INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

AI-DRIVEN PHYSIOTHERAPY REHABILITATION FRAMEWORK USING WEARABLE MOTION SENSORS, REAL-TIME MOVEMENT ANALYSIS, AND PERSONALIZED EXERCISE PRESCRIPTION FOR FUNCTIONAL RECOVERY AND PATIENT OUTCOME OPTIMIZATION

Uzair Hassan[*1], Muhammad Ilyas Khan[2], Dr. Asad Shabih[3], Dr. Fizza Nazar[4], Dr. Hanan Azfar[5]

[*1] *Doctor of Physical Therapy (DPT), Abasyn University Islamabad Campus, Pakistan*
[2]*Senior Lecturer, DPT, MS Sports Physical Therapy (MS-SPT), Abasyn University Islamabad Campus, Pakistan*
[3]*Doctor of Physical therapy DPT, United Medical and Dental College (UMDC), Karachi, Pakistan*
[4]*Doctor of Physical Therapy (House Officer), Jinnah Postgraduate Medical Centre, Ziauddin University, Karachi, Pakistan*
[5]*DPT, MS-OMPT, PhD Scholar (Superior University), Consultant Physiotherapist/Orthopedic Manual Therapist at Medline Healthcare, Gujranwala, Pakistan*

[*1]hassanuzair781@gmail.com, [2]info@drilyasphysiotherapyclinic.com, [3]asdshabih@gmail.com, [4]fizzahnazar0@gmail.com, [5]hananazfar@medline.pk

DOI: https://doi.org/10.5281/zenodo.24152858

Keywords
Wearable sensors, inertial measurement units, physiotherapy rehabilitation, movement analysis, exercise recognition, personalized exercise prescription, digital twin, functional recovery

Article History
Received: 10 August 2026
Accepted: 24 October 2026
Published: 12 November 2026

Copyright @Author
Corresponding Author: *
Uzair Hassan

Abstract

Physiotherapy rehabilitation requires appropriate movement and an exercise dose matched to functional capacity. However, home sessions are rarely observed, compensatory movements may go undetected, and fixed protocols may not reflect individual recovery. This study presents a closed-loop, artificial intelligence (AI)-driven framework integrating wearable inertial measurement units (IMUs), movement recognition, exercise-quality assessment, clinician decision support, and personalized weekly exercise prescription. Because no dataset combining sensor signals, prescribed dose, and longitudinal outcomes was available, the evaluation used a literature-informed synthetic design, not human recordings. Sixty virtual patients completed 5,040 repetitions of six lower-limb exercises, recorded by IMUs on the thigh, shank, and trunk. Four model families were assessed using patient-grouped validation and bootstrap intervals. Gradient boosting achieved 0.994 accuracy for exercise recognition, a macro-F1 of 0.857 for movement-form classification, and an area under the receiver operating characteristic curve (AUC) of 0.9999 for compensation detection. Exercise-quality scoring achieved R^2 = 0.846 and a mean absolute error of 4.65 points.

The three-sensor configuration outperformed individual sensors and sensor pairs, while sampling rates as low as 12.5 Hz maintained similar classification performance. A 30° sensor rotation reduced macro-F1 from 0.848 to 0.720; rotation augmentation improved it to 0.800. Feature extraction and inference required approximately 6 ms per window on one server core. In a 12-week simulation with 150 virtual patients per arm, the combined framework improved the Timed Up and Go test by 3.11 seconds, compared with 1.99 seconds for a fixed protocol and 2.76 seconds for simulated clinician adjustment. Eighty percent met the assumed composite responder definition, compared with 51% under the fixed protocol. Feedback

and dose personalization both contributed to improvement, while measurement noise reduced the benefit of personalization. These results describe framework behavior under stated assumptions, not clinical efficacy: sensor signals and recovery trajectories were simulated, so model performance represents an upper-bound estimate. Prospective evaluation using real patient data is necessary to establish generalizability, safety, clinical utility, and effects on functional recovery.

1. Introduction

Physiotherapy rehabilitation depends on a simple but demanding principle: patients recover function when they perform the right exercises, with acceptable movement quality, at a dose that matches their current capacity, and when they do so often enough to matter. Delivering that principle at scale is difficult. A therapist can supervise a handful of sessions each week, whereas recovery after joint replacement, ligament surgery, stroke or chronic musculoskeletal pain unfolds over months and mostly takes place at home, where nobody observes the repetitions. Home programmed therefore suffer from three persistent weaknesses, namely uncertain adherence, undetected compensatory movement, and prescriptions that remain fixed even as the patient improves. Digital and sensor-based programmed have been proposed as a remedy, and several recent trials report encouraging signals on adherence, retention and functional measures [1]–[3].

Wearable inertial measurement units are the most practical sensing layer for this purpose. They are inexpensive, worn on the limb or trunk, independent of lighting and camera placement, and able to stream accelerometer and gyroscope signals that describe how a movement unfolds rather than only how many times it occurred. A recent systematic review and meta-analysis examined how closely wearable inertial units agree with optical motion capture and force plates for lower-limb motion during walking [4], and a systematic review of wearable sensors with machine learning in rehabilitation training screened 1,490 records to arrive at 32 eligible studies [5]. Surveys devoted to stroke rehabilitation reach a similar conclusion, that wearable sensing combined with machine learning offers a route to remote monitoring and automated assessment [6].

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

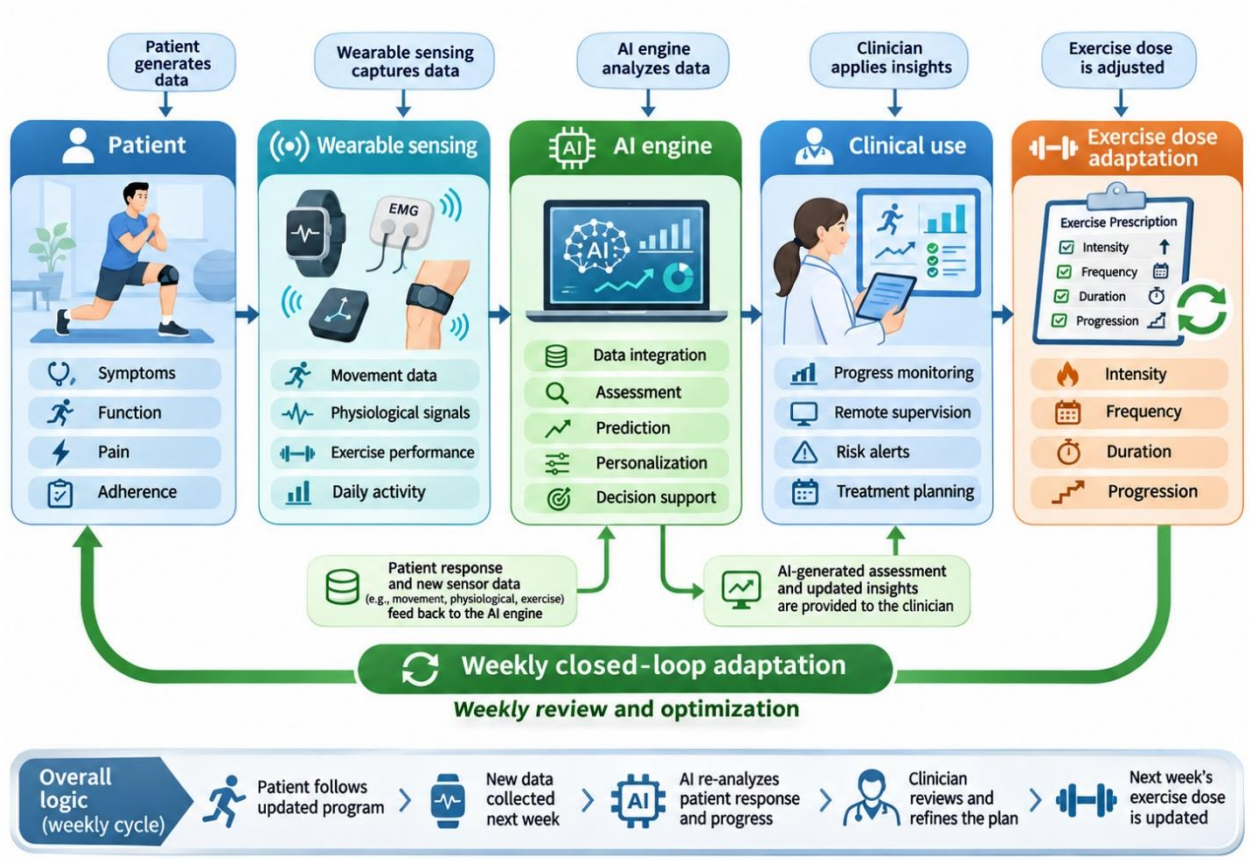


Fig. 1. Conceptual framework of the AI-driven physiotherapy rehabilitation system, linking patient, wearable sensing, AI engine, clinical use and a weekly closed-loop adaptation of exercise dose.

The machine learning layer has progressed quickly. Models that recognize which exercise is being performed and whether it is executed correctly have been built from lower-limb sensors and recurrent networks, with a reported average fitness-activity accuracy of 97.4 percent in one device study [7]. A two-stage approach that first recognizes exercise and then judges its execution reported accuracies between 88.4 and 96.2 percent across three scenarios [8], and hierarchical models have been applied to a single waist-worn sensor to monitor an exercise programmed for older adults at home and in the laboratory [9]. Gait-based prediction of future falls with wearable data is a parallel line of work [10]–[12]. Yet high accuracy within one study does not guarantee that a classifier will behave well for a new patient, a re-positioned sensor or a noisier device, and the literature on domain shift in sensor-based activity recognition shows that models often fail to generalize across subjects, devices and datasets [13], [14].

Personalization is the second open problem. Exercise dose is usually set once, by protocol, although the tolerable and effective dose changes as function returns. Reinforcement learning and contextual bandit methods have been proposed to tailor exercise recommendations, including a randomized crossover trial of app-based prescription using reinforcement learning [15] and a simulation study of an online bandit agent for N-of-1 exercise recommendation [16]. Digital twin concepts extend this idea by maintaining a continuously updated virtual model of the patient, and several recent reviews describe their potential in rehabilitation and musculoskeletal care [17]–[20]. What remains unclear is how much of the benefit of an AI-driven programmed comes from better feedback on movement quality, how much from a better-chosen dose, and how much from their combination.

Adoption introduces a further layer of difficulty. A mixed-methods survey of 236 physical therapists found that cost and resources were the main reported barriers to using artificial intelligence, together with limited knowledge of the technology [21]. A cluster randomised trial of a case-based reasoning decision support system in primary physiotherapy care randomised 44 physiotherapists and included 724 patients, and it reported no significant between-group difference in the primary global perceived effect outcome [22]. These findings are a reminder that technical accuracy is necessary but not sufficient, and that tools must be judged by outcomes, workload and trust as well as by classification scores.

This study therefore asks three questions. First, can a wearable motion-sensing pipeline recognise exercises, classify movement form and score exercise quality for patients it has never seen, and with what confidence? Second, how sensitive is that pipeline to the number of sensors, the sampling rate, the window length, sensor rotation and signal noise, and does it meet a real-time feedback latency budget? Third, when recognition, scoring and dose personalisation are combined in a closed loop, how does the resulting programme compare with fixed, clinician-adjusted and partial-AI alternatives over twelve weeks?

Because no open patient-level dataset combining raw sensor streams, prescribed dose and weekly functional outcomes was available to the authors, the study uses a fully synthetic cohort whose structure is informed by the published literature. This choice allows every assumption to be stated and varied, and it permits experiments, such as controlled sensor rotation, that would be impractical in a first clinical study. It also means that the numerical results describe the behaviour of the framework under stated assumptions and should not be read as clinical effect estimates. The contribution of the paper is a transparent, reproducible evaluation design, a quantified view of robustness and latency, and a staged comparison of the components of a closed-loop physiotherapy system.

The remainder of the paper is organised as follows. Section 2 reviews related work and identifies gaps. Section 3 describes the synthetic cohort, features, models, validation, robustness tests and prescription simulation. Section 4 presents results and discussion for each figure and table. Section 5 concludes, and Section 6 discusses limitations and future work.

2. Literature Review

This section synthesizes 32 sources published between 2021 and 2026, organized into six interconnected research streams that establish the evidence base for the proposed AI-driven physiotherapy rehabilitation framework. As illustrated in Fig. 2, these streams comprise wearable sensing, devices and reviews (n = 4); exercise recognition and quality assessment (n = 5); generalization, privacy and edge learning (n = 5); gait, falls and risk prediction (n = 4); personalization and digital twins (n = 6); and clinical trials, adoption and decision support (n = 8). Each stream addresses a complementary aspect of technology-assisted rehabilitation, ranging from sensor validity and movement assessment to functional outcome prediction, adaptive exercise prescription, clinical implementation and patient-centered decision support.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

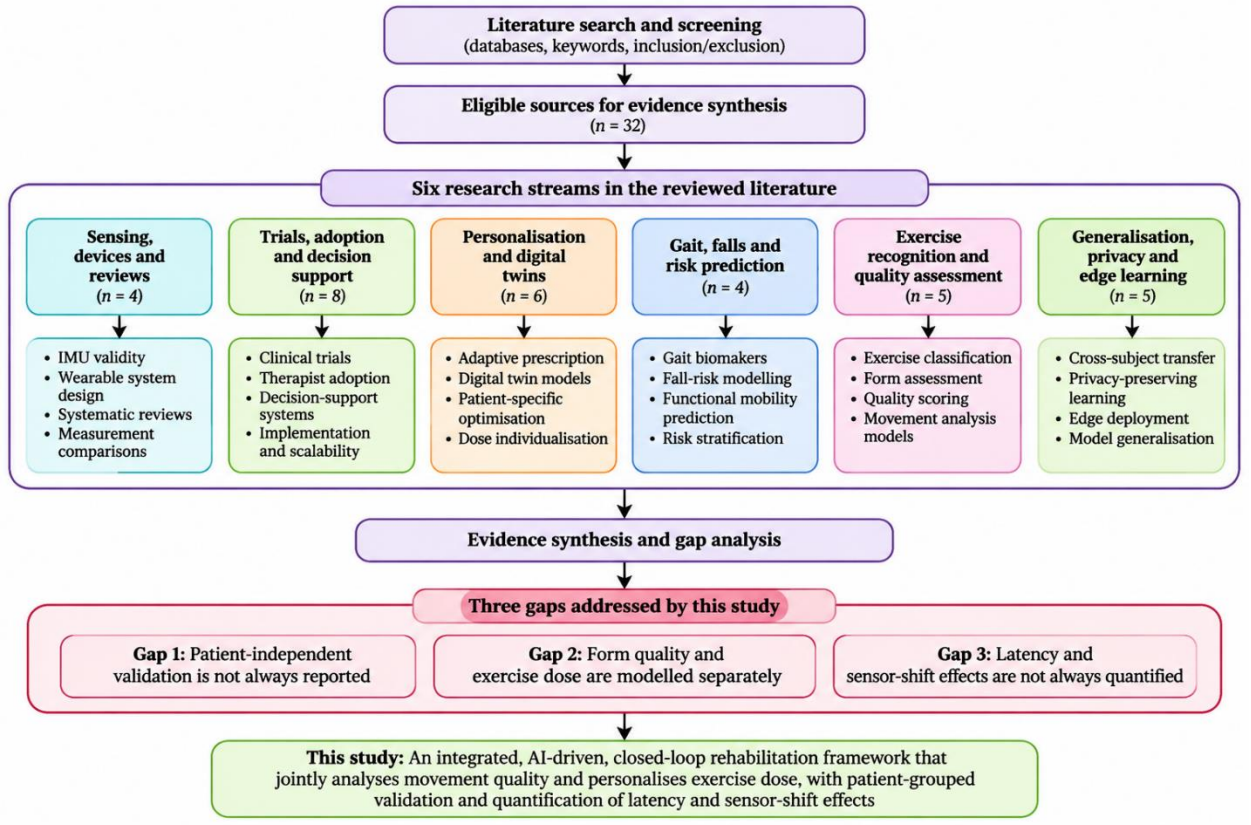


Fig. 2. Evidence landscape of the reviewed literature, illustrating six research streams, their respective source counts (n = 32), the evidence synthesis process, and three research gaps addressed by the proposed AI-driven physiotherapy rehabilitation framework.

The six streams are closely interconnected rather than independent areas of investigation. The accuracy and reliability of wearable sensing determine the quality of movement data available for exercise recognition and assessment. These outputs subsequently support compensation detection, exercise-quality scoring and functional risk prediction, while generalization, privacy and edge-learning approaches influence the reliability and feasibility of deployment beyond controlled settings. Furthermore, personalized exercise prescription and digital twin models depend on reliable patient-level information and continuous assessment of functional progress. As highlighted in Fig. 2, the literature reveals three important research gaps: limited reporting of patient-independent validation, the separate modelling of movement quality and exercise dose, and insufficient quantification of computational latency and sensor-shift effects. The present study addresses these gaps through an integrated framework that combines movement analysis, exercise-quality assessment and personalized dose adaptation within a weekly closed-loop rehabilitation process.

2.1 Wearable sensing, devices and reviews

The measurement foundation of the field is the inertial measurement unit. The meta-analytic evidence on its concurrent validity compares wearable inertial units with optical motion capture for sagittal-plane ankle, knee and hip motion during walking, and with force plates for ground reaction force estimates [4]. Such validation work matters because every downstream classifier inherits the measurement error of its sensors. Device-level studies show how sensing is being packaged for rehabilitation. One group built a system of five radio-frequency inertial devices worn on the calves, thighs and abdomen to capture home rehabilitation and fitness data, and paired it with a two-layer long short-term memory network [7]. Reviews of

wearable sensing and machine learning for rehabilitation training [5] and for automated stroke rehabilitation [6] summarise sensor types, algorithms and clinical targets, and both emphasise remote monitoring as the main attraction. For the upper limb, a study of nineteen stroke survivors wearing a wrist sensor during four functional tasks in the laboratory and at home supported the feasibility of classifying movement quality outside the clinic [23].

2.2 Exercise recognition and quality assessment

Recognizing which exercise was performed is a necessary first step, but clinical value lies in judging how well it was performed. A study using four inertial units tested six machine learning classifiers on recognition and evaluation together, and reported recognition accuracies of 88.4 and 91.4 percent in two initial scenarios, 96.2 percent in a two-stage scenario, and evaluation accuracies between 93.6 and 100 percent [8]. A hierarchical design with a single waist-worn unit first separated prescribed exercises from daily activities in a ten-minute window and then classified sub-types in a shorter window, using data from eighteen older adults [9]. Deep residual architectures that combine electromyography with inertial signals have also been proposed to recognize gym exercise activities with relevance to training and rehabilitation [24]. Human-centered work has asked how such assessments should reach clinicians, and an interactive system supporting collaborative decision making between an AI model and therapists was presented for rehabilitation assessment [25]. Table 1 collects representative studies from these streams and the main quantities each reported.

Table 1: Representative studies and the evidence they reported.

Study	Focus	Design and scale	Reported evidence
[1]	Wearable-supported training after hip arthroplasty	Randomised trial, 240 patients, 24 weeks	Hip outcome difference of 1.298 points, below the minimal clinically important difference named by the authors; timed up-and-go faster by 1.198 s at 6 weeks; cost difference not significant
[2]	Sensor-based telerehabilitation after knee surgery	Registry, 604 users, 10,311 measurements	Adherence of 86% against 74% in matched controls
[3]	Digital care versus conventional care for chronic low back pain	Randomised trial, 70 patients per group	Dropout of 15.7% against 34.3%; comparable disability recovery
[26]	Wearable remote rehabilitation after stroke	Pilot randomised trial, 120 inpatients, 3 weeks	Upper-extremity Fugl-Meyer gain of 11.28 points against 7.45 in the control group
[8]	Joint exercise recognition and evaluation	Four inertial units, six classifiers	Recognition accuracy 88.4% to 96.2%; evaluation accuracy 93.6% to 100%
[7]	Lower-limb fitness activity recognition	Five wireless inertial devices, two-layer LSTM	Average accuracy of 97.4%
[12]	Fall risk in Parkinson's disease	Real-world inertial gait data	Merged two-week real-world data outperformed unsupervised gait tests, which reached 68.0% balanced accuracy

Study	Focus	Design and scale	Reported evidence
[22]	Case-based decision support in primary physiotherapy	Cluster trial, 44 physiotherapists, 724 patients	No significant difference in global perceived effect

2.3 Gait, falls and risk prediction

Wearable gait analysis has a longer history in fall-risk prediction than in exercise coaching, and its methods transfer readily. A random forest model built on linear and nonlinear gait features from a wearable unit was developed with 171 community-dwelling older adults and then tested on 44 further adults followed for six months [10]. Another study applied machine learning to kinematics recorded by wearable sensors during the Timed Up and Go test to predict fall risk [11]. In Parkinson's disease, a comparison of aggregation strategies found that merging two weeks of real-world gait data predicted fall risk better than unsupervised gait tests [12]. A camera-based alternative extracted temporal gait features through pose estimation and reported that a gradient-boosted model reached 96 percent accuracy, while noting that wearable sensors carry inconvenience and cost for participants [27]. These studies show the value of simple models on well-chosen gait features, and they also illustrate a recurring limitation, namely that performance is often reported for one sample and one set-up.

2.4 Generalization, privacy and edge learning

Models trained on one group of people or devices often degrade on another. A comparison of handcrafted features and deep representations for domain generalisation stated that current activity recognition approaches cannot generalise across subjects, devices or datasets with acceptable performance [13]. Adversarial subject translation, which adapts a model from annotated source subjects to unlabelled target subjects, improved high-level activity recognition metrics by up to 13, 4 and 13 percent on three public datasets [14]. Privacy adds a second constraint, because sensor streams from patients are sensitive. Surveys of federated learning for human activity recognition describe how decentralised training can keep raw data on the device [28], and a review of federated learning and deep learning in sensor-based recognition also covers explainability [29]. A further review notes that centralised training raises privacy concerns, that federated learning has seen limited use in sensor-based recognition, and that dedicated reviews are scarce [30]. A review of wearables and digital twins for rehabilitation extends the discussion to data collection, security and edge intelligence for local, low-latency feedback [20].

2.5 Personalization and digital twins

Personalised exercise prescription has been approached with reinforcement learning. A randomised crossover trial of app-based exercise prescription using reinforcement learning examined satisfaction and exercise intensity [15], and a study of an online contextual bandit agent for N-of-1 exercise recommendation concluded from simulation that personalised adaptive intervention is feasible and can help even when few observations are available, while adding design complexity [16]. Digital twin research proposes a more general structure, a virtual patient model that is continuously updated with sensor data and used to test options before they are applied. A narrative review of digital twins in physiotherapy and rehabilitation searched PubMed, IEEE Xplore, Scopus and Web of Science [17], a technological review examined digital twins with artificial intelligence for predictive healthcare with emphasis on movement rehabilitation [18], and a current concepts review addressed musculoskeletal applications [19]. These reviews are largely conceptual in character, and controlled comparisons that separate the contribution of the twin from that of other programme features remain to be reported.

2.6 Clinical trials, adoption and decision support

Clinical trials of digital and sensor-assisted programmes give the most direct evidence of

benefit, and their effects are modest but consistent in adherence. A trial in 240 patients after hip arthroplasty found small hip-outcome differences at 24 weeks that the authors placed below the minimal clinically important difference, together with faster timed up-and-go at 6 weeks and a cost difference that was not significant [1]. A registry of 604 users of a sensor-based knee telerehabilitation device reported adherence of 86 percent against 74 percent in matched controls [2]. For chronic low back pain, a digital programme showed disability recovery comparable to conventional care with a lower dropout rate of 15.7 percent against 34.3 percent [3]. A pilot trial in 120 stroke inpatients reported a larger upper-extremity Fugl-Meyer gain with a wearable remote system [26]. Decision support has had a more mixed record. The case-based reasoning system SupportPrim PT was built from a case base of 105 patients with 29 weighted attributes [31], yet its cluster randomised evaluation found no significant difference in the primary outcome [22]. Practitioners also hesitate, since 236 physical therapists identified cost and resources as the major barriers [21]. Finally, any claim of benefit needs a yardstick, and responsiveness and minimal clinically important change have been estimated for common patient-reported and performance-based measures in knee osteoarthritis [32].

2.7 Research gaps

Three gaps follow from this review and are marked in Fig. 2. First, patient-independent validation is not consistently reported alongside random-split results, so the optimism introduced by repetition-level leakage can remain hidden. Second, movement quality and exercise dose are usually modelled in separate studies, which prevents any estimate of what each contributes when they operate together. Third, deployment properties such as latency, sensor rotation and noise tolerance, which decide whether a classifier can give feedback during a repetition, are not always quantified. The present study is designed to address these gaps in a transparent synthetic setting before clinical data are collected.

Recent advancements in artificial intelligence, machine learning, intelligent monitoring, and real-time computational systems provide important technological foundations for developing advanced physiotherapy rehabilitation frameworks. Deep-learning-based medical imaging applications demonstrate the potential of AI-driven pattern recognition and automated diagnostic assessment, while predictive neural network architectures highlight opportunities for intelligent clinical decision support and continuous patient monitoring [33], [34]. Edge-computing technologies further demonstrate the importance of reducing computational latency and enabling real-time processing in distributed sensing environments, which is particularly relevant to wearable rehabilitation devices and immediate movement-quality feedback [35]. Advanced linear algebra and mathematical optimization techniques provide computational foundations for high-dimensional feature analysis, machine-learning optimization, and the development of predictive models for complex movement data [36]. Furthermore, integrated radio-over-fiber technologies offer complementary perspectives on high-speed communication infrastructure and reliable data transmission, which are relevant considerations for connected wearable sensing and remote rehabilitation systems [37]. Research on voltage stability, automatic regulation, and intelligent turbine monitoring further illustrates engineering approaches to system stability, feedback control, sensor-based monitoring, and operational optimization, offering complementary perspectives for designing reliable real-time rehabilitation systems [38]–[40]. Furthermore, multilabel machine-learning approaches demonstrate the applicability of computational classification techniques to complex diagnostic patterns, while research on ethical and regulatory considerations in AI-based medical imaging emphasizes responsible implementation, reliability, and governance in healthcare technologies [41], [42]. Collectively, these developments provide complementary methodological and technological perspectives for advancing wearable sensor integration, automated movement recognition, exercise-quality assessment, personalized exercise prescription, and closed-loop physiotherapy rehabilitation.

3. Research Methodology

Fig. 3 presents the overall methodological workflow of the study, beginning with synthetic cohort generation and IMU signal acquisition, followed by signal preprocessing, segmentation and feature extraction. The extracted features are used to develop four learning tasks: exercise recognition, movement-form classification, compensatory movement detection and exercise-quality score regression. Four model families are compared for each task, followed by patient-independent validation using 12-fold patient-grouped splits and patient-level bootstrap confidence intervals. Additional evaluations examine model robustness to sensor configuration, sampling frequency, window length, sensor rotation and measurement noise, alongside deployment-related considerations, including inference latency, alert burden and feature-group importance.

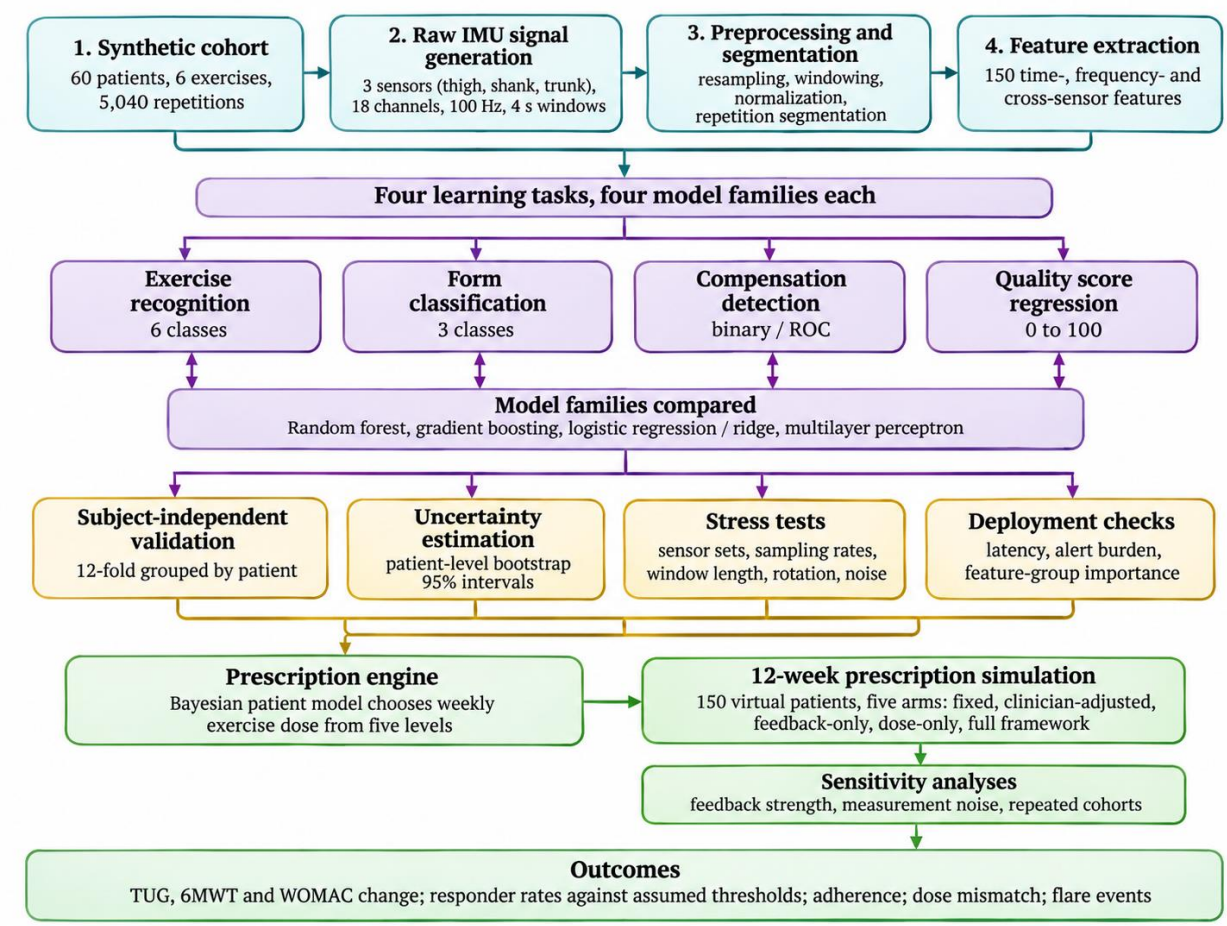


Fig. 3. Overall methodological workflow showing synthetic data generation, feature extraction, model development, validation, robustness testing, deployment assessment and 12-week personalized exercise-dose simulation.

The methodology follows a sequential evaluation strategy in which movement recognition and quality assessment are first evaluated independently to establish their predictive performance, reliability and computational feasibility. Their outputs subsequently inform the Bayesian prescription engine, which uses a patient-specific model to select weekly exercise doses from five predefined levels. The prescription strategy is then assessed through a 12-week simulation involving 150 virtual patients across five experimental arms: fixed protocol, clinician-adjusted, feedback-only, dose-only and the complete framework. Sensitivity analyses further examine the influence of feedback strength, measurement noise and repeated cohort generation. Functional outcomes, including Timed Up and

Go (TUG), six-minute walk test (6MWT) and Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) changes, are evaluated alongside responder rates, adherence, dose mismatch and flare events. All computational experiments were implemented in Python using scikit-learn, NumPy and SciPy on a two-core cloud processor. Random processes were consistently seeded to support computational reproducibility and facilitate independent replication of the reported results.

3.1 Synthetic cohort and ethical position

The study is entirely computational and uses no data from human participants, so ethical approval and consent were not required. The cohort comprises 60 simulated patients whose ages were drawn from a normal distribution with a mean of 58 years and a standard deviation of 13 years, truncated between 24 and 84 years. Each patient received an impairment severity drawn from a beta distribution with a mean of 0.44, and this severity governed range of motion, movement tempo, tremor amplitude, and the probability of poor exercise form. Patients with higher severity moved more slowly, reached a smaller fraction of the nominal range, produced noisier signals, and were more likely to compensate. Each patient also received an individual sensor noise factor and a small random mounting offset for each sensor, with a standard deviation of six degrees about each axis, to represent imperfect placement.

Each patient performed six lower-limb exercises commonly used after knee and hip surgery or in neurological rehabilitation, namely squat, sit-to-stand, heel raise, straight leg raise, step-up and hip abduction. Fourteen repetitions of each exercise were generated per patient, giving 5,040 repetitions. Each repetition carried one of three form labels, namely correct, compensatory or incomplete range. The probability of a compensatory repetition rose from 0.16 at zero severity to 0.46 at maximal severity, and the probability of an incomplete-range repetition rose from 0.10 to 0.32 over the same range. A continuous exercise quality score between 0 and 100 was also derived for each repetition. The score fell with both the severity of compensation and the deficit in range, with compensation weighted more heavily than range deficit, and Gaussian noise with a standard deviation of two points was added to represent rater disagreement.

3.2 Signal generation

Three virtual inertial units were placed on the thigh, shank and trunk, each producing three accelerometer and three gyroscope channels, for 18 channels in total. Signals were generated at 100 Hz in windows of four seconds, a length that always contained at least one complete repetition. For each exercise a primary-plane joint angle trajectory was defined for each segment, with exercise-specific amplitudes, phase lags between segments and nominal durations. Angular velocity was obtained by differentiating the trajectory, and accelerometer channels combined the projection of gravity with tangential and centripetal terms that depend on the distance of the sensor from the joint. Compensatory repetitions added secondary-plane motion at the trunk, representing lateral lean, and at the thigh, representing knee valgus wobble, together with extra high-frequency jitter. Incomplete-range repetitions reduced the primary amplitude by up to 42 percent and shortened the repetition. Patient-level mounting rotations, sensor bias and white noise were then applied, and repetition-to-repetition variability of amplitude and duration was drawn at about 14 and 13 percent respectively. The amplitudes of similar exercise pairs, such as squat and sit-to-stand, or step-up and straight leg raise, were deliberately made close, so that recognition could not be solved by a single threshold.

3.3 Feature extraction

Each four-second window was reduced to 150 features after resampling to 50 Hz unless stated otherwise. For each of the 18 channels, seven descriptive statistics were computed, namely mean, standard deviation, range, root mean square, mean absolute first difference as a jerk proxy, skewness and kurtosis, giving 126 features. For each sensor, the range of the integrated angle on each gyroscope axis was added, together with the standard deviation of accelerometer magnitude, the dominant frequency of gyroscope magnitude, and the fraction of the window in which angular speed

exceeded a quarter of its peak. Finally, correlations between the gyroscope signals of sensor pairs on two axes were added to capture inter-segment coordination, which is an indicator of compensation. Feature extraction used vectorised array operations so that a single window could be processed within a few milliseconds.

3.4 Learning tasks and models

Four tasks were defined. Exercise recognition assigned each window to one of six exercises. Form classification assigned each repetition to one of three form classes. Compensation detection was the binary version of form classification, scored by the predicted probability of the compensatory class. Quality scoring was a regression task on the 0 to 100 score. Four classifier families were compared for the classification tasks, namely a random forest with 200 trees, histogram-based gradient boosting with 120 iterations, regularised logistic regression on standardised features, and a multilayer perceptron with two hidden layers of 96 and 48 units. Three regressor families were compared for scoring, namely a random forest, gradient boosting and ridge regression. Hyperparameters were fixed in advance at conventional values and were not tuned on the evaluation folds, so that no optimistic selection effect enters the results.

3.5 Validation and uncertainty

The primary validation was subject-independent. Patients were split into twelve groups of five, and each group was held out in turn while models were trained on the remaining 55 patients, so that no patient contributed repetitions to both training and testing. For comparison, a conventional stratified twelve-fold split at repetition level was also run, in which repetitions from the same patient appear on both sides of the split. The gap between the two quantifies the optimism that repetition-level leakage would introduce. Accuracy, macro-averaged F1, area under the receiver operating characteristic curve, the coefficient of determination and mean absolute error were computed from the pooled out-of-fold predictions. Uncertainty was expressed as 95 percent percentile intervals from a cluster bootstrap that resampled patients with replacement, 600 times for most metrics and 300 times for the area under the curve, which respects the correlation of repetitions within a patient. Operating points for compensation alerts were chosen at fixed specificities of 90 and 95 percent.

3.6 Stress tests and deployment checks

Six sensor configurations were compared, namely each single sensor, thigh plus shank, thigh plus trunk, and all three sensors, at four sampling rates of 12.5, 25, 50 and 100 Hz obtained by anti-aliased decimation, using six-fold patient-grouped validation with a random forest of 120 trees. Window length was varied from 1.0 to 4.0 seconds by truncating each window. Robustness to sensor displacement was tested by rotating every sensor frame by a fixed angle between 0 and 30 degrees about a random axis at test time. Two training regimes were compared, a standard regime and an augmented regime in which the training set was doubled with copies rotated by random angles of up to 20 degrees. Robustness to noise was tested by adding Gaussian noise whose standard deviation was between 0 and 0.8 of each channel standard deviation. Feature-group importance was estimated by permuting groups of features, defined by sensor and channel, in the held-out fold three times and recording the mean fall in macro-F1. Inference latency was measured as the median and 95th percentile of 200 single-window calls to feature extraction and to each trained model on one processor core.

3.7 Prescription engine and twelve-week simulation

The prescription layer was evaluated in a simulated twelve-week programme. Each virtual patient had a latent functional capacity between 0 and 1, which determined the measured outcomes. The Timed Up and Go time fell from 22 seconds at zero capacity to 6 seconds at full capacity, the six-minute walk distance rose from 150 to 570 metres, and the WOMAC total score fell from 90 to 0, each with measurement noise. Baseline capacity was drawn between 0.12 and 0.70. Exercise dose was set weekly at one of five levels. Each patient had an unknown optimal dose that rose with capacity and differed

between patients, and the weekly gain in capacity was greatest when the prescribed dose matched this optimum, declined smoothly as it departed from it, and shrank in proportion to the remaining distance to full capacity. A dose more than 1.2 levels above the optimum carried a 35 percent weekly risk of a flare, which produced a small loss of capacity and lowered adherence in the following week. Adherence for each patient and week was drawn around an individual base rate, raised by real-time feedback and by a well-matched dose. Compensatory movement reduced the weekly gain, and real-time feedback reduced the compensation rate by up to 55 percent over the first weeks.

Five arms were compared on the same virtual patients and the same random streams. Arm A prescribed a fixed middle dose throughout. Arm B simulated clinician adjustment, with the dose reviewed every four weeks using an estimate of the optimum that carried an error with a standard deviation of 0.6 levels. Arm C gave real-time feedback on movement quality but kept the dose fixed. Arm D personalised the dose without feedback, using a Bayesian patient model that maintained a probability distribution over each patient's unknown optimal-dose offset on a grid, updated it every week from the observed gain, adherence and measured capacity, and selected the dose with the highest expected match. Arm E combined both components. In arms D and E the first two weeks used the fixed dose while the model collected information. The primary comparison was the twelve-week change in the Timed Up and Go time, and secondary outcomes were the changes in six-minute walk distance and WOMAC score, mean adherence, mean dose mismatch, flare count, and the share of patients reaching illustrative responder thresholds of a 2.0 second improvement in Timed Up and Go, a 50 metre gain in walk distance and a 12 point fall in WOMAC. A composite responder met at least two of the three thresholds. These thresholds are assumptions chosen for illustration and were not taken from a particular patient population.

3.8 Statistical analysis, sensitivity and reproducibility

For the prescription simulation, 150 virtual patients were generated per arm, and group means were reported with 95 percent bootstrap intervals from 2,000 resamples. Arms were compared with the fixed arm using the Mann-Whitney test and the standardised mean difference with small-sample correction. To judge how much the conclusions depend on one cohort, the whole simulation was repeated on 30 independent cohorts of 100 virtual patients, and the paired difference between arms was summarised. Sensitivity analyses varied the strength of the feedback effect on compensation, with reductions of 25, 55 and 85 percent, and the noise of the measured capacity used by the Bayesian model, with standard deviations of 0.012 and 0.03. Table 2 lists the principal settings of the study. Raw simulated signals, feature matrices, out-of-fold predictions and recovery trajectories were saved to disk, and all random generators were seeded.

Table 2: Principal settings of the synthetic study.

Component	Setting
Cohort	60 simulated patients; ages 24 to 84 years; impairment severity from a beta distribution
Exercises and repetitions	Six lower-limb exercises; 14 repetitions each; 5,040 repetitions in total
Sensors and signals	Thigh, shank and trunk units; 18 channels; 100 Hz base rate; 4 s windows
Features	150 features per window after resampling to 50 Hz
Classifiers	Random forest, gradient boosting, logistic regression, multilayer perceptron
Validation	Twelve patient-grouped folds; stratified repetition-level split for comparison; patient-level bootstrap

Component	Setting
Stress tests	6 sensor sets by 4 sampling rates; 5 window lengths; 6 rotation angles; 5 noise levels
Prescription simulation	Five arms; 12 weeks; 150 virtual patients per arm; 30 replicate cohorts of 100
Responder thresholds (assumed)	Timed Up and Go 2.0 s; six-minute walk 50 m; WOMAC 12 points; composite at least two of three

4. Results and Discussion

This section reports the results of the synthetic study in the order of the pipeline, from recognition and scoring, through validation and stress tests, to the twelve-week prescription simulation. All performance figures for the learning tasks come from patient-grouped validation unless stated otherwise, and all intervals are 95 percent patient-level bootstrap intervals. Because the cohort is synthetic, the values describe how the framework behaves under the assumptions of Section 3 and should not be read as clinical effect sizes. The cohort contains 5,040 repetitions from 60 simulated patients. Of these, 2,622 repetitions (52.0 percent) were correct, 1,423 (28.2 percent) were compensatory and 995 (19.7 percent) had incomplete range, and the mean quality score was 81.3 points with a standard deviation of 16.1.

4.1 Overall classification performance

Table 3 summarizes the performance of the four classifier families on exercise recognition and form classification, together with the same metric under a leaky repetition-level split. Exercise recognition was solved almost completely by all families, with gradient boosting reaching an accuracy of 0.994 (0.991 to 0.996) and the multilayer perceptron lowest at 0.972. Form classification was harder, and gradient boosting again led with a macro-F1 of 0.857 (0.844 to 0.868).

Table 3: Subject-independent performance for exercise recognition and form classification, with the repetition-level split shown for comparison.

Task	Model	Accuracy (95% CI)	Macro-F1 (95% CI)	Macro-F1, leaky split
Exercise (6 classes)	Random forest	0.992 (0.989 to 0.994)	0.992 (0.989 to 0.994)	0.991
Exercise (6 classes)	Gradient boosting	0.994 (0.991 to 0.996)	0.994 (0.991 to 0.996)	0.993
Exercise (6 classes)	Logistic regression	0.987 (0.984 to 0.989)	0.987 (0.984 to 0.989)	0.989
Exercise (6 classes)	Neural network (MLP)	0.972 (0.966 to 0.977)	0.972 (0.966 to 0.977)	0.975
Form (3 classes)	Random forest	0.882 (0.871 to 0.893)	0.850 (0.831 to 0.865)	0.856
Form (3 classes)	Gradient boosting	0.883 (0.869 to 0.894)	0.857 (0.844 to 0.868)	0.872
Form (3 classes)	Logistic regression	0.850 (0.836 to 0.860)	0.819 (0.804 to 0.831)	0.823

Task	Model	Accuracy (95% CI)	Macro-F1 (95% CI)	Macro-F1, leaky split
Form (3 classes)	Neural network (MLP)	0.858 (0.843 to 0.870)	0.825 (0.810 to 0.838)	0.842

The remaining exercise recognition were concentrated in one pair. Of the 840 squat windows, 14 were assigned to step-up, and 9 of 840 step-up windows were assigned to squat, which agrees with the deliberately similar amplitude profiles of the two movements. Every other exercise was recognised with at most three errors in 840 windows. The near-ceiling performance should be interpreted with care. Real exercise sessions contain pauses, partial repetitions and unscripted movements that a simulator does not generate, so the figures above are best read as an upper bound for clean, segmented repetitions and not as a forecast for free-living recognition.

4.2 Exercise quality scoring

Fig. 4 compares the predicted exercise quality score with the reference score for every repetition, using out-of-fold predictions from the best regressor. The density of points follows the identity line closely across the full range, from poor repetitions near 40 points to good repetitions near 100.

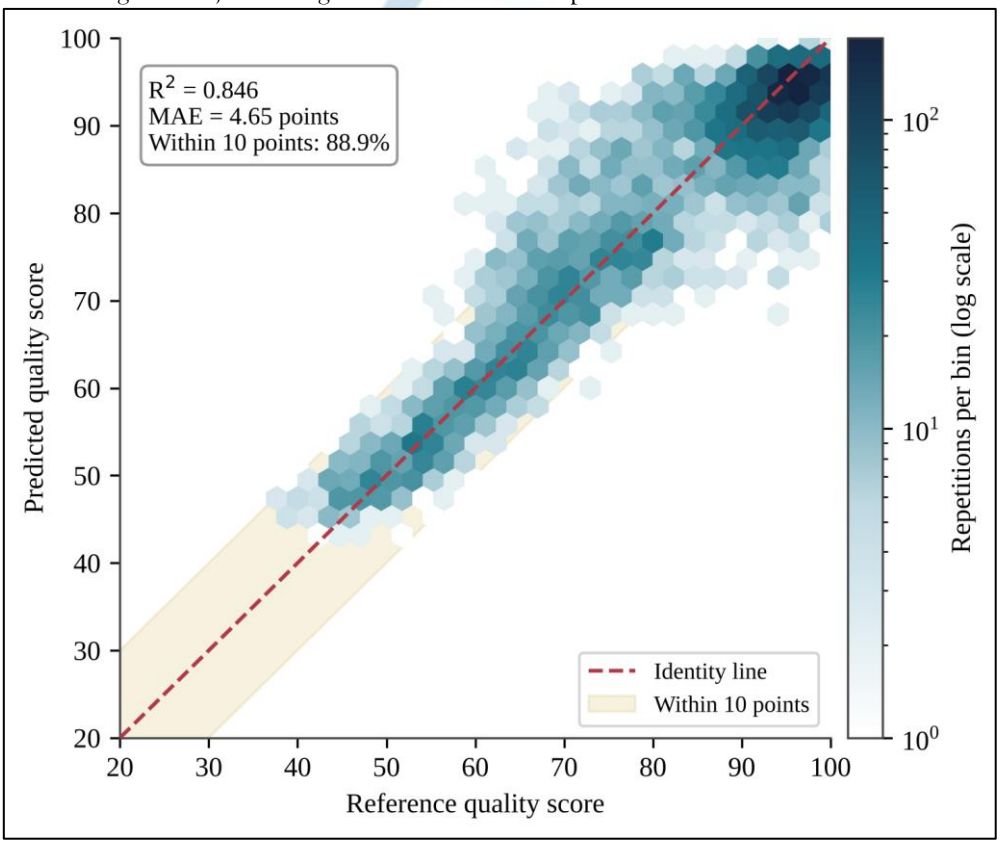


Fig. 4. Agreement between predicted and reference exercise quality scores for gradient boosting under patient-grouped validation, with the identity line and a band of plus or minus 10 points.

Gradient boosting achieved a coefficient of determination of 0.846 (0.830 to 0.858) and a mean absolute error of 4.65 points (4.47 to 4.89). The random forest was close behind, with a coefficient of determination of 0.842 and a mean absolute error of 4.75 points, whereas ridge regression, which cannot represent interactions among features, reached only 0.735 and 6.51 points. The mean prediction error was 0.20 points, so the model was essentially unbiased, and the 95 percent limits of agreement ran from -12.2 to 12.6 points. A total

of 66.1 percent of repetitions were scored within 5 points of the reference and 88.9 percent within 10 points.

The scatter shows two systematic features. Predictions near the top of the scale are slightly compressed, because a model trained to minimize squared error shrinks extreme values towards the mean, and the spread of predictions at any reference level widens for mid-range scores, where compensation and range deficit combine in many ways. For clinical use, a score with limits of agreement of about 12 points is better suited to tracking trends over sessions and triggering feedback at coarse thresholds than to ranking individual repetitions finely, and a clinician-facing display should show the score with a band or traffic-light category rather than a single decimal value.

4.3 Form classification

Fig. 5 shows the confusion matrix for form classification with gradient boosting, normalized by the reference class. The figure separates the three form categories and shows where the classifier errs.

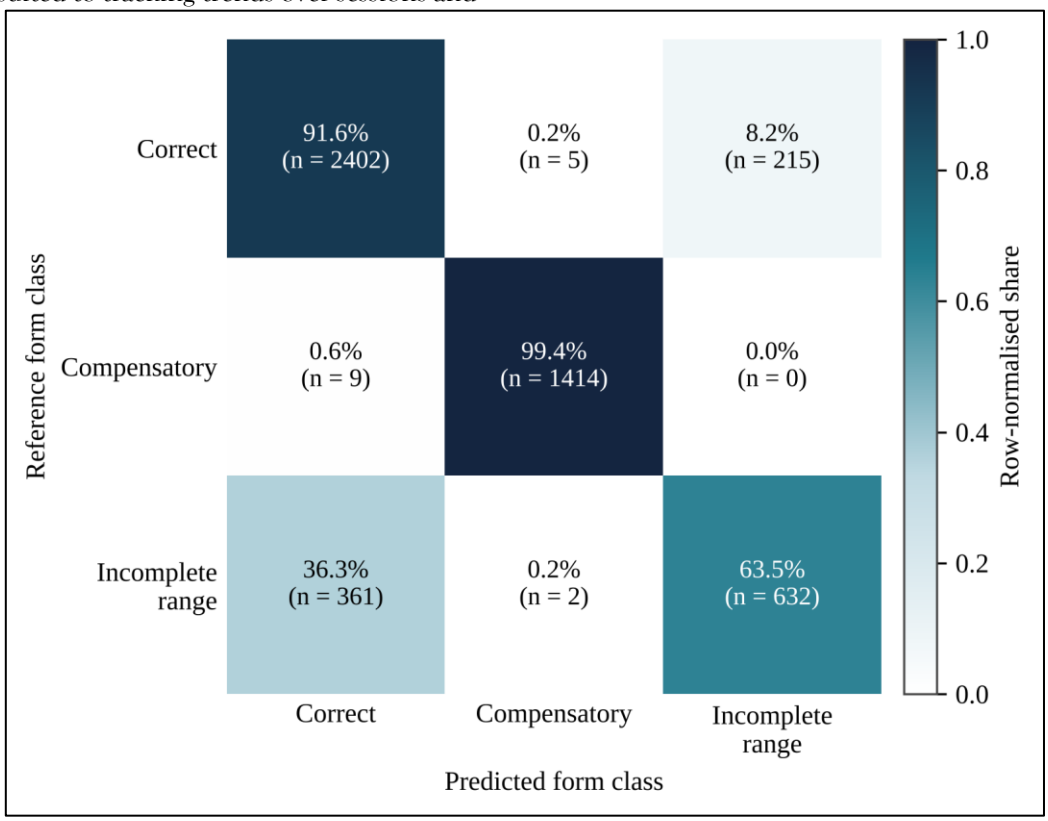


Fig. 5. Row-normalized confusion matrix for three-class form classification with gradient boosting under patient-grouped validation.

Compensatory repetitions were identified with a precision of 0.995 and a recall of 0.994, for an F1 of 0.994, and only 9 of 1,423 compensatory repetitions were labelled as correct. Correct repetitions were recognized with a recall of 0.916 and a precision of 0.867. Incomplete-range repetitions were the weak point, with a recall of 0.635, a precision of 0.746 and an F1 of 0.686. A total of 361 of 995 incomplete-range repetitions (36.3 percent) were classified as correct, and 215 correct repetitions were classified as incomplete.

This pattern has a clear explanation in the simulator. Compensation produces a distinctive secondary-plane signature at the trunk and thigh that no other form of error shares, so it is separable. A reduced range, by contrast, is a matter of degree, and a patient with a naturally small range of motion can produce a correct repetition that resembles a mildly truncated one from a patient with a larger range. The labels were set at a severity threshold, so repetitions near the threshold are intrinsically ambiguous. In practice, the same ambiguity would affect

human raters, and a deployed system should treat range feedback as advisory, using a tolerance band that is personalized to the patient's own baseline range, while reserving firm alerts for compensatory patterns.

4.4 Subject-independent validation compared with a leaky split

Fig. 6 compares the macro-F1 of the four classifier families for form classification under patient-grouped validation with the value obtained when repetitions from the same patient are allowed on both sides of the split.

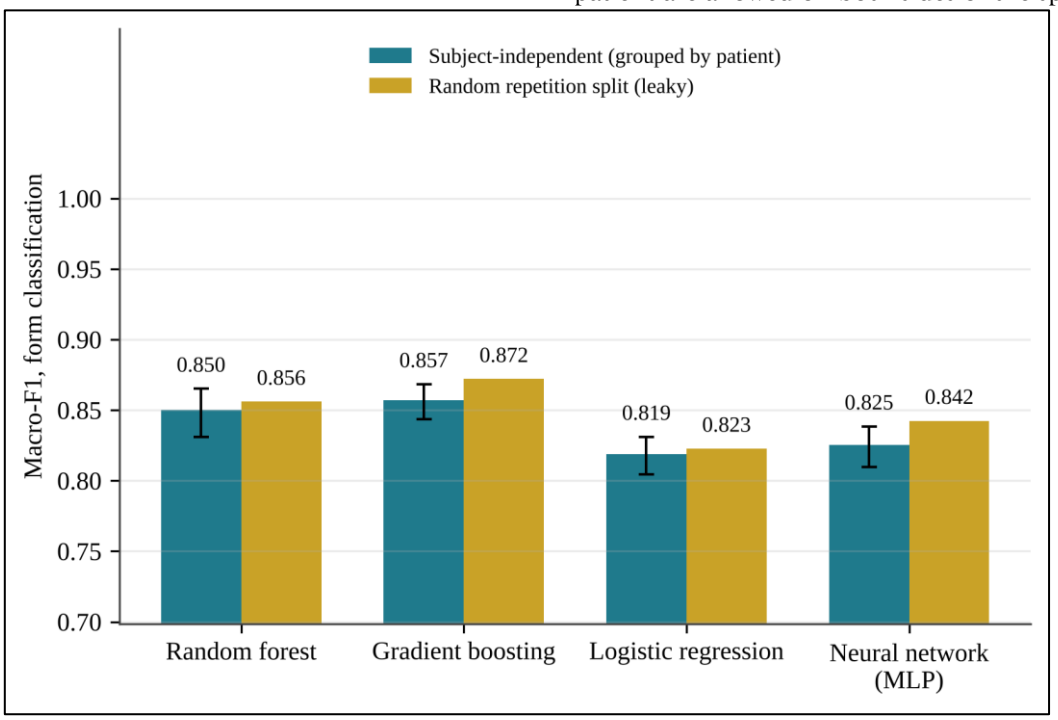


Fig. 6. Macro-F1 for form classification under subject-independent validation, with 95% patient-level bootstrap intervals, compared with a leaky repetition-level split.

The repetition-level split gave higher values for every family, but the inflation was modest. It was 0.006 for the random forest, 0.015 for gradient boosting, 0.004 for logistic regression and 0.017 for the multilayer perceptron. The ordering of the families was the same under both schemes, with the two tree ensembles ahead of the multilayer perceptron and logistic regression, and the intervals for the two ensembles overlapped, so the data do not support a firm ranking between them.

The small size of the gap should not be taken as evidence that leakage is harmless. In the simulator, a patient's idiosyncrasies are limited to a handful of parameters, namely range, tempo, tremor, noise and mounting offset, and these change features smoothly. Real patients differ in many more ways, including body habitus, footwear, surgical history, strapping and habitual strategy, and a model can memorise these through repeated exposure to the same person. The leakage gap observed here is therefore a lower bound, and subject-independent validation remains the appropriate reporting standard for any clinical claim.

4.5 Compensation detection and alert behaviour

Fig. 7 shows the receiver operating characteristic curves for detecting compensatory repetitions, drawn on a logarithmic false-positive axis so that the differences between models, which are invisible on a linear axis, become visible.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

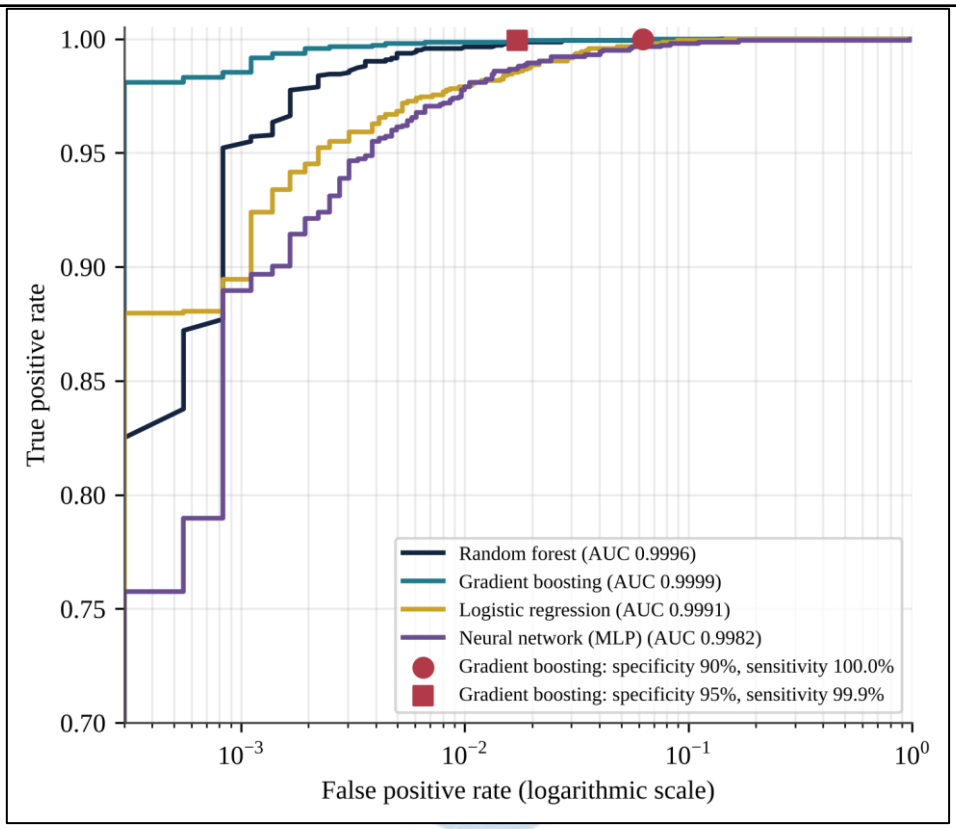


Fig. 7. Receiver operating characteristic curves for compensation detection under patient-grouped validation, with a logarithmic false-positive axis and operating points of gradient boosting at specificities of 90 and 95 percent.

All four models achieved an area under the curve above 0.998. Gradient boosting reached 0.9999 (0.9998 to 1.0000), the random forest 0.9996, logistic regression 0.9991 and the multilayer perceptron 0.9982. The logarithmic axis shows that gradient boosting separated the classes at false-positive rates below 0.1 percent, where the random forest and logistic regression lost sensitivity. At the operating point chosen for a specificity of 90 percent, the achieved specificity was 93.8 percent and the sensitivity was 100.0 percent, and at a specificity target of 95 percent the achieved specificity was 98.3 percent with a sensitivity of 99.9 percent.

These values are unusually high and reflect the strength of the compensatory signature in the simulator, which adds lateral trunk motion that is several times larger than the baseline secondary-plane movement. Real compensation is subtler and varies between people, and published recognition accuracies for real recordings are lower. What the result does show is that the pipeline, once the signal exists, does not lose it, and that the operating point can be set with a clear trade-off. At the 90 percent specificity setting the system raised alerts on 32.7 of every 100 repetitions, of which 4.5 were false alarms, against a true compensation prevalence of 28.2 percent. A programmed with about four false alarms per hundred repetitions is tolerable for an audio or vibration cue, but alerts that interrupt exercise too often are likely to reduce adherence, so the threshold should be treated as a tunable clinical parameter and not as a fixed constant.

4.6 Sensor configuration and sampling rate

Fig. 8 presents the macro-F1 for form classification for six sensor configurations at four sampling rates, using a random forest with six-fold patient-grouped validation.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

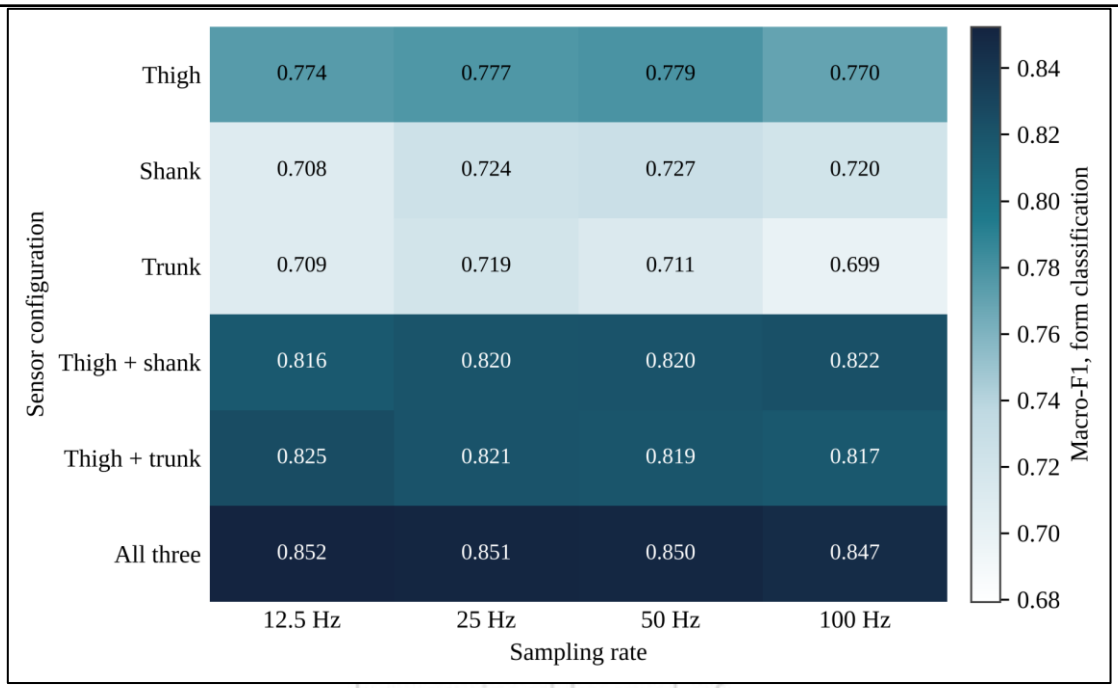


Fig. 8. Macro-F1 for form classification across sensor configurations and sampling rates, under six-fold patient-grouped validation with a random forest.

Adding sensors improved performance in a stepwise manner. The best single sensor, the thigh, reached 0.779 at 50 Hz, whereas the shank and the trunk reached 0.727 and 0.711. Pairs of sensors reached about 0.820 for thigh plus shank and 0.819 for thigh plus trunk, and all three sensors reached 0.850. The compensation detection area under the curve for the trunk sensor alone was 1.000, which explains why the trunk was the poorest sensor for three-class form classification but the best for detecting compensation. Range deficits are expressed mainly by the limb sensors, while compensation is expressed by the trunk, and the combination serves both.

Sampling rate had almost no effect. For all three sensors, the macro-F1 was 0.852 at 12.5 Hz, 0.851 at 25 Hz, 0.850 at 50 Hz and 0.847 at 100 Hz, which differs by less than the width of the bootstrap intervals reported elsewhere. The movements in this study have most of their spectral content below about 3 Hz, so a rate of 12.5 Hz is sufficient for features of this kind. Practical consequences follow. A lower sampling rate reduces battery use and radio traffic, and it shortens feature extraction, and a two-sensor thigh-plus-trunk garment would sacrifice about three points of macro-F1 relative to a three-sensor set while removing one device, a trade-off that clinicians and manufacturers can weigh against cost and wearing burden.

4.7 Which signals carry the information

Fig. 9 ranks groups of features by the fall in macro-F1 when the group is randomly permuted in held-out data across six patient-grouped folds, which shows how much the classifier depends on each sensor and channel. The left panel maps this dependence onto the body, with one rose for each sensor in which wedge area is proportional to the fall in macro-F1, and the right panel ranks the twelve most important groups on a logarithmic scale with 95% intervals and the cumulative share of importance they carry.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

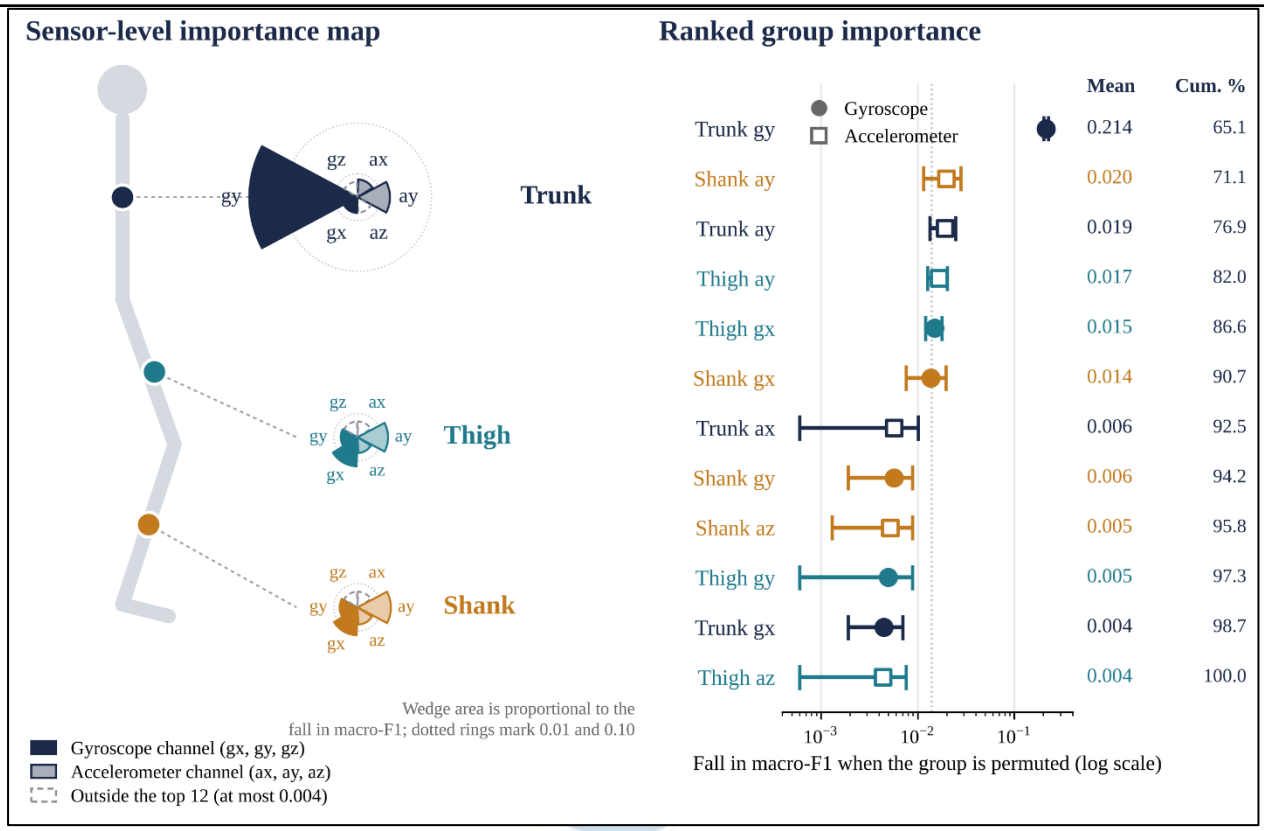


Fig. 9. Permutation importance of feature groups, expressed as the mean fall in macro-F1 with 95% intervals across six patient-grouped folds. (a): sensor-level importance map, with wedge area proportional to the fall in macro-F1 and dashed wedges for groups outside the top twelve. (b): the twelve most important groups on a logarithmic scale, with cumulative share.

One group dominated. Permuting the trunk gyroscope channel in the frontal plane reduced macro-F1 by 0.214 (95% interval of ±0.013), which is 10.8 times the effect of the next most important group, shank ay, with a fall of 0.020, and which alone accounts for 65.1% of the importance carried by the twelve leading groups. The wedge for this group dwarfs every other wedge in the trunk rose, and the trunk as a whole carry about 74.0% of the importance in the top twelve, against about 13.5% for the shank and 12.5% for the thigh. This is consistent with the simulator, in which lateral trunk lean is the principal compensatory pattern and appears as frontal-plane angular velocity at the trunk. The next five groups, namely shank ay, trunk ay, thigh ay, thigh gx and shank gx, each contributed between 0.014 and 0.020, and their 95% intervals overlap, so their relative order should not be over-interpreted; together they raise the cumulative share to 90.7%, and they carry the range and tempo information needed to separate correct from incomplete repetitions. The remaining six groups in the top twelve each contributed 0.006 or less, with the lower bounds of several intervals approaching zero, and the six groups outside the top twelve did not exceed 0.004. Gyroscope and accelerometer groups were equally represented among the twelve leading groups (6 each), but the gyroscope groups accounted for roughly 78.4% of the total because of the trunk gyroscope term.

For a clinician, this ranking is the basis of an explanation that can be shown on a dashboard. A compensation alert can be explained as excess sideways trunk motion, and a range warning as reduced thigh or shank swing, which are statements that a therapist can verify by watching the patient. Group-level permutation is cheaper and more stable than attributing importance to individual correlated features, but it describes the model and not the patient, and local explanations for single repetitions would need

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

an additional method before they are used in decisions about individual care.

4.8 Robustness to sensor rotation and noise

Fig. 10 shows how macro-F1 for form classification changes when every sensor frame is rotated by an additional angle at test time, for a model trained in the standard way and for a model trained on a rotation-augmented training set. The left panel is a rotation fan in which the angular position is the rotation angle and the radius is macro-F1, with 95% bands around each curve, and the right panel gives the change in macro-F1 per degree of rotation between successive angles.

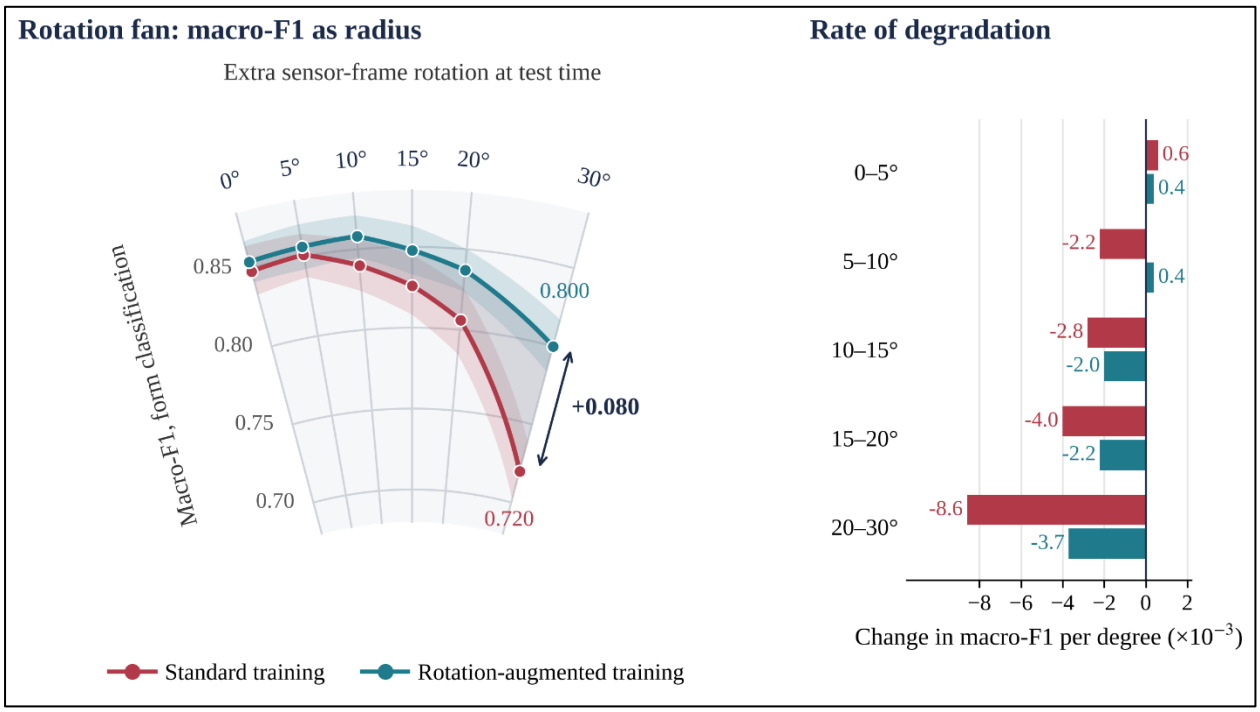


Fig. 10. Fig. 10. Macro-F1 for form classification as a function of extra sensor-frame rotation at test time, for standard and rotation-augmented training, with 95% intervals across six patient-grouped folds. (a): rotation fan, in which the angular position is the rotation angle and the radius is macro-F1. (b): change in macro-F1 per degree of rotation between successive angles.

The standard model was stable up to 10 degrees, with a macro-F1 of 0.848 at 0 degrees and 0.840 at 10 degrees, but it declined to 0.826 at 15 degrees, 0.806 at 20 degrees and 0.720 at 30 degrees, a loss of 0.128 from the unrotated value. The augmented model lost much less, falling from 0.854 to 0.848 at 15 degrees, 0.837 at 20 degrees and 0.800 at 30 degrees. The gap between the two curves widened steadily, from 0.006 at 0 degrees to 0.018 at 10 degrees, 0.031 at 20 degrees and 0.080 at 30 degrees, and the bands of the two models were clearly separated at 30 degrees and only just touched at 20 degrees. The right panel shows that the loss accelerated for both models but far more for the standard one: its macro-F1 fell by 4.0×10^{-3} per degree between 15 and 20 degrees and by 8.6×10^{-3} per degree between 20 and 30 degrees, compared with 2.2×10^{-3} and 3.7×10^{-3} for the augmented model, which is about 2.3 times slower over the final interval.

Noise had a similar but slower effect. With zero added noise the macro-F1 of the random forest was 0.847, and it stayed within 0.018 of that value up to a noise level of 0.2 times each channel standard deviation, then fell to 0.798 at 0.4 and 0.685 at 0.8. Together, the two stress tests suggest practical guidance. Garment design and a short placement check should keep sensor misalignment under about 10 degrees, rotation augmentation should be routine in training because it costs little and protects against larger shifts, and a signal-quality indicator should suppress feedback when noise exceeds roughly

0.4 times the signal spread, since classifier output beyond that level is no longer reliable.

4.9 Window length and latency

Table 4 relates the window length to classification performance and to the earliest time at which feedback can be given, and it uses the measured computing times of the pipeline. Feature extraction for a four-second window took a median of 3.19 ms (95th percentile 3.66 ms) on one core of the cloud processor. Model inference took a median of 2.69 ms for gradient boosting, 10.41 ms for the random forest and 0.22 ms for logistic regression.

Table 4: Effect of window length on form classification and earliest feedback time (random forest, six-fold patient-grouped validation; computing time measured for gradient boosting on one server core).

Window (s)	Accuracy	Macro-F1	Compensation AUC	Earliest feedback (s)
1.0	0.765	0.615	0.9868	1.006
1.5	0.806	0.692	0.9953	1.506
2.0	0.840	0.774	0.9972	2.006
3.0	0.872	0.835	0.9992	3.006
4.0	0.882	0.850	0.9997	4.006

Compute time was negligible compared with the window itself. The total of feature extraction and inference was about 5.9 ms for gradient boosting, 13.6 ms for the random forest and 3.4 ms for logistic regression, so the earliest feedback time was set almost entirely by the window length. Shortening the window cost accuracy for three-class form, with macro-F1 falling from 0.850 at 4 seconds to 0.774 at 2 seconds and 0.615 at 1 second. Compensation detection was far more tolerant, with an area under the curve of 0.9868 at 1 second and 0.9972 at 2 seconds. A two-tier design is therefore attractive, in which a short window raises a rapid compensation cue within about a second, and the full repetition window is used afterwards for range and quality feedback.

The measurements were made on a server processor and not on a wearable or phone, so they show that the model itself is light, and they do not show that an edge device would meet the same budget. Feature extraction and tree inference are inexpensive operations, and a slower mobile processor would be expected to remain well inside a one-second budget, but this expectation was not tested and should be verified on target hardware before any claim of real-time performance on a device is made.

4.10 Twelve-week prescription simulation

Fig. 11 shows the reduction in Timed Up and Go time from baseline for the five prescription arms over the twelve weeks, for 150 virtual patients per arm that share the same baseline characteristics and random streams. The arms are ordered by their improvement at week 12, the dots mark the weekly means, and the contour lines join points of equal improvement.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

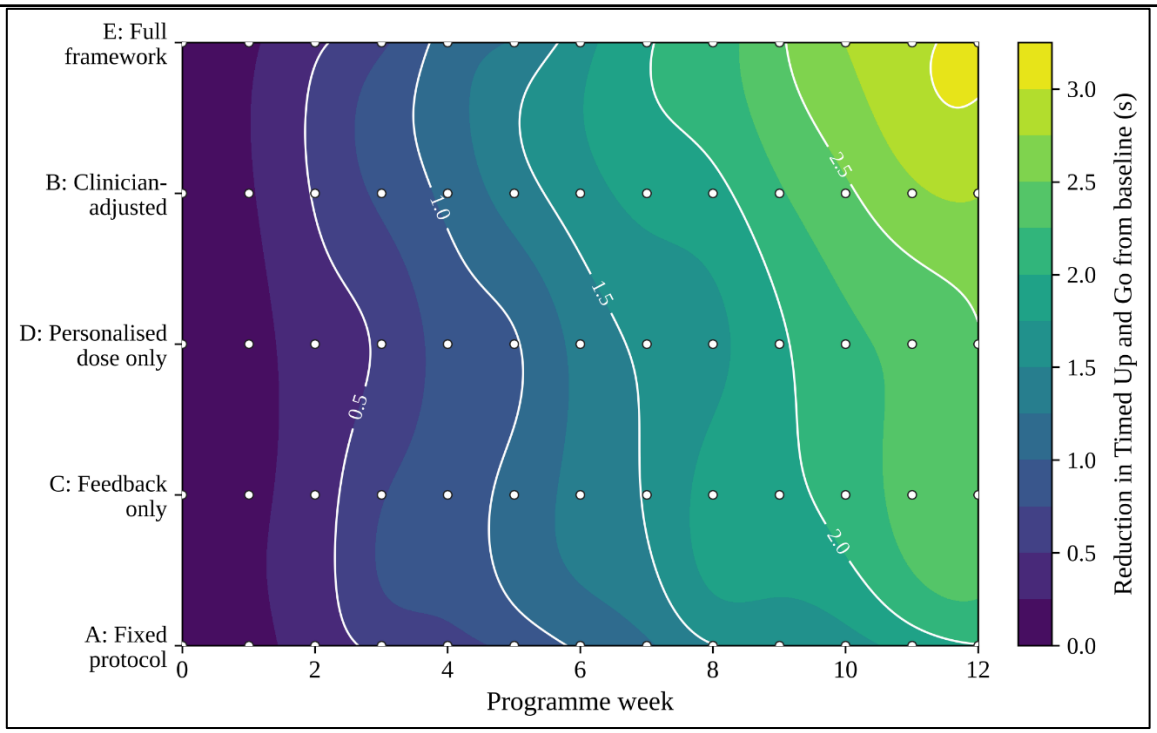


Fig. 11. Reduction in Timed Up and Go time from baseline over twelve weeks for the five prescription arms, ordered by week-12 improvement; dots mark weekly means and contours join points of equal improvement.

All arms improved, from a common baseline of 14.8 seconds. By week 12 the full framework (arm E) reached 11.7 seconds, against 12.8 seconds for the fixed protocol (arm A), 12.0 for clinician adjustment (arm B), 12.3 for feedback only (arm C) and 12.3 for personalized dose only (arm D). The trajectories separated gradually. The arms were close during the first three weeks, because the personalized arms used the fixed dose for the first two weeks, and they diverged from about week 4 as the patient model accumulated information and feedback reduced compensation. The 2.0-second contour shows the difference most clearly: the full framework crossed it at about week 7, clinician adjustment at about week 8, feedback only and personalized dose only at about week 9, and the fixed protocol only at week 12.

Table 5 gives the twelve-week changes. The mean improvement in Timed Up and Go was 3.11 seconds (2.87 to 3.36) for the full framework, compared with 1.99 seconds (1.67 to 2.32) for the fixed protocol, an advantage of 1.12 seconds with a standardised difference of 0.63. Walking distance rose by 78.1 metres against 48.7 metres, and the WOMAC total fell by 17.6 points against 11.3 points. The share of patients meeting at least two of the three assumed responder thresholds was 80 percent for the full framework and 51 percent for the fixed protocol.

Table 5: Twelve-week outcomes by prescription arm (mean with 95% bootstrap interval; 150 virtual patients per arm).

Arm	Change in TUG (s)	Change in 6MWT (m)	Change in WOMAC	Adher. (%)	Mis-match	Flares	Resp. (%)
A: Fixed protocol	1.99 (1.67 to 2.32)	48.7 (41.7 to 55.8)	11.3 (9.7 to 13.0)	66.6	0.79	0.95	51
B: Clinician-adjusted	2.76 (2.55 to 2.98)	71.2 (65.5 to 77.1)	15.8 (14.5 to 17.0)	70.2	0.51	0.13	75

Arm	Change in TUG (s)	Change in 6MWT (m)	Change in WOMAC	Adher. (%)	Mis-match	Flares	Resp. (%)
C: Feedback only	2.48 (2.14 to 2.83)	60.8 (52.8 to 68.9)	14.2 (12.4 to 16.0)	78.2	0.79	0.89	65
D: Personalized dose only	2.48 (2.24 to 2.74)	62.4 (56.0 to 68.6)	14.1 (12.9 to 15.4)	67.3	0.64	0.26	65
E: Full framework	3.11 (2.87 to 3.36)	78.1 (71.8 to 84.4)	17.6 (16.3 to 19.0)	79.5	0.61	0.25	80

The comparison among arms is informative about mechanisms. Feedback alone (arm C) raised adherence to 78.2 percent from 66.6 percent and cut the final compensation rate from 26.2 to 12.0 percent, but it left the dose mismatch at 0.79 levels and the flare count at 0.89 per patient. Personalised dose alone (arm D) reduced the mismatch to 0.64 levels and the flare count to 0.26 per patient, but did not change adherence or compensation. Each component therefore improved the primary outcome by a similar amount, about 0.48 seconds for feedback and 0.49 seconds for dose, and their combination yielded 1.12 seconds, close to the sum of the parts. The simulated clinician (arm B) performed strongly, with an improvement of 2.76 seconds, because its dose estimate was assumed to be reasonably accurate and was refreshed every four weeks, and the full framework exceeded it by 0.35 seconds.

The result is not specific to a single cohort. Across 30 independent cohorts of 100 virtual patients, the full framework exceeded the fixed protocol in every cohort, by a mean of 1.20 seconds (95 percent interval 0.97 to 1.44), and exceeded the clinician-adjusted arm in 100 percent of cohorts. Table 6 shows how sensitive the conclusion is to two assumptions that are hard to know in advance, namely the strength of the feedback effect and the noise of the measured capacity that feeds the patient model.

Table 6: Sensitivity of the mean twelve-week improvement in Timed Up and Go (seconds) to the feedback effect and to measurement noise.

Feedback effect	Measurement noise (SD)	A	B	C	D	E
25%	0.012	1.99	2.73	2.37	2.48	2.99
25%	0.03	1.99	2.73	2.37	2.27	2.57
55%	0.012	1.99	2.76	2.48	2.48	3.11
55%	0.03	1.99	2.76	2.48	2.27	2.71
85%	0.012	1.99	2.78	2.59	2.48	3.24
85%	0.03	1.99	2.78	2.59	2.27	2.88

The full framework stayed above the fixed protocol in all six settings, but its margin over the clinician-adjusted arm did not hold. With low measurement noise the full framework led the clinician arm by 0.25 seconds at a feedback effect of 25 percent and by 0.45 seconds at 85 percent. With the higher noise level of 0.03 the full framework fell below the clinician arm at the two weaker feedback effects, at 2.57 against 2.73 seconds for 25 percent and 2.71 against 2.76 seconds for 55 percent, and it exceeded the clinician arm only at the strongest feedback effect, at 2.88 against 2.78 seconds. Dose personalization alone, which depends entirely on the quality of the measured capacity, fell from 2.48 to 2.27 seconds as noise increased. The practical message is that the value of automated dose adjustment depends on measurement

quality, which is a further reason to monitor signal quality and to keep the clinician in the loop as a check on the recommendation.

Several limits of this simulation need to be stated plainly. The dynamics of recovery, the dose-response relationship, the flare mechanism and the size of the feedback effect are assumptions and were not estimated from patients. The responder thresholds are illustrative. The arms share an identical underlying mechanism by construction, so the study can show how components interact if the assumptions hold, but it cannot show that they hold. The correct reading of Table 5 is therefore as a statement about the logic of a closed-loop system, and the size of the benefit in real patients must be established in a prospective trial.

5. Conclusion

This study evaluated a closed-loop physiotherapy framework that connects wearable motion sensing, real-time movement analysis, exercise quality scoring and personalized dose prescription, using a transparent synthetic cohort of 60 patients and 5,040 repetitions. Under patient-grouped validation, gradient boosting recognized exercises with an accuracy of 0.994, classified movement form with a macro-F1 of 0.857, detected compensatory repetitions with an area under the curve of 0.9999, and scored exercise quality with a coefficient of determination of 0.846 and a mean absolute error of 4.65 points.

The stress tests showed which design choices matter. Three sensors outperformed any single sensor or pair, with a macro-F1 of 0.850 against 0.779 for the best single sensor, and sampling rates down to 12.5 Hz sufficed. A sensor rotation of 30 degrees reduced macro-F1 from 0.848 to 0.720 for standard training, whereas rotation augmentation limited the value to 0.800. The computing time of 5.9 ms per window was negligible against the window itself, so feedback timing is governed by window length.

In the twelve-week simulation, the full framework improved Timed Up and Go by 3.11 seconds against 1.99 seconds for a fixed protocol, and 2.76 seconds for clinician adjustment, with real-time feedback and dose personalization each contributing about half of the gain over the fixed protocol. Because the cohort and the recovery dynamics are synthetic, these results demonstrate the behavior of the framework under stated assumptions and a reproducible evaluation design. They do not establish clinical efficacy, which requires validation on recordings from real patients.

6. Discussion and Future Work

The principal strength of the study is its transparency. Every assumption is stated, every random process is seeded, and the same pipeline can be re-run with different assumptions, which allowed controlled experiments on sensor rotation, noise and measurement error that a first clinical dataset could not easily provide. The study also separates the contributions of feedback and dose, which clinical trials of complete digital programmed rarely do, and it reports uncertainty with patient-level bootstrap intervals.

The principal limitation is equally clear. The signals, the compensatory patterns and the recovery dynamics were simulated, so the high accuracies, especially for exercise recognition and compensation detection, reflect the separability built into the simulator and are upper bounds. Real recordings contain segmentation errors, unscripted movements, soft-tissue artefact, varied footwear and clothing, and patients whose compensation strategies are individual. The leakage gap observed here is a lower bound for the same reason. The responder thresholds were assumed, a single clinical profile of lower-limb rehabilitation was simulated, adherence was modelled as a simple function of dose match and feedback, and no model of pain, fear of movement or motivation was included. Hyperparameters were fixed and not tuned, which protects against optimism but may understate what each model family can achieve.

Several directions follow. The first is validation on real data, ideally a prospective study in which patients wear the sensors during supervised sessions and at home, with exercise labels and quality ratings supplied by blinded therapists, so that agreement between the model and human raters can be compared with agreement among raters. The second is external validation across sites, devices and body types, with the rotation and noise experiments repeated on real placement variability. The third is on-device

implementation, measuring latency, memory and battery use on target wearables or phones, and exploring privacy-preserving training across clinics so that models can improve without centralizing patient signals. The fourth is the extension of the prescription engine to richer patient models that include pain, fatigue and psychosocial factors, with safety constraints that limit the rate of dose change and require clinician approval of increases. The fifth is explanation, since a dashboard needs explanations of single repetitions and weekly recommendations that therapists find trustworthy, and these need to be tested with clinicians for usability and workload. Finally, a pragmatic randomized trial that compares feedback alone, dose personalization alone and the combination against usual care would test the mechanism suggested by the simulation and would provide the effect estimates needed for sample-size planning and cost-effectiveness analysis.

REFERENCES

Jiang, C., Shen, Y., Huang, L., Ma, Y., Lu, S., & Yin, J. (2026). The efficiency and cost-effectiveness of wearable sensors in a digital physiotherapeutic total hip arthroplasty-specific training system for patients after total hip arthroplasty: Randomized controlled trial. *JMIR mHealth and uHealth, 14*(1), Article e93050. https://doi.org/10.2196/93050

Höher, J., Lischke, B., Petersen, W., Mengis, N., Niederer, D., Stein, T., Stoffels, T., Prill, R., & Schmidt-Lucke, C. (2023). Sensor-based telerehabilitation system increases patient adherence after knee surgery. *PLOS Digital Health, 2*(2), Article e0000175. https://doi.org/10.1371/journal.pdig.0000175

Cui, D., Janela, D., Costa, F., Molinos, M., Areias, A. C., Moulder, R. G., Scheer, J. K., Bento, V., Cohen, S. P., Yanamadala, V., & Correia, F. D. (2023). Randomized-controlled trial assessing a digital care program versus conventional physiotherapy for chronic low back pain. *npj Digital Medicine, 6*, Article 121. https://doi.org/10.1038/s41746-023-00870-3

Bissenov, A., Zhubi, E., Engh, M. A., Gresits, O., Tóth, R., & Terebessy, T. (2025). Concurrent validity of wearable IMUs for sagittal plane lower-limb range of motion during walking and estimated ground reaction forces: A systematic review and meta-analysis. *Journal of Orthopaedic Surgery and Research, 20*, Article 891. https://doi.org/10.1186/s13018-025-06322-w

Wei, S., & Wu, Z. (2023). The application of wearable sensors and machine learning algorithms in rehabilitation training: A systematic review. *Sensors, 23*(18), Article 7667. https://doi.org/10.3390/s23187667

Sengupta, N., Rao, A. S., Yan, B., & Palaniswami, M. (2024). A survey of wearable sensors and machine learning algorithms for automated stroke rehabilitation. *IEEE Access, 12*, 36026–36054. https://doi.org/10.1109/ACCESS.2024.3373910

Huang, Y.-J., Chang, C.-S., Wu, Y.-C., Han, C.-C., Cheng, Y.-Y., & Chen, H.-M. (2024). Development of wearable devices for collecting digital rehabilitation/fitness data from lower limbs. *Sensors, 24*(6), Article 1935. https://doi.org/10.3390/s24061935

García-de-Villa, S., Casillas-Pérez, D., Jiménez-Martín, A., & García-Domínguez, J. J. (2022). Simultaneous exercise recognition and evaluation in prescribed routines: Approach to virtual coaches. *Expert Systems with Applications, 199*, Article 116990. https://doi.org/10.1016/j.eswa.2022.116990

Shang, M., Dedeyne, L., Dupont, J., Vercauteren, L., Amini, N., Lapauw, L., Gielen, E., Verschueren, S., Varon, C., De Raedt, W., & Vanrumste, B. (2024). Otago exercises monitoring for older adults by a single IMU and hierarchical machine learning models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 32*, 462–471. https://doi.org/10.1109/TNSRE.2024.3355299

Lockhart, T. E., Soangra, R., Yoon, H., Wu, T., Frames, C. W., Weaver, R., & Roberto, K. A. (2021). Prediction of fall risk among community-dwelling older adults using a wearable system. *Scientific Reports, 11*, Article 20976. https://doi.org/10.1038/s41598-021-00458-5

Roshdibenam, V., Jogerst, G. J., Butler, N. R., & Baek, S. (2021). Machine learning prediction of fall risk in older adults using Timed Up and Go test kinematics. *Sensors, 21*(10), Article 3481. https://doi.org/10.3390/s21103481

Ullrich, M., Roth, N., Küderle, A., Richer, R., Gladow, T., Gaßner, H., Marxreiter, F., Klucken, J., Eskofier, B. M., & Kluge, F. (2023). Fall risk prediction in Parkinson's disease using real-world inertial sensor gait data. *IEEE Journal of Biomedical and Health Informatics, 27*(1), 319–328. https://doi.org/10.1109/JBHI.2022.3215921

Bento, N., Rebelo, J., Barandas, M., Carreiro, A. V., Campagner, A., Cabitza, F., & Gamboa, H. (2022). Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition. *Sensors, 22*(19), Article 7324. https://doi.org/10.3390/s22197324

Soleimani, E., Khodabandelou, G., Chibani, A., & Amirat, Y. (2022). Generic semi-supervised adversarial subject translation for sensor-based activity recognition. *Neurocomputing, 500*, 649–661. https://doi.org/10.1016/j.neucom.2022.05.075

Doherty, C., Lambe, R., O'Grady, B., O'Reilly-Morgan, D., Smyth, B., Lawlor, A., Hurley, N., & Tragos, E. (2024). An evaluation of the effect of app-based exercise prescription using reinforcement learning on satisfaction and exercise intensity: Randomized crossover trial. *JMIR mHealth and uHealth, 12*, Article e49443. https://doi.org/10.2196/49443

Meier, D., Ensari, I., & Konigorski, S. (2023). Designing and evaluating an online reinforcement learning agent for physical exercise recommendations in N-of-1 trials. In *Proceedings of the 3rd Machine Learning for Health Symposium* (Proceedings of Machine Learning Research, Vol. 225, pp. 340–352). PMLR. https://proceedings.mlr.press/v225/

Olawade, D. B., Modum, E. R., Olawuyi, O. F., Olasilola, O. R., Makanjuola, B. D., & Alabi, J. O. (2026). The role of digital twin technology in physiotherapy and rehabilitation practice. *Virtual Reality & Intelligent Hardware, 8*(1), 71–86. https://doi.org/10.1016/j.vrih.2026.01.002

Chaparro-Cárdenas, S. L., Ramirez-Bautista, J.-A., Terven, J., Córdova-Esparza, D.-M., Romero-Gonzalez, J.-A., Ramírez-Pedraza, A., & Chavez-Urbiola, E. A. (2025). A technological review of digital twins and artificial intelligence for personalized and predictive healthcare. *Healthcare, 13*(14), Article 1763. https://doi.org/10.3390/healthcare13141763

Diniz, P., Grimm, B., Garcia, F., Fayad, J., Ley, C., Mouton, C., Oeding, J. F., Hirschmann, M. T., Samuelsson, K., & Seil, R. (2025). Digital twin systems for musculoskeletal applications: A current concepts review. *Knee Surgery, Sports Traumatology, Arthroscopy, 33*(5), 1892–1910. https://doi.org/10.1002/ksa.12627

Piechowiak, M., Goch, A., Panas, E., Masiak, J., Mikołajewski, D., Rojek, I., & Mikołajewska, E. (2025). The global importance of machine learning-based wearables and digital twins for rehabilitation: A review of data collection, security, edge intelligence, federated learning, and generative AI. *Electronics, 14*(23), Article 4699. https://doi.org/10.3390/electronics14234699

Alsobhi, M., Sachdev, H. S., Chevidikunnan, M. F., Basuodan, R., K U, D., & Khan, F. (2022). Facilitators and barriers of artificial intelligence applications in rehabilitation: A mixed-method approach. *International Journal of Environmental Research and Public Health, 19*(23), Article 15919. https://doi.org/10.3390/ijerph192315919

Granviken, F., Meisingset, I., Bach, K., Bones, A. F., Simpson, M. R., Hill, J. C., van der Windt, D. A., & Vasseljen, O. (2025). Personalised decision support in the management of patients with musculoskeletal pain in primary physiotherapy care: A cluster randomised controlled trial (the SupportPrim project). *Pain, 166*(5), 1167–1178. https://doi.org/10.1097/j.pain.0000000000003456

Seo, N. J., Coupland, K., Finetto, C., & Scronce, G. (2024). Wearable sensor to monitor quality of upper limb task practice for stroke survivors at home. *Sensors, 24*(2), Article 554. https://doi.org/10.3390/s24020554

Mekruksavanich, S., & Jitpattanakul, A. (2024). A residual deep learning method for accurate and efficient recognition of gym exercise activities using electromyography and IMU sensors. *Applied System Innovation, 7*(4), Article 59. https://doi.org/10.3390/asi7040059

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., & Bermúdez i Badia, S. (2021). A human-AI collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445472

Guo, L., Wang, J., Wu, Q., Li, X., Zhang, B., Zhou, L., & Xiong, D. (2023). Clinical study of a wearable remote rehabilitation training system for patients with stroke: Randomized controlled pilot trial. *JMIR mHealth and uHealth, 11*, Article e40416. https://doi.org/10.2196/40416

Lim, Z. K., Connie, T., Goh, M. K. O., & Saedon, N. I. B. (2024). Fall risk prediction using temporal gait features and machine learning approaches. *Frontiers in Artificial Intelligence, 7*, Article 1425713. https://doi.org/10.3389/frai.2024.1425713

Aouedi, O., Sacco, A., Khan, L. U., Nguyen, D. C., & Guizani, M. (2024). Federated learning for human activity recognition: Overview, advances, and challenges. *IEEE Open Journal of the Communications Society, 5*, 7341–7367. https://doi.org/10.1109/OJCOMS.2024.3484228

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2025). Deep learning and federated learning in human activity recognition with sensor data: A comprehensive review. *Computer Modeling in Engineering & Sciences, 145*(2), 1389–1485. https://doi.org/10.32604/cmes.2025.071858

Ting, X. W., Pang, Y. H., Ooi, S. Y., Lim, Z. Y., & Hiew, F. S. (2025). A review on federated learning on sensor-based human activity recognition. *HighTech and Innovation Journal, 6*(3), 1079–1103.

Granviken, F., Vasseljen, O., Bach, K., Jaiswal, A., & Meisingset, I. (2024). Decision support for managing common musculoskeletal pain disorders: Development of a case-based reasoning application. *JMIR Formative Research, 8*, Article e44805. https://doi.org/10.2196/44805

Mostafaee, N., Pirayeh, N., & Fakoor, M. (2024). Responsiveness and minimal clinically important changes of common patient-reported and performance-based outcome measures of physical function in patients with knee osteoarthritis. *Physiotherapy Theory and Practice, 40*(11), 2661–2669. https://doi.org/10.1080/09593985.2023.2269241

Khalil, A., Hussain, M., Majeed, M. K., Hamza, A., Ali, A., Ajaz, K., et al. (2025). Artificial intelligence in neuro-oncology: Integrating advanced machine learning techniques for accurate and early detection of brain tumors through MRI imaging. Spectrum of Engineering Sciences, 413–435.

Khalil, A., Hussain, A. W., Khan, A. H., Majeed, M. K., Mir, S. Z., Abbasi, M. D., et al. (2025). Transforming heart transplantation with AI: Deep neural networks for predictive analytics and real-time monitoring in clinical decision support systems. Pakistan Journal of Medical & Cardiological Review, 4(3), 390–411.

Haq, R. U., Aman, F., Majeed, M. A., Raza, S., Khan, A., Hussain, R., & Majeed, M. K. (2025). Developing edge computing solutions for IoT devices to reduce latency and enhance real-time decision-making. Spectrum of Engineering Sciences, 961–970.

Majeed, M. K., Dilshad, W., Essa, R., Ali, I., Majid, M., & Zia, A. (2026). Advanced linear algebra and mathematical optimization techniques for high-dimensional data analysis and machine learning applications. Spectrum of Engineering Sciences, 4(5), 2225–2249.

Majeed, M. K. (2019). Integrated radio over fiber technologies for 5G and smart LTE-Advanced systems. Asian Journal of Engineering, Sciences & Technology, 9(2).

Ahmad, B., Majeed, M. K., Soomro, A. A., & Jillani, S. A. (2026, January). Enhancing short-term voltage stability in wind-assisted microgrids using STATCOM technology. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1–6). IEEE.

Ahmad, B., Muneer, M., Jillani, S. A., & Ahmed, S. (2026, January). Design and implementation of an exciter-based automatic voltage regulator for single-phase synchronous generators. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1–6). IEEE.

Ahmad, B., Jillani, S. A., Rafique, M., & Soomro, A. A. (2026, January). Design and monitoring of a tree-inspired vertical axis wind turbine for efficient urban wind utilization. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1–6). IEEE.

Nadeem, G., Ul Haque, M. F., Ahmed, H., Ebrahim, M., Ahmad, S., de la Torre Díez, I., et al. (2025). Improving bovine disease detection through multilabel classification. Scientific Reports, 15(1), Article 34766.

Nadeem, G., & Anis, M. I. (2025). Ethical and regulatory consideration in AI-based medical imaging. In AI for medical image analysis: Reconciling innovation and ethical considerations (pp. 265–290). Springer Nature Switzerland.

International Journal of Advanced Research