2025 IEEE 13th International Conference on Healthcare Informatics (ICHI) | 979-8-3315-2094-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICHI64645.2025.00009

Comparative Machine Learning Analysis Highlights Novel Predictive Capability of Deep Neural Decision Forest for Ulnar Collateral Ligament Reconstruction in Baseball Athletes

Rohan Raj Butani
The Overlake School
Redmond, United States
rohanbutani81@gmail.com

Kedar Chintalapati
The Overlake School
Redmond, United States

Anirudh Sridharan
The Overlake School
Redmond, United States

Theodore Ioan Oltean
The Overlake School
Redmond, United States

Arjunn Shastri
Redmond High School
Redmond, United States

***Abstract*—Ulnar collateral ligament (UCL) reconstruction, commonly known as Tommy John surgery, has become increasingly prevalent among Major League Baseball (MLB) pitchers. We hypothesized that advanced machine learning (ML) techniques combining deep learning with ensemble decision trees could accurately predict the likelihood of UCL reconstruction using performance metrics from public databases. To test this, we employed a matched-pair design, selecting MLB pitchers with a minimum of 1500 career pitches and compiling diverse player profiles from Baseball Savant and Baseball Reference. We developed and compared nine ML models, including logistic regression, K-nearest neighbors, support vector machine, decision tree, random forest, XGBoost, artificial neural network (ANN), Deep Neural Decision Tree (DNDT), and Deep Neural Decision Forest (DNDF). Our findings indicate that while traditional models like logistic regression and K-nearest neighbors achieved accuracies around 62.5% and random forest reached 71.9%, the DNDF model significantly outperformed others with an accuracy of 79.2%. Feature importance analysis, using Shapley values, revealed that total pitch count was the most critical predictor, emphasizing its biomechanical impact on elbow stress. These results suggest that the DNDF model might effectively identify predictive patterns in low-dimensional tabular data, outperforming other ML approaches in UCL injury risk prediction. Future research may focus on integrating additional performance metrics, including pre-MLB pitch counts and real-time sensor data, to enhance prediction accuracy further and promote equitable healthcare interventions for athletes.**

***Index Terms*—Baseball, Machine Learning, AI, Tommy John, UCL, Healthcare Equity**

I. Introduction

Among baseball players, injuries to the ulnar collateral ligament have become increasingly prevalent due to the biomechanical stresses involved. Tommy John surgery, named after the pitcher who first underwent the procedure in 1974, involves reconstructing the UCL, a critical ligament for stabilizing the arm during throwing motions. Over recent decades, the incidence of Tommy John surgery among baseball players has risen significantly. In Major League Baseball (MLB), the number of surgeries increased from approximately 20 per year in 2010 to over 30 per year by 2020 (Conte et al. 2016), reflecting growing strain on pitchers' arms due to higher throwing velocities and spin rates.

In 2018, the prevalence of UCL reconstruction in major and minor league baseball combined was around 20%, representing a substantial proportion of professional baseball players (Leland, D.P. et al. 2019). Understanding the epidemiology of UCL injuries is crucial for developing targeted injury prevention and management strategies. Machine learning has emerged as a powerful tool in sports analytics, offering insights into injury risk assessment and player performance optimization. By analyzing large datasets encompassing player biometrics, performance metrics, and injury histories, researchers can develop predictive models that better inform injury prevention protocols (Claudino et al. 2021). Applications of predictive analyses of UCL reconstruction present unique benefits that are not only related to the monetary advantages of not recovering from surgery for an extended period of time but also due to key limitations of UCL reconstruction and consequent rehabilitation. Over 40% of pitchers who underwent UCL reconstruction were unable to fully return to peak performance in a case-control study conducted by Fury et al. in 2021.

This paper utilizes publicly available Major League Baseball (MLB) metric databases (Baseball Savant and Baseball Reference), incorporating player performance data and advanced analytics to train and test multiple machine learning-driven approaches to predict UCL reconstruction. This paper explores nine separate ML approaches, ranging from standard models to more advanced techniques. The specific models used, along with a brief overview of the characteristics that motivated inclusion of this model, are found in the Methodology: Model

2575-2634/25/$31.00 ©2025 IEEE
DOI 10.1109/ICHI64645.2025.00009

Descriptions section.

This model comparison is the first to compare deep learning and machine learning models to assess UCL reconstruction. The robust performances of our ANN, DNDT, and DNDF models may refute the notion that deep learning is not feasible for the binary classification of UCL reconstruction with low-dimensional tabular data.

II. Related Works

Prior literature reflects past efforts, albeit relatively few, to accurately deploy binary classification to predict UCL reconstruction in MLB players.

Most similarly, SA Rendar tested a myriad of machine-learning models to predict future UCL reconstruction for rookie pitchers specifically trained with sampled career data of MLB pitchers and tested with rookie pitcher data. Rendar found peak accuracy using an XGBoost algorithm with a train/test split of 80/20 and an accuracy of around 56%. However, the cumulative collection of features such as pitches, etc., presents potential issues in data viability due to the effects of UCL injury on pitching mechanics.

Machine learning has also been used to analyze and determine key predictors of UCL reconstruction. Whiteside et al. employed Naive Bayes and SVM classifiers in a 2016 report that were able to obtain accuracies of 72% and 75%, respectively. Our paper hopes to expand upon these results through the inclusion of an additional 4 models and varied parameter selection. Chalmers et al. also used information from public databases to assess key predictors for UCL reconstruction, and they found in their 2016 report that fastball velocity was the strongest predictor. Key features we would like to expand upon are the separation of pitch velocity and spin rate (both used by Whiteside et al.) by their individual pitch types as defined by StatCast. Furthermore, our study looks to analyze the differential effect on model accuracy through training with cumulative career data as opposed to by year, the prevalent method in Whiteside et al.'s model.

Outside of UCL construction specifically, ML has been used in general baseball injury prediction. Karnuta et al. used 6 different ML models in a comparative analysis versus regression modeling to predict future injury in MLB players.

Past literature provides a strong, albeit small, framework for the future of injury prevention and modeling in MLB athletes and baseball in general.

III. Biomechanics

Pitching a baseball requires the rapid generation of force from contractile tissues and involves muscles throughout the entire body. This motion engages muscles in the legs, hips, torso, and arms, all of which contribute to the necessary acceleration of the arm (Werner et al., 2008). This whole-body coordination is vital for effective pitch delivery, but it also subjects various joints, particularly the elbow and shoulder, to high levels of stress.

Pitchers face unique challenges due to the repetitive biomechanical stresses involved in throwing. High-speed pitches exert tremendous force on the arm, increasing susceptibility to injuries such as UCL (Ulnar Collateral Ligament) tears. Seroyer et al. (2010) found that "During acceleration, the elbow initially flexes from 90° to 120°, then rapidly extends to near 25° just before ball release, resulting in immense stretching of the UCL." These injuries not only impact individual player careers but also pose significant financial and performance-related risks to teams (Ahmad et al., 2020). The precise throwing motion of a baseball pitcher has been under the intense focus of scientific literature for several decades, with particular interest in understanding how specific mechanics contribute to injury.

Biomechanical analyses of a baseball pitch categorize the throwing motion into a sequence of "phases" (Christoffer et al., 2019), namely wind-up, stride, cocking, acceleration, deceleration, and follow-through (Chalmers et al., 2017). Each phase requires a different combination of muscular contractions and force generation. For example, the cocking phase, in which the arm is drawn backward, places substantial torsional stress on the shoulder and elbow, while the acceleration phase, which culminates in ball release, subjects the UCL to maximum strain. These distinct phases also highlight the different risk factors associated with each part of the motion.

The throwing motion present in baseball involves a combination of forces and torques that act on the elbow and shoulder joints of the athlete (Trasolini et al., 2022). During a pitch, the UCL experiences a valgus stress that pushes the forearm away from the body, countered by muscular forces that attempt to stabilize the joint. If the force exceeds the UCL's tensile strength, microtears can develop, eventually leading to a rupture. Scientific literature has established a clear correlation between elbow torque and injury risk, indicating that higher torque values are associated with a greater likelihood of UCL injury (Khalil et al., 2021). This principle connection forms the foundation of this paper's feature selection process.

Understanding the biomechanical component of pitching is critical in developing injury prevention and rehabilitation strategies. For example, altering the mechanics of a pitch to reduce torque on the elbow or emphasizing strength training in the surrounding musculature can help mitigate injury risks. Mitigation efforts may focus on limiting elbow torque through adaptations in training and pitching mechanics, reflecting principles similar to those explored in this model. Additionally, rehabilitation protocols often focus on reestablishing muscular balance and optimizing mechanics to reduce further stress on the joint, helping athletes return to play more effectively. By analyzing these biomechanical factors, the study aims to assist in the creation of models that can predict injury risk and inform the design of personalized training and recovery programs for athletes.

IV. Methodology

A. Model Descriptions

Below are brief descriptions of the learning models used in this paper. All models used were not pre-trained and trained from scratch:

1) Logistic Regression: Logistic Regression is a statistical method used for binary classification problems that models the probability that a given input belongs to a particular class by applying the logistic function to a linear combination of input features.

The hyperparameters tuned for logistic regression were C, a parameter that determines the strength of regularization, and the type of regularization (LASSO vs Ridge).

2) K-Nearest Neighbors (K-NN): K-Nearest Neighbors is a simple, non-parametric, and instance-based learning algorithm used for classification and regression tasks. It operates by identifying the 'k' closest data points in the feature space to a new input sample based on a chosen distance metric like Euclidean distance.

The hyperparameter tuned for K-NN was n-neighbors, which controlled how many neighboring data points were considered when making predictions.

3) Support Vector Machine (SVM): A Support Vector Machine is a supervised learning algorithm primarily used for classification tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a high dimensional space.

The hyperparameters tuned for SVM were C, a parameter that balances the trade-off between margin size and classification error; gamma, which controls the influence of individual data points; and kernel, which defines how data is transformed into higher dimensions.

4) Decision Tree: A Decision Tree is a flowchart-like tree structure used for making decisions based on input features. Each internal node represents a test on an attribute, each branch represents the outcome of that test, and each leaf node represents a class label or continuous value (for regression)

The hyperparameters tuned for Decision Tree were max_depth, which limits the trees depth to prevent overfitting; min_samples_split, which sets the minimum number of samples required to split a node; and min_samples_leaf, which determines the minimum samples allowed in a leaf node to avoid overly small branches.

5) Random Forest: A Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

For the Random Forest model, the tuned hyperparameters included max_depth, which limits tree depth to prevent overfitting; n_estimators, number of trees for stability and accuracy; max_features, features considered per split to balance complexity and performance; min_samples_leaf and min_samples_split, minimum samples for leaf nodes and splits to reduce overfitting; and criterion, Gini Impurity vs. Entropy.

6) Artificial Neural Network (ANN): An Artificial Neural Network is a computational model inspired by the neural networks of the human brain. It consists of layers of interconnected nodes (neurons), where each connection has an associated weight.

For the Artificial Neural Network, the tuned hyperparameters included the number of hidden layers/neurons per layer, to control the depth and capacity of the network and regularization techniques like L1 and L2 regularization to help prevent overfitting by constraining the model's complexity.

7) Deep Neural Decision Tree (DNDT): A Deep Neural Decision Tree integrates the concept of decision trees within a deep learning framework. Unlike traditional decision trees that use hard thresholds to split data, deep neural decision trees use soft, differentiable decision functions.

As a result of long runtimes, DNDT and DNDF were hand-tuned. The hyperparameters tuned were num_trees, depth, and used_features_rate.

8) Deep Neural Decision Forest (DNDF): A Deep Neural Decision Forest is a hybrid model that combines the representation learning capabilities of deep neural networks with the decision-making structure of decision forests.

As a result of long runtimes, DNDT and DNDF were hand-tuned. The hyperparameters tuned were num_trees, depth, and used_features_rate.

9) XGBoost: XGBoost is a scalable, tree-based ensemble learning method derived from the gradient boosting framework. It builds a series of weak learners (decision trees) in a sequential manner, where each subsequent tree focuses on correcting the errors made by the previous ones.

The hyperparameters used for XGBoost were learning rate, which determines the contribution of each new tree added during the boosting process (determines step size for updating weights) and other parameters similar to a Random Forest, such as max_depth, min_samples_split, and min_samples_leaf. These hyperparameters help control model complexity, reduce overfitting, and improve predictive accuracy.

B. Data Collection

Data for this model was collected using the Baseball Savant and Baseball Reference databases, which provide a database of player statistics and biometric data with the ability to filter search based on a number of selection criteria. The data used in the models was collected and verified by Statcast, a state-of-the-art tracking technology used throughout the MLB, and then uploaded to Baseball Savant. This method of data collection allowed for easily accessible, reproducible data that has been standardized throughout the MLB. Baseball Reference was used to obtain player height and weight. The full collection of data was written in tabular format.

Determination of selected features was influenced by both availability and potential to influence elbow injury outcomes, affirmed by the presence of scientific literature highlighting predictors of elbow injury in baseball pitching. Model training utilized balanced a data set consisting of 50% positive and 50% negative for UCL reconstruction in order to perform a matched-pair design. A minimum of 1500 career pitches was established as selection criteria to ensure statistic viability in conjunction with a 1500 pitch minimum to accurately track spin rate. Additionally, only pitchers who threw the aforementioned 1500 pitches from the 2015 season and later and who sustained their UCL injury in the MLB will be considered due to advanced metric availability. Hyperparameter tuning of all models was performed through a gridsearch with 5-fold cross-validation.

In total, 78 pitchers were both positive for UCL reconstruction and satisfied the selection criteria. In sampling for pitchers negative for UCL reconstruction, it was determined that of the 1633 pitchers that satisfied the selection criteria, 239 were excluded due to having received UCL reconstruction. 78 of the remaining 1394 pitchers were sampled at random for data collection to eliminate potential bias in grouping. Thus, the remaining pitchers were sampled at random until 78 unique, viable data points were obtained.

1) Pair Selection: The nuances of modeling and identifying age and cumulative statistics in ML stem from the risk of confounding variables. Due to the potential negative performance impacts of UCL reconstruction, data collection in individuals positive for UCL reconstruction was taken until the last data points taken before the procedure, thus preventing any risk of data contamination. However, preliminary halting of data collection in individuals negative for UCL reconstruction was also necessary to prevent age from acting as a confounding variable. Our model introduces a paired-control design in order to account for age-related discrepancies. UCL reconstruction-positive individuals were matched by age with negatives, ensuring that matches were based on age, data availability, and retirement age (ensuring that players were still active at the matched age).

C. Feature Selection

The determination of selected features was influenced by both availability and potential to influence elbow injury outcomes, affirmed by the presence of scientific literature highlighting predictors of elbow injury in baseball pitching. Thus, explained below are our features selected, along with hypothesized justifications for their respective influences on the presence of UCL reconstruction.

1) Fastball Velocity: Fastball velocity represents the average speed at which a pitcher's fastball travels. This feature was selected due to its correlated impact on the forces exerted on the elbow joint. Hurd et al. found in 2012 in young baseball pitchers that increased pitch velocity resulted in increased susceptibility to elbow injury. A higher fastball velocity will correspond to greater forces applied to the joint, including the UCL. Therefore, as greater forces on the elbow joint increase the amount of biomechanical stress placed on the UCL, fastball velocity should have a positive correlation with UCL reconstruction.

2) Fastball Percentage: Fastball percentage represents the proportion of a pitcher's throws that are classified as fastballs. This exerts increased stress on the elbow joint, as defined above.

3) Fastball Spin Rate: Fastball spin rate measures the rate of rotation of a pitcher's fastballs as they travel through the air. A higher spin rate is primarily caused by a downwards wrist motion, but also by an increased ”snap” of the elbow, requiring increased power output from the UCL.

4) Breaking Ball Velocity: Breaking ball velocity represents the average speed at which a pitcher's breaking ball travels. The name stems from the nature of each of these pitches to ”break,” or move away from the strike zone due to a difference of pressure on opposite sides of the rapidly spinning baseball resulting in a force perpendicular to the path of motion, also known as the Magnus effect. (Seifert 2012)

The generation of a breaking ball's high angular velocity is done by the arm and body of the pitcher. This motion requires greater wrist and elbow flexion (Dun et al. 2007). Additionally, Manzi et al. found in 2023, in professional baseball pitchers, that there was an increase in shoulder horizontal adduction torque in curveballs, a type of breaking ball. Thus, the throwing motion of a breaking ball may place the elbow joint of the pitcher in a more compromised position (Tamate et al. 2019). However, there is currently a debate among scientific communities if the mechanics of a breaking ball itself contribute to an increased risk of injury, and this debate is reflected by mixed results produced in scientific literature.

Furthermore, many official guides advising caution regarding curveballs and other breaking balls are guided by expert opinion rather than scientific literature.

5) Breaking Ball Percentage: Breaking ball percentage represents the proportion of a pitcher's throws that are classified as breaking balls. A higher breaking ball percentage signifies a greater percentage of a pitcher's throws, which places the arm in the aforementioned position.

6) Total Pitches: The total pitch quantity represents the total number of pitches made in a pitcher's career until the last point of data collection.

Each pitch represents an exertion of torque and force on the elbow joint, so a higher pitch count will correspond with more total force applied to the elbow joint, assuming all other variables are held constant.

Thus, given the strong correlation between biomechanical forces on the elbow joint and elbow injury, the current scientific literature supports a positive correlation between elbow joint overuse and UCL injury, especially within the context of pitching (Fortenbaugh et al. 2009, Oyama 2012).

7) Breaking Ball Spin Rate: Breaking ball spin rate measures the rate of rotation of a pitcher's breaking balls as they travel through the air. A higher spin rate is primarily caused by increased ”snap” of the elbow.

8) Weight: Higher weight may signify an average of increased applied forces to a baseball. (Forsythe et al. 2017) Furthermore, TA Miller reported in the *NCSA's Guide to Tests and Assessments* that ”additional weight (in the form of nonessential fat) provides greater resistance to athletic motion thereby forcing the athlete to increase the muscle force of contraction per given workload.” Thus, the impact of weight on applied musculoskeletal force may extend beyond only the reported output.

Werner et al. found in a 2008 study found a strong correlation between weight and ball velocity in collegiate baseball pitchers. This finding further establishes weight as a potential predictor for pitch velocity, which then corresponds to increased force on the UCL.

9) Height: Downs et al. found in a 2021 study that within youth pitchers, upper arm length correlated strongly with increased elbow torque and force applied. Furthermore, height is strongly correlated with upper arm length (Sarma et al. 2020) but is a more readily available statistic.

Thus, generally, greater height in a pitcher correlates to greater torque applied to the elbow joint when force is identical, further asserted by the equation for torque $\tau = F \times d$, as increased arm length (strongly correlated with height) results in a longer lever-arm.

D. Results

Table I presents the descriptive statistics for feature differences across player pairs, reported as mean ± standard deviation. These values highlight the variability in key pitching and physical characteristics, such as pitch counts, velocity, spin, and body measurements, within the dataset (n = 156).

TABLE I
DESCRIPTIVE STATISTICS FOR FEATURES (DIFFERENCES ACROSS PAIRS)

n = 156

Feature	Pair Differences
Total Pitches	3167.4 ± 7634.6
Fastball Percentage	13.03 ± 19.0
Fastball Velocity	0.3115 ± 3.98
Fastball Spin	-16.41 ± 210.3
Breaking Percentage	4.135 ± 14.02
Breaking Velocity	-0.0282 ± 5.05
Breaking Spin	-84.68 ± 381
Height	0.231 ± 7.734
Weight	1.962 ± 13.64

*Values are presented as mean ± standard deviation.

TABLE II
POINT-BISERIAL CORRELATIONS BETWEEN FEATURES AND UCL RECONSTRUCTION

Feature	Test Statistic	P-Value
Total Pitches	0.247000	**0.001817**
Fastball%	0.425776	**0.000000**
Fastball Velocity	0.066652	0.406879
Fastball Spin	-0.049089	0.541503
Breaking%	0.177838	**0.025861**
Breaking Velocity	0.000392	0.996116
Breaking Spin	-0.142420	0.075184
Height	0.029701	0.711935
Weight	0.092388	0.249799

*$\alpha = 0.05$

A point-biserial correlation coefficient (r_{pb}) was utilized to quantify the strength and direction of the association between continuous features and the binary outcome variable (UCL reconstruction). Significant positive correlations were observed for Total Pitches (r = 0.247, p = 0.0018), Fastball Percentage (r = 0.426, p ¡ 0.0001), and Breaking Ball Percentage (r = 0.178, p = 0.0259), indicating these features are significantly associated with UCL reconstruction at the = 0.05 level.

Feature importance was assessed using Shapley values for the Deep Neural Decision Forest (DNDF). In the future, to strengthen the interpretability of feature contributions across models, Shapley analysis should be extended to other models, providing a more holistic comparison of feature impact. This multi-model perspective would allow for a deeper understanding of feature importance, which may be further validated through consultation with domain experts to ensure alignment with real-world knowledge.

V. Analysis

A. Accuracy

TABLE III
Model Types and Their Accuracies

Model Type	**Accuracy (%)**
Logistic Regression (LR)	62.5
K-Nearest Neighbors (KNN)	62.5
Support Vector Machine (SVM)	66.7
Decision Tree (DT)	66.7
Random Forest (RF)	71.9
XGBoost	66.7
Artificial Neural Network (ANN)	68.0
Deep Neural Decision Tree (DNDT)	70.8
Deep Neural Decision Forest (DNDF)	79.2

Table III summarizes the accuracy of various machine learning models used for UCL reconstruction prediction. The models range from traditional approaches like Logistic Regression (62.5%) to advanced architectures like the Deep Neural Decision Forest (79.2%), highlighting the performance differences across methods.

Deep Neural Decision Forest (DNDF) achieved the highest accuracy at 79.2%, indicating its strong ability to capture complex relationships and make robust predictions in this dataset.

Random Forest (RF) outperformed many models with an accuracy of 71.9%, benefiting from its ensemble approach of combining multiple decision trees.

Deep Neural Decision Tree (DNDT) achieved an accuracy of 70.8%, highlighting the strength of deep learning-based tree structures.

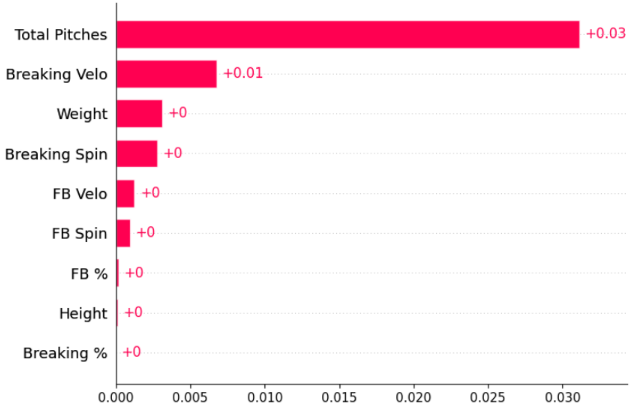
Artificial Neural Network (ANN) attained an accuracy of 68.0%, demonstrating the model's ability to handle complex, non-linear data.

Support Vector Machine (SVM), XGBoost, and Decision Tree (DT) achieved 66.7% accuracy, reflecting their ability to manage non-linear relationships but likely being constrained by the dataset's size or feature complexity.

Logistic Regression (LR) and K-Nearest Neighbors (KNN) had the lowest accuracies at 62.5%, suggesting that these simpler models struggled to capture the complexity of the data.

B. Feature Importance

Fig. 1. SHapley Additive exPlanations



SHAP (SHapley Additive exPlanations) is applied to quantify feature contributions and evaluate how the DNDF makes decisions, enhancing the interpretability of its predictions on low-dimensional tabular data. Greater relative feature importance (horizontal axis) represents a higher impact of a specific feature on the model. Consequently, future projects should prioritize studying features like total pitches and breaking velocity, which were highlighted as important by SHAP analysis, for a deeper understanding of UCL reconstruction.

The SHAP analysis also reinforces the point-biserial correlations utilized in section III, particularly the high significance of total pitches, low importance of height, and moderate relevance of breaking spin. However, the DNDF model diverged from the point-biserial findings by entirely rejecting fastball percentage, which had been identified as significant, and instead emphasized breaking velocity, a feature that the point-biserial analysis did not find important. These two differences were the most pronounced, while other feature importance discrepancies were relatively minor.

VI. Discussion

A. Model Comparison

The Artificial Neural Network's (ANN's) modest accuracy of 68% is not unexpected given the nature of the dataset. While ANNs excel in domains with large sample sizes with patterns that are high-dimensional and non-linear, the relatively low-dimensional tabular data in this experiment may limit the advantages of deep neural architectures. In such cases, ANNs can struggle with overfitting or fail to efficiently capture pivotal dynamics for decision-making in more straightforward, structured datasets. Despite tuning the architecture and employing regularization techniques, the ANN's performance was constrained by its inherent reliance on learning complex representations, which may be unnecessary for this dataset.

Conversely, the random forest (RF) demonstrated slightly better performance, with an accuracy of 71.9%. This is consistent with the well-known strength of tree-based models on tabular data, where feature splits and ensemble learning allow for robust decision boundaries. RF is particularly adept at handling feature interactions and non-linear relationships, two characteristics that are often present in tabular data. Its reliance on bagging and the random selection of features per split results in models that are less prone to overfitting and more resilient to noise. The marginal improvement in performance over the ANN could potentially be attributed to the RF's ability to better capture the structure of the data, despite the simplicity of the dataset.

The Deep Neural Decision Tree (DNDT) achieved an accuracy of 70.8%, positioning its performance between the ANN and RF, but notably below the Deep Neural Decision Forest (DNDF). As a hybrid model combining neural network-based feature learning with a decision tree architecture, the DNDT benefits from its ability to capture complex patterns within the data, while maintaining the interpretability of a tree-based model. However, its performance lagged behind the DNDF, likely due to the absence of ensemble learning that is fundamental to the DNDF's success. Unlike DNDF, which leverages multiple trees to reduce variance and improve robustness, the DNDT operates as a single tree structure. This makes it more vulnerable to overfitting and less capable of fully exploring diverse feature interactions. Additionally, while the neural component of DNDT aids in learning representations, it lacks the depth and breadth of representation learning seen in DNDF, which can fine-tune these representations across multiple trees. This combination of factors likely explains why the DNDT did not achieve the same level of performance, as it was unable to capitalize on the strengths of ensemble methods that are critical in enhancing decision-making processes in complex models like DNDF.

The modest performance of DNDT relative to RF can be attributed to two primary factors. First, while the neural network embedded within DNDTs can capture hierarchical patterns, in this case, the tabular data may not have benefited substantially from these learned representations, limiting the potential of the neural feature extraction process. Second, as DNDT structures are generally less complex than Deep Neural Decision Forests, they may not capitalize as fully on ensemble learning, which can reduce variance and improve robustness. The DNDT's reliance on a single tree-like decision process, combined with learned representations, may explain why its performance did not surpass that of the RF, which effectively utilizes multiple trees to reduce overfitting.

The DNDF achieved an accuracy of 79.2%, which can potentially be explained by its hybrid nature, combining the representational power of neural networks with the structured decision-making capabilities of forests. The architecture of DNDF incorporates deep neural networks to learn abstract representations from the data, which are then passed through decision trees that fine-tune these representations and make final predictions. This allows the DNDF to leverage the strengths of both approaches, mitigating their individual weaknesses.

The DNDF's ability to outperform the ANN, RF, and DNDT may arise for 2 key reasons: (1) By integrating neural networks into the framework, the DNDF can learn hierarchical feature representations, capturing complex relationships in the data that a traditional RF might miss. These learned representations, once passed to the decision trees, allow for more precise splits and enhanced decision-making. The decision trees, in turn, are less susceptible to the overfitting challenges most often faced by neural networks, as they are constrained by the structure of the decision-making process. This combination of learning feature representations and structured decisions likely contributes to the improved accuracy observed. (2) The ensemble nature of the DNDF, where multiple trees work in tandem, also capitalizes on the strength of decision forests in reducing variance through aggregation. Each tree can make use of the representations learned by the neural network, while maintaining the individual trees' robustness against overfitting. This combination of feature learning and decision-making likely explains why the DNDF was able to outperform the RF, which lacks the representational power of a neural network, and the ANN, which does not have the ensemble mechanism of decision trees to reduce variance. Furthermore, the DNDF's ability to refine and aggregate multiple learned representations allows it to capture more complex patterns in the data, giving it an edge over the simpler approach of the DNDT.

B. Modeling Limitations

Although our presented models display robust levels of accuracy, they currently possess key limitations that may inhibit their ability to be used as fully accurate predictors in injury analysis contexts.

(1) The data collected is unable to account for player metrics recorded prior to entering the MLB (Minor League, NCAA, High School, etc.).

(2) Feature selection was largely dependent on available metrics on the Baseball Savant. Most advanced metrics were only available starting in 2015 with the creation of Statcast, leading to the exclusion of a considerable number of data points.

C. Data Collection Challenges and Solutions

Data availability poses a significant challenge in predicting Tommy John surgery risks, e.g., for high school, college, and Minor League Baseball (MiLB) players. Comprehensive injury and performance data are often lacking in these leagues. As a result, only data collected from a pitcher's career in the MLB is viable to be used by the models. Improving data collection methodologies through the improvement and development of standardized injury reporting and the potential integration of wearable sensor technologies is essential for enhancing predictive modeling accuracy across all levels of play.

Feature importance analysis highlighted critical predictors influencing surgery risk, which are validated not only by the training algorithm but also by an analysis of each feature's effect on the elbow joint. The most important features, total pitches, was triple the importance of the next most important feature, breaking velocity. Biomechanically, this could be explained by the significant effects of overuse on stress to the elbow joint. Total pitches are a cumulative metric of elbow use and thus may be the best correlator to UCL reconstruction. Breaking velocity, therefore, may also be an indicator of the comparative stress of breaking pitches, which could potentially be explained by the elbow joint being subject to higher torques.

Weight and Breaking Spin all had similar, albeit comparatively insignificant, relative importance. The lack of relative importance could then be explained by the potential weakness of the correlation between these features and elbow torque.

Fastball velocity and spin both had similar yet even less significant impacts on model output. This result for fastball velocity does appear to contrast with current standings of scientific literature and our hypothesized feature significance, and it is thus recommended that this be investigated further.

Fastball percentage, breaking percentage, and height all had negligible effects on model output. The results of fastball percentage and breaking percentage feature importance may signify one of two conclusions.

(1) All pitch types exert similar biomechanical torques on the elbow joint.

(2) Differences in biomechanical torques were either not pronounced enough in the dataset or were vastly overshadowed by other selected features.

Breaking velocity and breaking percentage were the two least important features. This may be related to a lack of definitive scientific and anatomical conclusions relating breaking balls to increased elbow injury. Furthermore, the comparative value of a breaking ball within baseball is largely based on spin, rather than a fastball which is conversely based on speed.

D. Healthcare Equity Considerations

Equitable healthcare access remains a critical issue in the United States of America, particularly concerning negative health outcomes and limited options for injury treatment among underrepresented communities (Kliethermes S. A. et al. 2024). Racial disparity in access to surgery is prevalent in the medical field, such that racial and ethnic minority patients are less able to receive proper medical diagnosis and surgical treatment for a multitude of varying procedures (de Jager, E. et al. 2019). Furthermore, from 2004 to 2014, the cost of recovery for Tommy John surgery among MLB pitchers was on average $1.9 million per year, a cost that may disproportionately affect professional athletes from more underprivileged backgrounds (Meldau, J. et al. 2019). Additionally, a significant percentage of patients are unable to return to their previous level of competition after undergoing Tommy John surgery (Cain E. L. et al, 2010). By utilizing machine learning for the prediction of early injury risk, this study contributes to enhancing healthcare equity in professional sports, ensuring all players receive timely intervention and support.

E. Future Directions + Feature Selections

Continued advancements in machine learning, particularly with the integration of real-time data streams from wearable technologies, provide exciting opportunities to further refine injury prediction models. A key direction for future research lies in addressing the impact of pitch counts prior to a player's MLB career. Although current models predominantly rely on data from professional baseball, cumulative pitch counts in the years leading up to the major leagues may be equally significant in determining injury risk. Estimating these pre-MLB pitch counts, taking into account player experience and age, could offer a more comprehensive understanding of elbow stress over a player's career. This would require expanding the datasets to include information from amateur leagues, potentially improving predictive accuracy and helping to inform early intervention strategies for young athletes.

As a future direction, refining model evaluation through metrics such as F1-score, precision, and recall could enhance the reliability of injury prediction models. Precision and recall are essential for understanding the trade-off between correctly identifying injury risks and avoiding false negatives that could lead to missed predictions. While precision measures the accuracy of positive predictions, recall focuses on the model's ability to correctly identify all true injury cases. F1-score, the harmonic mean of precision and recall, is particularly useful for balancing these two metrics, especially when the dataset is imbalanced, as it might be in injury prediction tasks where most athletes do not suffer from injuries. Incorporating these metrics into model optimization would ensure that the prediction models prioritize both identifying high-risk players accurately and minimizing false alarms, which could ultimately lead to more actionable insights for prevention strategies in sports.

Another important direction for future work is the refinement of feature selection. In our study, total pitch count emerged as the most important feature in predicting UCL reconstruction. However, there are other factors, such as pitch velocity, spin rate, and recovery time between appearances, that could further enhance model performance. While the current study has validated the impact of total pitch count, future research should continue to evaluate and validate these additional features, ensuring that only the most scientifically significant predictors are incorporated into the models. This iterative process of feature refinement will help ensure that injury prediction models remain robust and applicable across different player populations and environments. Feature importance methods such as SHAP could be applied to additional high-performing models to evaluate the consistency of predictive factors identified to be impactful.

While age and its associated effects on UCL injury risk have not been a primary focus in most predictive models, there is a growing recognition that age-related declines in collagen synthesis play a significant role in ligament and tendon integrity. Collagen, particularly type I collagen, is essential for maintaining the structural strength and resilience of the UCL. As athletes age, the body's natural production of collagen decreases, reducing tissue elasticity and increasing vulnerability to injuries. While our current model does not explicitly factor in age, future studies could explore how age-related collagen depletion influences injury risk, particularly in the context of cumulative stress from pitching. By incorporating age as a feature, future models could offer more personalized predictions that account for the biological processes contributing to UCL injuries. This would also open the door to potential interventions, such as regenerative medicine therapies aimed at improving collagen synthesis, which could enhance recovery and reduce re-injury rates in older athletes.

Additionally, examining the timing of UCL surgeries and the likelihood of pitchers returning to pre-surgery form is an area ripe for investigation. Surgical techniques for Tommy John surgery have advanced significantly, leading to faster recovery times and higher rates of return to the field. However, more research is needed to understand how the timing of the surgery, both in terms of when it is performed and how soon after the injury, is linked to recovery outcomes. By analyzing trends in recovery and performance post-surgery, future studies could develop predictive models that not only estimate injury risk but also forecast the likelihood of a player returning to their previous level of performance. This would provide invaluable insights for athletes, coaches, and medical teams, enabling more informed decisions regarding treatment and rehabilitation strategies.

These areas of future research will be crucial for enhancing the accuracy and applicability of injury prediction models, not only in baseball but also across a wide range of sports. By refining our understanding of the complex factors influencing UCL injuries, we can develop more effective strategies for injury prevention and rehabilitation, ultimately improving athlete health and performance.

VII. Conclusion

Our approach is the first to demonstrate the efficacy of a Deep Neural Decision Forest (DNDF) model over a Random Forest model in predicting UCL tear risk among baseball pitchers. By leveraging comprehensive tabular biomechanical player data and advanced analytics, we illustrate that a machine learning model with a scientifically justified feature set can predict the likelihood of UCL reconstruction surgery with high probability. This capability facilitates the early identification of injury risks, supporting proactive healthcare strategies and promoting player welfare and performance outcomes. Our paper advances existing research by showcasing the potential of DNDFs to provide a novel, data-driven approach to injury prediction.

Among the nine models tested, the DNDF achieved the highest accuracy at 79.2%, surpassing the previous best accuracy of 75% reported by Whiteside et al. using a support vector machine. This finding establishes the DNDF as the most effective model for predicting UCL reconstruction to date. Total pitch count emerged as the most important predictive feature, underscoring its relevance in UCL injury risk. The simplicity of the training process using low-dimensional tabular data highlights the ease of feature selection and data collection, as well as the feasibility of pulling data from large public databases. This simplicity, combined with the predictive power of DNDF, may emphasize the broader applicability of our approach to other datasets and sports where injury prediction and prevention strategies can lead to cost-effective, health-promoting interventions.

References

[1] Ahmad, C. S. et al. (2020). Ulnar collateral ligament reconstruction in Major League Baseball pitchers: Epidemiology, performance, and career longevity. Journal of Shoulder and Elbow Surgery, 29(8), 1611-1619.

[2] Cain, E. L. et al. (October 7, 2010). Outcome of ulnar collateral ligament reconstruction of the elbow in 1281 athletes: Results in 743 athletes with minimum 2-year follow-up. The American Journal of Sports Medicine, 38(12): 2426-2434.

[3] Chalmers, P. N., Wimmer, M. A., Verma, N. N., Cole, B. J., Romeo, A. A., Cvetanovich, G. L., Pearl, M. L. (2017). The relationship between pitching mechanics and injury: A review of current concepts. Sports Health, 9(3):216-221. doi:10.1177/1941738116686545.

[4] Claudino, J. G. et al. (2021). Machine learning applications in sports: A systematic review. Journal of Sports Sciences, 39(5), 477-494.

[5] Conte, S. et al. (2016). Epidemiology of ulnar collateral ligament reconstruction in Major and Minor League Baseball pitchers: Comprehensive report of 1429 cases. Journal of Shoulder and Elbow Surgery, 25(6), 872-879.

[6] de Jager, E. et al. (March 1, 2019). Disparities in surgical access: A systematic literature review, conceptual model, and evidence map. Journal of the American College of Surgeons, 228(3): 276-298.

[7] Downs, J. L., Wasserberger, K. W., Barfield, J. W., Saper, M. G., Oliver, G. D. (2021). Increased upper arm length and loading rate identified as potential risk factors for injury in youth baseball pitchers. The American Journal of Sports Medicine, 49(11), 3088-3093. doi:10.1177/03635465211028555.

[8] Dun, S., Loftice, J., Fleisig, G. S., Kingsley, D., Andrews, J. R. (2008). A biomechanical comparison of youth baseball pitches: Is the curveball potentially harmful? The American Journal of Sports Medicine, 36(4): 686-692. doi:10.1177/0363546507310074.

[9] Fortenbaugh, D., Fleisig, G. S., Andrews, J. R. (2009). Baseball pitching biomechanics in relation to injury risk and performance. Sports Health, 1(4): 314-320. doi:10.1177/1941738109338546.

[10] Forsythe, C. M., Crotin, R. L., Greenwood, M., Bhan, S., Karakolis, T. (2017). Examining the influence of physical size among major league pitchers. Journal of Sports Medicine and Physical Fitness, 57(5), 572-579. doi:10.23736/S0022-4707.16.06355-6.

[11] Fury, M. S., Oh, L. S., Linderman, S. E., Wright-Chisem, J., Fury, J. N., Scarborough, D. M., Berkson, E. M. (2021). Return to performance after ulnar collateral ligament reconstruction in Major League Baseball pitchers: A case-control assessment of advanced analytics, velocity, spin rates, and pitch movement. Orthopedic Journal of Sports Medicine, 9(9): 23259671211035753. doi:10.1177/23259671211035753.

[12] Karnuta, J. M., Luu, B. C., Haeberle, H. S., et al. (2020). Machine learning outperforms regression analysis to predict next-season Major League Baseball player injuries: Epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000-2017. Orthopedic Journal of Sports Medicine, 8(11). doi:10.1177/2325967120963046.

[13] Khalil, L. S., Jildeh, T. R., Abbas, M. J., Klochko, C. L., Scher, C., Van Holsbeeck, M., Muh, S. J., Makhni, E. C., Moutzouros, V., Okoroha, K. R. (2021). Elbow torque may be predictive of anatomic adaptations to the elbow after a season of collegiate pitching: A dynamic ultrasound study. Arthroscopy, Sports Medicine, and Rehabilitation, 3(6), e1843-e1851. https://doi.org/10.1016/j.asmr.2021.08.012.

[14] Leland, D. P. et al. (September 29, 2019). Prevalence of medial ulnar collateral ligament surgery in 6135 current professional baseball players: A 2018 update. *Orthopedic Journal of Sports Medicine, 7(9).*

[15] Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, *22*(5), 717-727

[16] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, *35*, 507-520.

[17] Kannus, P. (2000). Structure of the tendon connective tissue. *Scandinavian Journal of Medicine & Science in Sports*.

[18] Kliethermes SA, Asif IM, Blauwet C, et al. (2023). Focus areas and methodological characteristics of North American-based health disparity research in sports medicine: a scoping review. *Br J Sports Med* Epub ahead of print: doi:10.1136/ bjsports-2023-107607

[19] Kontschieder, P., Fiterau, M., Criminisi, A., & Bulo, S. R. (2015). Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision* (pp. 1467-1475).

[20] Manzi, J. E., Dowling, B., Trauger, N., Hansen, B. R., Quan, T., Dennis, E., Fu, M. C., Dines, J. S. (2023). The relationship between maximum shoulder horizontal abduction and adduction on peak shoulder kinetics in professional pitchers. Sports Health, 15(4), 592-598. doi:10.1177/19417381221104038.

[21] Mays, P. K., McAnulty, R. J., Campa, J. S., Laurent, G. J. (1991). Age-related changes in collagen synthesis and degradation in rat tissues: Importance of degradation of newly synthesized collagen in regulating collagen production. Biochemical Journal, 276(Pt 2), 307-313. doi:10.1042/bj2760307.

[22] Meldau, J. et al. (2020). Cost analysis of Tommy John surgery for Major League Baseball teams. Journal of Shoulder and Elbow Surgery, 29(1), 121-125.

[23] Miller, T. A. (2017). NSCA's Guide to Tests and Assessments. Retrieved from https://www.nsca.com/education/articles/kinetic-select/sport-performance-and-body-composition/.

[24] Oyama, S. (2012). Baseball pitching kinematics, joint loads, and injury prevention. Journal of Sport and Health Science, 1(2), 80-91. doi:10.1016/j.jshs.2012.06.004.

[25] Rendar, S. A., Ma, F. (2022). Predicting ulnar collateral ligament injury in rookie Major League Baseball pitchers. arXiv preprint, arXiv:2207.00585.

[26] Sarma, A., Barman, B., Das, G. C., Saikia, H., Momin, A. D. (2020). Correlation between the arm-span and the standing height among males and females of the Khasi tribal population of Meghalaya state of North-Eastern India. Journal of Family Medicine and Primary Care, 9(12), 6125-6129. https://doi.org/10.4103/jfmpc.jfmpc.1350.20

[27] Seroyer, S. T., Nho, S. J., Bach, B. R., Bush-Joseph, C. A., Nicholson, G. P., Romeo, A. A. (2010). The kinetic chain in overhand pitching: Its potential role for performance enhancement and injury prevention. Sports Health, 2(2), 135-146. doi:10.11771941738110362656.

[28] Tamate TM, Garber AC. Curveballs in Youth Pitchers: A Review of the Current Literature. Hawaii J Health Soc Welf. 2019 Nov;78(11 Suppl 2):16-20.

[29] Trasolini, N. A., Nicholson, K. F., Mylott, J., Bullock, G. S., Hulburt, T. C., Waterman, B. R. (2022). Biomechanical analysis of the throwing athlete and its impact on return to sport. Arthroscopy, Sports Medicine, and Rehabilitation, 4(1), e83-e91. doi:10.1016/j.asmr.2021.09.027.

[30] Varani, J., Perone, P., Fligiel, S. E., et al. (2006). Inhibition of collagen synthesis by chronic exposure to free fatty acids. *Journal of Investigative Dermatology*

[31] Werner, S. L., Suri, M., Guido, J. A., Meister, K., Jones, D. G. (2008). Relationships between ball velocity and throwing mechanics in collegiate baseball pitchers. Journal of Shoulder and Elbow Surgery.

[32] Whiteside, D., Martini, D. N., Lepley, A. S., Zernicke, R. F., Goulet, G. C. (2016). Predictors of ulnar collateral ligament reconstruction in Major League Baseball pitchers. The American Journal of Sports Medicine, 44(9), 2202-2209. doi:10.1177/0363546516643812.