Leveraging Machine Learning to Predict Injury Risk in NBA Players Using Performance and Workload Data

Peter Sarkis, DePaul University, School of Computing, psarkis@depaul.edu
Thejus Kannoth, DePaul University, School of Computing, tkannoth@depaul.edu
Matthew Smolzer, DePaul University, School of Computing, msmolzer@depaul.edu
Mounika Munigala, DePaul University, School of Computing, mmunigal@depaul.edu

ABSTRACT

We present research examining injuries that play a role in the performance and longevity of NBA players, affecting team success and financial investments. This study leverages machine learning to predict injury risk using player performance, workload, and injury history. Analyzing NBA data from 2013–2023, we applied models such as Logistic Regression, Decision Trees, Random Forest, Principal Component Analysis (PCA), and Gradient Boosting. SMOTE addressed class imbalance, and Support Vector Machine (SVM) was implemented with preprocessing techniques like Standard Scaling and Label Encoding. Hyperparameter tuning was applied for optimization. Findings provide insights to optimize player management and reduce injuries.

KEYWORDS: NBA player injuries, Injury prediction, Machine Learning, Player workload, Performance metrics

INTRODUCTION

In professional basketball, injuries can significantly impact a team's performance, player longevity, and financial stability. As the league continues to evolve, the ability to anticipate and mitigate injury risks has become increasingly valuable for teams looking to maintain a competitive edge. This study explores how machine learning can be applied to predict injury likelihood in NBA players by analyzing historical data from 2013 to 2023. By examining key performance metrics, workload management patterns, and prior injury history, this research aims to identify trends that contribute to injury susceptibility and provide data-driven insights for better player management.

This research is driven by the growing need for data-driven decisions in sports management. Traditionally, injury prevention has relied on subjective assessments by coaches and medical staff. However, with machine learning advancements, teams can now use data to make more objective decisions. Beyond individual player management, predictive modeling can help organizations refine training programs, improve long-term health strategies, and make more informed roster-building decisions.

Accurately predicting injuries brings several key benefits. First, it helps teams manage player performance by identifying those who need workload adjustments to avoid strain. Second, it contributes to player longevity by enabling the implementation of preventative strategies that reduce wear and tear over time. Third, injuries can be costly, leading to lost playing time and reduced return on investment for top players. Minimizing them protects team assets and keeps the roster strong throughout the season. Lastly, a predictive model provides a competitive edge

by allowing teams to make data-driven decisions on player rest, conditioning, and game-time strategies.

By addressing these aspects, this project seeks to bridge the gap between sports analytics and medical decision-making in the NBA. The insights gained can help coaches, trainers, and team management create better injury prevention plans, improving player careers and team performance. Additionally, these findings could enhance sports medicine research and set the foundation for more advanced, real-time injury prediction models, making data-driven player health management more effective.

To guide this research, we focus on four key questions: (1) What are the primary factors contributing to NBA player injuries? (2) Can machine learning models accurately predict injury risk based on player statistics? (3) How does a player's workload, including minutes played and back-to-back games, influence their injury risk? (4) Do injury risks vary based on a player's position or age group? These questions will help uncover critical insights that can enhance injury prevention strategies in the NBA.

LITERATURE REVIEW

There is a growing body of research exploring the role of machine learning (ML) in sports injury prediction as it relates to key factors contributing to injuries, the impact of workload, injury risk variations by player demographics, and the accuracy of machine learning (ML) models. These studies help establish a foundation for applying ML techniques to NBA injury prediction.

Key Factors Contributing to Injuries

Several studies focus on identifying injury risk factors. One of the early contributions was the METIC deep learning model to predict future injuries using player data over time (Cohan et al., 2021). The model significantly outperformed traditional machine learning techniques, demonstrating its potential in reducing injury rates and associated costs for sports teams. The application of ML to detect early warning signs of injury was examined by analyzing subtle changes in player movements and physical conditions, enabling teams to optimize training and recovery schedules (Gálvez et al., 2025). Similarly, XGBoost predicted lower extremity muscle strains in NBA players, identifying key factors such as age, injury history, and game load, which have critical implications for injury prevention strategies (Lu et al., 2022).

Workload and physical data are consistently found to be significant predictors. Another study applied various ML algorithms, including Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting, to predict injuries based on physical features like body fat percentage and height (Farghaly & Deshpande, 2024). The Random Forest model achieved the highest accuracy, further supporting the role of ML in player health management. In line with this, the utilization of GPS data was combined with player workload, position, and velocity to predict injuries in football players (Freitas et al., 2025). Recovery states and mental fatigue in young football players was displayed using wearable data and achieved impressive classification accuracy with several ML models, emphasizing the importance of monitoring player wellness in injury prevention (Teixeira et al., 2024).

The Impact of Workload

Studies have also explored the impact of workload and training intensity on injury risk. Combining internal load data such as wellness questionnaires with external load data like GPS tracking, significantly improves injury prediction accuracy (Vallance et al., 2020). Also, key

variables like body mass and sprint distance could predict recovery states, further highlighting the relationship between training loads and injury risk (Teixeira et al., 2024).

Injury Risk Variations by Player Demographics

The impact of past injuries on future injury risk is another area of focus. Player demographics, such as age and position, are key factors in injury prediction. It was found that players with higher salaries tend to have longer recovery times, which underscores the need for tailored rehabilitation strategies for star players (Sarlis et al., 2024). A distinct model using physical tests like knee movement and flexibility was developed to predict injury risk in youth soccer players, offering practical tools for injury prevention at a grassroots level (Robles-Palazón et al., 2023).

Age and player position are also critical factors in injury prediction. An additional study applied ML to lower-extremity injury prevention in futsal and youth soccer players, using pre-season data and neuromuscular performance tests (Robles-Palazón et al., 2023). Likewise, another study demonstrated how simple, yet effective screening methods can help identify players at risk and enable coaches and staff to prevent injuries before they occur (Ruiz-Pérez et al., 2021). Moreover, it was discovered that tracking wellness and training factors for volleyball players over a span of 24 weeks can help identify patterns related to injury risk, emphasizing the importance of monitoring both physical and subjective wellness data (De Leeuw et al., 2022). Also, analyzing runners' training logs with ML could help predict injuries, particularly by using features such as training load and past injury history (Lövdal et al., 2021).

The Accuracy of Machine Learning (ML) Models

The integration of ML into sports injury prediction spans across various sports. ML models were used to predict injuries in baseball players, achieving accuracy rates of 76% for position players (Karnuta et al., 2020). Combining screening and monitoring data helped predict non-contact injuries in football players (Hecksteden et al., 2023). The application of ML to NHL players, outperformed traditional methods and achieved high accuracy (Luu et al., 2020). Similarly, another study explored how blockchain and ML can enhance injury data management in football, highlighting improvements in both efficiency and security (Pu et al., 2023).

Several other studies have focused on biomechanical factors in injury prediction. One study applied ML to analyze athletic load and physical data to predict injuries in soccer players (Naglah et al., 2018), while another used LSTM neural networks to predict performance peaks based on subjective wellness data, offering valuable insights into managing training loads and preventing injuries (Wiik et al., 2019). A different study investigated lower-limb injury risks in basketball players by analyzing jumping mechanics, contributing to a better understanding of injury prevention at the biomechanical level (Wu & Wang, 2022).

Two studies explored the use of machine learning to analyze player workload, with one focusing on soccer (Vallance et al., 2020) and the other on tennis (Whiteside et al., 2017). Their findings highlight how monitoring player movements and workload can optimize training and reduce injury risks. Another study developed a ML model to optimize movement patterns in basketball training robots, which could be adapted to improve player movement and prevent injury during practice sessions (Xu & Tang, 2021).

Furthermore, one study highlighted the potential of ML to improve decision-making in sports science (Richter et al., 2024). Despite challenges in developing effective ML models, the

integration of coding and statistical expertise in training programs could further enhance ML applications in injury prediction across all sports.

DATA

The dataset used in this study is sourced from Kaggle and contains historical NBA player data from 2013 to 2023, covering over 1,200 players. It includes performance metrics, workload indicators, and injury records to facilitate injury risk prediction. The data is structured in tabular format, with each row representing a player-season and various statistical attributes.

Explanatory and Outcome Variables

The independent variables used in the model include player workload and performance metrics such as Days Missed, Post Touches, Games Played (GP), Usage Percentage (USG PCT), Age, Player Weight, Player Height, Pace, Average Speed, Minutes, and Drives, among others. The dependent variable is Injured Type, which categorizes the type of injury sustained by the player.

Data Preprocessing

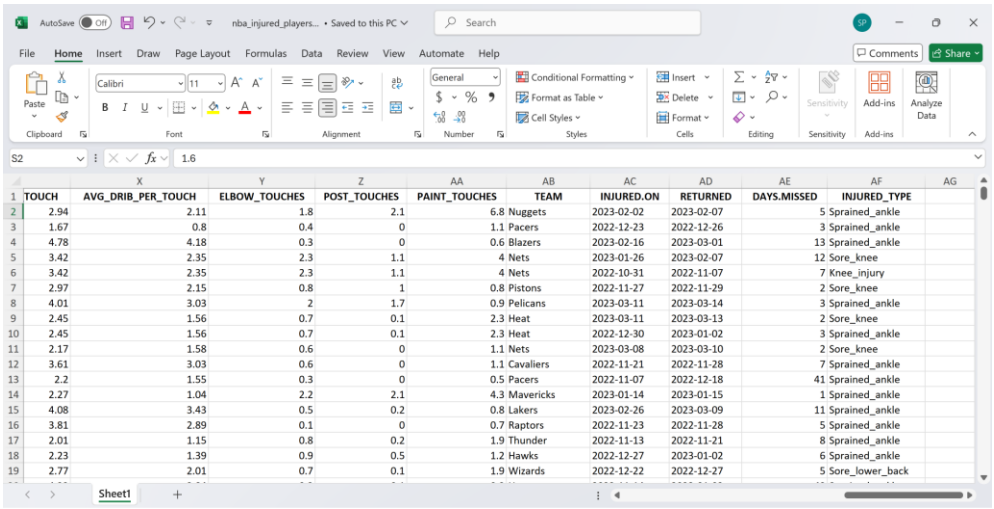
Several preprocessing steps were performed to clean and prepare the dataset for analysis. Missing values were identified and appropriately addressed to maintain data integrity. Duplicate player records were removed to ensure unique observations. Players were filtered to include only those with recorded injuries, ensuring a relevant dataset for injury prediction. Categorical features such as player names, teams, and positions were processed to be usable in machine learning models. Continuous numerical variables were normalized and standardized to ensure consistency and improve model performance.

Feature Selection

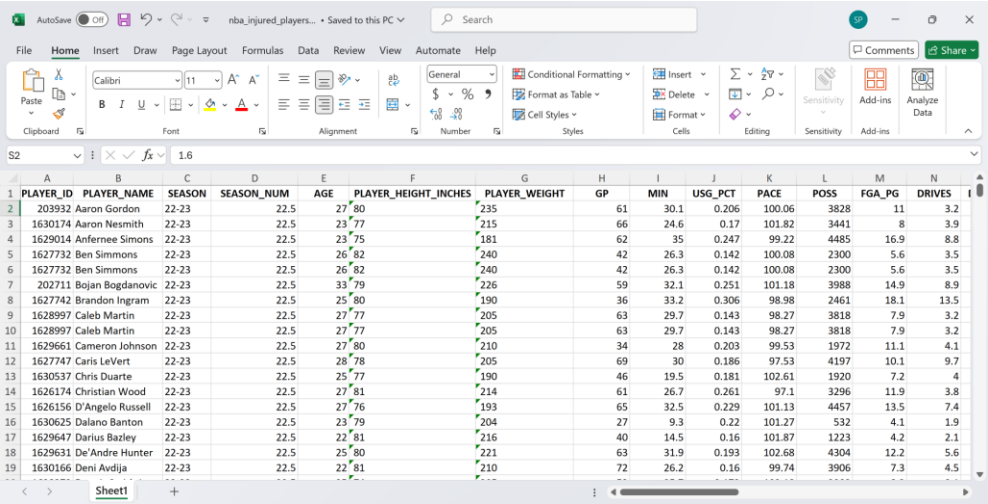
A ranking of variable importance was conducted to identify the most influential factors in injury risk prediction. The accuracy of the predictive model can potentially be improved by removing less important variables. Although no new features were explicitly engineered in this phase, further feature selection techniques may be applied during model optimization.

A statistical summary and visual snapshot of the dataset can be seen below to give insight into key trends and distributions before model implementation.

Snapshot of Dataset



TOUCH	AVG_DRIB_PER_TOUCH	ELBOW_TOUCHES	POST_TOUCHES	PAINT_TOUCHES	TEAM	INJURED.ON	RETURNED	DAYS.MISSED	INJURED_TYPE
2.94	2.11	1.8	2.1	6.8	Nuggets	2023-02-02	2023-02-07	5	Sprained_ankle
1.67	0.8	0.4	0	1.1	Pacers	2022-12-23	2022-12-26	3	Sprained_ankle
4.78	4.18	0.3	0	0.6	Blazers	2023-02-16	2023-03-01	13	Sprained_ankle
3.42	2.35	2.3	1.1	4	Nets	2023-01-26	2023-02-07	12	Sore_knee
3.42	2.35	2.3	1.1	4	Nets	2022-10-31	2022-11-07	7	Knee_injury
2.97	2.15	0.8	1	0.8	Pistons	2022-11-27	2022-11-29	2	Sore_knee
4.01	3.03	2	1.7	0.9	Pelicans	2023-03-11	2023-03-14	3	Sprained_ankle
2.45	1.56	0.7	0.1	2.3	Heat	2023-03-11	2023-03-13	2	Sore_knee
2.45	1.56	0.7	0.1	2.3	Heat	2022-12-30	2023-01-02	3	Sprained_ankle
2.17	1.58	0.6	0	1.1	Nets	2023-03-08	2023-03-10	2	Sore_knee
3.61	3.03	0.6	0	1.1	Cavaliers	2022-11-21	2022-11-28	7	Sprained_ankle
2.2	1.55	0.3	0	0.5	Pacers	2022-11-07	2022-12-18	41	Sprained_ankle
2.27	1.04	2.2	2.1	4.3	Mavericks	2023-01-14	2023-01-15	1	Sprained_ankle
4.08	3.43	0.5	0.2	0.8	Lakers	2023-02-26	2023-03-09	11	Sprained_ankle
3.81	2.89	0.1	0	0.7	Raptors	2022-11-23	2022-11-28	5	Sprained_ankle
2.01	1.15	0.8	0.2	1.9	Thunder	2022-11-13	2022-11-21	8	Sprained_ankle
2.23	1.39	0.9	0.5	1.2	Hawks	2022-12-27	2023-01-02	6	Sprained_ankle
2.77	2.01	0.7	0.1	1.9	Wizards	2022-12-22	2022-12-27	5	Sore_lower_back



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	PLAYER_ID	PLAYER_NAME	SEASON	SEASON_NUM	AGE	PLAYER_HEIGHT_INCHES	PLAYER_WEIGHT	GP	MIN	USG_PCT	PACE	POSS	FGA_PG	DRIVES
2	203932	Aaron Gordon	22-23	22.5	27	80	235	61	30.1	0.206	100.06	3828	11	3.2
3	1630174	Aaron Nesmith	22-23	22.5	23	77	215	66	24.6	0.17	101.82	3441	8	3.9
4	1629014	Anfernee Simons	22-23	22.5	23	75	181	62	35	0.247	99.22	4485	16.9	8.8
5	1627732	Ben Simmons	22-23	22.5	26	82	240	42	26.3	0.142	100.08	2300	5.6	3.5
6	1627732	Ben Simmons	22-23	22.5	26	82	240	42	26.3	0.142	100.08	2300	5.6	3.5
7	202711	Bojan Bogdanovic	22-23	22.5	33	79	226	59	32.1	0.251	101.18	3988	14.9	8.9
8	1627742	Brandon Ingram	22-23	22.5	25	80	190	36	33.2	0.306	98.98	2461	18.1	13.5
9	1628997	Caleb Martin	22-23	22.5	27	77	205	63	29.7	0.143	98.27	3818	7.9	3.2
10	1628997	Caleb Martin	22-23	22.5	27	77	205	63	29.7	0.143	98.27	3818	7.9	3.2
11	1629661	Cameron Johnson	22-23	22.5	27	80	210	34	28	0.203	99.53	1972	11.1	4.1
12	1627747	Caris LeVert	22-23	22.5	28	78	205	69	30	0.186	97.53	4197	10.1	9.7
13	1630537	Chris Duarte	22-23	22.5	25	77	190	46	19.5	0.181	102.61	1920	7.2	4
14	1626174	Christian Wood	22-23	22.5	27	81	214	61	26.7	0.261	97.1	3296	11.9	3.8
15	1626156	D'Angelo Russell	22-23	22.5	27	76	193	65	32.5	0.229	101.13	4457	13.5	7.4
16	1630625	Dalano Banton	22-23	22.5	23	79	204	27	9.3	0.22	101.27	532	4.1	1.9
17	1629647	Darius Bazley	22-23	22.5	22	81	216	40	14.5	0.16	101.87	1223	4.2	2.1
18	1629631	De'Andre Hunter	22-23	22.5	25	80	221	63	31.9	0.193	102.68	4304	12.2	5.6
19	1630166	Deni Avdija	22-23	22.5	22	81	210	72	26.2	0.16	99.74	3906	7.3	4.5

EDA Statistical Summaries

```
 Summary statistics
           PLAYER_ID   SEASON_NUM          AGE  PLAYER_HEIGHT_INCHES  \
count  1.214000e+03  1214.000000  1214.000000           1214.000000
mean   7.845084e+05    19.278418    26.687809             78.896211
std    7.099509e+05     2.220974     4.132859              3.500903
min    2.037000e+03    13.500000    19.000000             70.000000
25%    2.023300e+05    17.500000    23.000000             76.000000
50%    2.038990e+05    19.500000    26.000000             79.000000
75%    1.628377e+06    21.500000    30.000000             82.000000
max    1.631110e+06    22.500000    39.000000             89.000000

       PLAYER_WEIGHT           GP          MIN      USG_PCT         PACE  \
count    1214.000000  1214.000000  1214.000000  1214.000000  1214.000000
mean      219.747117    55.697694    25.148023     0.202437   100.240041
std        24.981038    17.097655     7.856119     0.059407     3.106698
min       169.000000     1.000000     1.000000     0.026000    91.030000
25%       200.000000    46.000000    19.425000     0.160000    98.270000
50%       220.000000    60.000000    26.600000     0.192000   100.240000
75%       240.000000    68.000000    31.900000     0.245000   102.217500
max       311.000000    81.000000    37.900000     0.375000   120.910000

              POSS  ...    AVG_SPEED  PULL_UP_FGA      TOUCHES  \
count  1214.000000  ...  1214.000000  1214.000000  1214.000000
mean   3062.188633  ...     4.185362     2.750165    45.427842
std    1396.901822  ...     0.215552     2.690347    20.514543
min       3.000000  ...     3.580000     0.000000     1.500000
25%    2020.000000  ...     4.060000     0.500000    29.600000
50%    3165.000000  ...     4.190000     1.900000    43.000000
75%    4202.250000  ...     4.330000     4.400000    60.600000
max    6037.000000  ...     5.180000    12.900000    97.400000

       FRONT_CT_TOUCHES  AVG_SEC_PER_TOUCH  AVG_DRIB_PER_TOUCH  ELBOW_TOUCHES
count       1214.000000        1214.000000         1214.000000    1214.000000
mean          24.257084           2.868204            2.042636       1.419934
std           10.328467           1.314478            1.558033       1.558144
min            0.500000           1.130000            0.000000       0.000000
25%           16.200000           1.720000            0.700000       0.400000
50%           24.000000           2.475000            1.560000       0.800000
75%           31.200000           3.960000            3.240000       1.900000
max           56.200000           6.340000            6.710000       8.500000


      POST_TOUCHES  PAINT_TOUCHES  DAYS.MISSED
count   1214.000000    1214.000000  1214.000000
mean       1.301400       2.550165     8.752059
std        2.207039       2.511706    18.313393
min        0.000000       0.000000     1.000000
25%        0.000000       0.800000     2.000000
50%        0.300000       1.600000     4.000000
75%        1.600000       3.875000     8.000000
max       14.500000      13.900000   282.000000

[8 rows x 25 columns]
```

Summary statistics like the mean, standard deviation, min, and max were computed for numerical variables such as minutes played, usage percentage, player age, etc.

HYPOTHESIS/MODEL

Our hypothesis is that player workload and performance metrics such as minutes played, usage percentage, and average speed are significant predictors of injury risk in NBA players. By analyzing these factors, we aim to determine the most influential contributors to injury occurrences. To predict injury risk, we will explore various machine learning models, including Logistic Regression for binary classification to determine whether a player is at risk of injury based on workload and performance metrics, Random Forest and Decision Trees for determining key injury risk factors and improving interpretability, Gradient Boosting for improved predictive accuracy through ensemble methods, and Principal Component Analysis (PCA) for reduced dimensionality and identifying key patterns in player performance and injury trends.

To address class imbalance, we will apply the Synthetic Minority Over-sampling Technique (SMOTE). Feature selection will be performed to identify the most impactful predictors of injury risk. Support Vector Machine (SVM) will be implemented with preprocessing techniques such as Standard Scaling and Label Encoding to enhance classification performance. Hyperparameter tuning using GridSearchCV and RandomizedSearchCV will also be conducted to optimize model performance. The selected model will be evaluated based on performance metrics such as accuracy, precision, recall, and F1-score, which ensures reliable predictions for injury risk assessment in NBA players.

METHODS

Research Questions

This study aims to answer the following key research questions:

(1) What are the key factors that contribute to NBA player injuries?

(2) Can machine learning models accurately predict injury risk using player statistics?

(3) How does a player's workload (minutes played, back-to-back games, etc.) influence their injury risk?

(4) Do injury risks vary based on a player's position or age group?

These questions will be addressed through statistical summaries, exploratory data analysis (EDA), correlation analysis, and machine learning modeling. Various visualizations and model evaluations will be used to provide insights into injury risk prediction in NBA players.

Exploratory Data Analysis (EDA)

To gain a deeper understanding of the NBA dataset, we conducted an exploratory data analysis (EDA) that includes statistical summaries, visualizations, and correlation analysis. Besides statistical summaries, various visualizations were created to examine injury trends and relationships between workload metrics and injuries.

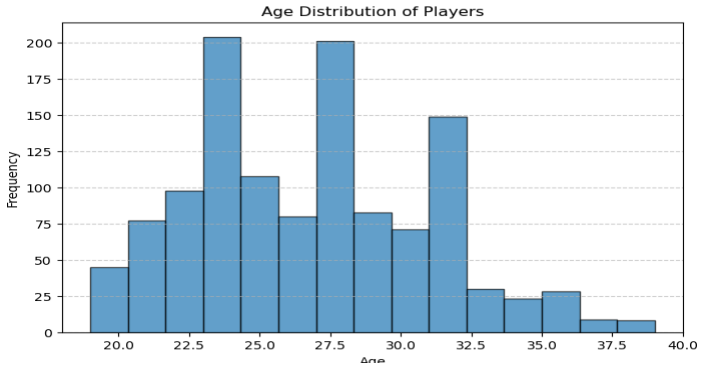


Figure 1. Histogram to identify the most common age range among players and potential trends related to injury risk in the NBA.

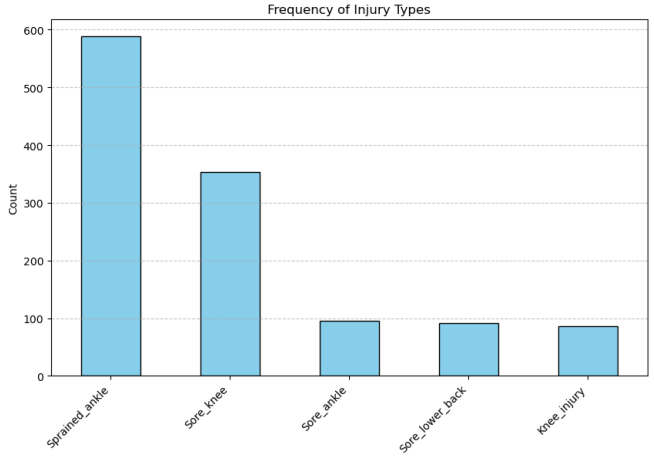


Figure 2. Bar chart to identify injury patterns and potential risk factors, highlighting which injuries are most common among NBA players.

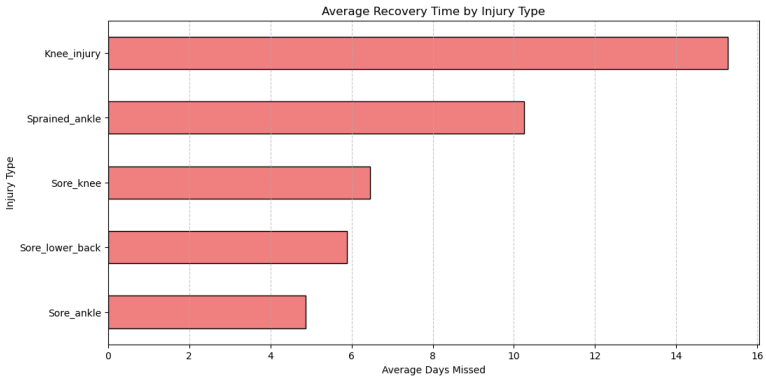


Figure 3. Bar chart illustrates the average recovery time by injury type, measured in average days missed. This shows which injuries typically require the longest recovery periods, helping to assess their impact on player availability.

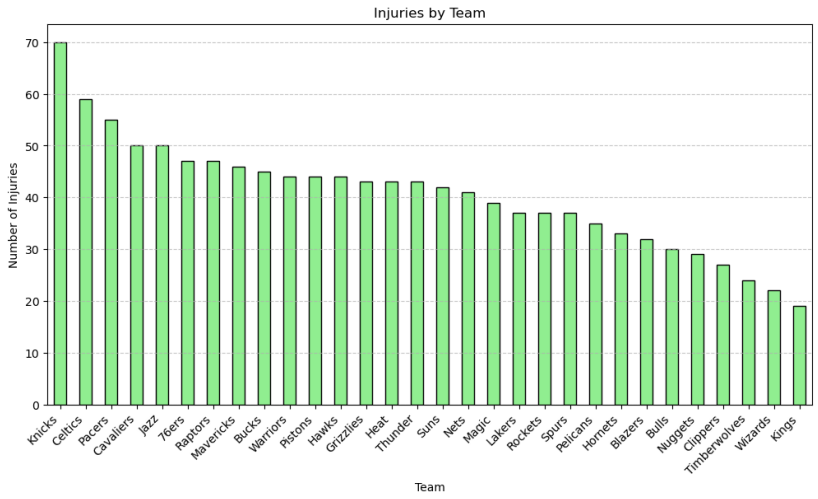


Figure 4. Bar chart shows the average number of injuries per team, highlighting which teams experience more injuries on average.

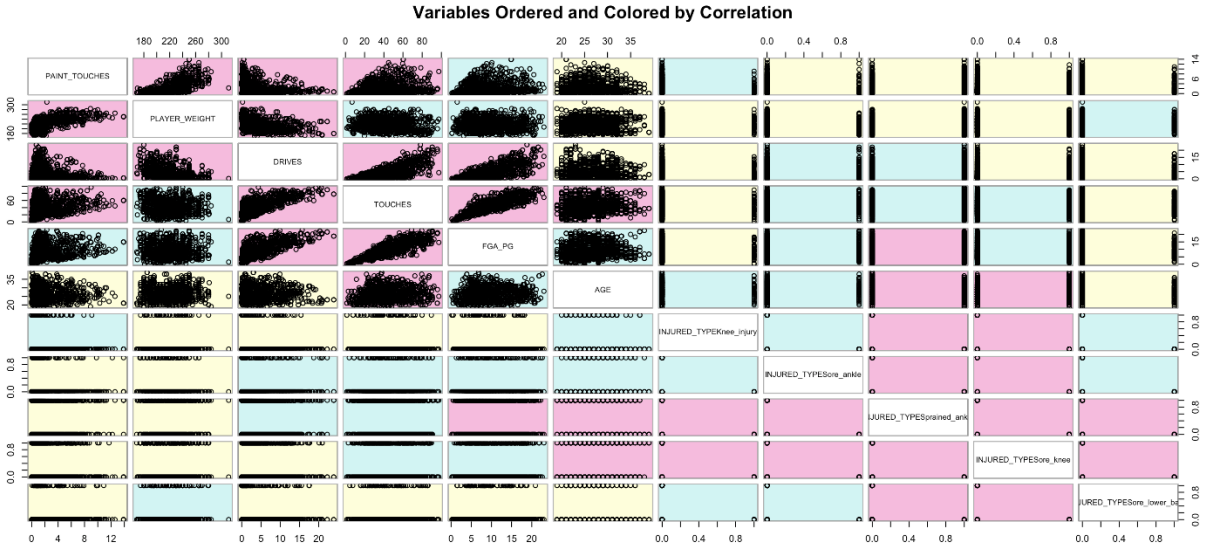


Figure 5. This scatter matrix shows the correlation between our response variable (injury type) and our most significant predictors. The matrix is color coded to best show the strength of correlations with red being more correlated, and green being least correlated.

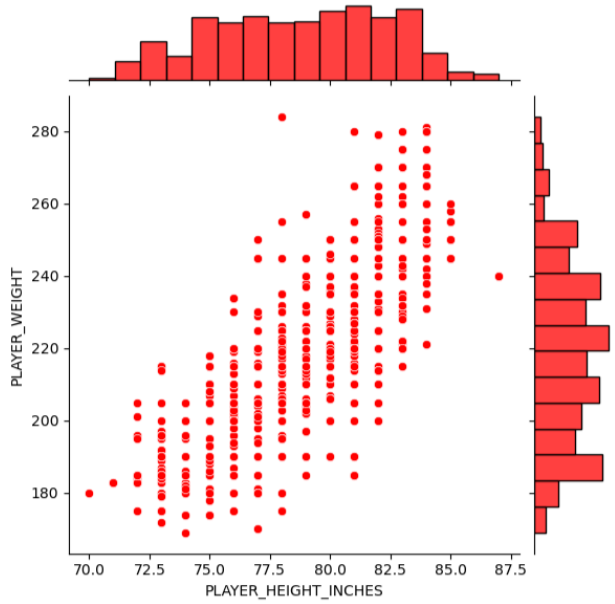


Figure 6. This joint plot illustrates the relationship between player height and weight, showing a positive correlation, meaning taller players tend to weigh more.

Data Splitting into Training and Testing

The dataset was divided into 80% for training and 20% for testing to evaluate the model effectively. Cross-validation was performed within the training set using a train-validation split to optimize model performance and prevent overfitting. The validation set was used for hyperparameter tuning, which helped to refine the models before final testing.

Machine Learning (ML) Models

Several machine learning models were tested to find the best approach for predicting injury risk. Logistic Regression served as a baseline for classifying injured vs. not injured players. Random Forest and Decision Trees were used for feature importance and interpretability. Gradient Boosting (XGBoost) was used to improve prediction accuracy with ensemble learning. Principal Component Analysis (PCA) helped reduce dimensionality to identify key injury-related patterns. Support Vector Machine (SVM) was also tested, using preprocessing techniques like Standard Scaling and Label Encoding, to enhance classification performance. K-Nearest Neighbors (KNN) was applied as a distance-based classifier, using Standard Scaling to normalize features and improve performance in identifying injury patterns.

Model Evaluation Metrics

Model performance was evaluated based on accuracy, which measures the overall prediction correctness. Precision, Recall, and F1-scores to assess the balance between false positives and false negatives. A Confusion Matrix to help break down correct and incorrect classifications. The results of these evaluations will determine the most effective model for predicting injury risk in NBA players.

Overall Model Building Process

The model building process follows a structured approach to ensure accuracy and reliability in injury prediction:

1. **Exploratory Data Analysis (EDA):** Initial analysis of data distribution, correlations, and key trends using visualizations.

2. **Data Splitting:** The dataset is divided into training (80%) and testing (20%) sets to evaluate model performance.

3. **Model Training and Validation:** Multiple models are trained using cross-validation techniques, with hyperparameter tuning performed to optimize accuracy.

4. **Feature Selection and Dimensionality Reduction:** Key features are selected based on importance ranking, and PCA is applied to improve model efficiency.

5. **Final Model Testing:** The best-performing model is evaluated on the test set, with performance measured using accuracy, precision, recall, and F1-score.

6. **Model Interpretation and Insights:** The results are analyzed to understand the key injury risk factors and provide actionable insights for NBA teams.

RESULTS

The XGBoost model using cross-validation achieved the highest accuracy, outperforming Decision Trees, Random Forest and Logistic Regression in terms of accuracy, precision, and F1-score. The most significant predictor variables aligned with expectations, including Days Missed, Post Touches, Possessions, Games Played, and Age.

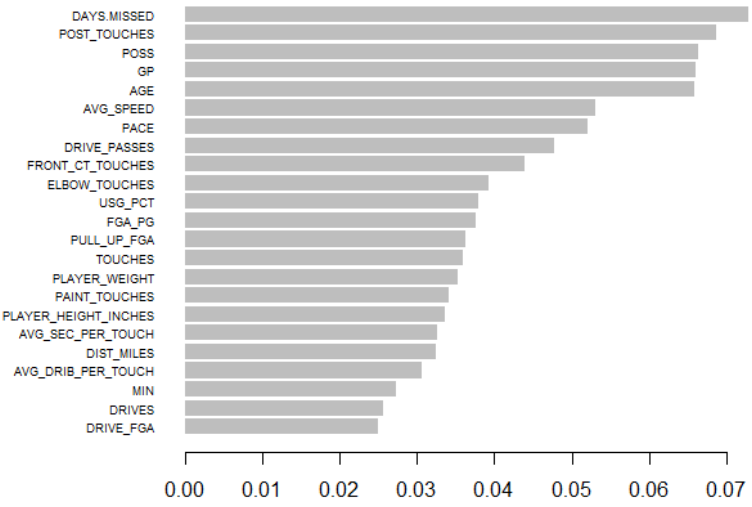


Figure 7. This chart shows the importance of each of the variables that are used within the final XGBoost model. The five most significant being Days Missed, Post Touches, Possessions, Games Played, and Age.

The model itself works well, as allowing for the learning from multiple rounds and using multiple trees, the model can become more predictive. In this case, with basketball injuries, there are similarities in the way players get injured. By figuring out what is causing these injuries, preventative measures can be made to reduce injury without taking away from the sport. The XGBoost final model had an accuracy of 97.97%, being able to predict injuries well given the necessary statistics.

```
                Reference
Prediction       Knee_injury Sore_ankle Sore_knee Sore_lower_back Sprained_ankle
  Knee_injury             19          0         1               0              0
  Sore_ankle               0         18         0               0              2
  Sore_knee                0          0        68               0              1
  Sore_lower_back          0          0         0              22              0
  Sprained_ankle           0          0         1               0            114

Overall Statistics

               Accuracy : 0.9797
                 95% CI : (0.9532, 0.9934)
    No Information Rate : 0.4756
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.97

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Knee_injury Class: Sore_ankle Class: Sore_knee Class: Sore_lower_back Class: Sprained_ankle
Sensitivity                     1.00000           1.00000          0.9714                1.00000               0.9744
Specificity                     0.99559           0.99123          0.9943                1.00000               0.9922
Pos Pred Value                  0.95000           0.90000          0.9855                1.00000               0.9913
Neg Pred Value                  1.00000           1.00000          0.9887                1.00000               0.9771
Prevalence                      0.07724           0.07317          0.2846                0.08943               0.4756
Detection Rate                  0.07724           0.07317          0.2764                0.08943               0.4634
Detection Prevalence            0.08130           0.08130          0.2805                0.08943               0.4675
Balanced Accuracy               0.99780           0.99561          0.9829                1.00000               0.9833
```

Figure 8. This is the output of the model into a confusion matrix. This allows for the accuracy to be calculated as well as a visual of all the correct predictions on the test set of the data.

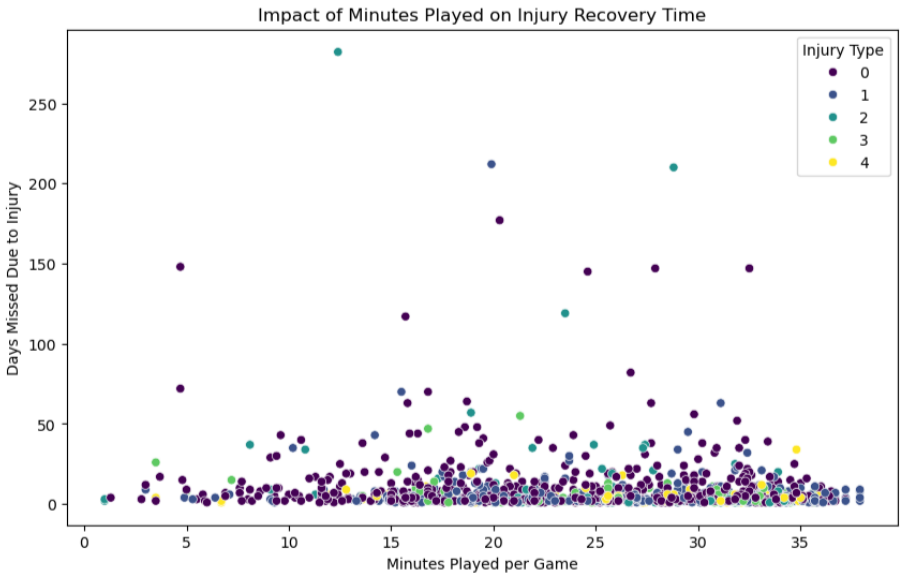


Figure 9. This scatter plot shows that while most players recover quickly, some with lower playing time face prolonged recovery, suggesting injury severity and player condition matter more than workload.

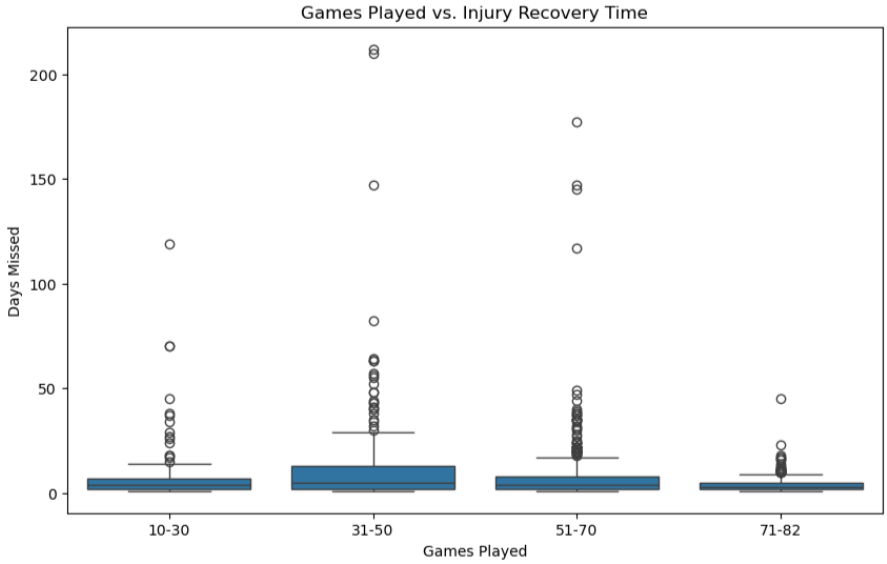


Figure 10. This box plot shows that players in the 10-50 game range have higher variance in missed days, with outliers suggesting severe injuries limit total games played.

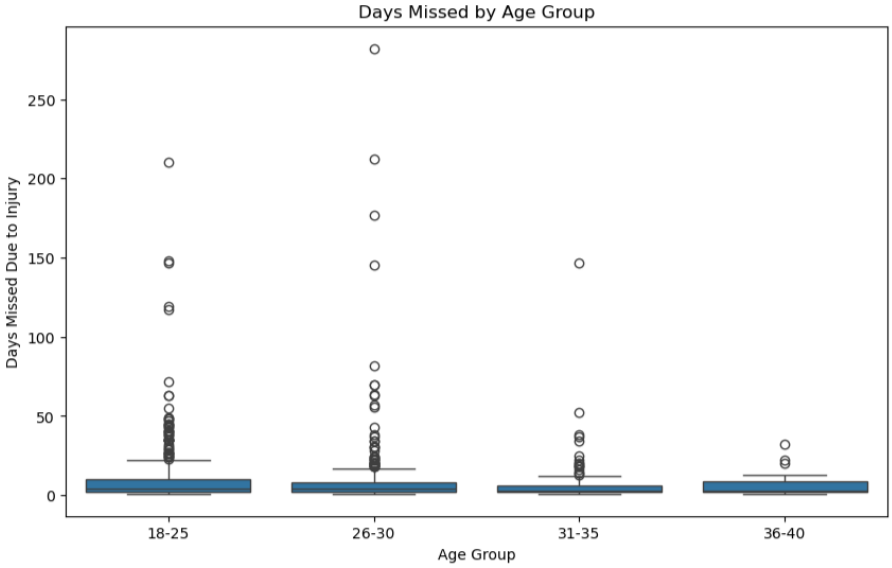


Figure 11. This box plot shows age-related injury recovery, with younger players (18-30) having more extreme outliers, suggesting severe injuries may impact them more due to high intensity play.

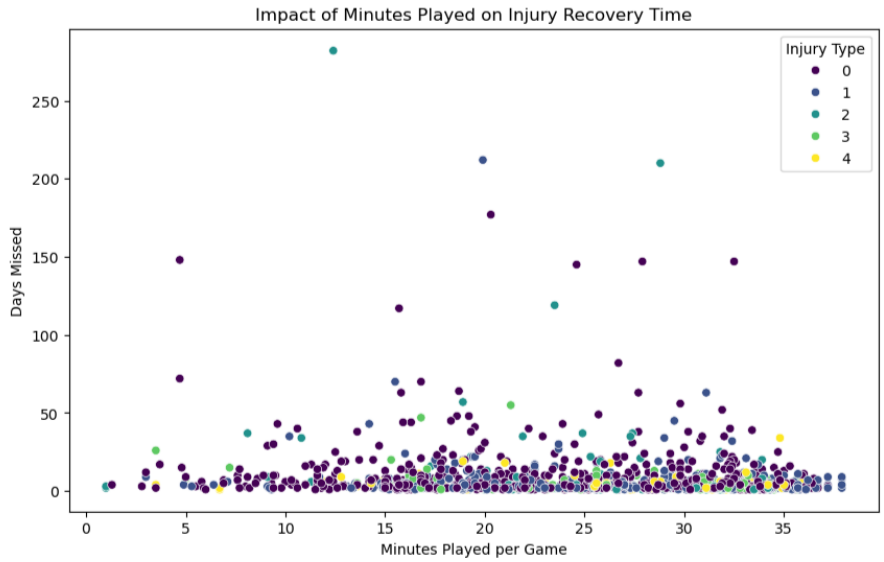


Figure 12. This box plot shows that players in the 10-50 game range have higher variance in missed days, with outliers suggesting severe injuries limit total games played.

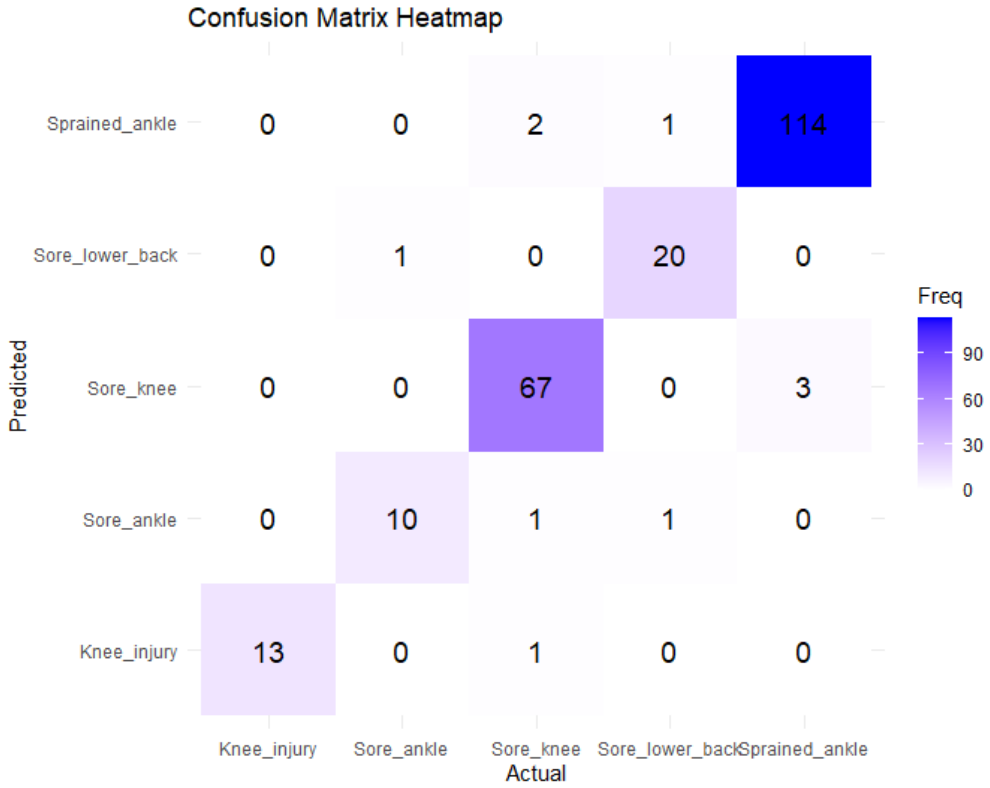


Figure 13. This heatmap is the visualization of the confusion matrix made by the model. The shading is based on the frequency of each prediction compared to the actual injury. While there are significantly more sprained ankles, the heat map shows the high accuracy of the model.

Injury recovery times are influenced by factors beyond just workload. While most players miss only a few days regardless of playing time, some with moderate minutes experience significantly longer recovery periods, suggesting that injury severity and individual conditioning play a larger role. Players who participate in fewer games tend to have greater variation in days missed, indicating they may have suffered more severe injuries requiring prolonged recovery. Additionally, younger athletes show more extreme outliers in recovery time, possibly due to high intensity play styles and increased workload at a younger age. Overall, injury severity appears to depend on multiple factors, including conditioning, injury type, and play style, rather than just total minutes played.

The results confirm that player workload and performance metrics play a crucial role in injury risk prediction. Machine learning models, particularly ensemble methods like Gradient Boosting, provide valuable insights for injury prevention strategies in professional basketball. By having this model that is highly accurate in predicting injury, new data can be added to further increase the accuracy and adaptability of the model into other injury types.

With additional time, further improvements could be made, including gathering more extensive data, experimenting with additional machine learning techniques such as stacking or deep learning, and further refining hyperparameter tuning methods to enhance model accuracy.

DISCUSSION AND CONCLUSIONS

The findings of this study confirm that player workload and prior injuries are strong predictors of future injuries in the NBA. The XGBoost model demonstrated the highest predictive accuracy, highlighting the value of ensemble learning techniques in injury risk assessment. Key features such as Days Missed, Games Played, Age, Pace, and Possessions played a crucial role in determining injury likelihood, aligning with existing research. These insights reinforce the importance of managing player workload since high-intensity and excessive minutes played increase the risk of player injuries. By leveraging machine learning, teams can adopt a proactive approach that identifies players at high-risk before injuries occur. Additionally, implementing targeted interventions like load management strategies and personalized recovery programs can significantly reduce player injuries. Predictive analytics integration into team decision-making can help enhance player availability and longevity, which ultimately impacts overall team success and financial investments in player contracts.

Beyond improving injury prediction, this study bridges the gap between sports analytics and sports medicine, providing a data-driven foundation for optimizing player health and performance. While the models performed well, challenges such as data limitations, potential biases in injury reporting, and the complexity of injury mechanisms highlight areas for future research. Expanding the dataset to include real-time data metrics could further refine predictions and improve model generalizability. As machine learning continues to evolve, its integration into sports will become increasingly valuable, transforming how teams manage player health and competitive performance. This research highlights the impact of analytics in professional basketball and paves the way for more advanced applications in sports injury prevention.

FUTURE WORK

As this field continues to evolve, further research is needed to refine these models and expand their applicability to other sports and player demographics. This study can be enhanced through

more datasets to other areas where models can be applied including additional professional leagues such as the National Football League (NFL) and the National Hockey League (NHL). The integration of real-time injury tracking data and the development of interactive dashboards for injury risk visualizations could provide teams with immediate insights, allowing for proactive decision-making regarding player workload, rest schedules, and rehabilitation strategies. Feature engineering can be further utilized for the development of new features that capture injury trends, such as rolling averages of workload metrics. Additionally, all studies mentioned in this paper collectively underscore the significant potential of machine learning in predicting injuries across various sports, including basketball. By analyzing diverse data sources such as player workload, physical metrics, and wellness information, ML models can help optimize training, enhance player recovery, and ultimately reduce injury risks.

AUTHOR CONTRIBUTIONS

Peter Sarkis chose the dataset, prepared initial data cleaning, put together the report including the abstract, introduction, conclusion, some literature review, the presentations, and all updates. Mounika Munigala helped with the literature review and future work. Thejus Kannoth and Mattew Smolzer performed further data cleaning and preparation, developed all machine learning algorithms, and created visualizations for the results and methods sections. All authors contributed to writing this paper.

APPENDIX 1: Feature Selection Criteria

The following table displays the significant features considered in the machine learning model for predicting NBA injuries, along with their descriptions and reasons for inclusion.

Feature Name	Description	Reason for Inclusion
Days Missed	Games missed due to injury	Direct indicator of injury impact
Post Touches	Number of times a player touches the ball in the post	Post play may lead to increased physical contact
GP	Total games played in a season	High workload may increase injury risk
Poss	Number of offensive possessions a player has	More possessions can indicate higher activity level
Age	Player's age	Older players may have increased injury risk
Pace	Average pace of play	Fast-paced teams may fatigue and are prone to injuries
Drive Passes	Passes towards the basket	Frequent drives could lead to more contact and potential injuries

Average Speed	Player's average speed on court	High-speed movements can increase injury risk
Front CT Touches	Touches in the frontcourt	Measures offensive involvement and physicality
Paint Touches	Touches inside the paint area	More paint touches may increase physical contact
USG PCT	Percentage of how much a player is used on the court	Higher usage may lead to fatigue and overuse injuries
Pull Up FGA	Field goal attempts off dribble	Shot attempts may impact lower body strain
Elbow Touches	Ball touches by free-throw line	Involvement in mid-range areas
FGA FG	Field goals attempted and made	Workload and potential fatigue
Player Weight	Player's body weight	Heavier players may experience higher joint stress
Player Height	Length of the player	Taller players may have different injury risk factors than shorter
Touches	When a player touches the ball	Higher involvement may correlate with injury risk
Avg Sec Per Touch	Average seconds a player holds the ball	Longer possession time could indicate playing style and workload
Min	Minutes played	Higher minutes may increase fatigue and risk of injury
Dist Miles	Distance covered per game	More movement could lead to fatigue-related injuries
Avg Drb Per Touch	Average dribbles per touch	Measures ball-handling frequency and movement
Drive FGA	Field goal attempts taken on drives	May increase contact-related injuries by driving to the basket
Drives	Number of drives to the basket	More drives may indicate higher physicality

APPENDIX 2: Model Performance Metrics

Below are the evaluation metrics for different machine learning models used in the study. These metrics indicate that XGBoost with SMOTE and cross-validation outperformed other models, as it displayed a better balance between precision and recall for injury prediction, and had the highest accuracy.

Machine Learning Algorithm	Pre-processing Techniques	Accuracy	Precision	F1-Score	Recall
Logistic	**SMOTE**	**46.89**	**39.5**	**0.47**	**0.47**
Linear	**PCA**	**36.9**	**39.9**	**0.37**	**0.37**
XGBoost	**SMOTE**	**94.7**	**93.9**	**0.94**	**0.94**
XGBoost	**SMOTE w/ Cross Validation**	**97.97**	**95.9**	**0.97**	**0.97**
K-Nearest Neighbors (KNN)	**Standard Scaling**	**52.1**	**0.54**	**0.48**	**0.52**
Random Forest (Optimized)	**Standard Scaling, Hyperparameter Tuning**	**57**	**0.55**	**0.57**	**0.53**
Random Forest (Optimized)	**Standard Scaling, Hyperparameter Tuning, SMOTE**	**53**	**0.51**	**0.52**	**0.53**
Support Vector Machine (SVM)	**Standard Scaling, SMOTE**	**44.44**	**0.49**	**0.46**	**0.44**

REFERENCES

Cohan, A., Schuster, J., & Fernandez, J. (2021). A deep learning approach to injury forecasting in NBA basketball. Journal of Sports Analytics, 7(4), 277–289. https://doi.org/10.3233/jsa-200529

De Leeuw, A., van der Zwaard, S., van Baar, R., & Knobbe, A. (2022). Personalized machine learning approach to injury monitoring in elite volleyball players. European Journal of Sport Science : EJSS., 22(4), 511–520. https://doi.org/10.1080/17461391.2021.1887369

Farghaly, O., & Deshpande, P. (2024). Leveraging Machine Learning to Predict National Basketball Association Player Injuries. 2024 IEEE International Workshop on Sport, Technology and Research (STAR), 216–221. https://doi.org/10.1109/STAR62027.2024.10636005

Freitas, D. N., Mostafa, S. S., Caldeira, R., Santos, F., Fermé, E., Gouveia, É. R., Morgado-Dias, F., & Dwyer, D. (2025). Predicting noncontact injuries of professional football players using machine learning. PloS One., 20(1). https://doi.org/10.1371/journal.pone.0315481

Gálvez, A., Chan, V.S., Pérez-Carabaza, S., Iglesias, A. (2025). Artificial Intelligence and Machine Learning-Based Data Analytics for Sports: General Overview and NBA Case Study. In: Blondin, M.J., Fister Jr., I., Pardalos, P.M. (eds) Artificial Intelligence, Optimization, and Data Sciences in Sports. Springer Optimization and Its Applications, vol 218. Springer, Cham. https://doi.org/10.1007/978-3-031-76047-1_5

Hecksteden, A., Schmartz, G. P., Egyptien, Y., Aus der Fünten, K., Keller, A., & Meyer, T. (2023). Forecasting football injuries by combining screening, monitoring and machine learning. Science & Medicine in Football., 7(3), 214–228. https://doi.org/10.1080/24733938.2022.2095006

Karnuta, J. M., Luu, B. C., Haeberle, H. S., Saluan, P. M., Frangiamore, S. J., Stearns, K. L., Farrow, L. D., Nwachukwu, B. U., Verma, N. N., Makhni, E. C., Schickendantz Mark, S., & Ramkumar, P. N. (2020). Machine Learning Outperforms Regression Analysis to Predict Next-Season Major League Baseball Player Injuries: Epidemiology and Validation of 13,982 Player-Years From Performance and Injury Profile Trends, 2000-2017. Orthopaedic Journal of Sports Medicine, 8(11)https://doi.org/10.1177/2325967120963046

Lövdal, S. S., Den Hartigh, R. J., & Azzopardi, G. (2021). Injury Prediction in Competitive Runners With Machine Learning. International Journal of Sports Physiology and Performance., 16(10), 1522–1531. https://doi.org/10.1123/ijspp.2020-0518

Lu, Y., Pareek, A., Lavoie-Gagne, O. Z., Forlenza, E. M., Patel, B. H., Reinholz, A. K., Forsythe, B., & Camp, C. L. (2022). Machine learning for predicting lower extremity muscle strain in National Basketball Association athletes. Orthopaedic Journal of Sports Medicine, 10(7), 23259671221111742. https://doi.org/10.1177/23259671221111742

Luu BC, Wright AL, Haeberle HS, et al. Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017. Orthopaedic Journal of Sports Medicine. 2020;8(9). doi:10.1177/2325967120953404

Naglah, A., Khalifa, F., Mahmoud, A., Ghazal, M., Jones, P., & Murray, T. "Athlete-Customized Injury Prediction using Training Load Statistical Records and Machine Learning," 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, 2018, pp. 459-464, doi: 10.1109/ISSPIT.2018.8642739.

Pu, C., Zhou, J., Sun, J., & Zhang, J. (2023). Football Player Injury Full-Cycle Management and Monitoring System Based on Blockchain and Machine Learning Algorithm. International Journal of Computational Intelligence Systems., 16(1). https://doi.org/10.1007/s44196-023-00217-6

Richter, C., O'Reilly, M., & Delahunt, E. (2024). Machine learning in sports science: challenges and opportunities. Sports Biomechanics /, 23(8), 961–967. https://doi.org/10.1080/14763141.2021.1910334

Robles-Palazón, F. J., Puerta-Callejón, J. M., Gámez, J. A., De Ste Croix, M., Cejudo, A., Santonja, F., Sainz de Baranda, P., & Ayala, F. (2023). Predicting injury risk using machine learning in male youth soccer players. Chaos, Solitons, and Fractals., 167. https://doi.org/10.1016/j.chaos.2022.113079

Ruiz-Pérez, I., López-Valenciano, A., Hernández-Sánchez, S., Puerta-Callejón, J. M., De Ste Croix, M., Sainz de Baranda, P., & Ayala, F. (2021). A Field-Based Approach to Determine Soft Tissue Injury Risk in Elite Futsal Using Novel Machine Learning Techniques. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.610210

Sarlis, V., Papageorgiou, G., & Tjortjis, C. (2024). Leveraging Sports Analytics and Association Rule Mining to Uncover Recovery and Economic Impacts in NBA Basketball. Data, 9(7), 83. https://doi.org/10.3390/data9070083

Teixeira, J. E., Encarnação, S., Branquinho, L., Ferraz, R., Portella, D. L., Monteiro, D., Morgans, R., Barbosa, T. M., Monteiro, A. M., & Forte, P. (2024). Classification of recovery states in U15, U17, and U19 sub-elite football players: a machine learning approach. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1447968

Vallance, E., Sutton-Charani, N., Imoussaten, A., Montmain, J., & Perrey, S. (2020). Combining Internal- and External-Training-Loads to Predict Non-Contact Injuries in Soccer. Applied Sciences., 10(15). https://doi.org/10.3390/app10155261

Whiteside, D., Cant, O., Connolly, M., & Reid, M. (2017). Monitoring Hitting Load in Tennis Using Inertial Sensors and Machine Learning. International Journal of Sports Physiology and Performance., 12(9), 1212–1217. https://doi.org/10.1123/ijspp.2016-0683

Wiik, T., Johansen, H. D., Pettersen, S.-A., Baptista, I., Kupka, T., & Johansen, D. (2019). "Predicting Peak Readiness-to-Train of Soccer Players Using Long Short-Term

Memory Recurrent Neural Networks," 2019 International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 2019, pp. 1-6. doi: 10.1109/CBMI.2019.8877406.

Wu, H., & Wang, L. (2022). Analysis of lower limb high-risk injury factors of patellar tendon enthesis of basketball players based on deep learning and big data. The Journal of Supercomputing., 78(3), 4467–4486. https://doi.org/10.1007/s11227-021-04029-3

Xu, T., & Tang, L. (2021). Adoption of Machine Learning Algorithm-Based Intelligent Basketball Training Robot in Athlete Injury Prevention. Frontiers in Neurorobotics., 14. https://doi.org/10.3389/fnbot.2020.620378