Using Decision Tree Classifier to Increase Screening Test Sensitivity for the Prediction of ACL Retear

Tanishk Govil
The University of Virginia
Charlottesville, Virginia, USA
ygf3yv@virginia.edu

Tarek Hamid
The University of Virginia
Charlottesville, Virginia, USA
pve8nt@virginia.edu

Kimberly Helm
The University of Pennsylvania
Philadelphia, Pennsylvania, USA
kimhelm@seas.upenn.edu

Elliot Greenberg
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, USA
greenberge@chop.edu

Kevin Landrum
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, USA
landrumk1@chop.edu

J. Todd R. Lawrence
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, USA
lawrencej@chop.edu

Theodore J. Ganley
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, USA
ganley@chop.edu

Amanda Watson
The University of Virginia
Charlottesville, Virginia, USA
aawatson@virginia.edu

Figure 1: Measured physical therapy exercises included as features in the dataset. Left: Biodex Testing. Middle: Squat and Jump Testing. Right: Static Hold Testing.

ABSTRACT

Screening tests are often used in medicine to assess whether a patient is at a high risk of contracting a disease. Recent literature has proposed prediction algorithms for Anterior Cruciate Ligament (ACL) retears that aim to achieve high accuracy. However, these models fail to reach an adequate sensitivity to function as effective screening tests. In such cases, model sensitivity is sacrificed for heightened specificity. Misclassifying patients who will eventually go on to retear their ACL as low-risk patients prevents them from obtaining necessary therapeutic support and is not appropriate for a clinical setting. In this study, we implement a Decision Tree Classifier as a screening test to evaluate a patient's risk of retearing their ACL six months after surgery, before the patient is released to activity. By incorporating a machine learning-based screening technique, we hope to minimize false negatives and create a tool that can readily be adopted in clinical practice.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Human-centered computing** → **Ubiquitous and mobile computing**; •

UbiComp Companion '24, October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1058-2/24/10
https://doi.org/10.1145/3675094.3680116

Computing methodologies → Machine learning; **Feature selection**; **Supervised learning**; **Cross-validation**; **Classification and regression trees**.

KEYWORDS

Anterior Cruciate Ligament, Machine Learning, Sensitivity, Specificity, Anterior Cruciate Ligament Reinjury, Retear, Risk Factors

ACM Reference Format:
Tanishk Govil, Tarek Hamid, Kimberly Helm, Elliot Greenberg, Kevin Landrum, J. Todd R. Lawrence, Theodore J. Ganley, and Amanda Watson. 2024. Using Decision Tree Classifier to Increase Screening Test Sensitivity for the Prediction of ACL Retear. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3675094.3680116

1 INTRODUCTION

Screening tests have long been used in medicine to assess a patient's likelihood of diagnosis [3]. To create reliable predictions, screening tests must have high sensitivity. In other words, they must correctly classify almost all patients who will go on to retear as high-risk for reinjury. In recent years, machine learning has gained popularity in clinical applications to diagnose, treat, and monitor illness [5]. Machine learning algorithms can be more effective than traditional prediction methods by using larger data sets to create more robust clinical models. By implementing a machine learning-based screening test, we aim to increase prediction accuracy and offer tailored rehabilitation recommendations.

In this paper, we seek to answer the following research questions:

RQ1: How can we improve on the sensitivity of previous machine learning implementations for predicting ACL retear?

RQ2: How does our approach compare to previous retear prediction systems?

We implemented a screening test using the Decision Tree Classifier machine learning algorithm to predict ACL retear with 88.0 percent sensitivity and 77.9 percent specificity. Further, we evaluated our system using five-fold cross-validation and performed a comparison against similar implementations. Specifically, our contributions are summarized as follows:

(1) A machine learning model that identifies patients at high risk of retear with a high degree of sensitivity and specificity.
(2) An evaluation of our system and a comparison against similar implementations.

The remainder of our paper is structured as follows: Section 2 will describe our machine learning algorithm and personalized recommendation system in detail. In Section 3 we evaluate our implementation. Finally, in Section 4 we conclude our findings and describe the direction of future work.

2 METHODS

This section describes the methods used for cleaning and imputing data, as well as model selection. First, we will describe the features and characteristics of our data set. Next, we will explain the and the methodology for selecting the optimal imputation method for missing data. Finally, we will summarize the model selection process and optimal algorithm.

2.1 Data Set

The original data set was composed of 1063 patients who tore their ACL from 2009 to 2020 and were between the ages of 8 and 21 years old. For each patient, fifty-six features were analyzed, including age at surgery, delay to surgery, type of graft, sex, body mass index, relatives with ACL tears, and a range of surgical and physical therapy data. Data was further split into categories of demographic, injury, surgical, recovery, and rehab information six months after surgery. Figure 1 displays the measurement of some of the rehabilitation data that were included as features in our final model. These data include Biodex testing, which measures a patient's isokinetic strength by collecting data such as angular velocity and generated torque[6]. These features were collected during in-clinic assessments and post-surgical follow-ups; wearable devices such as EMG or inertial sensing devices that capture physical activity and markers of physical performance can also calculate these features, enabling at-home, continuous monitoring of potential retears. Patients who had less than 50% of data on record were excluded from the data set. The remaining 591 patients included in the analysis consisted of 305 males and 286 female athletes, and 112 went on to retear their ACL (18.95%). In order to evaluate model proficiency, 10% of the data was removed to be used as a holdout set. To separate the holdout set, we stratified by whether or not the patient went on to retear their ACL. This was to ensure that the holdout set had a representative number of retears as the training set. The remaining data was split into five folds (20% of the data each) for cross-validation. To split the data into five folds, we stratified by age, gender, and whether or not the patient went on to retear. This was to ensure representative demographics in each fold. Figure 2 shows the process flow for data reduction techniques.

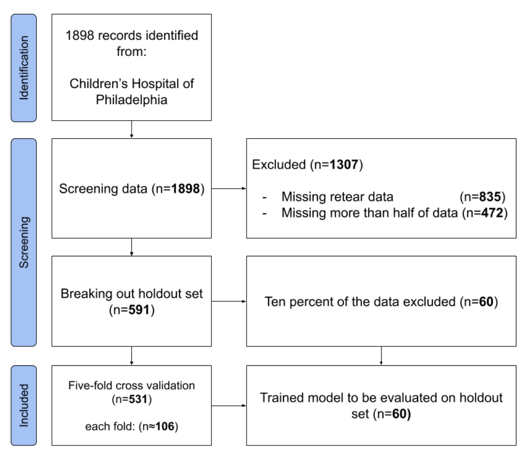

Figure 2: Process Flow of Data Reduction

2.2 Imputation Methods

When deciding whether to impute missing values, a trade-off exists between the risk of overfitting due to imputation or excluding patients with an incomplete data set and evaluating too few data points. Imputation may lead to overfitting as a result of a portion of the data being a derivative of collected attributes, effectively increasing the weight of a smaller subset of data in the final model.

Similarly, using too few data points can lead to overfitting by using a skewed subset of data that isn't representative of the tested population. Of the 591 patients included in the final analysis, just 4 had a complete data set (0.68%) and 12,585 of 33,687 values were missing (37.4%). In previous work by Watson, et al.[7], missing data points were ignored, potentially leading to patients who went on to retear being incorrectly classified as low-risk. We aim to overcome this challenge by imputing missing values. Multiple imputation methods were implemented and compared, including replacing missing data with an integer (9999); the mean, median, and mode of each feature; and a range of imputation regressors including Bayesian Ridge, Decision Tree, Extra Tree, K Neighbors, and XG Boost [1]. Each imputation method was evaluated on the sensitivity, or true positive rate, of the machine learning algorithm utilized.

2.3 Exploration of Models Tested

Similarly to finding the appropriate imputation method, several machine learning algorithms were compared: Naive Bayes, Adaboost, Support Vector Machines, Decision Tree Classifier, and Random Forest [2]. The algorithms implemented range from using conditional probability in a multi-dimensional sample space via Naive Bayes to resampling algorithms such as Random Forests and Adaboost. All five algorithms were carried out with each of the nine imputation techniques, for forty-five total models. The final analysis was implemented using the imputation technique and machine learning algorithm with the lowest false positive rate. Ultimately, the Decision Tree Classifier was implemented with Decision Tree Regressor imputation, achieving a sensitivity of 88.0% and a specificity of 77.9% in the cross-validation. Figure 3 shows the process flow for testing and evaluation.

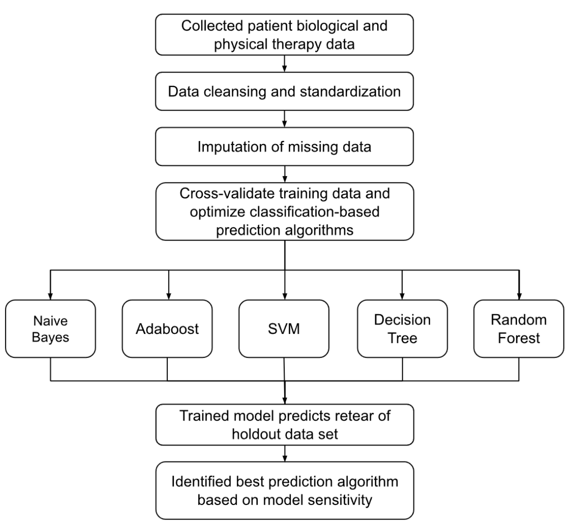

Figure 3: Process Flow of Model Selection

2.4 Addressing Class Imbalance

Less than 20% of the patients in the dataset would go on to retear their ACL, making the data heavily imbalanced. In the current state, the low-risk predictions would significantly outweigh the high-risk predictions, creating a model sacrificing sensitivity for heightened specificity. We evaluated multiple methods for addressing this imbalance, including Synthetic Minority Over-sampling Technique (SMOTE) and the assignment of class weights. Ultimately, we chose to assign class weights rather than execute SMOTE algorithms because we achieved superior results using the existing data rather than generating new samples.

3 EVALUATION

In this section, we'll explain the algorithm selection process, the chosen model, and the model's performance. First, we'll compare the tested models and imputation methods based on their achieved sensitivity. Then, we'll describe why the algorithm selected is the preferred machine-learning model for this study.

3.1 Model Evaluation

In order to evaluate the Decision Tree Classifier as a prediction mechanism, we calculated the sensitivity and specificity of the model, or the degree to which the model can correctly categorize true positives and true negatives respectively.

$$sensitivity = \frac{TP}{TP + FN} \quad (1) \qquad specificity = \frac{TN}{TN + FP} \quad (2)$$

In the evaluation of the Decision Tree Classifier as a screening test, we prioritize the sensitivity: its ability to categorize high-risk patients correctly. Table 1 lists the sensitivity and specificity statistics of each tested machine learning algorithm during cross-validation, using the imputation method that yielded the highest sensitivity. To address class imbalance, we used manual tuning to assign class weights. We found that to yield the highest sensitivity, the retear class should have a weight of around 0.90 for all the models. We prioritize sensitivity over specificity because, from a clinical standpoint, it is crucial to classify everyone who will go on to retear their ACL as high-risk patients. For a screening test, we want to ensure that everyone who will go on to retear their ACL is flagged as a high-risk patient. In this case, false negatives are much more important to mitigate than false positives, making sensitivity a critical measure of our model.

Table 1: Sensitivity and Specificity for Tested ML Models

Model	Imputation	Sensitivity	Specificity
Decision Tree	Decision Tree Reg	0.880	0.779
SVM	XG Boost Reg	0.821	0.551
Naive Bayes	Extra Tree Reg	0.790	0.482
Adaboost	Decision Tree Reg	0.758	0.882
Adaboost	Extra Tree Reg	0.701	0.821

The Decision Tree implementation yields a sensitivity of 88.0% on the cross-validation and holds across all five folds. When tested on the holdout set, the sensitivity is 90.9%, correctly classifying 10 of 11 true positives. The high sensitivities across all testing sets make the Decision Tree Classifier a useful algorithm for screening. The specificity of the Decision Tree Classifier implementation averages 77.9% on the five-fold cross-validation and 67.3% on the holdout set. Although these are lower than the sensitivities, this is not a deterrent to implementation as a screening test. Screening tests aim to identify as many subjects that will retear as possible. Therefore, tests with high sensitivity tend to be effective for screening, as they rarely produce false negatives [3]. We will focus on using hybrid models in the future to balance sensitivity and specificity better. Table 2 lists the sensitivity and specificity for all five folds during the cross-validation and the holdout. The mean sensitivity during the cross-validation is 0.880, with a standard deviation of 0.0865. These results show that the model holds across all five folds.

Table 2: Sensitivity and Specificity for Each Fold

Fold	Sensitivity	Specificity
One	0.905	0.791
Two	0.933	0.835
Three	0.737	0.885
Four	0.870	0.723
Five	0.957	0.663
Holdout	0.909	0.673

3.2 Decision Tree Feature Importance

Our decision tree classifier model is trained on surgical, demographic, injury, and rehabilitation data. This model is to be used as a screening test before athletes return to activity, around six months after surgery. Therefore, to train our model, we only selected the data that was available six months after surgery, ignoring any rehabilitation data recorded later. Using manual tuning, we found that a weight of 0.885 for the retear class produced the highest sensitivity while maintaining a high specificity. The decision tree's five most important features, calculated by Gini Importance, were as follows:

(1) Involved limb vertical hop distance
(2) Delay to surgery
(3) Involved limb triple hop distance normalized to body height
(4) Uninvolved limb single leg hop
(5) Any relative with ACL tear

We find that in general, patient rehab and recovery data are the most predictive of their risk of retear. More specifically, Biodex and hop test data are the most important features. Paterno et al.[4] confirm our findings. Their Classification and Regression Tree (CART) analysis recognizes triple hop distance normalized to body height as one of the primary predictors of ACL retears, and our model does as well. However, our model places a large emphasis on other types of hop tests and Biodex isokinetic strength testing, whereas the CART model from Paterno et al.[4] values demographic information such as age and sex. Our model has more potential benefits to the athletes since strength and hop tests can be improved through training, whereas age and sex, are much more difficult to modify. This will help high-risk patients minimize their chance of retearing their ACL.

3.3 Comparison to Similar Implementations

Recently, Watson, et al.[7] and Paterno et al.[4] have independently implemented machine learning algorithms using demographic, biological, and physical therapy data to predict the occurrence of ACL retear. Testing data showed that both models categorized low-risk patients with high accuracy; however, our Decision Tree Classifier implementation significantly outperformed in terms of sensitivity, while still maintaining high specificity. In their analysis, Watson et al. implemented a clinician-guided machine learning algorithm. After collecting patient data, their model used clinician feedback to determine optimal ranges and weights for each feature. They formulated a weighting function to predict whether each patient should be classified as high or low risk of retear. Testing their algorithm on an unseen set of data produced a sensitivity of 40.0% and a specificity of 100%. In other words, their algorithm correctly classified 40% of patients that retore and 100% of patients who didn't. Paterno et al. conducted a similar study using CART analysis to determine feature importance and create a prediction model. Their methodology excluded patients with missing data, so imputation was unnecessary. Their prediction model correctly classified 66.7% of patients who retore and 72.0% of patients who avoided reinjury. Table 3 compares the sensitivity and specificity of each implementation compared to our model.

Table 3: Comparison to Similar Implementations

Study	Algorithm	Sensitivity	Specificity
Current	Decision Tree Classifier	0.880	0.779
[4]	Decision Tree	0.667	0.720
[7]	Clinician-Guided Alg.	0.40	1.00

Our model outperforms both models overall, as it has the highest sensitivity and maintains a high specificity. Note that although our model's sensitivity is far greater than the other models, which is the most important metric for screening tests, it was still important to maintain a high specificity. From a clinical standpoint, low specificity would cause unnecessary intervention and increased anxiety among ACL tear patients. Nevertheless, the 88.0% sensitivity and 77.9% specificity show that our model is a significantly more effective screening test than existing implementations.

4 CONCLUSION AND FUTURE WORK

Previous prediction and recommendation systems for ACL reinjury have focused on overall model accuracy. Our model focuses on optimizing model sensitivity, or minimizing the false negative rate. This model can be used as a screening test to minimize the number of patients who are incorrectly classified as low-risk for retear and subsequently provided less therapeutic support. By implementing a Decision Tree Classifier on a data set containing features with low correlation, our system has a sensitivity of 88.0% and a specificity of 77.9%, significantly outperforming similar state-of-the-art metrics. As a future direction of our work, we plan to implement a recommendation system that is tailored to each patient. This will allow clinicians to quickly understand why someone is classified as a high-risk patient, and how the recovery plan can be modified to minimize this risk. By proposing an algorithm with increased accuracy and sensitivity, clinical resources may be more efficiently allocated. More robust therapeutic strategies could be used to support patients who would have otherwise experienced reinjury.

REFERENCES

[1] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 53 (2020), 1487–1509.

[2] Batta Mahesh. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]* 9 (2020), 381–386.

[3] L Daniel Maxim, Ron Niebo, and Mark J Utell. 2014. Screening tests: a review with examples. *Inhalation toxicology* 26, 13 (2014), 811–828.

[4] Mark V. Paterno, Bin Huang, Staci Thomas, Timothy E. Hewett, and Laura C. Schmitt. 2017. Clinical Factors That Predict a Second ACL Injury After ACL Reconstruction and Return to Sport. *The Orthopaedic Journal of Sports Medicine,* 5, 12 (2017).

[5] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine* 380, 14 (2019), 1347–1358.

[6] Nigel AS Taylor, Ross H Sanders, E Ian Howick, and Stephen N Stanley. 1991. Static and dynamic assessment of the Biodex dynamometer. *European journal of applied physiology and occupational physiology* 62 (1991), 180–188.

[7] Amanda Watson, Pengyuan Lu, Elliot Greenberg, J. Todd R. Lawrence, Theodore J. Ganley, and Insup Lee. 2021. RT-ACL: Identification of High-Risk Youth Patients and their Most Significant Risk Factors to Reduce Anterior Cruciate Ligament Reinjury Risk. *ACM/IEEE Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (2021).