*Applied Mathematics on Science and Engineering, 2(2):28–36, 2025.*

**Applied Mathematics on Science and Engineering**

*Journal homepage: https://furthersci.org/journal-details.php?journal-id=49*

ISSN (E): 3048-8362

***Research article***

# Classification of Injury Risk Among Basketball Players: A Statistical Perspective

**Krishtina Baruah**[1], **Tanusree Deb Roy**[1] **and Ruhiteswar Choudhury**[1*]

[1]*Department of Statistics, Assam University, Silchar 788011, India.*
[*]*Corresponding author: sonu.choudhury2015@gmail.com*

**Article Info**

***Keywords:*** *Classification, Basketball, Linear Discriminant, and Injury Risk.*

***Received:*** *21.08.2025*
***Accepted:*** *20.10.2025*
***Published:*** *10.11.2025*



**Abstract**

Injuries are a major concern in basketball, often caused by fatigue, high training loads, and insufficient recovery. This study developed a predictive model using player data-including demographics, training intensity, recovery patterns, and fatigue levels-to assess injury risk. Among several methods tested, binary logistic regression performed best, achieving 65% sensitivity, 78% specificity, and an AUC of 0.726. Key predictors included fatigue score, training hours, and recovery days. Fatigue emerged as the strongest risk factor, increasing injury odds by 3.52 times per unit rise, while each additional recovery day reduced the risk by 92%. Anthropometric variables like age, height, and weight showed no significant influence. Linear Discriminant Analysis (LDA) further confirmed moderate separation between injured and non-injured players. These findings highlight the importance of managing fatigue and recovery to reduce injury rates in competitive basketball. Future research can improve model accuracy through real-time monitoring and integration of advanced machine learning techniques.

## 1. Introduction

Basketball is a high-intensity sport characterized by rapid accelerations, abrupt decelerations, frequent jumps, and sudden directional changes, all of which place significant stress on athletes' musculoskeletal systems [1, 2]. Also, Basket Ball ranked $7^{th}$ among the most famous and loved sport across the world with around 825 Million fans around the globe. The rules of the game is as described in Figure 1. Due to these physical demands, basketball players are at a heightened risk of acute and overuse injuries, including ankle sprains, anterior cruciate ligament (ACL) tears, patellar tendinopathy, and muscle strains [3]. Studies indicate that nearly 60% of competitive basketball players sustain at least one injury per season, with lower extremity injuries accounting for the majority of cases [4]. These injuries not only sideline athletes but also lead to long-term health consequences, financial burdens, and reduced team performance.

At the professional level (e.g., National Basketball Association), injury rates are well-documented, with studies reporting 3–5 injuries per 1,000 hours of exposure [5]. However, amateur and collegiate players face similarly high risks, with ankle sprains alone accounting for $\sim$ 25% of all basketball injuries [6]. These injuries result in missed games, long-term rehabilitation, and, in severe cases, career-ending consequences. Beyond physical tolls, injuries impose financial burdens on players, teams, and healthcare systems, emphasizing the need for proactive injury prevention strategies. Injury risk represents the probability of an athlete experiencing physical harm during athletic activities [3].

Injuries among basketball players can lead to short-term performance declines, long-term health complications, and substantial financial burdens on sports organizations and healthcare systems [7]. Despite advances in sports science and injury prevention strategies, the rate of musculoskeletal injuries - particularly in the lower extremities (ankles, knees) - remains high [8]. Identifying and classifying injury risk factors is therefore crucial for developing targeted prevention programs, optimizing player performance, and extending athletic careers Source: [9].

**How To Cite:** K. Baruah, T.D. Roy, R. Choudhury. Classification of Injury Risk Among Basketball Players: A Statistical Perspective. *Applied Mathematics on Science and Engineering* , 2(2):28–36, 2025.

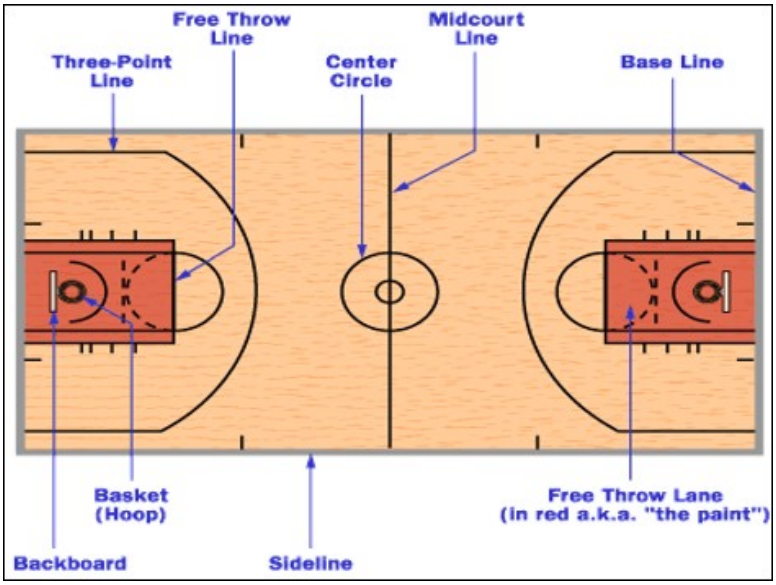


**Figure 1:** Image of Standard Basketball Court with rules

Proposes a framework that deals with descriptive, predictive as well as casual measures to underline the misuse of regression techniques on such biostatistical measures [10]. Shows that Women NBA (WNBA) players' have higher injury risk over Men NBA, however in both the cases the top injury spot is "Ankle Sprain" [11]. Suggested age, size and experience won't have much influence on injury rates [6].

Found sport-specific Countermovement Jump CMJ patterns Figure 2, distinguishing soccer and basketball players via logistic regression [12]. [11] used LDA to predict ACL injury risk, which eventually lead to better prevention and treatment. [13] found high postural sway during single-leg standing significantly increases ankle injury risk, balance training may help in such aspect Source: [9].

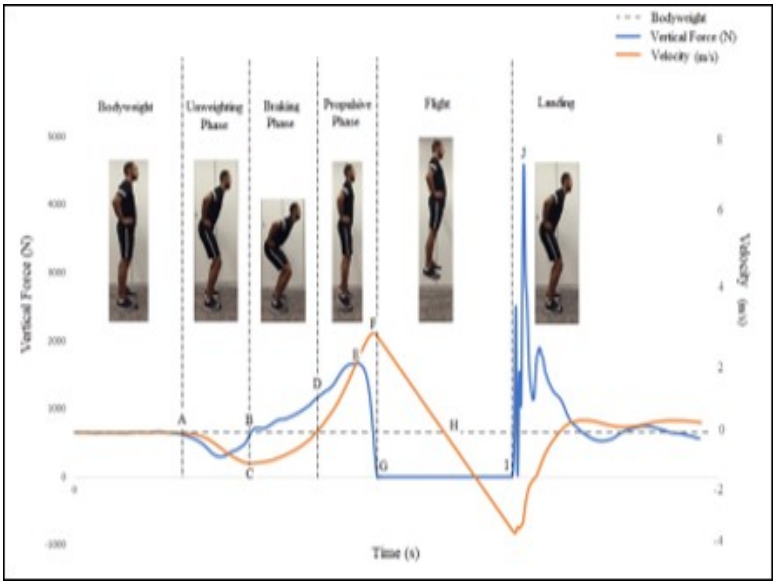


**Figure 2:** Countermovement (CMJ) Jump

Previous studies have explored various biomechanical, physiological, and situational factors contributing to basketball injuries. Studies have shown that previous injury history, muscle imbalances, fatigue, and improper landing mechanics are key predictors of future injuries [14, 15]. Additionally, external factors such as playing surface, footwear, and competition level play a significant role [16, 17]. However, existing injury risk assessment models often rely on isolated variables rather than integrating multiple risk factors into a comprehensive classification system [15]. This gap limits the ability of coaches, medical staff, and sports scientists to implement personalized injury prevention strategies effectively.

## 2. Methodology

### 2.1. Description of Data

The dataset used in this study has been collected from a secondary source [18], a widely used platform for open-access datasets. The dataset comprises of Demographic Variables (Age, Gender), Physical Attributes (Height(cm), Weight(kg)), Position (Guard, Forward, Center).

Training lead (Intensity, Weekly Training hours), Recovery metrics (No. of Days, Rest), Fatigue & Performance Score & Target Variable ("0" indicates "Not Injured", "1" indicates "Inured").

### 2.2. Binary Logistic Regression

Binary Logistic Regression is a statistical method used to model the relationship between a binary dependent variable and one or more independent variables(predictors). This technique is specially designed for situations where the outcome variable is binary. The core of binary logistic regression is the logistic function, which maps any real number into the range [0,1].

Model – Suppose $x_1, x_2, x_3, \ldots, x_n$ be "n" independent variables, then logistic regression estimates the probability (P) of an event occurring:

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

where,

- P(Y=1) is the probability of a player being injured.
- $x_1, x_2, x_3, \ldots, x_n$ are the predictor variables.
- $\beta_0$ is the intercept, and $\beta_1$, $\beta_2$,...$\beta_k$ are coefficients.

Some important assumptions of Binary Logistic Regression

- The outcome variable should have only two possible categories.
- There should be no multicollinearity (highly correlated) among the independent variables.
- The relationship between the independent variables and the log-odds of the dependent variable should be linear.
- Data Points should not be dependent on each other.

### 2.3. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a widely used technique in machine learning and statistics, primarily for classification tasks and reducing the complexity of data. Developed by Ronald Fisher in the 1930s [19, 20] LDA helps separate different categories or classes by finding the best linear combination of features. Unlike methods such as logistic regression, which directly predict class probabilities, LDA assumes that the data follows a normal distribution and that different classes share similar variability patterns [20].

Given a dataset (X,y), where each sample $x_i \in \mathbb{R}^p$ belongs to one of K classes, the data is partitioned into K groups $\pi_k$, each with $n_k$ samples. LDA seeks a linear transformation $q_i = G^T x_i$, mapping the data to a lower-dimensional space $\mathbb{R}^r$ with r<p [21].

The class mean of class k is defined as:

$$\mu_k = \frac{1}{n_k} \sum_{x_i \in \pi_k} x_i$$

The within-class scatter matrix is:

$$s_\omega = \sum_{k=1}^{k} \sum_{x_i \in \pi_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

and the between-class scatter matrix is:

$$S_\omega = \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^T$$

After projection, the scatter matrices become:

$$\overline{S_\omega} = G^T S_\omega G, \overline{S_b} = G^T S_b G$$

The LDA optimization problem is formulated to maximize the Fisher criterion, which is the ratio of the between-class scatter to the within-class scatter:

$$\max \frac{|\overline{S_b}|}{|\overline{S_\omega}|} = \max \frac{|G^T S_b G|}{|G^T S_\omega G|}$$

Assuming $S_\omega$ is non-singular, the optimal transformation $G^*$ is obtained by selecting the top r eigenvectors corresponding to the largest eigenvalues of the matrix $S_\omega^{-1} S_b$ [22].

Alternative formulations, like large margin LDA [23], maximize the minimal distance between each class center and the overall mean, requiring non-convex optimization but sometimes solvable through sequences of convex quadratic programs.

### 2.4. Model Evaluation

#### Area Under ROC Curve (AUC)

AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. Mathematically [24],

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN}$$

AUC is computed as the integral of TPR over FPR:

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

It ranges from 0 to 1, where 1 indicates perfect classification and 0.5 indicates random guessing.

#### Precision (Positive Predictive Value)

Precision quantifies how many of the instances predicted as positive are actually positive [25]:

$$Precision = \frac{TP}{TP+FP}$$

A high precision means few false positives.

#### Sensitivity

Sensitivity measures the proportion of actual positives correctly identified [25]:

$$Sensitivity = \frac{TP}{TP+FN}$$

High sensitivity means few false negatives.

#### F1-Score

The F1-score is the harmonic mean of precision and sensitivity, balancing both metrics [26]:

$$F1 - Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

It's useful when you want a single metric that balances false positives and false negatives.

#### Accuracy

Accuracy gives the overall proportion of correctly classified instances (both positive and negative) [25]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

It's a general measure but can be misleading if the dataset is imbalanced.

## 3. Results and Discussion

The study included data from [135 basketball players] basketball players, with 46.5% male and 53.5% female athletes. Basic player characteristics such as age, height, and weight showed no significant differences between injured and non-injured groups Table 1.

**Table 1:** Descriptive Statistics of Participant Variables

| Variables | Mean±Sd | Range |
|---|---|---|
| Age | 21.17±1.99 | 18-24 |
| Fatigue | 4.9 ± 2.55 | 1-9 |
| Training | 5.10 ± 2.49 | 1-9 |
| Height | 180.8±11.5 | 160-199 |
| Weight | 67.8 ± 12.2 | 40-95 |

### 3.1. Binary Logistic Regression

Since, the indicator variable is binary (Injured or Not Injured) hence, this variable is taken as the dependent variable and the rest other variables are considered to be independent variables. The fitted model under-scores various important points as the factor Fatigue-Score significantly increased injury odds, as a unit raise in Fatigue have a huge impact on injury risk (specifically by 3.5 units), indicates how crucial is the recovery for the athletes. However, the rest days per week showed a strong precautionary measure, as each added rest day reduced the injury odds by 92%. Also, the impact of training hours per week slightly increases the injury risk by 29%, shows the importance of workload management.
*Only the significant factors have been shown in the Table 2.

**Table 2:** Logistic Regression Results

| Predictors | Estimate (Std. Error) | Odds-Ratio | 95% CI | P-Value |
|---|---|---|---|---|
| Training Intensity | 0.46089 (0.21553) | 1.585 | (1.039-2.418) | 0.032 |
| Training hrs/week | 0.25677 (0.12202) | 1.293 | (1.018-1.642) | 0.035 |
| Recovery Days | -2.54538 (0.79553) | 0.078 | (0.016-0.373) | 0.001 |
| Fatigue Score | 1.25840 (0.40627) | 3.520 | (1.587-7.806) | 0.002 |

Figure 3 visually represents the odds ratio along-with their 95% Confidence Interval.

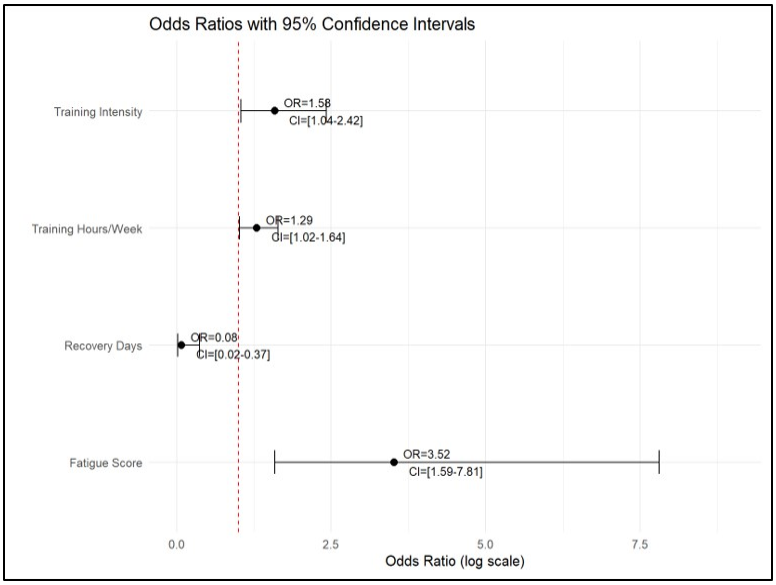


**Figure 3:** Forest Plot for Odds Ratios and 95% CI for Significant Predictors

The performance metrices for the Binary Logistic Regression Model Table 3.

**Table 3:** "Performance of the BLR model" to be included

| AUC | PRECISION | SENSITIVITY | F1-SCORE | ACCURACY |
|---|---|---|---|---|
| 0.78 | 68.4% | 65.2% | 66.7% | 72% |

The model performance measures achieved good discriminative ability with AUC value as 0.78, precision of 68.4% indicating reliable positive predictions. Sensitivity score of 65.2% indicating two-third of the injured players were correctly identified, similar to F1-score and an accuracy of 72% signifies solid overall performance of the model. The visual inspection has been carried out by the following ROC curve.

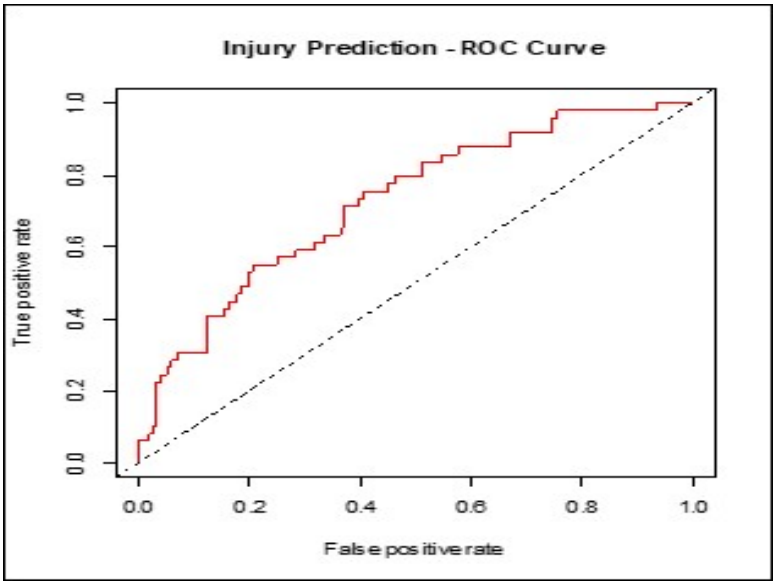


**Figure 4:** ROC Curve for Binary Logistic Regression

Figure 4 carries out the ROC analysis validates the model's capacity to identify players at elevated risk of injury with acceptable accuracy. While further optimization and validation using a larger or more diverse dataset are warranted, the current findings establish a foundation for the integration of predictive analytics in sports injury prevention protocols.

To understand the pattern of the curve, further smoothed curve has been given in the following, to get a clear picture about the ROC behavior Figure 5.

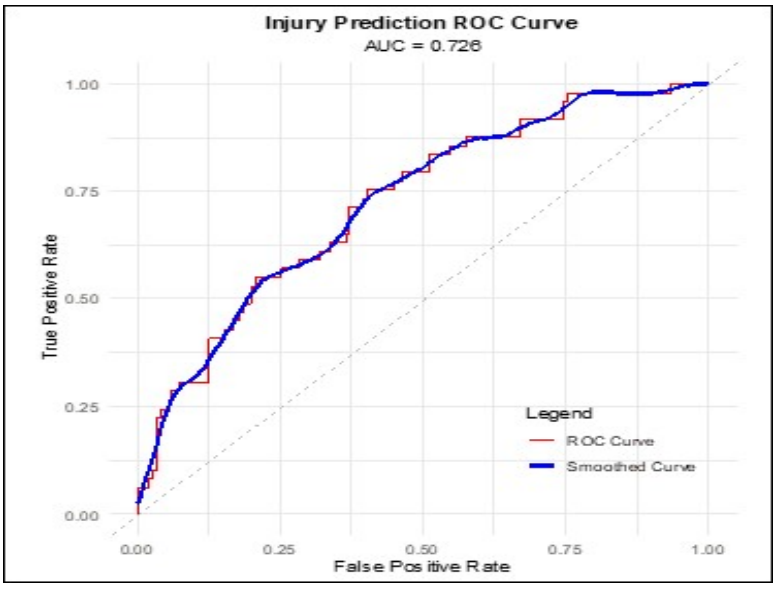


**Figure 5:** Smoothed ROC Curve for Binary Logistic Regression

## 3.2. Linear Discriminant Analysis

To further investigate injury prediction, we applied Linear Discriminant Analysis (LDA) to the same dataset. The LDA model aimed to classify players as either "injured" or "not injured" based on the available predictors. Below, we present the key results, including the discriminant coefficients, the classification table, and visual plots showing group separation and model performance. These outputs help assess which variables most effectively separate the two groups and how well the model performs compared to logistic regression Figure 6.

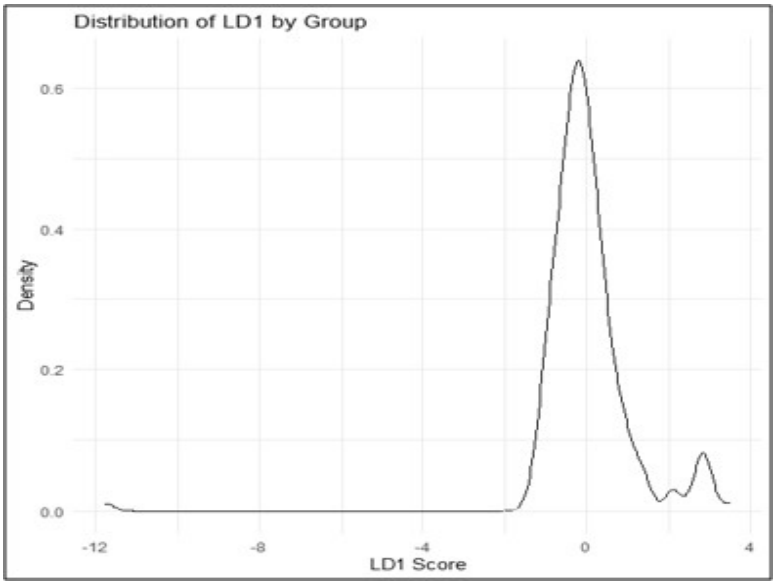


**Figure 6:** Distribution of Variables for Linear Discriminant Analysis Model

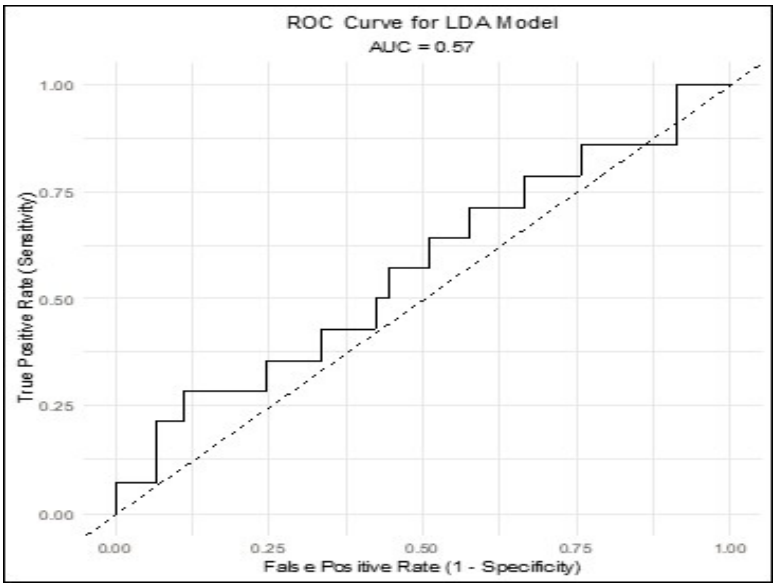


**Figure 7:** ROC curve for Linear Discriminant Analysis Model

The ROC curve Figure 7 with an AUC of 0.57 in Table 4 indicates the LDA model has limited predictive power for injuries - only performing slightly better than random chance (AUC=0.50). The nearly diagonal curve shape shows the model can't reliably differentiate between injured and uninjured athletes. While there's a minimal improvement over guessing, this performance level isn't strong enough for practical use in injury prevention. To make the model clinically or operationally useful, you'd need to improve its accuracy by adding more relevant features, collecting additional data, or trying alternative modeling techniques that might better capture the injury risk patterns.

The feature importance plot Figure 8 reveals key insights into injury risk factors based on the logistic regression model. The strongest predictors of higher injury risk are playing in the Forward position, being Male, and having an elevated ACL_Risk_Score, as indicated by their large positive coefficients. Conversely, more Recovery_Days_Per_Week, better Performance Score, and higher Team Contribution Score appear protective, showing negative associations with injury. Interestingly, while Fatigue Score moderately increases risk, factors like Age and Training_Intensity show relatively weaker effects. Notably, Height_cm and Loa_Balance_Score have minimal influence, as their coefficients hover near zero. These findings highlight that positional demands, biological sex, and ACL risk are critical injury determinants, while recovery time and performance metrics may help mitigate risk. The model suggests targeted injury prevention should prioritize load management for forwards and male athletes, while maintaining adequate recovery periods.

**Table 4:** Performace of the models

| Model | AUC | Precision | Sensitivity | F1 | Accuracy |
|---|---|---|---|---|---|
| BLR | 0.78 | 68.4% | 65.2% | 66.7% | 72% |
| LDA | 0.57 | 40% | 35% | 37% | 60% |

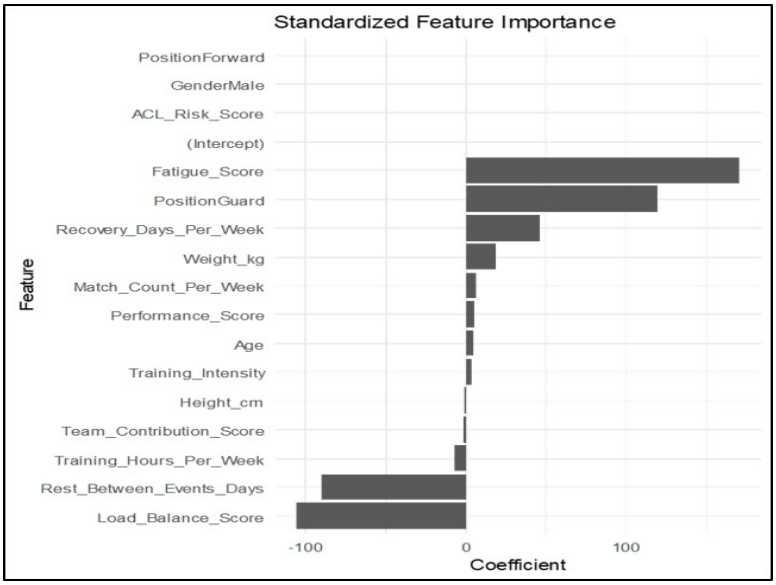


**Figure 8:** Feature Importance Plot

## 4. Conclusion

This study showed that using a binary logistic regression model can help identify which basketball players are more likely to get injured, based on key factors like fatigue, rest, and training load. We found that fatigue greatly increased the risk of injury, while adding rest days reduced that risk significantly. Interestingly, female players were found to have a higher chance of injury compared to male players, suggesting that injury prevention programs may need to be tailored by gender. Surprisingly, common characteristics like age, height, and weight were not strong predictors of injury risk.

Overall, the model worked fairly well, correctly identifying injured players about 65% of the time and healthy players about 78% of the time, with an AUC score of 0.726, showing it could reliably separate high-risk from low-risk players. Other models we tested, like LDA and QDA, performed worse.

These findings suggest that injury prevention efforts in basketball should focus more on managing player fatigue, ensuring enough rest, and carefully monitoring training intensity. While the model is a good start, it could be improved by adding more detailed information, such as past injury records or real-time movement data, possibly using wearable devices. By using data in smarter ways, we can help players stay healthier, play longer, and enjoy the game more safely.

### Article Information

**Disclaimer (Artificial Intelligence):** The author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.), and text-to-image generators have been used during writing or editing of manuscripts.

**Competing Interests:** Authors have declared that no competing interests exist.

## References

[1] A. Singh. The most popular sports in the world. *WorldAtlas*, February 2025. https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

[2] K. Franzén, H. Wäsche, and C. Pfeiffer. High-impact details of play and movements in female basketball game. *International Journal of Sports Science  Coaching*, 16(5):1050–1062, 2021.

[3] Y. Huang, C. Li, Z. Bai, Y. Wang, X. Ye, Y. Gui, and Q. Lu. The impact of sport-specific physical fitness change patterns on lower limb non-contact injury risk in youth female basketball players: a pilot study based on field testing and machine learning. *Frontiers in physiology*, 14, 2023. Article 1182755.

[4] J. Agel, D. E. Olson, R. Dick, E. A. Arendt, S. W. Marshall, and R. S. Sikka. Descriptive epidemiology of collegiate men's basketball injuries: National collegiate athletic association injury surveillance system, 1988–1989 through 2003–2004. *Journal of Athletic Training*, 42(2):194–201, 2007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1941297/.

[5] G. S. Bullock, T. Ferguson, J. Vaughan, D. Gillespie, G. Collins, and S. Kluzek. Temporal trends and severity in injury and illness incidence in the national basketball association over 11 seasons. *Orthopaedic Journal of Sports Medicine*, 9(6):1–9, 2021. doi:10.1177/23259671211004094.

[6] V. Sarlis and C. Tjortjis. Sports analytics: data mining to uncover nba player position, age, and injury impact on performance and economics. *Information*, 15(4):242, 2024.

[7] C. C. Chan, P. S. H. Yung, and K. M. Mok. The relationship between training load and injury risk in basketball: A systematic review. *Healthcare*, 12(18):1829, September 2024. MDPI.

[8] S. D. Stephenson, J. W. Kocan, A. V. Vinod, M. A. Kluczynski, and L. J. Bisson. A comprehensive summary of systematic reviews on sports injury prevention strategies. *Orthopaedic Journal of Sports Medicine*, 9(10), 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8558815/.

[9] O. Walker. Countermovement jump (cmj). *Science for Sport*, July 2016. https://www.scienceforsport.com/countermovement-jump-cmj/.

[10] J. B. Carlin and M. Moreno-Betancur. On the uses and abuses of regression models: a call for reform of statistical practice and teaching. *Statistics in Medicine*, 44(13-14):e10244, 2025.

[11] G. D. Myer, N. J. Bates, G. S. Bullock, C. Chassanidis, and T. E. Hewett. Linear discriminant analysis successfully predicts knee injury outcomes from mechanical impact simulations. *The American Journal of Sports Medicine*, 48(8):1926–1934, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7566284/.

[12] C. Chalitsios, T. Nikodelis, V. Panoutsakopoulos, C. Chassanidis, and I. Kollias. Classification of soccer and basketball players' jumping performance characteristics: A logistic regression approach. *Sports*, 7(7):163, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6681078/.

[13] C. Wang, N. Zheng, and K. M. Mok. Risk-factor analysis of high school basketball-player ankle injuries: A prospective controlled cohort study evaluating postural sway, ankle strength and flexibility. *Journal of Athletic Training*, 41(1):23–30, 2006. https://pubmed.ncbi.nlm.nih.gov/16731218/.

[14] V. Sarlis, G. Papageorgiou, and C. Tjortjis. Injury patterns and impact on performance in the nba league using sports analytics. *Computation*, 12(2):36, 2024.

[15] N. Aksović, S. Bubanj, B. Bjelica, M. Kocić, L. Lilić, M. Zelenović, others, and C. Sufaru. Sports injuries in basketball players: a systematic review. *Life*, 14(7):898, 2024.

[16] J. S. Dufek, J. A. Mercer, B. T. Bates, and G. A. Paletta. Influence of sports flooring and shoes on impact forces and performance during jump tasks. *PLOS ONE*, 13(3):e0193917, 2018. doi:10.1371/journal.pone.0193917.

[17] J. Rivera, P. Mentele, N. Wells, X. Smith, E. Rimer, and A. Stamatis. Breaking the injury cycle: The role of comprehensive integration in injury recurrence among female collegiate athletes. *International Journal of Exercise Science: Conference Proceedings*, 2(17):131, 2025.

[18] Find. open datasets and machine learning projects — kaggle. https://www.kaggle.com/datasets.

[19] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi:10.1111/j.1469-1809.1936.tb02137.x.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning: With applications in R*. Springer, 2013. doi:10.1007/978-1-4614-7138-7.

[21] H. Wang, C. Ding, and H. Huang. Multi-label linear discriminant analysis. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6316. , Berlin*. Springer, Heidelberg, 2010. doi:10.1007/978-3-642-15567-3_10.

[22] K. Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[23] Y. Chen and J. Yang. Zhang, d., liang. *J. Complete large margin linear Discriminant Analysis using mathematical programming approach. Pattern Recognition*, 46(6):1579–1594, 1995.

[24] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a reciever operating characteristic (roc) curve. *Radiology*, 143 (1):29–36, 1982. doi:10.1148/radiology.143.1.7063747.

[25] D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[26] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA. USA, 2nd edition, 1979.