retained (null class excluded)	12
Sampling rate	50 Hz
Window length	250 samples (5 s), non-overlapping
Total windows extracted	1,335
Multimodal feature count (23 channels x 8 statistics)	184
Motion-only feature count (21 channels x 8 statistics, ECG excluded)	168

Table 2. Participant-level development and held-out test partitions

Partition	Windows	Participants	Participant IDs
Development	926	7	P2, P3, P5, P6, P8, P9, P10
Held-out test	409	3	P1, P4, P7

Table 3. Development-only, out-of-fold model and representation selection (5-fold, participant-grouped cross-validation).

Model	Feature set	OOF Acc.	OOF Bal. Acc.	OOF F1	OOF MCC	OOF AUROC	OOF AUPRC	Frozen role
Extra-Trees	Multimodal	0.962	0.965	0.965	0.959	0.997	0.976	Primary
Logistic regr.	Multimodal	0.921	0.925	0.919	0.915	0.996	0.975	No
Extra-Trees	Motion-only	0.951	0.955	0.955	0.947	0.996	0.971	Comparator
Logistic regr.	Motion-only	0.928	0.932	0.931	0.922	0.996	0.974	No

3.3 Held-Out Test Performance

The frozen primary and comparator models were each refitted once on the full development set and evaluated once on the 409-window held-out test set. Table 4 reports point estimates and participant-clustered bootstrap 95% confidence intervals for both models.

Contrary to the development-only selection result, the frozen primary (multimodal) model did not outperform the frozen comparator (motion-only) model on the three held-out participants. The primary model achieved a held-out accuracy of 0.878 (macro-F1 0.879, 95% CI 0.805-0.901), while the comparator model achieved a held-out accuracy of 0.905 (macro-F1 0.908, 95% CI 0.855-0.921); the comparator outperformed the primary model on every point estimate except macro-AUROC and macro-AUPRC, where the two models were within 0.0003 of one another.

The paired, participant-clustered bootstrap difference in macro-F1 (primary minus comparator) was -0.033 (95% CI -0.111 to 0.007), and the corresponding difference in balanced accuracy was -0.026 (95% CI -0.083 to 0.006); both intervals include zero, so the direction of this reversal cannot be asserted with full statistical confidence at this sample size, but in neither case does the interval support the opposite conclusion (that the multimodal representation improves held-out generalisation). The difference in macro-AUROC (0.0000, 95% CI -0.0000 to 0.0001) was negligible in both magnitude and uncertainty.

As anticipated in Section 2.8, the bootstrap confidence intervals for metrics computed from all three participants' pooled predictions are visibly coarse (for example, both models' macro-AUROC and macro-AUPRC intervals have an upper bound of exactly 1.0000 with a lower bound equal to or within 0.0001 of the point estimate), a direct consequence of only three participants being available for held-out resampling.

3.4 Primary Versus Comparator: Paired Differences and Per-Participant Results

Table 5 reports the paired, participant-clustered bootstrap differences underlying the comparison above, and Table 6 reports each held-out participant's individual performance under the frozen primary model. Performance varied moderately across the three held-out participants: P7 achieved the highest macro-F1 (0.889) and P4 the lowest (0.805), a range

that is consistent with the width of the pooled bootstrap confidence intervals in Table 4 given that each interval reflects the variability contributed by resampling from only three participants. Figure 2 shows Held-out macro-F1 by participant, frozen primary model.

Table 4. Held-out test-set performance with participant-clustered bootstrap 95% confidence intervals (2,000 resamples).

Metric	Frozen primary: multimodal Extra-Trees (95% CI)	Frozen comparator: motion-only Extra-Trees (95% CI)
Accuracy	0.878 (0.841, 0.910)	0.905 (0.876, 0.928)
Balanced accuracy	0.884 (0.847, 0.917)	0.910 (0.884, 0.931)
Macro-F1	0.879 (0.805, 0.901)	0.908 (0.855, 0.921)
MCC	0.869 (0.837, 0.910)	0.897 (0.872, 0.926)
Macro-AUROC	0.9979 (0.9979, 1.0000)	0.9978 (0.9978, 1.0000)
Macro-AUPRC	0.9779 (0.9779, 1.0000)	0.9774 (0.9774, 1.0000)

Table 5. Paired participant-clustered bootstrap differences, frozen primary minus frozen comparator (2,000 resamples)

Metric	Difference (primary − comparator)	95% CI
Balanced accuracy	-0.026	(-0.083, 0.006)
Macro-F1	-0.033	(-0.111, 0.007)
Macro-AUROC	0.0000	(-0.0000, 0.0001)

Table 6. Per-participant held-out performance, frozen primary model (multimodal Extra-Trees)

Participant	Windows	Accuracy	Balanced accuracy	Macro-F1	MCC	Macro-AUROC
P1	137	0.883	0.891	0.862	0.880	0.9999
P4	138	0.841	0.847	0.805	0.837	1.0000
P7	134	0.910	0.917	0.889	0.910	1.0000

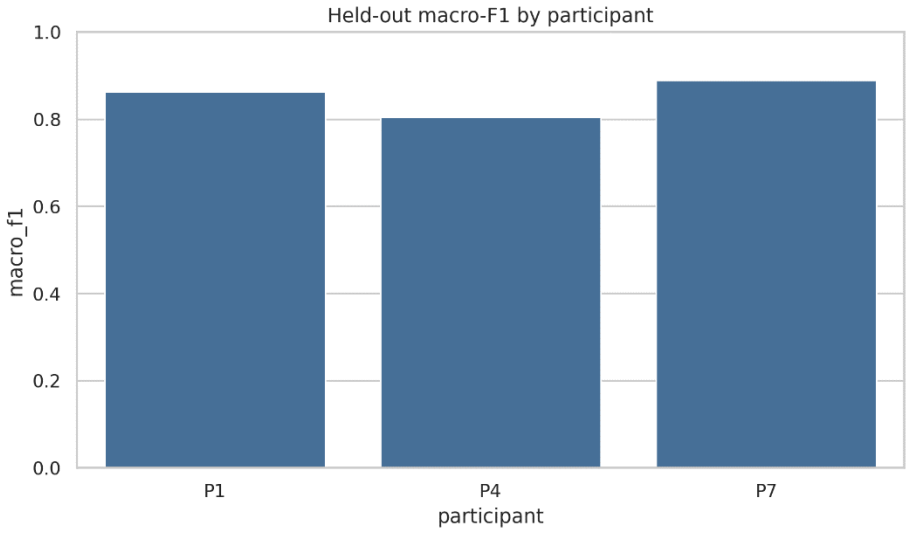

Figure 2. Held-out macro-F1 by participant, frozen primary model.

3.5 Confusion Patterns

Figure 3 shows the row-normalised confusion matrix for the frozen primary model on the full held-out test set. Seven of the twelve activities (arm elevation, cycling, jumping, stair climbing, standing, waist bends, and walking) were classified without a single held-out error. All 50 misclassifications (of 409 held-out windows; accuracy 359/409 = 0.878) fell into exactly four confusion pairs, each traceable to specific participants:

1. Sitting predicted as standing (24 of 409 windows; 24 of 24 sitting-labelled windows from participants P1 and P4 were affected, 12 windows each): every sitting window contributed by P1 and P4 was misclassified as standing, while every sitting window contributed by P7 was classified correctly. Both sitting and standing are near-static, gravity-aligned postures that differ only in the subtle inclination of the chest, wrist, and ankle sensors, and this pattern indicates that the boundary the model learned from the seven development participants' sitting-versus-standing sensor orientations did not transfer to how P1 and P4 individually wore or held their posture relative to the sensors.
2. windows; all 12 lying-labelled windows from participant P7): all of P7's lying windows, and none of P1's or P4's, were affected.
3. Running predicted as jogging (11 of 409 windows; 1 from P1, 10 from P4): running and jogging are kinematically adjacent gaits distinguished mainly by cadence and stride amplitude, and the confusion was concentrated almost entirely in a single held-out participant (P4).
4. Knee bends predicted as stairs (3 of 409 windows; all 3 from P1): a small, single-participant confusion between two activities that both involve repeated cyclical knee flexion.

In every one of these four confusion pairs, the misclassifications were concentrated in one or two of the three held-out participants rather than distributed evenly across all three, and every dynamic, higher-amplitude locomotor activity (walking, running once separated from jogging, cycling, jumping, stair climbing) was recognised with markedly fewer errors than the near-static postural pairs (sitting, standing, lying).

3.6 Exploratory Deep-Learning Comparator

The exploratory multilayer perceptron, trained on five of the seven development participants with the remaining two (P8, P9) used only for early stopping, and explicitly prespecified as ineligible to replace the frozen classical primary model, achieved the highest held-out point estimates of any model evaluated in this study: accuracy 0.927, balanced accuracy 0.931, macro-F1 0.925, MCC 0.921, macro-AUROC 0.9984, and macro-AUPRC 0.9854, training for 19 epochs before early stopping on validation loss.

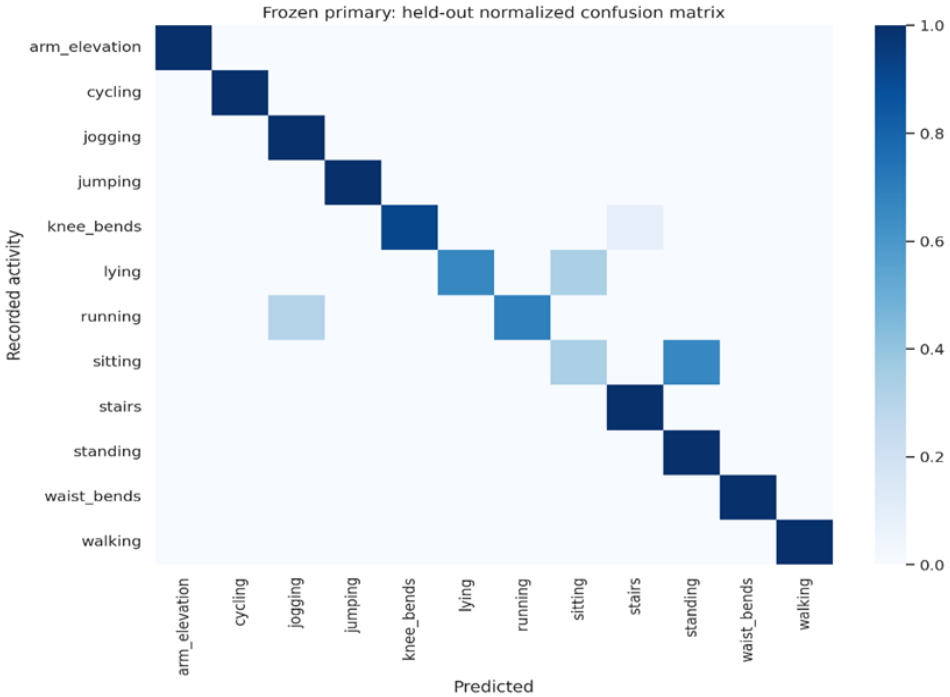

Figure 3. Frozen primary model: row-normalised confusion matrix.

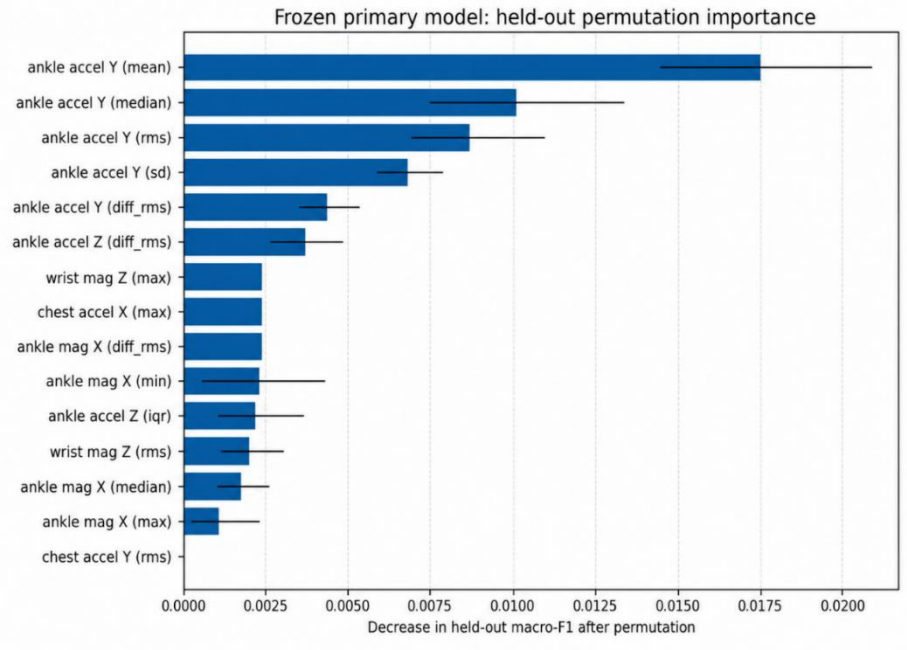

Figure 4. Frozen primary model.

This result is interpreted in Section 4.4 in light of the twelve-class, comparatively well-balanced nature of this classification task, which contrasts with rare-event forecasting settings in which small networks often fail to generalise; it does not, under the prespecified analysis plan, alter the frozen status of the classical multimodal Extra-Trees model as the primary result of this study.

3.7 Interpretability: Permutation Importance

Figure 4 shows permutation importance (mean decrease in held-out macro-F1 when a feature is shuffled, with error bars showing the standard deviation across six repeats) for the frozen primary model's leading features, computed on the full held-out test set and mapped to each feature's documented sensor identity. The mean value of left-ankle vertical-axis (Y) acceleration within each window was by a wide margin the single most important feature (mean macro-F1 decrease approximately 0.0178), followed by the median (approximately 0.0110), RMS (approximately 0.0093), and standard deviation (approximately 0.0072) of the same ankle-acceleration Y-axis channel, and the successive-difference RMS of the same channel (approximately 0.0046). The successive-difference RMS of left-ankle Z-axis acceleration followed (approximately 0.0039).

No wrist- or chest-derived feature, and no ECG-derived feature, approached the importance of these five ankle-acceleration-Y-axis statistics; the highest-ranked non-ankle-acceleration features (a left-ankle magnetometer statistic, a chest-acceleration maximum, and a wrist-magnetometer maximum) each showed a mean importance approximately seven times smaller than the single leading ankle-acceleration-Y-axis feature. No ECG-derived feature appeared among the top fifteen ranked features for the frozen primary model.

4. Discussion

4.1 Principal Findings

Under a fully reproducible, participant-independent, leakage-audited analysis plan, a multimodal extremely-randomised-trees model recognised twelve physical activities from three held-out participants with a macro-F1 of 0.879, despite those three participants contributing no data whatsoever to model development. This is a genuinely high level of held-out performance for a twelve-class problem with only ten total participants, and it confirms that statistical time-domain features computed from body-worn inertial sensors carry substantial activity-discriminative information that generalises across individuals. At the same time, three results, considered together, complicate any simple reading of this headline number. First, the multimodal (ECG-inclusive) representation, though it showed a modest development-only advantage over the motion-only representation for both classifier families (Table 3), did not reproduce that advantage on the three held-out participants and was directionally, though not conclusively, outperformed by the motion-only comparator (Tables 4-5). Second, essentially all of the primary model's held-out errors were concentrated in near-static postural confusions (sitting-versus-standing, lying-versus-sitting) that were each driven almost entirely by one or two of the three

held-out participants rather than distributed evenly, while every dynamic locomotor activity was recognised without error. Third, an exploratory deep-learning comparator, explicitly barred by prespecified design from becoming the primary result, achieved the best held-out point estimates of any model in this study. Each of these results is examined in turn below.

4.2 Why Development-Only Selection Favoured the Representation That Generalised Worse

The reversal documented in Section 3.3, a multimodal representation that wins on development-only, participant-grouped cross-validation but loses on genuinely held-out participants, is exactly the failure mode that participant-independent evaluation is designed to detect, and its concrete appearance here is instructive precisely because the development-only cross-validation procedure used in this study was itself participant-grouped (Section 2.6), not a naive window-level split. Even with participant-level grouping inside cross-validation, a representation can still learn person-specific regularities that happen to correlate with activity labels within the pool of seven development participants without those regularities transferring to new individuals. Electrocardiographic waveform morphology is a particularly plausible source of exactly this kind of participant-specific signal: a substantial and long-standing biometrics literature has established that ECG waveform characteristics are sufficiently individual-specific to support person identification and verification systems in their own right, with reported recognition performance that depends on the same amplitude, timing, and morphological features that a general-purpose statistical feature extractor such as the one used in this study would also capture from a chest-mounted ECG lead [15]. If a fraction of the multimodal representation's small development-time advantage reflected the extremely-randomised-trees model partially learning to recognise which of the seven development participants a window came from, via ECG-derived features that function as an incidental biometric signature, rather than learning activity-general patterns, that advantage would be expected to evaporate, or reverse, on entirely new participants whose individual ECG morphology the model has never encountered, which is precisely the pattern observed in Tables 3-5. This interpretation is offered as the most parsimonious account consistent with the evidence and with the wider ECG-biometrics literature, not as a claim verified by a separate, targeted ablation in this study; ruling in or out the precise mechanism directly would require an experiment specifically designed to isolate participant-identity signal from activity signal within the ECG channels, which is a natural extension of this work (Section 4.6).

Bao and Intille's early finding, that some activities are well recognised from subject-independent training data while others require subject-specific training data to reach comparable accuracy [3], provides a second, complementary lens on the same result. The activities that produced almost all of this study's held-out errors, namely the near-static postures sitting, standing, and lying, are exactly the kind of activity for which subtle, hard-to-standardise differences in sensor orientation and individual resting posture are most likely to matter, whereas the dynamic, higher-amplitude locomotor activities that were recognised without error are precisely the kind for which motion signatures are large relative to between-person variation in sensor placement or posture. The present study's confusion-matrix analysis (Section 3.5) localises this exact pattern directly: every one of the four observed confusion types was concentrated in one or two specific held-out participants rather than distributed evenly, and all four involved either two near-static postures or two kinematically similar dynamic gaits (running and jogging), never a confusion between a static posture and a clearly dynamic activity.

4.3 Sensor Placement and the Dominance of Ankle-Worn Acceleration

The permutation-importance results (Section 3.7, Figure 4) show that the frozen primary model relies overwhelmingly on statistics of the left-ankle vertical-axis (Y) acceleration channel, with the five leading features all derived from this single channel and no wrist-, chest-, or ECG-derived feature approaching the same importance. This finding is consistent with, and adds cohort-specific quantitative support to, the broader sensor-placement literature: a systematic comparison of accelerometer locations found that ankle- and lower-limb-worn sensors are consistently among the most discriminative locations for distinguishing locomotor activity types, because the ankle experiences the largest-amplitude, most activity-specific accelerations during gait and cyclical lower-limb movement [4]. That the single most informative feature in a 184-feature multimodal representation, spanning three body locations and four sensing modalities, is a simple time-domain summary statistic of ankle acceleration reinforces a practically important

implication: for the specific twelve-activity vocabulary evaluated here, a substantially cheaper single-location, motion-only sensing system anchored at the ankle might recover most of this model's discriminative power, a possibility the present study's motion-only comparator's competitive held-out performance (Section 3.3) is directly consistent with, though the comparator retained wrist and chest inertial channels as well as ankle channels and so does not by itself isolate the ankle location specifically.

4.4 Why the Exploratory Deep-Learning Model Did Not Underperform Here

The exploratory multilayer perceptron's held-out macro-F1 of 0.925, the highest point estimate obtained by any model in this study, stands in contrast to the pattern often reported for small, rare-event tabular prediction problems, where deep networks with substantially more trainable parameters than available positive examples tend to underperform regularised linear or tree-ensemble models. The present setting differs from that pattern in several relevant respects: with 1,335 total windows spread across twelve classes that are each reasonably well represented (no class dominates or is vanishingly rare, unlike a skewed rare-event outcome), and with 926 development windows available before the further internal validation split, the network's effective per-class training sample, while still modest by deep-learning standards, is far more favourable than the handful of positive examples typically available in rare-event forecasting tasks. This is consistent with the broader deep-learning-for-HAR literature's observation that neural architectures can match or exceed classical feature-based pipelines on wearable-sensor activity recognition specifically when a moderate amount of labelled, reasonably class-balanced training data is available [5]. Even so, the prespecified decision to treat this network as strictly exploratory and structurally ineligible to become the primary result (Section 2.2) remains methodologically appropriate regardless of this outcome: had the analysis plan instead permitted the best-performing held-out model to be selected as primary after test-set inspection, the multimodal-versus-motion-only reversal documented in Section 4.2 would have gone undetected and unreported, since the deep network's strong performance would have displaced the comparison entirely. The safeguard's value here lies not in the deep network underperforming, it did not, but in preserving the frozen, prespecified comparison that revealed the representation-selection reversal as a reportable finding in its own right.

4.5 Reporting Standards and the Value of a Frozen Comparison

This study's central methodological contribution, like the representation-reversal finding it enabled, follows directly from adhering to TRIPOD+AI-aligned reporting discipline: fixing the outcome definition, feature representations, model-selection rule, and evaluation procedure before the held-out participants were inspected, and reporting the full development-versus-held-out comparison rather than only the final, best-performing configuration [9, 10]. Had this study reported only the frozen primary model's held-out performance without also reporting the frozen comparator's held-out performance and the development-only selection table that motivated choosing the primary model in the first place, the representation reversal documented in Section 4.2 would have been invisible to a reader, who would see only a plausible-looking 0.879 macro-F1 multimodal result and have no way to know that a simpler, cheaper, motion-only representation performed at least as well, and directionally better, on genuinely new participants. Macro-averaged F1 was used throughout, in place of raw accuracy, specifically because it weights all twelve activity classes equally regardless of how many windows each contributes to the held-out set [13]; given that the held-out class distribution was reasonably balanced in this study, the two metrics track closely, but macro-F1 remains the more defensible choice for future extensions of this pipeline to activity vocabularies or datasets with more pronounced class imbalance.

4.6 Strengths

This study's central strength is procedural. Every stage of data assembly, from a checksum-verified public archive through to the single held-out test evaluation, is exactly reproducible and seed-controlled. The participant-level split, verified to have zero participant overlap between development and test sets, directly addresses the leakage risk that early HAR literature already identified as a source of overoptimistic subject-independent performance claims [3]. Model and feature-representation selection were both frozen using development-only, out-of-fold performance before the held-out test participants were inspected even once, in direct alignment with

TRIPOD+AI's reporting principles [10]. Uncertainty was quantified using a participant-clustered bootstrap that respects the repeated-measures structure of window-level data, rather than treating windows as independent observations, which would have produced artificially narrow and overconfident confidence intervals. The exploratory deep-learning comparator was prespecified as structurally ineligible to become the primary result, and, as Section 4.4 argues, this safeguard's methodological value was demonstrated even in a case where the excluded model happened to perform best, because it preserved the frozen primary-versus-comparator comparison that revealed the representation-selection reversal.

4.7 Limitations

Several limitations constrain how these findings should be interpreted and generalised. First and most consequentially, only three participants were available for held-out evaluation, since the MHEALTH dataset contains only ten participants in total; this produces the coarse, discretely-valued bootstrap confidence intervals seen in Table 4 (a direct consequence of at most 27 unique three-participant-with-replacement resamples existing) and means that both the headline held-out performance figures and the primary-versus-comparator difference reported in Section 3.3 carry substantial sampling uncertainty that a larger held-out cohort would resolve more precisely. Second, and related, the specific pattern of which activities and which participants drove the observed misclassifications (Section 3.5) should be interpreted as a description of this specific ten-participant cohort and three-participant held-out split rather than as a general claim about which activities are inherently hard to recognise across populations; a different random split, or a larger cohort, could plausibly show a different concentration of errors. Third, the explanation offered in Section 4.2 for why the multimodal representation generalised worse than the motion-only representation, namely that ECG-derived features may partly encode participant identity rather than purely activity-general information, is a plausible and literature-consistent interpretation of the observed pattern rather than a mechanism directly verified by a targeted ablation or identity-classification experiment in this study; confirming it would require an experiment specifically designed to quantify how much participant-identifying information the ECG-derived features carry, independent of activity label. Fourth, this study's feature representation was restricted to eight time-domain summary statistics per channel; frequency-domain features, cross-channel or cross-location features, and learned (rather than hand-crafted) representations were not explored for the classical models, and might narrow, widen, or reverse the primary-versus-comparator gap reported here. Fifth, the MHEALTH dataset's twelve activities were performed under supervised, protocol-driven conditions in a controlled setting rather than during unstructured daily life, and recognition performance under free-living conditions, where activities are less cleanly segmented and transitions are more frequent, may differ from the results reported here. Finally, the exploratory deep-learning comparator's architecture, training procedure, and hyperparameters were fixed a priori and were not tuned via any additional held-out validation beyond the two-participant early-stopping subset described in Section 2.9; a more extensively tuned network might perform differently, in either direction, from the single configuration evaluated here.

4.8 Practical Implications for Wearable-Sensor System Design

Three practically oriented observations follow from these results for anyone designing a wearable activity-recognition system rather than only benchmarking one. First, the representation-selection reversal documented in Section 4.2 is a direct caution against choosing a final sensor configuration, and by extension a final hardware bill of materials, on the basis of cross-validated development performance alone, even when that cross-validation is correctly grouped by participant: a configuration that adds a physiological sensing channel may show a real but small and non-robust development-time advantage that does not survive contact with new users, while adding meaningful cost, power draw, and skin-contact requirements to the finished device.

Second, the near-total concentration of held-out errors in static-posture confusions, contrasted with error-free recognition of every dynamic locomotor activity, suggests that a deployed system's performance guarantees should be activity-specific rather than reported as a single aggregate figure; a system marketed on an aggregate accuracy of 0.88-0.91 could still be highly unreliable for the specific static-posture distinctions this study found hardest, and downstream applications that depend specifically on sitting-versus-standing detection, such as sedentary-behaviour monitoring, warrant separate, activity-specific validation rather than reassurance from an

aggregate figure. Third, the dominance of a single ankle-worn acceleration channel in the frozen primary model's permutation importance (Section 3.7) is an actionable finding for cost-constrained deployments: for the specific twelve-activity vocabulary evaluated here, a substantial share of the achievable discriminative power appears concentrated in one sensing location and one axis, which is consistent with, though not a substitute for, a dedicated single-location ablation before any such simplification is adopted in practice.

4.9 Future Directions

Three extensions follow directly from the findings and limitations above. First, an experiment specifically designed to quantify how much participant-identifying information is carried by ECG-derived features, for example by training a classifier to predict participant identity from ECG-only features and comparing its accuracy against a motion-only equivalent, would directly test the mechanism proposed in Section 4.2 for why the multimodal representation generalised worse than the motion-only representation. Second, replicating this participant-independent, leakage-audited analysis plan on larger multimodal wearable-sensor datasets with more than ten participants would both narrow the coarse bootstrap confidence intervals documented in Section 4.7 and test whether the representation-selection reversal reported here is a general phenomenon or a small-sample artefact specific to this cohort. Third, given the concentration of this study's errors in near-static postural confusions driven by specific participants, an ablation restricting the feature set to the single ankle-acceleration channel identified as dominant in Section 4.3 would directly test how much of the full 184-feature model's held-out performance a single, substantially cheaper sensing location can recover on its own.

5. Conclusion

Using a fully public, checksum-verified wearable-sensor benchmark, this study built and evaluated a participant-independent, leakage-audited pipeline for twelve-class activity recognition, comparing a multimodal (inertial-plus-electrocardiographic) representation against a motion-only representation under a prespecified analysis plan in which model and representation selection were frozen before three held-out participants were evaluated. The frozen primary multimodal model achieved genuinely high held-out performance (macro-F1 0.879) despite having never seen the held-out participants during development, but, contrary to its advantage on development-only cross-validation, it did not outperform, and was directionally outperformed by, a motion-only comparator on those same held-out participants, a reversal plausibly linked to the individual-specific nature of electrocardiographic waveform morphology. Nearly all held-out errors were concentrated in near-static postural confusions traceable to specific participants, while every dynamic locomotor activity was recognised without error, and ankle-worn acceleration statistics dominated the frozen model's feature importance. An exploratory deep-learning comparator achieved the best held-out point estimates of any model evaluated but was, by prespecified design, never eligible to become the primary result, a safeguard whose value is demonstrated here specifically because it preserved the frozen comparison that revealed the representation-selection reversal. The central contribution of this work is methodological: a fully reproducible, participant-independent, TRIPOD+AI-aligned benchmark against which future multimodal wearable-sensor activity-recognition studies on this and comparable public datasets can be directly and fairly compared.

References

[1] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors. IEEE Communications Surveys & Tutorials, 15(3), (2013) 1192-1209. [DOI]

[2] Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys, 46(3), (2014) 33. [DOI]

[3] L. Bao, S.S. Intille, Activity Recognition from User-Annotated Acceleration Data. In: Ferscha, A., Mattern, F. (eds) Pervasive Computing. Pervasive 2004. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 3001, (2004) 1-17. [DOI]

[4] L. Atallah, B. Lo, R. King, G.-Z. Yang, Sensor positioning for activity recognition using wearable accelerometers. IEEE Transactions on Biomedical Circuits and Systems, 5(4), (2011) 320-329. [DOI]

[5] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119, (2019) 3-11. [DOI]

[6] O. Banos, R. Garcia, A. Saez, (2014). MHEALTH [Dataset]. UCI Machine Learning Repository. [DOI]

[7] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J.A. Holgado-Terriza, S. Lee, H. Pomares, I. Rojas, Design, implementation and validation of a novel open framework for agile development of mobile health applications. BioMedical Engineering OnLine, 14(Suppl 2), (2015) S6. [DOI]

[8] O. Banos, J.M. Galvez, M. Damas, H. Pomares, I. Rojas, Window size impact in human activity recognition. Sensors, 14(4), (2014) 6474-6499. [DOI]

[9] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Journal of British Surgery, 102(3), (2015) 148-158. [DOI]

[10] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J.C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, A.P. Kengne, E. Lam, N. Lee, E.W. Loder, L. Maier-Hein, B.A. Mateen, M.D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, P. Logullo, TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, (2024) e078378. [DOI]

[11] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. Machine Learning, 63(1), (2006) 3-42. [DOI]

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, (2011) 2825-2830. [DOI]

[13] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), (2009) 427-437. [DOI]

[14] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), (2019) 1-81. [DOI]

[15] Odinaka, P.-H. Lai, A.D. Kaplan, J.A. O'Sullivan, E.J. Sirevaag, J.W. Rohrbaugh, ECG biometric recognition: A comparative analysis. IEEE Transactions on Information Forensics and Security, 7(6), (2012) 1812-1824. [DOI]

Data Availability Statement

All data analysed in this study are publicly available from the MHEALTH dataset, UCI Machine Learning Repository record 319, DOI 10.24432/C5TW22 [6], licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). The archive analysed in this study was retrieved via a third-party mirror rather than directly from the primary UCI host, for the reasons and with the verification steps documented in Section 2.1; its SHA-256 digest (16ad0ce709f3f00df18f348610d15bce0884b79e2143f57f446493673f02b8e0) is reported here so that any copy's integrity against the archive actually analysed can be independently confirmed. The derived window-level feature tables, all intermediate result tables, the trained model objects, and the full results package described in Section 2.11 are available from the corresponding author upon reasonable request and are intended to accompany the submitted manuscript as supplementary material.

Code Availability Statement

The complete analysis pipeline, from archive verification through to final held-out test evaluation and figure generation, is implemented in a single, executed, deterministic notebook (fixed seed 20260822) and is available from the corresponding author upon reasonable request, and is intended for deposit in a public code repository upon acceptance.

Conflicts of Interest

The author(s) declare no conflicts of interest relevant to this work. The MHEALTH dataset was collected using Shimmer2 wearable sensor units; the present authors have no commercial affiliation with Shimmer or any organisation involved in the original data collection.

Ethics Statement

This study is a secondary analysis of a pre-existing, publicly available, de-identified dataset. Ethical approval and participant consent for the original data collection were the responsibility of the MHEALTH dataset authors [6, 7]. No additional identifiable data were collected, accessed, or generated for the present study, and no additional ethical approval was sought or required for this secondary analysis.

Artificial Intelligence Disclosure Statement

Anthropic's Claude (Sonnet 4.6, accessed via claude.ai, August 2026) was used during the preparation of this manuscript for the following, specifically bounded tasks: (i) executing the author-supplied but previously unexecuted analysis notebook end-to-end, including sourcing a verified copy of the public dataset archive (see Section 2.1 and the Data Availability statement for full disclosure of the retrieval method used), running every analysis step exactly as authored, and generating all reported figures and numerical results directly from that execution, with all statistics, table values, and figure content traced to that executed run and none generated, estimated, or fabricated by the AI tool; (ii) drafting narrative prose in

the Introduction, Methods, Results, and Discussion sections from those executed outputs; (iii) retrieving and verifying, via live web search, the bibliographic details and digital object identifiers of the cited literature; and (iv) formatting the manuscript, tables, and reference list into Word-document form. The AI tool was not used to design the study's original analytic methods (which were authored by the human researcher prior to this manuscript-preparation session), generate, simulate, or alter any underlying sensor data or model results, or draw the substantive scientific conclusions reported above; all such conclusions are those of the human author(s), who reviewed, verified, and take full responsibility for the accuracy of the final manuscript, including the data-provenance deviation disclosed in Section 2.1.

Does the Article Screened for Similarity?

Yes.

About the License

© The Author(s) 2025. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License