REVIEW ARTICLE

DOI: 10.54392/ijpefs2642

International Journal of
PHYSICAL EDUCATION, FITNESS AND SPORTS

Explainable Machine Learning for Physical Activity Intensity Classification using Multimodal Wearable Sensors and Sensor Ablation Analysis

Archibald Danquah-Amoah [a, *], Viena Sembaiyan Gunasekaran [b], Alex Abrokwa Fosuhene [c], Mensah Sitti [d]

[a] Department of Medical Laboratory Technology, Biomedical Equipment Engineering, Accra Technical University, Ghana.
[b] Department of Biomedical Engineering, Hochschule Furtwangen University, Jakob-Kienzle-Straße 17, 78054 Villingen-Schwenningen, Germany.
[c] Department of Science Laboratory Technology, Accra Technical University, Ghana.
[d] Department of Computer Science and Engineering, University of Mines and Technology, Tarkwa, Ghana.
* Corresponding Author E-mail: adanquah-amoah@atu.edu.gh
DOI: https://doi.org/10.54392/ijpefs2642
Received: 17-05-2026; Revised: 02-10-2026; Accepted: 19-10-2026; Published: 13-11-2026

Abstract: Wearable sensors are used to assess whether individuals reach moderate or vigorous physical activity, but many classification studies test models on people who also supplied training data and report accuracy without any statement of uncertainty. This study classified activity intensity in people withheld from training and asked which sensors are needed and how reliably predictions generalise to new participants. Previously collected recordings of twelve activities from eight adults wearing wrist, chest and ankle inertial sensors and a heart-rate monitor were labelled low, moderate or vigorous using the 2024 Adult Compendium of Physical Activities. Logistic regression, random forest and gradient boosting were evaluated by leave-one-participant-out validation with all tuning nested inside the training participants, across eleven sensor combinations. A heart-rate normalisation that needs no labelled calibration session was compared with raw heart rate, and prediction sets were built with conformal methods calibrated on unseen participants. With all sensors, random forest achieved a macro-F1 score of 0.954 and gradient boosting achieved 0.951. A wrist sensor alone reached 0.87 and adding heart rate raised this to 0.92, an improvement seen in seven of eight participants. The personalised heart-rate measure gave no clear gain over raw heart rate. Prediction sets calibrated using held-out participants achieved approximately 90% marginal coverage, whereas sets calibrated on training data covered only 73% of cases. Vigorous activity was under-covered until coverage was calibrated separately for each intensity class, which raised it from 77% to 90%. A wrist sensor combined with heart-rate monitoring is a promising configuration for further evaluation, and stated uncertainty should be calibrated on people outside the training set.

Keywords: Physical Activity Intensity, Wearable Sensors, Heart Rate, Conformal Prediction, Subject-Independent Validation, Explainable Machine Learning

1. Introduction

Physical activity guidelines for children, adults and older people are written in terms of intensity. They recommend minimum amounts of moderate and vigorous activity each week, and the evidence behind them links these amounts to cardiorespiratory fitness and cardiovascular health (Franklin *et al.,* 2022). Duration alone therefore tells a physical educator, a coach or a researcher only half of what they want to know; how hard the person worked matters just as much. Intensity is normally expressed in metabolic equivalents (MET), the ratio of the energy cost of an activity to resting energy expenditure, and the 2024 Adult Compendium of Physical Activities provides updated MET values for hundreds of activities, from lying quietly to competitive sport (Herrmann *et al.,* 2024). Conventional bands place activities below 3 MET

as light, those between 3 and 6 MET as moderate, and those at 6 MET or more as vigorous.

Wearable devices are widely used to monitor activity outside the laboratory. Consumer and research devices with accelerometers, gyroscopes and optical or chest-strap heart-rate sensors are worn by millions of people, and a living umbrella review of systematic reviews has documented how accurate they are for steps, heart rate, energy expenditure and other outputs. In sport and exercise medicine, wearables are used to monitor heart rhythm and training response (Beavers & Chung, 2022), and artificial intelligence applied to wearable and motion-analysis data has become a fast-growing part of sports biomechanics (Souaifi *et al.,* 2025). For intensity in particular, large studies have compared cadence thresholds with machine-learning classifications of wrist accelerometer data and found that the two can agree at group level while differing substantially for individuals (Wei *et al.,* 2024b, Wei *et al.,* 2024a).

The individual level is where most practical decisions are made. A teacher who uses a wrist device to check whether a pupil reached vigorous effort, or a coach who monitors an athlete's easy days, needs the classification to be correct for that person as well as on average. Two weaknesses of the current machine-learning literature make it hard to know how often this is the case. First, many activity-recognition studies split windows from the same participants into training and test sets. Because movement patterns are highly personal, the test then rewards familiarity with the person, and the reported accuracy does not describe performance for a new user. Leakage of this kind is now recognised as a widespread cause of over-optimistic results in machine-learning research (Kapoor & Narayanan, 2023), and a recent study of wearable stress detection showed how much performance and calibration change when evaluation becomes subject-independent (Papamentzelopoulos & Nikoletseas, 2026). There is a second problem. A classifier will give a label to every window, even an ambiguous one, and seldom says how often that label is likely to be wrong.

In sport and physical education these weaknesses have real consequences, since decisions are taken about individual people and often on the spot. Consider a teacher checking that a lesson gave each pupil enough vigorous activity, a strength and conditioning coach making sure a recovery session really was light, or a researcher counting moderate-to-vigorous minutes across a school week. Every one of them is affected when a window is placed on the wrong side of an intensity boundary, and a tool that never admits doubt encourages them to trust it too far. Arguably, a classifier tested fairly on new people, and able to state its uncertainty in calibrated terms, is worth more than a marginally more accurate one that offers neither.

Heart rate raises a separate issue. An accelerometer records movement; heart rate reflects what that movement costs the body, so it ought to tell apart activities that look alike to a motion sensor yet differ in effort, brisk walking and stair climbing being the obvious example. The difficulty is that absolute heart rate varies from person to person with fitness, age and other factors, and a value that signals moderate effort in one individual can signal vigorous effort in another. The usual remedy is to scale heart rate to each person's resting and maximal values, but that normally needs a supervised resting measurement, which is rarely possible in the field.

Four aims followed from this. The first was to classify intensity in participants kept out of all model development. The second was to find out, by sensor ablation, what each body location and heart rate adds. The third was to test a heart-rate normalisation built without labels from each person's own recording. The fourth was to compare conformal prediction sets calibrated on unseen participants with naive sets calibrated on the training data, looking at coverage overall and within each intensity class. Three contributions are claimed. First, the study gives a subject-independent benchmark for intensity classification on a public dataset that many groups already use. Second, it puts a figure on the shortfall of a single wrist device, with and without heart rate, against a three-sensor system. Third, it shows that uncertainty statements from wearable classifiers require calibration on participants excluded from model fitting, with coverage interpreted marginally under exchangeability assumptions, and that the vigorous class needs class-conditional calibration if it is to be protected. The literature is reviewed next, followed by the materials and methods, the results with their discussion, and the conclusions.

2. Literature Survey

2.1 Measuring physical activity intensity

The traditional route to intensity is a set of cut-points applied to accelerometer counts or to acceleration magnitude. For wrist-worn devices in children and adolescents, a systematic review found

that calibrated thresholds varied widely and achieved only moderate accuracy on independent validation, and a separate evaluation of wrist cut-points in young people came to much the same view (Trost *et al.*, 2022). Methods for capturing human movement that have been checked against direct observation were brought together in a scoping review, which also described when each of them tends to work well (Letts et al., 2024). Models trained on raw acceleration offer another option. In one example, deep neural networks identified both the wear site and the activity intensity from raw acceleration in children and adults; portable indirect calorimetry supplied the reference metabolic equivalents, and accuracy exceeded 90% (Ng *et al.*, 2024). At population scale, machine-learning intensity estimates have been compared with cadence-based methods in tens of thousands of UK Biobank participants, again showing acceptable group-level but poor individual-level agreement.

Studies in this area also differ in what they treat as the true intensity. Measured oxygen uptake gives the most direct reference but is costly and limits studies to a few dozen people in a laboratory. The alternative is to assign each activity its Compendium value, which allows existing activity datasets to be reused for intensity classification but treats every person performing an activity as working at the same relative effort. The present study takes the second route, because it allows an open, reproducible analysis, and examines how sensitive the conclusions are to the coding of activities that lie near an intensity boundary.

2.2 Machine Learning for Human Activity Recognition

Human activity recognition from wearable sensors has a large and active literature. Reviews have summarised the use of deep learning with wearable sensors (Zhang *et al.*, 2022b), the state of the art in machine-learning techniques for activity recognition (Hossen & Abas, 2025) and the use of deep learning for smartphone-based recognition (Stampfler *et al.*, 2023). Public datasets with several body-worn sensors, including the one used here, have been used to test a long series of architectures: residual networks combined with bidirectional recurrent layers (Li & Wang, 2022), multi-scale convolutional and gated recurrent fusion models (Zhang *et al.*, 2022a), conformer models borrowed from speech recognition (Kim *et al.*, 2022), dual-branch interactive networks for multichannel time series (Tang *et al.*, 2022), cross-attention pyramid networks (Pang *et al.*, 2024) and sparse transformers with channel attention (Cao & Wang, 2023). Other work has improved efficiency through early exiting and gradient boosting (Yu *et al.*, 2024) or through purely convolutional models small enough for edge devices (Xu *et al.*, 2026). These studies typically report accuracy between 90% and 98% for recognising the activities themselves.

Most of this work relies on deep architectures that learn features from raw signals. Hand-crafted features computed from short windows, such as the mean, spread, percentiles and spectral content of acceleration, remain competitive when data are limited, and they have the practical advantage that their meaning is clear to a physiologist or coach. Tree ensembles trained on such features are also fast to train and to evaluate repeatedly, which matters when every model must be refitted many times for nested validation and for subject-level calibration. The present study therefore uses hand-crafted features with standard learners, in order to place the emphasis on evaluation design, sensor choice and uncertainty rather than on architecture.

Two observations from this literature motivate the present secondary analysis. The first is that the target is almost always the activity label, not its intensity. For health and physical education, the activity label is a means to an end, what matters is whether the person reached moderate or vigorous effort, and errors between two activities of the same intensity are harmless while errors across an intensity boundary are not. The second is that model complexity has grown faster than evaluation rigour. The validation design is often not stated in abstracts, and comparisons between architectures rarely include simple, well-tuned baselines such as gradient boosting on hand-crafted features.

2.3 Subject-independent evaluation, personalization and leakage

Evaluation on unseen people is the realistic test for any wearable classifier intended for new users. Leave-one-subject-out validation was used to examine whether gyroscopes add value to accelerometers, and the answer depended on the activity and the mounting site (Shang *et al.*, 2022). A lightweight convolutional model that performed well under leave-one-subject-out validation on its own resistance-training dataset showed lower macro F1 scores on public daily-activity datasets, which the authors attributed to broader activity diversity and differences between sensor domains. Domain adaptation has been proposed to transfer models between sensors placed on different body parts (Zhu *et*

al., 2025), and personalised federated learning has been used to cope with differences between users' data distributions (Cheng *et al.,* 2023). Each of these approaches accepts that people differ and that models must be judged on people they have not seen.

The size of the effect of ignoring this principle is now well documented outside activity recognition. Guidance has been published on how to avoid data leakage in biological machine learning (Bernett *et al.,* 2024), and in neuroimaging, leakage was shown to inflate prediction performance substantially (Rosenblatt *et al.,* 2024). In wearable stress detection, leakage between windows from the same people inflated macro F1 by 27 to 61 points and corrupted model calibration. Movement data are especially prone to this problem because they carry a strong personal signature, wrist accelerometry recorded during walking can identify individuals, much as a fingerprint does (Koffman *et al.,* 2023). A classifier evaluated on windows from people it has already seen can therefore succeed partly by recognising the person.

2.4 Heart Rate and Physiological Signals

Heart rate has long been combined with motion sensors to estimate energy expenditure. Individualised regression models based on heart-rate reserve predicted energy expenditure during treadmill protocols more accurately than models based on absolute heart rate, especially at submaximal intensities. Machine learning has been used to predict oxygen uptake kinetics during heavy-intensity cycling from heart rate and other signals (Hedge *et al.,* 2023), to predict relative intensity from a contact-free radar heart-rate sensor combined with a three-dimensional accelerometer (Biró *et al.,* 2023), and to estimate energy expenditure from electrocardiogram and acceleration signals with interpretable deep learning (Song *et al.,* 2025). Sport-specific applications include energy expenditure during badminton training (Yan *et al.,* 2025), during assisted and non-assisted gait using inertial, electromyographic and heart-rate data (Lopes *et al.,* 2022) and during dance (Tao *et al.,* 2025). A common thread is that physiological signals add most when movement alone is ambiguous about effort, and that normalisation to the individual helps when people differ in fitness. Most of this work, however, requires a calibration session to obtain resting or maximal values, and few studies have tested whether a label-free personal baseline is good enough.

2.5 Explainability in Wearable Classification

Explainable machine learning aims to show which inputs drive a model's decisions. In activity recognition, it has been used to interpret classifiers built on knee-worn sensors (Elkhouly *et al.,* 2026), and feature-importance analyses appear in several of the studies above. For a sensor-ablation question, explanations at the level of individual features are less useful than explanations at the level of sensor groups, because the practical decision is which device to wear. Grouped permutation importance, in which all features from one sensor are shuffled together on held-out people, answers this question directly and avoids the problem that correlated features from the same sensor share importance between them.

2.6 Uncertainty and Conformal Prediction

Clinical machine learning has begun to adopt formal methods for stating uncertainty. A scoping review of uncertainty-aware deep learning in healthcare described Bayesian, ensemble and other approaches and noted how seldom they are evaluated in practice (Loftus *et al.,* 2022). Conformal prediction is a simple, model-agnostic alternative that turns any classifier's scores into prediction sets with a guaranteed long-run coverage, provided that calibration data and new cases are exchangeable (Vazquez & Facelli, 2022). Recent work has extended the idea to control other kinds of risk. The exchangeability condition is the weak point for wearable data. If calibration scores come from people whose data were also used for training, they will be more confident than scores for new people, and the resulting sets will be too small. To the author's knowledge, the consequences of this for wearable activity or intensity classification have not been quantified.

A second weak point concerns imbalanced classes. Standard conformal prediction guarantees coverage on average over all cases, so when one class is rare, the average can be met while that class is covered poorly. Class-conditional, or Mondrian, conformal prediction addresses this by computing a separate threshold for each class from calibration cases of that class only. In intensity classification, vigorous activity is usually the least frequent class and also the one that physical activity guidelines care most about, so class-conditional coverage is the natural requirement. Whether the extra protection costs larger, less informative prediction sets is an empirical question that depends on how well the underlying classifier separates the classes.

2.7 Comparison of existing studies and research gap

Table 1 summarises representative studies close to the present one, using the information reported in their abstracts. The studies differ in data, targets and metrics, so the table is meant to show design features rather than to rank methods.

Four gaps follow from this review. Studies on the public multi-sensor datasets have concentrated on activity labels rather than intensity. Validation on unseen people is not universal, and when it is used, the drop in performance is often large. Heart-rate normalisation without a calibration session has not been tested in a subject-independent design. And no study in Table 1 attached calibrated uncertainty to activity or intensity predictions or checked whether stated coverage holds for each new person. The present study addresses each of these gaps.

3. Materials and Methods

3.1 Dataset and participants

For this secondary analysis, the publicly available PAMAP2 physical activity monitoring dataset (Reiss & Stricker, 2012) was obtained from the UCI Machine Learning Repository. In the original data collection, nine participants wore inertial measurement units on the dominant wrist, the chest and the dominant ankle, each sampling three-axis acceleration, angular velocity and magnetic field at 100 Hz, together with a chest-strap heart-rate monitor sampling at about 9 Hz, while they performed a protocol of up to twelve activities. Participant 109 contributed only twelve usable windows from one activity and no valid resting heart-rate estimate, and was excluded before any modelling, as planned. The analysis therefore used eight participants (seven men, one woman, age 26.8 ± 3.2 years, height 181 ± 8 cm, mass 83 ± 9 kg). Participant 103 performed eight activities, participants 104 and 107 eleven, and the others all twelve. The data were collected with informed consent and are distributed for reuse. Figure 1 outlines the analysis.

3.2 Intensity Labels

Each activity was assigned a MET value from the 2024 Adult Compendium and grouped as low (below 3), moderate (3 to 5.9) or vigorous (6 or more), as in Eq. 1.

Table 1. Representative studies on wearable activity and intensity classification

Study	Data and sensors	Target	Testing on unseen people	Uncertainty reported
Zhang *et al.*(2022)	WISDM, UCI-HAR, PAMAP2	Activity	Not stated in abstract	No
Kim *et al.* (2022)	WISDM, UCI-HAR, PAMAP2, oversampling	Activity	Not stated in abstract	No
Shang *et al.* (2022)	UCI-HAR, WISDM, accelerometer and gyroscope	Activity	Leave-one-subject-out	No
Cheng *et al.* (2023)	Four datasets including PAMAP2	Activity, federated	Per-client personalisation	No
Ng *et al.* (2024)	54 participants, hip, wrist, chest accelerometers	Intensity (measured MET)	Not stated in abstract	No
Biró *et al.* (2023)	Radar heart rate and 3D accelerometer	Intensity, RPE, MET	Leave-one-out and out-of-sample	No
Zhu *et al.* (2025)	DSADS, PAMAP2, several body sites	Activity, cross-domain	Cross-body-part transfer	No
Xu *et al.* (2026)	Smart dumbbell, PAMAP2, MHEALTH	Exercise and activity	Leave-one-subject-out	No
Papamentzelopoulos *et al.* (2026)	WESAD, SWELL-KW, HRV and EDA	Stress	Leave-one-subject-out	Calibration error
Present study	PAMAP2, wrist, chest, ankle IMU and heart rate	Intensity (Compendium MET)	Leave-one-participant-out	Conformal sets with per-person coverage

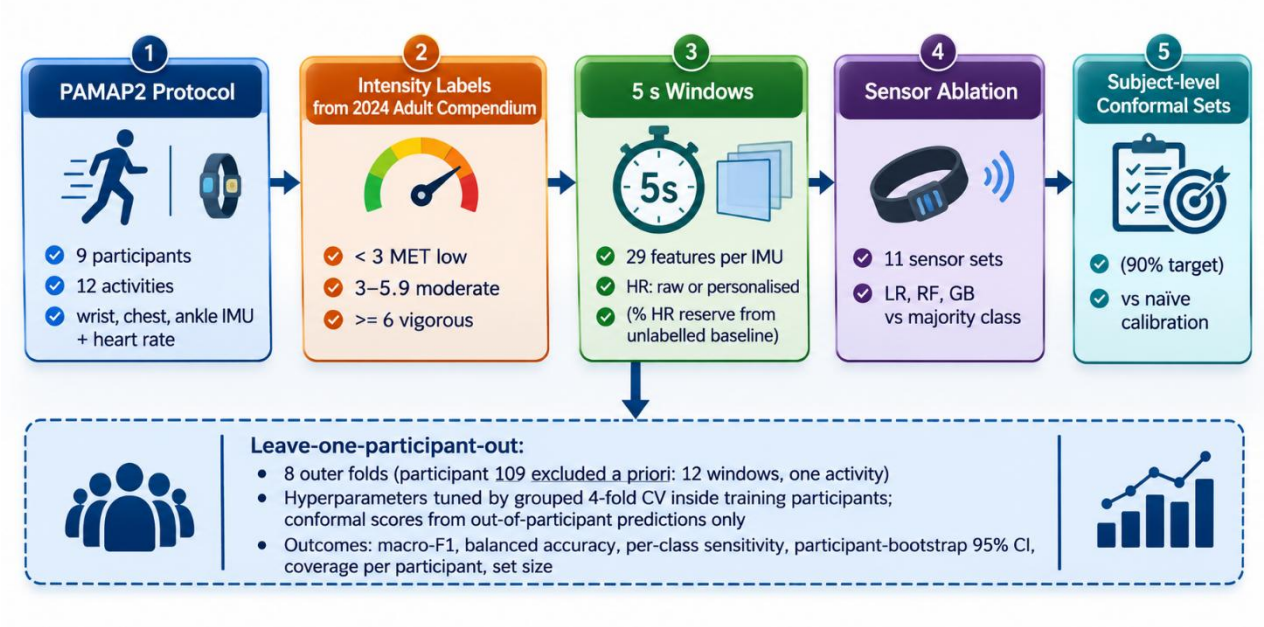


Figure 1. Study workflow.

Lying and sitting (1.0), standing (1.3) and ironing (1.8) were low, vacuum cleaning (3.0), descending stairs (3.5), walking (3.8), cycling (4.0) and Nordic walking (4.3) were moderate, and ascending stairs (6.8), running (8.5) and rope jumping (11.8) were vigorous. Because some activities sit near a boundary or have several possible Compendium codes, three alternative mappings were tested as sensitivity analyses.

$$c(a) = \{\text{low if MET}(a) < 3,\ \text{moderate if } 3 \leq \text{MET}(a) < 6,\ \text{vigorous if MET}(a) \geq 6 \quad (1)$$

3.3 Windowing and Features

Signals were cut into 5 s windows that never crossed an activity boundary, giving 3,819 windows (1,595 low, 1,719 moderate and 505 vigorous). Windows with more than 20% missing inertial data were discarded and shorter gaps were interpolated linearly. For each sensor, 29 features were computed from the 16 g accelerometer and the gyroscope. The acceleration magnitude (Eq. 2) is independent of sensor orientation, which matters for a wrist device that rotates with the forearm.

$$a(t) = \sqrt{a_x^2 + a_y^2 + a_z^2} \quad (2)$$

The mean, standard deviation and 10th and 90th percentiles were computed for acceleration magnitude, angular velocity magnitude and each acceleration axis. From the power spectral density P(f_k) of the acceleration magnitude, estimated with Welch's method, the dominant frequency, the log power in three bands (0.5 to 3 Hz, 3 to 8 Hz and 8 to 15 Hz) and the spectral entropy (Eq. 3) were derived, where p_k is the power at frequency k normalised to sum to one. Correlations between acceleration axes and the mean jerk (Eq. 4), the average absolute change of the magnitude per second with sampling frequency f_s, completed the set.

$$H_{\text{spec}} = -\sum_k p_k \ln p_k,\ p_k = \frac{P(f_k)}{\sum_j P(f_j)} \quad (3)$$

$$\text{jerk} = \frac{f_s}{n-1}\sum_{t=1}^{n-1}|a(t+1) - a(t)| \quad (4)$$

Heart rate was carried forward over gaps of up to 2 s. Raw heart-rate features were the window mean and the slope of a straight line fitted to heart rate within the window, which indicates whether heart rate is still rising after a change in activity.

3.4 Label-Free Heart-Rate Personalisation

For each participant, a resting value HRrest was estimated as the 5th percentile of all heart-rate samples in that participant's protocol recording, without reference to activity labels, and maximal heart rate HRmax was predicted from age (Tanaka *et al.*, 2001). Each window's mean heart rate was then expressed as a percentage of heart-rate reserve (Eq. 5) and as the difference from the resting value. Because the estimate uses no labels, it can be computed for a new user from an initial period of ordinary wear.

$$\%\text{HRR} = 100\,\frac{HR_{\text{window}} - HR_{\text{rest}}}{HR_{\max} - HR_{\text{rest}}},\ HR_{\max} = 208 - 0.7 \times \text{age} \quad (5)$$

The estimated resting values correlated with those listed in the dataset documentation (r = 0.80) but

sat 6.1 beats per minute higher on average. That is not surprising: the protocol contained only a few minutes of lying, so the lowest part of each recording is a little above a genuine resting value.

3.5 Learners and Sensor Sets

The learners were multinomial logistic regression with class weighting, random forest (300 trees with class weighting) and histogram gradient boosting (150 iterations), compared with a majority-class reference that always predicts the most frequent class. Gradient boosting builds an additive model of shallow trees, each fitted to the gradient of the multinomial log-loss of the current ensemble (Eq. 6), where h_m is the tree added at step m and η the learning rate.

$$F_m(x) = F_{m-1}(x) + \eta\, h_m(x),\ \hat{p}_k(x) = \frac{\exp F_k(x)}{\sum_j \exp F_j(x)} \quad (6)$$

Eleven sensor sets were evaluated with gradient boosting: wrist, chest and ankle alone, wrist and ankle with personalised heart rate, wrist with raw heart rate, all three inertial sensors with no heart rate, raw heart rate or personalised heart rate, and heart rate alone, raw or personalised. Logistic regression and random forest were evaluated on the full set of inertial sensors with personalised heart rate.

3.6 Validation and Performance Measures

Evaluation used leave-one-participant-out cross-validation with eight outer folds, so that every prediction was made for a person whose data had not been used in training or tuning. Inside each fold, hyperparameters (regularisation strength, minimum leaf size, learning rate and tree depth) were chosen by grouped four-fold cross-validation on the training participants, and imputation and scaling were fitted on training data only. The primary outcome was the macro-averaged F1 score (Eq. 7), the unweighted mean of the class-wise harmonic means of precision P_k and sensitivity S_k over K = 3 classes, which gives the small vigorous class the same weight as the others. Balanced accuracy (Eq. 8), the mean of the class-wise sensitivities, and per-class sensitivity were also reported.

$$\text{macro} - \text{F1} = \frac{1}{K}\sum_{k=1}^{K} \frac{2\,P_k\,S_k}{P_k + S_k} \quad (7)$$

$$\text{BA} = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \quad (8)$$

Macro-F1 was also computed separately for each held-out participant, and 95% confidence intervals for the mean per-participant score were obtained from 2,000 bootstrap resamples of participants (Eq. 9), where the bracketed quantities are the 2.5th and 97.5th percentiles of the bootstrap means. Paired comparisons of per-participant F1 between sensor sets used the Wilcoxon signed-rank test, with eight participants, the smallest attainable two-sided p value is 0.008.

$$\text{CI}_{95\%} = \left[\bar{F}^*_{(0.025)}, \bar{F}^*_{(0.975)}\right] \quad (9)$$

To describe what the models relied on, grouped permutation importance was computed on each held-out participant. All features of one sensor group G (wrist, chest, ankle or personalised heart rate) were shuffled together ten times and the fall in macro-F1 was recorded (Eq. 10).

$$I_G = \text{macro} - \text{F1}(X) - \frac{1}{B}\sum_{b=1}^{B} \text{macro} - \text{F1}\left(X_G^{(b)}\right) \quad (10)$$

3.7 Conformal Prediction Sets

Prediction sets were formed from the nonconformity score s = 1 − $\hat{p}_y$, where $\hat{p}_y$ is the predicted probability of the true class. A class k was included in the set for a new window x when its score fell at or below a threshold $\hat{q}$ (Eq. 11). The threshold is the finite-sample-corrected quantile of n calibration scores (Eq. 12), which guarantees coverage of at least 1 − α when calibration and test cases are exchangeable.

$$\Gamma(x) = \left\{k : 1 - \hat{p}_k(x) \le \hat{q}\right\} \quad (11)$$

$$\hat{q} = \text{Quantile}\left(s_1, \ldots, s_n, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right) \quad (12)$$

Three ways of obtaining calibration scores were compared at α = 0.10. In naive calibration, scores came from the model's own predictions on its training data. In subject-level cross-conformal calibration, every training participant was held out in turn, the model was refitted on the remaining training participants with the hyperparameters already selected, and scores were collected for the held-out participant, so that all calibration scores came from people the model had not seen. The class-conditional (Mondrian) version used the same scores but set a separate threshold $\hat{q}_k$ for each intensity class, computed only from calibration windows whose true class was k (Eq. 13). This protects the coverage of each class, at the cost of using fewer scores per threshold.

$$\hat{q}_k = \text{Quantile}\left(\{s_i : y_i = k\}, \frac{\lceil (n_k+1)(1-\alpha) \rceil}{n_k}\right) \quad (13)$$

Coverage, the share of windows whose set contained the true class (Eq. 14), was reported overall, by class and for each held-out participant, together with the mean set size and the share of empty sets. All analyses used Python 3.13 and scikit-learn 1.9 with fixed random seeds.

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[y_i \in \Gamma(x_i)] \quad (14)$$

4. Results and Discussion

4.1 Classification in unseen participants

With all inertial sensors and personalised heart rate, the three learners performed similarly in participants they had not seen (Table 2). Random forest reached a macro-F1 of 0.954, gradient boosting 0.951 and logistic regression 0.939, against 0.207 for the majority-class reference. The overlapping confidence intervals show that, once evaluation is honest and features are sensible, the choice among standard learners matters little for this task, a pattern also seen when several learners were compared under leave-one-subject-out validation for activity recognition. Gradient boosting was used for the remaining analyses because it handles missing values natively, supports the ablation and conformal analyses at modest cost and performed as well as the best alternative.

A word on logistic regression is in order. Its macro-F1 was only about 0.01 to 0.02 lower than that of the tree ensembles, and it had the highest sensitivity for the low class of all learners (0.967). A linear model of this kind can be inspected directly, its coefficients can be reported, and it can run on very small devices, which is the same concern that has driven work on lightweight models for edge-based fitness tracking. Where transparency or computing resources matter more than the last percentage point of accuracy, it is a reasonable choice for intensity classification from good features.

Errors were concentrated in the vigorous class, of which 91% of windows were recognised, the misclassified vigorous windows were mostly assigned to moderate activity (Figure 2a). Low and moderate windows were recognised in 96% and 97% of cases. Errors of this kind, between neighbouring bands, are the least harmful in practice, and almost no low window was called vigorous or the reverse. By activity, rope jumping, ironing and stair ascent were the hardest to place (Figure 2b).

Table 2. Leave-one-participant-out classification of intensity with all sensors and personalised heart rate

Model	Macro-F1	Balanced accuracy	Participant F1 (95% CI)	Sensitivity low	Sensitivity moderate	Sensitivity vigorous
Majority class	0.207	0.333	0.206 (0.196–0.215)	0.000	1.000	0.000
Logistic regression	0.939	0.941	0.940 (0.926–0.953)	0.967	0.939	0.915
Random forest	0.954	0.950	0.951 (0.926–0.969)	0.952	0.969	0.929
Gradient boosting	0.951	0.945	0.947 (0.926–0.962)	0.957	0.968	0.911

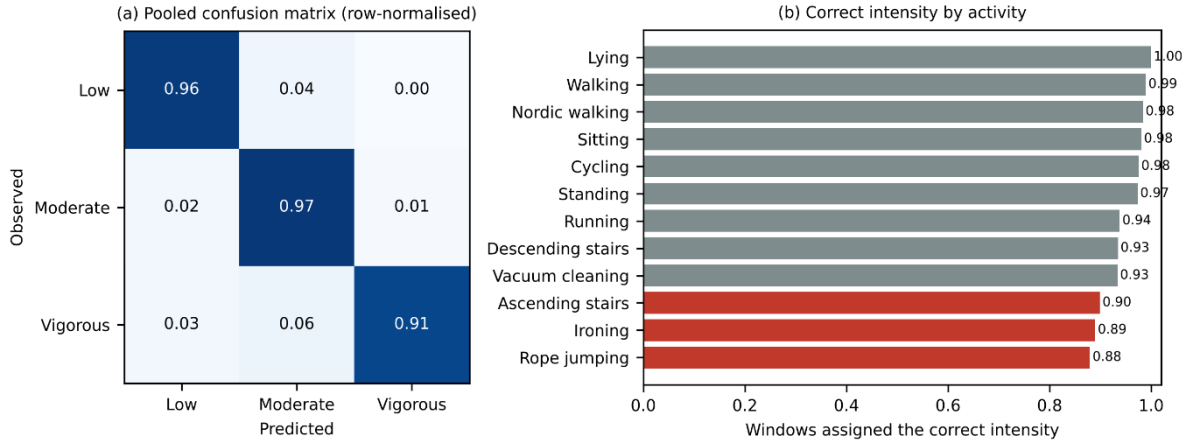


Figure 2(a). Pooled row-normalised confusion matrix for gradient boosting with all sensors and personalised heart rate, and **(b)** proportion of windows assigned the correct intensity for each activity.

Rope jumping produced the fewest windows of any activity, so the models saw little of it, ironing involves arm movement similar to light household tasks of higher intensity, and stair ascent at a self-selected pace lies close to the vigorous boundary, so some windows may be physiologically moderate even though the Compendium places the activity above 6 MET.

Performance varied between people. Per-participant macro-F1 for gradient boosting ranged from 0.88 to 0.97, and participant 104 was the hardest case for both random forest and gradient boosting. Variation of this size is hidden when windows from the same person are allowed in both training and test data, because movement carries a personal signature and leakage inflates apparent performance. It is also the variation that matters for a teacher or coach, because a tool that works well on average can still mislead for a particular pupil or athlete. Reporting per-participant results, as done here, gives readers a fairer impression of what a new user can expect.

4.2 Comparison with Existing Studies

The studies summarised in Table 1 differ from the present one in target, data and validation, so their numbers cannot be set side by side with the results above. Most recognised the activity rather than its intensity, and several did not state in their abstracts whether test participants were excluded from training. Two contrasts are nevertheless informative. Where studies in Table 1 compared evaluation designs or moved to new datasets, their authors reported that performance fell, in some cases sharply, which is consistent with the leakage effects described in the literature survey, one lightweight model, for example, kept useful class ranking on PAMAP2 but lost macro F1 compared with its own dataset, and leakage between windows inflated stress-detection scores by tens of points. Studies that classified intensity against measured or perceived effort reported accuracies above 90%, in line with the present results. The macro-F1 of about 0.95 obtained here for three intensity classes in unseen participants is in the same range as the best subject-independent results for activity classes, but intensity is a simpler target, and the advantage should not be read as superiority of method. More important is what the present secondary analysis adds that the earlier work did not: per-person coverage of calibrated prediction sets, a sensor ablation built around a single wrist device, and a test of label-free heart-rate personalisation.

4.3 Sensor Ablation

All three inertial sensors together gave the best result (0.954), and adding heart rate to them did not improve it further (difference -0.006, p = 0.64, Figure 3). With three body locations, the movement of arms, trunk and legs already separates the activities well, and heart rate, which lags behind changes in activity by tens of seconds, has little left to add. No single sensor matched the combination. Chest and ankle each came in at roughly 0.90. The wrist sensor alone reached 0.867, below all inertial sensors combined in every one of the eight participants (p = 0.008), and its main weakness was vigorous activity (sensitivity 0.75). This agrees with the moderate accuracy reported for wrist cut-points (Clanchy *et al.*, 2024) and with the poor individual-level agreement between wrist-based intensity methods in large cohorts.

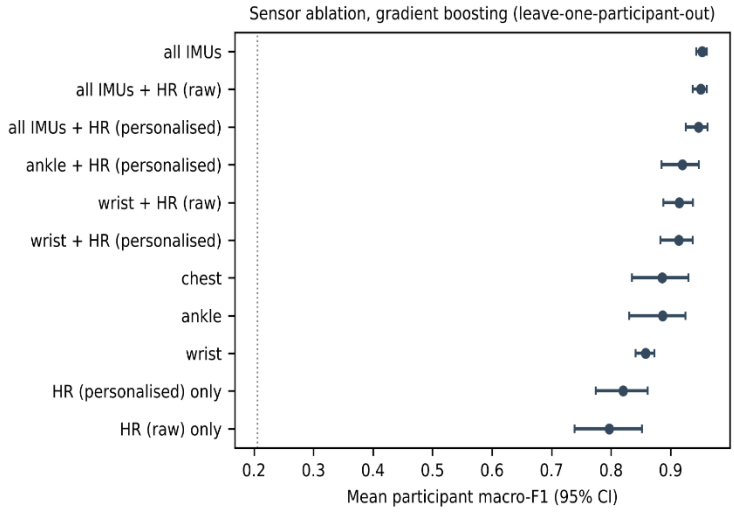


Figure 3. Mean participant macro-F1 (95% CI) for gradient boosting with each sensor set. The dotted line marks the majority-class reference.

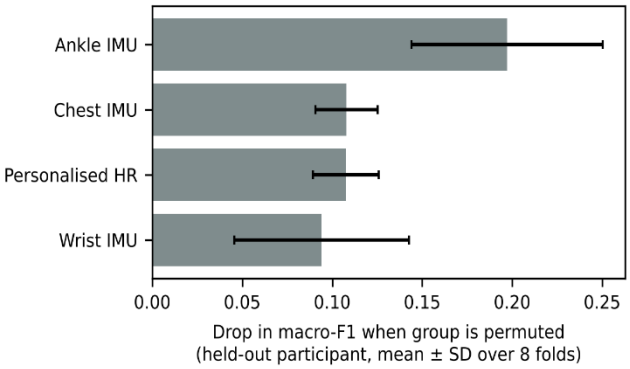


Figure 4. Grouped permutation importance on held-out participants for gradient boosting with all sensors and personalised heart rate.

Adding heart rate improved wrist-only classification. With heart rate added, macro-F1 rose to 0.919 and vigorous sensitivity to 0.87. Seven of the 8 participants improved (mean +0.056, p = 0.016). At the ankle, heart rate helped in much the same way. Heart rate alone, without any motion sensor, reached 0.838, which shows that physiological load carries a great deal of intensity information by itself, as earlier work using heart rate, heart-rate reserve or other cardiac signals, alone or with acceleration, has also found.

The grouped permutation analysis agreed with the ablation: when all sensors were available, shuffling the ankle features reduced macro-F1 most (0.20), followed by the chest (0.11), personalised heart rate (0.11) and the wrist (0.09) (Figure 4). The ankle records the impact and cadence of locomotion, which separate walking, stair climbing, running and rope jumping, whereas the wrist mostly records arm movement, which is common to many activities of different intensity. The value of a sensor has likewise been found to depend on the activity and the mounting site in subject-independent recognition.

The chest sensor occupied a middle position. It records trunk movement and the vertical bounce of the body during locomotion, and it was the best single sensor for recognising low-intensity windows. A chest-worn device has a practical advantage as well: heart-rate straps are already worn on the chest, so a strap that also contained an inertial sensor would combine the second-best motion signal with heart rate in a single device. The variation between participants was, however, larger for the chest and ankle sensors than for the wrist, which suggests that their placement and the way people move their trunk and legs differ more between individuals than wrist movement does.

For physical education and sport settings the practical message is clear. Most people will wear a single wrist device rather than three sensors. A wrist accelerometer alone confuses vigorous with moderate effort, because the arm moves in similar ways during brisk walking, stair climbing and running. Heart rate fills in the physiological information that the wrist lacks, and a wrist device recording both signals ends up within about 0.04 of the full three-sensor set-up. Most consumer watches already carry a wrist motion sensor and an optical heart-rate sensor, and the accuracy of their heart-rate readings has been reviewed in detail (Doherty *et al.*, 2024). For monitoring intensity in schools, clubs and recreational sport, this is the combination people will actually use. It follows that intensity estimates based only on wrist accelerometry, which remain common in research, are likely to miss some vigorous activity.

Once motion sensors were present, the label-free personalisation was no better than raw heart rate. With the wrist, macro-F1 was 0.919 against 0.922 (p = 0.95); with all sensors it was 0.951 against 0.952. Used on its own, personalised heart rate produced a higher pooled macro-F1 (0.838 against 0.819), yet the gain did not hold across participants (3 of 8 participants, p = 0.84). The hypothesis that personalisation would help therefore found no support.

The small and relatively homogeneous sample may explain this result. Eight young adults of similar fitness also have similar resting and maximal heart rates. Raw heart rate is then already on something close to a shared scale, and the model can take up whatever offsets remain. In addition, age-predicted maximum heart rate hardly changes over such a narrow age range, so the reserve calculation alters the scale only a little. The picture should be different in more mixed groups, for instance children and older adults in one class, athletes alongside sedentary people, or anyone on heart-rate-lowering medication. There the same raw

heart rate can mean very different levels of effort, and relative measures ought to count for more; in a treadmill study, heart-rate reserve models improved energy expenditure estimates across sex, body size and resting heart rate (Seo *et al.*, 2025). The label-free approach should be tried in those groups next. Its practical appeal remains: it needs nothing beyond a period of ordinary wear, and the estimated resting values tracked the documented ones closely, so the method is at least safe to use even where it adds little.

4.4 Prediction-Set Coverage for New Participants

Naive calibration on training data badly overstated certainty. Its sets contained the true class in only 73% of windows instead of the intended 90%, and in 50% of vigorous windows (Table 3, Figure 5). Seven of the eight participants had coverage below 0.85. The reason is that a model is far more confident on people it was trained on than on a new person, so thresholds learned in-sample are too tight. Calibrating on out-of-participant predictions restored overall coverage to 0.899, essentially the target, for the full system and to 0.896 for the wrist with heart rate. This is the central methodological result of the study: in wearable classification, conformal calibration must respect the grouping of data by person, just as validation must. The coverage guarantee of conformal prediction rests on exchangeability between calibration and new cases, and in-sample scores violate it, a similar loss of calibration under leakage was reported for wearable stress detection. A practitioner who is told that a label is 90% reliable should be able to rely on that statement for a new pupil or athlete, and only person-level calibration delivers it.

Overall coverage, however, hid a shortfall in the class that matters most for activity guidelines.

Table 3. Coverage of conformal prediction sets at a 90% target (gradient boosting)

Sensors and calibration	Overall	Low	Moderate	Vigorous	Lowest participant	Mean set size	Empty sets (%)
All sensors, naive	0.734	0.758	0.778	0.505	0.617	0.74	26.4
All sensors, subject-level marginal	0.899	0.914	0.925	0.766	0.786	0.92	8.2
All sensors, subject-level class-conditional	0.902	0.900	0.903	0.905	0.772	0.92	7.9
Wrist + HR, naive	0.729	0.800	0.702	0.598	0.649	0.74	25.9
Wrist + HR, subject-level marginal	0.896	0.919	0.898	0.818	0.751	0.94	6.0
Wrist + HR, subject-level class-conditional	0.896	0.895	0.893	0.913	0.742	0.97	3.9
Wrist, subject-level class-conditional	0.901	0.893	0.905	0.917	0.846	1.15	1.6

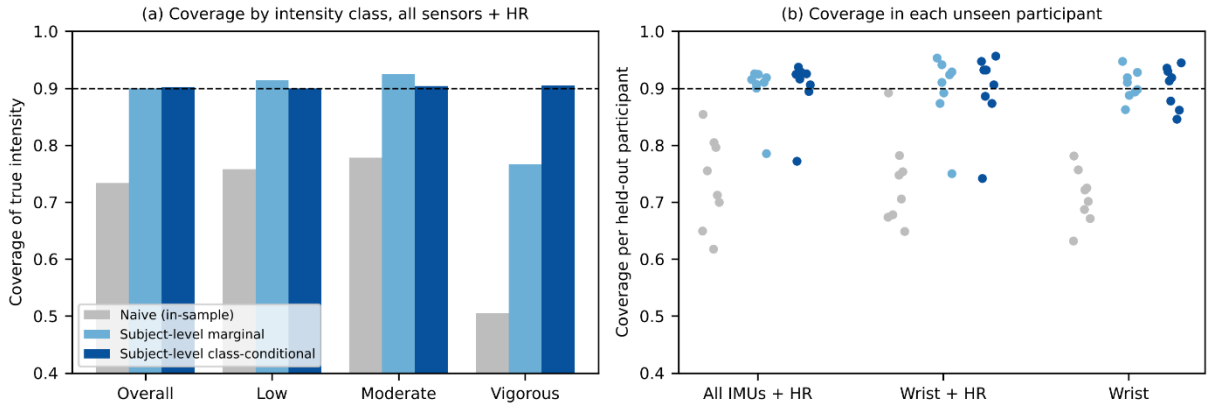


Figure 5. Coverage of the true intensity by class **(a)** and in each held-out participant **(b)** for naive, subject-level marginal and subject-level class-conditional conformal sets. The dashed line marks the 90% target.

With a single threshold, vigorous windows were covered only 77% of the time, while low and moderate windows were slightly over-covered. A single threshold is dominated by the more numerous low and moderate windows, and the model is least certain about vigorous ones. Setting a separate threshold for each class brought vigorous coverage to 90%, with low and moderate coverage close to 90% and no increase in average set size for the full system. Without heart rate, the wrist's class-conditional sets grew larger and more often held two classes. This is what one would expect from a weaker sensor: in some windows the method admits that it cannot separate moderate from vigorous effort, instead of guessing.

When all sensors were used, nearly every non-empty set held just one class, while about 8% of sets were empty. An empty set means that no class was plausible enough to include, which in practice should be treated as a flag that the window is unusual, for example a transition between activities or a sensor artefact, rather than given a label. The guarantee is an average over people, and coverage in individual participants ranged from 0.77 upwards, with one participant below 0.85. Eight participants are too few to estimate person-level coverage precisely, and larger samples are needed before claims about the worst-case person can be made. Extensions of conformal methods that control other risks (Angelopoulos, 2026), and the broader set of uncertainty methods used in clinical machine learning, could be explored for per-person guarantees.

4.5 Sensitivity Analyses

Windowing choices made little difference (Table 4). Overlapping 5 s windows changed macro-F1 by less than 0.02, and 2 s windows lowered it by 0.01 to 0.03, most for the wrist alone, which gains from a longer view of the movement. The order of sensor sets did not change, and raw and personalised heart rate stayed within 0.01 of each other throughout. Recoding cycling as vigorous or vacuum cleaning as low also left the comparisons intact. Recoding stair ascent as moderate had a larger effect. The vigorous class then held only running and rope jumping, some participants had few or no vigorous windows, and participant-level F1 became unstable (for example, a 95% interval of 0.75 to 0.94 for the wrist with heart rate). Under that coding the advantage of adding heart rate to the wrist was small in pooled macro-F1 (0.928 against 0.902) and absent at participant level, so the wrist-plus-heart-rate result depends partly on stair climbing being treated as vigorous, as the Compendium does.

4.6 Practical Implications

Activity guidelines are written around moderate and vigorous minutes, so mistakes at the moderate-vigorous boundary are the ones that count most. Three recommendations can be offered to people who develop or use wearable intensity tools. First, evaluation should always hold out whole people, and results should be reported per person as well as pooled.

Table 4. Sensitivity of gradient-boosting macro-F1 to windowing and to intensity coding

Analysis	All IMUs + HR (pers.)	All IMUs + HR (raw)	Wrist + HR (pers.)	Wrist + HR (raw)	Wrist
5 s windows, no overlap (primary)	0.951	0.952	0.919	0.922	0.867
5 s windows, 50% overlap	0.959	0.963	0.925	0.929	0.885
2 s windows, no overlap	0.942	0.947	0.901	0.903	0.839
Cycling coded vigorous (7.0 MET)	0.951	0.954	0.913	0.910	0.882
Stair ascent coded moderate (4.5 MET)	0.954	0.941	0.928	0.935	0.902
Vacuum cleaning coded low (2.9 MET)	0.962	0.965	0.918	0.921	0.863

Second, a wrist device should include heart rate if vigorous activity is to be detected reliably, motion alone is not enough at the wrist. Third, any statement of confidence attached to a classification should be calibrated on people outside the training set, and separately for each intensity class if the rarest class matters. Conformal prediction makes this straightforward: it needs no change to the underlying model, only a different way of collecting calibration scores, and it produces outputs that a teacher or coach can understand, such as "moderate", "vigorous" or "moderate or vigorous".

Take a physical education lesson as an example. A teacher could use a wrist device with heart rate to estimate how much of the lesson each pupil spent in moderate-to-vigorous activity, and could look at the prediction sets to see which of those estimates are uncertain. Windows whose sets are empty or contain two classes could then be left out of summaries, or reported apart, so that any conclusion depends only on windows the system is sure about. This is more honest than reporting a single number with no indication of its reliability, and it requires only modest computation that consumer devices can already support.

The cost of subject-level calibration lies in development, not in use. In development, out-of-participant scores require one refit of the model per training participant, so training time grows in proportion to the number of participants. After the thresholds have been set, handling a new window takes nothing more than comparing its predicted probabilities with three stored numbers. So a manufacturer or a research group could calibrate once, on a varied development sample, and supply the thresholds together with the model. The thresholds should be re-estimated whenever the model, the sensors or the target population change, because coverage depends on all three.

4.7 Limitations

The sample is small and homogeneous: eight young adults, seven of them men, performing a scripted laboratory protocol. Each activity was given its Compendium value instead of a measured energy expenditure, as obtained in studies using portable calorimetry; anyone who performed an activity harder or more gently than average therefore received the same label. Because the vigorous class was the smallest, its estimates are the least certain. The resting heart rate was estimated from the same recording that was classified, in practice it would come from earlier wear. The protocol activities lasted for long bouts. In free-living conditions, where bouts are short and mixed and transitions frequent, performance will probably be lower. Finally, the conformal guarantee is marginal over people and does not promise coverage for every individual, which matters when decisions are made for a single person. Only hand-crafted features and three standard learners were tested. Deep models trained on raw signals could be more accurate, though the conclusions about calibration hold for any classifier that outputs probabilities.

5. Conclusion

On people the models had never seen, wrist, chest and ankle units together classified activity intensity with a macro-F1 of close to 0.95. A wrist sensor with heart rate reached about 0.92 and clearly beat the wrist-worn sensor alone, largely because heart rate helped with vigorous effort. When all sensors were available, the ankle contributed the most information. Scaling heart rate to a label-free personal baseline added no accuracy in this fairly uniform sample, even though the baseline agreed well with the documented resting values. Prediction sets calibrated on training data gave false confidence, covering only about three quarters of cases against a 90% target, whereas sets calibrated on unseen participants achieved their stated coverage, and class-conditional calibration was needed to protect the vigorous class. Wearable intensity tools for physical education, sport and health monitoring should report subject-independent performance and attach uncertainty that has been calibrated in new people. The next step is to test these methods in larger and more varied groups, under free-living conditions and against measured energy expenditure.

References

Angelopoulos, A.N. (2026). Conformal risk control for non-monotonic losses. *Philosophical Transactions of the Royal Society,* 27(384), 20250068. [DOI] [PubMed]

Beavers, D.L., Chung, E.H. (2022). Wearables in sports cardiology. *Clinics in Sports Medicine,* 41(3), 405-423. [DOI] [PubMed]

Bernett, J., Blumenthal, D.B., Grimm, D.G., Haselbeck, F., Joeres, R., Kalinina, O.V., List, M. (2024). Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, 21(8), 1444-1453. [DOI] [PubMed]

Biró, A., Szilágyi, S. M., Szilágyi, L., Martín-Martín, J., Cuesta-Vargas, A.I. (2023). Machine learning on prediction of relative physical activity intensity using medical radar sensor and 3D accelerometer. *Sensors*, *23*(7), 3595. [DOI] [PubMed]

Cao, K., Wang, M. (2023). Human behavior recognition based on sparse transformer with channel attention mechanism. *Frontiers in physiology*, 14, 1239453. [DOI] [PubMed]

Cheng, D., Zhang, L., Bu, C., Wang, X., Wu, H., Song, A. (2023). ProtoHAR: Prototype guided personalized federated learning for human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, *27*(8), 3900-3911. [DOI] [PubMed]

Clanchy, K., Stanfield, M., Smits, E., Liimatainen, J., Ritchie, C. (2024). Calibration and validation of physical behaviour cut-points using wrist-worn ActiGraphs for children and adolescents: A systematic review. *Journal of Science and Medicine in Sport*, *27*(2), 92-104. [DOI] [PubMed]

Doherty, C., Baldwin, M., Keogh, A., Caulfield, B., & Argent, R. (2024). Keeping Pace with Wearables: A Living Umbrella Review of Systematic Reviews Evaluating the Accuracy of Consumer Wearable Technologies in Health Measurement: C. Doherty et al. *Sports Medicine*, 54(11), 2907-2926. [DOI] [PubMed]

Elkhouly, A., Asghar, U., Raj, G. (2026). Explainable Machine Learning for Human Activity Recognition Using Auxetic cTPU Knee-Worn Sensors. *Sensors*, 26(17), 5548. [DOI] [PubMed]

Franklin, B. A., Eijsvogels, T. M., Pandey, A., Quindry, J., & Toth, P. P. (2022). Physical activity, cardiorespiratory fitness, and cardiovascular health: A clinical practice statement of the American Society for Preventive Cardiology Part II: Physical activity, cardiorespiratory fitness, minimum and goal intensities for exercise training, prescriptive methods, and special patient populations. *American Journal of Preventive Cardiology*, 12, 100425. [DOI] [PubMed]

Hedge, E.T., Amelard, R., Hughson, R.L. (2023). Prediction of oxygen uptake kinetics during heavy-intensity cycling exercise by machine learning analysis. *Journal of Applied Physiology*, 134(6), 1530-1536. [DOI] [PubMed]

Herrmann, S.D., Willis, E.A., Ainsworth, B.E., Barreira, T.V., Hastert, M., Kracht, C.L., Schuna Jr, J.M., Cai, Z., Quan, M., Tudor-Locke, C., Whitt-Glover, M.C., Jacobs, D.R.J, (2024). 2024 Adult Compendium of Physical Activities: A third update of the energy costs of human activities. Journal of Sport and Health Science, 13(1), 6-12. [DOI] [PubMed]

Hossen, M. A., & Abas, P. E. (2025). Machine learning for human activity recognition: State-of-the-art techniques and emerging trends. Journal of Imaging, 11(3), 91. [DOI] [PubMed]

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. [DOI] [PubMed]

Kim, Y.W., Cho, W.H., Kim, K.S., Lee, S. (2022). Inertial-measurement-unit-based novel human activity recognition algorithm using conformer. *Sensors*, *22*(10), 3932. [DOI] [PubMed]

Koffman, L., Zhang, Y., Harezlak, J., Crainiceanu, C., Leroux, A. (2023). Fingerprinting walking using wrist-worn accelerometers. *Gait & Posture*, 103, 92-98. [DOI] [PubMed]

Letts, E., Jakubowski, J.S., King-Dowling, S., Clevenger, K., Kobsar, D., Obeid, J. (2024). Accelerometer techniques for capturing human movement validated against direct observation: a scoping review. *Physiological Measurement*, *45*(7), 07TR01. [DOI] [PubMed]

Li, Y., Wang, L. (2022). Human activity recognition based on residual network and BiLSTM. *Sensors*, *22*(2), 635. [DOI] [PubMed]

Loftus, T.J., Shickel, B., Ruppert, M. M., Balch, J. A., Ozrazgat-Baslanti, T., Tighe, P. J., Efron, P.A., Hogan, W.R., Rashidi, P., Upchurch, G.R., Bihorac, A. (2022). Uncertainty-aware deep learning in healthcare: a scoping review. PLOS digital health, 1(8), e0000085. [DOI] [PubMed]

Lopes, J.M., Figueiredo, J., Fonseca, P., Cerqueira, J.J., Vilas-Boas, J.P., Santos, C.P. (2022). Deep learning-based energy expenditure estimation in assisted and non-assisted gait using inertial, EMG, and heart rate wearable sensors. *Sensors*, *22*(20), 7913. [DOI] [PubMed]

Ng, J.Y., Zhang, J.H., Hui, S.S., Jiang, G., Yau, F., Cheng, J., Ha, A.S. (2024). Development of a multi-wear-site, deep learning-based physical activity intensity classification algorithm using raw acceleration data. *Plos one*, 19(3), e0299295. [DOI] [PubMed]

Pang, H., Zheng, L., Fang, H. (2024). Cross-attention enhanced pyramid multi-scale networks for sensor-based human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, *28*(5), 2733-2744. [DOI] [PubMed]

Papamentzelopoulos, S., Nikoletseas, S. (2026). Trustworthy Stress Detection from Wearable Sensor Signals for Preventive Anxiety Monitoring: Subject-Independent Accuracy and Calibrated Confidence with the Autonomic-Branch Quantum-Graph Network. *Sensors*, 26(20), 6425. [DOI] [PubMed]

Reiss, A., Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, *IEEE*, UK. [DOI]

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature communications*, *15*(1), 1829. [DOI] [PubMed]

Seo, Y., Lee, Y., Lee, D.T. (2025). Comparing Heart Rate and Heart Rate Reserve for Accurate Energy Expenditure Prediction Against Direct Measurement. *International Journal of Environmental Research and Public Health*, 22(10), 1539. [DOI] [PubMed]

Shang, M., De Raedt, W., Varon, C., Vanrumste, B. (2022). Are gyroscopes an added value in leave-one-subject-out activity recognition with IMUs?. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, United Kingdom. [DOI] [PubMed]

Song, Y., Wang, Z., Wang, H., Sun, G., Gong, B., Zhang, F. (2025). Interpretable deep learning for personalized energy expenditure prediction using ECG and acceleration signals in incremental exercise. *Scientific Reports*, 15(1), 36277. [DOI] [PubMed]

Souaifi, M., Dhahbi, W., Jebabli, N., Ceylan, H.İ., Boujabli, M., Muntean, R.I., Dergaa, I. (2025). Artificial intelligence in sports biomechanics: A scoping review on wearable technology, motion analysis, and injury prevention. *Bioengineering*, *12*(8), 887. [DOI] [PubMed]

Stampfler, T., Elgendi, M., Fletcher, R.R., Menon, C. (2023). The use of deep learning for smartphone-based human activity recognition. *Frontiers in public health*, 11, 1086671. [DOI] [PubMed]

Tanaka, H., Monahan, K.D., Seals, D.R. (2001). Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology,* 37(1), 153-156. [DOI] [PubMed]

Tang, Y., Zhang, L., Wu, H., He, J., Song, A. (2022). Dual-branch interactive networks on multichannel time series for human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(10), 5223-5234. [DOI] [PubMed]

Tao, K., Meng, K., Gao, B., Yang, J., Qiu, J. (2025). Triple-E Principle: Leveraging Occam's Razor for Dance Energy Expenditure Estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 33, 2088-2096. [DOI] [PubMed]

Trost, S.G., Brookes, D.S., Ahmadi, M.N. (2022). Evaluation of wrist accelerometer cut-points for classifying physical activity intensity in youth. *Frontiers in digital health*, 4, 884307. [DOI] [PubMed]

Vazquez, J., Facelli, J.C. (2022). Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3), 241-252. [DOI] [PubMed]

Wei, L., Ahmadi, M.N., Biswas, R.K., Trost, S.G., Stamatakis, E. (2024a). Comparing Cadence vs. Machine Learning-Based Physical Activity Intensity Classifications: Variations in the Associations of Physical Activity with Mortality. *Scandinavian Journal of Medicine & Science in Sports*, *34*(9), e14719. [DOI] [PubMed]

Wei, L., Ahmadi, M.N., Hamer, M., Blodgett, J.M., Small, S., Trost, S., Stamatakis, E. (2024b). Comparing cadence-based and machine learning based estimates for physical activity intensity classification: the UK Biobank. *Journal of Science and Medicine in Sport*, 27(8), 551-556. [DOI] [PubMed]

Xu, Y., Li, J., Huang, Z., Cheng, T., Chen, C., Tan, B., Tan, Z., Chen, H., Zhou, Y. (2026). DeM-FCN: an ultra-lightweight and purely convolutional framework for edge-native human activity recognition in wearable fitness tracking. *Frontiers in Sports and Active Living,* 8, 1858408. [DOI] [PubMed]

Yan, X., Yang, J., Dai, J., Tao, K. (2025). Energy-Expenditure Estimation During Aerobic Training Sessions for Badminton Players. *Sensors*, 25(19), 6257. [DOI] [PubMed]

Yu, J., Zhang, L., Cheng, D., Huang, W., Wu, H., Song, A. (2024). Improving human activity recognition with wearable sensors through bee: Leveraging early exit and gradient boosting. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32, 3452-3464. [DOI] [PubMed]

Zhang, C., Cao, K., Lu, L., Deng, T. (2022a). A multi-scale feature extraction fusion model for human activity recognition. *Scientific Reports*, 12(1), 20620. [DOI] [PubMed]

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., Alshurafa, N. (2022b). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4), 1476. [DOI] [PubMed]

Zhu, T., Dong, Y., Zhou, Y., Zhu, C., Cao, L. (2025). Cross-domain human activity recognition via domain adaptation and fused attention. *IEEE Journal of Biomedical and Health Informatics*, *29*(8), 5394-5404. [DOI] [PubMed]

Author Contributions Statement

Archibald Danquah-Amoah: Conceptualization, Methodology, Investigation, Writing – original draft. Viena Sembaiyan Gunasekaran: Formal analysis, Validation, Writing – review & editing. Alex Abrokwa Fosuhene: Data curation, Visualization, Writing – review & editing. Mensah Sitti: Conceptualization, Supervision, Project administration, Writing – review & editing. All the authors read and approved the final version of this manuscript.

Acknowledgments

The authors thank A. Reiss and D. Stricker for making the PAMAP2 dataset publicly available.

Other Ethics Statements

This secondary analysis used a public, de-identified dataset collected with informed consent in the original study, the present secondary analysis did not recruit or assess participants. The PAMAP2 dataset is available from the UCI Machine Learning Repository and the analysis code is available from the author. Generative AI tools assisted with drafting and proofreading. The authors reviewed and verified the manuscript's content, including the reported results, and take full responsibility for the published work.

Funding Source

This study received no external funding.

Does this article pass screening for similarity?

Yes

About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.