Journal of Big Data

RESEARCH Open Access

Data-driven approaches for predicting Tommy John Surgery risk in major league baseball pitchers

Bosuk Kang[1,2], Minsu Park[3*], Angel P. del Pobil[4] and Eunil Park[3,4*]

*Correspondence:
mspark501@g.skku.edu;
eunilpark@skku.edu

[1] Department of Semiconductor and Display Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[2] Samsung Electronics, Suwon 16419, Republic of Korea
[3] Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, Republic of Korea
[4] Robotic Intelligence Laboratory, Jaume I University, Castellon 12071, Spain

Abstract

Injury management is critical in all sports, directly impacting player performance. Baseball players are particularly susceptible to injuries, as players often compete in 5 to 7 games per week, placing continuous strain on their bodies. Among various injuries, Tommy John Surgery (TJS) poses a notable risk for Major League Baseball (MLB) pitchers. Traditional TJS prediction methods required sensors or video-based motion capture, which are impractical during actual games and limited in making predictions too close to the injuries, such as within 30 pitches. To address these challenges, this study proposes a deep learning (DL) framework that utilizes both classification and regression tasks. Using MLB pitching data (2016–2023), the classification model detects injury risk up to 100 days in advance with a high prediction performance of 0.73 F1-score, while the regression model estimates the time remaining until the player's last pre-surgery game with R^2 of 0.79. In addition, to enhance our model's applicability, we employ an explainable artificial intelligence technique to analyze the impacting mechanical features, such as a lowered four-seam fastball release point, that accelerate UCL deterioration, increasing TJS risk. These findings provide a practical foundation for early intervention strategies, potentially preserving pitcher health and reducing the need for complex surgical procedures.

Keywords: Injury prediction, Tommy John Surgery (TJS), Bigdata, Baseball, Deep learning (DL), Classification, Regression

Introduction

Effective injury management is paramount for optimizing player performance across all sports. In team sports, injuries can have significant effects on individual performance and overall team success, often leading to substantial competitive and financial setbacks.

For instance, a study of the English Premier League (EPL) [1], one of the world's most elite soccer leagues, estimated that injuries cost teams approximately £45 million per season. Additionally, every 136 player-absence days correlates to a one-point loss in league standings, underscoring the critical importance of injury prevention and management in preserving team competitiveness and mitigating economic losses.

Baseball, however, presents even greater challenges in injury-related management compared to soccer due to its demanding schedule. The EPL season spans from August to May, with teams playing 38 matches-roughly one game per week [2]. In contrast, Major League Baseball (MLB) runs from March to October, with teams playing 162 games per season, averaging 5 to 7 games per week [3]. This significantly condensed schedule affords MLB players far less recovery time, increasing the risk of cumulative fatigue and injury (Table 1). Moreover, the distinct specialization of player roles, such as pitchers and hitters, introduces unique biomechanical stressors, necessitating position-specific injury prevention strategies. These factors make injury management in baseball not only more complex but also more critical to the longevity and success of players and teams.

A study on hitter injuries in Major League Baseball (MLB) and Minor League Baseball (MiLB) revealed that 62.2% of shoulder injuries-the most common type of injury among hitters-result from sudden movements or acute trauma [4]. In contrast, research on MLB and MiLB pitchers highlighted that elbow injuries are the most prevalent, with serious cases leading to more than five days of roster exclusion. Specifically, 59.1% of starting pitchers and 58.3% of relief pitchers reported elbow injuries, underscoring the unique vulnerability of this player group [5].

The distinct types and causes of injuries across baseball positions necessitate tailored approaches to injury prediction and management. For hitters, the predominance of sudden, trauma-induced injuries makes tracking specific metrics for prevention particularly challenging. Moreover, the relatively short average recovery time of 22 days for hitter injuries further limits the utility of predictive models [4]. However, the repetitive, chronic nature of many pitcher injuries provides greater potential for identifying early warning signs, allowing more targeted interventions.

For pitchers, elbow injuries, the most common type, are mainly caused by repetitive stress over extended periods [6]. As such, continuous monitoring and identification of key contributing factors are critical for effective injury management. If an elbow injury progresses beyond the point of rehabilitation, it can result in ulnar collateral ligament (UCL) damage, requiring Tommy John Surgery (TJS). This procedure involves reconstructing the damaged ligament using a tendon typically harvested from the wrist or hamstring.

TJS is a severe, long-term injury with an average recovery period of approximately 20.5 months [7], imposing substantial economic costs. A study of 194 MLB pitchers who underwent TJS estimated an average financial loss of $1.9 million per pitcher [8]. Moreover, as shown in Figure 1, the number of MLB pitchers undergoing TJS has steadily increased since its introduction. Consequently, understanding and predicting pitcher injuries associated with TJS is critical to mitigating financial losses and safeguarding the health and careers of players while preserving team performance.

Table 1 Comparison of match frequency between EPL and MLB

League	Duration(months)	Total matches	Weekly games
EPL	9.5	38	About 1
MLB	7.5	162	About 6

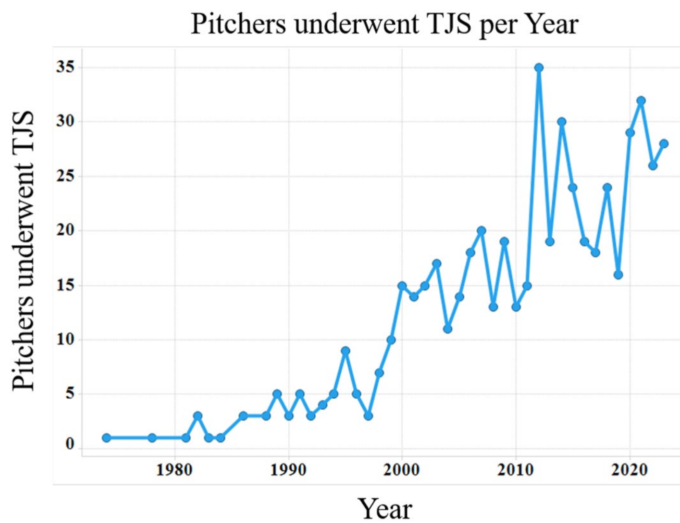


Fig. 1 Annual number of TJS performed on MLB pitchers since the procedure's introduction in 1974

Previous research on managing and predicting elbow injuries in pitchers mainly depended on pre-season pitching analysis or medical imaging. For instance, several studies examined pitching mechanics using video analysis to identify injury-related factors [9, 10], while others utilized advanced imaging techniques, such as stress ultrasonography, to predict UCL injuries based on ligament morphology [11, 12]. However, these approaches are limited by their inability to account for real-time biomechanical changes during actual games. In addition, the reliance on medical data restricts the practicality of these methods due to accessibility challenges.

Baseball has long been a leader in data-driven analytics, a movement catalyzed by Bill James's seminal work, *Baseball Abstract* [13], and the subsequent evolution of Sabermetrics-a rigorous, evidence-based framework underpinned by advanced statistical modeling. Modern MLB games now generate comprehensive, high-resolution data for every pitch thrown, recording an array of metrics with unparalleled precision. Building on this wealth of information, we hypothesize that pitchers who undergo TJS exhibit distinct and identifiable patterns compared to those who do not. Leveraging these granular in-game data, this study seeks to predict and elucidate the underlying factors contributing to TJS, providing a data-driven foundation for early intervention and injury prevention.

We constructed a comprehensive dataset using the Pybaseball package in Python to develop deep learning models for injury prediction [14]. The dataset includes 5,537,981 pitches, encompassing 94 distinct pitching metrics, and is organized into 620 player cases classified as injury or non-injury. To predict injuries up to 100 days in advance, we implemented various DL models, including Long Short-Term Memory (LSTM), CNN-LSTM, Transformer-Encoder, ResNet, and Vision Transformer (ViT). Among these, the ViT model outperformed the others, achieving an F1-score of 0.73, demonstrating its effectiveness in handling complex multivariate time-series data.

For estimating the remaining days until an injury event, we employed a 1D-CNN regression model, which achieved a R^2 of 0.79. Furthermore, we utilized the explainable

AI (XAI) tool SHAP to interpret the regression model's outputs, identifying a downward release point of the four-seam fastball as a critical indicator of injury risk. This finding was kinetically validated, showing strong alignment with established injury mechanisms. Our results offer valuable insights by leveraging granular, real-world player data, confirming that DL model interpretations can align with existing biomechanical injury theories. These findings contribute to advancing the predictive and preventive strategies for pitcher injuries in baseball.

The key contributions of our research are as follows:

- This study is the first, to our knowledge, to specifically predict injuries leading to TJS in MLB pitchers using widely available in-game data, rather than focusing on general elbow injuries or relying on limited-access medical datasets.
- Our approach enables the prediction of injury risk up to 100 days in advance. By incorporating explainable AI (XAI) techniques, the study identifies critical changes in pitching mechanics that can inform early intervention and prevention strategies.
- XAI reveals that a lowered four-seam fastball release point signals impending injury, guiding preventive pitching measures.

Related work

Injury prevention is a critical priority for athletes, as injuries can significantly diminish performance, shorten careers, and escalate healthcare costs [15–17]. In baseball, elbow injuries requiring TJS are particularly severe, with extended recovery periods averaging 20.5 months [7] and substantial financial implications, including increased healthcare expenditures and rehabilitation costs [8, 18, 19]. Given the high stakes, researchers have pursued various strategies to mitigate injury risk, ranging from traditional medical and biomechanical approaches to modern, data-driven methods. These efforts aim to predict and prevent injuries more effectively, ensuring athletes' sustained health and performance.

Before looking at previous studies on traditional elbow injury research (Sect. 2.1) and more recent machine learning or deep learning approaches (Sect. 2.2), it is helpful to understand why injury prevention matters so much. By combining insights from medical science, biomechanics, and big data analysis, we can work toward tools that predict injuries early and guide prevention strategies. This not only helps athletes stay healthy and maintain a high level of performance but also benefits teams and leagues by reducing the overall impact of injuries.

Traditional studies about pitchers' elbow injury

Previous research explored the biomechanics of elbow injuries in professional baseball pitchers, mainly depending on non-game-related data. One study [9] analyzed preseason video recordings of 23 professional pitchers and monitored their elbow injuries over three seasons. The findings revealed a significant correlation between injury occurrence and specific pitching mechanics, such as elevated elbow valgus torque and high shoulder external rotation torque, particularly among players who sustained injuries or underwent surgery. Similarly, another study [10] investigated the 3D motion of 69 adult pitchers to identify biomechanical factors contributing to elbow injuries. This study highlighted six key variables, including elbow flexion, significantly influencing valgus torque on the elbow. Notably, changes in elbow flexion at the release point were shown to correlate with variations in valgus torque, underscoring the intricate relationship between pitching mechanics and the kinetic transfer of forces through the arm.

Medical imaging has also been employed to predict elbow injuries in pitchers. One study [11] utilized ultrasound to measure medial joint laxity in 70 professional pitchers during spring training and assessed their subsequent risk of UCL injuries. The findings indicated that pitchers with an ulnohumeral joint gap exceeding 5.6 mm in their dominant arm had a sixfold higher likelihood of UCL rupture during the season (AUC=0.77, p=0.02). These results suggest that ultrasound evaluation of UCL morphology could serve as a valuable tool for identifying pitchers at elevated risk of injury, thereby enhancing player assessments and informing targeted preventive strategies.

Another study [12] conducted a comparative analysis of 203 professional pitchers with and without TJS using ultrasound examinations. The results revealed that pitchers who underwent TJS displayed a significantly larger resting ulnohumeral joint space (median, 0.5 mm vs. 0.2 mm, $p = 0.006$) and a higher prevalence of hypoechoic foci (57.9% vs. 30.4%, $p = 0.03$). These findings further highlight the potential of medical imaging to identify structural markers associated with injury risk, offering actionable insights for early intervention.

The majority of prior research mainly focused on analyzing and predicting pitcher elbow injuries using non-game-related data, which fails to capture the dynamic, real-time mechanics and stressors encountered during actual competition. To address these limitations, we seek to utilize real-game data, enabling a more accurate assessment of injury risk. By doing so, we aim to guide effective injury management strategies and provide a deeper understanding of how in-game conditions impact pitcher elbow health.

ML/DL-based studies on pitchers' injuries

Several scholars utilized real-game data to predict pitchers' injuries using machine learning (ML) and deep learning (DL) techniques. For example, a study [20] analyzed game data from 4657 players, sourced from databases such as Baseball-Reference [21], FanGraphs [22], and Baseball Savant [23], to predict player placements on the 10-day, 15-day, or 60-day Disabled List in the following season. The study employed an ensemble model combining Random Forest [24], Logistic Regression [25], and XGBoost [26], which outperformed traditional single-model approaches, achieving F1-scores of 0.70 for hitters' injuries and 0.55 for pitchers' injuries. However, when focusing on more specific injury categories, such as elbow injuries, the model's performance significantly declined, with F1-scores of only 0.07 for hitters and 0.17 for pitchers. This underscores the model's limitations in predicting more granular injury types.

Another study [27] utilized a DL approach, analyzing broadcast footage of 20 MLB pitchers who were injured during the 2017 season. [28] trained an Inflated 3D ConvNet (I3D) model, initially designed to predict various types of pitcher injuries. When applied to predicting UCL tears, the model achieved F1-scores of 0.74 for left-handed pitchers and 0.69 for right-handed pitchers. Although the original evaluation focused on the final 20 pitches before an injury, an ablation study revealed that extending the analysis window to 30 pitches further improved the F1-score, indicating that incorporating a slightly longer pitch history can enhance the model's predictive accuracy.

Despite the success of these studies in utilizing real-game data, they present notable limitations. The ML-based study [20] faced challenges with low F1-scores when predicting specific elbow injuries. In contrast, the DL-based approach [27], while highly accurate, made predictions too close to the injury event (within 30 pitches), limiting its practical utility for proactive injury prevention. Additionally, these models did not interpret how their predictions related to the underlying kinetic factors contributing to injuries, leaving a significant gap in understanding the biomechanical drivers behind the results.

To address these challenges, this study introduces a DL framework that integrates both classification and regression tasks. The proposed approach aims to predict TJS injuries up to 100 days in advance and estimate the remaining time until injury, allowing for a more comprehensive analysis of the contributing biomechanical factors.

Dataset

Data source

In this study, we utilized MLB pitching data from the 2016 to 2023 seasons [23], characterized by the extensive use of '*Trackman*' technology to capture detailed pitch metrics. The data was gathered through the Pybaseball [14] package in Python, resulting in a dataset of 5,537,981 pitches with 94 distinct attributes, including metrics such as pitch velocity, spin rate, release angle, and pitch type. The dataset includes only regular-season games.

Players were categorized into two groups based on injury status: those who had undergone TJS and those who had not. Data for players who underwent TJS were obtained from a publicly available database [29], which has tracked Tommy John surgeries across both MLB and MiLB players since 2012. For the TJS group, data from their last two seasons before surgery were used. Only players with a minimum of two game appearances during those seasons were included. Similarly, non-injury players were selected if they participated in at least three out of their most recent four seasons, as shown in Table 2. In total, 620 samples were included, with 101 injury and 519 non-injury cases. The overall statistics for the pitches are summarized in Table 3.

Table 4 presents the pitch metrics used in the analysis. We choose to exclude data associated with simple fatigue factors (e.g., total pitch counts, player age), as these exhibit substantial individual differences and are not consistently tracked across official and unofficial (e.g., Spring Training) games. In particular, pitch counts can vary dramatically outside regular-season play, and players may attain starting roles at

Table 2 Normal vs. injured cases by number of seasons represented

Category	Normal			Injured		
Seasons represented	4	3	2	4	3	2
Cases	518	1	–	44	53	4
Total		519			101	

Table 3 Metrics comparison by handedness and injury status

Category	Normal		Injured	
Handedness	Left-handed	Right-handed	Left-handed	Right-handed
Cases (n)	149	370	25	76
Height (cm)	189.5 ± 4.6	189.2 ± 5.4	188.9 ± 5.0	188.5 ± 5.6
Weight (kg)	97.7 ± 8.9	98.5 ± 9.3	98.1 ± 9.7	97.6 ± 8.7
Fastball speed (mph)	92.6 ± 2.7	93.9 ± 2.5	92.8 ± 2.6	94.9 ± 2.2
Spin rate (rpm)	2222 ± 171	2289 ± 159	2222 ± 131	2315 ± 173

Data represent physical and performance metrics averaged over all players grouped by handedness and injury status

Table 4 These metrics, extracted from Pybaseball, are defined by Baseball Savant[14].

Pitching metrics	Description
Ax, Ay, Az	"The acceleration of the pitch in the X, Y, Z dimensions."
Effective_Speed	"Derived speed based on the pitcher's release (from hitter's perspective)"
Pfx_X, Z	"Movement in the X and Z dimensions from the catcher's perspective."
Plate_X, Z	"Position of the ball in the X and Z dimensions when it crosses home plate."
Release_Extension	"Release extension of the pitch in feet, as tracked by Statcast."
Release_Pos_X, Z	"Release position in the X and Z dimensions from the catcher's perspective."
Release_Speed	"Pitch velocity at the release point."
Release_Spin_Rate	"Spin rate of pitch."
Spin_Axis	"Spin axis in the 2D X-Z plane in degrees (from 0 to 360)."
Vx, Vy, Vz	"Velocity of the pitch in the X, Y, Z dimensions."

Table 5 Pitch types extracted from the Pybaseball data with less than 90% missing values

Pitch type	Description
CH	Change-up
CU	Curve
FC	Cutter
FF	Four-seam fastball
SI	Sinker (two-seam fastball)
SL	Slider

different ages, complicating the creation of a generalized model. Furthermore, prior research [30] indicates that TJS cases also occur among players under 18, suggesting that age alone may not be a reliable predictor in this context. By removing these highly individualized variables, we focused on more consistent mechanical metrics that better capture potential injury-related changes.

Table 5 summarizes the pitch types included in the study. By combining the pitch-specific metrics from Table 4 with the pitch types from Table 5, we created a single feature for each pitch type, resulting in a total of 102 features per pitcher. Using this dataset, we aim to assess whether injury risk can be predicted based on pitch-type-specific metrics.

Preprocessing

We pivoted the records extracted using Pybaseball by pitch type, allowing us to use pitch-type-specific metrics as features for each pitcher. To account for differences between left-handed and right-handed pitchers, we standardized all pitching metrics to align with those of right-handed pitchers.

Missing values in the pitching metrics were handled using two approaches. For pitchers with entirely missing metrics, we replaced the missing values with the average values across all pitchers. For partially missing data, we used linear interpolation to estimate the missing values. Furthermore, to account for individual variations even among pitchers in similar conditions, we applied differencing based on each player's average values from the first season (among the four seasons analyzed), rather than using absolute values.

Outliers were removed by excluding data points that deviated by more than $\pm 4.7\sigma$ from the mean for each pitching metric. To standardize the time-series data, we used the last game for each pitcher (for injured players, the last game before surgery) as the reference point. Since the pitching dates varied across pitchers, we reordered the prior pitching records in reverse chronological order relative to this reference date, ensuring that the time axis represented the number of days remaining until the last game. This approach resulted in time-series data with daily intervals, as shown in Fig. 2.

In the classification task, we reduced variation in each metric by aggregating the data into five-day intervals and using the average values. For the regression task, which predicts the number of days remaining until injury, applying the same aggregation method as in the classification task would have reduced the number of samples, complicating the training process. Therefore, to maintain an appropriate granularity for the target variable, we kept all time points but still maintained a five-day interval for the input data.

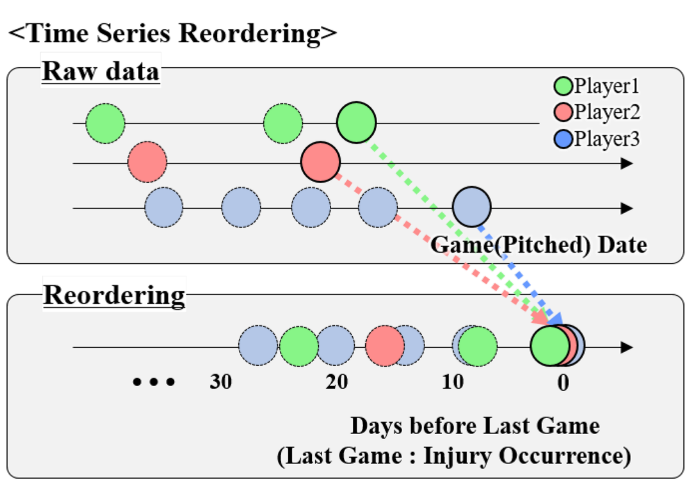


Fig. 2 Time-series reordering process. The last pitching date for both injured and non-injured groups is set as day 0, with prior pitching dates arranged in reverse chronological order relative to this reference point

Method

We employed a multivariate time-series dataset to conduct experiments using a range of DL models, including LSTM, CNN-LSTM, Transformer-Encoder, ResNet, ViT, and 1D-CNN. These models were selected for their ability to capture the complex, non-linear patterns and periodicity inherent in each player's data. Their architectures are well-suited to learning intricate dependencies across multiple variables over time, enabling more accurate and personalized pattern recognition.

First, we analyzed data up to 100 days prior to injury to classify injured versus non-injured players. The choice of a 100-day window was informed by practical considerations drawn from existing literature on the nonoperative treatment of partial UCL tears, where rehabilitation periods typically last around 3 months (approximately 12 weeks) for players to return to play without surgery [31, 32]. By identifying potential injury risks at least 100 days in advance, coaches, trainers, and medical staff have a sufficient timeframe to implement preventive strategies, adjust pitching workloads, or initiate early interventions, potentially reducing the likelihood of progression to a stage requiring surgical intervention.

Moreover, to predict the remaining days until injury, we developed regression models using a 1D-CNN with data spanning up to 550 days prior to the injury event. This extended timeframe was chosen based on preliminary experiments suggesting that incorporating longer historical data enhanced the model's predictive stability and accuracy.

Lastly, we applied XAI techniques in the regression analysis to explore whether pitching metrics correlate with injury occurrences in a biomechanical context. This approach not only deepens our understanding of the key contributing factors but also identifies the metrics with the most significant predictive influence when estimating the remaining time until injury. By combining a 550-day regression window with XAI methods, our framework provides a comprehensive perspective on early injury timeline estimation and offers valuable biomechanical insights that can inform preventive strategies.

Classification framework for TJS prediction

The dataset used in this study consisted of multivariate time-series data, with each pitcher's game schedule varying, leading to inherently irregular sampling intervals. While we resampled the data to regular intervals, some irregularities persisted due to differences in player availability and event frequency. To address the temporal nature of this data, we first employed well-established time-series models-LSTM, CNN-LSTM, and Transformer-Encoder. Specifically, LSTM targets long-term dependencies, CNN-LSTM combines local and temporal features, and the Transformer-Encoder's parallel attention mechanism captures extended interactions within the pitching metrics.

Considering the findings of prior research [33], which demonstrated that vision-based models can outperform traditional methods in handling irregularly sampled time-series data, we extended our approach to incorporate ResNet and ViT. For these models, we transformed the time-series data into single-channel images, with time on one axis and pitching metrics on the other. This transformation allows the vision models to interpret

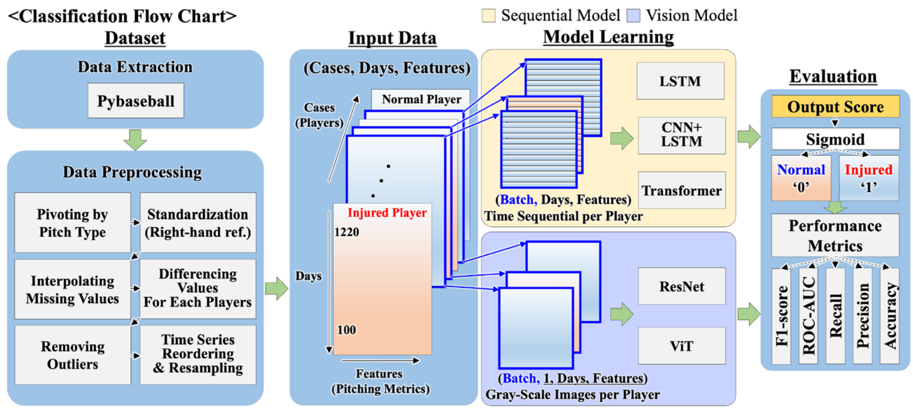


Fig. 3 Summary of the classification task

Table 6 Overview of model architectures and key configurations

Models	Configuration
ViT	Vision transformer (Patch16, input size 224), pre-trained on ImageNet
ResNet	ResNet-50 pre-trained on ImageNet
Transformer Encoder	512-dim. embeddings, 16 attention heads, 12 encoder layers
CNN+LSTM	Two 1D-CNN layers followed by a 4-layer LSTM (512-dim. embeddings)
LSTM	A 4-layer LSTM with 512-dim. embeddings

the data as a two-dimensional 'feature-time space,' enabling them to capture complex cross-feature interactions and subtle temporal patterns that may be difficult to detect using purely sequential architectures. In essence, while conventional models focus on the evolution of individual features over time, vision-based models can simultaneously exploit the relationships among multiple features across different time points, effectively uncovering intricate spatial (feature-to-feature) and temporal (feature-to-time) patterns within the data. Figure 3 illustrates this process, and Table 6 provides the detailed configurations for each model.

- *LSTM* [34]: LSTM, an extension of Recurrent Neural Networks [35], employs a memory cell and gating structure that allows it to effectively retain and update context information over extended sequences. This design addresses issues like the vanishing gradient problem, enabling the network to better preserve long-term temporal dependencies. In our approach, we adopt a six-layer, bidirectional LSTM architecture, allowing the model to incorporate information from both preceding and following time steps in the pitching sequence. Through this bidirectional structure, we aim to more comprehensively capture intricate temporal patterns, ultimately improving the accuracy of injury prediction.
- *CNN-LSTM* [36, 37]: A CNN-LSTM model integrates the spatial pattern recognition abilities of Convolutional Neural Networks with the temporal sequence modeling strengths of LSTMs. In our case, a 1D-CNN first distills key local patterns from the pitching metrics, and these condensed representations are then fed into LSTM layers to capture temporal dynamics. By combining these two perspectives, we aim

to detect nuanced relationships that span both feature space and time, ultimately enhancing the model's capacity to forecast injury risk.

- *Transformer* [38]: The Transformer-Encoder leverages self-attention to simultaneously consider all positions in a sequence, effectively revealing long-distance relationships without relying on step-by-step processing. This parallelized attention mechanism not only improves computational efficiency but also facilitates the discovery of subtle interactions among pitching metrics. By applying the Transformer-Encoder to our data, we expect a more refined interpretation of how multiple factors evolve and influence injury outcomes over extended periods, potentially leading to more accurate predictions.
- *ResNet* [39]: ResNet, a deep CNN architecture incorporating skip connections, enables stable training even at substantial depths, facilitating the extraction of intricate visual features. In our work, we represented the pitching time-series as single-channel images, then utilized a ResNet-50 model pre-trained on ImageNet [40] from PyTorch. This image-based encoding allows ResNet to identify subtle spatial patterns, potentially reflecting underlying biomechanical signals related to injury risks.
- *ViT* [41]: ViT, originally developed for image classification, segments an image into patches and employs self-attention to discover both granular and holistic patterns. In our implementation, we adapted the ViT Base Patch16 224 model (pre-trained on ImageNet [40] via the TIMM library) by converting the pitching time-series data into single-channel images and applying padding (−5) to standardize input dimensions. Through this patch-based, attention-driven analysis, ViT can potentially uncover delicate spatial relationships and subtle indicators of injury risk that conventional sequential models might overlook.

Regression framework for predicting days until injury

Since each player's pitch data is irregularly sampled, we initially tested LSTM, CNN+LSTM, and Transformer models for the regression task, using interpolation to handle missing time points, as done in our classification approach. However, after 10 runs with different random seeds, none of these sequential models achieved an R^2 score above 0.1, indicating fundamental limitations given our dataset conditions. We also explored vision-based architectures (ResNet, ViT) referenced from our classification framework but found them unsuitable for many-to-many regression tasks.

Drawing on insights from Sateesh Babu et al. [42], which emphasize the stability of convolutional approaches for sequence modeling, we adopted a single-channel 1D-CNN. Instead of modeling each player's time series separately, this approach aggregates all pre-injury data from injured players to capture generalizable patterns leading up to an injury. By predicting the number of days remaining until injury, our 1D-CNN provides a more robust quantitative understanding of how subtle variations in pitching metrics may signal an impending injury event.

- *1D CNN Regression* [42]: As shown in Fig. 4, our 1D-CNN regression framework processes time-series pitching metrics through four convolutional layers, each

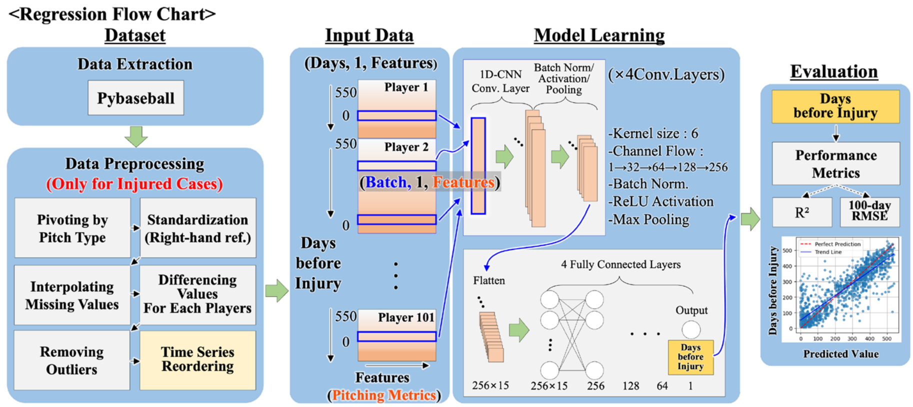


Fig. 4 Summary of the regression task

followed by batch normalization, a ReLU activation, and max-pooling. This standard sequence ensures stable training, enhances representational capacity, and reduces data dimensionality. After these convolutional blocks, the extracted features are flattened and passed through four fully connected layers to produce a single scalar output: the number of days remaining until injury. This design balances model complexity and interpretability, enabling the network to capture critical temporal patterns in the pitching metrics while maintaining computational efficiency. In doing so, our approach seeks to identify subtle signals that may precede the onset of injury.

Experiment

In this paper, we employed both classification and regression approaches to predict outcomes related to TJS for MLB pitchers. To evaluate the overall performance of the models in both tasks, we conducted experiments using 10 different random seeds.

All experiments were conducted with an NVIDIA RTX4090 GPU (22.5 GB memory) and an Intel(R) Xeon(R) CPU @ 2.20 GHz. We used Python 3.11.11 for training, implementation, validation, and testing procedures. It enabled the efficient handling of our extensive MLB pitching dataset and significantly accelerated the deep learning processes.

Classification task

Model training and loss function

For the classification task, the same random seeds were applied across all models to ensure consistency. The dataset was split into training, validation, and test sets with a 6:2:2 ratio, using stratified sampling to preserve the proportions of injured and non-injured cases.

To address the class imbalance in injury cases, we experimented with both a SMOTE-based upsampling approach [43] and a Weighted Binary Cross-Entropy (Weighted BCE) technique, given the relatively small number of injured players. In trials using our best-performing model (ViT), Weighted BCE improved the injured-class F1-score by at least 0.03 compared to both the standard (no upsampling) and SMOTE-based approaches, as shown in Fig. 5. This improvement led us to adopt Weighted BCE as our final loss function.

Specifically, we modified the binary cross-entropy (BCE) loss function, as shown in Eq. 1. To enhance the contribution of the minority class (injury cases), we introduced a weighting factor, defined in Eq. 2 [44]. This factor was proportional to the imbalance ratio, allowing the model to better account for underrepresented samples during training.

$$\mathrm{BCE}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right) \tag{1}$$

$$\text{Weighted BCE}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left(w_1 y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right) \tag{2}$$

- w_1: weight for the positive class(class *1* : Injured case)
- y_i: ground truth label for the i-th sample
- $\hat{y}_i$: predicted probability for the i-th sample
- N: total number of samples

Evaluation metrics

Considering our binary classification task involves imbalanced classes, with non-injured cases outnumbering injured cases by approximately 5:1, relying solely on accuracy would not provide a comprehensive assessment of model performance. To address this, we utilized multiple evaluation metrics that capture different aspects of predictive quality:

- *Precision*: Focuses on how many of the predicted positives are truly positive, thus indicating the model's ability to avoid false alarms. This is particularly important when a false positive (predicting an injury when there is none) has significant implications.

	ViT			ResNet		
	Weighted	SMOTE	Normal	Weighted	SMOTE	Normal
Count	10	10	10	10	10	10
Avg	0.73	0.64	0.70	0.64	0.60	0.63
StdDev	0.09	0.11	0.06	0.10	0.10	0.07

Fig. 5 Injured-class F1-score distributions for each classification model across three class-imbalance strategies: Non-sampling with weighted BCE (blue), SMOTE upsampling plus BCE (yellow), and non-sampling with standard BCE (red). Results are shown for 10 repeated runs per model

- *Recall*: Highlights the proportion of actual positives correctly identified, reflecting the model's effectiveness in detecting true injuries. A high recall suggests that the model misses fewer actual injury cases, which is critical in a scenario where missing an injury could lead to player health risks.
- *F1-Score*: Serves as a harmonic mean of precision and recall, providing a single-value measure of balance between false alarms and missed positives. The F1-score is well-suited for imbalanced data, as it ensures both dimensions of predictive performance are fairly represented.
- *Receiver operating characteristic-area under the curve (ROC-AUC)*: Evaluates the model's ability to distinguish between classes across various decision thresholds. A high ROC-AUC indicates that the model ranks positive instances higher than negatives most of the time, offering a threshold-independent view of discriminative power.

Hyperparameters for classification

A key aspect of improving our classification performance involved addressing data imbalance. After experimenting with both SMOTE and weighted BCE, we found that assigning higher loss weights (Pos Weight) to the minority class was more effective than upsampling. We first optimized the weight factor, determining that a value of 5-reflecting the approximate 5:1 imbalance ratio-yielded the best results. In addition, all models were standardized using MinMaxScaler and trained with the Adam optimizer, combined with a Cosine Annealing scheduler [45] to ensure stable convergence. We then conducted a grid search to fine-tune hyperparameters such as learning rates, batch sizes, and regularization terms, as detailed in Table 7. This process allowed us to identify the optimal configuration that maximized performance metrics while effectively mitigating data imbalance.

Regression task

Model training and loss function

For the regression task, the dataset of injured players was divided into training, validation, and test sets in a 6:2:2 ratio, aggregating all recorded dates without

Table 7 Summary of hyperparameter settings for the classification models

Models	Batch size	Learning rate	L1 Reg.	L2 Reg.	Pos weight
ViT	**16**,32,64,128	1.0e-6~5.0e-6	0~3.0e-6	0~3.0e-5	
		4.0e-6	**0**	**3.0e-5**	
ResNet	**16**,32,64	1.0e-5 ~ 8.0e-05	0~3.0e-6	0~1.e-05	1,**5**
		6.e-05	**1.0e-6**	**0**	
Transformer	32,**64**,128	1.e-06~5.e-06	0~3.0e-7	0~1.0e-6	
		3.0e-6	**1.0e-7**	**1.0e-7**	
CNN+LSTM	16,**32**,64	9.0e-7~5.0e-6	**0**	**0**	
		2.e-05			
LSTM	32,64,**128**	6.0e-6~1.0e-5	**0**	**1.0e-5**	
		8.0e-5			

The text in bold indicates the numerical values used in the experiment distinguishing between individual players. We employed the mean squared error (MSE), as shown in Eq. 3, as the loss function to train the regression model, with the goal of predicting the number of days remaining before an injury. This approach prioritized capturing general patterns across all available data points by treating each instance independently, thus simplifying the model.

$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

- y_i: ground truth label for the i-th sample
- $\hat{y}_i$: predicted probability for the i-th sample
- N: total number of samples

Evaluation metrics on regression

For the regression task, we evaluated model performance using two metrics: the Coefficient of Determination (R^2) and a custom 100-day accuracy measure. To ensure robustness and generalizability, we conducted 10 experiments per model.

- *R^2 (coefficient of determination)*: R^2 quantifies how much of the variation in the time-to-injury outcome is explained by the input features. A value approaching 1 indicates that the model closely fits the observed data, capturing key patterns that relate pitching metrics to the injury timeline. This measure provides a baseline sense of model adequacy, offering insight into its explanatory power.
- *100-day RMSE*: After the classification stage identifies players who are likely to be injured within 100 days, we employ the Root Mean Squared Error (RMSE) metric specifically for these players. By focusing on individuals whose injuries are imminent, the 100-Day RMSE quantifies how closely the model's predictions align with the actual remaining days until injury. A lower RMSE indicates more accurate and reliable predictions, providing a clearer basis for timely interventions and preventative strategies when it matters most.

Hyperparameters for regression

Preliminary experiments showed that a four-layer network architecture provided stable performance for our regression task. Based on this, we fixed the network depth to four layers and performed a grid search to optimize key hyperparameters, including the learning rate, batch size, and regularization terms. Moreover, all inputs were standardized using MinMaxScaler, and the model was trained with the Adam optimizer combined with a Cosine Annealing scheduler to ensure more consistent convergence. As shown in Table 8, this configuration effectively balanced model complexity while optimizing factors critical to improving regression accuracy.

Table 8 Summary of hyperparameter settings for the regression model

Model	Batch size	Learning rate	L1 Reg.	L2 Reg.	Dropout
1D-CNN	**16**,32,64,128	1.0e-4 ~ 3.0e-3 **1.0e-3**	0 ~ 1.0e-4 **0**	0 ~ 9.0e-6 **0**	**0**,0.1,0.2

The text in bold indicates the numerical values used in the experiment

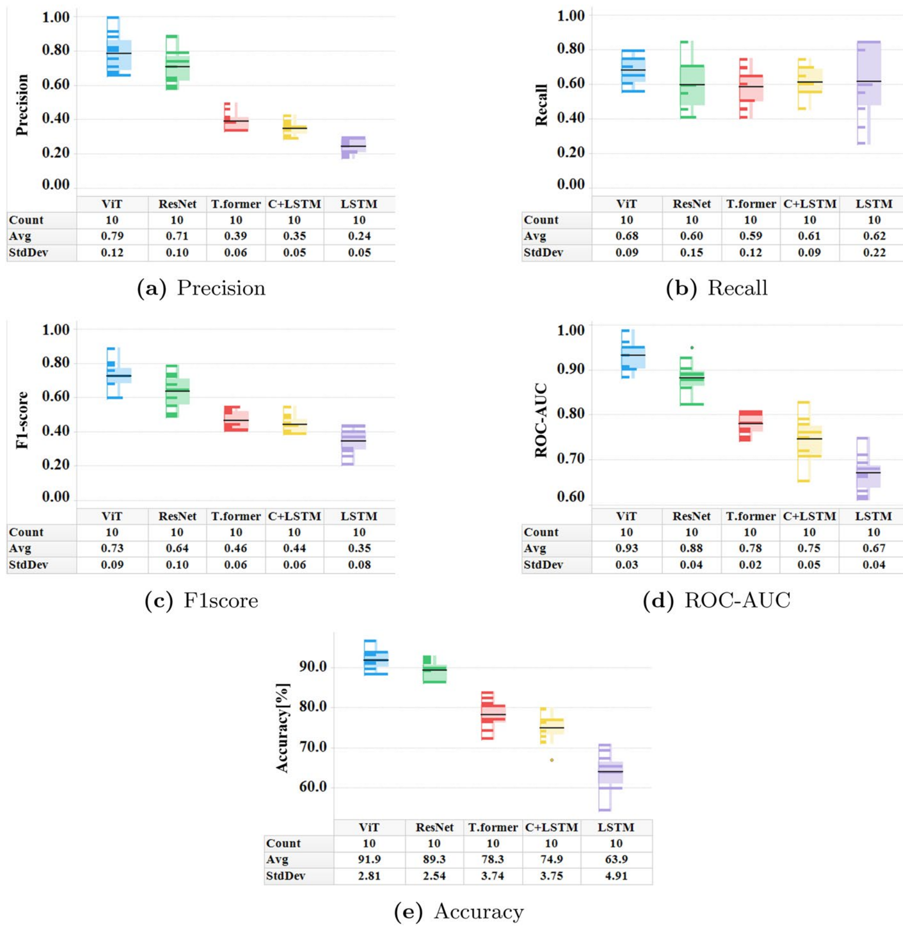


Fig. 6 Classification performance metrics (injured-class) of the 5 classification DL models across 10 experiments. **a** Precision, **b** Recall, **c** F1score, **d** ROC-AUC and **e** Accuracy

Results

Results on classification

Figure 6 and Table 9 present the results of the five classification models used in this study. Given the imbalance between injured and non-injured cases, we assessed performance using four key metrics: F1-score and ROC-AUC. The models, ranked in descending order of classification performance, were ViT, ResNet, Transformer-Encoder, CNN-LSTM, and LSTM. The ViT model achieved an average F1-score of 0.73 across 10 experiments, with a ROC-AUC of 0.93. These results suggest that, despite the small sample size of 620 and the imbalanced data, the ViT model reliably distinguishes between injured and non-injured cases. Additionally, the methodology proposed in previous research [33], which highlights the effectiveness of vision models in capturing both spatial and temporal features in irregularly sampled datasets, such as medical data, was also validated in our research.

Additionally, Fig. 7 presents the epoch-by-epoch training and validation loss curves for each classification model. Notably, the pre-trained vision models (ResNet, ViT) tended to converge more rapidly than the others, prompting us to reduce the batch size to allow for more fine-grained updates and potentially mitigate overfitting. By examining these curves, one can assess how quickly each model converges and detect any signs of instability or overfitting throughout the training process.

In contrast to ViT, ResNet-a vision-based model-demonstrates lower predictive capability, likely due to fundamental differences in their data-processing approaches. While ViT uses self-attention to capture complex, long-range dependencies by transforming time-series signals into patch-based representations, ResNet primarily relies on convolutional filters and residual connections. This convolutional framework is effective at extracting localized spatial features but struggles to capture the extended temporal and contextual relationships present in the transformed time-series data. In contrast, ViT's self-attention mechanism allows for a more comprehensive analysis by integrating both spatial and temporal information across patches. This broader perspective likely explains ViT's superior performance compared to ResNet.

However, when comparing the computational costs of these vision-based models in our classification task, ViT requires about three times the training time, seven times the inference time, and nearly four times the GPU usage of ResNet, as shown in Table 9. Despite these overheads, we find ViT to be a rational choice for the following reasons. First, it delivers over a 0.09 improvement in F1-score compared to the next best model. Second, although its training and inference times are indeed higher (roughly 3 × and 7× those of ResNet, respectively), using a pre-trained ViT rather than a from-scratch approach helps mitigate the training burden. Third, the total inference time for 124 players amounts to just 0.64 seconds, making it feasible in real-world scenarios-such

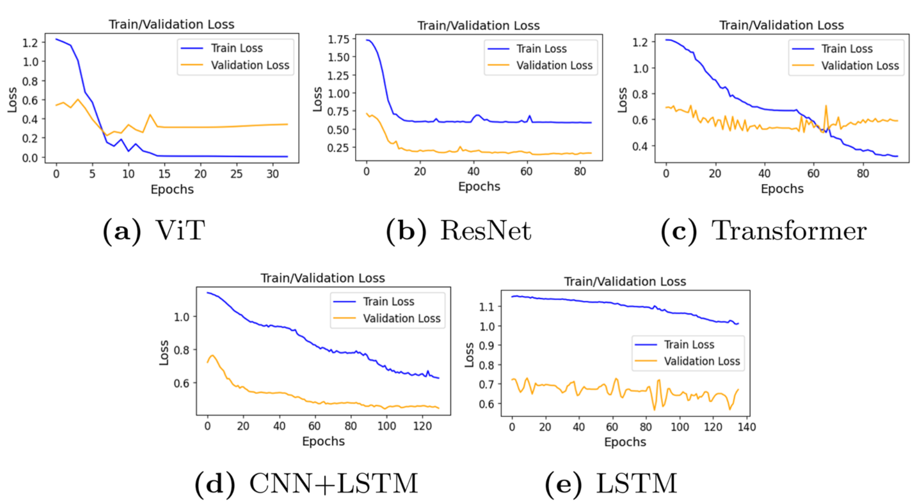


Fig. 7 Epoch-by-epoch training and validation loss curves for each classification model, illustrating stepwise performance development and convergence trends

Table 9 Model performance and resource usage metrics

Model	Class	Precision	Recall	F1-score	ROC-AUC	Acc.	Train time (s)	Inference time (s)	GPU usage (GB)
ViT*	**Normal**	**0.94**	**0.96**	**0.95**	**0.93**	**90.8**	217	0.64	5.18
	Injured	**0.79**	**0.68**	**0.73**					
ResNet*	Normal	0.93	0.95	0.94	0.88	89.3	61	0.09	**1.29**
	Injured	0.71	0.60	0.64					
T.former	Normal	0.91	0.82	0.86	0.78	78.3	355	0.36	6.77
	Injured	0.39	0.59	0.46					
CNN+ LSTM	Normal	0.91	0.78	0.84	0.75	74.9	**45**	**0.02**	2.79
	Injured	0.35	0.61	0.44					
LSTM	Normal	0.90	0.64	0.75	0.67	63.9	135	0.07	5.74
	Injured	0.24	0.62	0.35					

*Indicates a pretrained model All metrics are averaged over 10 experiments. The test set consists of 104 non-injured ("Normal") and 20 injured ("Injured") cases. "Train Time(s)" and "Inference Time(s)" refer to the total time in seconds for model training and inference, respectively, while "GPU Usage (GB)" indicates the maximum GPU memory consumption in gigabytes during training. Results are reported separately for each class to highlight class-specific performance. ViT consistently demonstrates superior predictive ability for injury cases, particularly with respect to F1-score and ROC-AUC, suggesting its effectiveness in distinguishing injury-related patterns

as running the model on a per-game basis. Consequently, we consider ViT's improved classification performance well worth the additional computational cost in practical MLB settings.

Figure 8 presents a visualization of the ViT model's attention map, highlighting the specific metrics the model focuses on when making predictions. This visualization emphasizes the interpretability of the model, demonstrating that it prioritizes certain features in predicting injury outcomes. However, these results alone do not establish a direct link between injuries and specific pitching metrics. This limitation prompted us to conduct additional experiments using a regression model to further investigate this relationship.

Results on regression

To assess the average performance of the regression model, we conducted ten runs of the 1D-CNN regression model, as shown in Fig. 9c. The R^2 value demonstrated an average correlation of 0.79. Additionally, the 100-Day RMSE was calculated to be 95.7 on average, as presented in Fig. 9a and b. Since all trials yielded meaningful R^2 values, we proceeded with an XAI analysis to gain a deeper understanding of how the model interprets the relationships between the days leading up to injuries and the pitching metrics.

Moreover, as shown in Fig. 9c, an inverse relationship was observed between the 100-Day RMSE and the R^2 value, indicating that higher R^2 values correspond to lower prediction errors for players within the critical 100-day window before injury. This suggests that the R^2 value is a key indicator of the model's ability to accurately predict injury risk for this high-risk subgroup of players.

Interpretation using SHAP values

SHAP [46] was selected over other XAI tools, such as LIME, due to its ability to provide both global and local explanations, as well as its foundation in game theory, which

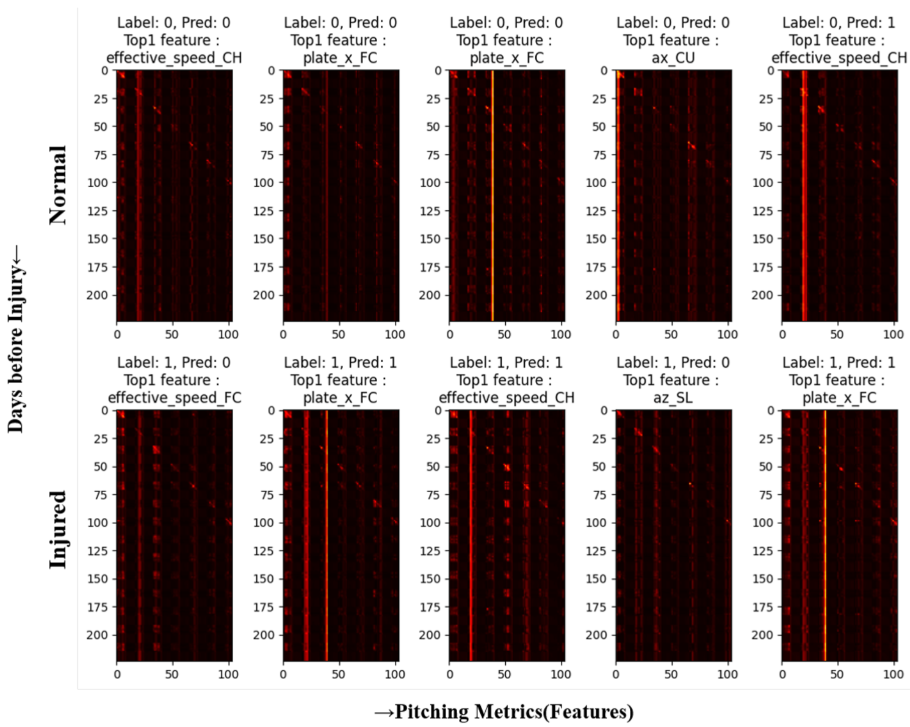
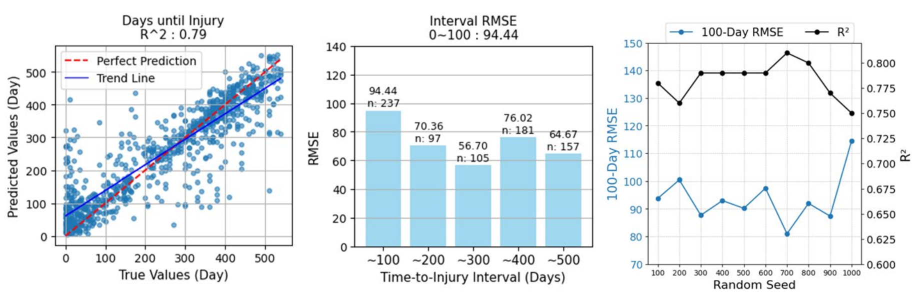


Fig. 8 Example of attention maps for class-specific predictions, with class *1* representing injured cases. Each attention map highlights the top feature emphasized during classification, offering insights into the model's focus on specific pitching metrics for injury prediction.



(a) R^2 (b) 100-Day RMSE (c) R^2, 100-Day RMSE

Fig. 9 Evaluation results of the model from 10 experiments. **a** R^2 value (0.79), **b** 100-Day RMSE (95.7), and **c** R^2 and 100-Day RMSE values. **a** and **b** represent the results of a model that demonstrated average performance across the 10 experiments.

allows for a comprehensive analysis of feature importance. While LIME excels at local interpretability, it is limited in capturing global relationships within the data [47]. For example, LIME cannot generate dependence plots, which are crucial for visualizing how feature values influence predictions across the dataset and for analyzing feature interactions.

SHAP, on the other hand, excels in this regard by providing dependence plots and other global interpretability tools, making it particularly well-suited for tasks that require a holistic understanding of the data. This global perspective is crucial for our study, as the objective is not only to predict TJS but also to identify key contributing factors across the dataset. These insights are essential for developing actionable strategies in player management and injury prevention.

Building on the global interpretability capabilities of SHAP, Gradient SHAP enhances the standard framework by incorporating gradient information from DL models [48]. By introducing slight variations in input values and measuring the resulting gradient responses, Gradient SHAP enables more accurate attribution of feature importance within complex, non-linear architectures. This gradient-based approach is particularly valuable for deep learning models, as it captures how subtle changes in input features influence the model's output. As a result, Gradient SHAP offers a nuanced and comprehensive perspective on feature contributions, even in high-dimensional spaces where intricate feature interactions can significantly affect predictive performance. For implementation details and further documentation, we direct the reader to the official SHAP GitHub repository [49].

We first conducted 10 experiments to identify the top 10 pitching metrics with the highest SHAP values for each model, as summarized in Table 10 and illustrated in Fig. 10. This analysis revealed 20 distinct features that consistently emerged as the most influential contributors. To verify that these 20 features genuinely played a critical role in enhancing model interpretability and predictive performance, we performed an additional 10 experiments using the same random seeds, but with training and evaluation restricted solely to these selected features. This validation

Table 10 Count of pitching metrics (features) selected among the top 10 SHAP values across 10 experiments

Pitching metric	Top10	Top5	Top3
Release_Speed_FF	**10**	**10**	**8**
Spin_Axis_FF	**9**	**10**	**10**
Release_Extension_SL	**9**	**9**	**6**
Release_Pos_X_SI	9	8	2
Spin_Axis_SI	9	2	0
Release_Spin_Rate_FF	8	5	2
Release_Speed_SL	7	3	1
Release_Extension_FF	4	1	1
Release_Pos_X_SL	4	0	0
Release_Pos_X_FF	4	0	0
Pfx_Z_CH	3	1	0
Release_Pos_X_CH	3	0	0
Spin_Axis_SL	3	0	0
Release_Pos_Z_CH	2	0	0
Release_Extension_CU	2	0	0
Release_Pos_Z_CU	1	1	0
Release_Pos_Z_SI	1	0	0
Release_Spin_Rate_SL	1	0	0
Vy0_SL	1	0	0
Release_Speed_CH	1	0	0

The table displays how frequently each feature was ranked within the top 10, top 5, and top 3 most important features

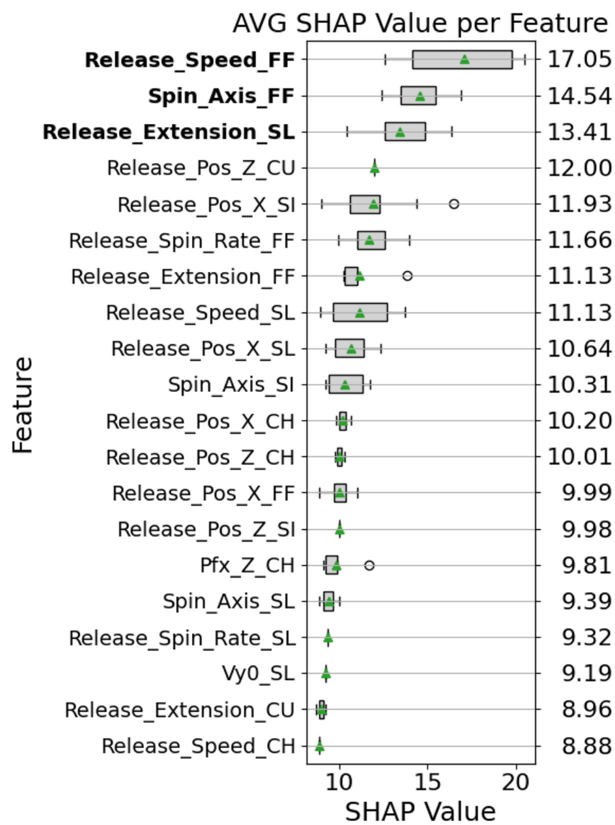


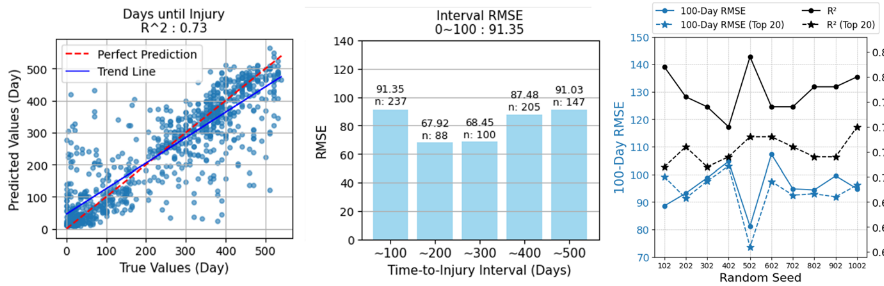
Fig. 10 Average SHAP values of each feature from 10 experiments. This bar chart highlights the relative importance of each feature, with higher SHAP values indicating a stronger influence on the model's predictions. Description of each feature is in Table 11

Table 11 Examples of pitching metrics for each pitch type were extracted from TOP 5 SHAP values in Fig. 10

Pitch type	Description
Spin_Axis_FF	Four-seam Fastball's spin axis(direction)
Release_Speed_FF	Four-seam Fastball's release speed
Release_Extension_SL	Slider's release extension from pitching rubber
Release_Spin_Rate_FF	Four-seam Fastball's spin rate in releasing position
Release_Pos_X_SI	Sinker's hoizental release position

step ensured that the features highlighted by the SHAP analysis were both stable and interpretable, reinforcing the credibility of our feature selection process.

By comparing the results of the original full-feature model with the model restricted to the top 20 SHAP-identified features, we observed only a minor decrease in R^2 (0.06) and a modest decrease in 100-Day RMSE (2.2), as shown in Fig. 11c. This relatively small performance gap suggests that the selected 20 features not only capture the model's most critical predictive signals but also preserve a strong level of accuracy. In other words, despite reducing the dimensionality of the input space, these features continue to capture the essential factors influencing injury prediction, reinforcing their validity as key predictive variables.



(a) R^2 (b) 100-Day RMSE (c) R^2, 100-Day RMSE

Fig. 11 Evaluation results of a refined model incorporating 20 key features derived from previous experiments. **a** R^2 value (0.73), **b** 100-Day RMSE (93.6), and **c** a comparative plot illustrating R^2 and 100-Day RMSE outcomes alongside those obtained using the full feature set. Results in **a** and **b** represent the model's average performance across the 10 experiments.

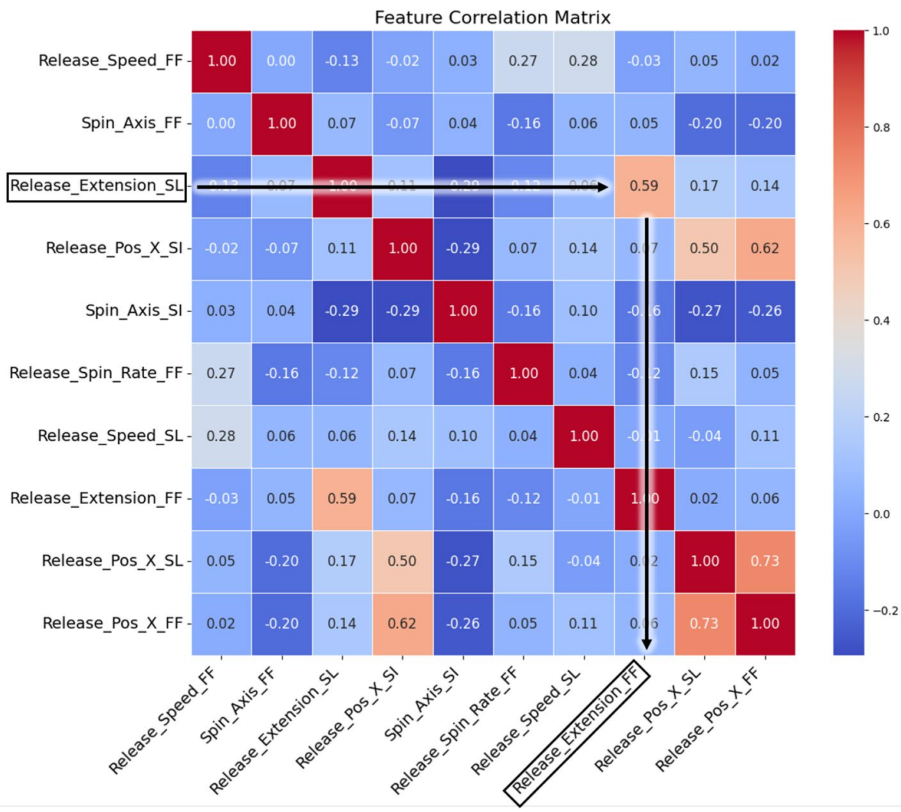


Fig. 12 Autocorrelation results for the 20 key features identified from the first 10 experiments. Among these features, Release_Extension_SL exhibits the strongest correlation with Release_Extension_FF

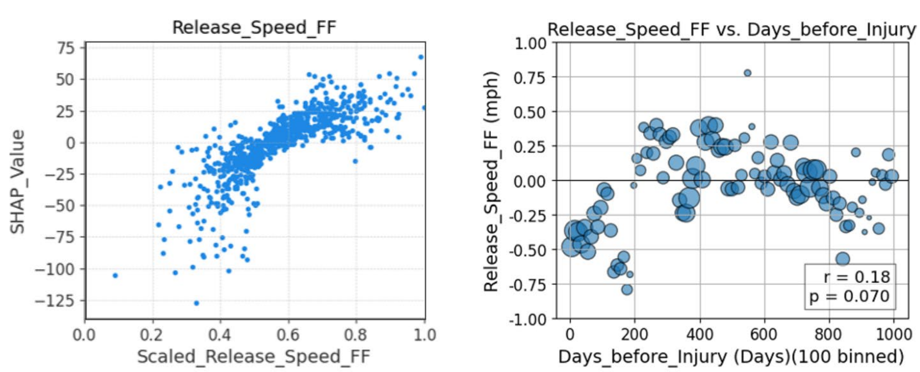
Among the 20 features, we identified three with significantly higher average SHAP values compared to the others, as shown in Fig. 10. Upon examining their autocorrelation matrices, as described in Fig. 12, we found that all three features (Release_Speed_FF, Spin_Axis_FF, Release_Extension_SL) were associated with the Four-seam Fastball. Based on these findings, we aim to interpret the results from two perspectives: the movement of the ball and the physical changes in the pitcher.

The movement of the ball

- *Release_Speed_FF*: This metric represents the speed of the four-seam fastball at the time of the pitch. The time-series plot in Fig. 13b shows a trend of decreasing values as the injury date approaches. This trend is consistent with the dependence plot as shown in Fig. 13a derived from SHAP values, indicating that the model interprets the ball's speed slowing as the injury event nears. As illustrated in Fig. 13b, the Pearson correlation analysis between Release_Speed_FF and the days remaining until injury (r=0.18, p=0.07) suggests a small but not statistically significant association at the 0.05 level [50]. Nonetheless, a noticeable decline in Release_Speed_FF was observed during the final pre-injury season, indicating that despite the non-significant p-value, this metric may still hold practical relevance for injury risk assessment.

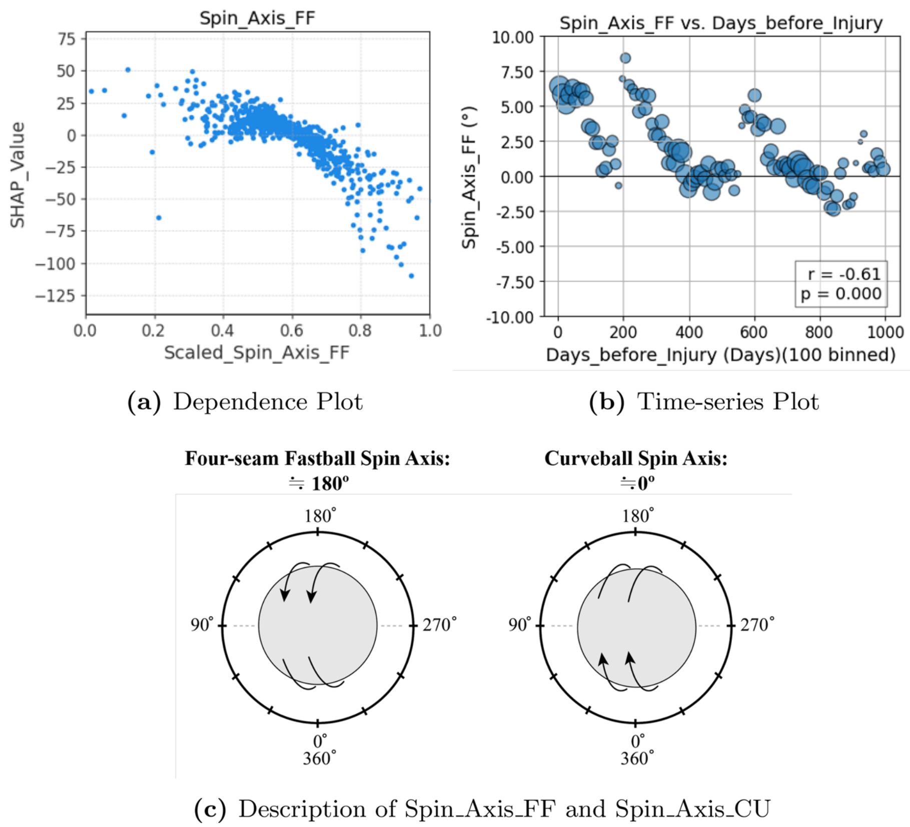
The physical changes in the pitcher

- *Spin_Axis_FF*: This metric represents the four-seam fastball's axis of rotation. Since all pitch metrics are derived from right-handed pitchers, an increase in this axis value indicates a more horizontal spin orientation for such pitchers, as illustrated in Fig. 14c. The dependence plot in Fig. 14a shows that as this metric increases, the remaining days until injury decrease. In addition, Fig. 14b confirms a rising trend in Spin_Axis_FF as the injury date approaches, culminating in a 6° increase by the final pre-injury game. A Pearson correlation analysis reveals a significant inverse relationship (r=−0.61, p <0.05) between Spin_Axis_FF and the days remaining until injury, indicating that this metric is meaningfully associated with impending injury. A more detailed interpretation of this finding is provided in Sect. 6.3.3.
- *Release_Extension_SL*: This metric represents the distance from the pitching rubber to the release point for a slider. The dependence plot in Fig. 15a suggests that an increase in this metric corresponds to a closer proximity to the injury event, and the time-series plot in Fig. 15b similarly indicates an approximately 0.2 feet increase as the injury date nears. A Pearson correlation analysis (r = 0.70, p < 0.05) further



(a) Dependence Plot (b) Time-series Plot

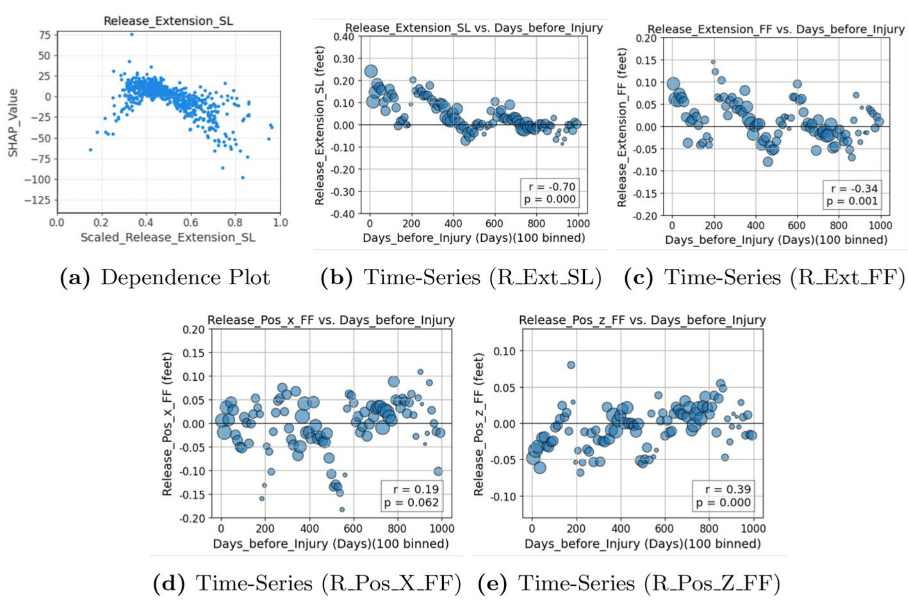
Fig. 13 SHAP value dependence plot and time-series analysis for Release_Speed_FF (**a**) and the binned (100) average time-series plot for days before injury (**b**)



(a) Dependence Plot **(b)** Time-series Plot

(c) Description of Spin_Axis_FF and Spin_Axis_CU

Fig. 14 SHAP value dependence plot and time-series analysis for Spin_Axis_FF. **a** Displays the dependence plot for SHAP values related to changes in Spin_Axis_FF. **b** Shows a binned (100) average time-series plot of Spin_Axis_FF. **c** Is a description of Spin_Axis. The Spin_Axis in the 2D X-Z plane in degrees from 0 to 360, such that 180 represents a pure backspin fastball and 0 degrees represents a pure topspin (12-6) curveball.



(a) Dependence Plot **(b)** Time-Series (R_Ext_SL) **(c)** Time-Series (R_Ext_FF)

(d) Time-Series (R_Pos_X_FF) **(e)** Time-Series (R_Pos_Z_FF)

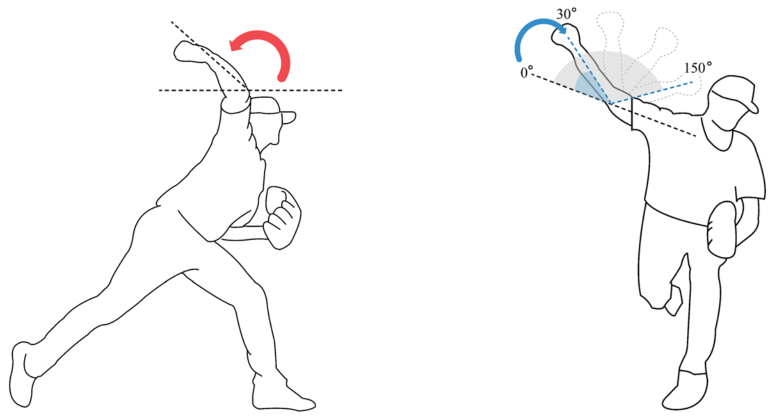
Fig. 15 SHAP value dependence plot and time-series analysis for Release Extension and Position metrics. **a** Shows the SHAP value dependence plot for Release_Extension_SL. **b–e** display binned (100) average time-series plots for Release_Extension_SL, Release_Extension_FF, Release_Pos_X_FF, and Release_Pos_Z_FF, illustrating how these metrics evolve as the injury date approaches suggests that higher values of this metric are associated with a closer injury date. Moreover, because the slider-related metric demonstrates a meaningful correlation ($R^2 = 0.59$) with the four-seam fastball metric (Fig. 12) and exhibits a similar time-series trend (Fig. 15b, c), we have chosen to analyze Release_Extension_FF instead of the original metric to maintain interpretive consistency with previously examined indicators (Release_Speed_FF, Spin_Axis_FF). In order to conduct additional analysis Release_Extension_FF, we further segmented the data using Release_Pos_x_FF and Release_Pos_z_FF as reference axes. As illustrated in Fig. 15d, there is no discernible trend along the x-axis release point, with a p-value of 0.62 indicating no statistically significant relationship. In contrast, Fig. 15e reveals that the time-series trend for Release_Pos_z_FF runs counter to what is shown for Release_Extension_FF in Fig. 15c. Moreover, the Pearson correlation analysis between Release_Pos_z_FF and the days remaining until injury yields $r = 0.39$ and $p < 0.05$, demonstrating a meaningful association in which Release_Pos_z_FF decreases as the injury date approaches. This observed pattern indicates that the pitcher may be lowering their forearm during the release, which could help explain the previously noted increase in Spin_Axis_FF. A prior study by Jinji [51] noted that while the trajectory of a four-seam fastball is influenced by the direction of the palm, the spin axis itself is not directly affected by other wrist-related kinematic factors. This implies that the arm's overall motion must be examined more closely. If the wrist does not substantially affect the spin axis at the release point, then the lowered Release_Pos_Z_FF found in this study's regression model may serve as indirect evidence that changes in arm positioning are contributing to the observed increase in Spin_Axis_FF.

Relating to TJS

In our study, we used a 1D-CNN regression model to predict the number of days remaining until injury for pitchers who had undergone TJS. Our findings revealed that as pitchers approached the injury event, the vertical release point of their four-seam fastball (Release_Pos_Z_FF) decreased, while their spin axis (Spin_Axis_FF) became more horizontal and release speed (Release_Speed_FF) decreased. These changes were interpreted within the broader context of elbow injuries. Specifically, alterations in a pitcher's mechanics-such as a decrease in the release point (Release_Pos_Z_FF) and changes in the spin axis (Spin_Axis_FF)-are presented as factors that increase valgus torque (Fig. 16a). This added stress on the ulnar collateral ligament (UCL) can progressively weaken it, leading to a loss of velocity (Release_Speed_FF). If these mechanical issues persist, the cumulative damage may ultimately require TJS.

Supporting this interpretation, a previous study [9] assessed 23 professional pitchers prior to spring training to identify injury risk factors. The study found significantly higher elbow valgus torque in pitchers who later sustained elbow injuries or required TJS compared to their non-injured counterparts (injured group: 91.6 N·m; non-injured group: 74.7 N·m; $p = 0.013$). Elevated valgus torque places additional stress on the UCL, increasing its susceptibility to injury.

Further evidence supporting this interpretation comes from another study [10] involving 69 collegiate and professional pitchers, which found a significant negative



(a) Elbow valgus torque during pitching. Valgus torque is the outward rotational force applied to the elbow joint, stressing the UCL.

(b) Elbow flexion during pitching. Flexion refers to the angle formed between the upper arm and the forearm at the elbow joint.

Fig. 16 Comparison of elbow valgus torque and elbow flexion during the pitching motion

correlation between elbow valgus torque and elbow flexion angle at fastball release ($r = -0.36$; $p = 0.04$), as shown in Fig. 16b. Specifically, a lower flexion angle at ball release was associated with increased valgus torque on the elbow joint. This suggests that when pitchers lower their forearm during the release phase, they experience greater valgus torque, further exacerbating stress on the UCL.

Integrating these findings, we propose that the observed decrease in Release_Pos_Z_FF as the injury approaches reflects a lowering of the forearm during the release phase. This mechanical adjustment likely reduces the elbow flexion angle at ball release. According to the studies mentioned, a decreased elbow flexion angle correlates with increased elbow valgus torque, which in turn elevates stress on the UCL. This heightened stress accelerates elbow damage, ultimately increasing the likelihood of injuries that may require procedures such as TJS.

Discussion and conclusions

This study proposed a method for predicting the risk of TJS among MLB pitchers using various DL models. The ViT model achieved the highest performance, with an F1-score of 0.73 and a ROC-AUC of 0.93, using real-time, game-derived data to make predictions up to 100 days before an injury event. These results highlight the potential of DL in capturing complex time-series data and identifying non-linear, periodic patterns in pitching metrics. Moreover, the 1D-CNN regression model demonstrated strong predictive power in estimating the days remaining until injury, achieving an R^2 of 0.79, indicating a robust alignment between predictions and actual outcomes.

Building on these results, we applied SHAP values from the regression model-which predicts the number of days remaining before an injury-to enhance interpretability. This analysis revealed notable shifts in key pitching mechanics: the vertical release point of the four-seam fastball dropped by approximately 0.2 feet, and its spin axis became about 6° more horizontal. These mechanical changes are closely linked to increased stress on the UCL, suggesting that even subtle alterations in pitching mechanics can serve as early indicators of heightened injury risk. By clarifying how specific aspects of pitching mechanics influence the likelihood of TJS, these findings provide a practical basis for early intervention strategies aimed at preserving pitcher health and potentially reducing the need for complex surgical procedures.

Limitations and future works

Despite these promising results, the study has several limitations. The dataset comprised only 620 cases, which may limit the generalizability of the models. The small sample size contributed to notable variance in key evaluation metrics, such as the F1-score, suggesting that the models are sensitive to data variability. Furthermore, because the dataset was limited to MLB pitchers, the findings may not fully generalize to other baseball leagues. This limitation highlights several avenues for future research. First, expanding the dataset to include pitchers from Minor League Baseball and international leagues would enhance the reliability and generalizability of the models, ensuring broader applicability across different competitive levels. Second, integrating richer, multimodal data–such as wearable sensor readings [10], biomechanical measurements [9], and actual game footage [27]–could provide a deeper understanding of how injury risks develop and evolve over time, offering a more comprehensive view of pitching mechanics and performance under real-world conditions. Finally, employing advanced explainable AI methods, including causal inference techniques[52], could help distinguish between factors that are merely correlated with injury risk and those that truly drive it. By pursuing these directions, future studies could develop more effective, evidence-based tools that enable coaches, trainers, and medical staff to intervene early and strategically, ultimately safeguarding pitcher health.

Ethical considerations The findings of this work introduce a new dimension to TJS management. From a managerial perspective, coaches can proactively identify potential injuries in pitchers and adjust pitch counts, pitch types, or even game appearances accordingly. However, this approach may negatively impact a pitcher's performance if the detection model inaccurately signals an injury risk, leading to premature removal from play when the pitcher is still capable of competing. Therefore, it is essential that our work is applied in real-life scenarios with careful consideration, recognizing that expert knowledge must be integrated into decision-making processes based on our model's predictions.

Real-time applicability and future directions Our regression model processes data on a day-by-day basis, which suggests that real-time predictions become feasible whenever new player data are updated (e.g., after each game). For instance, if the model identifies that a pitcher's risk window is under 100 days, trainers and medical staff could focus on the key metrics flagged by the model (e.g., four-seam fastball speed, vertical release point) and adjust upcoming training or rehabilitation sessions accordingly. However, the model does not currently learn from follow-up data related to post-injury care or long-term recovery strategies. In future research, integrating such information-along with automated suggestions for intervention-could substantially enhance the model's practical utility.

Clinical applicability and complementary diagnostics While our proposed method provides valuable insights into potential injury risks-particularly by highlighting metrics such as Spin_Axis_FF-these findings should be viewed as supportive rather than definitive. Final decisions regarding a player's readiness or need for further medical evaluation must ultimately reside with trainers, medical staff, and specialists. In particular, Spin_Axis_FF offers an additional clue not commonly addressed in previous research, potentially complementing established diagnostic techniques such as MRI scans. As a result, we recommend that sports practitioners employ our approach alongside traditional evaluations, ensuring that suspected injuries receive the thorough clinical investigation they require.

Author contributions
B. Kang contributed to the research's design, implementation, and analysis by examining the manuscript. A. P. del Pobil and E. Park supervised the research. B. Kang, M. Park, A. P. del Pobil and E. Park wrote and revised the manuscript. All authors approved the final version of the manuscript.

Funding
This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (RS-2023-00259497), the ITRC (Information Technology Research Center) program (IITP-2025-RS-2024-00436936), the Digital Field program (RS-2024-00419073), and the Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This paper describes research conducted at UJI Robotic Intelligence Laboratory. Support for this laboratory is provided in part by Universidad Jaume I and by Ministerio de Ciencia e Innovacion (PLEC2023-010360).

Availability of data and materials
The dataset and code used in this paper can be accessed at https://github.com/dxlabskku/TJS_Prediction

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 15 December 2024 Accepted: 25 March 2025
Published online: 11 April 2025

References
1. Eliakim E, Morgulev E, Lidor R, Meckel Y. Estimation of injury costs: financial damage of english premier league teams underachievement due to injuries. BMJ Open Sport Exerc Med. 2020;6(1):000675.
2. Premier League: English Premiar League (EPL) Official Final Ranking. 2023. https://www.premierleague.com/tables?co=1&se=578&ha=-1.
3. Major League Baseball: Major League Baseball (MLB) Official final ranking. 2023. https://www.mlb.com/standings/mlb/2023?tableType=regularSeasonExpanded.
4. Marigi EM, Conte S, Reinholz AK, Steubs JA, Knudsen ML, Krych AJ, Camp CL. Shoulder injuries in professional baseball batters: analysis of 3,414 injuries over an 8-year period. Arthrosc Sports Med Rehabil. 2022;4(3):1119–26.
5. Carr JB, McElheny KD, Corrigan A, Rowe D, Ma K, Curriero FC. The most common type, severity, and expected frequency of injuries vary by defensive position in professional baseball players. Am J Sports Med. 2022;50(9):2534–41.
6. Cain EL Jr, Dugas JR, Wolf RS, Andrews JR. Elbow injuries in throwing athletes: a current concepts review. Am J Sports Med. 2003;31(4):621–35.
7. Erickson BJ, Gupta AK, Harris JD, Bush-Joseph C, Bach BR, Abrams GD, San Juan AM, Cole BJ, Romeo AA. Rate of return to pitching and performance after tommy john surgery in major league baseball pitchers. Am J Sports Med. 2014;42(3):536–43.
8. Meldau JE, Srivastava K, Okoroha KR, Ahmad CS, Moutzouros V, Makhni EC. Cost analysis of tommy john surgery for major league baseball teams. J Should Elbow Surg. 2020;29(1):121–5.
9. Anz AW, Bushnell BD, Griffin LP, Noonan TJ, Torry MR, Hawkins RJ. Correlation of torque and elbow injury in professional baseball pitchers. Am J Sports Med. 2010;38(7):1368–74.

10. Aguinaldo AL, Chambers H. Correlation of throwing mechanics with elbow valgus load in adult baseball pitchers. Am J Sports Med. 2009;37(10):2043–8.
11. Shanley E, Smith M, Mayer BK, Bailey LB, Thigpen CA, Tokish JM, Kissenberth MJ, Noonan TJ. Using stress ultrasonography to understand the risk of ucl injury among professional baseball pitchers based on ligament morphology and dynamic abnormalities. Orthop J Sports Med 2018;6(8).
12. Hanna AJ, Popper HR, Sonnier JH, Erickson BJ, Jack RA, Cohen SB. The utility of stress ultrasound in identifying risk factors for elbow ulnar collateral ligament tear: a longitudinal study of 203 professional baseball players. Am J Sports Med. 2024;52(4):1060–7.
13. James B. Baseball abstract. 1977.
14. Introduction of Pybaseball. https://pypi.org/project/pybaseball/2.0.0/
15. Smith R, Lombardo DJ, Petersen-Fitts GR, Frank C, Tenbrunsel T, Curtis G, Whaley J, Sabesan VJ. Return to play and prior performance in major league baseball pitchers after repair of superior labral anterior-posterior tears. Orthop J Sports Med. 2016;4(12):2325967116675822.
16. Mody KS, Fletcher AN, Akoh CC, Parekh SG. Return to play and performance after anterior cruciate ligament reconstruction in national football league players. Orthop J Sports Med. 2022;10(3):23259671221079636.
17. Trofa DP, Miller JC, Jang ES, Woode DR, Greisberg JK, Vosseller JT. Professional athletes return to play and performance after operative repair of an achilles tendon rupture. Am J Sports Med. 2017;45(12):2864–71.
18. Conte S, Camp CL, Dines JS. Injury trends in major league baseball over 18 seasons: 1998–2015. Am J Orthop. 2016;45(3):116–23.
19. McMichael FO. Estimating economic impacts of tommy john surgery on major league baseball pitchers. PhD thesis, Montana State University-Bozeman, College of Agriculture. 2018.
20. Karnuta JM, Luu BC, Haeberle HS, Saluan PM, Frangiamore SJ, Stearns KL, Farrow LD, Nwachukwu BU, Verma NN, Makhni EC, et al. 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. Orthop J Sports Med. 2020;8(11):2325967120963046.
21. Baseball Reference: Baseball Reference. https://www.baseball-reference.com/
22. FanGraphs: FanGraphs. https://www.fangraphs.com/
23. Baseball Savant: Baseball Savant. https://baseballsavant.mlb.com/
24. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
25. Hosmer DW, Lemeshow S. Applied logistic regression. (No title), 2000;31.
26. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining. 2016;785–794.
27. Piergiovanni A, Ryoo MS. Early detection of injuries in mlb pitchers from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.
28. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;6299–6308.
29. Roegele J. Tommy John Surgery list. https://docs.google.com/spreadsheets/u/0/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/htmlview. 2012.
30. Griffith TB, Duralde XA. Successful performance after ulnar collateral ligament reconstruction: an analysis of 88 drafted professional baseball pitchers with a matched comparison cohort. Orthop J Sports Med. 2019;7(11):2325967119880820.
31. Podesta L, Crow SA, Volkmer D, Bert T, Yocum LA. Treatment of partial ulnar collateral ligament tears in the elbow with platelet-rich plasma. Am J Sports Med. 2013;41(7):1689–94.
32. Kato Y, Yamada S, Chavez J. Can platelet-rich plasma therapy save patients with ulnar collateral ligament tears from surgery? Regen Therapy. 2019;10:123–6.
33. Li Z, Li S, Yan X. Time series as images: vision transformer for irregularly sampled time series. Adv Neural Inf Process Syst. 2024;36.
34. Hochreiter, S. Long short-term memory. Neural Comput MIT-Press. 1997.
35. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1989;1(2):270–80.
36. Qiao Y, Wang Y, Ma C, Yang J. Short-term traffic flow prediction based on 1dcnn-lstm neural network structure. Mod Phys Lett B. 2021;35(02):2150042.
37. Sun H, Zhao S. Fault diagnosis for bearing based on 1dcnn and lstm. Shock Vibr. 2021;2021(1):1221462.
38. Vaswani A. Attention is all you need. Adv Neural Inf Process Syst. 2017.
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016;770–778.
40. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee;2009. 248–255.
41. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. 2021. https://openreview.net/forum?id=YicbFdNTTy
42. Sateesh Babu G, Zhao P, Li X-L. Deep convolutional neural network based regression approach for estimation of remaining useful life. In: Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I 21, (2016);214–228. Springer.
43. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
44. Aurelio YS, De Almeida GM, Castro CL, Braga AP. Learning from imbalanced data sets with weighted cross-entropy function. Neural processing letters. 2019;50:1937–49.
45. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017). https://openreview.net/forum?id=Skq89Scxx

46. Scott M, Su-In L, et al. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
47. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, Menegaz G. A perspective on explainable artificial intelligence methods: shap and lime. Adv Intell Syst. 2024;2400304
48. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International Conference on Machine Learning, 2017;3319–3328. PMLR
49. SHAP: Explainable AI. https://github.com/shap/shap/tree/master
50. J. Cohen, Statistical Power Analysis for the Behavioral Sciences. routledge, 2013.
51. Jinji T, Sakurai S. Throwing arm motion to determine spin axis of pitched baseball. In: ISBS-Conference Proceedings Archive 2006.
52. Breuer NO, Sauter A, Mohammadi M, Acar E. Cage: Causality-aware shapley value forÂ global explanations. In: Longo L, Lapuschkin S, Seifert C, editors. Explainable Artificial Intelligence. Cham: Springer; 2024. p. 143–62.

Publisher's Note