Ensemble Algorithm for Risk Prediction of Clinical Failure After Anterior Cruciate Ligament Reconstruction

Tianlun Zhang,* MD, Zipeng Ye,* MD, Jiangyu Cai,* MD, Jiebo Chen,* MD, Ting Zheng,* MD, Junjie Xu,*† MD, and Jinzhong Zhao,*† MD
Investigation performed at the Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

Background: Patient-specific risk profiles of clinical failure after anterior cruciate ligament reconstruction (ACLR) are meaningful for preoperative surgical planning and postoperative rehabilitation guidance.

Purpose: To create an ensemble algorithm machine learning (ML) model and ML-based web-based tool that can predict the patient-specific risk of clinical failure after ACLR.

Study Design: Cohort study; Level of evidence, 3.

Methods: Included were 432 patients (mean age, 26.8 ± 8.4 years; 74.1% male) who underwent anatomic double-bundle ACLR with hamstring tendon autograft between January 2010 and February 2019. The primary outcome was the probability of clinical failure at a minimum 2-year follow-up. The authors included 24 independent variables for feature selection and model development. The data set was split randomly into training sets (75%) and test sets (25%). Models were built using 4 ML algorithms: extreme gradient boosting, random forest, light gradient boosting machine, and adaptive boosting. In addition, a weighted-average voting (WAV) ensemble model was constructed using the ensemble-voting technique to predict clinical failure after ACLR. Concordance (area under the receiver operating characteristic curve [AUC]), calibration, and decision curve analysis were used to evaluate predictive performances of the 5 models.

Results: Clinical failure occurred in 73 of the 432 patients (16.9%). The 8 most important predictors for clinical failure were follow-up period, high-grade preoperative knee laxity, time from injury to ACLR, participation in competitive sports, posterior tibial slope, graft diameter, age at surgery, and medial meniscus resection. The WAV ensemble algorithm achieved the best predictive performance based on concordance (AUC, 0.9139), calibration (calibration intercept, –0.1806; calibration slope, 1.2794; Brier score, 0.0888), and decision curve analysis (greatest net benefits) and was used to develop an web-based application to predict a patient's clinical failure risk of ACLR.

Conclusion: The WAV ensemble algorithm was able to accurately predict patient-specific risk of clinical failure after ACLR. Clinicians and patients can use the web-based application during preoperative consultation to understand individual prediction outcomes.

Keywords: anterior cruciate ligament reconstruction; clinical failure; machine learning; artificial intelligence; ensemble method; open-access application

The anterior cruciate ligament (ACL) is a commonly injured ligament in the knee, especially in athletic populations.[8,17,29] Regardless of surgical technique, treatment aims to increase anteroposterior and rotational stability, restoring the native knee biomechanics in terms of tibiofemoral load bearing during movement as much as possible.[3] Despite the recent advances in surgical techniques, ACL reconstruction (ACLR) does not come without adverse outcomes, including residual rotatory laxity and graft rupture.[2,13,17,24,28,48] Persistent rotatory laxity and graft rupture are associated with poor functional outcomes and the subsequent need for revision surgery.[4]

Recent studies have indicated that the risk factors of clinical failure after ACLR include younger age at surgery,[23,36] increased posterior tibial slope,[56] meniscal

The Orthopaedic Journal of Sports Medicine, 12(8), 23259671241261695
DOI: 10.1177/23259671241261695

deficiency,[45] small graft size,[32] and allograft reconstruction.[34] One drawback is that these studies provide only relevant information on a general level and are not able to accurately determine the specific risk for each patient in order to deliver individualized prediction conclusions. Patient-specific risk is challenging to estimate and quantify because of the complicated interactions between numerous characteristic factors. In this context, it is important to provide more comprehensive knowledge and user-friendly tool to pinpoint the types of patients who might encounter clinical failure of ACLR and determine their specific risk factors, especially in patients with high return-to-sports expectations and functional demands.[27]

Machine learning (ML) can increase the predictive power and viability as an emerging method of health care research.[10,16,22,27] ML algorithms have been shown to learn from adequate instances and adapt their internal parameters (weights) and reinforce pertinent associations, thus increasing the correctness of a specific mathematical model.[39] Furthermore, compared with conventional statistics, ML algorithms can handle more complex interactions in large data sets and increase prediction accuracy.[27] Previous reviews have shown that numerous orthopaedic surgeons have been quite interested in predicting some clinical outcomes of different procedures using ML algorithms, including ACLR,[25] dissatisfaction after primary total knee arthroplasty,[26] ACLR revision,[36] and overnight hospital admission after ACLR.[30] However, only a single ML model was selected for risk factor prediction in these studies, such as the elastic-net penalized logistic regression,[25] the Cox lasso,[36] and the random forest (RF) algorithm,[26] and every algorithm model has its individual disadvantages.[54] According to previous studies, a nested ensemble method showed the best performance among the candidate ML models.[30,54] An ensemble algorithm combines $\geq$2 single algorithm models with varying strengths and weaknesses to build a more sustainable model with better performance.[16]

To our knowledge, no study has used an ensemble algorithm for risk prediction of clinical failure after the ACLR, although its predictive effectiveness and accuracy have been well documented recently in some other areas of medicine.[9,30,40,54] Therefore, in this study, we aimed to develop an ML ensemble model and a user-friendly, ML-based, web-based application for this purpose. We hypothesized that an ensemble ML algorithm with superior accuracy to the single ML algorithms would be a trustworthy tool to predict the clinical failure of patients after ACLR.

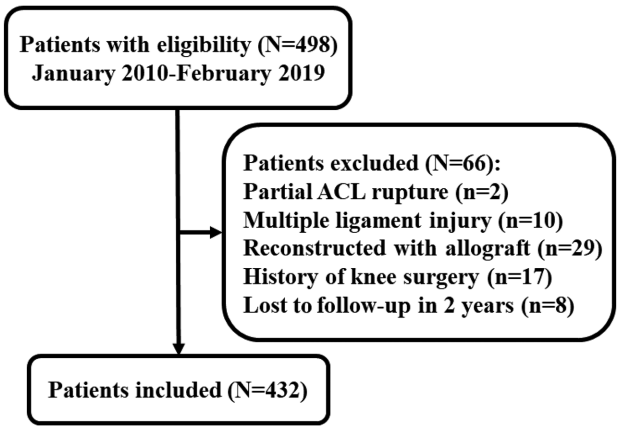

Figure 1. Flowchart of patient selection. ACL, anterior cruciate ligament.

METHODS

Study Population

The present study was carried out in accordance with the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guidelines and the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research.[7,31] The protocol for this study was approved by our institutional review board; informed consent was waived. We reviewed the records of 498 patients who underwent primary anatomic double-bundle ACLR in our institution between January 2010 and February 2019. Exclusion criteria consisted of (1) partial ACL rupture; (2) combined ligament injury, skeletal immaturity, or fracture; (3) reconstruction with allograft or artificial ligament; (4) history of ipsilateral knee surgery; and (5) incompleteness of data provided for 2-year outcomes. Of the initial 498 patients, we excluded 66 (13.3%), leaving 432 included in our study (Figure 1).

Primary Study Outcome

The primary outcome was ACLR clinical failure, a composite measure of rotatory laxity or a graft rupture.[15] Rotatory laxity was defined as a moderate or severe (grade 2 or 3) asymmetric pivot shift at any follow-up visit or a persistent (detected at $\geq$2 visits) mild asymmetric pivot shift (grade

†Address correspondence to Jinzhong Zhao, MD, Department of Sports Medicine, Shanghai Sixth People's Hospital, 600 Yishan Road, Shanghai, 200233, China (email: jzzhao@sjtu.edu.cn); and Junjie Xu, MD, Department of Sports Medicine, Shanghai Sixth People's Hospital, 600 Yishan Road, Shanghai, 200233, China (email: david_jj_xu@hotmail.com).

*Department of Sports Medicine, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China.

T.Z. and Z.Y. contributed equally to this article.

J.X. and J.Z. contributed equally to this article.

Final revision submitted December 4, 2023; accepted January 10, 2024.

One or more of the authors has declared the following potential conflict of interest or source of funding: Support for this study was received from the Science and Technology Commission of Shanghai Municipality (grant 22dz1204700). AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.

Ethical approval for this study was obtained from Shanghai Sixth People's Hospital (ref No. 2020-KY-030 [K]).

1).[15] Graft rupture was defined as discontinuity of the graft confirmed by either magnetic resonance imaging or arthroscopic examination.[15]

Candidate Covariates

A total of 24 preoperative and intraoperative covariates prospectively obtained in our department were tested for predictive value. Patient characteristics included age at surgery, sex, body mass index, participation in competitive sports, injured side, limited range of motion, limited sports ability, joint pain, and time from injury to surgery. The posterior tibial slope was obtained by true lateral knee radiographs.[47] The anterior drawer test, Lachman test, and pivot-shift test were performed manually under anesthesia according to the International Knee Documentation Committee (IKDC) guidelines.[14] The presence of a grade 3 anterior drawer test, Lachman test, or pivot-shift test was considered as high-grade preoperative knee laxity.[33,43] Additionally, subjective clinical assessments, including the IKDC score, Lysholm score, and Tegner activity score, were obtained from patients before surgery. Graft diameter, meniscal injury treatments, and graft length in the tunnel were recorded intraoperatively. The postoperative follow-up period was also recorded.

Model Development and Creation

Feature Extraction. The significant predictive features were selected by applying recursive feature elimination (RFE) with RF algorithms.[18] The RFE created a model with all features, sorting them by the importance score. After eliminating low-ranked features with the lowest scores, another unique model was built. The process was repeated until a subset of predictors with the best predictive performance was determined.

Base ML and Ensemble Algorithm. Four well-accepted base ML algorithms, including extreme gradient boosting (XGBoost), RF, light gradient boosting machine (LightGBM), and adaptive boosting (Adaboost), were used to develop predictive models with the selected features mentioned above in the current study.[16] These ML methods were compared according to their inherent characteristics.[42] An ensemble algorithm was combined using an ensemble voting technique based on weighted-average voting (WAV) to predict clinical failure after ACLR.[54] Suppose there are M number of classification algorithms (CAs); each is denoted as CA_1, CA_2, …, CA_M. The WAV ensemble did not simply sum up the posterior probabilities obtained for each sample from the M classification algorithms. Briefly, the performance of each model on the validation data set could be evaluated after optimizing their hyperparameters. Then, we assigned weights, to each model to create our WAV ensemble model according to their performances (between 0 and 1 and could add up to 1). The final predictive result ($\hat{y}$) of the voting ensemble algorithm was computed using the following equation[54]:

$$\hat{y} = W_1 \times Pr(CA_1(X)) + W_2 \times Pr(CA_2(X)) + \ldots + W_M \times Pr(CA_M(X))$$

such that $\Sigma W_1 + W_2 + \ldots + W_M = 1$ if $\hat{y} > 0.5$; the prediction result of the sample was clinical failure. In these equations, W (weight) indicates the percentage of trust on a particular classifier based on its performance in the training data set, Pr indicates the posterior probability, and X represents the independent variables.

Model Training and Hyperparameter Optimization. The data set with selected features was randomly split into 2 parts: 75% of the data set for training and 25% for testing.[25,26,35,54] Each base model was optimized on the training set utilizing 10-fold cross-validation with 10 repetitions and was tuned according to the training errors and mean of results in the training process.[25,26,35,54] Briefly, the training data set was randomly divided into 10 equally sized pieces. We used the exhaustive grid search algorithm to predict clinical failure with 9 pieces to tune the hyperparameters of each model, while the remaining piece was used as the validation set to evaluate each model's performance with the hyperparameters mentioned before. Thus, the cross-validation procedure made different subsets of samples available for optimizing the model, which made the model more robust and reduced errors as much as possible. The most optimal hyperparameters for each model were obtained in this process. Model evaluation on Monte Carlo cross-validation was performed using bootstrapping with 1000 resampled data sets in the independent testing set. A flowchart of the process is shown in Figure 2.

Model Performance Assessment

Model performance was assessed by calculating the predicted probabilities for clinical failure after ACLR in the independent testing data set (remaining 25%) (Figure 2). The performance of the 4 base ML models and the WAV ensemble algorithm in optimal prediction was evaluated in terms of concordance, calibration, and Brier score. Concordance was measured using the area under the receiver operating characteristic curve (AUC), which was applied to evaluate the accuracy of the model by considering its specificity and sensitivity; the value of AUC ranges from 0 to 1, with 1 indicating perfect concordance. Calibration referred to the accuracy of the predicted probabilities, which compared the predicted outcomes with the actual observed outcomes. The mean misclassification in each predicted risk was summarized in a calibration plot, with a straight line with an intercept of 0 and a slope of 1 representing perfect concordance of model expectation to observed frequencies (ie, ideal model calibration). The Brier score is a proper scoring function assessing the overall performance and an extension of calibration and discrimination. The Brier score of each model is the mean squared difference between the true observed outcomes and the model prediction probabilities. In general, lower Brier scores indicate better calibration of predictions, with 0 being perfect performance and calibration.

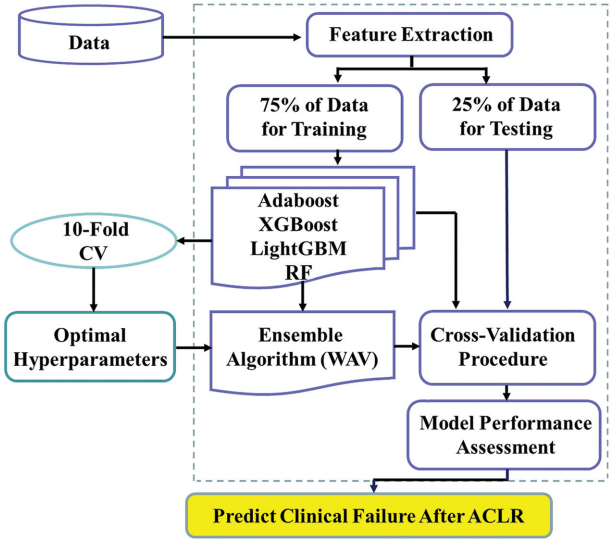

Figure 2. Flowchart visualizing the machine learning algorithm development and evaluation process. ACLR, anterior cruciate ligament reconstruction; CV, cross-validation; LightGBM, light gradient boosting machine; RF, random forest; WAV, weighted-average voting ensemble; XGBoost, extreme gradient boosting.

A comparison was made between the Brier score of each algorithm model and the null model (the null model–predicted probabilities were equivalent to the true outcome prevalence in the entire study cohort).

Decision Curve Analysis

Decision curve analysis is able to reveal the benefit of applying the predictive algorithm in clinical practice. There is a cutoff point to determine the occurrence of interested clinical events (also known as the threshold value) in the AUC. The cutoff point is the specific value of an indicator used to judge the occurrence of interested clinical events (denoted as A). The threshold probability refers to the constituent ratio after classification according to the cutoff point (*number of persons* $\geq A$/*total number of persons*). The net benefit value represents the increased profit of changing clinical management at this threshold probability. The calculation of the net benefit is shown in the following equation[52]:

$$\text{Net benefit} = \text{True positive rate} - (\text{False positive rate} \times \text{Weighting factor})$$

where the weighting factor is calculated as *Threshold probability*/(1 – *Threshold probability*).

The relationship between the threshold probability and net benefit can also be drawn as a line graph, which is the decision curve. Making a plot of the threshold probability and net benefit for the different decision models and putting them together creates a decision curve for comparison of different prediction models. The analysis of the decision curve in this study included 4 curves: the complete WAV ensemble algorithm model, a simplified ensemble model with only 2 optimal predictors (follow-up period and high-grade preoperative knee laxity), and 2 baselines (no treatment for all patients and treatment for all patients). The farther the curve of the decision model is from the baseline, the higher the net benefit of the decision is.

Individual-Level Interpretations and Digital Application

Local interpretable model-agnostic explanations (LIME) is a method that enables patient-specific interpretations of a model.[46,51] LIME samples local input variable distributions using a predefined number of permutations and evaluates the effect of values within a specific range for each predictor feature on the primary outcome.[25] The importance of each feature is computed and carried forward based on similarities between the features and the model predictions.[25,44] LIME can provide a visual explanation of how each feature contributes to the global predictions, showing how each feature either supports (increases the probability of clinical failure) or contradicts (decreases the probability of clinical failure) the prediction.[25,30]

With the use of LIME, we transformed the candidate prediction model with the best performance into an open-access risk calculator website application accessible on desktops and smartphones. With this application, users are able to get visible predictions of clinical failure after ACLR with accompanying explanations.

RESULTS

Patient Characteristics

The mean age of the 432 study patients was 26.8 ± 8.4 years, 74.1% were male, and 54.2% participated in sports. The full preoperative and intraoperative data are presented in Table 1. In the study population, 73 (16.9%) were classified as having experienced clinical failure (16 experienced graft rupture) during a mean follow-up period of 72.2 ± 37.3 months.

Feature Selection Outcomes

There were no missing data for any of the candidate covariates. Features were selected using RFE with RF algorithms, and the 8 most important predictors for clinical failure in the model were follow-up period, high-grade preoperative knee laxity, time from injury to surgery, participation in competitive sports, posterior tibial slope, graft diameter, age at surgery, and medial meniscus resection (Figure 3).

Predictive Model Performance and Comparison

The AUCs of the 4 base models with 10-fold cross-validation in the training data set (75% of data) were as follows: XGBoost, 0.932 (95% CI, 0.905-0.959); RF, 0.894 (95%

TABLE 1
Baseline Characteristics and Intraoperative Findings for Patients Included in the Machine Learning Analysis (N = 432)[a]

Variable	Value
Sex, male/female, n	320/112
Age at surgery, y	26.8 ± 8.4
Body mass index, kg/m^2	24.4 ± 3.4
Follow-up period, mo	72.2 ± 37.3
Preinjury Tegner score	7.1 ± 1.4
Preoperative Lysholm score	64.1 ± 20.7
Preoperative IKDC score	54.0 ± 19.0
Posterior tibial slope, deg	10.6 ± 2.3
Physical examination findings, n	
Anterior drawer test, 0/1/2/3	0/348/84/0
Lachman test, 0/1/2/3	0/26/300/106
Pivot-shift test, 0/1/2/3	0/207/158/67
High-grade preoperative knee laxity	112 (25.9)
Meniscal lesions	
Medial meniscal tear	91 (21.1)
Lateral meniscal tear	118 (27.3)
Bimeniscal tear	68 (15.7)
Medial meniscus resection	72 (16.7)
Lateral meniscus resection	90 (20.8)
Graft length in the tunnel, mm	24.8 ± 2.5
Time from injury to surgery, mo	22.3 ± 39.4
Graft diameter, mm	9.9 ± 0.7
Participation in competitive sports	234 (54.2)
Joint pain	186 (43.1)
Limitation of joint movement	332 (76.9)
Limited motor ability	147 (34.0)

[a]Data are presented as mean ± SD or n (%) unless otherwise indicated. AMB, anteromedial bundle; IKDC, International Knee Documentation Committee; PLB, posterolateral bundle.

CI, 0.861-0.926); LightGBM, 0.935 (95% CI, 0.913-0.956); and Adaboost, 0.945 (95% CI, 0.931-0.958) (Figure 4). Based on the model from the training data set, we predicted the patient-specific risk of clinical failure after ACLR in the testing data set (remaining 25% of data). The receiver operating characteristic curves for the testing data set indicated that the WAV ensemble model exhibited the best AUC (0.9139) among all of the predictive models (Figure 5).

Details of the performance assessment for the 4 base ML models and the WAV ensemble model are shown in Table 2. The WAV ensemble model achieved the best calibration and Brier score (calibration intercept, –0.1806; calibration slope, 1.2794; Brier score, 0.0888) (Table 2 and Figure 6). Calibration plots for the 4 base ML models on the testing data set are displayed in Figure 6.

Decision Curve Analysis

The results of the decision curve analysis are shown in Figure 7. It could be observed that the simplified ensemble model curve was above the baseline of the 2 hypotheses (treat all and treat none). The complete WAV ensemble algorithm had a higher rise compared with the simplified ensemble model curve in almost all interval ranges, implying that more true-positive results could be identified without increasing the false-positive rate for patients who expected a prediction when we applied the WAV ensemble model. This indicated that the WAV ensemble algorithm had potential application value. Moreover, using more information related to patient characteristics could increase the clinical application value of an ensemble model.

Digital Application and Individual Explanations

An open-source digital application (https://zorthoapps.shinyapps.io/aclr) was built to be accessible on desktop computers, tablets, and smartphones (Figure 8). This application generated a risk assessment of the clinical failure after ACLR on the practical implementation for each patient. We included 2 examples to show our open-source website application (Figure 9). Patient 1 was 25 years old, had preoperative high-grade knee laxity, and needed medial meniscus resection; at one point, he participated in competitive sports (Figure 9A). Four features contradicted experiencing clinical failure: posterior tibial slope,

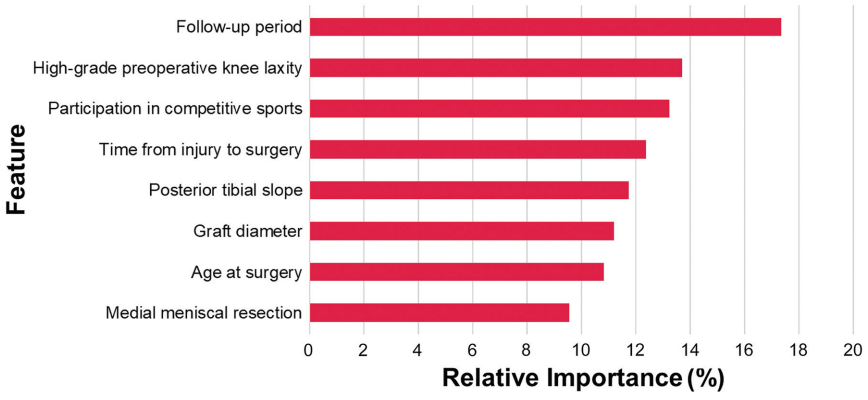

Figure 3. The 8 most important predictor variables for clinical failure after anterior cruciate ligament reconstruction as selected using recursive feature elimination with random forest algorithms.

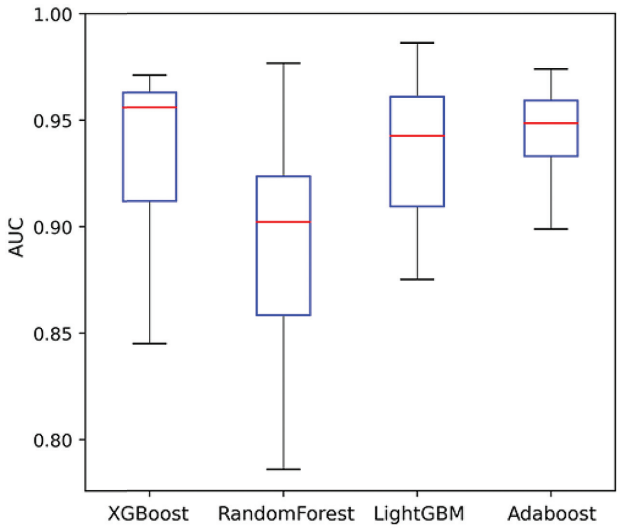

Figure 4. Box plot of area under the receiver operating characteristic curve (AUC) for machine learning models with 10-fold cross-validation in the training data set (75% of data). The red line indicates the mean, the top and bottom of the box indicate the 95% CI, and the whiskers indicate values in the dataset not exceeding 1.5 times interquartile range. Adaboost, adaptive boosting; LightGBM, light gradient boosting machine; XGBoost, extreme gradient boosting.

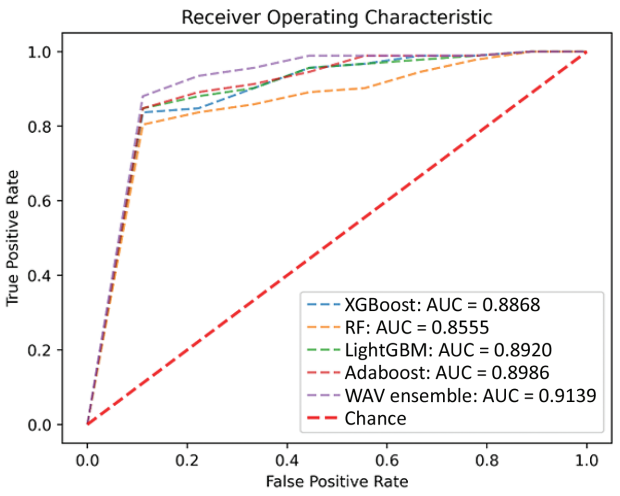

Figure 5. Receiver operating characteristic curves of machine learning models in the testing data set (25% of data). Adaboost, adaptive boosting; AUC, area under the receiver operating characteristic curve; LightGBM, light gradient boosting machine; RF, random forest; WAV, weighted-average voting ensemble; XGBoost, extreme gradient boosting.

TABLE 2
Model Evaluation on Monte Carlo Cross-Validation in the Testing Data Set (25% of Data)[a]

Model	AUC	Calibration Intercept	Calibration Slope	Brier Score[b]
RF	0.8555 (0.8549 to 0.8561)	–0.1482 (–0.1494 to –0.1470)	1.2969 (1.2952 to 1.2985)	0.1384 (0.1381 to 0.1387)
LightGBM	0.8920 (0.8915 to 0.8925)	0.0068 (0.005 to 0.008)	1.004 (1.002 to 1.006)	0.1080 (0.1074 to 0.1085)
Adaboost	0.8986 (0.8982 to 0.8991)	–0.0750 (–0.0769 to –0.0737)	1.2100 (1.2083to 1.2117)	0.0926 (0.0922 to 0.0929)
XGBoost	0.8868 (0.8863 to 0.8874)	–0.0007 (–0.0023 to –0.0008)	1.0221 (1.0195 to 1.0247)	0.1032 (0.1027 to 0.1037)
WAV ensemble	0.9139 (0.9135 to 0.9141)	–0.1806 (–0.1816 to –0.1795)	1.2794 (1.2778 to 1.2809)	0.0888 (0.0885 to 0.0891)

[a]Data are presented as mean (95% CI). Adaboost, adaptive boosting; AUC, area under the receiver operating characteristic curve; LightGBM, light gradient boosting machine; RF, random forest; WAV, weighted-average voting; XGBoost, extreme gradient boosting.
[b]Null model Brier score, 0.1367.

graft diameter, follow-up period, and time from injury to surgery (Figure 9B). Once these parameters were input, the program calculated a 75% probability that the patient would experience a clinical failure after >2 years postoperatively (Figure 9B). In detail, the probability of a random event is between 0 and 1, and the prediction model uses 0.5 as the discriminant boundary point. A probability of 75%, a tendency >0.5 of clinical failure based on our model, does not mean exact quantified failure risk. Attention should be paid to surgical techniques and additional postoperative rehabilitation interventions.

Unlike patient 1, patient 2 did not experience medial meniscus resection. No preoperative high-grade knee laxity was found. However, patient 2 participated in competitive sports (Figure 9C). Ranked by weight, the other factors supporting the clinical failure experienced by patient 2 were follow-up period, graft diameter, age at surgery, posterior tibial slope, and time from injury to surgery. These factors made little contribution to the result. The probability that patient 2 will experience clinical failure at >2 years postoperatively was calculated at 12% (Figure 9D).

DISCUSSION

The important findings of this study were that (1) the ensemble model validated the superior accuracy compared with the 4 single ML algorithms, showing the best predictive performance based on an AUC of 0.9139, calibration intercept of –0.1806, calibration slope of 1.2794, and Brier score of 0.0888; and (2) the ensemble algorithm was systematically developed as a web-based application to predict patients' risk of clinical failure of ACLR, which could allow

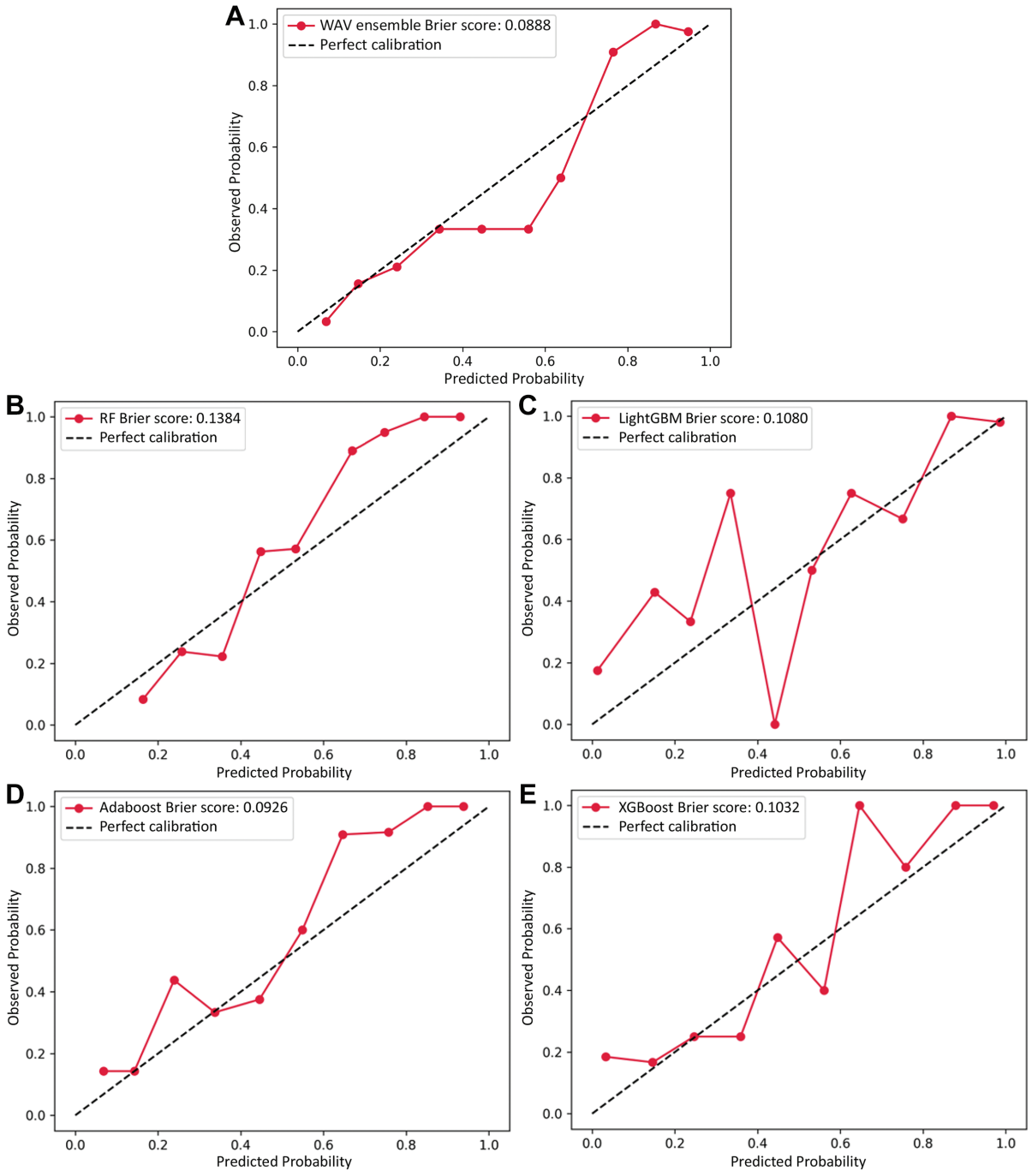

Figure 6. Calibration plots of the Brier scores for the models in the testing data set: (A) weighted-average voting (WAV) ensemble, (B) random forest (RF), (C) light gradient boosting machine (LightGBM), (D) adaptive boosting (Adaboost), and (E) extreme gradient boosting (XGBoost). The y-axis displays the true observed proportion of those who experienced the clinical failure, while the x-axis displays the corresponding predictions made by the ensemble model. The dashed line represents perfect prediction.

more convenient and accurate preoperative consultation. Compared with traditional statistical methods, the strength of ML is that it prioritizes repeatability and accurate predictions of models rather than just providing interpretability.[5,25] In some previous studies, the concordance of the single model was reported to be moderate, with a lower AUC ranging between 0.69 and 0.82.[25,35] The ensemble model has been proven to be an effective way of improving clinical outcome prediction accuracy over single algorithm models.[10,16,54] In this context, our ensemble ML algorithm model showed the greatest AUC of 0.9139 compared with all single models, indicating that ensemble

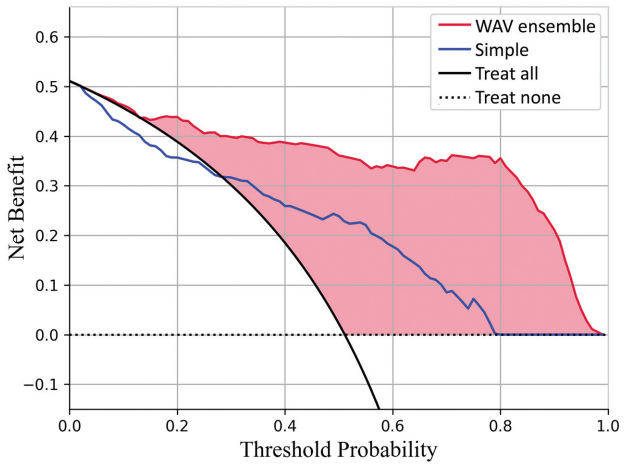

Figure 7. Decision curve analysis comparing the complete weighted-average voting (WAV) ensemble algorithm with a simple ensemble model using only 2 optimal predictors (follow-up period and high-grade preoperative knee laxity). The x-axis demonstrates risk thresholds for clinical failure, while the y-axis shows the standardized net benefit of changing management. The oblique black line labeled ''treat all'' plots the net benefit from the default scheme of changing management for all patients, while the dotted horizontal line labeled ''treat none'' represents the scheme of changing management for none of the patients (net benefit is zero at all thresholds). The further the curve of the decision model is from the baseline, the higher the net benefit of the decision is.

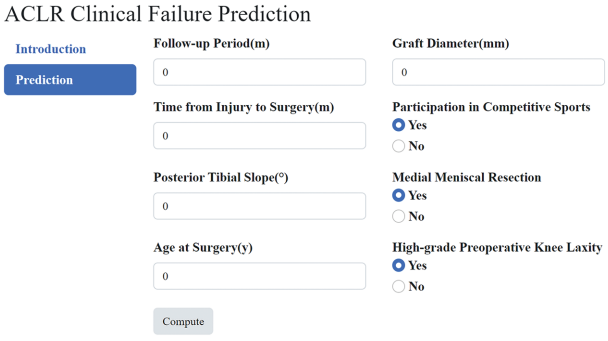

Figure 8. Prediction page on the web-based application (https://zorthoapps.shinyapps.io/aclr) for predicting clinical failure after anterior cruciate ligament reconstruction (ACLR). Users can input a patient's variables and click the Compute button to receive a risk assessment.

ML has better potential to predict clinical failure after ACLR from the individual patient level.

ACLR is a standard procedure to treat ACL deficiency, aiming to improve knee stability and restore normal biomechanics. However, the clinical failure of ACLR has been reported to result in limited postoperative function, swelling, and pain.[13,28,48] Persistent rotatory laxity or a graft rupture has been shown to correlate with inferior clinical scores and the subsequent need for revision surgery, and it can be especially devastating for young athletes.[15,57,58] All 8 variables were considered important for postoperative clinical failure of ACLR in our WAV ensemble model, which is consistent with previous studies.[12,21,32,38,41,49,50,58] Decreasing age significantly increased the risk of graft rupture or early graft revision after ACLR.[13,21,23,31,38,41,50,58] Preoperative high-grade knee laxity was associated with greater odds of ACLR graft rupture.[12,33,38] Greater tibial slope as a risk factor for graft rupture,[12] participation in competitive sports as a predictor of ipsilateral graft failure,[23] and the associations between decreased graft diameter with early graft revision all have been documented.[32] A previous study have indicated that time within 3 months from injury to surgery was associated with an increased risk of ACL revision.[49] Jacquet et al[21] concluded that the meniscus resection was predictive of residual pivot shift after a long follow-up. These studies explored risk factors for clinical failure rather than predicting an individual's risk of clinical failure after surgery. The factors affecting postoperative clinical failure varied among individual patients. Accurate prediction and quantification of the patient-specific risk were challenging given the complex relationships between these factors.[36] In general, the risk that a patient may experience adverse outcomes after ACLR is routinely assessed based on traditional statistical prediction models.[10,19,21,41] Traditional statistical method forecasting may not be able to achieve accurate quantitative prediction, which could be attributed to the inherent limitations of such models.[25,27] Developing a practical and convenient predictive tool for clinical failure of primary ACLR can be beneficial to optimize the surgical decision and postoperative rehabilitation plan.[25,30,36]

The most common type of ML is called "supervised learning." This approach consists of algorithms that analyze the relationship between predictors and outcomes.[27] Compared with traditional statistics, ML algorithms are data-driven, without an a priori hypothesis of the relationship between the available data and outcome. Conventional statistical techniques such as linear or logistic regression have a potential pitfall: The relationship between input variables and output is user-chosen and may result in a suboptimal (less accurate) prediction model. In reality, the relationship between input and output is nonlinear when a large amount of input variables are involved.[27] These algorithms could learn to complete feature selection and accurate prediction based on training and could then be optimized for better performance.[5,19] When an ML algorithm can highlight the correlation between the predictors and the outcome of interest; it can then be developed into a calculator capable of reliably predicting the outcome for a wider range of patients.[35] Recently, some ML models have been applied in orthopaedics for clinical outcome prediction, such as the clinically meaningful improvement after ACLR (AUC, 0.82),[25] dissatisfaction after primary total knee arthroplasty (AUC, 0.77),[26] ACLR revision (AUC, 0.69),[36] and outcome after surgery for degenerative cervical myelopathy (AUC, 0.70).[37] The AUC range of the single model in our study is between 0.8555 and 0.8968. These studies have shown

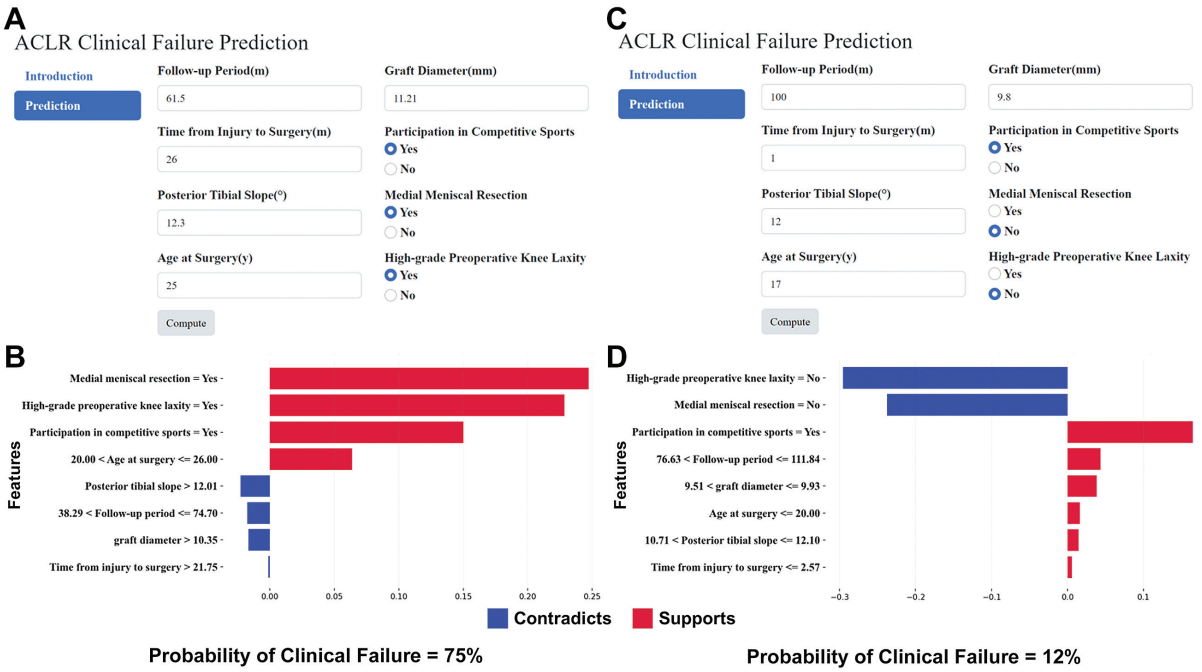

Figure 9. Two examples of individual patient-level explanations for ensemble algorithm predictions of clinical failure. (A and B) Prediction for patient 1. (C and D) Prediction for patient 2. The top row shows the inputted patient variables on the prediction page of the website. In the bottom row, the red bars indicate features that put the patient at risk of clinical failure, and the blue bars indicate features that contradict clinical failure. ACLR, anterior cruciate ligament reconstruction.

acceptable prediction abilities, despite different prediction targets, data sources, and candidate single models. It should be noted that although these studies ultimately chose appropriate single models for target prediction, they might be unsuitable for horizontal comparison. Ghasemieh et al[16] outlined several limitations of single-based ML, such as limited data set size, dependence on patient participation, unstable training, no real-world validation, lack of diversity in samples, and potential for bias. In addition, a single model is more likely to cause overfitting, and the ensemble model can improve the normalization ability, while the prediction ability of a single model is not high.[16] Ensemble-based techniques are introduced to overcome the limitations of single-based methods and obtain a robust prediction classification.[10,16] The ensemble model has been proven to be the most accurate way of improving the accuracy of prognosis and diagnosis over single classifier models.[10,16,22,30,54] Our ensemble ML algorithm model showed the greatest AUC of 0.9139 when compared with all single models.

Relevant studies widely used decision curve analysis to show the clinical value of their predictive models.[25,30] The net benefit of each model across different threshold probabilities was revealed directly by decision curve analysis. It also identified the range of threshold probabilities in which a model could bring benefit. It visually demonstrated that intraoperative decision-making and postoperative management changes based off our WAV ensemble model could bring greater net benefits (lower probability of clinical failure). Decision curve analysis combined the accuracy metrics and clinical applicability, which was a suitable method for evaluating postoperative clinical failure of ACLR risk prediction and perioperative period strategies.[55] In summary, we provided an ensemble ML practical prediction model for orthopaedic surgeons to determine the failure probability of a particular patient before surgery, considering the risk of postoperative failure, optimizing surgical decision-making, and guiding postoperative rehabilitation plan.

We incorporated the WAV ensemble model into an open-access web application using the LIME method. Current evidence has shown that LIME is a model-independent interpretability approach that can explain the predictions of any model.[46,51] LIME provided individual explanations for the behavior of the WAV ensemble model. It can help individuals comprehend the "black box" prediction process of clinical failure after ACLR by explaining a localized example.[11,44] Clinicians can understand the basis of prediction outcomes and analyze whether the prediction results are reliable using LIME.[44] In our website application, users are able to get visible predictions with accompanying explanations of clinical failure after ACLR based on LIME. The benefits of a web-based application for clinicians are clear: user-friendly visual interface, easy operation for clinicians, and convenient access. As far as patients are concerned, preoperative predictive evaluation can help them participate in making surgical decisions to improve patient satisfaction. Furthermore, the open-access network can continuously incorporate new data to train the model and improve its accuracy.[30,36] Subsequently, it could also

incorporate different ethnic data and further establish a global database, which can help physicians improve the postoperative prediction accuracy of ACLR as well as surgical decision-making ability and further develop a more scientific rehabilitation plan to achieve better functional rehabilitation in different populations worldwide.

Limitations

Some limitations in the current study need to be addressed. First, in order to ensure the consistency and standardization of the data set, we collected data of patients who underwent primary anatomic double-bundle ACLR at a single center, although the external generalizability of our study may decline. We look forward to predicting clinical outcomes based on widely used surgical techniques and multicentric clinical data in future research. Second, the study needed a larger sample size to reduce the risk of overfitting during the predictive model development. More patients should be included in our follow-up work. Furthermore, we considered 8 widely demonstrated features: age at surgery, time from injury to surgery, participation in competitive sports, graft diameter, medial meniscus resection, posterior tibial slope, high-grade preoperative knee laxity, and follow-up period. Other factors such as sex[20] and body mass index,[53] which were demonstrated to be associated with clinical failure or revision after ACLR, did not show significance in the current study. Notably, there were controversies over whether sex could affect the clinical outcomes after ACLR.[1,23,41] Ageberg et al[1] suggested that female patients showed worse clinical outcomes compared with male patients before and at 1 and 2 years after ACLR, which was statistically significant. No impact of sex on clinical failure outcomes after ACLR was found in the studies by Persson et al[41] and Kaeding et al.[23] We look forward to training the predictive model with more sex-balanced data in future research and further exploring this issue. Third, the rotatory laxity proportion in the clinical failure after ACLR was 78% (57 in 73 clinical failure). Some mechanisms of rotatory laxity were not explored in this study, although it is necessary to explain this status completely.[6] We look forward to comprehensively analyzing the mechanism of clinical failure after ACLR in our further research. Last, although the ensemble algorithm was well calibrated and the concordance excellent, the results of our study might not apply to other races in other countries because they represented data from only the Asian population. Continuous learning and rigorous external validation must be undertaken before widespread clinical implementation. However, this ensemble ML model provided potential for clinically useful predictions.

CONCLUSION

Our study determined that the WAV ensemble algorithm accurately predicted the patient-specific risk of clinical failure after ACLR based on preoperative, intraoperative, and postoperative factors, showing greater AUC and Brier scores compared with the other 4 models. Furthermore, our web-based application has the potential to help clinicians and patients understand the basis of prediction outcomes and individual explanations, which could provide more convenient and accurate preoperative consultation.

ORCID iDs

Tianlun Zhang https://orcid.org/0000-0002-0377-7099
Zipeng Ye https://orcid.org/0000-0002-6960-4887
Jiangyu Cai https://orcid.org/0000-0003-4819-1160
Jiebo Chen https://orcid.org/0000-0003-2778-4418
Junjie Xu https://orcid.org/0000-0001-9353-0331
Jinzhong Zhao https://orcid.org/0000-0003-2265-1878

REFERENCES

1. Ageberg E, Forssblad M, Herbertsson P, Roos EM. Sex differences in patient-reported outcomes after anterior cruciate ligament reconstruction: data from the Swedish knee ligament register. *Am J Sports Med*. 2010;38(7):1334-1342.
2. Ahlden M, Samuelsson K, Sernert N, et al. The Swedish National Anterior Cruciate Ligament Register: a report on baseline variables and outcomes of surgery for almost 18,000 patients. *Am J Sports Med*. 2012;40(10):2230-2235.
3. Andriacchi TP, Koo S, Scanlan SF. Gait mechanics influence healthy cartilage morphology and osteoarthritis of the knee. *J Bone Joint Surg Am*. 2009;91(suppl 1):95-101.
4. Ayeni OR, Chahal M, Tran MN, Sprague S. Pivot shift as an outcome measure for ACL reconstruction: a systematic review. *Knee Surg Sports Traumatol Arthrosc*. 2012;20(4):767-777.
5. Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *J Arthroplasty*. 2018;33(8):2358-2361.
6. Chen JL, Allen CR, Stephens TE, et al. Differences in mechanisms of failure, intraoperative findings, and surgical characteristics between single- and multiple-revision ACL reconstructions: a MARS cohort study. *Am J Sports Med*. 2013;41(7):1571-1578.
7. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
8. Delaloye JR, Murar J, Koch PP, Sonnery-Cottet B. Combined anterior cruciate ligament and anterolateral ligament lesions: from anatomy to clinical results. *Ann Joint*. 2018;3:82.
9. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019;19(1):211.
10. Dutta A, Hasan MK, Ahmad M, et al. Early prediction of diabetes using an ensemble of machine learning models. *Int J Environ Res Public Health*. 2022;19(19):12378.
11. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak*. 2019;19(1):146.
12. Firth AD, Bryant DM, Litchfield R, et al. Predictors of graft failure in young active patients undergoing hamstring autograft anterior cruciate ligament reconstruction with or without a lateral extra-articular tenodesis: the Stability Experience. *Am J Sports Med*. 2022;50(2):384-395.
13. Frobell RB, Roos HP, Roos EM, et al. Treatment for acute anterior cruciate ligament tear: five year outcome of randomised trial. Br J Sports Med. 2015;49(10):700.
14. Irrgang JJ, Ho H, Harner CD, Fu FH. Use of the International Knee Documentation Committee guidelines to assess outcome following

anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 1998(6):107-114.
15. Getgood AMJ, Bryant DM, Litchfield R, et al. Lateral extra-articular tenodesis reduces failure of hamstring tendon autograft anterior cruciate ligament reconstruction: 2-year outcomes from the STABILITY Study randomized clinical trial. *Am J Sports Med*. 2020;48(2):285-297.
16. Ghasemieh A, Lloyed A, Bahrami P, Vajar P, Kashef R. A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. *Decis Analyt J*. 2023;7:100242.
17. Grassi A, Pizza N, Al-Zu'bi BBH, et al. Clinical outcomes and osteoarthritis at very long-term follow-up after ACL reconstruction: a systematic review and meta-analysis. *Orthop J Sports Med*. 2022; 10(1):23259671211062238.
18. Han Y, Huang L, Zhou F. A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. *Bioinformatics*. 2021;37(15):2183-2189.
19. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med*. 2020;13(1):69-76.
20. Hewett TE, Myer GD, Ford KR, Paterno MV, Quatman CE. Mechanisms, prediction, and prevention of ACL injuries: cut risk with three sharpened and validated tools. *J Orthop Res*. 2016;34(11):1843-1855.
21. Jacquet C, Pioger C, Seil R, et al. Incidence and risk factors for residual high-grade pivot shift after ACL reconstruction with or without a lateral extra-articular tenodesis. *Orthop J Sports Med*. 2021;9(5): 23259671211003590.
22. Jonnalagadda A, Rajvir M, Singh S, et al. An ensemble-based machine learning model for emotion and mental health detection. *J Inf Knowl Manag*. 2022;22(02).
23. Kaeding CC, Pedroza AD, Reinke EK, et al. Risk factors and predictors of subsequent ACL injury in either knee after ACL reconstruction: prospective analysis of 2488 primary ACL reconstructions from the MOON cohort. *Am J Sports Med*. 2015;43(7):1583-1590.
24. Kamath GV, Murphy T, Creighton RA, et al. Anterior cruciate ligament injury, return to play, and reinjury in the elite collegiate athlete: analysis of an NCAA Division I cohort. *Am J Sports Med*. 2014;42(7): 1638-1643.
25. Kunze KN, Polce EM, Ranawat AS, et al. Application of machine learning algorithms to predict clinically meaningful improvement after arthroscopic anterior cruciate ligament reconstruction. *Orthop J Sports Med*. 2021;9(10):23259671211046575.
26. Kunze KN, Polce EM, Sadauskas AJ, Levine BR. Development of machine learning algorithms to predict patient dissatisfaction after primary total knee arthroplasty. *J Arthroplasty*. 2020;35(11):3117-3122.
27. Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(3):753-757.
28. Lind M, Menhert F, Pedersen AB. Incidence and outcome after revision anterior cruciate ligament reconstruction: results from the Danish registry for knee ligament reconstructions. *Am J Sports Med*. 2012;40(7):1551-1557.
29. Liu S, Li H, Tao H, et al. A randomized clinical trial to evaluate attached hamstring anterior cruciate ligament graft maturity with magnetic resonance imaging. *Am J Sports Med*. 2018;46(5):1143-1149.
30. Lu Y, Forlenza E, Cohn MR, et al. Machine learning can reliably identify patients at risk of overnight hospital admission following anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 2021;29(9):2958-2966.
31. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323.
32. Magnussen RA, Lawrence JT, West RL, et al. Graft size and patient age are predictors of early revision after anterior cruciate ligament reconstruction with hamstring autograft. *Arthroscopy*. 2012;28(4): 526-531.
33. Magnussen RA, Reinke EK, Huston LJ, et al. Effect of high-grade preoperative knee laxity on 6-year anterior cruciate ligament reconstruction outcomes. *Am J Sports Med*. 2018;46(12):2865-2872.
34. Maletis GB, Inacio MC, Funahashi TT. Risk factors associated with revision and contralateral anterior cruciate ligament reconstructions in the Kaiser Permanente ACLR registry. *Am J Sports Med*. 2015;43(3):641-647.
35. Martin RK, Wastvedt S, Pareek A, et al. Machine learning algorithm to predict anterior cruciate ligament revision demonstrates external validity. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(2):368-375.
36. Martin RK, Wastvedt S, Pareek A, et al. Predicting anterior cruciate ligament reconstruction revision: a machine learning analysis utilizing the Norwegian Knee Ligament Register. *J Bone Joint Surg Am*. 2022;104(2):145-153.
37. Merali ZG, Witiw CD, Badhiwala JH, Wilson JR, Fehlings MG. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One*. 2019;14(4):e0215133.
38. MOON Knee Group; Spindler KP, Huston LJ, et al. Anterior cruciate ligament reconstruction in high school and college-aged athletes: does autograft choice influence anterior cruciate ligament revision rates? *Am J Sports Med*. 2020;48(2):298-309.
39. Myers TG, Ramkumar PN, Ricciardi BF, et al. Artificial intelligence and orthopaedics: an introduction for clinicians. *J Bone Joint Surg Am*. 2020;102(9):830-840.
40. Osamor VC, Okezie AF. Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis. *Sci Rep*. 2021;11(1): 14806.
41. Persson A, Fjeldsgaard K, Gjertsen JE, et al. Increased risk of revision with hamstring tendon grafts compared with patellar tendon grafts after anterior cruciate ligament reconstruction: a study of 12,643 patients from the Norwegian Cruciate Ligament Registry, 2004-2012. *Am J Sports Med*. 2014;42(2):285-291.
42. Qiu H, Luo L, Su Z, et al. Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. *BMC Med Inform Decis Mak*. 2020;20(1):83.
43. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649.
44. Robnik-Šikonja M, Štrumbelj E, Kononenko I. Efficiently explaining the predictions of a probabilistic radial basis function classification network. *Intell Data Anal*. 2013;17(5):791-802.
45. Saita Y, Schoenhuber H, Thiebat G, et al. Knee hyperextension and a small lateral condyle are associated with greater quantified antero-lateral rotatory instability in the patients with a complete anterior cruciate ligament (ACL) rupture. *Knee Surg Sports Traumatol Arthrosc*. 2019;27(3):868-874.
46. Salami D, Sousa CA, Martins M, Capinha C. Predicting dengue importation into Europe, using machine learning and model-agnostic methods. *Sci Rep*. 2020;10(1):9689.
47. Salmon LJ, Heath E, Akrawi H, et al. 20-year outcomes of anterior cruciate ligament reconstruction with hamstring tendon autograft: the catastrophic effect of age and posterior tibial slope. *Am J Sports Med*. 2018;46(3):531-543.
48. Schlumberger M, Schuster P, Schulz M, et al. Traumatic graft rupture after primary and revision anterior cruciate ligament reconstruction: retrospective analysis of incidence and risk factors in 2915 cases. *Knee Surg Sports Traumatol Arthrosc*. 2017;25(5):1535-1541.
49. Snaebjornsson T, Hamrin Senorski E, Svantesson E, et al. Graft fixation and timing of surgery are predictors of early anterior cruciate ligament revision: a cohort study from the Swedish and Norwegian knee ligament registries based on 18,425 patients. *JB JS Open Access*. 2019;4(4):e0037.
50. Snaebjornsson T, Svantesson E, Sundemo D, et al. Young age and high BMI are predictors of early revision surgery after primary anterior cruciate ligament reconstruction: a cohort study from the Swedish and Norwegian knee ligament registries based on 30,747 patients. *Knee Surg Sports Traumatol Arthrosc*. 2019;27(11):3583-3591.
51. Speiser JL, Callahan KE, Houston DK, et al. Machine learning in aging: an example of developing prediction models for serious fall

injury in older adults. *J Gerontol A Biol Sci Med Sci*. 2021;76(4):647-654.
52. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.
53. van Eck CF, Schkrohowsky JG, Working ZM, Irrgang JJ, Fu FH. Prospective analysis of failure rate and predictors of failure after anatomic anterior cruciate ligament reconstruction with allograft. *Am J Sports Med*. 2012;40(4):800-807.
54. Velusamy D, Ramasamy K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Comput Methods Programs Biomed*. 2021;198:105770.
55. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2016;26(6):565-574.
56. Webb JM, Salmon LJ, Leclerc E, Pinczewski LA, Roe JP. Posterior tibial slope and further anterior cruciate ligament injuries in the anterior cruciate ligament-reconstructed patient. *Am J Sports Med*. 2013;41(12):2800-2804.
57. Webster KE, Feller JA, Klemm HJ. Second ACL injury rates in younger athletes who were advised to delay return to sport until 12 months after ACL reconstruction. *Orthop J Sports Med*. 2021;9(2):2325967120985636.
58. Webster KE, Feller JA, Leigh WB, Richmond AK. Younger patients are at increased risk for graft rupture and contralateral injury after anterior cruciate ligament reconstruction. *Am J Sports Med*. 2014;42(3):641-647.