CS5823 Project Report

Issac Alvarez EFL646

Abstract

In professional sports, player injuries can dramatically affect team performance and player careers. This project aims to develop an explainable AI model to predict the probability of a professional NBA player getting injured using historical data and play style features. The motivation stems from being able to create a transparent tool that can help coaches and sports analyst undertand not just when an injury might occur but why. Combining various datasets from injury records to game level statistics such as (but not limited to) minutes played, fouls drawn, and shot proximity to defenders in order to generate a player risk profile. Using XG-boost and SHAP & LiME for model interpretability, we highlight key contributors to injury risk in a way that is both data driven and understandable for non technical users. The preliminary findings suggest that aggressive playstyles and high usage rates can correlate with elevated injury probabilities. Future work will explore time series modeling to incorporate fatigue accumulation and game context with the goal of building a better injury risk monitoring tool.

float

1 Introduction

In the world of professional basketball, player injuries can have significant consequences for team performance, player longevity, team revenue, and fan engagement. Predicting injuries is a complex task due to the unpredictability but also dynamic nature of in game events, player roles, play styles, and physical conditions. While machine learning models have been used to predict performance metrics, there isn't a publicly available model to find the likelihood of a player getting injuries, more particularly in a way that provides actionable and interpretable insights.

The scale of the problem is clear when looking at current NBA data. As of the 2024–2025 season, teams like the Indiana Pacers and Milwaukee Bucks have seen as many as 10 injured players each. More than half of NBA teams have had at least 8 players injured, emphasizing how widespread and routine injuries have become across the league. However, it's not just about the number of players affected. When considering severity, teams like the Utah Jazz and New Orleans Pelicans lead the league in total days lost to injuries, each surpassing 400 missed days. This reflects not just frequency but the long-term impact certain injuries can have on a roster.

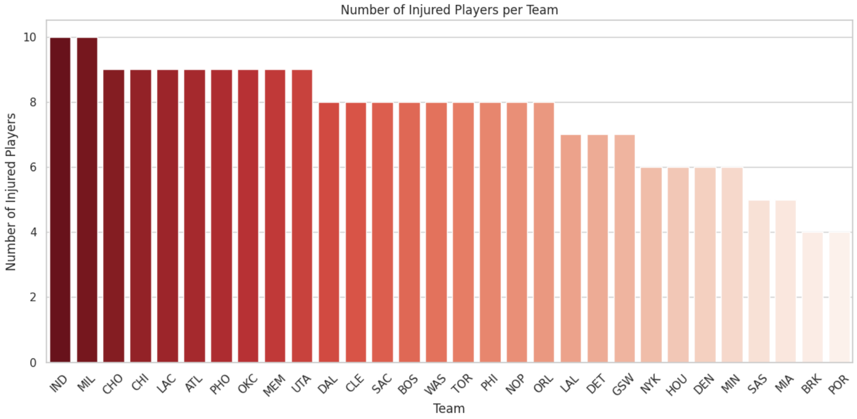


Figure 1: Number of Injured Players and Total Missed Days by Team

At the player level, individual cases further highlight the issue. Taylor Hendricks and De Anthony Melton have each missed over 150 days this season alone. Others like Ben Simmons, LaMelo Ball, and Ja Morant have dealt with repeated injuries logging over 10 distinct incidents each. These patterns suggest not only high physical demand but potential underlying risk factors that need to be modeled and understood.

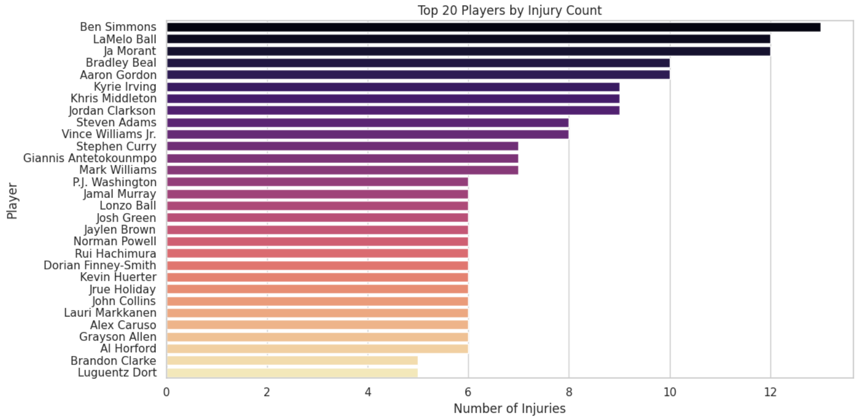


Figure 2: Top 20 Players by injury count

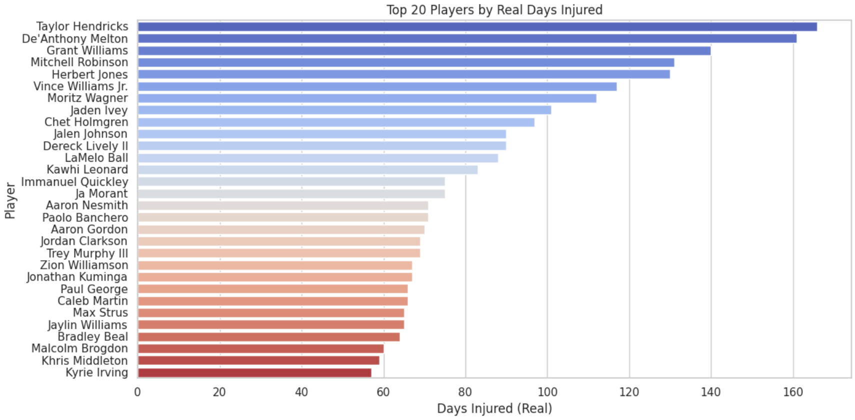


Figure 3: Top 20 Players by Days Injured

This project focuses on predicting the likelihood of NBA player injuries using historical injury logs combined with per game performance statistics. Features included physical exertion (ex. games played, minutes played), and play style (ex points per game, fouls drawn, contested shots), and historical injury patterns. Due to the vast amount of data available, and keeping in mind advances in sports medicine, the decision was made to only focus on the current 2024-2025 NBA season as to have the most accurate data possible.

Predictions alone are not sufficient. In high impact domains like sports medicine and performance management, explainability is crucial. Coaches, analysts, and medical staff need to understand not only which players are at risk but why a model makes such predictions. This is where XAI becomes essential. By leveraging tree based models like XG Boost and explainable methods such as SHAP and LiME, the aim is to provide transparent, interpretable insights into the factors contributing to injury risk.

2 Models

The objective of this project was to predict the severity of injury risk for NBA players based on performance statistics, biometric features, and historical injury records. This was formulated as a multi-class classification task with four possible classes: None, Minor, Moderate, and Severe, based on the number of days a player was injured.

The approach for this project was to train a custom XGBoost (Extreme Gradient Boosting) model from scratch, without using any pre-trained models. XGBoost was selected due to its strong performance on structured datasets, its ability to handle missing data, and its support for multi-class classification via softmax probability outputs.

The model was trained using a resampled version of the dataset with SMOTE to address class imbalance, particularly the under representation of the “Severe” injury class. The dataset was split into training and testing sets using stratified sampling to preserve class distribution. I performed manual hyperparameter tuning on learning rate, tree depth, and number of estimators to optimize performance.

Model performance was evaluated using precision, recall, F1-score, and a confusion matrix, with particular emphasis on the model’s ability to correctly identify moderate and severe injury cases. To enhance interpret-ability, SHAP, and LiME were used to visualize the contribution of features to each prediction to offer transparency at both the global and individual player level.

3 Explanation Methods

To interpret the predictions made by the injury severity scores, two popular explanation methods were utilized: SHAP and LIME.

3.1 SHAP

SHAP was particularly suited for this project because it is optimized for tree-based models like XGBoost and ensures consistency in feature attributions.

SHAP generated both global summary plots (showing overall feature importance) and local explanations for individual players using waterfall plots. These explanations highlighted which player statistics (such as minutes played or past injury frequency were most responsible for a specific injury risk prediction)

3.2 LIME

LIME was used to generate intuitive, local explanations for individual predictions. This model helped explain the prediction logic in a format that is accessible to non technical audiences.

4 Experiments

This section should describe datasets, research problems, results, and observations.

4.1 Datasets

The project relied primarily on two custom-compiled datasets: 1. Pro Sports Transaction Dataset. 2. NBA Player Statistics Dataset.

4.1.1 Pro Sports Transaction Dataset

This dataset was scraped from ProSportsTransactions.com and contains historical injury records for NBA players dating back to the league's inception. For this project, the focuses was on the 2024–2025 NBA season. Each record includes: - The date a player was placed on or removed from the injured list - Descriptive notes indicating the injury category or nature - The player's name and affiliated team at the time of injury

The raw injury log was cleaned to resolve inconsistencies in player names and to remove ambiguous or inactive records. Injury durations were calculated by subtracting return dates from injury start dates. Special handling was implemented for long-term injuries ("out for season") where return dates were not provided.

4.1.2 NBA Player Statistics

This dataset was created by scraping and aggregating season averages and game-level statistics for NBA players from public sports data websites. The focus on the top 10 players from each team based on performance rankings, resulting in approximately 300 players. This decision was made since typically the top 10 players from each team are the players who will offer significant statistics available. Each player's record includes:

-Biometric data: age, height, weight

-Performance stats: points, minutes, field goal percentage, rebounds, assists, steals, blocks, turnovers, fouls, etc.

-Historical injury frequency

The two datasets were merged by player name, and features such as total injuries, injury categories, and total days injured were added to the player statistics dataset.

4.2 Preprocessing

To prepare the data for predictive modeling, several preprocessing steps were applied, including cleaning, merging, feature engineering, and most importantly, labeling injury outcomes. Since predicting whether an injury will occur at all is too binary and simplistic given the complexity of basketball injuries, the data was labeled by injury severity and injury type, allowing for a more simple and medically relevant prediction task.

4.2.1 Labeling Injury Severity

Each player in the dataset was categorized into one of four severity classes based on the number of games missed due to injury over the 2024–2025 season:

None: 0 games missed

Minor: 1–3 games missed

Moderate: 4–9 games missed

Severe: 10 or more games missed

As shown in the chart titled Injury Severity Distribution Among Top 300 Players, the majority of players fell into the Severe category (135 out of 300), highlighting just how common long-term injuries are in the NBA. Meanwhile, only 73 players had no reported injuries this season, suggesting the importance of robust injury modeling in nearly every roster.

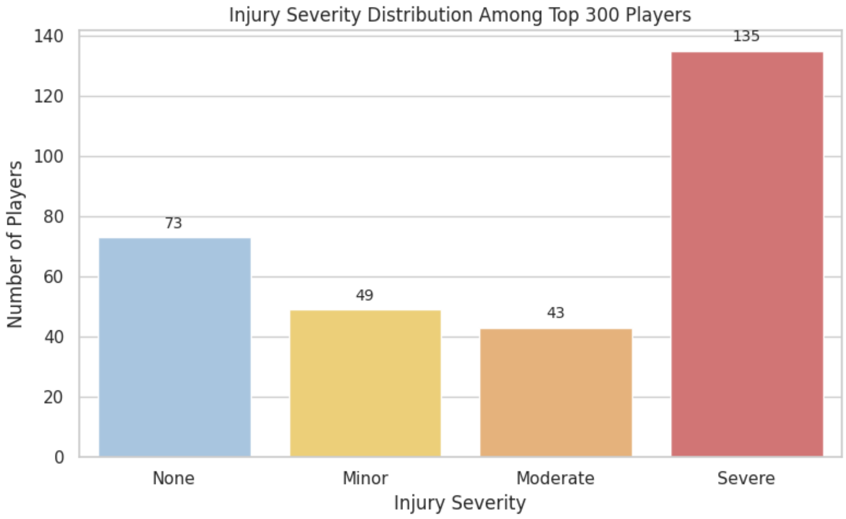


Figure 4: Injury severity among top 300 players

4.2.2 Physical Attributes vs Severity

To explore whether injury severity correlated with physical traits like height or weight, box plots were created. The Height Distribution by Injury Severity and Weight Distribution by Injury Severity graphs show that while average physical metrics were fairly consistent across severity levels, players with more severe injuries showed slightly more variability in both height and weight. This suggested some potential predictive value in physical build but was not strong enough to be conclusive on its own.

Labeling Injury Type (Initially Considered) In addition to severity, each injury was also categorized by injury type, including labels such as:

Lower: Legs, knees, ankles, feet

Upper: Shoulders, arms, elbows

Hand: Fingers, hands, wrists

Major Injury: Season-ending or surgical injuries

Non-body: Illness, personal leave

Other: Miscellaneous or uncategorized injuries

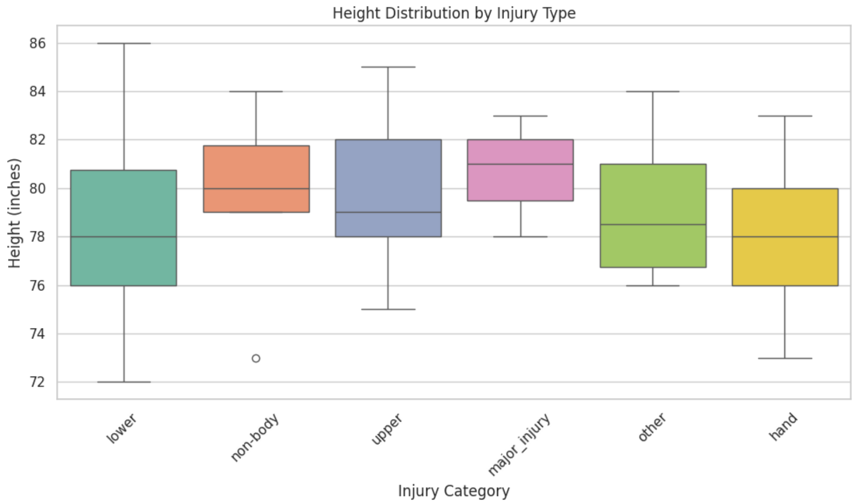


Figure 5: Height vs Type of Injury

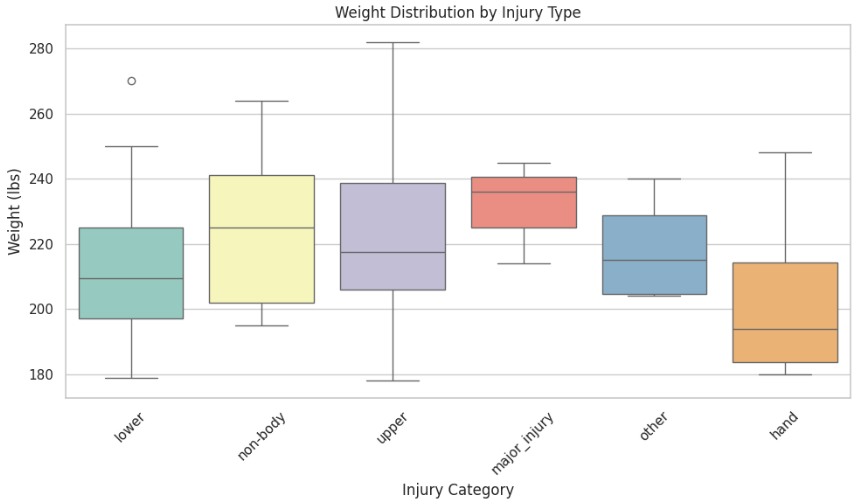


Figure 6: Weight vs Type of Injury

This labeling was done using keyword matching in the injury description field. Visualizations like Height Distribution by Injury Type and Weight Distribution by Injury Type were used to explore trends across categories.

However, while this additional categorization revealed some interesting patterns, incorporating injury type as a prediction target or classification scheme added complexity without meaningful performance gains in early experiments. It also introduced additional sparsity and imbalance, especially in less frequent categories. As a result, injury type was ultimately used only as an exploratory variable and not as a primary label or feature in the final model. The focus remained on predicting injury severity, which was more actionable and better supported by the dataset.

4.2.3 The Correlation Matrix

To better understand the relationships between player performance metrics and physical attributes, a correlation matrix was computed across key features, including minutes played (MP), field goal percentages, rebounds, assists, fouls, height, weight, and more.

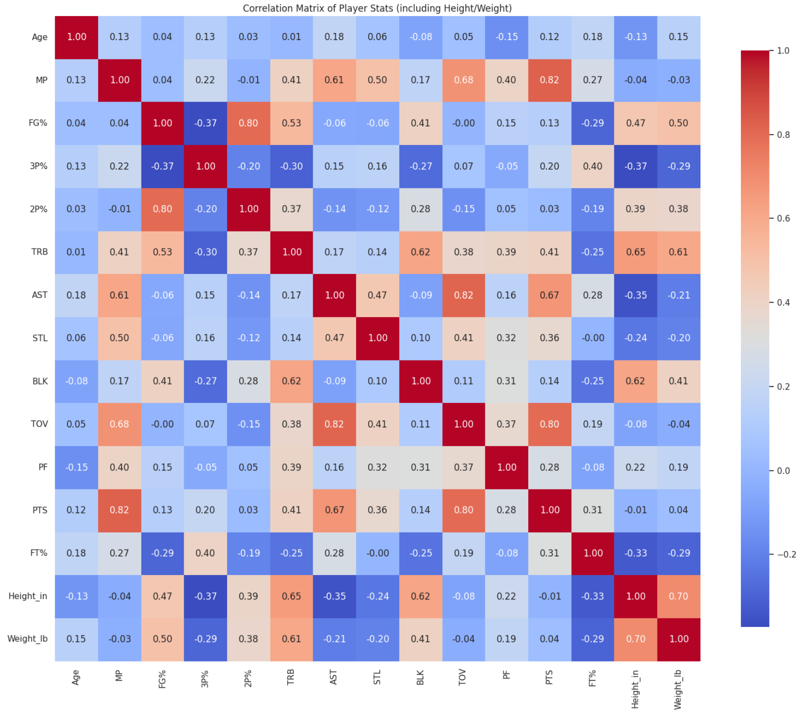


Figure 7: Correlation Matrix

This correlation matrix helped guide feature selection by identifying which variables were redundant and which contributed unique information. For example, while both height and weight were retained for modeling, variables like 2P% and FG% were checked for multicollinearity due to their high mutual correlation (0.80). This analysis was also helpful in identifying which features might provide insight into injury risk—for example, players with high minutes and high rebounds (often forwards/centers) may face increased physical wear, which was later reflected in model outcomes.

4.3 Research Problems

The primary research question driving this project was:

Can we use explainable AI techniques to see what factors contribute most to player injuries?

This led to several sub-questions:

What player traits or patterns are most predictive of injury risk?

How does performance (ex minutes played, shooting efficiency) correlate with injury severity?

Are some injury types or durations more prevalent for specific types of players?

Observations from the modeling phase also led to additional data considerations, such as refining injury labeling thresholds and assessing whether cumulative injury history should be weighted more heavily in future iterations.

5 Results

The final model was evaluated on a validation set of 104 samples, using a four-class classification task for injury severity: None (0), Minor (1), Moderate (2), and Severe (3). Overall, the model achieved an accuracy of 71%, with balanced performance across most classes.

5.1 The Confusion matrix

The confusion matrix shows that the model performed best on the None and Minor categories, correctly identifying 24 out of 26 and 20 out of 26 players, respectively. Moderate injuries were the most difficult to classify, with noticeable confusion across all other categories, especially Severe. This is likely due to overlapping features such as minutes played and physical stats.

Severe injuries were predicted reasonably well, with 18 out of 26 correctly classified. However, the model often misclassified them as Moderate or Minor, highlighting the difficulty of precisely identifying long-term injury risk based solely on available features.

The confusion between Moderate and Severe reflects a real-world gray area between players who return quickly and those who miss extended time, even when their statistical profiles are similar.

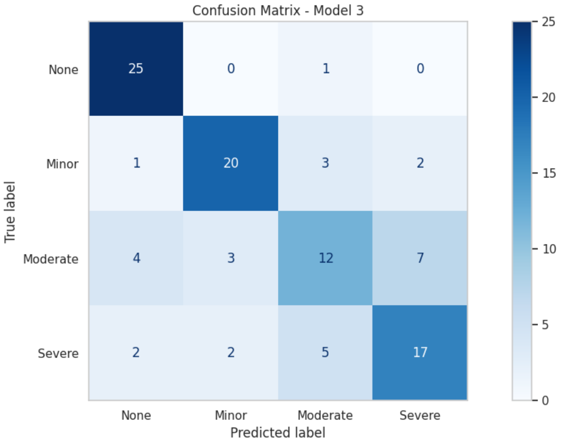


Figure 8: Confusion Matrix Best performing model

5.2 Precision, Recall and F1 Scores

According to the classification report, precision was highest for Minor injuries at 0.83, followed by None at 0.77. This suggests the model was more reliable when it predicted those classes.

Recall was highest for the None class at 0.92, meaning most players who were not injured were correctly identified. In contrast, the Moderate class had the lowest F1-score at 0.52, confirming it was the hardest to predict accurately.

The macro average F1-score was 0.70, showing that the model maintained balanced performance across all four classes. The weighted average was also 0.70, indicating that class imbalance did not significantly distort the results.

5.3 Summary of the Results

The model shows strong ability to distinguish between healthy players and those with minor or severe injuries. However, Moderate injuries remain a challenge, likely due to overlapping patterns with both adjacent classes. Future improvements could come from more granular features, such as contextual injury notes or time-series trends. Despite this, the current model provides a solid foundation for flagging potential injury risks and can already support coaching and medical staff in identifying at-risk players with reasonable confidence.

6 XAI Results

6.1 SHAP

6.1.1 Feature Importance Graph

. To better understand which features were most influential in the injury severity prediction model, SHAP values were computed for all inputs. The bar chart above shows the mean absolute SHAP value for each feature, grouped by class (injury severity levels: 0 = None, 1 = Minor, 2 = Moderate, 3 = Severe).

The higher the SHAP value, the higher impact the feature had on the models output for that class.

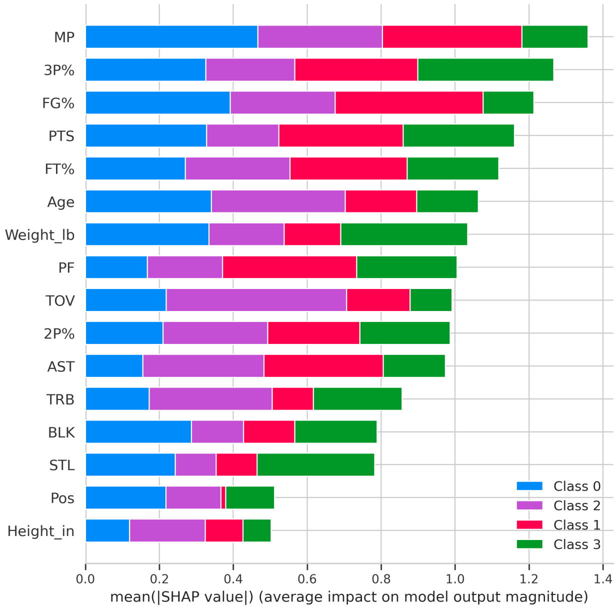


Figure 9: SHAP feature importance bar

Insights:

- Minutes Played (MP) had the highest overall influence across all classes, particularly for predicting Class 1 (Minor) and Class 2 (Moderate) injuries. This makes sense, as increased court time raises exposure to fatigue and collisions.
- Three-Point Percentage (3P%), Field Goal Percentage (FG%), and Points (PTS) were also highly influential. These may act as proxies for offensive workload and usage, contributing to physical strain.
- Free Throw Percentage (FT%), Age, and Weight also played important roles, especially for predicting more severe injuries (Class 3). Older and heavier players may carry greater physical risk.
- Fouls (PF) and Turnovers (TOV) were moderately impactful and likely reflect chaotic or aggressive play, which could increase injury risk.
- Position (Pos) and Height (Height inch) had relatively lower influence in comparison, though still contributed modestly to certain classes.

This SHAP analysis helps validate that the model is learning from meaningful, intuitive relationships. It also gives domain experts transparency into why a prediction was made. For instance, if a player is flagged as high risk for severe injury, the model can point to long minutes, scoring load, age, or physical build as the key contributing factors

6.1.2 Feature Importance Waterfall None

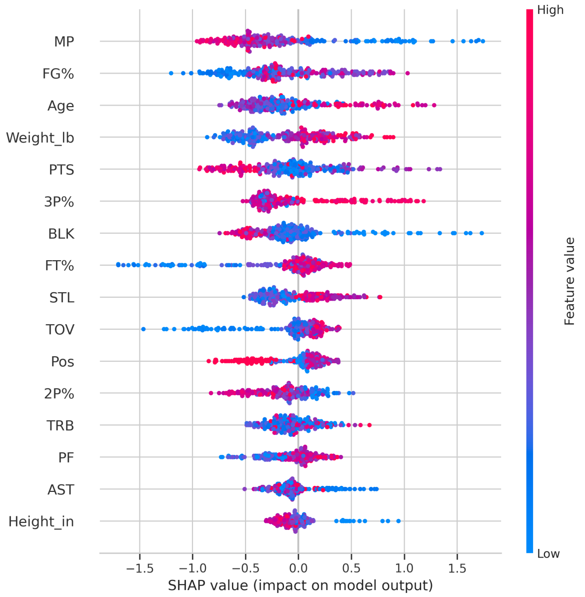


Figure 10: SHAP feature importance None

The model found that players with more minutes played were less likely to be classified as injury-free. At first, this seems confusing. In reality, players play a lot because they are not injured. The model sees high minutes as a sign of physical wear, even though it is actually a result of being healthy. Other features like age, weight, and high points per game also lowered the chance of being in the ”None” class, likely because they are linked with more physical strain.

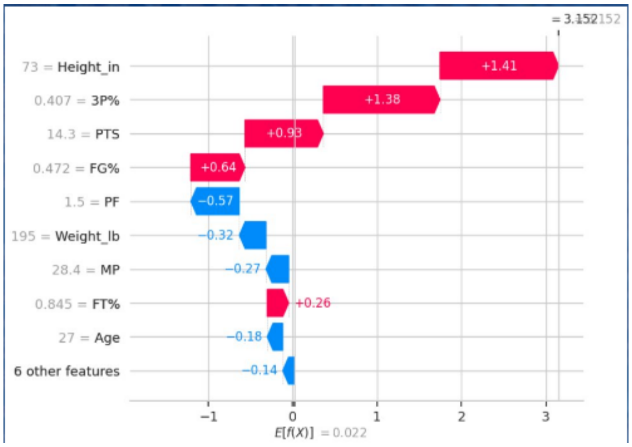


Figure 11: SHAP local feature importance None

To focus on localized explanations, for this case, the model predicted a player would not get injured. The top positive contributors to that prediction were low personal fouls (PF) and lower weight, which pushed the score down and helped the model lean toward a ”None” label. On the other hand, features like height, three-point percentage (3P%), and points scored pushed the prediction upward, suggesting some mild risk. This might seem contradictory, but the model sees taller players and high scorers as more active, which can increase exposure. However, in this case, the protective factors outweighed the risks.

6.1.3 Feature Importance Waterfall Minor

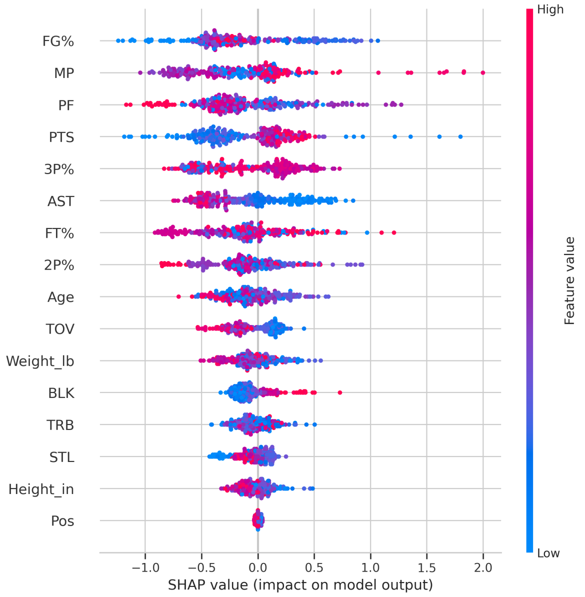


Figure 12: SHAP feature importance Minor

Players with high field goal percentage, lots of minutes, and more personal fouls were more likely to fall into the minor injury group. That might seem strange since these stats usually belong to top-performing players. But high usage often means more physical activity and contact, which can lead to small injuries. The model picks up on this connection between being very active and having short-term risks.

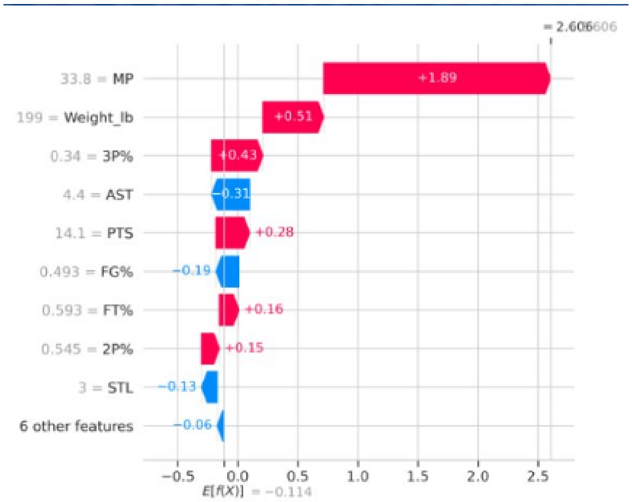


Figure 13: SHAP local feature importance Minor

This player was predicted to have a minor injury, with minutes played and weight contributing most to the decision. These features pushed the prediction higher because more minutes and higher body weight are often linked with physical stress. Additional features like three-point percentage and points also nudged the model toward the minor injury class. While some values like assists and steals pulled the prediction down slightly, they weren't strong enough to change the outcome.

6.1.4 Feature Importance Waterfall Moderate

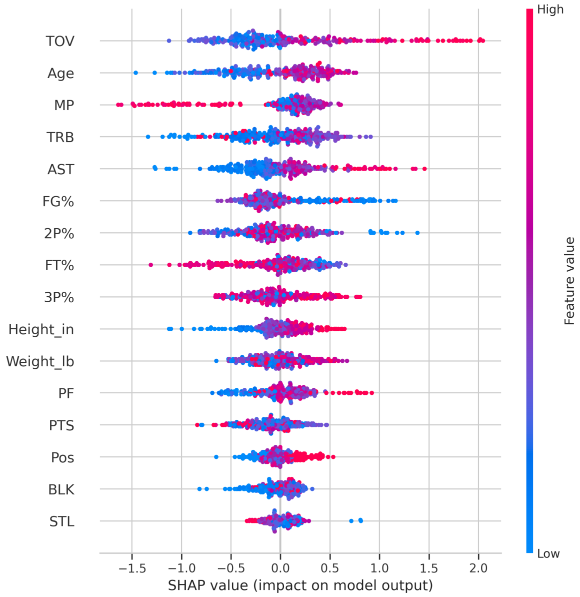


Figure 14: SHAP feature importance Moderate

For moderate injuries, the top signals were more turnovers and older age. This makes sense because older players or those handling the ball more might be under more stress. What's a bit surprising is that scoring or minutes were not as important here. That's likely because players with moderate injuries often fall in the middle — not high-usage stars but still active enough to get hurt. The model seems to notice subtle signs like declining efficiency or fatigue.

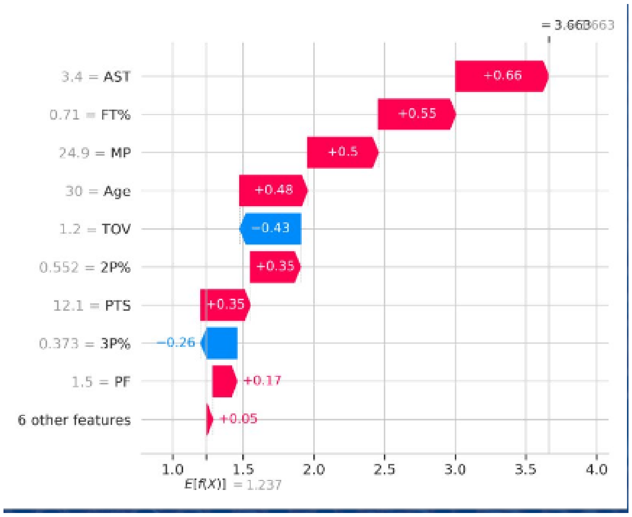


Figure 15: SHAP local feature importance Moderate

For this moderate injury prediction, several features had a positive influence. Assists, free throw percentage, and minutes played all added to the injury likelihood. These suggest a high-usage player with offensive responsibility and consistent playing time. The model also factored in age and turnovers, which slightly reduced the risk but were outweighed by the others. This mix of workload and efficiency stats likely flagged the player as moderately at risk.

6.1.5 Feature Importance Waterfall Severe

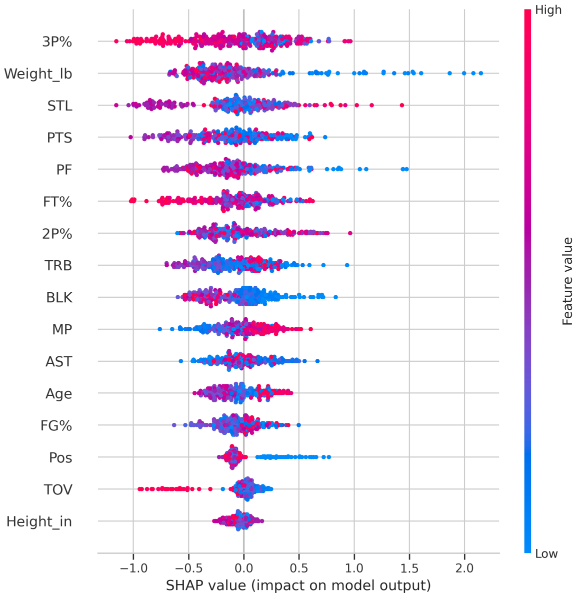


Figure 16: SHAP feature importance Severe

For severe injuries, the most important features were high three-point percentage and weight. This may seem unexpected. But shooting lots of threes often means jumping and landing in traffic, which increases injury risk. Heavier players also carry more load on their joints. Minutes played didn't matter as much here, possibly because severely injured players already play fewer minutes after getting hurt, so the model sees less signal in that feature.

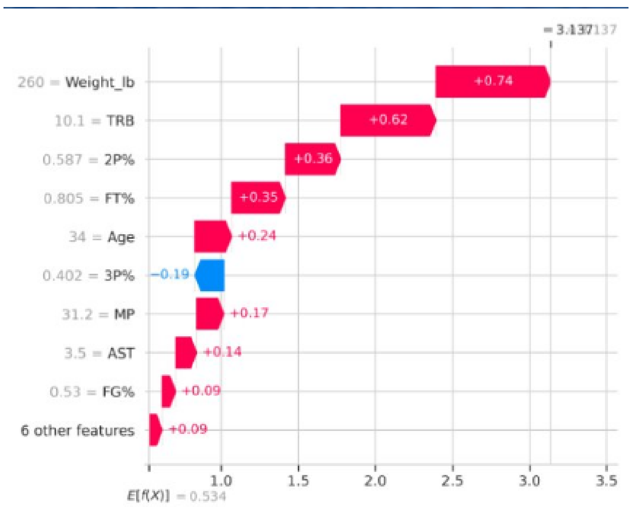
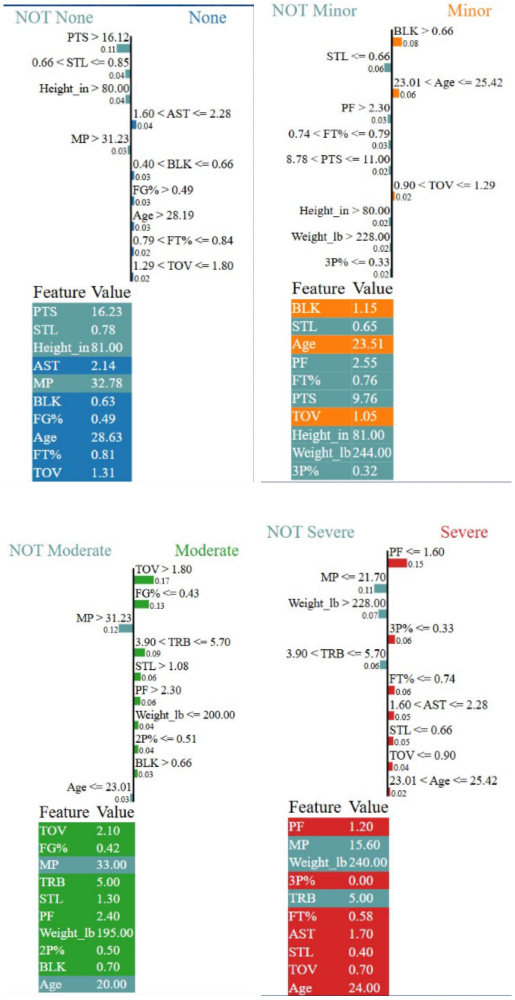


Figure 17: SHAP local feature importance Severe

This prediction was driven by a combination of physical and workload factors. The model strongly responded to weight, rebounds, and 2P%, all of which increased the predicted risk. These features suggest a physically engaged player, possibly a forward or center, who is active near the basket. High free throw percentage, age, and minutes played added further support for a severe injury classification. Although a slightly lower three-point percentage pulled the score down a bit, the rest of the inputs pushed the risk score confidently into the severe category.

6.2 LIME



Feature	Value
PTS	16.23
STL	0.78
Height_in	81.00
AST	2.14
MP	32.78
BLK	0.63
FG%	0.49
Age	28.63
FT%	0.81
TOV	1.31

Feature	Value
BLK	1.15
STL	0.65
Age	23.51
PF	2.55
FT%	0.76
PTS	9.76
TOV	1.05
Height_in	81.00
Weight_lb	244.00
3P%	0.32

Feature	Value
TOV	2.10
FG%	0.42
MP	33.00
TRB	5.00
STL	1.30
PF	2.40
Weight_lb	195.00
2P%	0.50
BLK	0.70
Age	20.00

Feature	Value
PF	1.20
MP	15.60
Weight_lb	240.00
3P%	0.00
TRB	5.00
FT%	0.58
AST	1.70
STL	0.40
TOV	0.70
Age	24.00

Figure 18: LIME explanations for each injury severity class: None, Minor, Moderate, and Severe

The LIME visualizations provide local explanations for why the model predicted each specific injury class (None, Minor, Moderate, Severe) for individual players. Each plot shows how certain features either pushed the prediction toward a given class or away from it. For example, in the ”None” prediction, lower points, shorter height, and lower minutes played increased the chance of being injury-free. In contrast, the ”Minor” prediction was driven by higher blocks, higher fouls, and being within a younger age range, suggesting more active physical engagement with moderate exposure. For ”Moderate” injuries, features like high turnovers, lower field goal percentage, and moderate minutes played were strong signals. Lastly, the ”Severe” injury prediction was explained by factors like low minutes, high weight, and low three-point percentage, all of which align with a profile of physically burdened players at risk. Overall, LIME helps break down how a combination of thresholds and values leads the model to classify a single prediction, making the decision process more interpretable for each case.

7 Conclusion

This project explored the use of machine learning and explainable AI to predict the likelihood and severity of NBA player injuries using the 2024–2025 season as a case study. By combining historical injury data with per-game performance statistics and physical attributes, the model was able to identify patterns associated with different levels of injury risk. Features such as minutes played, weight, field goal percentage, and age were consistently impactful across multiple model explanations, with SHAP and LIME helping reveal both global trends and individual predictions.

One key insight was that players with high usage or offensive efficiency were often associated with increased injury risk, especially in the minor and moderate categories. However, severe injuries were more influenced by physical strain indicators like weight and rebounding activity. These findings align with real-world expectations in some areas but also surfaced counterintuitive relationships, such as high three-point percentages or minutes played being treated as risk factors when they may actually reflect availability bias.

Challenges included class imbalance, overlapping feature behavior across severity levels, and the subjective nature of injury severity itself. Additionally, some features like injury type were considered but excluded due to added complexity and limited impact on predictive performance.

For future work, incorporating time-series data, external medical context, or biomechanical data could improve prediction accuracy and real-world applicability. Expanding to multi-season datasets or integrating game context (e.g., back-to-backs, travel) may also strengthen the model's generalization and utility in load management strategies.

Ultimately, this project demonstrates that explainable injury risk modeling is both feasible and valuable in sports analytics. With continued refinement and improvements, such tools could support coaches, sports trainers, and front offices in making proactive decisions that enhance player health and team performance.

References

[1] Basketball Reference. Nba player stats and game logs, 2025.

[2] L. A. Borowski, E. E. Yard, S. K. Fields, and R. D. Comstock. Injury in the national basketball association, 2005–2007. *The American Journal of Sports Medicine*, 36(12):2328–2335, 2008.

[3] Pro Sports Transactions. Nba injury history database, 2025.

[4] Xiaoyang Wei, Junbo Zhang, Peng Xu, and Li Cao. Injury risk prediction of professional basketball players using machine learning techniques. *IEEE Access*, 12:42294–42307, 2024.