

Article

Interpretable Machine Learning Distinguishes Correct from Incorrect Rehabilitation Movement with Fewer Wearable Sensors

Georgios Bouchouras ^{1,2,*} , Georgios Sofianidis ² , Evangelos Kontaxakis ²  and Konstantinos Kotis ^{1,*} 

¹ Intelligent Systems Lab, Department of Cultural Technology and Communication, University of the Aegean, 81100 Mytilene, Greece

² Center of Interdisciplinary Education and Research in Rehabilitation, Metropolitan College, 54625 Thessaloniki, Greece; gsofianidis@mitropolitiko.edu.gr (G.S.); vkontaxakis@mitropolitiko.edu.gr (E.K.)

* Correspondence: gbouchouras@mitropolitiko.edu.gr (G.B.); kotis@aegean.gr (K.K.)

Featured Application

This study shows that ML-based wearable movement assessment can distinguish correct from incorrect movement, explain why the movement is wrong, remain useful with fewer Inertial Measurement Units (IMUs), and support a proof-of-concept path toward corrective feedback.

Abstract

Wearable systems for rehabilitation monitoring often rely on complex sensor configurations and produce outputs that are difficult to interpret. This limits their practical use. This study investigates whether movement quality assessment can be achieved accurately and transparently using a reduced set of signals, meaning fewer body segment inertial measurement units and therefore smaller task-specific feature tables. Using wearable sensor data from lower-limb rehabilitation tasks performed under correct and intentionally erroneous conditions, we extracted a small set of rotation-based features and evaluated them within a supervised machine learning framework. The results show that these features can reliably distinguish correct from incorrect movement, with classification accuracy around 0.70, while preserving clear biomechanical interpretation. Reduced sensor configurations retained, and in some cases improved, performance, with balanced accuracy reaching 0.947 and 0.917 in the examined tasks. A proof-of-concept real-time formulation further showed that movement deviations can be detected early within repetitions, while limiting false feedback on correct executions to approximately 9%. Overall, the findings show that movement quality assessment can be achieved with minimal sensing, while also supporting early error detection and practical feedback. These properties are relevant to wearable rehabilitation systems, including internet of things applications that depend on efficient sensing, interpretable analysis, and timely feedback.



Academic Editor: Francesco Zirilli

Received: 17 April 2026

Revised: 29 April 2026

Accepted: 5 May 2026

Published: 7 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: movement quality assessment; wearable sensors; inertial measurement units; IMU; rehabilitation monitoring; early error detection; interpretable features

1. Introduction

Assessing movement quality is an important task in rehabilitation monitoring. Clinicians need not only to detect altered movement, but also to understand how and why it deviates from the intended pattern. Wearable sensing systems can help address this need. They combine wearable sensors with computational analysis of collected sensor data to support

clinician review, patient feedback, and remote follow-up. Their value is especially clear in home-based rehabilitation, where frequent in-person assessment is difficult and low-burden monitoring is needed [1,2]. Recent studies have shown growing interest in combining inertial sensing with machine learning (ML) to extend rehabilitation assessment beyond the clinic [3,4].

These developments are also relevant to applications of artificial intelligence (AI) in Internet of Things (IoT) environments. In such settings, wearable devices can support continuous monitoring and data-driven decision-making. However, practical deployment depends on more than predictive accuracy alone. Sensor complexity, usability, and interpretability also matter. These factors influence user acceptance and long-term use in healthcare applications [5].

Despite the progress in the field, two main challenges remain. The first is sensor burden. Multi-sensor configurations can capture detailed movement information, but they are harder to use in routine care or home settings because they require extra configuration and deployment effort. Simpler configurations are easier to deploy, but they may miss movement variations that are distributed across several body segments. The second challenge is interpretability. Many models can detect movement deviations, but they do not clearly indicate (explain to decision makers) what has actually changed. This limits their value for clinical interpretation and patient-facing feedback [6]. In practice, useful systems should reduce sensing complexity while still producing meaningful and understandable outputs.

In this study, we investigate whether movement quality assessment can be achieved with a reduced set of signals derived from wearable sensors/Inertial Measurement Units (IMUs). Here, a “reduced set of signals” refers primarily to reducing the number of instrumented body segments and therefore reducing the corresponding segment–feature channels available to the classifier, rather than to a separate feature-selection stage within a fixed sensor configuration. Using data from lower-limb rehabilitation exercises performed under correct and instructed erroneous conditions, we apply an ML framework to analyse and understand movement discrimination, sensor reduction, interpretability, and feedback-oriented monitoring. We also extend the analysis to a proof-of-concept setting for early error detection and feedback based on short time windows. The proposed approach combines efficient IMU-based sensing, ML, and readable feature-based outputs to advance practical wearable systems for rehabilitation.

The objectives of this study are the following:

1. Distinguish correct from incorrect (altered) movement signals using simple and interpretable measures;
2. Provide explainability output regarding the reasons the movement is altered, using the same set of measures;
3. Identify smaller sensor sets that are more suitable for wearable system deployment;
4. Test whether this approach can support a simulated real-time corrective-feedback setting.

The paper is structured as follows: Section 2 (Materials and Methods) presents the dataset, preprocessing pipeline, feature extraction, subject-level aggregation, and the statistical and ML analyses used in this study. Section 3 (Results) reports the main findings, including effect size analysis, classification performance, and sensitivity analyses. Section 4 (Discussion) interprets the results, highlights the clinical and methodological implications, and addresses the limitations of the study. Finally, Section 5 (Conclusions) summarizes the main contributions and outlines directions for future work.

2. Materials and Methods

2.1. Dataset and Tasks

We analyzed GAITEX, a publicly available dataset of impaired gait and rehabilitation exercises recorded using inertial and optical sensors [7,8]. We focused on two unilateral lower-limb tasks derived from Xsens (wearable motion-capture system [9]) body segment orientations. The repeated dynamic task (RD) contained the labels *rd_correct*, *rd_pronation*, *rd_supination*, and *rd_toes*. The right-leg stance task (RGS) contained *rgs_correct*, *rgs_abduction*, *rgs_flexion*, and *rgs_stork*. After preprocessing and subject-level aggregation, 19 subjects were available for RD and 18 for RGS.

The two rehabilitation tasks analyzed in this study are illustrated in Figure 1. Both exercises are performed in a standing position and were designed to probe lower-limb control under clinically relevant conditions. The repeated dynamic task (RD) consists of a cyclic lower-limb movement emphasizing controlled ankle and foot motion. The right-leg stance task (RGS) focuses on postural stability while balancing on the right leg and controlling compensatory body movements. For each task, the dataset provides a standardized start position, one correct execution, and three representative erroneous movement patterns commonly observed in individuals with foot drop. These task variants form the basis for the movement quality analyses performed in this study. The instrumented body segments used in the analyses are summarized in Figure 2. This schematic indicates the main IMU-derived segment channels considered for the subject-level and window-level analyses, including lower-limb, pelvis, sternum, and right-hand segments. The overall analysis workflow is summarized in Figure 3. After defining the rehabilitation tasks and sensor-derived segment channels, the pipeline proceeds from preprocessing and interpretable feature extraction to subject-level classification, sensor subset evaluation, readable movement summaries, and the proof-of-concept feedback-oriented analysis.

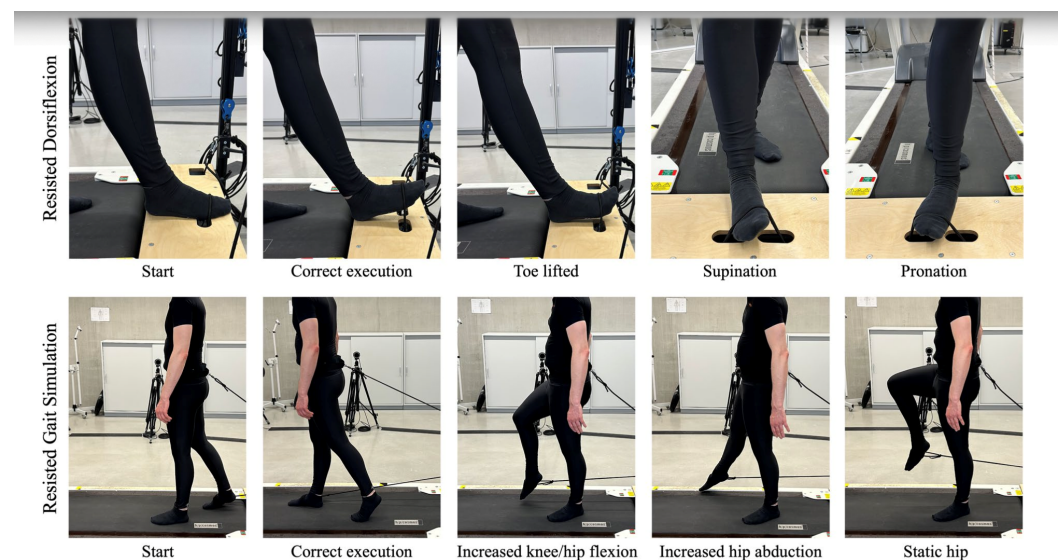


Figure 1. Overview of the two rehabilitation tasks and their execution variants, adapted from Spilz et al. (2025) [7]. Both exercises are performed in a standing position. The repeated dynamic task (RD) is shown in the upper row, and the right-leg stance task (RGS) in the lower row. For each task, the first image illustrates the standardized start position, followed by the correct execution and three representative erroneous movement patterns commonly observed in individuals with foot drop.

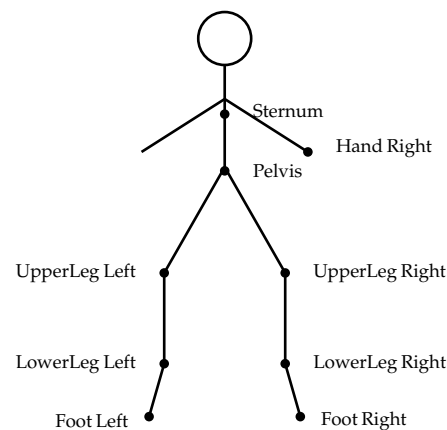


Figure 2. Simple schematic of the main body-mounted IMU placements used in the subject-level and window-level analyses. The analyzed segment set comprised the feet, lower legs, upper legs, pelvis, sternum, and right hand. RD also included one additional non-body segment channel (Prop1) in the full unconstrained configuration when available in the processed feature table.

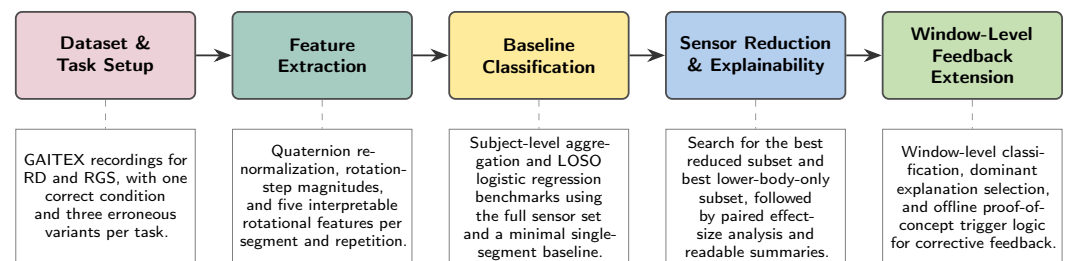


Figure 3. Overview of the proposed workflow. The study starts from GAITEX rehabilitation tasks, extracts interpretable rotational IMU features, evaluates subject-level correct-versus-incorrect movement classification, identifies task-specific reduced sensor configurations and readable explanations, and extends the analysis to an offline proof-of-concept window-level feedback setting.

2.2. Feature Extraction

For each repetition and each instrumented segment, the Xsens orientation stream was treated as a quaternion time series $\mathbf{q}_t = (w_t, x_t, y_t, z_t)$, which was first re-normalized so that $\|\mathbf{q}_t\| = 1$ at every sample. Consecutive orientations were then converted into relative rotations using

$$\mathbf{q}_t^{\text{rel}} = \text{conj}(\mathbf{q}_{t-1}) \otimes \mathbf{q}_t, \quad (1)$$

where $\text{conj}(\cdot)$ is quaternion conjugation and $\otimes$ is the Hamilton product. The corresponding rotation-step magnitude was computed as

$$\theta_t = 2 \arccos(|w_t^{\text{rel}}|), \quad \theta_t \in [0, \pi], \quad (2)$$

with w_t^{rel} denoting the scalar component of the relative quaternion. This yields a scalar angular displacement between two consecutive orientation samples. Given the inter-sample time difference $\Delta t_t = t_t - t_{t-1}$, angular-speed magnitude was then defined as

$$\omega_t = \frac{\theta_t}{\Delta t_t}, \quad (3)$$

after excluding non-positive time differences. Angular acceleration was estimated from the finite difference of angular speed using

$$\alpha_t = \frac{\omega_t - \omega_{t-1}}{\text{median}(\Delta t)}, \quad (4)$$

using the median within-repetition sampling interval for robustness.

From these time series, we extracted five interpretable rotational features for each segment and repetition:

- Mean angular speed:

$$\text{mean_speed} = \frac{1}{N} \sum_{t=1}^N \omega_t \quad (5)$$

- RMS angular speed:

$$\text{rms_speed} = \sqrt{\frac{1}{N} \sum_{t=1}^N \omega_t^2} \quad (6)$$

- Peak angular speed:

$$\text{peak_speed} = \max_t \omega_t \quad (7)$$

- RMS angular acceleration:

$$\text{rms_accel} = \sqrt{\frac{1}{M} \sum_{t=1}^M \alpha_t^2} \quad (8)$$

- Rotational range:

$$\text{rot_range} = \max_t \theta_t - \min_t \theta_t. \quad (9)$$

Here, N is the number of valid angular-speed samples within the repetition and M is the number of valid angular-acceleration samples. We intentionally kept the feature set simple. No latent embeddings, spectral transforms, or deep learned features were used. This served two purposes. First, it allowed us to test whether straightforward kinematic summaries were already sufficient for useful discrimination between correct and incorrect movement. Second, it supported later interpretation, because model behavior could be related directly to segment-level biomechanical quantities rather than to hidden latent variables [4,10]. Repetition-level features were then averaged to the subject level separately for each subject, task, condition, and segment. Each subject-level sample therefore represents the average feature profile for one subject under one task condition.

2.3. ML Pipeline

The ML component of this study was implemented as a supervised pipeline operating on the interpretable segment-level feature tables. For the binary movement quality task, the target label was *correct* versus *incorrect* movement, where the incorrect class pooled the three instructed erroneous variants within each task. For the retained multiclass benchmark, the target label was the four-condition task label consisting of the correct condition plus the three specific erroneous variants.

We used logistic regression as the classifier because it is appropriate for small tabular datasets, yields directly auditable linear decision functions, and keeps the relation between learned model behavior and the underlying feature space transparent. Unless stated otherwise, each classification pipeline used median imputation for missing values, z-score standardization, and logistic regression with the `liblinear` solver and `max_iter` = 5000. Performance was estimated with leave-one-subject-out (LOSO) cross-validation so that each test fold contained data from a subject not used for model fitting.

For binary movement quality classification, we report accuracy and balanced accuracy. Accuracy is the proportion of correctly classified held-out subject-level samples. Balanced accuracy is the arithmetic mean of sensitivity and specificity in the binary correct-versus-

incorrect task, computed per held-out LOSO fold and then averaged across folds. For the retained multiclass benchmark, we report accuracy. Classification was performed using a logistic regression model, whereas interpretation was derived directly from the same kinematic features used for classification.

2.4. Baseline Benchmarks

The baseline analysis began with subject-level benchmarking using the full sensor set. Binary discrimination between correct and incorrect movement was evaluated with the logistic regression pipeline described above under LOSO cross-validation. This full sensor set benchmark was compared with a minimal single-segment baseline, defined as the simplest one-segment representation used as a low-information reference point.

We also retained a multiclass LOSO benchmark using the same model family to test whether the same interpretable feature space preserved information about the specific erroneous variant type. These baseline analyses establish the first methodological question clearly: do interpretable IMU features separate correct from incorrect movement before any sensor-reduction step is introduced?

2.5. Sensor Configuration Analysis

To study the trade-off between sensing burden and analytical performance, we evaluated many sensor subsets of size 1 to 4 for each task. For every subset, we

1. Built a subject-level wide feature table from the selected segments;
2. Trained a binary logistic regression model with median imputation, z-score standardization, the `liblinear` solver, and `max_iter = 5000`;
3. Evaluated LOSO accuracy and balanced accuracy.

For each sensor count, we retained the best-performing subset according to mean LOSO balanced accuracy. Throughout this paper, we refer to these as the *best reduced subsets*. We report two versions of this analysis:

- An unconstrained analysis across all available segments;
- A lower-body-only analysis including feet, lower legs, upper legs, and pelvis.

The unconstrained analysis identifies which segments carry the most information for movement quality classification. The lower-body-only analysis identifies the best lower-body-only subset, which is more realistic for wearable rehabilitation or home-monitoring use.

2.6. Cross-Task Compact Subset Check

Because the best reduced subset was task-specific, we also carried out a targeted robustness check using one compact subset that was common to the best unconstrained solutions of both tasks. The overlapping segments were the right hand and left lower leg. We therefore evaluated a shared two-segment configuration (Hand Right + LowerLeg Left) for RD and RGS using the same subject-level LOSO logistic regression pipeline. This analysis was not used to define the main best-subset results; it was included to test whether one small shared setup could remain practically usable across both tasks.

2.7. Model Family Sensitivity Analysis

To test whether logistic regression was competitive relative to other still-auditable baselines, we ran an additional subject-level sensitivity analysis on the full representation and on the shared compact subset described above. Using the same LOSO protocol and feature tables, we compared logistic regression against a shallow decision tree, Gaussian naive Bayes, and a shallow random forest. The purpose of this comparison was not to

optimize a new main model, but to check whether the main conclusions depended strongly on the use of logistic regression alone.

2.8. Explainability and Readable Feedback Layer

We also added an explainability layer based on the same interpretable feature space used for classification. For each task variant, we computed segment–feature deviations relative to the correct condition using subject-level paired differences and paired effect sizes (d_z). We then translated the strongest deviations into short natural-language summaries, such as “higher typical rotational speed at the right lower leg” or “lower rotational speed at the left foot.”

We generated two forms of summaries. First, we produced variant-level summaries describing the dominant deviations associated with each erroneous condition. Second, we generated subject-specific example summaries relative to each subject’s own correct baseline.

To make the verbal labels reproducible, we mapped absolute paired effect sizes to magnitude bands as follows: trivial for $|d_z| < 0.20$, small for $0.20 \leq |d_z| < 0.50$, moderate for $0.50 \leq |d_z| < 0.80$, large for $0.80 \leq |d_z| < 1.20$, and very large for $|d_z| \geq 1.20$. This layer was deliberately kept simple. It was designed to be readable by humans and suitable for future reporting tools, not to serve as a validated conversational feedback system. The present work therefore treats the text summaries as a structured rendering of the strongest measured segment–feature deviations, not as clinically validated advice. Their current validation is internal and methodological: the summaries are generated directly from the same paired effect-size tables used in the quantitative analysis, and the dominant phrases can be traced back to explicit segment–feature differences and their d_z values. The explainability claim in this study is therefore specific: the same interpretable segment–feature space used for movement quality classification can also indicate which body region and which movement quantity changed most strongly in each variant. In a wearable monitoring setting, that kind of output could support dashboards, clinician summaries, or future feedback interfaces without relying on opaque model behavior alone. External validation against clinician judgments remains future work.

2.9. Proof-of-Concept Closed-Loop Extension

To extend the methodology, we reformulated the same tasks at the window level. Each repetition was segmented into overlapping windows, and the same five interpretable rotational features were extracted per segment and per window. Window-level feature tables were then built for three sensor configurations per task:

- The full sensor set;
- The best reduced subset from the subject-level unconstrained search;
- The best lower-body-only subset from the subject-level constrained search.

For each window-level sensor configuration, we trained binary logistic regression models with LOSO cross-validation using the same preprocessing logic as in the subject-level analysis. This allowed us to test whether correct-versus-incorrect discrimination remained available before the full repetition had ended. We summarized this stage using window-level accuracy, window-level balanced accuracy, and median first-detect progress within incorrect repetitions.

We then derived a window-level explanation layer. For windows predicted as incorrect, each segment–feature value was compared with its task-specific correct reference. The dominant explanation for a window was defined as the segment–feature deviation with the largest standardized departure from the correct condition. This yielded interpretable keys such as LowerLeg Right + mean speed + higher or Pelvis + mean speed + higher. Variant-level summaries were then obtained by aggregating the most frequent dominant explanation across incorrect windows.

Finally, we simulated a simple corrective-feedback trigger. A cue was triggered only when three conditions were met: the window was predicted as incorrect, the incorrect-class probability exceeded a threshold, and the same dominant explanation persisted for a predefined number of consecutive windows. We also required the dominant explanation to exceed a minimum deviation-score threshold. Because a permissive default rule was too sensitive, we selected the reported task-specific rules through a small grid search over probability threshold, persistence length, and deviation-score threshold. The rule-selection criterion was to maximize trigger rate on incorrect repetitions subject to a false-trigger rate on correct repetitions of at most 10%. Under that criterion, the retained rules were as follows: for RD, probability threshold 0.95, persistence 2 windows, and deviation-score threshold 1.0; for RGS, probability threshold 0.95, persistence 3 windows, and deviation-score threshold 1.0. This proof-of-concept extension was evaluated as an offline replay of a possible closed-loop policy rather than as a live hardware deployment.

The code (re)used for this study was implemented in Python 3.13 and is publicly available in the GitHub repository (v1.0.0) at Machine-Learning-with-Minimal-IMUs (<https://github.com/GiorgosBouh/Machine-Learning-with-Minimal-IMUs>, accessed on 13 April 2026). The repository includes the scripts used for data loading, repetition segmentation, interpretable feature extraction, subject-level movement quality classification, sensor subset evaluation, readable explainability outputs, and the proof-of-concept window-level feedback extension. It also includes a README file that explains the repository structure, software requirements, expected data layout, and the main steps needed to reproduce the analyses.

3. Results

3.1. Baseline Results

The baseline analyses showed that the feature set contained useful movement quality information. Full representations outperformed minimal single-segment baselines for both tasks. In RD, mean LOSO accuracy was 0.697 ± 0.214 for the full representation and 0.447 ± 0.214 for the minimal representation ($\Delta = 0.25$, sign-flip $p < 0.001$). In RGS, the corresponding values were 0.694 ± 0.220 and 0.472 ± 0.190 ($\Delta = 0.222$, $p = 0.003$).

Multiclass LOSO classification remained above chance in both tasks, with mean accuracy 0.461 in RD and 0.403 in RGS (chance = 0.25). This means that the same simple features also retain some information about which variant was performed. The baseline summary is shown in Table 1.

Table 1. Baseline subject-level benchmarks. Binary results compare the full representation with the minimal single-segment representation. “Full acc.” and “Minimal acc.” report mean subject-level LOSO accuracy $\pm$ SD. Δ is the paired mean accuracy difference (full minus minimal), tested with a two-sided sign-flip permutation test. Multiclass accuracy refers to four-class LOSO classification (correct plus three variants).

Task	Full Acc.	Minimal Acc.	Δ (Full–Min.)	Multiclass Acc.
RD	0.697 ± 0.214	0.447 ± 0.214	0.250 ($p < 0.001$)	0.461
RGS	0.694 ± 0.220	0.472 ± 0.190	0.222 ($p = 0.003$)	0.403

Figure 4 shows the segment ablation results. It highlights that classification performance is not uniform across body regions and that the most informative segments differ between RD and RGS. This supports the practical conclusion that sensor placement should be task-specific.

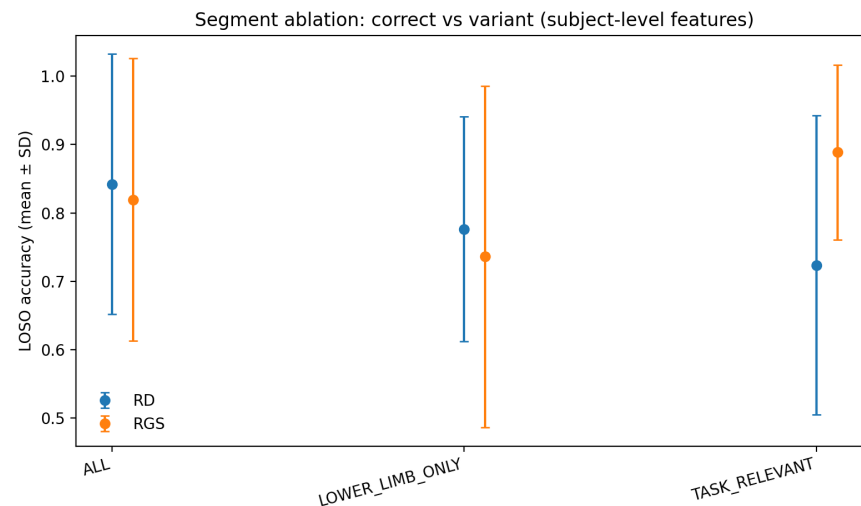


Figure 4. Segment ablation results for subject-level binary classification. Each bar corresponds to a single segment used alone as input to the LOSO logistic regression benchmark, and the plotted value is mean LOSO accuracy across held-out subjects (error bars denote SD across folds/subjects). The panel therefore shows how much information each individual body segment carries for correct-versus-incorrect discrimination in RD and RGS. Performance differs across body regions, indicating that the most informative segments depend on the task.

3.2. Explainable Results and Readable Movement Summaries

The explainability layer converted the strongest segment–feature changes into short movement summaries. The point here is not that the system produces generic text, but that the text is anchored in concrete, task-specific kinematic deviations from the correct condition. In RD pronation, the clearest pattern was higher right lower-leg speed across mean, RMS, and peak metrics ($d_z = 1.79$ – 1.87). In RD toes, the strongest changes shifted to the right hand, where peak speed and rotational range exceeded $d_z = 2.3$, suggesting a strong compensatory arm response. In RGS abduction, the dominant pattern was reduced left-foot speed together with increased pelvis speed. In RGS stork, the pattern combined left-foot reductions with increased right upper-leg motion. These are interpretable movement fingerprints rather than opaque importance scores [10].

These summaries do not replace expert interpretation, but they show why the study makes a limited explainability claim. The same feature space used for classification also identifies which segment and which rotational quantity changed most strongly in each variant, and those changes can be rendered in language without leaving the interpretable feature domain. At the present stage, validation is internal rather than clinical: each summary is traceable to explicit paired segment–feature deviations and corresponding effect sizes, but formal comparison with clinician judgments has not yet been performed. In practice, the summaries can therefore be used as structured, auditable descriptors or as input for future reporting tools in wearable system platforms, rather than as standalone clinical recommendations. Representative text outputs are listed in Table 2.

The readable summaries are also consistent with the feature-relevance analysis in Figure 5. These task-specific maps support the same explainability point: the strongest discriminative patterns are localized to interpretable segment–feature combinations rather than hidden latent factors. In other words, the explanation layer is not an extra cosmetic addition after classification; it is derived from the same transparent structure that makes the classifier output auditable in the first place.

Table 2. Examples of readable movement summaries generated from the strongest segment–feature deviations. Verbal magnitude labels follow the predefined $|d_z|$ bands described in the Methods Section, and each row reports the largest contributing paired effect size.

Variant	Readable Summary
RD pronation	Very large increases in right lower-leg typical, average, and peak rotational speed (largest $d_z = 1.87$).
RD supination	Very large increases in right-hand typical and average rotational speed, plus increased right lower-leg rotational excursion (largest $d_z = 1.55$).
RD toes	Very large increases in right-hand peak rotational speed, rotational excursion, and typical rotational speed (largest $d_z = 2.38$).
RGS abduction	Very large reductions in left-foot rotational speed with simultaneous pelvis speed increase (largest $ d_z = 1.77$).
RGS flexion	Very large reductions in left-foot speed and lower left-leg mean speed (largest $ d_z = 1.29$).
RGS stork	Very large increase in right upper-leg mean speed with strong reductions in left-foot speed metrics (largest $d_z = 1.74$).

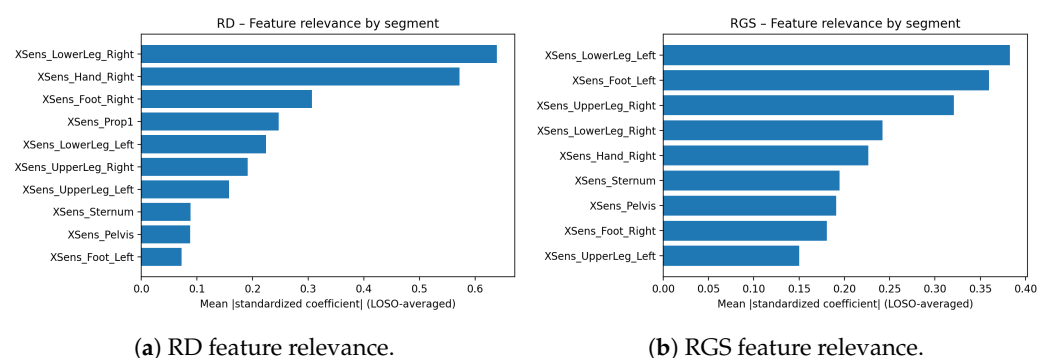


Figure 5. Feature-relevance maps for RD and RGS retained from the earlier analysis. The maps highlight which segment–feature combinations contributed most strongly to discrimination in each task.

3.3. Sensor Configuration Results

The sensor subset analysis addressed the third question of this study: how many sensors are actually needed once correct-versus-incorrect discrimination and explainability are already established? The answer was task-dependent, but the full sensor set was not always best. In RD, the best single segment was the right hand (balanced accuracy 0.746), while the best three-segment subset—right hand plus both lower legs—reached 0.947. This was higher than the full 10-segment configuration, which reached 0.886. In RGS, the best single segment was the left lower leg (0.731), the best three-segment subset (left foot, left lower leg, pelvis) reached 0.861, and the best four-segment subset (right foot, right hand, left lower leg, sternum) reached 0.917, again higher than the full nine-segment configuration (0.824).

These results suggest two main points. First, a full sensor suit is not always needed, which is important for wearable system deployment. Second, the best subset depends on the task and may include compensatory signals outside the main working limb. Figure 6 shows the performance frontier across sensor counts, and Table 3 lists the corresponding best subsets.

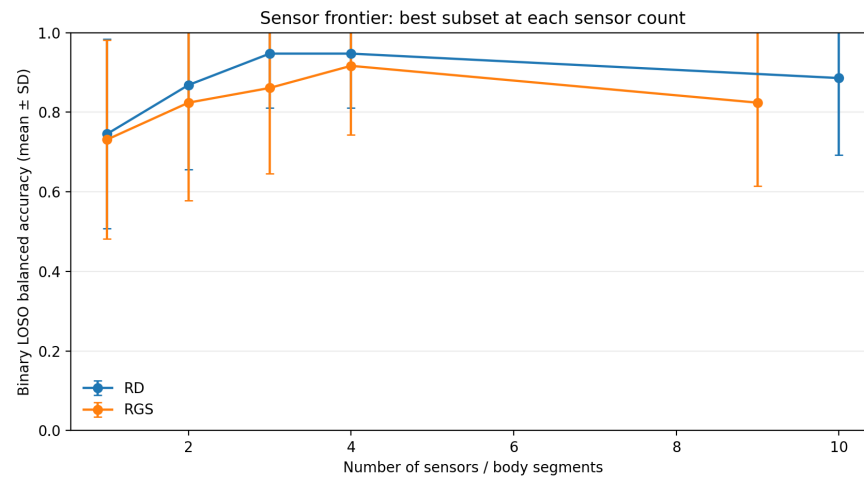


Figure 6. Performance of the best sensor subset at each subset size. For every possible number of sensors, the figure shows the subset that achieved the highest LOSO balanced accuracy (mean $\pm$ SD). For RD, the highest balanced accuracy was 0.947, achieved with Hand Right + LowerLeg Left + LowerLeg Right, and matched by a 4-sensor subset including Foot Left. For RGS, the highest balanced accuracy was 0.917, achieved with Foot Right + Hand Right + LowerLeg Left + Sternum. The exact sensor combinations are provided in Table 3. These results show that the optimal sensor configuration depends on the task.

Table 3. Best-performing unconstrained sensor subsets at each sensor count. Accuracy is mean LOSO accuracy across held-out subjects. Balanced accuracy is mean LOSO balanced accuracy, that is, the mean of sensitivity and specificity in the binary correct-versus-incorrect task, averaged across LOSO folds.

Task	<i>k</i>	Best Subset	Accuracy	Balanced Acc.
RD	1	Hand Right	0.829	0.746
RD	2	Hand Right + LowerLeg Right	0.908	0.868
RD	3	Hand Right + LowerLeg Left + LowerLeg Right	0.947	0.947
RD	4	Foot Left + Hand Right + LowerLeg Left + LowerLeg Right	0.947	0.947
RD	All	Full 10-segment set	0.908	0.886
RGS	1	LowerLeg Left	0.847	0.731
RGS	2	Foot Right + LowerLeg Left	0.875	0.824
RGS	3	Foot Left + LowerLeg Left + Pelvis	0.903	0.861
RGS	4	Foot Right + Hand Right + LowerLeg Left + Sternum	0.903	0.917
RGS	All	Full 9-segment set	0.847	0.824

3.4. Cross-Task Compact Subset and Model Family Sensitivity Results

The shared compact subset check showed that one small common setup can still retain a useful signal across both tasks, although it does not match the task-specific best subsets. Using the common two-segment configuration *Hand Right + LowerLeg Left*, RD reached a LOSO accuracy of 0.868 and balanced accuracy of 0.807, whereas RGS reached a LOSO accuracy of 0.847 and a balanced accuracy of 0.769. This suggests that one compact shared setup is feasible as a practical compromise, but the task-specific frontier remains stronger when the objective is maximal discrimination.

The model family sensitivity analysis did not overturn the main modeling choice. On the full representation, logistic regression remained the best-performing model among the tested interpretable baselines for both RD (accuracy 0.908, balanced accuracy 0.886) and RGS (accuracy 0.847, balanced accuracy 0.824). Shallow decision trees, Gaussian naive

Bayes, and shallow random forests all produced a lower balanced accuracy in this setting. The same pattern was preserved on the shared compact subset. These checks do not prove that logistic regression is universally optimal, but they do support the methodological claim that the reported conclusions are not an artifact of avoiding alternative transparent models.

3.5. Lower-Body-Only Sensor Sets

Because some of the best unconstrained solutions included upper-body signals, we also examined lower-body-only subsets. This gave a more practical picture. For RD, performance improved as lower-body sensors were added, reaching balanced accuracy 0.702 with four segments (both lower legs and both upper legs). For RGS, performance remained strong with a compact lower-body setup: a three-segment subset consisting of left foot, left lower leg, and pelvis reached a balanced accuracy of 0.861, and a four-segment subset with right foot, both lower legs, and pelvis reached 0.870.

The contrast between tasks is informative. RD benefited more from signals outside the lower body, while RGS worked well with a compact lower-body setup. This suggests that sensor placement should be chosen based on the task and the intended use case, not as one fixed rule for all movements. Table 4 summarizes the lower-body-only results.

Table 4. Best-performing lower-body-only subsets. Accuracy is mean LOSO accuracy across held-out subjects. Balanced accuracy is mean LOSO balanced accuracy for the binary correct-versus-incorrect task.

Task	k	Best Lower-Body Subset	Accuracy	Balanced Acc.
RD	1	LowerLeg Right	0.750	0.593
RD	2	LowerLeg Right + UpperLeg Right	0.776	0.640
RD	3	LowerLeg Left + LowerLeg Right + UpperLeg Right	0.803	0.675
RD	4	LowerLeg Left + LowerLeg Right + UpperLeg Left + UpperLeg Right	0.816	0.702
RGS	1	LowerLeg Left	0.847	0.731
RGS	2	Foot Right + LowerLeg Left	0.875	0.824
RGS	3	Foot Left + LowerLeg Left + Pelvis	0.903	0.861
RGS	4	Foot Right + LowerLeg Left + LowerLeg Right + Pelvis	0.889	0.870

3.6. Proof-of-Concept Closed-Loop Extension Results

The next question was narrower: if the same feature space is viewed as a stream of short windows rather than full trials, does it still support a plausible corrective-feedback workflow? The answer was cautiously positive. Window-level discrimination remained above chance in both tasks and did not disappear when compact subsets were used. In RD, the full configuration reached a window-level balanced accuracy of 0.601 and the best compact three-segment setup reached 0.605. In RGS, the full configuration reached 0.593, while the compact and lower-body setups remained lower but still above chance (both about 0.527). These values are lower than the subject-level results, which is expected because short windows provide less information than full repetitions, but they are still sufficient to justify a proof-of-concept real-time framing.

The second closed-loop result was timing. Incorrect repetitions were often detected early rather than only near trial completion. Median first-detect progress was about 0.093 of the repetition in RD and about 0.047 in RGS, with essentially the same timing for the compact RD setup and the compact or lower-body RGS setups. In practical terms, the model often crossed into “incorrect” territory within the first 5–10% of the movement. This does

not prove a finished real-time system, but it does show that the signal emerges early enough to make simulated feedback timing a meaningful research question.

The window-level explanation summaries were also consistent with the subject-level findings. At the variant level, RD pronation was most often linked to right lower-leg mean speed higher, RD supination to right lower-leg rotational range higher, RD toes to right-hand mean speed higher, RGS abduction to pelvis mean speed higher, RGS flexion to left lower-leg RMS acceleration higher, and RGS stork to right upper-leg mean speed higher. This matters because the explanation layer did not collapse when the task was reformulated in windows; it still pointed to concrete segment-feature deviations that could be mapped to a cue.

The feedback-trigger simulation needed more care. A permissive default rule produced too many triggers on correct repetitions and would have been difficult to defend. We therefore replaced the earlier hand-tuned description with a small grid-search rule-selection step that maximized the trigger rate on incorrect repetitions while constraining false triggers on correct repetitions to at most 10%. Under this criterion, the selected RD rule used a probability threshold of 0.95, persistence of 2, and deviation-score threshold of 1.0, whereas the selected RGS rule used 0.95, 3, and 1.0, respectively. With these selected rules, RD triggered in 9.95% of correct repetitions and 78.9% of incorrect repetitions, with median trigger progress of 0.639 for correct repetitions and 0.174 for incorrect repetitions. In RGS, the corresponding values were 8.8% and 64.6%, with median trigger progress of 0.619 for correct repetitions and 0.231 for incorrect repetitions. In other words, the policy remained conservative on correct movement while triggering substantially earlier and more often on incorrect RD repetitions. Table 5 summarizes this added layer.

Table 5. Proof-of-concept results for a real-time feedback version of the system. The upper part of the table shows how accurately movement errors can be detected from short time windows using three sensor setups (full, compact, and lower-body only). The lower part shows the performance of the final feedback rule, including the timing of the first corrective alert, the trigger rate on incorrect repetitions, and the false-trigger rate on correct repetitions.

Task	Configuration	Sensors	Window Acc.	Window Bal. Acc.	Median First Detect	Trigger Rate
RD	Full	10	0.772	0.601	0.093	—
RD	Best compact	3	0.800	0.605	0.093	—
RD	Best lower-body	4	0.783	0.557	0.091	—
RGS	Full	9	0.792	0.593	0.047	—
RGS	Best compact	4	0.781	0.527	0.047	—
RGS	Best lower-body	3	0.781	0.527	0.047	—
RD	Selected trigger, incorrect reps	10 *	—	—	0.174 [†]	0.789
RD	Selected trigger, correct reps	10 *	—	—	0.639 [†]	0.099
RGS	Selected trigger, incorrect reps	9 *	—	—	0.231 [†]	0.646
RGS	Selected trigger, correct reps	9 *	—	—	0.619 [†]	0.088

* The selected trigger rule was evaluated on the full-sensor prediction stream. The compact and lower-body configurations are shown in the upper part of the table for the window-level detection comparison. [†] For selected-trigger rows, this column reports the median progress of the first feedback trigger among repetitions where a trigger occurred.

4. Discussion

The main significance of this study is not only that correct and altered movement could be separated in two rehabilitation benchmarks, but that this could be done with an explicitly interpretable representation and without requiring the full available sensor configuration. This matters for rehabilitation wearables because practical deployment depends on setup burden, transparency, and actionability, not on predictive performance alone. The present results therefore support a deployment-oriented benchmark perspective: the relevant question is not whether a large sensor suit can classify movement, but whether a compact and auditable configuration can support a clinically meaningful judgement [11–15].

This positioning is important because much of the wearable rehabilitation literature has focused either on exercise recognition or on adherence monitoring. Those are valuable goals, but they are distinct from movement quality assessment, where the system must separate correct from incorrect execution and ideally support interpretation of the detected deviation. Reviews of lower-limb assessment wearables emphasize the gap between technical performance and clinically useful deployment [12,14,16]. In that context, the present study contributes a benchmark centered on movement quality, sensor parsimony, and explanation readiness rather than on activity recognition alone [15,16].

A second contribution is methodological. This study shows that explanation can remain anchored to the measured signal space rather than being appended as a post hoc layer over opaque latent representations. The reported summaries were generated directly from paired segment–feature deviations, which makes them auditable and easy to trace back to explicit kinematic quantities. That is an important property in healthcare-oriented AI, where trust and usefulness depend on whether a human reader can understand what changed and why it was flagged [10]. At the same time, this should be interpreted carefully: the present work demonstrates the internal consistency of the explanation layer with the quantitative feature analysis, but it does not yet establish agreement with clinician judgement or clinical usefulness in practice.

A third contribution is methodological rather than purely predictive. This study suggests that sensor reduction should not be treated only as a hardware simplification problem, but also as a representation problem. Earlier rehabilitation work has already shown that a single sensor, or a small number of sensors, can be sufficient for certain exercise recognition tasks [13,17,18]. Our results build on that direction by showing that reduced setups can still preserve enough information not only for discrimination, but also for explanation. This is especially relevant for movement monitoring systems intended for repeated use, home-based follow-up, or edge deployment, where setup burden and ease of use often matter as much as raw model capacity [5,14,15].

The proof-of-concept closed-loop extension also helps position this study relative to the broader rehabilitation technology literature. Prior studies have highlighted the need for sensor-based systems that do more than store recordings and instead support home rehabilitation, adherence monitoring, and timely feedback [5,11,13,19]. Reviews of interactive wearable rehabilitation systems and real-time gait detection similarly report a strong interest in IMU-based feedback, but also emphasize that clinical validation remains limited and that many systems stop short of strong patient-level evidence [20,21]. The present study does not yet deliver a validated real-time clinical feedback system, but it does show that the ingredients for such a system can be organized coherently: reduced sensing, early signal emergence, simple classification, and explanation in readable terms. In that sense, this study begins to bridge the gap between offline benchmarking and action-oriented monitoring [15,19,22].

Taken together, these findings support a broader practical principle: more sensing is not automatically more useful. In wearable rehabilitation, the goal is rarely to reconstruct everything that can be measured. It is to capture the subset of signals that best supports a clinically meaningful judgement. The literature increasingly points in this direction, emphasizing not only technical accuracy but also usability, interpretability, and translational fit in real settings [12,14,23]. From that perspective, reduced complexity is not a weakness of the approach. It is part of its intended value.

This study also has some limitations. First, the analysis is based on a relatively small cohort and on instructed movement deviations under controlled conditions. This makes the dataset suitable as a benchmark, but it does not capture the wider variability seen in clinical populations, where compensatory behaviour, fatigue, pain, and day-to-day inconsistency

can alter movement quality in less structured ways [7,24]. Second, although we added a sensitivity check with decision trees, Gaussian naive Bayes, and random forests, the modeling space explored here is still limited and should not be interpreted as an exhaustive model search. Third, this study focuses on benchmarking rather than full system validation. Sensor subset selection, performance estimation, explanation generation, and the feedback extension were all explored within the same experimental framework. Fourth, the readable summaries and closed-loop logic were not evaluated by clinicians or patients, and the current work therefore lacks an external clinical validation layer. This gap is important because prior rehabilitation system reviews have shown that patient-facing wearable feedback systems still require stronger clinical evidence and usability validation [20,22]. Future work should therefore move beyond technical feasibility and test whether the proposed outputs improve understanding, decision-making, or follow-up in real rehabilitation workflows [5,10].

5. Conclusions

This study presents a novel ML-based framework of wearables-enhanced analytics and movement quality assessment that distinguishes correct from altered movement using interpretable IMU features. The findings show that a restricted subset of meaningful signals is adequate for distinguishing correct from incorrect/altered movement, support interpretable analysis of movement deviations, and remain effective under reduced sensor configurations. In addition, the same feature set enables a proof-of-concept extension toward early error detection and corrective feedback. Overall, the proposed approach supports the development of wearables-enhanced rehabilitation systems that are accurate, interpretable, and practical for real-world use.

Author Contributions: Conceptualization, G.B. and G.S.; methodology, G.B.; software, G.B.; formal analysis, G.B.; investigation, G.B., G.S. and E.K.; writing—original draft preparation, G.B.; writing—review and editing, G.B., G.S., E.K. and K.K.; supervision, G.B. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GAITEX dataset is publicly available from its original source publication and repository (DOI: 10.5281/zenodo.15729055, <https://doi.org/10.5281/zenodo.15729055>), licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

Acknowledgments: The authors thank the creators of the GAITEX dataset for making the recordings publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Celik, Y.; Stuart, S.; Woo, W. L.; Godfrey, A. Gait analysis in neurological populations: Progression in the use of wearables. *Med. Eng. Phys.* **2021**, *87*, 9–29. [[CrossRef](#)] [[PubMed](#)]
2. Jatesiktat, P.; Lim, G.M.; Kuah, C.W.K.; Anopas, D.; Ang, W.T. Autonomous modeling of repetitive movement for rehabilitation exercise monitoring. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 175. [[CrossRef](#)] [[PubMed](#)]
3. Papi, E.; Murtagh, G.M.; McGregor, A.H. Wearable technologies in osteoarthritis and rehabilitation: A systematic review. *Sensors* **2022**, *22*, 2243. [[CrossRef](#)]
4. Peake, J.M.; Kerr, G.; Sullivan, J.P. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Front. Physiol.* **2018**, *9*, 743. [[CrossRef](#)] [[PubMed](#)]
5. Brennan, D.; Mawson, S.; Brownsell, S. Telerehabilitation: Enabling the remote delivery of healthcare, rehabilitation, and self-management. *Stud. Health Technol. Inform.* **2009**, *145*, 231–248. [[PubMed](#)]

6. Horst, F.; Lapuschkin, S.; Samek, W.; Müller, K.-R.; Schöllhorn, W. I. On the explainability of classification decisions with deep learning models in human movement analysis. In *Proceedings of the 42nd German Conference on Pattern Recognition, Dortmund, Germany, 24–27 September 2019*; Springer: Cham, Switzerland, 2019; pp. 15–28.
7. Spilz, A.; Oppel, H.; Werner, J.; Stucke-Straub, K.; Capanni, F.; Munz, M. GAITEX : Human motion dataset of impaired gait and rehabilitation exercises using inertial and optical sensors. *Sci. Data* **2026**, *13*, 11. [[CrossRef](#)] [[PubMed](#)]
8. Munz, M.; Spilz, A.; Oppel, H. *Dataset GAITEX: Human Motion Dataset of Impaired Gait and Rehabilitation Exercises Using Inertial and Optical Sensors, (1.0.0) [Data Set]*; Zenodo: Geneva, Switzerland, 2025. [[CrossRef](#)]
9. Xsens. Motion Capture. Available online: <https://www.xsens.com/products/motion-capture> (accessed on 11 April 2026).
10. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [[CrossRef](#)] [[PubMed](#)]
11. Huang, B.; Giggins, O.; Kechadi, T.; Caulfield, B. The limb movement analysis of rehabilitation exercises using wearable inertial sensors. In *Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016*; IEEE: New York, NY, USA, 2016; pp. 4686–4689. [[CrossRef](#)]
12. Porciuncula, F.; Roto, A.V.; Kumar, D.; Davis, I.; Roy, S.; Walsh, C.J.; Awad, L.N. Wearable movement sensors for rehabilitation: A focused review of technological and clinical advances. *PM&R* **2018**, *10*, S220–S232. [[CrossRef](#)] [[PubMed](#)]
13. Giggins, O.M.; Sweeney, K.T.; Caulfield, B. Rehabilitation exercise assessment using inertial sensors: A cross-sectional analytical study. *J. Neuroeng. Rehabil.* **2014**, *11*, 158. [[CrossRef](#)] [[PubMed](#)]
14. Regterschot, G.R.H.; Ribbers, G.M.; Bussmann, J.B.J. Wearable movement sensors for rehabilitation: From technology to clinical practice. *Sensors* **2021**, *21*, 4744. [[CrossRef](#)] [[PubMed](#)]
15. Phan, V.; Song, K.; Silva, R.S.; Silbernagel, K.G.; Baxter, J.R.; Halilaj, E. Seven things to know about exercise classification with inertial sensing wearables. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 3411–3421. [[CrossRef](#)] [[PubMed](#)]
16. Giggins, O.M.; Persson, U.M.; Caulfield, B. Wearable inertial sensor systems for lower limb exercise detection and evaluation: A systematic review. *Sports Med.* **2018**, *48*, 1221–1246. [[CrossRef](#)] [[PubMed](#)]
17. Giggins, O.; Sweeney, K.T.; Caulfield, B. The use of inertial sensors for the classification of rehabilitation exercises. In *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014*; IEEE: New York, NY, USA, 2014; pp. 2965–2968. [[CrossRef](#)]
18. Brennan, L.; Bevilacqua, A.; Kechadi, T.; Caulfield, B. Segmentation of shoulder rehabilitation exercises for single and multiple inertial sensor systems. *J. Rehabil. Assist. Technol. Eng.* **2020**, *7*, 2055668320915377. [[CrossRef](#)] [[PubMed](#)]
19. Bavan, L.; Surmacz, K.; Beard, D.; Mellon, S.; Rees, J. Adherence monitoring of rehabilitation exercise with inertial sensors: A clinical validation study. *Gait Posture* **2019**, *70*, 211–217. [[CrossRef](#)] [[PubMed](#)]
20. Wang, Q.; Markopoulos, P.; Yu, B.; Chen, W.; Timmermans, A. Interactive wearable systems for upper body rehabilitation: A systematic review. *J. Neuroeng. Rehabil.* **2017**, *14*, 20. [[CrossRef](#)] [[PubMed](#)]
21. Prasanth, H.; Caban, M.; Keller, U.; Courtine, G.; Ijspeert, A.; Vallery, H.; von Zitzewitz, J. Wearable sensor-based real-time gait detection: A systematic review. *Sensors* **2021**, *21*, 2727. [[CrossRef](#)] [[PubMed](#)]
22. de Lucena, D.S.; Rowe, J.B.; Okita, S.; Chan, V.; Cramer, S.C.; Reinkensmeyer, D.J. Providing real-time wearable feedback to increase hand use after stroke: A randomized, controlled trial. *Sensors* **2022**, *22*, 6938. [[CrossRef](#)] [[PubMed](#)]
23. Shenoy, A.; Samra, M.S.; Van Ooteghem, K.; Beyer, K.B.; Thomson, S.; McIlroy, W.E.; Eng, J.J.; Pollock, C.L. Evaluating the usability of inertial measurement units for measuring and monitoring activity post-stroke: A scoping review. *Sensors* **2025**, *25*, 3694. [[CrossRef](#)] [[PubMed](#)]
24. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **2018**, *180*, 68–77. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.