JMIR Publications
The Leading eHealth Publisher

Smartphone-Based Interpretable Machine Learning Framework for Classifying Knee Function from Single-Leg Squat Videos: Cross-Sectional Study

Si-hyun Kim, Kyue-nam Park

Submitted to: JMIR mHealth and uHealth
on: October 01, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Preprint
JMIR Publications

Smartphone-Based Interpretable Machine Learning Framework for Classifying Knee Function from Single-Leg Squat Videos: Cross-Sectional Study

Si-hyun Kim[1]; Kyue-nam Park[2]

[1] Department of Physical Therapy, Sangji University Wonju KR
[2] Department of Physical Education, Digital healthcare and Sports data science Lab., Yonsei University Seoul KR

Corresponding Author:
Kyue-nam Park

Department of Physical Education, Digital healthcare and Sports data science Lab., Yonsei University
50 Yonsei-ro, Seodaemun-gu
Seoul
KR

Abstract

Background: Accurate and interpretable classification of knee function is essential for evaluating movement quality in clinical and sports contexts but remains challenging without expert assessment or laboratory-based motion capture.

Objective: This study developed and validated an interpretable ensemble learning framework that classifies knee function into three levels (Good, Moderate, Poor) from single-smartphone, front-view single-leg squat (SLS) videos.

Methods: A dataset of SLS videos was labeled by physiotherapists into three categories, and kinematic features were extracted from 2D pose estimation. Engineered features, including ratios, nonlinear terms, and interaction variables, were selected using a hybrid statistical and recursive elimination approach. Six classifiers were trained with stratified 5-fold cross-validation and a held-out test set, and a soft voting ensemble was developed to enhance robustness. SHAP and LIME were applied for global and local interpretability.

Results: The soft voting ensemble combining Logistic Regression and LightGBM achieved high performance on the held-out test set (accuracy = 87.5%, F1-score = 87.6%, AUC = 0.938). SHAP analysis identified the Trunk × Knee interaction as the most influential feature for distinguishing poor classifications, while LIME provided instance-level explanations that clarified decision boundaries.

Conclusions: Global and local interpretability enhanced model transparency by illustrating why impaired movement patterns were detected, thereby supporting personalized feedback on knee function. This interpretable framework provides a practical and accessible solution for functional knee assessment, with direct applicability to rehabilitation, athletic screening, and remote monitoring. Clinical Trial: The study protocol was reviewed and approved by the Institutional Review Board of Jeonju University

(JMIR Preprints 01/10/2025:85126)
DOI: https://doi.org/10.2196/preprints.85126

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Preprint
JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Title

Smartphone-Based Interpretable Machine Learning Framework for Classifying Knee Function from Single-Leg Squat Videos: Cross-Sectional Study

Abstract

Background: Accurate and interpretable classification of knee function is essential for evaluating movement quality in clinical and sports contexts but remains challenging without expert assessment or laboratory-based motion capture.

Objective: This study developed and validated an interpretable ensemble learning framework that classifies knee function into three levels (Good, Moderate, Poor) from single-smartphone, front-view single-leg squat (SLS) videos.

Method: A dataset of SLS videos was labeled by physiotherapists into three categories, and kinematic features were extracted from 2D pose estimation. Engineered features, including ratios, nonlinear terms, and interaction variables, were selected using a hybrid statistical and recursive elimination approach. Six classifiers were trained with stratified 5-fold cross-validation and a held-out test set, and a soft voting ensemble was developed to enhance robustness. SHAP and LIME were applied for global and local interpretability.

Results: The soft voting ensemble combining Logistic Regression and LightGBM achieved high performance on the held-out test set (accuracy = 87.5%, F1-score = 87.6%, AUC = 0.938). SHAP analysis identified the Trunk × Knee interaction as the most influential feature for distinguishing poor classifications, while LIME provided instance-level explanations that clarified decision boundaries.

Conclusions: Global and local interpretability enhanced model transparency by illustrating why impaired movement patterns were detected, thereby supporting personalized feedback on knee function. This interpretable framework provides a practical and accessible solution for functional knee assessment, with direct applicability to rehabilitation, athletic screening, and remote monitoring.

Keywords: Explainable artificial intelligence (XAI), Knee function classification, Single-leg squat test, Smartphone-based assessment, Supervised machine learning, Voting classifier

JMIR Publications

Key points

We proposed an interpretable ensemble framework that accurately classified knee function into three levels from single-smartphone single-leg squat videos, achieving high performance while remaining clinically transparent.

By incorporating a trunk × knee interaction feature alongside knee angle and the knee-to-trunk ratio, the model introduced a novel perspective on movement patterns not captured by conventional evaluation criteria.

SHAP and LIME explanations enhanced clinical trust by clarifying model decisions, enabling practical applications for rehabilitation, athletic screening, and remote monitoring.

Preprint JMIR Publications

Introduction

Knee function is a key factor in mobility, athletic performance, and quality of life for active young adults, with nearly one-third experiencing knee-related problems [1-3]. For young populations, functional performance tests provide meaningful evaluations of knee function by indirectly estimating muscle strength, joint stability, flexibility, balance, and proprioception, thereby offering insights unattainable through static radiographic imaging [4]. Among these, the single-leg squat (SLS) is particularly effective for assessing dynamic knee function in both sports and clinical contexts, with movement quality classified on either a 2-point scale (Good vs. Poor) or a 3-point scale (Good vs. Moderate vs. Poor) [5]. SLS performance is typically evaluated via expert visual observation of segmental alignment or using laboratory-based three-dimensional (3D) motion capture to assess hip internal rotation, knee valgus, trunk lateral lean, and pelvic drop [6, 7]. 3D motion capture can measure movement with high accuracy, it requires costly equipment and longer processing time compared to visual assessment. [8]. Visual assessment, although less expensive, still necessitates expert analysis of recorded footage, making it time-consuming and susceptible to observer bias and limited reproducibility when classifying movement quality [7].

Recently, machine learning (ML) algorithms have been applied to evaluate and classify SLS performance without costly 3D motion capture systems or expert visual assessments, thereby reducing analysis time [8-10]. An IMU-based support vector machine (SVM) using three sensors achieved 96% accuracy for binary classification (Good vs. Poor) but dropped to 66% for three-class classification (Good vs. Moderate vs. Poor) based on hip and knee angles [9]. However, IMU accuracy depends on the number of sensors, and the requirement for multiple sensors limits their practicality in clinical settings [9]. More recent approaches using consumer-grade cameras with pose estimation algorithms have achieved moderate-to-high accuracy in three-class SLS classification, even with a single camera [8, 11]. For example, pose-estimated SLS videos captured with a single camera achieved 53%–77% agreement with expert assessments across six postural orientation errors, with knee valgus errors showing the highest agreement (77%) [8]. Similarly, a smartphone-based markerless motion capture study found that trunk, pelvic, and knee angles, together with the summated angle of these joints, enabled three-class classification with 76.9% accuracy using a decision tree model [11]. However, these studies primarily extracted features based on expert-defined SLS guidelines, such as trunk lateral leaning and knee valgus, whereas SLS performance involves overall movement patterns of the trunk and lower limbs. It is necessary for future ML classification models to generate features that capture inter-joint movement patterns, such as the knee-to-trunk ratio and amplified joint interactions during the SLS test.

While recent ML approaches have demonstrated promising accuracy in SLS classification, decision-making processes often remain opaque, limiting the adoption of such models in clinical practice [12]. In rehabilitation and sports medicine, clinicians and other professionals need to understand which specific kinematic factors

Preprint JMIR Publications

drive these predictive outcomes in order to design targeted interventions and ensure professional confidence [13]. Explainable artificial intelligence (XAI) methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), can provide explanatory, feature-level insights by quantifying the contribution of each joint movement or interaction to the model's decision [14]. Applying XAI to knee function assessment can classify performance levels and reveal key movement patterns, bridging algorithmic predictions with clinical feedback to foster trust and personalized rehabilitation planning.

The primary purpose of this study was to develop and validate an interpretable ensemble learning framework for the automatic classification of knee function into three categories (Good, Moderate, Poor) using SLS videos captured with a single smartphone camera. The secondary purpose was to apply novel biomechanical feature engineering and hybrid feature selection to identify features beyond conventional independent joint angles, aiming to reveal patterns undetected by human experts and enhance the explanatory power of subsequent XAI analyses. We hypothesized that an ensemble learning model with selected novel features would improve classification accuracy and robustness, and that applying SHAP and LIME would provide clinically relevant, class-specific, and instance-level explanations to facilitate targeted feedback in rehabilitation and fitness training.

Methods

1. Participants and Data Collection

1.1. Participants

This study included a total of 105 young active adults (aged 19–30 years) who were recruited in xxxxx city through university campus advertisements, social media, and local community bulletin boards. Participants were eligible for inclusion if they engaged in regular physical activity for at least 60 minutes per week. Exclusion criteria included a history of lower-extremity or low back pain within the previous three months, prior major knee injuries or surgeries, neurological disorders, or an inability to perform the SLS test [15]. All participants provided written informed consent prior to participation. The study protocol was reviewed and approved by the Institutional Review Board of xxxxx University.

1.2. Data Collection: Smartphone Video Recording of the Single-Leg Squat Test

Videos were recorded using an iPhone mounted on a tripod positioned 3 meters in front of the participant at pelvic height, capturing the entire body at 30 Hz in 1080p resolution. Only the frontal view was recorded, as clinical evaluation of SLS performance primarily relies on frontal plane deviations including trunk lean, pelvic tilt, and medial knee displacement [8]. Additionally, this single-camera configuration reflects a practical and scalable approach for real-world clinical and remote assessment settings, eliminating the need for multiple camera setups to capture side or top views [16].

Each participant performed the SLS on their dominant leg (defined as the leg used to kick a ball), wearing

standard gym clothing (e.g., shorts and T-shirts) to ensure full visibility of lower-limb motion. Participants were given a brief warm-up consisting of 5 minutes of stationary cycling and completed three practice trials under the supervision of a physiotherapist to ensure consistent technique. The SLS movement was performed at a comfortable, self-selected pace, avoiding excessively slow speeds that might dampen compensatory kinematic deviations such as knee valgus [17]. The standardized SLS test procedure was as follows: 1) Test began with participants standing with their feet shoulder-width apart and toes pointing forward; 2) They then placed their hands on their waist and extended the nondominant leg backward with approximately 90° of knee flexion; 3) From this position, participants descended into a squat as deeply as possible at a self-selected speed, bending the weight-bearing leg without losing balance, maintaining an upright posture, and facing the front camera; and 4) Finally, they fully extended the knee of the weight-bearing leg to return to the starting position [18].

Each participant performed three repetitions, with a one-minute rest between attempts. A repetition was considered invalid if the participant lost balance, allowed the legs to touch, placed the nondominant foot on the ground, or failed to maintain an upright posture facing the camera. In such cases, an additional repetition was performed at the end of the trial. If two or more repetitions within a trial were invalid, the entire trial was repeated after a sufficient rest period [17].

2. Labeling of Single-Leg Squat Movement Quality

Movement quality was evaluated independently by two licensed physiotherapists, each with over 10 years of clinical experience in musculoskeletal rehabilitation. Prior to labeling, the evaluator completed a structured two-hour training session on SLS assessment criteria, which included grading five representative SLS videos excluded from the main dataset [19]. A consensus discussion with one of the authors was held to standardize interpretation and rating procedures, ensuring consistent and objective classification.

Following training, the two evaluators independently assessed movement quality through visual observation while blinded to participant characteristics and uninvolved in other experimental procedures. Each participant's performance was reviewed using full-length SLS videos played at normal speed on a personal computer, without pausing or rewinding [19]. A multi-segmental approach was applied, simultaneously evaluating the trunk, pelvis, and knee throughout the three consecutive squats, and considering their movements in relation to one another [19].

SLS performance was graded into one of three categories based on observable frontal-plane deviations [20]:

[1] Class 0 (Good) – No lateral tilt of the trunk or pelvis, absence of visible knee valgus, and a stable, straight knee trajectory throughout the squat.

[2] Class 1 (Moderate) – Mild lateral tilt of the trunk or pelvis, slight knee valgus, and minor medial–lateral knee movement without notable instability.

[3] Class 2 (Poor) – Pronounced lateral trunk or pelvic tilt, obvious knee valgus, and clear medial–lateral displacement or shaking of the knee.

The final class label for each participant was determined by selecting the poorest observed performance among the three repetitions, and the two evaluators then reached agreement on this label; in cases of disagreement, consensus was achieved through discussion [20].

3. Pose Estimation and Feature Extraction

For consistency between labeling and analysis, the same poorest-rated trial used for labeling was also selected for 2D pose estimation and feature extraction. Pose estimation of the SLS smartphone video was performed using the MediaPipe Pose library (v0.10.21), which detects 33 body landmarks from RGB frames in real time. MediaPipe Pose operates without the need for depth sensors, GPU acceleration, or manual landmark initialization, making it well-suited for smartphone-based clinical assessments. Previous research has demonstrated that MediaPipe Pose provides highly valid estimates of lower-limb joint angles compared with gold-standard 3D motion capture (Pearson correlation = 0.87) and exhibits excellent test–retest reliability (ICC = 0.93–0.95), supporting its suitability for clinical biomechanics applications [21]. In accordance with established SLS test guidelines [20], the peak values of three primary joint angles were calculated to quantify frontal-plane deviations during the SLS test:

- Knee valgus: Computed as 180° minus the internal angle formed by the hip–knee–ankle points, representing frontal-plane knee alignment.
- Trunk lateral tilt: Defined as the deviation angle between the vertical axis and the line connecting the shoulder and hip midpoints, indicating upper-body lean.
- Pelvic tilt: Measured as the angle between the pelvic alignment and the vertical axis, reflecting pelvic obliquity.

While the primary joint angles align with conventional SLS evaluation criteria [20], they may overlook complex inter-segmental relationships and compensatory movement strategies. To address this limitation and identify potentially more informative indicators of knee function, the primary measures were further transformed into engineered features, resulting in a total of 14 interpretable kinematic variables. The engineered feature set comprised 11 interpretable kinematic variables:

- Ratio features: Knee-to-trunk ratio (Knee / Trunk), knee-to-pelvic ratio (Knee / Pelvic), trunk-to-pelvic ratio (Trunk / Pelvic), and knee-to-total-trunk ratio (Knee / (Trunk + Pelvic)), which express proportional relationships within the trunk–pelvis–knee complex and enable the detection of altered motor control strategies.
- Interaction features: Amplified joint interaction terms (Trunk × Knee, Trunk × Pelvic, and Pelvic × Knee), which capture synergistic or antagonistic segmental motion patterns that may indicate compensatory movement behaviors.
- Nonlinear features: Squared terms of trunk, pelvic, and knee angles ($Trunk^2$, $Pelvic^2$, $Knee^2$), which model curvilinear biomechanical effects.
- Composite feature: The summated angle (sum of trunk, pelvic, and knee), which represents the

overall deviation from neutral alignment during SLS.

14 features, including the three primary joint angles (trunk, pelvic, and knee), were incorporated to provide the model with both fundamental kinematic descriptors and higher-order derived parameters. The inclusion of ratio, interaction, nonlinear and composite features allows the model to capture subtle coordination deficits and compensatory strategies that may not be evident from raw joint angles alone, thereby enhancing its discriminative capacity and clinical interpretability.

4. Machine Learning Pipeline

4.1. Dataset Splitting and Balancing

Fig. 1 provides an overview of the workflow. The dataset was divided into training and test sets using stratified sampling, with 70% (n = 73) allocated for training and 30% (n = 32) for independent testing, ensuring proportional representation of the three movement quality classes (Good, Moderate, Poor). To address class imbalance in the training set (19 Good, 29 Moderate, 25 Poor), the Synthetic Minority Oversampling Technique combined with Tomek links (SMOTE-Tomek) was applied. This hybrid approach not only oversampled the minority classes but also removed borderline and noisy samples, improving class separation for model training. SMOTE-Tomek, a hybrid method combining oversampling and undersampling, was used to balance the training set by increasing minority samples and removing borderline instances [22].

4.2. Feature Selection

Feature selection was performed using a two-stage hybrid approach to ensure both statistical relevance and predictive performance. First, a filter-based method using an ANOVA F-test with false discovery rate correction was applied, resulting in the selection of 13 features that showed significant differences across the three movement quality classes ($p < 0.05$). Next, a wrapper-based method using recursive feature elimination with cross-validation with a Random Forest (RF) classifier was used to iteratively eliminate less informative features and identify those that maximized cross-validated performance. The final feature set was obtained by intersecting the results of both methods, resulting in three key features: knee valgus (Knee angle), knee-to-trunk ratio (Knee / Trunk), and amplified joint interaction (Trunk × Knee).

4.3. Classifiers

All implementations were conducted in Python 3.9 using libraries. For models that are sensitive to feature scaling, such as Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), Z-score normalization was applied within a pipeline, while tree-based models were trained on raw feature values. Six classifiers—LR, SVM, KNN, RF, AdaBoost, and LightGBM—were tuned via grid search with stratified 5-fold cross-validation, using F1-score as the optimization metric. Hyperparameter search ranges included the regularization type and strength for Logistic Regression, the kernel type and penalty parameters for SVM, the number of neighbors and distance metrics for KNN, the tree depth and number of estimators for ensemble tree methods, and the learning rate for boosting algorithms.

4.4. Voting Classifier

In addition to the individual models, a soft voting ensemble was constructed to integrate the predicted probabilities from individual models, thereby enhancing overall classification performance and robustness by mitigating the impact of potential misclassifications from any single model [23, 24]. To identify the optimal combination for soft voting, all pairwise combinations of the six base classifiers were evaluated using stratified 5-fold cross-validation on the training set, with the mean macro F1-score as the primary selection criterion. The combination of LR and LightGBM consistently achieved the highest mean macro F1-score across folds in the stratified 5-fold cross-validation. This pairing provided complementary predictive strengths, with LR offering robust performance for approximately linearly separable decision boundaries and LightGBM effectively modeling complex nonlinear patterns present in the kinematic features.

4.5. Model Evaluation

Model performance was evaluated using stratified 5-fold cross-validation on the training set and a held-out test set. Three primary metrics were used to assess classification performance: accuracy, macro-averaged F1-score, and multiclass area under the receiver operating characteristic curve (AUC). The macro-averaged F1-score was computed as the unweighted mean of the class-specific F1-scores, ensuring balanced evaluation across the three movement quality categories. Multiclass AUC was calculated using the one-vs-rest strategy, averaging the AUC scores obtained for each class. To assess the stability and reliability of the test set results, 95% confidence intervals (CIs) for each performance metric were estimated using 2000-iteration bootstrap resampling.

4.6. Model Explainability

Both SHAP and LIME analyses were conducted exclusively on the final trained model, ensuring consistency between the interpretability results and the model used for evaluation. This dual approach enabled the identification of class-specific feature importance at the global level and the extraction of actionable, instance-specific kinematic thresholds at the local level, supporting potential clinical and fitness applications [25, 26]. Global interpretability was assessed using SHAP applied to the soft voting ensemble. SHAP values were computed for each of the three selected features—knee angle, knee-to-trunk ratio (Knee / Trunk), and amplified joint interaction (Trunk × Knee)—to quantify their contribution to class predictions. Class-wise SHAP value aggregation was used to identify the most influential features for each movement quality category.

Local interpretability was examined using LIME to generate instance-level explanations. LIME was applied to representative samples from each class in the test set, with the number of perturbed samples set to 5000 to ensure stability of the local surrogate models. The output from LIME was used to identify kinematic threshold ranges for each of the selected features that were most predictive of class membership. These thresholds were derived from the feature values at the decision boundaries learned by the model, providing interpretable biomechanical markers for distinguishing between “Good,” “Moderate,” and “Poor” performance.

Results

1. Characteristics of the dataset

Participant's characteristics are summarized in Table 1. A total of 105 SLS videos were collected and annotated by expert physiotherapists into three functional categories based on movement quality: good (Class 0), moderate (Class 1), and poor (Class 2). The original training set comprised 73 samples, with an imbalanced distribution: 29 moderate, 25 poor, and 19 good. To address class imbalance, the training data were resampled using the SMOTE-Tomek method, resulting in a more balanced distribution: 27 good, 24 moderate, and 26 poor samples. The remaining 32 samples were reserved for testing, with the following class distribution: 12 moderate, 11 poor, and 9 good.

2. Classification Performance of Machine Learning Models

Six individual classifiers and a soft voting ensemble were evaluated using 5-fold cross-validation and a held-out test set (Table 2). Among the evaluated models, the soft voting ensemble (LR + LightGBM) demonstrated comparable accuracy (87.5% [95% CI: 75.0–96.9%]) and macro F1-score (87.6% [95% CI: 73.3–97.3%]) to the best-performing single model (LightGBM alone: accuracy 87.5% [95% CI: 75.0–96.8%] and macro F1-score 87.0% [95% CI: 74.2–97.2%]), but showed slightly superior AUC performance (0.938% [95% CI: 0.858–0.997%] vs. 0.932% [95% CI: 0.853–0.990%]) (Fig. 2). Because it combined the complementary strengths of LR and LightGBM, the soft voting model was selected as the final classifier for subsequent analysis. As illustrated in the confusion matrix (Fig. 3), the soft voting classifier exhibited strong class-wise performance on the test set. Specifically, it correctly classified 7 out of 9 Good, 11 out of 12 Moderate, and 10 out of 11 Poor samples. These results highlight the ensemble model's ability to reliably differentiate among varying levels of knee function using only kinematic features extracted from smartphone videos.

3. Key feature analysis and model interpretability

To interpret the model's decision-making process, we conducted both global and local feature importance analyses. SHAP analysis was applied to the soft voting model to assess class-wise contributions of the selected features (Fig. 4). Among the three features, the amplified joint interaction (Knee × Trunk) exhibited the highest average SHAP value, particularly contributing to the classification of Class 2 (Poor). The knee angle predominantly contributed to Class 0 (Good) and Class 1 (Moderate), while the knee-to-trunk ratio (Knee / Trunk), showed similar importance for Class 0 and Class 1. These findings suggest that the ensemble model captures distinct inter-joint dynamics that support interpretable, class-specific decision-making.

To enhance local interpretability, we applied LIME to visualize instance-level feature contributions. As illustrated in Figure 5, the classifier's decision boundaries aligned with interpretable kinematic thresholds. For example, movements classified as "Good" exhibited knee angles of –4.11 (below the threshold of –0.29), a knee-to-trunk ratio of –0.46 (below –0.02), and a trunk × knee interaction of –16.02 (less than –2.96) (Fig. 5A). "Moderate" movements showed intermediate values, such as trunk × knee of 8.29 (between –2.96 and 25.17), knee angle of 3.03 (between –0.29 and 3.55), and a knee-to-trunk ratio of 0.33 (between 0.18 and

0.48) (Fig. 5B). In contrast, “Poor” cases displayed markedly elevated values, with trunk × knee of 162.07 (above 72.74), a knee-to-trunk ratio of 3.12 (above 0.48), and a knee angle of 30.61 (above 10.29) (Fig. 5C). These numerical ranges align with the decision thresholds and confirm the model’s reliance on biomechanically relevant features for class-specific prediction.

Discussion

Principal Findings

This study developed and validated a robust, interpretable ensemble framework that classifies knee function into three levels (Good, Moderate, Poor) from single-smartphone, frontal-view SLS videos. Leveraging three engineered, biomechanically meaningful features, such as knee angle, knee-to-trunk ratio, and amplified Knee × Trunk interaction, the soft-voting ensemble of logistic regression and LightGBM achieved comparable performance to the best single model (LightGBM) on a held-out test set (accuracy 87.5%, F1-score 87.6% and AUC 0.938). Beyond accuracy, SHAP and LIME provided convergent global and local explanations that aligned the model’s predictions with clinically meaningful kinematic patterns, thereby bridging algorithmic decision-making with clinical reasoning. These explanations further enabled personalized feedback across the three functional grades of knee performance, derived entirely from single-smartphone video recordings.

Previous studies have reported that both wearable sensor–based and camera-based markerless models achieved moderate accuracy, typically ranging from 66% to 80% with an AUC of approximately 0.73, and often showed limited generalizability in three-class SLS classification [9, 11, 12]. In contrast, the present framework substantially improved performance, achieving 87.5% accuracy, an F1-score of 87.6%, and an AUC of 0.938 by integrating LR and LightGBM in a soft-voting ensemble trained on engineered biomechanical features. To move beyond guideline-based or composite measures, we generated a broader set of ratio-, interaction-, and nonlinear-based features to capture inter-joint coordination and applied a systematic selection process to identify the most informative predictors. The resulting features—knee angle, knee-to-trunk ratio, and Knee × Trunk interaction—provided complementary perspectives on movement quality that are not easily detected through conventional observational criteria, thereby extending the biomechanical basis for SLS assessment. Moreover, ML models must be trained on data obtainable in real-world settings; otherwise, even high-performing models remain clinically irrelevant when they depend on non-consumer sensors or multiple cameras [27]. The present framework overcomes this limitation by relying on a single, frontal smartphone video to extract easily obtainable features during routine SLS testing, enabling automatic classification of knee function levels and interpretable outputs that can be readily applied in practice.

Beyond predictive accuracy, the explainability analyses provided additional biomechanical insights into knee function. SHAP showed that the Knee × Trunk interaction had the strongest influence on poor classifications and contributed to moderate performance, whereas knee angle was more important for distinguishing good and moderate outcomes. Previous studies using individual joint or composite measures achieved only

moderate accuracy (66% with an SVM and 76.9% with a decision tree) [9, 11], without considering inter-segmental interactions. In contrast, our soft voting model reached 87.5% accuracy by incorporating a Knee × Trunk interaction feature, capturing simultaneous deviations that amplify coordination deficits and offering a more sensitive indicator than single-joint metrics. This observation aligns with biomechanical evidence that inter-joint coordination patterns are critical markers of movement quality during squatting and single-leg tasks [8], and supports the role of trunk–knee coupling as a central determinant of functional impairment [28]. Complementary LIME analyses further illustrated instance-specific decision boundaries, showing that markedly elevated Knee × Trunk values predicted poor classifications, whereas lower knee angles and ratios aligned with good outcomes. By clarifying why a patient is classified as having poor knee function, the XAI results allow clinicians and trainers to better trust the ML-based SLS evaluation. Moreover, they can guide expert decisions on whether management strategies should focus solely on the knee or also address altered trunk–knee coordination, thereby supporting more personalized rehabilitation and remote monitoring.

Limitations

Despite these promising results, several limitations should be acknowledged. The dataset size was relatively small (105 participants), which may increase the risk of overfitting. To mitigate this, stratified cross-validation, bootstrap resampling, and systematic feature selection were applied to assess the stability of the results [14]. In addition, the data were restricted to young healthy adults, which may limit generalizability to clinical populations or older adults. Future studies should therefore expand to more diverse cohorts, including older adults, injured patients, and athletes. Clinical trials are needed to evaluate whether the model improves real-world rehabilitation decision-making and patient outcomes. It will also be important to test ML- and XAI-based applications in clinical, fitness, and rehabilitation settings to determine whether automated knee function classification can reduce assessment time and improve efficiency. In addition, future research should examine whether integrating XAI feedback into personalized exercise prescriptions increases satisfaction among clinicians and users, and whether the framework can support long-term remote monitoring and prediction of treatment responses.

Conclusion

This study developed an interpretable ensemble learning framework that accurately classified knee function into three levels from single-smartphone SLS videos, achieving high performance (accuracy 87.5%, F1-score 87.6%, AUC 0.938) while maintaining transparency through SHAP and LIME analyses. By incorporating the Knee × Trunk interaction feature, along with knee angle and the knee-to-trunk ratio, the framework offered a novel perspective on movement patterns beyond conventional criteria and provided biomechanically meaningful explanations that clinicians and trainers can trust for personalized feedback. Overall, this smartphone-based approach offers a practical, accessible, and efficient solution for functional knee assessment, with XAI-based explanations enhancing clinical trust and supporting its applications in rehabilitation, athletic screening, and remote monitoring.

References

1. HASSEN A. Surgical options for the middle-aged patient with osteoarthritis of the knee joint. Intr Course Lect. 2001;50:499-511.

2. Filbay SR, Crossley KM, Ackerman IN. Activity preferences, lifestyle modifications and re-injury fears influence longer-term quality of life in people with knee symptoms following anterior cruciate ligament reconstruction: a qualitative study. J Physiother. 2016;62(2):103-10.

3. Ibeachu C, Selfe J, Sutton CJ, Dey P. Knee problems are common in young adults and associated with physical activity and not obesity: the findings of a cross-sectional survey in a university cohort. BMC Musculoskelet Disord. 2019 Mar 18;20(1):116. PMID: 30885176. doi: 10.1186/s12891-019-2487-2.

4. Narducci E, Waltz A, Gorski K, Leppla L, Donaldson M. The clinical utility of functional performance tests within one-year post-acl reconstruction: a systematic review. Int J Sports Phys Ther. 2011 Dec;6(4):333-42. PMID: 22163095.

5. Khuu A, Foch E, Lewis CL. Not all single leg squats are equal: A biomechanical comparison of three variations. Int J Sports Phys Ther. 2016 Apr;11(2):201-11. PMID: 27104053.

6. Ageberg E, Bennell KL, Hunt MA, Simic M, Roos EM, Creaby MW. Validity and inter-rater reliability of medio-lateral knee motion observed during a single-limb mini squat. BMC Musculoskelet Disord. 2010;11(1):265.

7. Ressman J, Grooten WJA, Barr ER. Visual assessment of movement quality in the single leg squat test: a review and meta-analysis of inter-rater and intrarater reliability. BMJ Open Sport Exerc Med. 2019;5(1).

8. Nae JÄ, Banega M, Kronström F, Jakobsson A, Creaby MW, Ageberg E. Evaluation of the ability of machine learning-models to assess postural orientation errors during a single-leg squat. JOSPT Methods. 2025;1(1):17-29.

9. Kianifar R, Lee A, Raina S, Kulić D. Automated assessment of dynamic knee valgus and risk of knee injury during the single leg squat. IEEE J Transl Eng Health Med. 2017;5:1-13.

10. Chariar M, Rao S, Irani A, Suresh S, Asha C. AI trainer: Autoencoder based approach for squat analysis and correction. IEEE Access. 2023;11:107135-49.

11. Park J, Kim Y, Kim S, Park K. Markerless kinematic data in the frontal plane contributions to movement quality in the single-leg squat test: a comparison and decision tree approach. J Sport Rehabil. 2024;1(aop):1-8.

12. Dajime PF, Smith H, Zhang Y. Automated classification of movement quality using the Microsoft Kinect V2 sensor. Comput Biol Med. 2020;125:104021.

13. Zheng X, Otten E, Reneman MF, Lamoth CJ. Explaining deep learning models for age-related gait classification based on acceleration time series. Comput Biol Med. 2025;184:109338.

14. Zheng X, Reneman MF, Echeita JA, Preuper RHS, Kruitbosch H, Otten E, et al. Association between central sensitization and gait in chronic low back pain: Insights from a machine learning approach. Comput Biol Med. 2022;144:105329.

15. Tak I, Wiertz W-P, Barendrecht M, Langhout R. Validity of a New 3-D Motion Analysis Tool for the Assessment of Knee, Hip and Spine Joint Angles during the Single Leg Squat. Sensors (Basel, Switzerland). 2020 2020-08-01;20. doi: 10.3390/s20164539.

16. Guo L, Chang R, Wang J, Narayanan A, Qian P, Leong MC, et al. Artificial intelligence-enhanced 3D gait analysis with a single consumer-grade camera. Journal of biomechanics. 2025;187:112738.

17. Rabello R, Bertozzi F, Brunetti C, Silva Zandonato L, Bonotti A, Rodrigues R, et al. The influence of task type and movement speed on lower limb kinematics during single-leg tasks. Gait Posture. 2022 Jul;96:109-16. PMID: 35635986. doi: 10.1016/j.gaitpost.2022.05.020.

18. Barker-Davies RM, Roberts A, Bennett AN, Fong DTP, Wheeler P, Lewis MP. Single leg squat ratings by clinicians are reliable and predict excessive hip internal rotation moment. Gait Posture.. 2018 Mar;61:453-8. PMID: 29486363. doi: 10.1016/j.gaitpost.2018.02.016.

19. Ressman J, Grooten WJA, Rasmussen-Barr E. Visual assessment of movement quality: a study on intra- and interrater reliability of a multi-segmental single leg squat test. BMC Sports Sci Med Rehabil 2021 Jun 8;13(1):66. PMID: 34099021. doi: 10.1186/s13102-021-00289-x.

20. Räisänen A, Pasanen K, Krosshaug T, Avela J, Perttunen J, Parkkari J. Single-leg squat as a tool to evaluate young athletes' frontal plane knee control. Clin J Sport Med. 2016 Nov;26(6):478-82. PMID: 26704374. doi: 10.1097/jsm.0000000000000288.

21. Lafayette TBG, Kunst VHL, Melo PVS, Guedes PO, Teixeira J, Vasconcelos CR, et al. Validation of angle estimation based on body tracking data from rgb-d and rgb cameras for biomechanical assessment. Sensors (Basel, Switzerland). 2022 Dec 20;23(1). PMID: 36616603. doi: 10.3390/s23010003.

22. Wang Z, Wu C, Zheng K, Niu X, Wang X. SMOTETomek-based resampling for personality recognition. IEEE access. 2019;7:129678-89.

23. Yüzkat M, Ilhan HO, Aydin N. Multi-model CNN fusion for sperm morphology analysis. Comput Biol Med. 2021;137:104790.

24. Cansiz B, Kilinc CU, Serbes G. Tunable Q-factor wavelet transform based lung signal decomposition and statistical feature extraction for effective lung disease classification. Comput Biol

Med. 2024;178:108698.

25. Azad M, Khan MFK, Abd El-Ghany S. XAI-Enhanced machine learning for obesity risk classification: a stacking approach with LIME explanations. IEEE Access. 2025.

26. Yue L, Chen W-g, Liu S-c, Chen S-b, Xiao S-f. An explainable machine learning based prediction model for Alzheimer's disease in China longitudinal aging study. Front aging neurosci. 2023;15:1267020.

27. Markowetz F. All models are wrong and yours are useless: making clinical prediction models impactful for patients. NPJ Precis Oncol. 2024;8(1):54.

28. Plummer HA, Oliver GD, Powers CM, Michener LA. Trunk lean during a single-leg squat is associated with trunk lean during pitching. Int J Sports Phys Ther. 2018 Feb;13(1):58-65. PMID: 29484242.

Preprint JMIR Publications

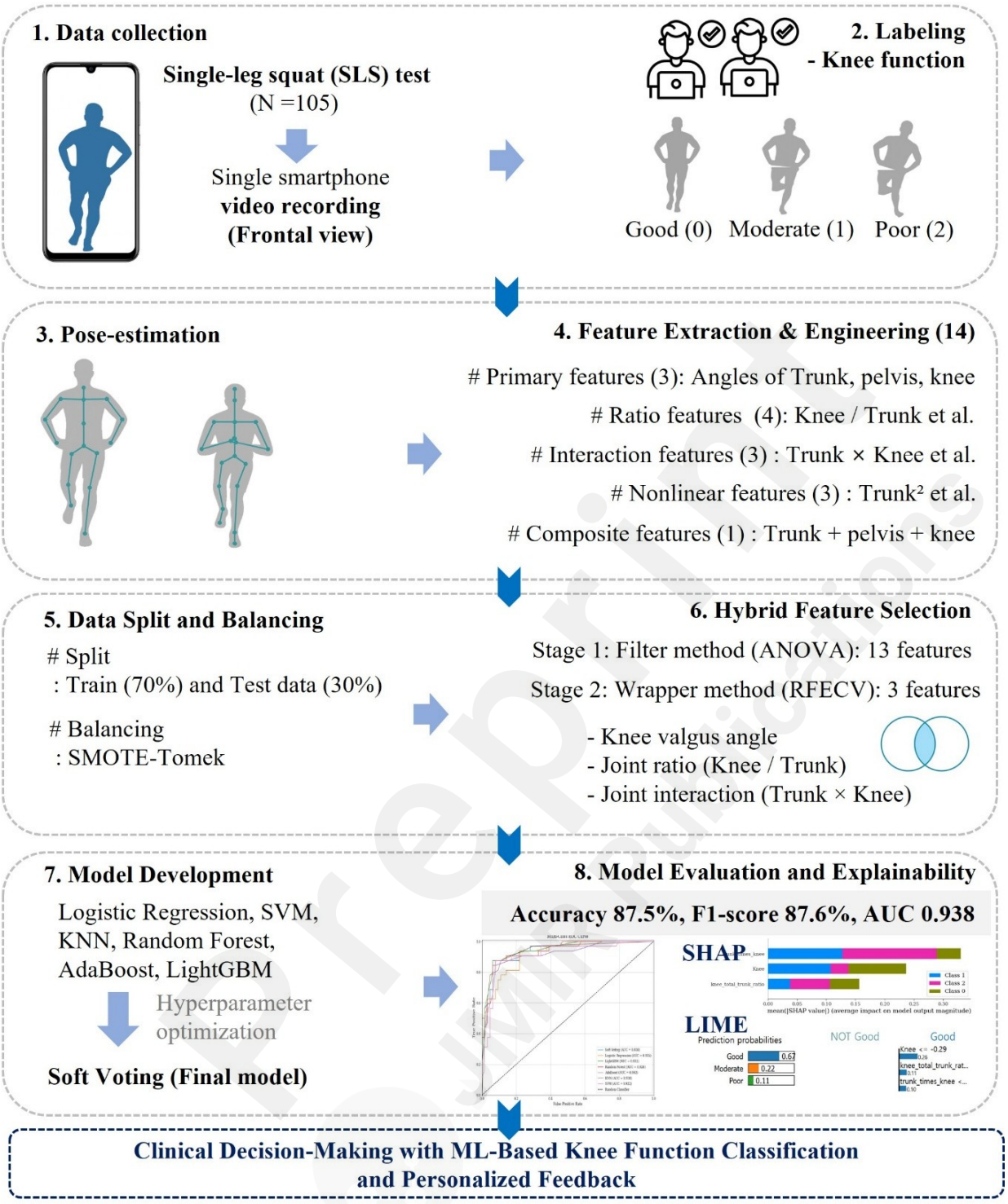


Fig. 1. Flowchart of machine learning model development for knee function classification and feedback.

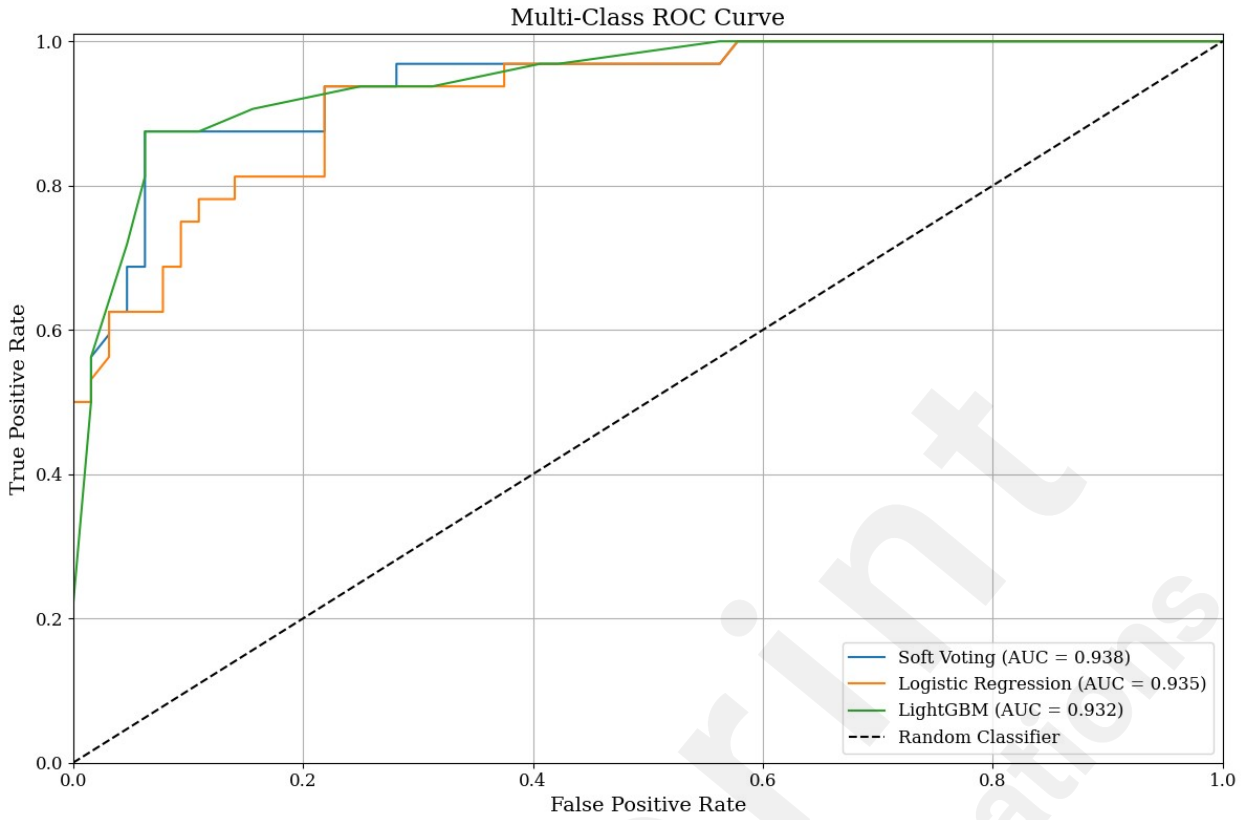


Fig. 2. The receiver operating characteristic (ROC) curves and area under the curve (AUC) values for soft voting ensemble with baseline classifiers (Logistic regression and LightGBM).

Preprint JMIR Publications



Fig. 3. Confusion matrices of the soft voting ensemble, logistic regression, and LightGBM models.

Preprint
JMIR Publications

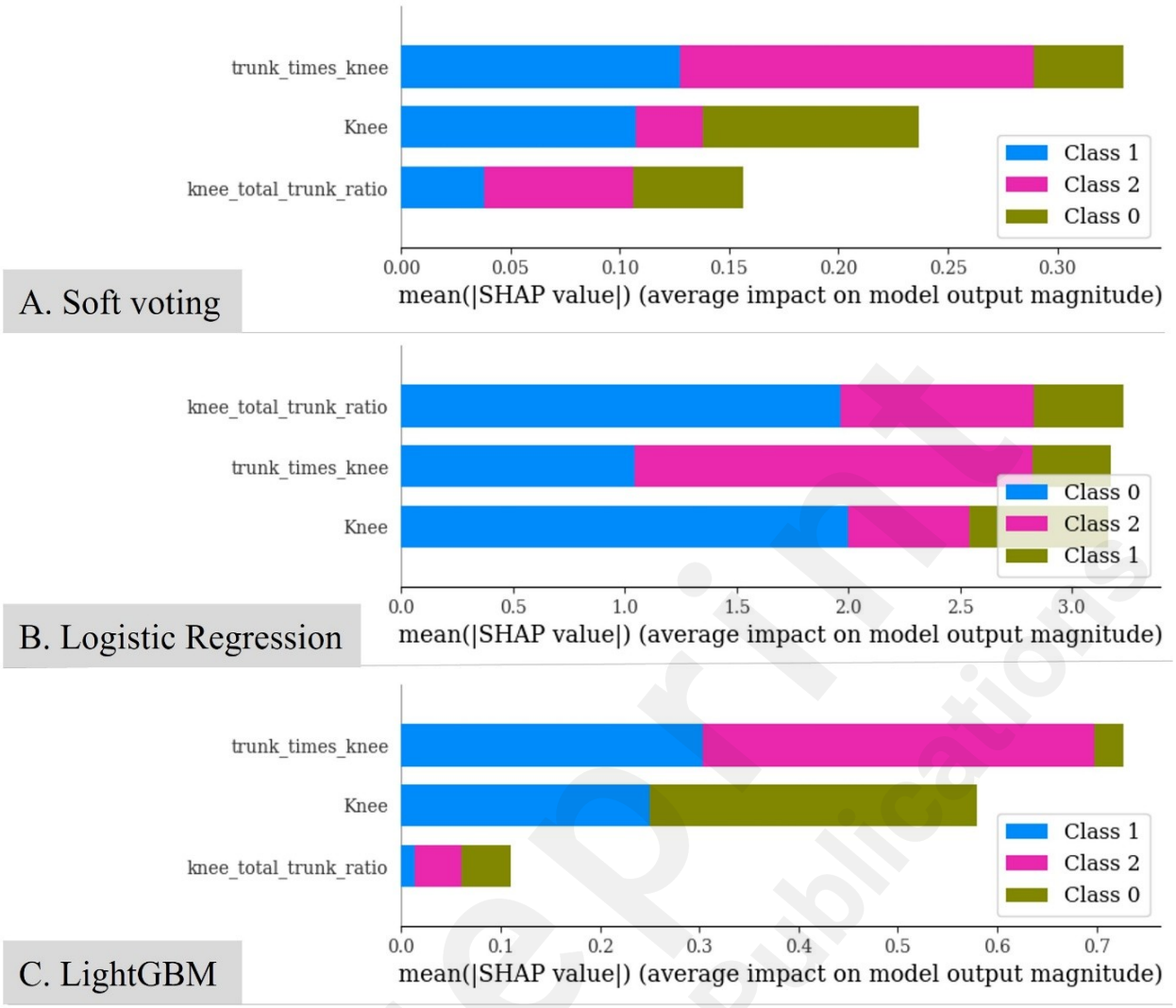


Fig. 4. SHAP summary plots for feature importance across models (A. Soft voting, B. Logistic Regression and C. LightGBM). Class labels correspond to knee function levels: Class 0 = Good, Class 1 = Moderate, Class 2 = Poor.

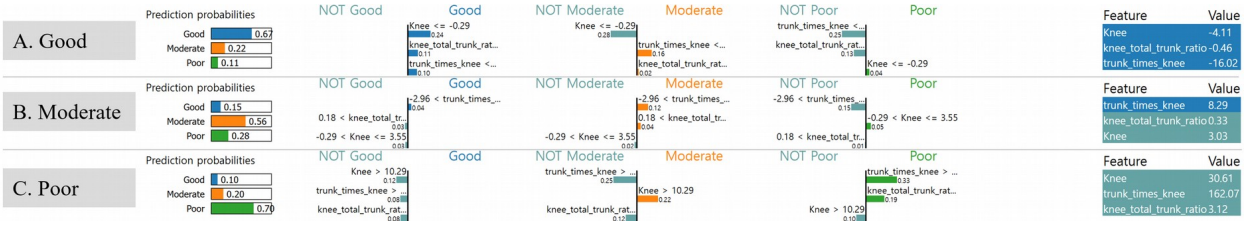


Fig. 5. LIME explanations of instance-level feature contributions, showing how trunk×knee interaction, knee-to-trunk ratio, and knee angle influenced classification across A. Good (Class 0), B. Moderate (Class 1), and C. Poor (Class 2).

Preprint JMIR Publications

Table 1. Participants characteristics (N=105)

	Knee function grading using the single-leg squat test (labeling)			
	Good (0)	Moderate (1)	Poor (2)	*P*
Number of participants	28	41	36	-
Gender (male/female)	13/15	21/20	19/17	-
Age (years)	22.36 ± 2.03[a]	22.46 ± 2.85	22.20 ± 2.47	0.75
Body mass index (m/kg^2)	22.38 ± 3.35	22.80 ± 3.59	23.22 ± 3.33	0.53
Exercise duration per week (min)	445.42 ± 286.15	353.25 ± 339.16	331.57 ± 215.35	0.38
Trunk lateral tilt (°)[b]	6.00 ± 2.47	7.24 ± 3.23	10.24 ± 6.34	<0.01
Pelvic lateral tilt (°)[b]	6.31 ± 1.93	8.50 ± 2.76	10.38 ± 4.94	<0.01
Knee valgus (°)[b]	-2.63 ± 2.55	3.67 ± 3.75	13.24 ± 9.71	<0.01

[a]Mean±standard deviation

[b]Peak joint angles measured during the single-leg squat test

Preprint JMIR Publications

Table 2. Performance metrics of classifiers.

Classifier	Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC
Logistic regression	78.1%	77.9%	78.5%	89.1 %	78.1%	0.935
SVM	84.4%	85.1%	84.0%	92.0%	84.4%	0.922
KNN	81.3%	85.9%	80.7%	90.1%	82.0%	0.938
Random forest	84.4%	86.8%	84.0%	91.8%	84.8%	0.928
AdaBoost	84.4%	86.7%	83.0%	91.8%	83.8%	0.902
LightGBM	87.5%	89.8%	86.7%	93.4%	87.6%	0.932
Soft voting[a]	87.5%	89.8%	86.7%	93.4%	87.6%	0.938

[a]The soft voting ensemble combined logistic regression and LightGBM.

Preprint JMIR Publications