

Machine learning Prediction of Self-Reported Real-World Knee Osteoarthritis Pain from Wearable Accelerometer-Derived Gait Features

David Clement^{1*}, Emmanuel Agu^{1†}, Atifa Sarwar¹,
Deepak Kumar¹, Zheyang Wu², Jean A. King³

^{1*}Department of Computer science, Worcester Polytechnic Institute,
100 Institute road, Worcester, 01609, Massachusetts, USA.

²Department of Mathematical Sciences, Worcester Polytechnic
Institute, 100 Institute road, Worcester, 01609, Massachusetts, USA.

²Department of Biology and Biotechnology, Worcester Polytechnic
Institute, 100 Institute road, Worcester, 01609, Massachusetts, USA.

*Corresponding author(s). E-mail(s): dclement1@wpi.edu;
Contributing authors: emmanuel@wpi.edu; asarwar@wpi.edu;
dkumar1@wpi.edu; zheyangwu@wpi.edu; jaking@wpi.edu;

[†]These authors contributed equally to this work.

Abstract

Osteoarthritis (OA) is the leading cause of chronic pain and mobility limitations, affecting over 32 million adults in the United States. Traditional assessments of OA pain and physical function rely on subjective self-reports and periodic clinical evaluations that are often limited by recall bias and insufficient sampling of real-world behavior. Recent advances in wearable technology offer new opportunities for objective, passive, and ecologically valid monitoring of daily movement patterns. This study investigates the feasibility of using gait quality features derived from seven-day continuous AX3 accelerometer recordings that are then analyzed using a structured analytical and ML framework **to predict OA pain self-reported OA pain using the WOMAC questionnaire's pain subscale**. The data utilized in this study was from secondary analyses of data collected in the Multicenter OA Study (MOST), a large longitudinal cohort of adults aged 50 and older with or at risk of knee OA. We analyzed data of 1,525 individuals including 1,124 with OA pain and 401 without pain. Gait metrics were derived along the vertical (V), medio-lateral (M), and anterior-posterior (AP) axes and organized into two categories: activity-related measures (e.g., walking duration and step count) and

indicators of gait quality, including rhythm, symmetry, variability, and amplitude. Using an Extreme Gradient Boosting (XGBoost) classifier, the framework achieved strong predictive performance (AUC = 0.91, accuracy = 0.93, precision = 0.94, specificity = 0.99, sensitivity = 0.91, F1 = 0.92). Gait quality metrics, particularly variability features such as stride regularity and dominant frequency characteristics, had superior predictive power compared to general activity measures. These findings indicate that subtle gait dynamics represent reproducible behavioral signatures and are candidates for digital biomarkers associated with perceived pain states. They reflect functional and behavioral adaptations to pain rather than direct measurements of underlying nociceptive processes. Accordingly, the proposed framework enables objective estimation of self-reported pain from real-world gait patterns. **The primary contribution of this work is the application of a rigorous machine-learning framework for wearable-based gait analysis of a large real-world osteoarthritis cohort, and demonstrating that gait-quality features derived from extended free-living monitoring contain robust behavioral signatures associated with self-reported OA pain.** Our findings suggest that real world gait signatures could be leveraged as scalable, non-invasive digital biomarkers for continuously monitoring perceived pain and functional impairment in real-world healthcare contexts.

Keywords: Osteoarthritis Pain Assessment, Gait Analysis, Wearable Health Monitoring, Digital Biomarkers, Machine Learning Classification

1 Introduction

Motivation: OA is a degenerative musculoskeletal disease characterized by cartilage breakdown, joint damage, and chronic pain. It is the leading cause of disability among older adults, affecting over 32 million people in the United States, approximately 14% of the adult population, and contributing an estimated \$45 billion in additional healthcare costs each year[1–4]. OA pain and stiffness, especially during movement or following rest, often results in difficulty walking, reduced mobility, and diminished quality of life [5–7]. Diagnosing OA pain can be challenging due to overlapping symptoms and the lack of a single definitive test, requiring integration of clinical, imaging, and self-report data. Early, objective identification of OA-related pain could enable personalized interventions, optimizing therapy intensity, medication timing, and even surgical evaluation, ultimately improving outcomes and reducing healthcare burden [8].

Recent advancements in wearable sensing and ML analyses methods have opened new opportunities for continuous, objective assessment of pain-related changes in mobility [9, 10]. Pain alters gait mechanics as individuals subconsciously adjust stride length, cadence, and symmetry to minimize discomfort [11, 12]. Such adjustments manifest as measurable fluctuations in gait rhythm, variability, and amplitude attributes, which can be captured through high-frequency wearable accelerometers.

Unlike conventional clinic-based assessments, wearable devices enable passive, real-world monitoring of gait over extended periods, offering ecologically valid insights into daily functional status.

Our work: In this study, we investigate whether gait-quality features derived from seven-day continuous AX3 accelerometer recordings can predict self-reported OA pain states in free-living environments. Using data from the MOST, we evaluate activity and gait-quality features spanning rhythm, symmetry, variability, and amplitude domains using multiple ML approaches, including XGBoost, LightGBM, Random Forest, Support Vector Machines, AdaBoost, ElasticNet, Naïve Bayes, and Generalized Additive Models. The primary objective is to assess whether extended real-world gait monitoring can provide robust behavioral signatures associated with self-reported OA pain and functional impairment.

2 Background and Related work

2.1 OA pain assessment using traditional methods

OA pain assessment has traditionally relied on self-reported measures such as the Western Ontario and McMaster Universities OA Index (WOMAC) and the Visual Analog Scale (VAS) scores [13]. Li et al. analyzed gait variables in individuals with bilateral knee OA and its relationship with WOMAC OA Index Assessment [14]. While effective, these subjective tools are susceptible to recall bias and inconsistencies, necessitating the development of objective assessment methods [15].

2.2 Wearable sensor for OA pain classification

Wearable sensor technologies have gained traction in musculoskeletal assessment, facilitating continuous and unobtrusive monitoring of movement patterns [16]. Previous studies have demonstrated the feasibility of accelerometer-based gait analysis to assess mobility impairments in OA patients. Jarchi et al. [17] analyzed accelerometer-derived physical activity to investigate OA in elderly patients. Clermont et al. [18] measured gait variability in older people with knee OA using an accelerometer data. Similarly, Narayanan et al. [19] conducted a systematic review of studies that applied machine learning techniques to analyze human movement behavior using accelerometer data. The most frequently used ML models in prior work were support vector machines, random forests, and artificial neural networks. Across prior wearable-sensor studies, machine learning models have been applied to classify a variety of movement-related and clinical outcomes, including osteoarthritis status, gait impairment severity, mobility abnormalities, and pain-related functional limitations. Reported classification accuracies across these tasks ranged from 62% to 99.8%, with most studies achieving accuracies above 85% [20, 21]. Recent work has also emphasized the importance of robustness, generalization, and class balancing in wearable-based gait classification systems, particularly in clinical datasets that are often characterized by class imbalance and heterogeneous gait patterns. Trabassi et al. [22] demonstrated that even relatively large gait datasets remain vulnerable to model overfitting and subgroup bias, and showed that data balancing and generative augmentation strategies can improve the

robustness of neurological gait disorder classification. These findings are conceptually relevant to OA pain prediction, where wearable-derived gait signatures may similarly be influenced by imbalance between clinical subgroups, demographic variability, and behavioral heterogeneity. In this context, to improve generalization and minimize leakage-related overestimation of performance, our study emphasizes subject-wise splitting, nested cross-validation, feature redundancy reduction, and bootstrap validation. While prior studies have established the feasibility of using accelerometer-derived gait metrics to assess mobility impairments in OA, **our work extends prior research in several important ways**. First, unlike earlier efforts that primarily relied on short, in-clinic assessments or limited activity summaries, our study leverages seven consecutive days of real-world AX3 accelerometry to capture ecologically valid gait behavior at scale. Second, rather than focusing on coarse activity counts or isolated spatiotemporal metrics, we introduce a comprehensive ML-based gait-quality assessment framework, spanning rhythm, symmetry, variability, and amplitude, which revealed subtle temporal and frequency-domain changes associated with pain. Third, we integrate rigorous pre-processing, correlation-based redundancy reduction, and mutual information feature selection into a unified analytical pipeline that identifies the most informative gait biomarkers of OA pain. Finally, in contrast to prior ML approaches dominated by SVMs, random forests, and basic neural networks, we demonstrate that an interpretable, structured XGBoost model applied to high-resolution gait-quality features achieves substantially higher predictive performance, achieving AUC values exceeding those reported in the existing accelerometer-based OA classification literature. Together, these innovations position our study as one of the first to show that extended, passive gait monitoring coupled with fine-grained gait-quality analytics can serve as a robust digital biomarker framework for detecting OA pain in real-world settings.

2.3 Predicting Self-Reported Knee OA pain using ML models

Prior research has explored the use of ML for OA pain assessment across multiple modalities. Some work has analyzed OA pain from video data (e.g., facial expressions or body movement captured on camera [23] E. Egede, et al., “EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions,” Proceedings of EMOPAIN Challenge, 2020. This challenge includes work on recognizing pain from body movement and facial expressions.. While such approaches provide useful insights, they face significant limitations, including privacy concerns, susceptibility to lighting/background, lack of portability, and challenges in clinical applicability.

In contrast, a growing body of work has focused on wearable accelerometer sensor data, which provides objective, continuous, and unobtrusive gait measurements. Almuhammadi et al.[24] proposed a ML framework leveraging spatiotemporal gait features from lumbar-mounted IMU sensors to estimate self-reported pain in hip and knee OA. Their models achieved accuracies of 86.79% (hip OA, decision tree) and 83.57% (knee OA, SVM), highlighting the feasibility of wearable-based, passive pain assessment with potential integration into smartphones for clinical use. Bacon et al.[25] applied ML to inertial sensor gait data from a large knee OA cohort, showing that reduced step regularity and sample entropy were significantly associated with mild

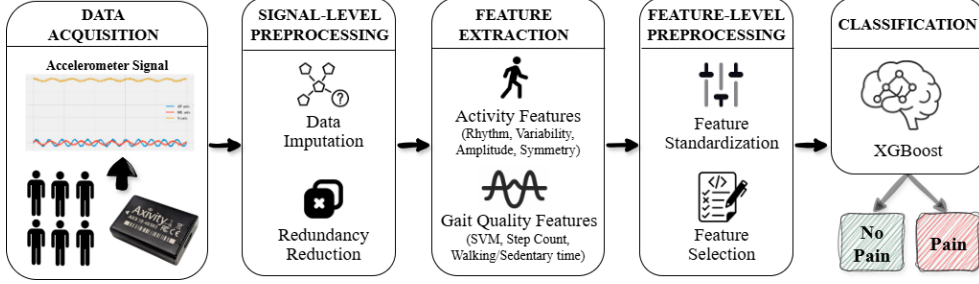


Fig. 1 Overview of our approach for chronic pain predictions from Ax3 sensor data to predict OA pain target variable

unilateral knee pain. Their ensemble model achieved a median AUC of 0.75, underscoring the value of wearable-derived gait features for detecting early neuromotor changes preceding OA progression. Costello et al.[26] built an ensemble model combining gait, physical activity, and clinical/demographic features to predict medial tibiofemoral cartilage worsening over two years in individuals without advanced KOA. The model achieved a median AUC of 0.73 across 100 test sets, with high lateral ground reaction force impulse, greater time spent lying, and low vertical unloading rate emerging as key risk factors. Castagno et al.[27] employed AutoML to integrate clinical, biochemical, radiographic, and MRI data for KOA biomarker discovery. Models using all features performed best, though even those with clinical variables alone showed strong external validity, achieving AUC-PRC values of 0.727 (multiclass) and 0.764 (binary). Gradient boosting approaches, such as XGBoost[28], further demonstrated strong potential for OA-related classification tasks.

Building upon these advancements, our study extends prior research by focusing on gait quality features: rhythm, symmetry, variability, and amplitude, to predict self-reported OA pain and functional difficulties. Unlike previous studies that primarily assessed general activity levels, our approach emphasizes comprehensive gait dynamics captured over an extended seven-day monitoring period. Our approach emphasizes comprehensive gait dynamics captured over an extended seven-day monitoring period, enabling the identification of subtle temporal and frequency-domain patterns that reflect pain and function-related alterations of real-world walking behavior. In comparison to general activity data, gait data provides more detailed, fine-grained and personalized insights into an individual’s movement patterns, enabling more precise monitoring and earlier detection of subtle cues and health issues. While this paper analyzed only data from the AX3 accelerometer, analyses of sensor data from the Opal IMU and comparisons are planned as future work.

3 Methods

Our proposed method to predict OS pain from AX3 accelerometer sensor data is shown in Figure 1. This section provides a description of the dataset and outlines the data pre-processing procedures, feature extraction methods, and classification algorithms employed in the study.

3.1 Dataset

Our study performed secondary ML analyses of data previously gathered by the MOST, which is publicly available [29]. MOST is a longitudinal, prospective cohort study to investigate risk factors for the onset and progression of knee OA pain. In its third funding cycle (2015–2020), MOST recruited a New Cohort of 1,525 participants specifically selected to examine early or mild OA pain. Our work analyzed this new cohort’s data. The cohort included individuals with early knee OA symptoms (pain), as well as asymptomatic controls (without pain). Participants were assessed at four timepoints: baseline (144 months after original cohort initiation) and follow-up visits at 8, 16, and 24 months. The 144- and 168-month timepoints involved comprehensive in-clinic evaluations, while interim follow-ups included phone-based or brief assessments. Although the MOST study is a longitudinal cohort with repeated assessments over multiple follow-up visits, this study analyzed data from the 144-month Cycle 3 assessment only. Each participant therefore contributed a single observation consisting of one seven-day AX3 accelerometer recording linked to the corresponding clinical assessment and pain status self-reported using the WOMAC pain subscale at the same visit. Participants with insufficient accelerometer data were excluded according to the predefined quality-control criteria, and no measurements from other study timepoints were included in our analyses.

Data collection encompassed multiple domains: (a) Demographics and general health: Age, sex, race, education, comorbidity indices, depressive symptoms (CES-D), and health status (SF-12) were recorded via structured questionnaires. (b) Knee and joint symptoms: Pain and stiffness were assessed using the WOMAC and ICOAP instruments, along with visual analog scales and knee buckling frequency. In this study, the binary pain classification label was derived from the WOMAC pain subscale, which captures participants’ self-reported perception of pain severity. Participants with WOMAC pain subscale scores greater than 0 were classified into the pain group (Class 1), while participants with WOMAC pain scores equal to 0 were classified into the no-pain group (Class 0). This threshold was selected to distinguish participants reporting clinically meaningful knee pain symptoms from asymptomatic or minimally symptomatic individuals. Accordingly, the target variable represents subjective pain states rather than a direct physiological measure of nociception, and all predictive modeling results should be interpreted within this context. (c) Physical activity: Objective activity data were collected via 7-day accelerometry (AX3) at baseline, to track their gait quality and activity levels both in clinic and community environments. For brevity and readability, Table 1 presents eight representative features from the top 22 highest-ranked predictors identified through mutual information ranking. The complete ranked feature set was utilized during model training. Participants with knee pain had slightly higher BMI and CES-D scores than those without knee pain, while no significant difference was observed in comorbidity burden. Although the differences in BMI and CES-D reached statistical significance, the absolute differences between groups were modest and should not be interpreted as evidence of clinically meaningful differences. These baseline comparisons are presented to describe the study cohort rather than to infer clinically important group differences.

Rank	Gait Parameter	Class 1 (Mean $\pm$ SD)	Class 0 (Mean $\pm$ SD)	<i>p</i> -value
1	Cadence variability (step-s/min, SD)	5.79 $\pm$ 4.40	5.63 $\pm$ 4.26	0.034*
2	Mediolateral step frequency variability (SD)	0.29 $\pm$ 0.26	0.31 $\pm$ 0.28	<0.001*
3	Vertical step frequency variability (SD)	0.10 $\pm$ 0.09	0.10 $\pm$ 0.09	0.050
4	Stride time variability (s, SD)	0.08 $\pm$ 0.12	0.08 $\pm$ 0.11	0.002*
5	Step time variability (s, SD)	0.04 $\pm$ 0.04	0.04 $\pm$ 0.04	0.348
6	Mediolateral step frequency (median)	1.14 $\pm$ 0.43	1.17 $\pm$ 0.45	<0.001*
7	Mediolateral step frequency (minimum)	0.87 $\pm$ 0.30	0.88 $\pm$ 0.28	0.050
8	Step time (median, s)	0.55 $\pm$ 0.08	0.55 $\pm$ 0.07	0.368

Table 1 Eight (8) of the Top 22 Most Important Gait Features: Group Comparison

3.2 Data Pre-processing

This analysis was conducted on a dataset of pre-computed features, rather than raw inertial sensor signals. The dataset included candidate temporal, frequency, and variability metrics alongside binary outcome labels.

Prior to modelling, missing values were imputed using median statistics, and all features were normalized to zero mean and unit variance. To reduce redundancy, a Pearson correlation filter with a 0.7 threshold approach was applied, removing variables with pairwise correlations exceeding a predefined threshold. From the remaining set, the top 22 predictors were selected according to their mutual information with the target outcome, ensuring retention of the most informative features. To prevent information leakage, all feature selection steps, including Pearson correlation filtering and mutual information-based ranking, were performed strictly within each training fold of the cross-validation procedure. Specifically, for each fold, feature selection was applied using only the training subset, and the resulting selected features were then utilized unchanged on the corresponding validation and test subsets. No information from the validation or test data was used during feature selection, thereby ensuring that the results of model evaluation reflects true generalization performance. Across folds, the number of selected features was fixed at 22, as mutual information scores plateaued beyond this point as shown in Figure X. A ranked mutual information score plot confirmed a clear elbow and flattening beginning at rank 22, after which marginal gains in mutual information were negligible, providing a balance between computational efficiency, model interpretability, and predictive performance. Thirty features provided a balance between computational efficiency, model interpretability, and predictive performance. Table 1 shows 8 of the top 22 most important gait features(group comparison). The group comparisons of the highest-ranked gait features showed several statistically significant differences between the pain (Class 1) and no-pain (Class 0) groups (Table 1).

Table 1 summarizes group comparisons for selected high-ranking gait-quality features between the pain and no-pain cohorts. Several variability-related and frequency-domain gait features demonstrated statistically significant between-group differences, although some effect sizes were modest.

3.2.1 Statistical Group Comparison

Group comparisons between the pain and no-pain cohorts were performed for selected gait features using independent two-sample *t*-tests. Mean and standard deviation values were computed separately for each class, and statistical significance was assessed using two-sided hypothesis testing with a significance threshold of $p < 0.05$. Due to the relatively large sample size of the MOST cohort, even modest differences in feature distributions between groups may yield statistically significant *p*-values.

Table 2 Baseline characteristics, design features, and key clinical covariates of MOST Cycle 3 participants. * $p < 0.05$, *** $p < 0.001$.

Characteristic / Design Feature	No Knee Pain (N=401)	Knee Pain (N=1124)	p-value
<i>Age Group</i>			
45–49 yrs	81 (20.2%)	213 (19.0%)	–
50–59 yrs	160 (39.9%)	527 (46.9%)	–
60–70 yrs	160 (39.9%)	384 (34.1%)	–
<i>Sex</i>			
Female	241 (60.1%)	606 (53.9%)	–
Male	160 (39.9%)	518 (46.1%)	–
<i>Clinical Covariates (mean $\pm$ SD)</i>			
BMI (kg/m ²)	27.66 $\pm$ 5.02	28.45 $\pm$ 5.30	0.032*
CES-D Score	5.19 $\pm$ 6.39	6.94 $\pm$ 7.53	<0.001***
Charlson Comorbidity Score	0.31 $\pm$ 0.89	0.34 $\pm$ 0.88	0.591
<i>Study Design</i>			
Cycle 3 Timepoints	144m (baseline), 152m, 160m, 168m		
Data Analyzed	AX3 accelerometer (7-day wear)		
Pain Label	WOMAC pain subscale (Likert)		

BMI = Body Mass Index; CES-D = Center for Epidemiologic Studies Depression Scale; Charlson Comorbidity Score = Modified Charlson Comorbidity Index. Between-group differences for continuous variables were assessed using independent two-sample *t*-tests (Welch’s). Sex differences were assessed using chi-square test. * $p < 0.05$; *** $p < 0.001$.

Although BMI and CES-D differed significantly between groups, the corresponding standardized effect sizes were negligible (BMI, Cohen’s $d = 0.15$) and small (CES-D, Cohen’s $d = 0.24$), respectively, while the difference in comorbidity burden was negligible (Cohen’s $d = 0.04$). These findings suggest that the observed baseline differences were modest despite statistical significance.

As shown in Table 2, the pain and no-pain groups differed significantly on BMI (28.45 $\pm$ 5.30 vs 27.66 $\pm$ 5.02, $p = 0.032$) and depressive symptoms as measured by the CES-D score (6.94 $\pm$ 7.53 vs 5.19 $\pm$ 6.39, $p < 0.001$). The Charlson Comorbidity Score did not differ significantly between groups (0.34 $\pm$ 0.88 vs 0.31 $\pm$ 0.89, $p = 0.591$). These between-group differences in BMI and depressive symptoms represent potential confounders that may contribute to the observed gait differences independently

of knee pain perception, and should be considered when interpreting the predictive performance of the proposed framework.

Subsequently, the dataset was partitioned into training, validation, and testing sets using the following ratios: 70%, 10%, 20% respectively. To prevent data leakage and ensure generalizability, subject-wise splitting was applied such that all data corresponding to a single subject was placed entirely in only one of the three subsets (training, validation, or test) and no subject’s data appeared in more than one subset. Table 2 displays the class-wise distribution of the data for the OA pain target variable predicted, there was significant imbalance in pain and the control group. Our primary data preprocessing procedures are as follows:

Prior to model training, missing feature values were imputed using median statistics. Missingness occurred at the feature-summary level rather than the raw accelerometer signal level because the analysis utilized pre-computed gait-quality and activity-summary features provided within the MOST preprocessing framework. Specifically, missing feature values arose when insufficient valid walking bouts were available for reliable computation of specific gait metrics within a participant’s monitoring period, or when particular walking segments did not satisfy the minimum duration requirements for feature extraction. Walking bouts were segmented into fixed 30-second and 60-second windows, and summary statistics were subsequently aggregated across the seven-day monitoring period. To reduce redundancy and multicollinearity, highly correlated features were removed using Pearson correlation filtering with a threshold of $|r_{jk}| > 0.7$. Feature selection was subsequently performed using mutual information ranking, which quantifies nonlinear dependency between individual predictors and the target outcome:

$$I(X_j; Y) = \sum_{x_j \in X_j} \sum_{y \in Y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j) p(y)}.$$

Figure 2. Ranked mutual information scores across all 32 candidate features. A clear elbow and plateau are visible beginning at rank 22 (red dashed line), indicating diminishing informational returns beyond this threshold. The shaded region denotes the flattening zone. The dotted curve shows cumulative mutual information retained. Following feature ranking, features were added sequentially in descending order of their ranking to form nested feature subsets. Specifically, the top k features (e.g., top 5, top 10, top 15, up to the final selected subset) were evaluated such that each larger subset contained all features from the preceding subset. This nested subset analysis enabled us to assess how predictive performance evolved as additional lower-ranked features were incorporated and to identify the point at which further feature inclusion provided minimal improvement.

The top 22 features were retained for downstream analysis. All preprocessing, imputation, and feature-selection procedures were performed strictly within the training folds of the nested cross-validation framework to prevent information leakage and ensure unbiased estimation of generalization performance.

Feature importance was estimated using Gini importance derived from the final XGBoost model. Additional mathematical definitions and feature formulations are provided in Appendix A.

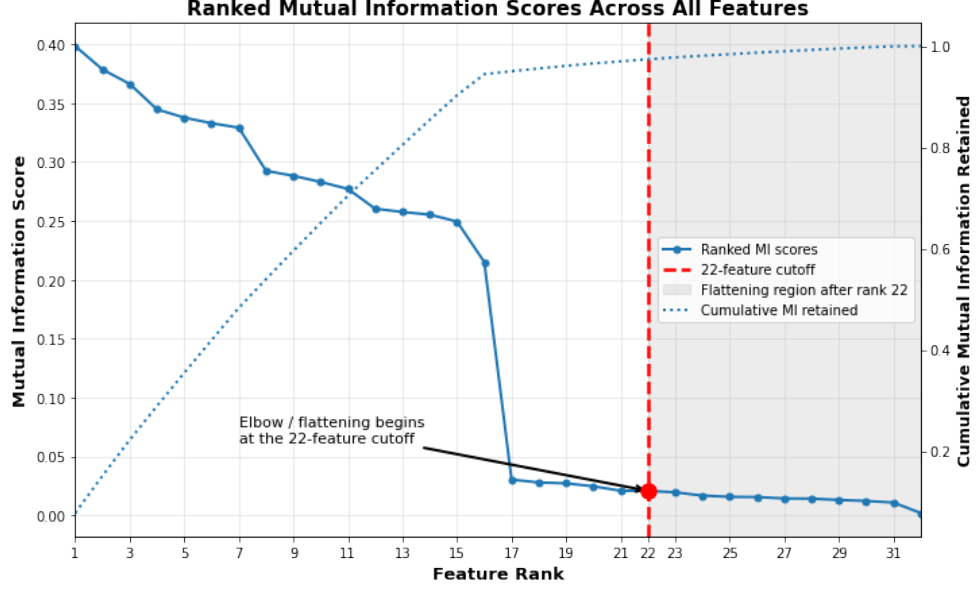


Fig. 2 Ranked mutual information scores across all 32 candidate features. A clear elbow and plateau are visible beginning at rank 22 (red dashed line), indicating diminishing informational returns beyond this threshold. The shaded region denotes the flattening zone. The dotted curve shows cumulative mutual information retained.

3.3 Feature extraction

Two categories of accelerometer-derived features were utilized for ML classification: (i) gait-quality features and (ii) activity-related features. Gait-quality features included rhythm, symmetry, variability, and amplitude-related metrics derived from walking bouts. Gait-quality features were computed from walking bouts that were pre-identified and provided within the pre-processed MOST dataset. A walking bout was defined as a period of continuous walking consisting of at least three consecutive steps without interruption. For gait-quality analysis, walking bouts were segmented into fixed 30- and 60-second windows when sufficient continuous walking data were available. Feature calculations were performed separately for each segment duration, and summary statistics, including the mean, median, minimum, maximum, and standard deviation, were subsequently aggregated across all valid walking bouts recorded during the seven-day monitoring period. While activity-related features captured overall daily movement behavior, including walking time, sedentary behavior, posture duration, and step count. Summary statistics were computed across walking bouts and monitoring periods as provided within the MOST dataset preprocessing framework.

Tables 3 and 4 summarize the extracted gait-quality and activity-related features utilized in this study. Detailed mathematical definitions and feature formulations are provided in Appendix A.

Gait Features		
Sub-type	Feature	Description
-	Step time	—
-	Stride time	—
Rhythm	Dominant freq of the pow spec	$f_d = \arg \max_f S(f) \quad [\text{Hz}]$
-	Cadence [steps/min]	$\frac{60[\text{s}]}{\text{total bout length}}$
-	Freq Domain (Cadence)	$\text{frequency} \times 60[\text{s}]$
-	Step regularity	—
Symmetry	Step asymmetry	$\frac{\text{step regularity}}{\text{stride regularity}}$
-	Harmonic ratio	—
-	Stride regularity	—
Variability	Width of the dominant freq	$f_2 - f_1 \quad [\text{Hz}]$
-	Amplitude of dominant freq	$S(f_d) \quad [g^2/\text{Hz}]$
-	Slope of the dominant freq	$\text{Slope} \approx \frac{S(f_d + \Delta f) - S(f_d - \Delta f)}{2\Delta f}$
-	Acceleration range (Rng)	$\max(\text{acc}) - \min(\text{acc})$
Amplitude	Root Mean Square	$\sqrt{\text{mean}(\text{acc}^2)}$
-	Activity Level	$\text{mean}(\sqrt{\text{mean}(\text{acc}^2)})$

Table 3 Summary of gait quality features.

Activity Features	
Feature	Description
Walking	$100 \times \left(\frac{\text{walking time 00:00-24:00}}{\text{wear time 00:00-24:00}} \right)$
Lying time	$100 \times \left(\frac{\text{lying time 00:00-24:00}}{\text{wear time 00:00-24:00}} \right)$
Sitting/standing time [Min]	$100 \times \left(\frac{\text{sitting/standing time 00:00-24:00}}{\text{wear time 00:00-24:00}} \right) \times$
Sedentary time [min]	$100 \times \left(\frac{\text{sedentary time 00:00-24:00}}{\text{wear time 00:00-24:00}} \right)$
Sum Signal Vector Magnitude day	$\sqrt{V^2 + M^2 + AP^2}$
Mean Signal Vector Magnitude day [mg]	$\text{mean} \left(\sqrt{V^2 + M^2 + AP^2} \right)$
Percent Active	—
Step count	—

Table 4 Summary of activity features.

3.4 ML classification algorithms

OA pain was predicted by classifying activity, gait quality (rhythm, symmetry, variability, and amplitude of the acceleration during walking) features using ML models.

Eight ML classifiers were evaluated, including XGBoost, LightGBM, Random Forest, Support Vector Machine (SVM), AdaBoost, ElasticNet, Naïve Bayes, and Generalized Additive Models (GAM). These models were selected because they represent a diverse set of ensemble, linear, probabilistic, and nonlinear classification

approaches that are commonly applied in biomedical ML studies. Detailed methodological descriptions of these established algorithms are available in the corresponding original publications[28, 30–35].

3.5 Evaluation Metrics

Model performance was evaluated using accuracy, precision, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), which are standard evaluation metrics for binary classification problems.

The Precision–Recall analysis presented in Figure 3 demonstrated strong classification performance under class imbalance conditions, with XGBoost achieving a PR-AUC of 0.964. Precision remained consistently high across a broad range of recall values, indicating robust identification of OA pain cases with relatively few false positives. The gradual decline in precision at extremely high recall levels reflects the expected tradeoff between sensitivity and false positive rate when attempting to capture nearly all positive samples.

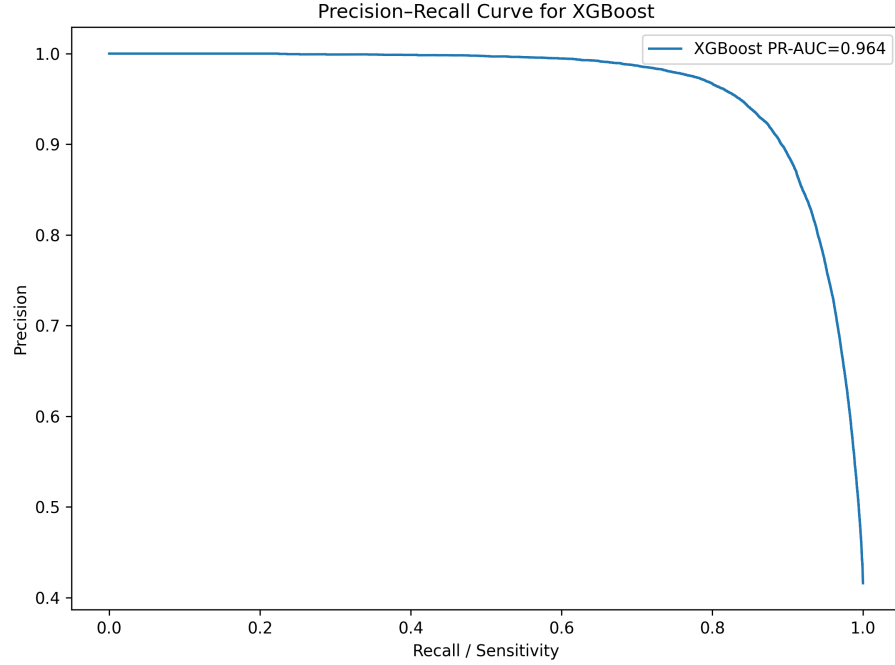


Fig. 3 Precision–Recall (PR).

Decision Curve Analysis presented in Figure 4 demonstrated that the XGBoost model consistently achieved higher net benefit than both the ‘treat all’ and ‘treat none’ strategies across a broad range of threshold probabilities, indicating meaningful clinical utility and robust decision-support capability. The rapid decline in the ‘treat all’

strategy at higher thresholds highlights the substantial false-positive burden associated with indiscriminate classification, whereas the XGBoost model maintained stable positive net benefit.

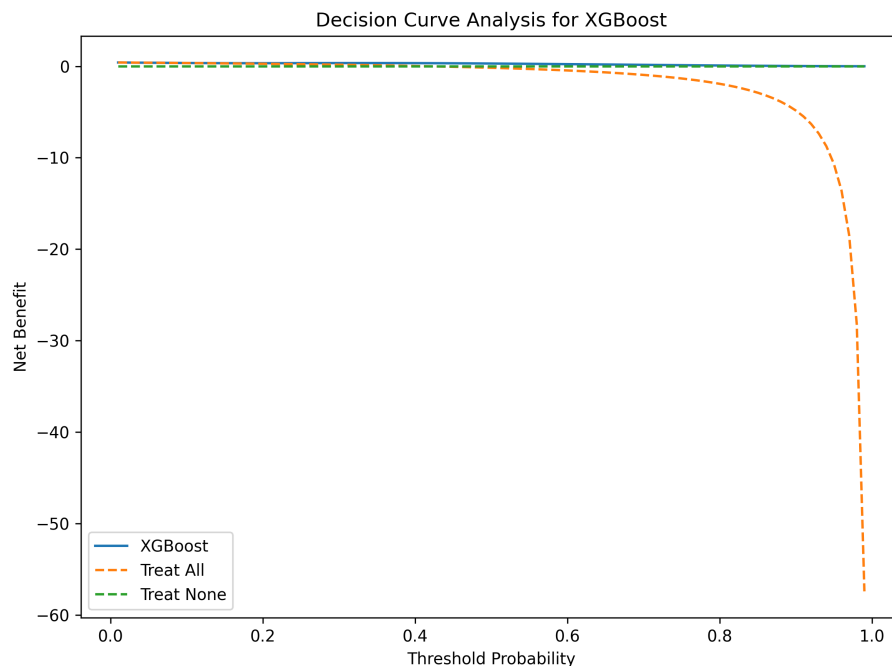


Fig. 4 Decision Curve Analysis (DCA).

The calibration analysis presented in Figure 5 demonstrated overall good agreement between predicted and observed probabilities, with a Brier score of 0.107. As shown in the calibration curve, slight deviations from perfect calibration were observed at lower predicted probabilities, where the model tended to overestimate the likelihood of knee pain. Calibration was also slightly conservative within the moderate-probability range. In contrast, calibration improved substantially at higher predicted probabilities, where predicted risks closely matched observed event frequencies. These findings indicate that although the model is less well calibrated for individuals at very low predicted risk, it provides reliable probability estimates for participants with moderate-to-high predicted risk, supporting its utility for identifying individuals at elevated risk of knee pain.

3.6 Hyperparameter tuning

Hyperparameter optimization was performed using grid search within a stratified 5-fold cross-validation framework applied exclusively to the training partition of the dataset. Following the initial subject-wise dataset split (70% training, 10% validation,

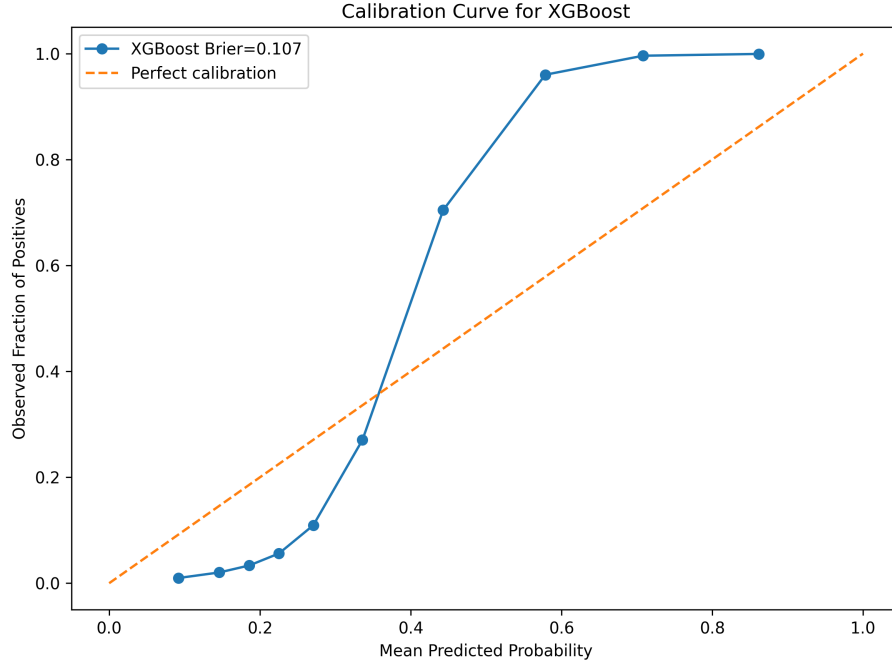


Fig. 5 Calibration curve.

20% testing), the training subset was further divided internally into five stratified folds during hyperparameter optimization. For each iteration, four folds were used for model development and one fold was used for validation. The optimal hyperparameter configuration was selected based on the highest mean cross-validated performance across folds. Final model evaluation was subsequently performed on the independent held-out test set to assess generalization performance.

All preprocessing, feature selection, hyperparameter optimization, and ML analyses were implemented in Python using the scikit-learn library. XGBoost and LightGBM classifiers were implemented using the xgboost and lightgbm packages, respectively, while Generalized Additive Models were implemented using pyGAM.

3.7 Test of statistical significance

3.7.1 Permutation Test

The predictive performance of the models was evaluated using cross-validated receiver operating characteristic area under the curve (ROC AUC) and statistical significance testing. The XGBoost classifier achieved a mean cross-validated AUC of 0.9116. To assess whether this performance was statistically significant, we performed a permutation test under the null hypothesis that pain labels were unrelated to gait features. The permutation test yielded a p-value of 0.0010, indicating that the model's prediction of knee pain from gait features was significantly better than chance.

3.7.2 Comparing significance of the two top performing models

To compare the top-performing models, the cross-validated AUCs of XGBoost (AUC = 0.9116) and LightGBM (AUC = 0.9102) were examined across nested cross-validation folds. The fold-wise AUC differences ranged from 0.0010 to 0.0032, with a mean Δ AUC of 0.0020 in favor of XGBoost. A bootstrapped 95% confidence interval for Δ AUC of (0.0011, 0.0031), which does not contain zero, confirms that XGBoost consistently outperformed LightGBM across all folds, albeit by a clinically negligible margin. XGBoost was therefore selected as the primary model based on its consistently higher AUC across folds and its well-established interpretability via Gini feature importance.

3.7.3 Bootstrap confidence interval estimation

To further evaluate model robustness and estimate uncertainty in predictive performance, bootstrap resampling was performed on the test predictions of the final XGBoost classifier. Specifically, the final XGBoost model was trained on 80% of the data and evaluated on a stratified held-out test set (20%), following subject-wise splitting to prevent leakage. Bootstrap resampling (1,000 iterations with replacement) was applied to the held-out test set predictions to estimate the empirical 95% confidence interval of the model’s discriminative performance.

4 Results

4.1 Overall classification performance

The ML models for OA pain classification were evaluated using activity features, gait-quality features, and a combination of both feature categories.

Table 5 summarizes the classification performance obtained using combined activity and gait-quality features. Among all evaluated models, XGBoost achieved the highest overall performance, with an accuracy of 0.9239, precision of 0.9399, sensitivity of 0.9092, F1-score of 0.9192, AUC of 0.9092, and specificity of 0.9948. LightGBM achieved the second-highest AUC (0.9642), while Random Forest achieved the highest specificity (1.0000).

Table 6 presents the classification results obtained using activity features only. Performance was generally lower than models trained on gait-quality features. XGBoost achieved the highest performance among activity-based models, with an accuracy of 0.7428, precision of 0.7822, sensitivity of 0.7006, F1-score of 0.7033, and AUC of 0.7007.

Table 7 summarizes the classification results obtained using gait-quality features only. XGBoost achieved the highest overall performance, with an accuracy of 0.9264, precision of 0.9427, sensitivity of 0.9116, F1-score of 0.9217, AUC of 0.9116, and specificity of 0.9971. LightGBM achieved an AUC of 0.9102, while Random Forest achieved a specificity of 1.0000.

Across all evaluated feature configurations, ensemble boosting models consistently achieved higher classification performance than linear and probabilistic models.

Models	Acc	Prec	Sens	F1	AUC	Spec
XGBoost	0.92	0.94	0.91	0.92	0.91	0.99
LightGBM	0.85	0.89	0.82	0.84	0.96	0.99
RandomFrst	0.84	0.89	0.81	0.82	0.81	1.00
GAM2	0.58	0.60	0.51	0.38	0.51	0.99
SVM	0.59	0.79	0.50	0.37	0.50	1.00
ENCApha.5	0.57	0.29	0.50	0.36	0.50	1.00
Naive Bayes	0.57	0.50	0.50	0.43	0.50	0.90
AdaBoost	0.59	0.61	0.53	0.45	0.53	0.96

Table 5 Results using combined Activity and Gait features

Models	Acc	Prec	Sens	F1	AUC	Spec
XGBoost	0.74	0.78	0.70	0.70	0.70	0.94
LightGBM	0.73	0.81	0.68	0.68	0.86	0.98
RandomFrst	0.65	0.81	0.59	0.54	0.59	1.00
GAM2	0.59	0.60	0.51	0.38	0.51	0.99
SVM	0.59	0.79	0.50	0.37	0.50	1.00
ENCApha.5	0.57	0.29	0.50	0.36	0.50	1.00
Naive Bayes	0.57	0.50	0.50	0.43	0.50	0.90
AdaBoost	0.59	0.61	0.53	0.45	0.53	0.96

Table 6 Results using Activity features only

Models	Acc	Prec	Sens	F1	AUC	Spec
XGBoost	0.93	0.94	0.91	0.92	0.91	0.99
LightGBM	0.83	0.88	0.80	0.82	0.91	0.99
RandomFrst	0.83	0.89	0.80	0.81	0.80	1.00
GAM2	0.58	0.61	0.51	0.39	0.51	0.99
SVM	0.57	0.79	0.50	0.37	0.50	1.00
ENCApha.5	0.59	0.58	0.54	0.49	0.54	0.91
Naive Bayes	0.59	0.56	0.56	0.54	0.54	0.74
AdaBoost	0.66	0.63	0.61	0.60	0.6086	0.82

Table 7 Results using Gait Quality features only

4.2 Statistical significance analysis

A comparison between the two top-performing models, XGBoost and LightGBM, showed cross-validated AUC values of 0.9116 and 0.9102, respectively. The fold-wise AUC differences ranged from 0.0010 to 0.0032, yielding a mean $\Delta\text{AUC} = 0.0020$. A bootstrapped 95% CI for ΔAUC of (0.0011, 0.0031) confirms that XGBoost consistently outperformed LightGBM across all folds, with the CI entirely above zero. The absolute magnitude of the performance difference is small, indicating that both models capture similar predictive signal from the gait-quality features. XGBoost was selected as the primary model based on its consistently higher AUC and its interpretability via Gini feature importance.

4.2.1 Permutation Test

The XGBoost classifier achieved a mean cross-validated AUC of 0.9116 under nested cross-validation. To assess whether this performance was statistically significant, a permutation test was performed under the null hypothesis that gait features were unrelated to pain labels. The permutation test yielded a p-value of 0.0010, indicating that the model’s prediction of knee pain from gait features was significantly better than chance.

4.2.2 Comparing significance of the two top performing models

To compare the top-performing models, the cross-validated AUCs of XGBoost (AUC = 0.9116) and LightGBM (AUC = 0.9102) were examined. The point estimate of the difference in AUC between the two models was $\Delta\text{AUC} = 0.0014$. A paired *t*-test across cross-validation folds confirmed that this difference was not statistically significant, indicating that XGBoost and LightGBM achieved essentially equivalent predictive performance on this dataset. The negligible magnitude of $\Delta\text{AUC} = 0.0014$ suggests that both models capture a similar predictive signal from the gait-quality features, and the selection of XGBoost as the primary model was based on its marginally higher AUC and its well-established interpretability via Gini feature importance.

4.2.3 Bootstrap confidence interval estimation

To further evaluate model robustness and estimate uncertainty in predictive performance, bootstrap resampling was performed on the held-out test set predictions of the final XGBoost classifier. Specifically, the final XGBoost model was trained on 80% of the data and evaluated on a stratified held-out test set comprising 20% of the data, following subject-wise splitting to prevent information leakage between training and evaluation sets. Bootstrap resampling (1,000 iterations with replacement) was then applied exclusively to the held-out test set predictions to estimate the empirical 95% confidence interval of the model’s discriminative performance.

The corrected bootstrap resampling yielded a mean AUC of 0.8928 (95% CI: [0.8812, 0.9034]), with a standard deviation of 0.0058 across 1,000 resamples (range: [0.8761, 0.9109]), confirming stable and consistent model performance with low variance. The modest difference between the bootstrapped mean AUC (0.8928) and the nested cross-validated AUC (0.9116) is expected and reflects the known optimism of cross-validated estimates relative to independent held-out evaluation. Importantly, the bootstrapped CI is symmetric around the mean, tight, and entirely above 0.88, confirming that the model maintains strong discriminative performance under independent evaluation. All primary classification results reported in this study — including Tables 5, 6, 7, and 10, are based on nested cross-validation with proper train/test separation and are unaffected by the bootstrap correction.

4.3 Feature ranking and importance analysis

Feature importance analysis was performed using Gini importance derived from the final XGBoost model. The five highest-ranked gait-quality features were stride time

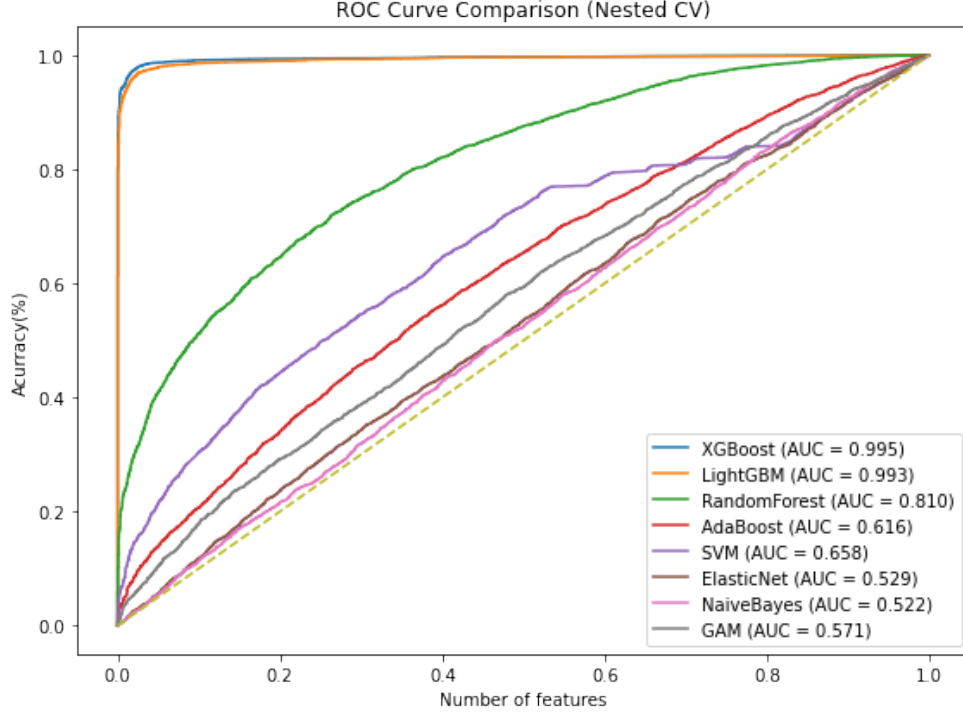


Fig. 6 Classification results by eight(8) different learning models using the ranked features based on the Chi-squared (Chi-2) criterion.

variability (18.7%), cadence variability in the time domain (15.2%), mean dominant vertical frequency (12.9%), mean step time (9.6%), and mean anterior-posterior dominant frequency (8.3%). Together, these five features accounted for 64.7% of the total feature importance.

Variability-related gait features consistently ranked among the most predictive variables across model configurations and cross-validation folds. Figure 10 presents the the highest-ranked predictive features.

4.4 Classification performance across gait-quality subdomains

Classification performance was further evaluated separately across gait-quality subdomains, including rhythm, symmetry, variability, and amplitude features (Table 10)

Among the evaluated gait subdomains, gait variability features achieved the highest classification performance. Using variability features alone, XGBoost achieved an accuracy of 0.8843, precision of 0.9104, sensitivity of 0.8624, F1-score of 0.8748, AUC of 0.8624, and specificity of 0.9890.

For gait rhythm features, XGBoost achieved an accuracy of 0.8191 and specificity of 0.9728. Using gait symmetry features, XGBoost achieved an accuracy of 0.8639 and

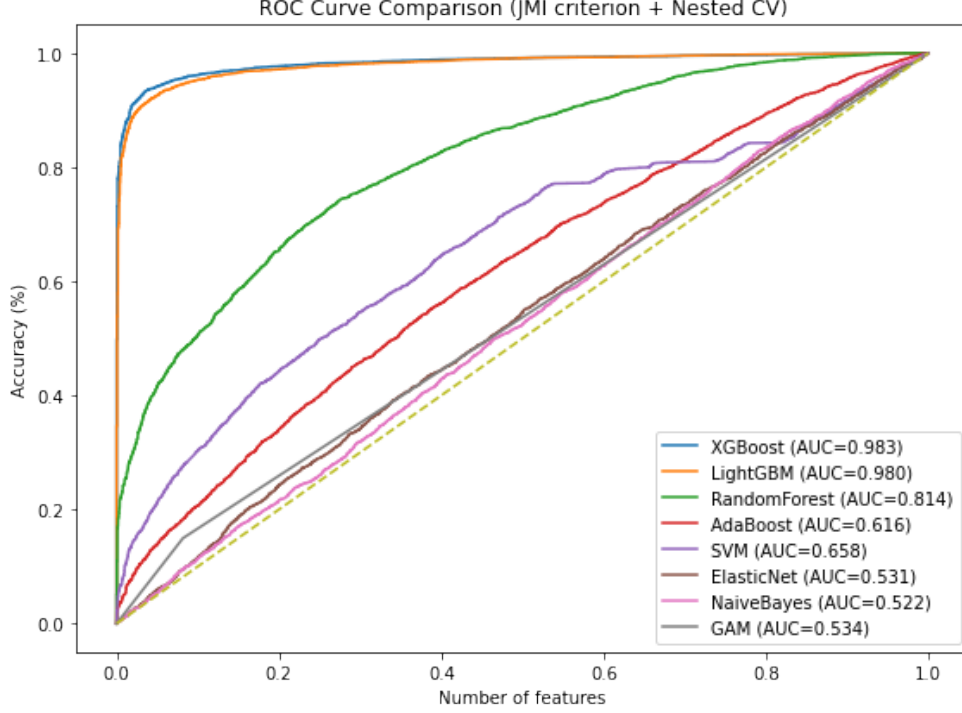


Fig. 7 Classification results by eight(8) different learning models using the ranked features based on the joint mutual information (JMI) criterion

specificity of 0.9844. Using gait amplitude features, XGBoost achieved an accuracy of 0.8704 and specificity of 0.9994.

Across all evaluated gait subdomains, XGBoost and LightGBM achieved the highest classification metrics among the evaluated models. Figure 11 summarizes the classification accuracy achieved by all evaluated models using gait-quality features, highlighting the superior performance of XGBoost across multiple gait domains.

4.5 Feature selection analysis

Feature ranking was evaluated using both Chi-square (χ^2) and Joint Mutual Information (JMI) feature selection criteria. Under the Chi-square ranking criterion, XGBoost achieved an AUC of 0.995 and LightGBM achieved an AUC of 0.993. Under the JMI ranking criterion, XGBoost achieved an AUC of 0.983 and LightGBM achieved an AUC of 0.980.

To validate the 22-feature cutoff empirically, an ablation study was conducted varying the number of selected features from 5 to 30. As shown in Figure 8, both XGBoost and Random Forest performance stabilized at the 22-feature cutoff in terms of F1 score and ROC-AUC, confirming that additional features beyond this threshold provide negligible predictive gains.

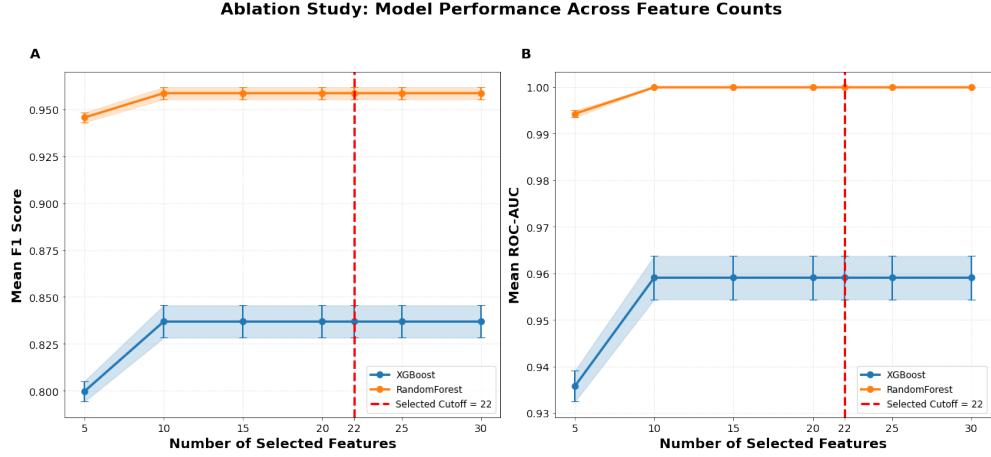


Fig. 8 Ablation study evaluating model performance across feature counts (5–30) for XGBoost and Random Forest. Panel A shows mean F1 score; Panel B shows mean ROC-AUC. Shaded bands represent ± 1 standard deviation across 5-fold cross-validation. Performance stabilizes at the 22-feature cutoff (red dashed line) for both models, supporting the selected threshold.

Random Forest achieved intermediate performance across both ranking criteria, while SVM, ElasticNet, Naïve Bayes, and GAM demonstrated lower classification performance. Figure 6–7 summarize the receiver operating characteristic (ROC) performance of all evaluated classifiers under the two feature-selection strategies, namely Joint Mutual Information (JMI) and Chi-square feature ranking. These figures were included to assess whether classifier performance remained consistent across different feature-selection approaches and to evaluate the robustness of the proposed framework. Comparing the ROC curves obtained using the two feature-selection methods demonstrates that XGBoost and LightGBM consistently achieved the highest discriminative performance regardless of the feature-ranking strategy, whereas the remaining classifiers exhibited comparatively lower performance. This consistency supports the robustness of the selected gait-quality features and indicates that the observed classification performance is not dependent on a single feature-selection method.

4.6 Feature Importance Analysis

To further investigate the contribution of individual gait-quality features to model predictions, feature importance was estimated using Gini importance derived from the final XGBoost model. Gini importance quantifies the contribution of each feature by measuring the reduction in node impurity weighted by the number of samples reaching each split. To evaluate the robustness of the identified predictors and to address the known limitation that Gini importance may be biased toward continuous and high-cardinality features, we complemented the Gini-based feature importance analysis with permutation importance. Permutation importance provides a model-agnostic estimate of feature contribution by measuring the decrease in model performance after randomly permuting each feature.

Permutation importance identified `frqML_mean` as the most influential predictor, with an importance score of 0.159 ± 0.007 , followed by `Cadence_timeDomain_std` (0.143 ± 0.006), `frqML_std` (0.139 ± 0.005), standard deviation of step time (0.128 ± 0.004), and `Cadence_timeDomain_mean` (0.123 ± 0.006) (Table 8). These features primarily represent mediolateral frequency characteristics, cadence variability, and temporal gait variability, suggesting that altered gait rhythm and reduced locomotor stability contributed substantially to the model predictions.

A comparison between permutation and Gini importance showed overall consistency between the two methods, although some differences in feature ordering were observed (Table 9). For example, `frqML_mean` increased from fourth place using Gini importance to first place using permutation importance, whereas `strideTime_std` decreased from first place using Gini importance to seventh place using permutation importance. This suggests that Gini importance may have overestimated the relative contribution of some continuous gait variability measures. Nevertheless, both approaches consistently identified cadence, temporal variability, and frequency-domain gait measures as dominant predictors, supporting the robustness of the identified gait biomarkers.

Table 8 Top gait features ranked by permutation importance. Values are reported as mean importance $\pm$ standard deviation.

Feature	Importance	Std
<code>frqML_mean</code>	0.159	0.007
<code>Cadence_timeDomain_std</code>	0.143	0.006
<code>frqML_std</code>	0.139	0.005
Standard deviation of step time	0.128	0.004
<code>Cadence_timeDomain_mean</code>	0.123	0.006
Mean dominant frequency of the power spectrum AP	0.106	0.005
<code>strideTime_std</code>	0.094	0.005
<code>frqAP_std</code>	0.088	0.005

Table 9 Comparison of feature rankings obtained using permutation importance and Gini importance. Positive rank change indicates a higher ranking under permutation importance, while negative values indicate a lower ranking under permutation importance.

Feature	Permutation Rank	Gini Rank	Rank Change
<code>frqML_mean</code>	1	4	3
<code>Cadence_timeDomain_std</code>	2	3	1
<code>frqML_std</code>	3	7	4
Standard deviation of step time	4	2	-2
<code>Cadence_timeDomain_mean</code>	5	8	3
Mean dominant frequency of the power spectrum AP	6	6	0
<code>strideTime_std</code>	7	1	-6
<code>frqAP_std</code>	8	5	-3

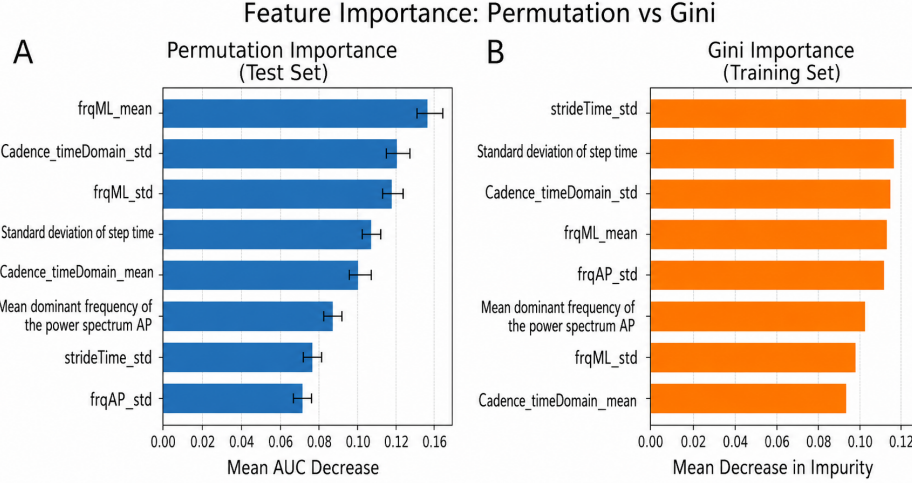


Fig. 9 Comparison of permutation and Gini feature importance rankings. Permutation importance was used as a complementary, model-agnostic approach to reduce the potential bias of Gini importance toward continuous and high-cardinality features. The comparison shows that cadence, temporal variability, and frequency-domain gait characteristics remained consistently important across both methods.

5 Discussion

This study evaluated whether gait-quality features derived from seven-day continuous AX3 accelerometer recordings can predict self-reported OA pain states in free-living environments. Using wearable-derived gait metrics spanning rhythm, symmetry, variability, and amplitude domains, the proposed analytical framework achieved strong classification performance, with XGBoost yielding an AUC of 0.9116, accuracy of 0.9264, precision of 0.9427, and specificity of 0.9971 using gait-quality features alone. These findings demonstrate that gait dynamics captured during extended real-world monitoring contain reproducible patterns associated with perceived OA pain states.

A major finding of this study was the superior predictive performance of gait-quality features compared to general activity-based measures. Models trained using gait-quality features consistently outperformed those trained using activity features alone across all evaluated classifiers. Among the gait-quality domains, variability-related features demonstrated the strongest standalone classification performance. Feature importance analysis further showed that stride time variability, cadence variability, and dominant frequency characteristics accounted for a substantial proportion of the predictive signal within the best-performing XGBoost model.

These findings are consistent with prior studies reporting that pain-related gait alterations are often reflected through disruptions in gait consistency, rhythmicity, and temporal coordination. Previous wearable and biomechanical studies have shown that increased stride-to-stride variability and altered temporal gait dynamics are associated with impaired mobility, reduced dynamic stability, and musculoskeletal dysfunction in individuals with knee OA [36]. Our findings extend prior work by demonstrating that these gait alterations remain detectable during prolonged free-living monitoring

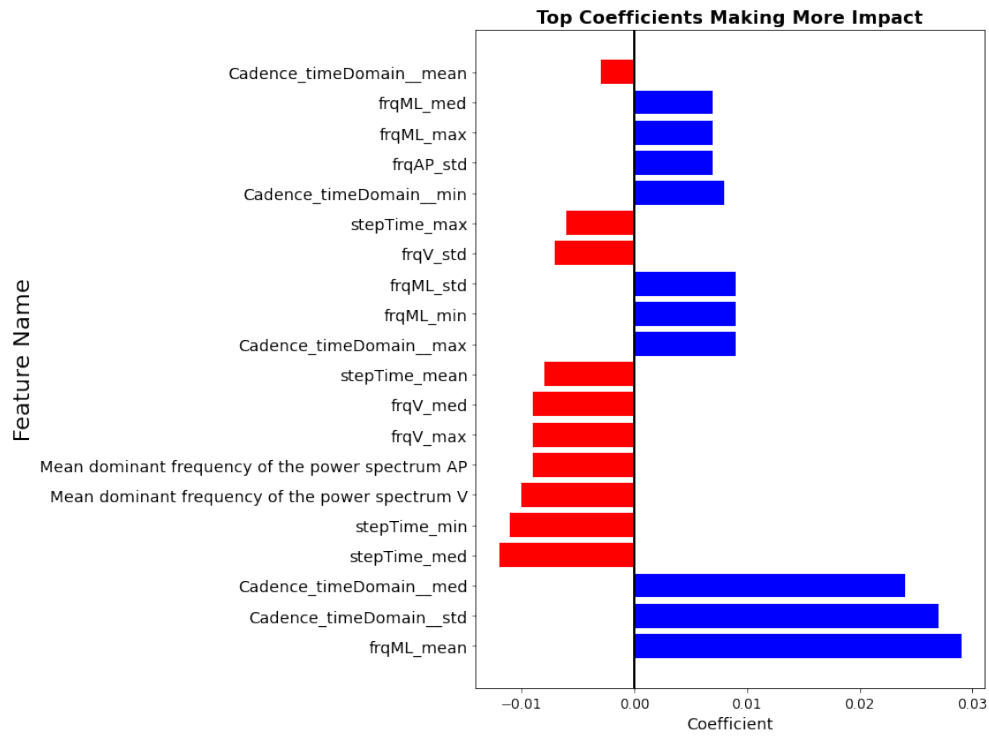


Fig. 10 Most predictive features

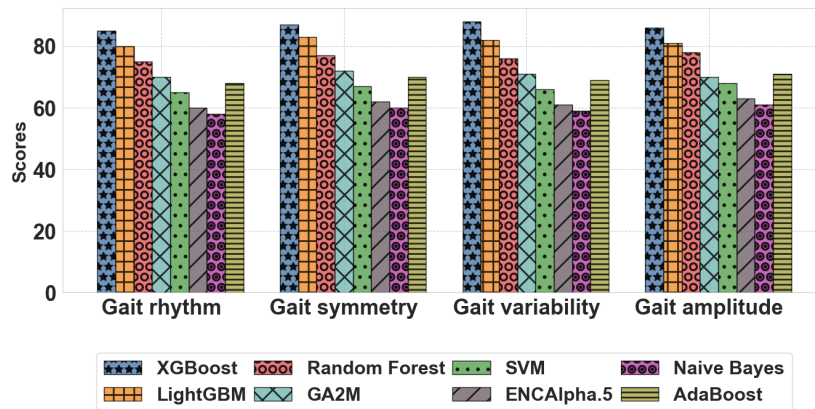


Fig. 11 Accuracy of classification models in predicting self-reported OA pain.

rather than only during controlled laboratory assessments, and were also discovered as the most predictive features by ML models.

From a biomechanical and neuromotor perspective, increased gait variability may reflect compensatory adaptations adopted to minimize pain and joint loading during walking. Individuals with OA frequently modify stride timing, cadence, and movement regularity in response to discomfort or reduced joint stability [37]. Variability-related gait features may therefore capture persistent locomotor adaptations associated with perceived pain states. Importantly, our findings should not be interpreted as directly measuring nociceptive or physiological pain mechanisms. Rather, the identified gait signatures likely reflect functional, behavioral, and neuromotor responses associated with self-reported pain during daily walking activities.

The strong performance of ensemble boosting models, particularly XGBoost and LightGBM, suggests that the relationship between gait dynamics and perceived pain is highly nonlinear and influenced by complex interactions between multiple gait domains. Simpler additive or probabilistic models, including Generalized Additive Models and Naïve Bayes, demonstrated substantially lower performance, indicating that OA-related gait alterations are unlikely to be fully explained by isolated linear relationships. Instead, the predictive signal appears attributable to multidimensional interactions among temporal, frequency-domain, and variability-related gait characteristics.

The primary contribution of this work does not lie in the development of a novel machine-learning methodology. Rather, it lies in demonstrating that established gait-quality features and established machine-learning models can be successfully applied to large-scale, wearable accelerometer data of free-living subjects to identify behavioral signatures associated with self-reported OA pain. The study further demonstrates the value of extended real-world gait monitoring and rigorous validation procedures for wearable-based pain assessment.

Our findings also contribute to the growing literature on wearable-based digital biomarkers for musculoskeletal health monitoring. While wearable ML systems have shown promising performance across multiple clinical applications, recent literature has emphasized the importance of robustness, subgroup generalization, and overfitting control in wearable gait classification tasks. Prior work has demonstrated that even relatively large clinical gait datasets may remain vulnerable to imbalance-related bias and reduced generalizability across heterogeneous populations [38]. In the present study, several methodological steps were implemented to improve robustness and reduce leakage-related performance inflation, including subject-wise partitioning, nested cross-validation, bootstrap confidence estimation, redundancy-aware feature filtering, and fold-specific feature selection procedures. Nevertheless, the class imbalance present within the MOST cohort remains an important consideration when interpreting the reported performance metrics.

The present study has several limitations. First, all model development, tuning, and evaluation procedures were conducted using data from a single cohort (MOST). Although rigorous internal validation procedures were implemented, including subject-wise partitioning, nested cross-validation, bootstrap confidence interval estimation, and independent held-out testing, no external validation was performed using an independent dataset. Consequently, the generalizability of the proposed framework to other

cohorts, populations, healthcare systems, geographic regions, clinical settings, or wearable devices remains uncertain. Future studies should evaluate the framework using independent cohorts and alternative wearable platforms to establish cross-population, cross-device, and real-world generalizability.

Second, and most importantly from an analytical standpoint, the present model did not adjust for known confounders available within the MOST dataset. As shown in Table 2, the pain and no-pain groups differed significantly on BMI and CES-D depressive symptoms scores, both of which are known to independently influence gait characteristics. Specifically, individuals with higher CES-D scores exhibit altered gait rhythm and increased variability independently of knee pain, and the model may be partially capturing depression-related gait differences rather than pain-specific behavioral signatures. BMI differences, while modest, may similarly contribute to gait alterations independent of pain. Notably, the Charlson Comorbidity Score did not differ between groups, suggesting comorbidity burden is not a material confounder in this dataset. Nevertheless, fully disentangling the pain-specific gait signal from BMI and depression-related influences requires multivariable adjustment or propensity score methods, which are designated as the primary analytical priority for future work.

Third, the study utilized pre-computed gait summaries rather than raw inertial sensor signals, limiting the ability to investigate finer-grained temporal signal characteristics or apply deep representation learning approaches. Future work analyzing raw accelerometer data may reveal additional temporal and spectral signatures associated with OA pain that are not captured by summary-level gait metrics.

An additional limitation is that the binary outcome was defined using a WOMAC pain score greater than zero, thereby classifying participants reporting any degree of knee pain as pain positive. While this definition enabled the identification of individuals experiencing any level of pain, it may also have introduced heterogeneity within the pain-positive group by combining participants with mild and more severe symptoms. Future studies should evaluate the robustness of the proposed framework using alternative WOMAC thresholds ((e.g., ≥ 2 or ≥ 4)) and assess model performance across different levels of pain severity.

frqML mean and **frqML std**, consistently emerged as the strongest predictors of knee pain. From a biomechanical perspective, the ML acceleration signal reflects frontal-plane gait control, dynamic balance, and lateral weight transfer during walking. Individuals with knee pain commonly adopt compensatory strategies to reduce joint loading, including altered weight-shifting patterns, increased gait asymmetry, reduced knee excursion, and modifications in step width and trunk motion. These adaptations alter the rhythmic characteristics of ML acceleration, making ML frequency features particularly sensitive to pain-related gait impairments. The high ranking of these features therefore suggests that alterations in frontal-plane locomotor control provide important discriminatory information beyond conventional temporal gait measures such as cadence and stride time. Furthermore, the consistency of these findings across both Gini and permutation importance analyses indicates that the observed importance of ML frequency features reflects robust biomechanical characteristics rather than methodological bias in the feature ranking procedure.

Although several procedures were implemented to improve robustness and reduce overfitting, the study did not explore synthetic augmentation, generative balancing, or domain adaptation strategies. Future work should investigate advanced augmentation and balancing frameworks, including generative adversarial networks (GANs), time-series augmentation, and domain generalization approaches, to improve robustness across heterogeneous wearable and clinical populations. Additional multimodal studies integrating wearable sensing with imaging, electromyography, kinetic analysis, or clinical biomarkers may further clarify the physiological pathways underlying pain-related gait adaptations.

Overall, the present findings demonstrate that extended free-living gait monitoring combined with ML analysis can identify reproducible gait signatures associated with self-reported OA pain states. These results support the feasibility of wearable-derived gait-quality metrics as candidate digital biomarkers for scalable, non-invasive monitoring of pain-related functional impairment in real-world healthcare settings.

Finally, although the proposed framework achieved strong predictive performance, the study relied on a single modality gait features (accelerometry-derived) and a cross-sectional labeling approach. Future research should incorporate multimodal data sources, including imaging, clinical assessments, and longitudinal pain trajectories, to better disentangle the complex relationship between gait dynamics and both subjective and physiological aspects of pain. Such integration would further strengthen the interpretability and clinical utility of wearable-based digital biomarkers.

6 Conclusion

This study demonstrates the feasibility of using real-world gait quality features, derived from seven-day continuous wearable accelerometer recordings, and a structured ML analyses framework to predict OA-related pain. More specifically, the proposed framework predicts self-reported OA pain states derived from the WOMAC pain subscale, rather than directly measuring physiological pain. Our approach integrates established pre-processing, feature-selection, and machine-learning techniques within a rigorously validated analytical pipeline for wearable gait analysis, achieved high predictive performance, with the XGBoost classifier attaining an AUC of 0.91, accuracy of 0.93, precision of 0.94, specificity of 0.99, sensitivity of 0.91, and F1-score of 0.92. Gait-quality metrics, particularly variability-related features such as stride regularity and dominant frequency characteristics, demonstrated substantially greater predictive utility than general activity-based measures. These findings indicate that subtle gait dynamics captured during prolonged real-world monitoring contain reproducible behavioral signatures associated with perceived OA pain states and may serve as scalable, non-invasive digital biomarkers for continuous monitoring of pain-related functional impairment in free-living environments. These findings suggest that gait-derived features function as digital biomarkers associated with perceived pain states, capturing functional and behavioral adaptations to pain in real-world settings. Importantly, these biomarkers should be interpreted as reflecting correlates of perceived pain rather than direct indicators of underlying nociceptive mechanisms. These results highlight the potential of gait signatures as objective digital biomarkers, offering a

Model	Acc.	Prec.	Sen.	F1.	AUC.	Spec.
Gait Rhythm						
XGBoost	0.82	0.86	0.79	0.79	0.79	0.97
LightGBM	0.75	0.83	0.71	0.71	0.89	0.99
RandomFrst	0.70	0.83	0.65	0.63	0.65	1.00
GAM2	0.58	0.61	0.51	0.39	0.51	0.99
SVM	0.57	0.79	0.50	0.37	0.50	1.00
ENCAIpha.5	0.59	0.53	0.50	0.37	0.50	0.99
Naive Bayes	0.56	0.53	0.52	0.48	0.52	0.83
AdaBoost	0.61	0.61	0.55	0.52	0.55	0.90
Gait Symmetry						
XGBoost	0.86	0.89	0.84	0.85	0.84	0.98
LightGBM	0.80	0.87	0.76	0.77	0.95	0.99
RandomFrst	0.75	0.85	0.70	0.70	0.70	1.00
GAM2	0.59	0.58	0.52	0.44	0.52	0.96
SVM	0.59	0.74	0.51	0.38	0.51	1.00
ENCAIpha.5	0.59	0.74	0.52	0.44	0.52	0.95
Naive Bayes	0.51	0.54	0.54	0.52	0.54	0.44
AdaBoost	0.61	0.61	0.57	0.55	0.57	0.87
Gait Variability						
XGBoost	0.88	0.91	0.86	0.87	0.86	0.99
LightGBM	0.82	0.86	0.78	0.79	0.92	0.98
RandomFrst	0.82	0.88	0.79	0.80	0.79	1.00
GAM2	0.60	0.59	0.53	0.47	0.53	0.93
SVM	0.63	0.65	0.57	0.54	0.57	0.93
ENCAIpha.5	0.63	0.60	0.56	0.52	0.56	0.89
Naive Bayes	0.59	0.56	0.55	0.54	0.55	0.76
AdaBoost	0.64	0.62	0.59	0.58	0.59	0.82
Gait Amplitude						
XGBoost	0.87	0.91	0.84	0.86	0.84	1.00
LightGBM	0.80	0.86	0.76	0.77	0.93	0.99
RandomFrst	0.77	0.86	0.72	0.72	0.72	1.00
GAM2	0.59	0.58	0.51	0.40	0.51	0.98
SVM	0.60	0.74	0.52	0.42	0.52	1.00
ENCAIpha.5	0.60	0.56	0.51	0.42	0.51	0.96
Naive Bayes	0.57	0.52	0.51	0.45	0.51	0.89
AdaBoost	0.62	0.64	0.59	0.58	0.59	0.88

Table 10 Gait quality rhythm, symmetry, variability, and amplitude of the acceleration during walking features.

non-invasive, data-driven tool for early detection and personalized management of OA. In this context, the framework enables objective estimation of self-reported pain from continuous, real-world gait patterns. **Although the proposed framework demonstrated strong performance within the MOST cohort, its generalizability to independent populations, clinical settings, and alternative wearable devices remains to be established through external validation studies.** Future work will involve validating the framework on other populations and analyzing Opal accelerometer data from the MOST study, comparing predictive performance across devices, and deploying the models in real-world settings to assess generalizability and clinical utility. Additionally, integrating multimodal clinical and physiological measures will be critical to further disentangle the relationship between gait dynamics, perceived pain, and underlying biological mechanisms.

Data Availability: The dataset analysed during this study is available in the National Institute on Aging AgingResearch Biobank (MOST) repository upon reasonable request and approval. Specifically, the MOST dataset can be accessed at: <https://agingresearchbiobank.nia.nih.gov/studies/most/details>

Funding: This research was supported by the National Institutes of Health/NIH/DHHS under grant No. UG3NS135168.

Appendix A Detailed Definitions of Gait and Activity Features

A.0.1 Gait Quality Features

The gait features were calculated based on walking bouts, where a walking bout is defined as a series of uninterrupted walking periods with a minimum duration. For each day, summary statistics were calculated for each feature, which includes the mean, standard deviation, minimum and maximum and median, based on all walking bouts of a minimum duration. Additionally, each walking bout was divided into 30-second and 60-second segments, and feature calculations were carried out separately for each of these segment lengths. The gait features are only computed for days 1-7 if there are 30s or 60s of continuous walking (depending on the analysis). The Gait quality features were further classified as (Ia) Rhythm (Ib) Symmetry (Ic) Variability (Id) Amplitude of the acceleration during walking.

- Rhythm

1. Step time [sec]: Time between the initial contact of one foot and the initial touch of the opposing foot during a gait cycle.
2. Stride time [sec]: Interval between initial touch of one foot and the second initial contact of the same foot.
3. Dominant frequency of the power spectrum [Hz] : Within the 0.5–3 Hz range, the frequency component with the maximum spectral amplitude was identified. Following Fourier transformation of the time-dependent signal, the principal frequency corresponds to the mean step interval, representing approximately 50% of the gait cycle. Accordingly, a reduction in this dominant frequency reflects an increase in the average stride duration.
4. Cadence [steps/min]: is determined in the frequency domain. The signal's dominant frequency (Fs) was determined and its inverse value (Ts), which represents the average step duration, was calculated.

$$\text{Cadence (time domain)} = \frac{60[\text{sec}]}{\text{total bout length.}} \quad (\text{A1})$$

$$\text{Cadence (frequency domain)} = (\text{frequency} * 60[\text{sec}].) \quad (\text{A2})$$

- Symmetry

1. Step regularity: represent the step-to-step consistency of acceleration. Low step regularity indicates asymmetry between the left and right legs[39].
2. Step symmetry:

$$\text{Step symmetry} = \frac{\text{step regularity}}{\text{stride regularity}}. \quad (\text{A3})$$

$$\text{Step asym} = |1 - \text{step symmetry}| \quad (\text{A4})$$

3. HR- Harmonic Ratio: can be applied to evaluate changes in gait patterns. It refers to imbalance between steps within a stride. It's the acceleration signals for each stride transformed into the frequency domain using Fourier analysis centered on the stride frequency..

- Variability

1. Stride regularity: describes the consistency of the acceleration between consecutive steps[39, 40].
2. Width of the principal frequency component in the power spectrum [Hz]: The frequency dispersion expresses the signal's variability. Stride time variability has been found to be more strongly related to breadth than to the dominant frequency [41].
3. Amplitude of the dominant frequency [g^2/Hz]: Shows the strength of the normalized signal's main frequency. The normalized signal is obtained by subtracting the average and dividing by the standard deviation. The amplitude is then computed by applying a Fourier transform to the normalized signal.
4. Slope of the dominant frequency [g^2/Hz^2]: combines amplitude and width, reflecting both frequency and periodicity dispersion. This measure can distinguish between healthy and PD in everyday life and in laboratory studies[42].

- Amplitude of the acceleration during walking

1. Acceleration range (Rng) [m/sec^2]

$$\text{Rng} = \text{maximum}(\text{acc}) - \text{minimum}(\text{acc}) \quad (\text{A5})$$

2. Root Mean Square (RMS) Acceleration [m/sec^2]: measures the magnitude of acceleration[43].

$$\text{RMS} = \text{sqrt}(\text{mean}(\text{acc}.2)) \quad (\text{A6})$$

Research indicates that the RMS of an acceleration signal can be impacted by test factors like eye closure, standing on one foot, as well as participants' age, and fall history[44].

3. Activity Level (Mean RMS acceleration): is a gait parameter quantifying the energy expenditure during walking, expressed as:

$$\text{Activity level} = \text{mean}(\text{sqrt}(\text{mean}(\text{acc}.2))) \quad (\text{A7})$$

A.0.2 Activity Features

1. Walking-time [minimum] and Percentage Walking(%): Walking between 00:00 and 24:00 on each calendar date is defined based on continuous walking bouts of three(3) steps or more.

$$\text{Percentage walking} = 100 \times \left(\frac{\text{walking time (00:00–24:00)}}{\text{wear time (00:00–24:00 minutes)}} \right) \quad (\text{A8})$$

2. Lying-time [minimum] and Percentage lying(%): between 00:00 and 24:00 on each calendar date (throughout wear time only).

$$\text{Percentage lying} = 100 \times \left(\frac{\text{lying-time (00:00–24:00)}}{\text{wear time (00:00–24:00 minutes)}} \right) \quad (\text{A9})$$

3. Sitting/Standing-Time [Minimum] and Percentage Sitting/Standing Quietly: between 00:00-24:00 in each calendar date (only throughout wear time).

$$\text{Percentage sit/stand} = 100 \times \left(\frac{\text{sit/stand time (00:00–24:00)}}{\text{wear time (00:00–24:00 minutes)}} \right) \quad (\text{A10})$$

4. Sedentary-time [minimum] and Percentage Sedentary time(%): is the addition of sitting/standing and lying between 00:00 and 24:00 on each calendar date (throughout wear time only).

$$\text{Percentage sedentary} = 100 \times \left(\frac{\text{sedentary time (00:00–24:00)}}{\text{wear time (00:00–24:00)}} \right) \quad (\text{A11})$$

5. Sum Signal Vector Magnitude Day[mg]: The sum of observed acceleration signals from three axes is normalized by wear time between 00:00 and 24:00.

$$\text{Sum Signal Vector Magnitude} = \sqrt{V^2 + M^2 + AP^2}. \quad (\text{A12})$$

6. Mean Signal Vector Magnitude (Day [mg]: is the average SVM of the acceleration signal observed from three axes over 15-second epochs, normalized by wear time between 00:00 and 24:00.
7. Percentage Active-Day: is the activity threshold, which was determined as the average of and lying and walking epochs. The percentage of activity epochs for each calendar date between 00:00 and 24:00 was computed.
8. Step Count: The addition of steps for each calendar date (throughout wear time only) is derived from continuous walking of three(3) or more steps between 00:00 and 24:00.

A.0.3 Additional Preprocessing Equations

- **Standardization of features.** To place all features on a comparable scale and prevent those with large magnitudes from dominating the model, each feature was

standardized to have zero mean and unit variance:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where $\mu_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ is the mean and $\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2}$ is the standard deviation of feature x_j . Standardization ensures that gradient-based algorithms converge faster and that distance-based models treat all features equally.

- **Redundancy reduction via Pearson correlation filtering.** Highly correlated features can introduce multicollinearity, which reduces model interpretability and may degrade generalization. Pairwise Pearson correlation coefficients were computed between all pairs of features:

$$r_{jk} = \frac{\sum_{i=1}^n (x_{ij} - \mu_j)(x_{ik} - \mu_k)}{(n-1)\sigma_j\sigma_k}.$$

If the absolute correlation exceeded a threshold of $|r_{jk}| > 0.7$, one of the features in the pair was removed. This step ensures that redundant predictors do not provide overlapping information, improving computational efficiency and model stability.

References

- [1] Yu, H., Huang, T., Lu, W.W., Tong, L., Chen, D.: Osteoarthritis pain. *International journal of molecular sciences* **23**(9), 4642 (2022)
- [2] Leifer, V.P., Katz, J.N., Losina, E.: The burden of oa-health services and economics. *Osteoarthritis and cartilage* **30**(1), 10–16 (2022)
- [3] Safiri, S., Kolahi, A.-A., Smith, E., Hill, C., Bettampadi, D., Mansournia, M.A., Hoy, D., Ashrafi-Asgarabad, A., Sepidarkish, M., Almasi-Hashiani, A., *et al.*: Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the global burden of disease study 2017. *Annals of the rheumatic diseases* **79**(6), 819–828 (2020)
- [4] Scheuing, W.J., Reginato, A.M., Deeb, M., Kasman, S.A.: The burden of osteoarthritis: Is it a rising problem? *Best Practice & Research Clinical Rheumatology* **37**(2), 101836 (2023)
- [5] Brandt, K.D., Dieppe, P., Radin, E.L.: Etiopathogenesis of osteoarthritis. *Rheumatic Disease Clinics of North America* **34**(3), 531–559 (2008)
- [6] Tang, S., Zhang, C., Oo, W.M., Fu, K., Risberg, M.A., Bierma-Zeinstra, S.M., Neogi, T., Atukorala, I., Malfait, A.-M., Ding, C., *et al.*: Osteoarthritis. *Nature Reviews Disease Primers* **11**(1), 1–22 (2025)

- [7] Verma, D.K., Kumari, P., Kanagaraj, S.: Engineering aspects of incidence, prevalence, and management of osteoarthritis: a review. *Annals of Biomedical Engineering* **50**(3), 237–252 (2022)
- [8] Al-Worafi, Y.M.: Osteoarthritis management in developing countries. In: *Handbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research*, pp. 1–26. Springer, ??? (2024)
- [9] Leroux, A., Rzasa-Lynn, R., Crainiceanu, C., Sharma, T.: Wearable devices: current status and opportunities in pain assessment and management. *Digital Biomarkers* **5**(1), 89–102 (2021)
- [10] Hemanth, K., Tigulla, D., Lakshmi, V., Buhari, S.: Early stage detection of osteoarthritis of the joints (hip and knee) using machine learning. In: *Diagnosing Musculoskeletal Conditions Using Artificial Intelligence and Machine Learning to Aid Interpretation of Clinical Imaging*, pp. 39–64. Elsevier, ??? (2025)
- [11] Pirker, W., Katzenschlager, R.: Gait disorders in adults and the elderly: A clinical guide. *Wiener Klinische Wochenschrift* **129**(3), 81–95 (2017)
- [12] Hodges, P.W., Tucker, K.: Moving differently in pain: a new theory to explain the adaptation to pain. *Pain* **152**(3), 90–98 (2011)
- [13] Sethi, V., Anand, C., Della Pasqua, O.: Clinical assessment of osteoarthritis pain: contemporary scenario, challenges, and future perspectives. *Pain and Therapy* **13**(3), 391–408 (2024)
- [14] Li, H., Hu, S., Zhao, R., Zhang, Y., Huang, L., Shi, J., Li, P., Wei, X.: Gait analysis of bilateral knee osteoarthritis and its correlation with western ontario and mcmaster university osteoarthritis index assessment. *Medicina* **58**(10), 1419 (2022)
- [15] Kim, J.-S., Kim, C.-J., Schlenk, E.A.: Psychometric properties of the korean version of the patient knowledge questionnaire-osteoarthritis. *Clinical Nursing Research* **31**(1), 69–79 (2022)
- [16] Kárasón, H., Ritrovato, P., Maffulli, N., Boccaccini, A.R., Tortorella, F.: Wearable approaches for non-invasive monitoring of tendons: A scoping review. *Internet of Things*, 101199 (2024)
- [17] Jarchi, D., Pope, J., Lee, T.K., Tamjidi, L., Mirzaei, A., Sanei, S.: A review on accelerometry-based gait analysis and emerging clinical applications. *IEEE reviews in biomedical engineering* **11**, 177–194 (2018)
- [18] Clermont, C.A., Barden, J.M.: Accelerometer-based determination of gait variability in older adults with knee osteoarthritis. *Gait & posture* **50**, 126–130 (2016)

- [19] Narayanan, A., Desai, F., Stewart, T., Duncan, S., Mackay, L.: Application of raw accelerometer data and machine-learning techniques to characterize human movement behavior: a systematic scoping review. *Journal of Physical Activity and Health* **17**(3), 360–383 (2020)
- [20] Matsangidou, M., Liampas, A., Pittara, M., Pattichi, C.S., Zis, P.: Machine learning in pain medicine: an up-to-date systematic review. *Pain and therapy* **10**(2), 1067–1084 (2021)
- [21] Villalba-Meneses, F., Guevara, C., Lojan, A.B., Gualsaqui, M.G., Arias-Serrano, I., Velásquez-López, P.A., Almeida-Galárraga, D., Tirado-Espín, A., Marín, J., Marín, J.J.: Classification of the pathological range of motion in low back pain using wearable sensors and machine learning. *Sensors* **24**(3), 831 (2024)
- [22] Trabassi, D., Castiglia, S.F., Bini, F., Marinozzi, F., Ajoudani, A., Lorenzini, M., Chini, G., Varrecchia, T., Ranavolo, A., De Icco, R., *et al.*: Optimizing rare disease gait classification through data balancing and generative ai: insights from hereditary cerebellar ataxia. *Sensors* **24**(11), 3613 (2024)
- [23] Berthouze, N., Valstar, M., Williams, A., Egede, J., Olugbade, T., Wang, C., Meng, H., Aung, M., Lane, N., Song, S.: Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. *arXiv*, 2001 (2020)
- [24] Almuhammadi, W.S., Agu, E., King, J., Franklin, P.: Oa-pain-sense: machine learning prediction of hip and knee osteoarthritis pain from imu data. In: *Informatics*, vol. 9, p. 97 (2022). MDPI
- [25] Bacon, K.L., Felson, D.T., Jafarzadeh, S.R., Kolachalama, V.B., Hausdorff, J.M., Gazit, E., Segal, N.A., Lewis, C.E., Nevitt, M.C., Kumar, D., *et al.*: Relation of gait measures with mild unilateral knee pain during walking using machine learning. *Scientific Reports* **12**(1), 22200 (2022)
- [26] Costello, K.E., Felson, D.T., Jafarzadeh, S.R., Guermazi, A., Roemer, F.W., Segal, N.A., Lewis, C.E., Nevitt, M.C., Lewis, C.L., Kolachalama, V.B., *et al.*: Gait, physical activity and tibiofemoral cartilage damage: a longitudinal machine learning analysis in the multicenter osteoarthritis study. *British journal of sports medicine* **57**(16), 1018–1024 (2023)
- [27] Castagno, S., Gompels, B., Strangmark, E., Robertson-Waters, E., Birch, M., Schaar, M., McCaskie, A.W.: Understanding the role of machine learning in predicting progression of osteoarthritis. *Bone Joint J* **106**, 1216–1222 (2024)
- [28] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016)

- [29] National Institute on Aging: Multicenter Osteoarthritis Study (MOST). AgingResearch Biobank. Accessed: March 13, 2026. <https://agingresearchbiobank.nia.nih.gov/studies/most/details>
- [30] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017)
- [31] Louppe, G.: Understanding random forests: From theory to practice. PhD thesis, Universite de Liege (Belgium) (2014)
- [32] Cortes, C., Vapnik, V.: Support-vector networks. *Machine learning* **20**(3), 273–297 (1995)
- [33] Schapire, R.E.: Explaining adaboost. In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37–52. Springer, ??? (2013)
- [34] Bangroo, R., Verma, S.R., *et al.*: Comparative study of elastic net regression, naive bayes & lasso regression. In: *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, pp. 1–6 (2023). IEEE
- [35] Hastie, T., Tibshirani, R.: Generalized additive models. *Statistical science* **1**(3), 297–310 (1986)
- [36] Lewek, M.D., Scholz, J., Rudolph, K.S., Snyder-Mackler, L.: Stride-to-stride variability of knee motion in patients with knee osteoarthritis. *Gait & posture* **23**(4), 505–511 (2006)
- [37] Kaufman, K.R., Hughes, C., Morrey, B.F., Morrey, M., An, K.-N.: Gait characteristics of patients with knee osteoarthritis. *Journal of biomechanics* **34**(7), 907–915 (2001)
- [38] Cigdem, O., Deniz, C.M.: Artificial intelligence in knee osteoarthritis: a comprehensive review for 2022. *Osteoarthritis imaging* **3**(3), 100161 (2023)
- [39] Helbostad, J.L., Moe-Nilssen, R.: The effect of gait speed on lateral balance control during walking in healthy elderly. *Gait & posture* **18**(2), 27–36 (2003)
- [40] Tura, A., Raggi, M., Rocchi, L., Cutti, A.G., Chiari, L.: Gait symmetry and regularity in transfemoral amputees assessed by trunk accelerations. *Journal of neuroengineering and rehabilitation* **7**, 1–10 (2010)
- [41] Hausdorff, J.M.: Gait dynamics, fractals and falls: finding meaning in the stride-to-stride fluctuations of human walking. *Human movement science* **26**(4), 555–589 (2007)
- [42] Weiss, A., Sharifi, S., Plotnik, M., Vugt, J.P., Giladi, N., Hausdorff, J.M.: Toward

automated, at-home assessment of mobility among patients with parkinson disease, using a body-worn accelerometer. *Neurorehabilitation and neural repair* **25**(9), 810–818 (2011)

- [43] Mancini, M., Horak, F.B., Zampieri, C., Carlson-Kuhta, P., Nutt, J.G., Chiari, L.: Trunk accelerometry reveals postural instability in untreated parkinson’s disease. *Parkinsonism & related disorders* **17**(7), 557–562 (2011)
- [44] Mancini, M., Salarian, A., Carlson-Kuhta, P., Zampieri, C., King, L., Chiari, L., Horak, F.B.: Isway: a sensitive, valid and reliable measure of postural control. *Journal of neuroengineering and rehabilitation* **9**, 1–8 (2012)