# Predicting NBA Player Injuries: A Data Driven Approach

**Emirhan Utku**[1] **Süleyman Yolcu**[1]

## Abstract

Injuries are a pervasive concern in professional sports, with the National Basketball Association (NBA) offering a unique context for examining this issue due to its high intensity and physical demands. This project aims to develop a predictive model for NBA player injuries using machine learning techniques. By integrating historical injury data with player performance metrics, we strive to identify patterns and factors that contribute to injury risk. Our preliminary findings suggest that decision tree models show promise in accurately classifying injury-prone players, thereby paving the way for improved injury prevention strategies, enhanced player longevity, and more informed team decision-making.

## 1. Introduction

Injuries in the National Basketball Association (NBA) significantly impact players, teams, and the league as a whole. From a player's perspective, injuries can derail promising careers and shorten playing longevity. For teams, unanticipated injuries to key athletes can drastically alter season trajectories and championship aspirations. At the league level, high-profile injuries can affect viewership, revenue, and the overall quality of competition. Consequently, the ability to predict which players are more susceptible to injuries and when they might occur has the potential to revolutionize player management strategies, reduce downtime, and improve preventative care. This project aims to develop a machine learning (ML)-based framework to predict NBA player injuries using historical injury records combined with in-game performance metrics. By identifying patterns and correlates of injury risk—such as workload, minutes played, player physical attributes, or specific types of injuries—coaches, trainers, and team decision-makers could proactively manage player rest, recovery protocols, and conditioning programs. Ultimately, the goal is to enhance player well-being, guide strategic rotational decisions, and improve fan engagement by maintaining healthier rosters throughout the season.

## 2. Related Work

Several studies and projects have explored injury prediction and prevention in professional sports through ML techniques. One pertinent work is the application of machine learning to NBA player injuries, leveraging algorithms like Random Forest and Gradient Boosting to highlight correlations between player attributes (e.g., height, weight, position) and injury proneness. This line of research often emphasizes the importance of player anthropometrics and workload indicators as key risk factors for injuries [1]. Although some literature focuses on other sports, such as soccer, the methodologies are often transferable. For example, studies examining the relationship between training load and injury occurrence in football suggest that ML can effectively identify overtraining thresholds and predict the onset of injuries using ensemble methods [2]. These insights are relevant to basketball, where intense and frequent gameplay similarly contributes to overuse injuries. Our project aligns with these works by merging diverse data sources—longitudinal injury data and detailed player performance statistics—and applying ML algorithms to predict injury risk. In extending these methods to the NBA context, we seek to refine the predictive accuracy and interpretability of models used for preventive strategies.

## 3. Methodology

### 3.1. Datasets

We utilized two primary datasets sourced from Kaggle:

1. **NBA Injuries 2010–2020:** Containing historical injury records from 2010 to 2018, this dataset includes the timing, frequency, and nature of injuries. Although we initially encountered notes including non-injury-related statuses (e.g., "activated," "returned"), we filtered these to focus exclusively on genuine injury instances. [Link to dataset]

2. **NBA Player Stats and Injured Data (2013–2023):** This dataset provides comprehensive player statistics, including minutes played, points, rebounds, assists, average speed, distance covered, and more. By merging these performance metrics with injury data, we aim to uncover meaningful correlations between workload and injury risk. [Link to

dataset]

### 3.2. Data Preprocessing

**1. Data Cleaning:** We removed rows containing vague or irrelevant notes (e.g., "placed on IL" without specifying the injury nature). Non-injury events such as activation or clearance to play were excluded.

**2. Feature Engineering:** We extracted injury-related keywords from the "Notes" column to categorize injuries by type (e.g., knee, ankle, hamstring). We also introduced a binary variable indicating severe injuries (players "out indefinitely") and grouped granular injury types into broader categories (head, upper body, lower body) for more robust pattern detection.

**3. Time and Seasonal Context:** We mapped injuries to their corresponding NBA seasons and excluded years outside the intersection of both datasets. This temporal alignment ensures that injury records correspond to the appropriate season-level player statistics.

**4. Merging Datasets:** After filtering out early seasons (2010–2012) not covered in the player stats dataset, we merged the two datasets on key fields (player name, team, and season). This resulted in a unified table containing injury indicators alongside performance metrics.

### 3.3. Exploratory Data Analysis (EDA)

We conducted extensive exploratory data analysis (EDA) to better understand the distribution and patterns of injuries:

**1. Common Injury Types:** Knee and ankle injuries emerged as the most frequent, underscoring the lower body's vulnerability in basketball due to repetitive jumping, cutting, and pivoting.

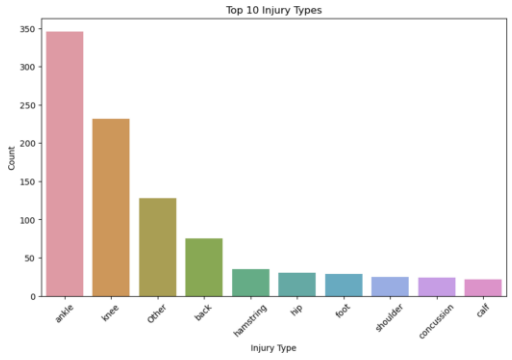


*Figure 1.* Top 10 Injury Types observed in the dataset.

**2. Yearly Trends:** Injuries generally increased from 2013 to 2019, as seen in Figure 2. This increase could be attributed to evolving playstyles, such as a faster pace of play and increased intensity, as well as improved monitoring and reporting of injuries. The slight drop in 2020 may be explained by the COVID-19 pandemic, which interrupted the regular season and led to fewer games being played, thus reducing the overall injury count.

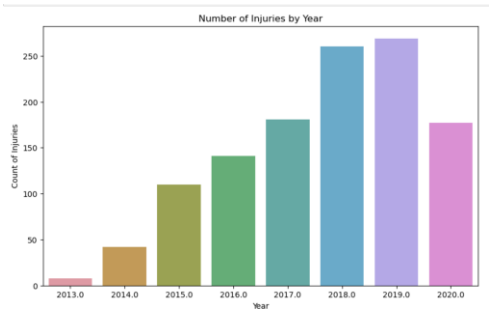


*Figure 2.* Number of injuries by year

**3. Severity Analysis:** As shown in Figure 3, knee-related injuries have the highest likelihood of resulting in indefinite absences, followed by ankle injuries. These findings underscore the critical nature of knee injuries in basketball and highlight the need for targeted interventions, such as improved knee support and conditioning programs. Other injury types, such as shoulder and foot injuries, also contribute significantly to long-term absences but at a lower frequency compared to knee and ankle injuries. This analysis reinforces the importance of proactive injury prevention measures focusing on these high-impact areas.

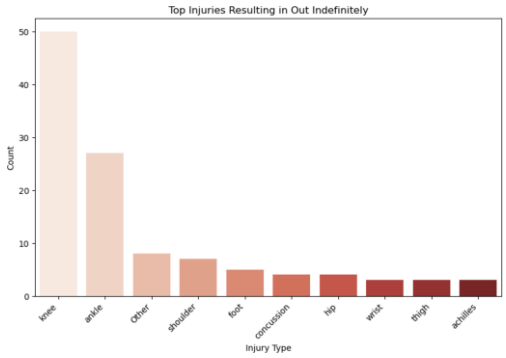


*Figure 3.* Top injuries resulting in indefinite absences.

**4. Injuries by Body Region:** Figure 4 displays the distribution of injuries across body regions, highlighting the legs as the most frequently injured part (713 occurrences), reflecting the high physical demands of basketball. Other significant areas include the back (76), foot (55), and hands (50), all linked to repetitive movements and physical contact. Facial and head injuries (64 combined) are less frequent but still notable, often resulting from collisions. The visualization emphasizes the need for targeted preventive measures, especially for the lower body and high-impact areas.

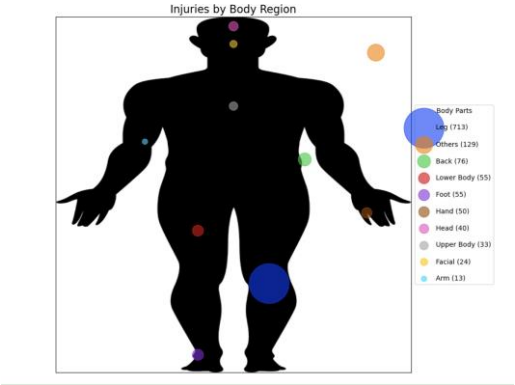


*Figure 4.* Injuries by body region

## 4. Experimental Evaluation

**Modeling Approach:** We developed models to predict whether a player would be "injury-prone" (1) or "not injury-prone" (0) during a season. This was based on historical injury records and performance metrics. The dataset was split into training (80%) and test (20%) sets, maintaining class stratification to ensure balanced evaluation.

**Algorithms Tested:**

1. *Logistic Regression:*

   **Performance:** This model achieved a baseline accuracy of **76.37%**. However, it struggled with recall for the injured class (**32%**), often biasing towards the majority class (not injured). Its weighted F1-score of **73%** indicated room for improvement in sensitivity for minority classes.

2. *Decision Tree Classifier:*

   **Performance:** The Decision Tree significantly outperformed Logistic Regression, achieving an accuracy of **86.67%**. It demonstrated a recall of **91%** for the injured class and a weighted F1-score of **87%**. The model's interpretable structure provided insight into factors contributing to injuries.

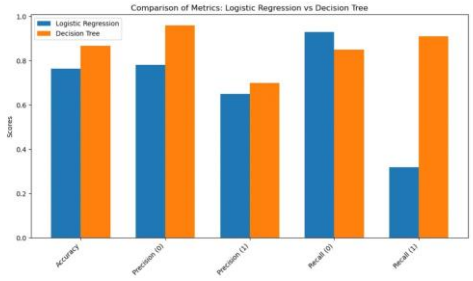


*Figure 5.* Classification metrics comparison

**Preliminary Results:**

- **Accuracy and Recall:** The Decision Tree model outperformed Logistic Regression across all metrics, particularly in recall for the injured class, highlighting its capability to identify injury-prone players effectively.
- **Feature Importance:** Tree-based models revealed key predictors of injuries, including playing time (*minutes played*), physical performance metrics (*average speed*, *distance covered*), and anthropometric attributes (*height*, *weight*).

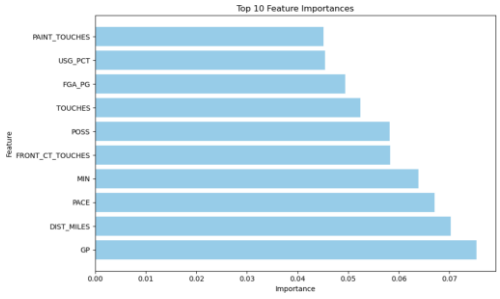


*Figure 6.* Top 10 features influencing injury predictions based on the Decision Tree model.

**Next Steps in Evaluation:**

- **Cross-Validation and Metrics:** We plan to incorporate cross-validation to ensure model robustness and evaluate performance using additional metrics such as F1-score and ROC-AUC.
- **Model Refinement:** Future experiments will explore ensemble methods, such as Gradient Boosted Trees, and advanced feature engineering (e.g., rolling averages of workload metrics) to enhance predictive accuracy.
- **Practical Implications:** Beyond predictive accuracy, interpreting model outputs will guide actionable strate-

gies like load management and tailored training programs for players at higher risk of injuries.

## 5. Conclusion

This progress report documents our efforts in building a machine learning system to predict NBA player injuries. From initial data cleaning and feature engineering to exploratory analyses and the development of baseline predictive models, we have laid the groundwork for a more refined and accurate predictive tool. Preliminary experiments show promising results with tree-based models, surpassing simpler linear classifiers in capturing complex injury patterns.

Our subsequent efforts will focus on further tuning model parameters, incorporating more nuanced features, and ensuring robust model evaluation. By doing so, we aim to approach a practical and reliable injury prediction framework. Ultimately, this project seeks to provide actionable insights to sports scientists, medical staff, and team management—fostering healthier and more competitive NBA seasons.

## References

1 Leveraging Machine Learning to Predict National Basketball Association Player Injuries. IEEE Xplore.Link

2 Machine Learning for Understanding and Predicting Injuries in Football. Sports Medicine — Open.Link