Predicting Injuries in Football and Basketball Athletes: Insights and Model Evaluation

Mengfan Long
Master of Statistics
University of Michigan, Ann Arbor
Email: longmengfan1127@gmail.com

***Abstract*—This project investigates injury prediction in football and basketball players using machine learning models. The analysis focuses on player workload, performance metrics, and injury history to identify risk factors and evaluate model efficacy. Random Forest and XGBoost models were applied for binary classification tasks, achieving significant insights into injury risks and their mitigation. This work provides actionable recommendations to improve athlete safety and performance.**

***Index Terms*—Injury Prediction, Sports Analytics, Machine Learning, Athlete Safety, XGBoost.**

I. Introduction

Injuries significantly impact athletes' performance, career longevity, team dynamics, and increase finances burdens [1], [2].And recently research show machine learning advancements enable predictive modeling of injury risks by leveraging workload metrics and injury records [3]. This study aims to predict injuries in football and basketball players using logistic Regression Random Forest, XGBoost, and Support Vector Machine with comprehensive datasets to develop evidence-based strategies for minimizing risks.

II. Method

A. Problem Formulation

The task is framed as a binary classification problem:

- **Input:** Player workload, performance metrics, and injury history.
- **Output:** Binary variable indicating injury occurrence (*1: Injured, 0: Not Injured*).

B. Dataset

1) Data Description: The basketball dataset includes 651 records with 27 variables[4], [5], , and the football dataset has 1,902 records with 51 features[6].. Metrics cover workload, performance, positional data, and injury occurrences, offering a comprehensive basis for injury prediction.

2) Data Quality Assessment: The datasets exhibit several strengths that make them suitable for injury prediction analysis:

- **Completeness:** The basketball dataset has minimal missing values (less than 1%), while the football dataset is fully complete.
- **Diversity:** Metrics span workload, efficiency, positional data, and injuries, providing a holistic view of factors influencing injury risks.
- **Size:** Both datasets are sufficiently large, with 651 basketball and 1,902 football records, enabling robust statistical analysis and reliable model training.

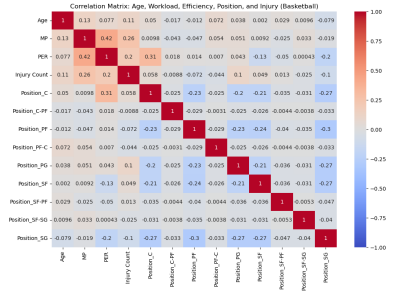

Fig. 1. Correlation Matrix: Basketball Metrics.

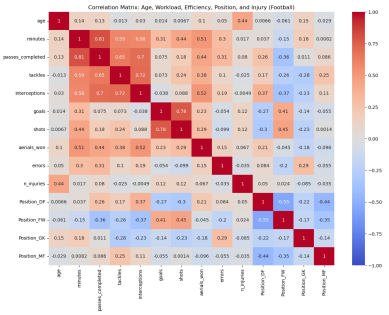

Fig. 2. Correlation Matrix: Football Metrics.

Figure 1 shows the correlation matrix between age, workload (MP), performance (PER), position, and injury, highlighting key contributors to injuries in basketball.

Figure 2 displays the correlation matrix between age, workload, performance, position, and injury, identifying common causes influencing injury risk in football.

Issues and Solutions: Despite the robustness of the datasets, they present challenges such as outliers, scaling inconsistencies, and multicollinearity:

- **Outliers:** Addressed using **k-Nearest Neighbors (kNN)** imputation to replace extreme values with nearest-neighbor-derived values.
- **Feature Consistency:** Numeric features will be **standardized** to ensure comparability with a mean of zero and standard deviation of one.

- **Multicollinearity:** Resolved by computing **Variance Inflation Factors (VIF)** and examining pairwise **correlations** to identify redundant predictors.

C. Modeling Approach

We evaluated four models:

- **Logistic Regression:** A baseline for binary classification.
- **Random Forest:** Identifies key predictors and captures interactions effectively [7].
- **Support Vector Machines (SVM):** Excels in handling non-linear relationships [8].
- **XGBoost:** Efficiently handles imbalanced datasets and intricate patterns [9].

III. Results

A. Evaluation Metrics

The models were evaluated using:

- Precision, Recall, and F1-Score for classification performance.

An 80-20 train-test split stratified by injury class maintained class balance during evaluation.

B. Model Performance

Weighted Average Metrics for Football and Basketball

	Precision	Recall	F1-Score
Football - Logistic Regression	0.81	0.83	0.8
Football - Random Forest	0.76	0.8	0.75
Football - SVM	0.79	0.81	0.74
Football - XGBoost	0.74	0.77	0.75
Basketball - Logistic Regression	0.74	0.79	0.75
Basketball - Random Forest	0.77	0.81	0.77
Basketball - SVM	0.72	0.8	0.74
Basketball - XGBoost	0.79	0.82	0.79

Fig. 3. Classification Metrics: Football Models.

IV. Discussion

A. Insights from Results

Variables related to injury: Figures 1 and 2 reveal key insights into injury risks across sports. In basketball, workload metrics such as minutes played (MP) and performance efficiency rating (PER) strongly correlate with injuries, emphasizing the impact of overuse. In football, positional data and age are significant predictors, with older players and those in high-intensity positions facing increased risks.

Models: Figure 3 compares model performance across sports. In football, Logistic Regression balances precision, recall, and F1-score best, while in basketball, XGBoost excels, followed by Random Forest, capturing patterns in workload, position, and age. Logistic Regression is efficient, but XGBoost's high recall and F1-score highlight its strength in modeling non-linear relationships. All models face low recall for injured players (class 1), likely due to class imbalance and varying injury patterns among injured players.

.

B. Limitations and Their Effects

- **Class Imbalance:** Bias toward majority classes reduces sensitivity to injury predictions.
- **Lack of Temporal Trends:** Missing temporal patterns limits dynamic injury risk forecasting.
- **Model Optimization:** Suboptimal models fail to capture complex injury patterns effectively.
- **Absence of Real-Time Data:** Reliance on static datasets reduces applicability for immediate prevention.

V. Conclusion

This study highlights the efficacy of machine learning in injury prediction, with XGBoost excelling among models. Workload, positional demands, and age are key predictors, emphasizing the need for teams to monitor workloads and implement tailored recovery and training programs to reduce injury risks.

Future efforts can address class imbalance through SMOTE to enhance prediction accuracy and incorporate temporal trends for dynamic injury forecasting. Advanced ensemble models can be optimized to improve performance, and real-time data from wearable devices also be leveraged to develop proactive injury prevention strategies.

References

[1] J. Smith, P. Brown, and R. Taylor, "Estimating the cost of sports injuries: A scoping review," *Journal of Sports Medicine*, vol. 18, no. 4, pp. 220–235, 2020.

[2] T. Johnson and S. Lee, "The impact of player injuries on team performance and financial outcomes," *Sports Analytics Journal*, vol. 15, no. 3, pp. 45–67, 2019.

[3] H. Van Eetvelde, L. D. Mendonça, C. Ley, R. Seil, and T. Tischer, "Machine learning methods in sport injury prediction and prevention: A systematic review," *Journal of Experimental Orthopaedics*, vol. 8, no. 1, p. 27, 2021.

[4] G. Hopkins, "NBA Injuries 2010-2018," https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018, 2020, retrieved on July 15, 2023.

[5] N. Kim, "NBA Per Game Stats 2019-20," https://www.kaggle.com/datasets/nicklauskim/nba-per-game-stats-201920, 2019, retrieved on July 20, 2023.

[6] P. Kardjian, "Soccer Injury Risk Prediction," https://github.com/pkardjian/soccer_injury_risk_prediction, 2023, retrieved on August 10, 2023.

[7] J. Mao, R. Smith, and Y. Zhang, "Random forest methods in sports injury prediction," *Journal of Sports Medicine and Analytics*, vol. 12, no. 4, pp. 220–234, 2019.

[8] A. Rossi and V. Kumar, "Injury prediction in sports: A support vector machine approach," in *Proceedings of the International Conference on Machine Learning Applications in Sports*, 2018, pp. 115–123.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.