Prediction of Injury Categories and Investigating the Key Factors Behind Injuries in Professional Basketball

Uttaran Bera
Masters in Computing
Dublin City University
Dublin, Ireland
uttaran.bera2@mail.dcu.ie

Omkar Sudhir Khaladkar
Masters in Computing
Dublin City University
Dublin, Ireland
omkar.khaladkar2@mail.dcu.ie

***Abstract*—Injuries, an unavoidable issue in sports, can significantly impact NBA players' careers and finances. This study addresses the limited use of machine learning in basketball injury prediction by developing a model to accurately forecast injury sites and identify key contributing factors. Focusing on professional male basketball players, the research classifies injuries based on player demographics, performance metrics, and injury history. By emphasizing class imbalance handling, the study aims to improve prediction accuracy for less common injury types. The findings will contribute to injury prevention strategies and player management in professional basketball, potentially mitigating the competitive and financial risks associated with injuries.**

***Index Terms*—Data Analytics, Machine Learning, Sports Injuries, Game Load, Body Fat Percentage, Rest Periods, Lower-Extremity Injuries, Oversampling, Injury Prevention, Player Performance, Ensemble Methods, Gradient Boosting, and Advanced Analytics.**

I. INTRODUCTION

Basketball, a globally beloved sport with over 450 million players worldwide [1], exemplifies the intricate relationship between athletic performance and injury risk. The sport's dynamic nature, characterized by rapid movements and frequent physical contact, contributes to its popularity but also increases injury susceptibility among players. In the multi-billion-dollar sports industry, optimizing athlete performance is paramount, yet injuries remain an persistent challenge with far-reaching consequences for players, teams, and the broader sports ecosystem [2], [3].

The NBA (National Basketball Association) and other professional leagues captivate global audiences, raising the stakes for all involved parties. The explosive nature of basketball gameplay predisposes players to a wide range of injuries, from acute traumas to chronic overuse conditions [4]. Despite advancements in sports science and medicine, effectively managing and preventing basketball injuries remains a formidable task. Basketball players are frequently subjected to intense physical demands that often lead to a variety of injuries. Lower extremity muscles, particularly the hamstring and adductor groups, are among the most commonly affected areas in basketball players. Shoulder and elbow injuries also pose significant challenges, with many players reporting long-term effects that persist until retirement or hinder their return to peak performance [5]. The incidence of shoulder and elbow injuries in basketball is reported at 0.82 per 1,000 athlete game exposures [6]. Notably, over 72% of NBA players who sustained shoulder injuries continued to experience ongoing issues until retirement, and 24% claimed they were unable to return to their previous level of play after the initial injury [7].

Given the substantial impact of injuries on player careers and team success, there is an urgent need for more sophisticated injury prediction and prevention strategies [8]. The emergence of machine learning (ML) techniques has revolutionized sports medicine, offering new possibilities for injury prediction. These advanced analytical methods leverage big data, integrating vast amounts of player information including physiological biomarkers, workload metrics, and comprehensive injury histories. ML algorithms can identify subtle patterns and correlations that might elude human analysts, potentially uncovering new insights into injury risk factors [9]. Meanwhile, conventional methods for monitoring athlete load have shown limited effectiveness in accurately predicting injury risks, primarily due to their focus on a restricted set of variables [10], [11]. The shift towards more advanced analytical approaches offers the potential to forecast injury-prone regions with greater precision and uncover critical contributing factors. This evolution in methodology is crucial for refining athlete management strategies and protecting the substantial investments made by sports organizations [12].

To address these challenges and capitalize on the potential of advanced analytics in sports medicine, our study makes several significant contributions to the field. The current study makes significant contributions to sports medicine and injury prevention in professional basketball. It develops and evaluates a ma-

chine learning model designed to predict injury sites in NBA players, leveraging a comprehensive dataset spanning ten seasons. This model aims to enhance injury prediction accuracy by incorporating physiological data, game statistics, and injury history, addressing limitations of conventional methods. Additionally, the research identifies crucial factors influencing injury occurrence in specific anatomical areas and tackles the challenge of class imbalance in predicting less common injury types.

The study's findings have important implications for player management in professional basketball. By pinpointing injury-prone areas and key contributing factors, it provides valuable information for implementing targeted preventative measures. This knowledge can optimize training regimens, game strategies, and recovery protocols, potentially extending players' careers and improving team performance. Thus, the research bridges the gap between advanced analytics and practical applications in sports medicine, contributing to the evolution of injury prevention and management in professional basketball.

The paper structure outlines the objectives of the study, related work in sports injury research, the methodology employed using a proper framework, the results and evalauation of feature importance in injury prediction models, and the conclusion highlighting the significance of the research findings and suggestions for further work in predicting injury types in athletes.

II. RELATED WORK

The field of sports injury prediction and prevention has seen significant advancements in recent years, with researchers employing various methods to understand and mitigate injury risks. [13] conducted a comprehensive analysis of injury data in the Australian Football League over a 21-year period, focusing on seasonal injury rates, prevalence, and recurrence rates. Their long-term surveillance approach and public release of annual injury data provided valuable insights for injury prevention strategies. Even so, the study did not explore the specific mechanisms or causes of injuries, limiting the depth of understanding regarding injury in Australian football.

While [13] focused on long-term data analysis, other researchers took different approaches to understand injury risks. [14] conducted a systematic review and analysis to identify risk factors for hamstring muscle strain injuries in sport. While their work was thorough in identifying risk factors, it did not assess the effectiveness of interventions aimed at reducing the risk of hamstring injuries.

Building upon these studies, [15] contributed to the field with a narrative review that aimed to synthesize and clarify existing knowledge. The study outlined commonly implemented methods for determining injury risk and highlighted the differences between association and prediction as they relate to injury. The authors tried to better understand injury factors and provided insights into the interpretation of various statistical measures used in injury risk assessment. However, it did not offer practical implementation guidance for the approaches discussed.

Recognizing the limitations of existing models, [16] introduced a novel perspective that challenged traditional linear approaches to injury risk. The paper proposed a new model that incorporates the consequences of repeated participation in sport, with and without injury, emphasizing that an athlete's risk factors are dynamic and can change frequently. This approach contrasted with previous models by suggesting that injury prevention strategies need to be highly individualized and responsive to an athlete's evolving risk profile rather than generalized for a group. At the same time, it did not provide a detailed critique of the limitations of previous models or specific examples of how risk factors can change over time.

The study of [5] provides an important foundation for understanding the complex nature of basketball injuries, as the sport has transformed into a more contact-oriented activity over time. This investigation, which monitored a total of 1094 players for an average of 3 seasons, highlights a notable gap in the existing literature on basketball-related injuries. Moving from theoretical models to practical applications, [17] conducted a specific study investigating the relationship between structure and injury in basketball players. The research examined 45 subjects participating in a community center basketball league, measuring various physical attributes and recording lower extremity injuries during a 16-game season. A logistic regression equation was developed using three structural variables to predict injury status, demonstrating a strong relationship between structural measures and lower extremity injury in basketball players. The predictive model correctly identified the injury status of 91% of the players and was applied prospectively to a different group of players. Yet, the study was limited by its small sample size and focus on only lower extremity injuries causing missed games, without considering other types of injuries or those that didn't result in missed games.

The study by [18] conducted a cross-sectional analysis of game load, fatigue, and injury risk in NBA athletes, using data from 627 players, 73,209 games, and 1,663 injuries across three NBA seasons (2012–2015). Logistic multilevel regression models revealed that higher game load and fatigue increased injury odds, while rest days reduced them. The results of this study are significantly different from those of previous research of [5] .

While [18] focused on recent seasons, earlier research sought to establish a long-term perspective on NBA injuries. [19] analyzed injuries and illnesses experienced by NBA players over a 10-year period from 1988-1989 to 1997-1998 seasons. The study tracked

the frequency of different types of injuries, time lost due to injuries, and game exposures, providing the first comprehensive 10-year perspective on injuries and illnesses in the NBA.

With a focus on advanced ML techniques [20] applied the same to analyze injuries in NBA data. The research collected injury data from various sources and applied the Random Forest method to build a model predicting injuries based on player performance.The findings concluded that concluded that a strong correlation exists between a player's recent performance and the likelihood of sustaining an injury in the ensuing match. This study introduced a pathway for analyzing data at both player and team levels, although it was limited by the scope of the data and the unbalanced nature of the dataset.

Recognizing the diversity of statistical approaches in the field, [21] provided a comprehensive review of different statistical methods used to model the risk of injuries in team sports.The study found that athletes who covered the most distance and made the most shot attempts in their last match were more likely to be injured in their next match. Although, the study lacked a quantitative analysis or a specific injury risk prediction model. Additionally, the dataset was highly imbalanced, with only 27 injury cases out of 13,975 data points, presenting a common challenge in predictive modeling.

The common challenge in injury prediction being the imbalanced nature of datasets, where injuries are relatively rare events researchers wanted to find a more uniform process to validate the results which have been done in other domains. Addressing this issue, [22] introduced the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic examples of the minority class rather than replicating existing ones. The researchers showed that combining SMOTE with under-sampling the majority class improved classifier performance compared to under-sampling alone. While influential in handling imbalanced datasets, the study did not thoroughly compare SMOTE to alternative methods.

Moving towards the prediction of injury risk in specific sites, [23] developed a deep learning approach to predict injuries in NBA basketball. The researchers collected year-long data on NBA player injuries and developed a model called METIC (Multiple bidirectional Encoder Transformers for Injury Classification) to predict future injuries. The authors found that multiple bidirectional encoder transformers performed significantly better than other model approaches for injury classification when considering metrics such as ROC, AUC, average precision, recall, and F1 score. However, the study was limited by the scope of available data and did not fully explore the interpretability of the deep learning model. Consistent with the findings of [18], this investigation also showed that a high playing-time load was associated with an increased incidence of all types of injuries, especially muscle and contusion-related injuries.

The relationship between previous injuries and subsequent injury risk has been a significant focus of research in sports medicine. [24] conducted a systematic review and meta-analysis to investigate the association between subsequent lower limb injuries and previous injuries. However, the study did not consider about the mechanisms underlying this association or explore the impact of different injury types on the risk of subsequent injuries.

Building on this research, [25] developed and validated a musculoskeletal readiness screening tool to predict future injuries in athletes. This tool integrated factors such as athlete concern and prior injuries for a comprehensive risk assessment. Data from 1,144 athletes across various sports showed its potential for broad use. However, the study did not examine the effects of specific training or rehabilitation programs on injury risk or compare the tool with other injury prediction models.

Expanding the scope of injury risk assessment, [26] investigated the relationship between previous injuries and subsequent injury risk in athletes, identifying the association and proposing potential neuro-muscular mechanisms. While it offered a holistic understanding, it failed to quantify the association's magnitude or explore the effectiveness of injury prevention strategies for athletes with prior injuries.

While previous studies focused on the direct relationship between past and future injuries, [27] took a different approach by investigating the association between sleep problems and injury risk among juveniles. The study pooled data from 10 observational studies involving 73,418 participants, offering a quantitative estimate of the increased injury risk associated with sleep problems. However, the findings were limited by lack of information on specific sleep problem types and injury outcomes.

Several studies have explored the integration of multiple factors in their analyses. [28] examined the relationship between subjective wellness, acute workloads, and injury risk in college football players. Data on subjective wellness (e.g., fatigue, muscle soreness, sleep quality) and training load metrics were collected from 45 Division I players over 2 years. This integration offered a comprehensive view of player readiness, training load, and injury risk. Since it focused solely on football the result however is difficult to generalize in other sports

Complementing the focus on subjective measures, [29] utilized objective data to investigate injury risk factors. This study examined the relationship between GPS and accelerometer-derived running loads and injury risk in elite Australian football players. The researchers collected data from 46 players over two competitive seasons, calculating various running load metrics and tracking the incidence of time-loss in-

juries. This novel approach allowed for a detailed quantification of player movement patterns and their association with injury occurrence. However, like the previous study, the research was confined to a single team, limiting the broader applicability of its findings.

The model developed by [23] utilized playing time as an input feature, but it did not incorporate player-specific match statistics, such as the number of field goals attempted or rebounds recorded. Also [20] showed a relationship between player-specific match statistics, such as field goals attempted and rebounds, and injury risk. Therefore, future studies should incorporate these player-level performance metrics, along with player demographics and game-level statistics, into predictive models of injury risk. This more comprehensive approach could lead to improved accuracy in forecasting injuries among basketball players.

To the best of our knowledge, no previous research has tried to use a combination of player characteristics, game metrics, and individual player match data to predict the risk of injuries at different body parts for NBA players. The goal of this study is to fill this gap in the existing literature.

III. DATA

To lay the foundation of our research, we had to take a methodical, multi-step approach. We knew we couldn't just rush in – we had to carefully consider what we were trying to achieve and how best to use the information at our disposal.

Our first task was to really get to grips with the data we had. We didn't just want to skim the surface; we needed to understand every detail and nuance hidden in those numbers. This meant diving deep into the stats, looking for patterns, and trying to make sense of the complex relationship between a player's performance, their physical characteristics, and their likelihood of getting injured.

It was like piecing together a complex puzzle, where each new bit of understanding could reshape our overall picture of NBA injuries.

A. Data Collection

As our study focused on NBA injuries from the 2010-2020 seasons, it was essential to utilize a comprehensive dataset that has been instrumental in recent research on player injuries and their impact on team performance [31]. This primary dataset, sourced from https://github.com/vsarlis/nbainjuryanalytics, formed the foundation of our investigation. To enhance our predictive models and incorporate player performance metrics and physiological differences, we utilized several key datasets. We integrated historical data from the **NBA Stats Since 1980** dataset, available on Kaggle, which provides over 60 years of season records and player statistics, offering a rich historical context for our analysis. Additionally, we included injury records from the WNBA, as detailed in the article "Who gets hurt, and how often - The Next" (The Next). This source provides insights into the frequency and types of injuries in the WNBA, helping to refine our understanding of injury patterns and their impact on player performance and team dynamics.

B. Exploratory Data Analysis

Following the data collection, we conducted thorough exploratory data analysis to verify consistency and assess the dataset's suitability for machine learning model training. This examination yielded several key insights:

- The dataset encompassed 11,225 injuries over the 10-year period, providing a substantial foundation for our training data.
- Musculoskeletal injuries emerged as the predominant category, accounting for 75.6% of all recorded injury types. Given the inherent complexity in predicting disease-caused injuries, our study focused exclusively on this musculoskeletal subset.
- Within the musculoskeletal category, lower extremity injuries proved to be the most prevalent, representing 69.82% of total injuries. This classification includes injuries to the hip, thigh, knee, calf, fibula, shin, ankle, heel, foot, and toes..
- Trunk injuries (involving the chest, abdominal, thoracolumbar, and pelvis areas) and upper extremity injuries (encompassing the shoulder, upper arm, forearm, elbow, hand, thumb, and fingers) each constituted approximately 14.5% of the total injuries.)
- While no distinct trend in injury numbers was observed over the decade, it's important to note the anomaly of the 2020 season, coinciding with the COVID-19 pandemic and the implementation of the NBA bubble, which saw a reduction in active players.

IV. METHODOLOGY

Throughout the research process, the study needed to go back and forth in multiple steps to get the best results to answer the research questions :

1) Can we accurately predict the type of injury a player is most likely to experience based on his performance and physiology?

2) What factors play the most important role in NBA injuries?

A. Data Processing and Preparation

The data integration process was challenging due to the heterogeneous nature of the datasets, each with unique metadata structures. We encountered numerous inconsistencies, including varying row and column formats and diverse representations of player names, necessitating a sophisticated approach to data cleaning and merging. Advanced dictionary-based techniques

and deep domain knowledge of basketball teams and players were employed to integrate the datasets successfully. The merging process utilized three primary keys: player names, ages, and team affiliations for respective seasons.

Upon deeper exploration of the data, we encountered a common challenge in research: class imbalance. The original classification system divided injuries into five categories based on anatomical precision. This resulted in statistical challenges as some injury types were more prevalent than others. For simplification, instead of focusing on detailed, specific categories, we consolidated injuries to the head, neck, and trunk into one broader category. This adjustment addressed the class imbalance problem while maintaining classification relevance to the body's physiology.

The decision to retain separate classifications for lower and upper extremity injuries was deliberate. These injury types differ significantly in their mechanisms, treatment approaches, and effects on player mobility. For example, a shoulder injury presents distinct challenges compared to a knee injury. Therefore, ensuring that the new categories were not only statistically balanced but also physiologically meaningful required a careful balance between statistical optimization and real-world applicability. The result was a streamlined, three-class system.

B. Feature Engineering

During feature analysis for the Machine Learning models, significant multicollinearity was observed between certain features. The study, involving 858 unique players with recorded injuries over a decade, revealed a high correlation of 82% between player height and weight. Similarly, minutes played and field goals attempted per game exhibited a high correlation value of 87.1%, which logically follows as more minutes on the court provide more opportunities for basket attempts.

To address these issues and prevent all correlated features to be used in the same model, we ultimately opted for a combination of features and reduce dimensionality. PCA was not employed due to the class imbalance present in the data [32].

To resolve the correlation between player height and weight, we evaluated NBA draft combine stats and other studies [33]. Based on this analysis, Body Fat Percentage was used as the standard combined measure in place of both height and weight. The Body Fat Percentage (BFP) was calculated using an empirical formula for adults, utilizing the Body Mass Index (BMI) with gender as a binary variable as follows [34]:

$$(\text{Adult BFP}) = 1.39\times\text{BMI}+0.16\times\text{age}-10.34\times\text{gender}-9 \quad (1)$$

where BMI is defined as:

$$\text{BMI} = \frac{\text{player weight (kg)}}{\text{player height (m)}^2} \quad (2)$$

Although this method for calculating Body Fat Percentage may not be state-of-the-art, it was applied consistently across all players, providing a standardized and valuable training feature. The resulting feature demonstrated a normal distribution, indicating its suitability for inclusion in our training model, as shown in Fig. 1 below, which illustrates the distribution of Body Fat Percentage among NBA players.

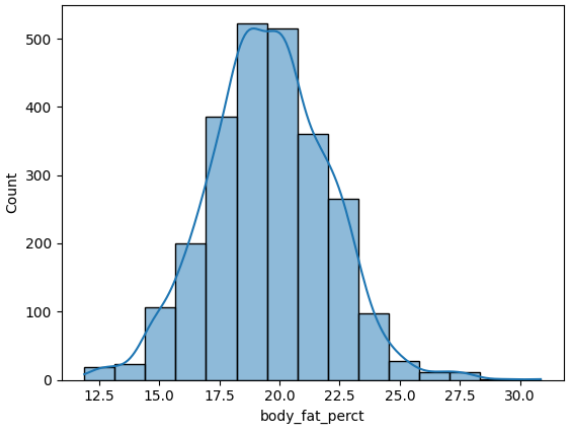

Fig. 1: Distribution of Body Fat percentage of players

To address the correlation between field goals attempted and minutes played, we introduced a linear combination of these variables, denoted as Field Goals Attempted per Minute (FGA.PM), calculated as:

$$\text{FGA.PM} = \frac{\text{Field Goals Attempted}}{\text{Minutes Played}} \quad (3)$$

This new feature effectively captures the player's shooting frequency while accounting for their time on the court, providing a more nuanced perspective on player performance.

The main novelty of our study lies in exploring how features like previous injury records and physiological characteristics along with in game statistics affect injury locations. Feature engineering was initiated to include injury history, incorporating information about past injuries in major anatomical areas (lower extremity, upper extremity, and head-neck-trunk). A new feature, total_injury_count, was created to capture overall injury history, combining Upper_ext_count, Lower_ext_count, and Head-neck-trunk_count. This historical data, along with player performance metrics such as age, body fat percentage, field goals attempted per minute, minutes played, and total rebounds, is crucial for accurately classifying injury areas.

C. Sampling Techniques

To address class imbalance in the datasets, various sampling techniques were employed to enhance model performance. Three primary methods were used: under-sampling, over-sampling, and SMOTE-NC.

1) Under-Sampling: Under-sampling was utilized to balance the dataset by reducing the number of samples from the majority class to match that of the minority class. While it simplifies the data points it may not represent the majority class effectively thus limiting the model's ability to generalize.

2) Over-Sampling: Over-sampling was applied to increase the number of samples in the minority class to achieve balance with the majority class. However this method is susceptible to over fitting while training.

3) SMOTE-NC: SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous data) [35], [36] was employed to address class imbalance by generating synthetic examples for the minority class through the analysis of nearest neighbors in the feature space. Unlike standard SMOTE, SMOTE-NC handles both continuous and categorical variables effectively, making it particularly suited to our dataset which includes numerical features like body fat percentage and minutes played, as well as categorical features like player position. By generating synthetic samples that closely resemble the real data, SMOTE-NC enhances the robustness of our predictive models. Handles datasets with mixed data types effectively, benefiting our dataset with both numerical (e.g., body fat percentage, minutes played) and categorical features (e.g., player position). This is the state of the art method and the final model results are the most realistic in this case.

D. Feature Selection

Effective feature selection is crucial for building better predictive models, as it helps identify the most relevant variables contributing to the model's performance. We employed two distinct approaches for feature selection based on the inclusion of injury statistics and encoding techniques for categorical variables.

1) Feature Selection Without Injury Stats: Initially, we focused on features excluding specific injury statistics. This subset of features includes: **Age**, **Body Fat Percentage (body_fat_perct)**, **Field Goals Attempted per Minute (FGA_per_MP)**, **Minutes Played (MP)**, **Total Rebounds (TRB)**.

2) Feature Selection With Injury History: In a more comprehensive approach, we incorporated injury statistics to enhance the model's ability to predict injury-related outcomes. The extended feature set includes:**Upper Extremity Injury Count (Upper_ext_count)**, **Lower Extremity Injury Count (Lower_ext_count)**, **Head-Neck-Trunk Injury Count (Head-neck-trunk_count)** together with all the previous features.

A separate model was also trained combining the injury records as one into **Total Injury count (prev_injuries_count)**

Including these injury statistics offers a deeper understanding of how past injuries might affect current performance and injury risks, leading to more accurate predictions.

E. Machine Learning Algorithms

Before training the models, it was essential to convert the target variable of Major Anatomical Category into discrete labels: "0 - Head-neck-trunk", "1 - Lower extremity", and "2 - Upper extremity". Additionally, player position, a crucial categorical variable, was one-hot encoded to be compatible with the ML models, and the target class was label encoded. The models were trained on under-sampled, over-sampled, and SMOTE-NC-created data for comparison and better evaluation. Weighted F1 scores, accuracy, and recall metrics were used to evaluate the models' performance, confirming the effectiveness of our approach.

1) Decision Tree: The Decision Tree algorithm without any hyper parameter tuning was selected as the baseline model for our analysis. It was chosen for its simplicity and interpretability, providing a clear visual representation of how decisions are made based on the features. Based on a 80-20 train-test split this model helped us evaluate the effectiveness of more complex algorithms by providing a straightforward baseline for comparison.

2) Random Forest: The Random Forest classifier was employed as an advanced model, leveraging ensemble methods to improve classification accuracy. We used a Random Forest with a parameter of 100 trees and entropy as the criterion for splitting. Random Forest was selected based on its proven efficiency in accurately classifying injuries in various sports domains, as highlighted in previous research . The ensemble nature of Random Forest allows it to effectively manage feature importance which was used later to identify the key contributing factors for different injuries.

3) XGBoost Classifier: The XGBoost Classifier was incorporated for its high performance and accuracy in boosting algorithms. With a default boosting of 100 times and mlogloss as the evaluation loss metric and multi:softprob for multi class classification as the hyperparameters this algorithm was included to address the challenges of classifying minor injury classes and to optimize performance. XGBoost has demonstrated its effectiveness in other sports and athlete injury classifications [37], making it a suitable choice for our study.

The choice of ensemble methods over supervised learning methods was driven by evidence showing their effectiveness in accurately classifying injuries across different sports and athletic contexts, as noted in [38].

F. Neural Network Architecture

We also implemented a Neural Network to enhance the prediction of injury outcomes, leveraging its capability to model complex patterns and interactions in data. The network was structured with an input layer of 128 units, followed by three hidden layers with 64, 32, and 16 units, respectively, all using ReLU

activation. The output layer utilized a softmax function to classify injuries into three categories: “Head-neck-trunk”, “Lower extremity”, and “Upper extremity.” The model was compiled using the AdamW optimizer with a learning rate of 0.001 and trained for 50 epochs with a batch size of 32. Data preprocessing included one-hot encoding of categorical features and concatenation with numerical features.

V. Results and Evaluation

Previous research has shown success in predicting and classifying injury areas among athletes, particularly basketball players. However, these studies were limited by suboptimal performance in F1 score, recall, and AUC ROC curve. This highlights that while accuracy can be valuable, it may not suffice in multiclass classification scenarios, where alternative assessments often prove more illuminating. As per our study the yearly injuries are biased a lot towards the lower extremity category over the duration of 10 years as shown in Fig. 2 below:

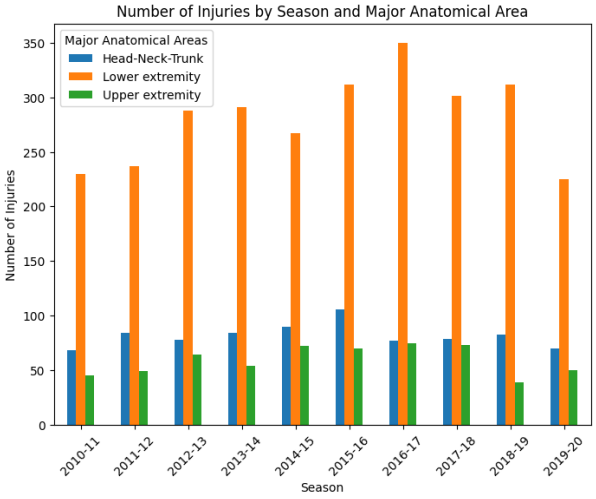

Fig. 2: Number of Injuries by Season and Major Anatomical Area

In such cases, models might show high accuracy in predicting the predominant class, but other critical factors like the area under the ROC curve for true and false positives could be diminished, signifying model bias [39]]. The present study aimed to mitigate this tendency towards majority class classification by training each model using three distinct sampling techniques, with outcomes presented in the following sections.

In the following tables the evaluation metrcis are laid out for the two scenarios the models were trained and tested on : 1) without considering previous injury records and 2) taking previous injury records specifically the injury location into account.

We can see that undersampling significantly reduced the dataset, making accurate test data prediction challenging. Nevertheless, boosting methods outperformed the baseline decision tree model in both accuracy and F1 score. Recall rates were 35% for no injury and 69% for injuries. Random Forest and XGBoost models demonstrated higher AUC than the decision tree, particularly in predicting upper extremity injuries. Including previous injuries as a feature notably improved performance metrics across all models, with XGBoost achieving the highest accuracy and F1 score, followed by Random Forest and Decision Tree. This feature proved valuable for enhancing model performance. While these findings are insightful, caution in interpretation remains necessary due to the effects of undersampling on the dataset.

TABLE I: UNDER SAMPLING

Models	Metric	Without Injury feature	With Injury feature
Decision Tree	Accuracy	34%	68%
	F1 Score	0.34	0.684
Random Forest	Accuracy	34.3%	69.6%
	F1 Score	0.34	0.695
XG Boost	Accuracy	36%	70.9%
	F1 Score	0.35	0.709

TABLE II: OVER SAMPLING

Models	Metric	Without Injury feature	With Injury feature
Decision Tree	Accuracy	76.3%	93%
	F1 Score	0.76	0.931
Random Forest	Accuracy	80.2%	91.6%
	F1 Score	0.80	0.904
XG Boost	Accuracy	76.6%	91.1%
	F1 Score	0.77	0.909

Including the previous injuries feature under oversampling conditions significantly improved performance across all three models. The Decision Tree showed the most dramatic improvement, with accuracy increasing from 76.3% to 93% and F1 Score from 0.76 to 0.931, reaching the highest values among the models. Random Forest and XG Boost also benefited from this feature, though less dramatically. Overall recall improved from 77% without previous injuries to 92% with them. However, these higher scores don’t necessarily indicate oversampling as the optimal technique, as it raises overfitting concerns due to duplicate records. Notably, despite the Decision Tree’s higher accuracy and F1 Score, XG Boost and Random Forest demonstrated better AUC ROC curves, suggesting that boosting and ensemble methods may be preferable for model selection. This underscores the importance of considering multiple performance metrics when evaluating models, especially in scenarios involving class imbalance and oversampling techniques.

TABLE III: SMOTE - NC

Models	Metric	Without Injury feature	With Injury feature
Decision Tree	Accuracy	60%	79%
	F1 Score	0.60	0.794
Random Forest	Accuracy	68%	85.1%
	F1 Score	0.68	0.841
XG Boost	Accuracy	70.6%	86.1%
	F1 Score	0.71	0.861

The application of SMOTE-NC, coupled with the inclusion of the previous injuries feature, significantly enhances model performance across all three models in this multi-class classification task. XG Boost emerges

as the top performer, achieving 86.1% accuracy and an F1 Score of 0.861, followed closely by Random Forest (85.1% accuracy, 0.841 F1 Score), while the Decision Tree improves but remains behind at 79% accuracy and 0.794 F1 Score. Recall rates show improvement, with 66% for no injuries and 83% for injuries. Although these results are lower than those from oversampling, they surpass undersampling outcomes, suggesting SMOTE-NC offers a balanced approach that mitigates overfitting risks while preserving valuable information and offers the most realistic results. For a better understanding the Aarea under the Curve in the ROC graph below demonstrates that the classification of the injury categories is best in the case of XG Boost.

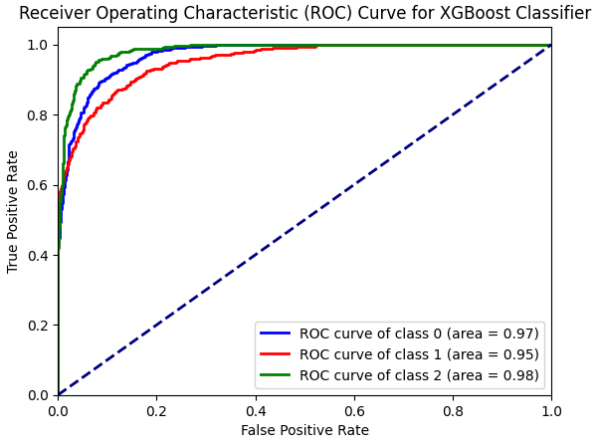

Fig. 3: Area Under Curve Values for the 3 injury classes

We had initially run only simple Machine Learning Models for better understanding the feature importance and interoperability. However this study extends to compare the results of a simple model with a deep neural network architecture the results of which are shown below:

TABLE IV: Neural Networks

Models	Metric	Without Injury feature	With Injury feature
N.Network	Accuracy	40%	79%
	F1 Score	0.38	0.79

The recall for without injury stats for Neural Networks comes around 40% and that for with injuries inclusion comes around 79%. This model is trained on the SMOTE-NC samples only using the similar test size of 20% of the total data points While this confirms that the injury history is a significant feature influencing subsequent future injuries, it also is interesting to note that the weighted average recall is much better for this neural network architecture than the other machine leraning models. On the downside however, it is not possible to interpret the importance of the features to better understand the contributing factors for injuries.

The models combining the injury history into one did not however fare better than the separate injuries thus confirming NBA players are more susceptible to similar kind of injuries in their playing career.

Key Injury Contributors:

While the answer to our first question of our research study was elaborated on by comparing different sampling techniques and machine learning models, the answer to the second question is two fold. The models trained without using injury history had significantly different factors responsible for identifying injury categories than those trained with injury records. Deriving the feature importance from random forest in all the 3 sampling scenarios it was observed that body fate percentage was the leading contributor in the models trained only with physiological and in game statistics with an average around 21% across the 3 models. The other feature which stood out was Field Goals Attempted per minute with an average of 18%. The total rebounds and minutes played were also significant as key injury factors. It is noteworthy however, the XG Boost model laid out that player positions were one of the more contributing factors signifying players at different positions suffer different kinds of injuries. Considering the injury history however there is a positive shift towards the specific injury counts as the most contributing feature. In all the sampling cases each injury category contributes on an average of 20% as the most important feature. The rest are similar to those of the previous case. Body fat percentage and field goals attempted with an average of more than 10% are the next most important features in case of Random Forest, while the player positions play the important role for Gradient Boosting with around 11% weight. This insight supports the development of tailored plans and load management strategies for players in different positions to prevent injuries.

5-Fold Cross Validation

In order to ensure the model isn't overfitting we carried out cross validation experiments. This technique involves partitioning the dataset into five equal-sized folds. In each iteration, four folds are used for training and one for validation, repeating this process five times to ensure every data point is used for both training and validation. The mean cross-validation results were 66% for undersampling, 87% for oversampling, and 78% for SMOTE-NC. This approach minimizes data variability and provides a reliable assessment of model performance on unseen data, confirming that advanced models and the inclusion of significant features like previous injuries enhance predictive performance in multiclass classification scenarios.

During this set of experiments it was observed that the recall for the Lower extremity classes were the least, as was the case before the cross validation. However it is quite significant to note why this situation arises. The one thing that can be related to this is that the synthetic samples are developed from the majority

class which is in fact the lower extremity class and hence the models aren't able to predict with certainty the same injury category. The other two classes have higher recall denoting the True Positives for the lower extremity injuries become difficult to predict as it is the majority class.

Finally after all the observations of the results we can say that , the accuracy and F1 score for each model drops from oversampled training data but is significantly higher than the undersampled one. This suggests this is the most consistent and reliable sampling technique to be used in this scenario for multi class classification as shown in Table III.This provides a valuable comparative study between various models and sampling techniques. The ROC curve in the case of SMOTE-NC sampled data is much higher for Random Forest and XG Boost (shown in fig 2.) as compared to the baseline model. This explains that the baseline models might classify most of the test data accurately, it doesn't train after a point and is unable to improve. While the advanced models might take longer time to train, they are more efficient while handling the test data and precisely classifying the major anatomical areas.

VI. Conclusion

This research conducts a comprehensive analysis of methods to address class imbalance in multi-class classification problems, particularly in predicting injury areas among professional basketball players. Utilizing machine learning models, we found that explainable features greatly improve the accuracy and precision of predicting common injury sites. Notably, including previous injuries yielded significantly better results than excluding them. This has important implications for optimizing player substitutions, managing rest periods, and preventing recurring injuries. However, challenges remain, such as using SHAP values with one-hot encoded categorical variables, indicating the need for advanced techniques to better understand feature importance. Future research should prioritize differentiating injuries based on player positions for more precise predictions.

Additionally, this study marks an important step towards forecasting injury types in athletes by considering various performance metrics, physical attributes, and player demographics. Expanding this research to predict both minor and major injuries in NCAA college basketball and the WNBA could yield deeper insights into the impact of player characteristics, such as body fat percentage influenced by gender, on injury risks. While this study lays a foundation for advanced injury prediction models, further research is essential to refine these models and explore additional factors. Such progress could greatly enhance injury prevention strategies and player management in professional basketball, contributing to longer and healthier athletic careers.

References

[1] Fédération Internationale de Basketball (FIBA) Quick Facts. 2019. Available online: www.fiba.com (accessed on 28 February 2021).

[2] J. W. Powell , T. P. Dompier, "Analysis of injury rates and treatment patterns for time-loss and non-time-loss injuries among collegiate student-athletes," *J. Athl Train*, vol. 39, no. 1, pp. 56–57, 2004.

[3] B. C. Werner *et al.*, "Acute gastrocnemius-soleus complex injuries in National Football League athletes," *Orthop J Sports Med*, vol. 5, no. 1, p. 2325967116680344, 2017.

[4] J. B. Taylor *et al.*, "Activity Demands During Multi-Directional Team Sports: A Systematic Review," *Sports Med*, vol. 47, no. 12, pp. 2533-2551, Dec. 2017. doi: 10.1007/s40279-017-0772-5.

[5] M. C. Drakos *et al.*, "Injury in the National Basketball Association: a 17-year overview," *Sports Health*, vol. 2, no. 4, pp. 284–290, 2010.

[6] Martin CL, Arundale AJH, Kluzek S, Ferguson T, Collins GS, Bullock GS, "Characterization of rookie season injury and illness and career longevity among National Basketball Association players," *JAMA Netw Open*, vol. 4, no. 10, p. e2128199, 2021.

[7] C. Gohal *et al.*, "Impact of shoulder injuries on quality of life for retired National Basketball Association players: a survey study," *Int J Sports Exerc Med*, vol. 5, p. 154, 2019.

[8] Y. Huang *et al.*, "A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning," *Front Physiol*, vol. 13, p. 937546, 2022. doi: 10.3389/fphys.2022.937546.

[9] H. Van Eetvelde *et al.*, "Machine learning methods in sport injury prediction and prevention: a systematic review," *J Exp Orthop*, vol. 8, no. 1, p. 27, Apr. 2021. doi: 10.1186/s40634-021-00346-x.

[10] F. M. Impellizzeri *et al.*, "Training Load and Its Role in Injury Prevention, Part I: Back to the Future," *J Athletic Training*, vol. 55, no. 9, pp. 885–892, 2020.

[11] A. Hulme *et al.*, "Towards a complex systems approach in sports injury research: simulating running-related injury development with agent-based modelling," *Br J Sports Med*, vol. 53, no. 9, pp. 560–569, 2019.

[12] M. Buchheit, "Houston, We Still Have a Problem," *Int J Sports Physiol Perform*, vol. 12, no. 8, pp. 1111–1114, 2017.

[13] J. W. Orchard, H. Seward, J. J. Orchard, "Results of 2 decades of injury surveillance and public release of data in the Australian Football League," *Am J Sports Med*, vol. 41, pp. 734–741, 2013. DOI: 10.1177/0363546513476270

[14] G. Freckleton, T. Pizzari, "Risk factors for hamstring muscle strain injury in sport: a systematic review and meta-analysis," *Br J Sports Med*, vol. 47, pp. 351–358, 2013. DOI: 10.1136/bjsports-2011-090664

[15] A. McCall, M. Fanchini, A. J. Coutts, "Prediction: the modern day sports science/medicine 'Quest for the Holy Grail'," *Int J Sports Physiol Perform*, vol. 10, pp. 1–11, 2017. DOI: 10.1123/ijspp.2017-0137

[16] W. H. Meeuwisse, H. Tyreman, B. Hagel, C. Emery, "A dynamic model of etiology in sport injury: the recursive nature of risk and causation," *Clin J Sport Med*, vol. 17, pp. 215–219, 2007. DOI: 10.1097/jsm.0b013e3180592a48

[17] J. P. Shambaugh *et al.*, "Structural measures as predictors of injury for basketball players," *Med Sci Sports Exerc*, vol. 23, no. 5, pp. 522-7, May 1991.

[18] M. Lewis, "It's a Hard-Knock Life: Game Load, Fatigue, and Injury Risk in the National Basketball Association," *J Athl Train*, vol. 53, no. 5, pp. 503-509, May 2018.

[19] C. Starkey, "Injuries and illnesses in the National Basketball Association: a 10-year perspective," *J Athl Train*, vol. 35, no. 2, pp. 161–167, 2000.

[20] W. Wu, "Injury Analysis Based on Machine Learning in NBA Data," *Journal of Data Analysis and Information Processing*, 2020.

[21] J. D. Ruddy, S. J. Cormack, R. Whiteley, M. D. Williams, R. G. Timmins, D. A. Opar, "Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches," *Front Physiol*, vol. 10, p. 829, Jul. 2019, doi: 10.3389/fphys.2019.00829.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[23] A. Cohan *et al.*, "A Deep Learning Approach to Injury Forecasting in NBA Basketball," pp. 277–289, 2021.

[24] L. A. Toohey *et al.*, "Is subsequent lower limb injury associated with previous injury? A systematic review and meta-analysis," *Br J Sports Med*, vol. 51, no. 23, pp. 1670–1678, 2017.

[25] A. C. Terry *et al.*, "The musculoskeletal readiness screening tool-athlete concern for injury & prior injury associated with future injury," *Int J Sports Phys Ther*, 2018. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6088131/.

[26] J. Fulton *et al.*, "Injury risk is altered by previous injury: a systematic review of the literature and presentation of causative neuromuscular factors," *Int J Sports Phys Ther*, vol. 9, no. 5, pp. 583–595, 2014.

[27] Y. B. Wang *et al.*, "Sleep problems and injury risk among juveniles: A systematic review and meta-analysis of observational studies," *Sci Rep*, vol. 7, no. 1, p. 9813, Aug. 2017.

[28] J. A. Sampson *et al.*, "Subjective Wellness, Acute: Chronic Workloads, and Injury Risk in College Football," *J Strength Cond Res*, vol. 33, no.

[29] M. J. Colby *et al.*, "Accelerometer and GPS-derived running loads and injury risk in elite Australian footballers," *J Strength Cond Res*, vol. 28, no. 8, pp. 2244-2252, Aug. 2014. doi: 10.1519/JSC.0000000000000362. PMID: 25054573.

[30] S. Studer, *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, pp. 392–413, 2021. DOI: 10.3390/make3020020

[31] V. Sarlis, V. Chatziilias, C. Tjortjis, and D. Mandalidis, "A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance," *Information Systems*, vol. 99, pp. 101750, 2021. DOI: 10.1016/j.is.2021.101750.

[32] T. Maruthi Padmaja, Bapi S. Raju, Rudra N. Hota, and P. Radha Krishna, "Class Imbalance and Its Effect on PCA Preprocessing," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 4, no. 3, pp. 272–294, 2014. DOI: 10.1504/IJKESDP.2014.065278.

[33] Jelena Popadic Gacesa,*et al.*, *Body mass index and body fat content in elite athletes*, *Exercise and Quality of Life*, vol. 3, pp. 43–48, 2011.

[34] Bruso, J. (2017, July 18). How to Convert BMI to Body Fat Percentage. Retrieved from : https://www.livestrong.com/article/173592-how-to-convert-bmi-to-body-fat-percentage/

[35] K. Li, W. Zhang, Q. Lu and X. Fang, "An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree," in *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, Beijing, China, 2014, pp. 34-38, doi: 10.1109/IIKI.2014.14.

[36] Dina Elreedy , Amir F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences*,vol. 505, pp. 32–64, 2019. DOI: 10.1016/j.ins.2019.07.070.

[37] M. Calderón-Díaz,,*et al.*, "Explainable Machine Learning Techniques to Predict Muscle Injuries in Professional Soccer Players through Biomechanical Analysis," *Sensors*, vol. 24, p. 119, 2024. DOI: 10.3390/s24010119.

[38] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, D. Bowman, "Application of machine learning to construction injury prediction," *Automation in Construction*, vol. 69, pp. 102–114, 2016.

[39] P. A. Flach , M. Kull, "Precision-Recall-Gain Curves: PR Analysis Done Right," in *Neural Information Processing Systems*, 2015. Available at: https://api.semanticscholar.org/CorpusID:937625.