International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Injury Prediction Using Machine Learning

Prof. Mamatha A[1], Vinay P Potdar[2], Srujan Jaka[3], Shravan K N[4], Vivek V Korti[5]

[1] Assistant Professor Computer Science and Engineering Dayananda Sagar Academy of Technology & Management Bengaluru, India
mamatha@dsatm.edu.in
[2] Student, 3th Year, B.E Computer Science and Engineering Dayananda Sagar Academy of Technology & Management Bengaluru, India
1dt22cs180@dsatm.edu.in
[3] Student, 3th Year, B.E Computer Science and Engineering Dayananda Sagar Academy of Technology & Management Bengaluru, India
1dt22cs161@dsatm.edu.in
[4] Student, 3th Year, B.E Computer Science and Engineering Dayananda Sagar Academy of Technology & Management Bengaluru, India
1dt22cs149@dsatm.edu.in
[5] Student, 3th Year, B.E Computer Science and Engineering Dayananda Sagar Academy of Technology & Management Bengaluru, India
1dt22cs185@dsatm.edu.in

ABSTRACT-

Sports Injury has been the biggest threat amongst the players, teams and sporting management since the advent of the sporting culture. Our objective here is to use this data to examine factors that may contribute to lower extremity injuries. In this paper, we present machine learning based solutions to predict the injury risk factor and level of injury to sportsman . by collecting data from the devices like smartwatches which Track your hikes, hunts and missions on smartwatches built to withstand abuse in the field. Predicting injury before hand would be a huge help to the players, ultimately revolutionizing the sports industry and knowing the resting period in advance, would help teams strategize in a better manner for can take prior measures in order to prevent those injuries.

Keywords- predictive modeling, classification, regression, multi-label classification, injury risk factors, injury prevention, recovery period estimation, vulnerable body parts,wearable devices, smartwatches, GPS, player history, environmental conditions, physiological data.

I. INTRODUCTION-

Billions of dollars are invested in the sports industry to enhance sports performance and reduce the risk of injuries. A lot of factors influence the risk factor in an injury. The given dataset helps us analyze the playing conditions like field type, weather, and temperature, along with the various plays of every single player, including their movements, position, turf, speed, etc. Our classifier takes into account all these factors to predict if an injury would occur or not. Our regressor model then estimates level of injury. Further, we also predict the how much injury happened whether it is minor major.
The sports industry invests billions of dollars in research and development to enhance athletic performance and reduce the risk of injuries. A variety of factors influence the likelihood of injuries, making it essential to analyze the interplay between playing conditions, player movements, and game dynamics. The dataset used for this purpose includes detailed information about field types, weather conditions, and temperature, alongside individual player data such as movements, positioning, speed, and interactions with the turf. By leveraging this data, advanced predictive models have been developed. The classifier focuses on determining the probability of an injury occurring by identifying patterns and conditions associated with injury risk. Meanwhile, the regressor estimates the severity of injuries, providing numerical values that help gauge their impact. Furthermore, the model also categorizes injuries into minor, moderate, and major, predicting the extent of damage and informing recovery strategies. These insights have significant applications, including injury prevention through optimized training protocols, player management by tailoring fitness plans and rest schedules, and infrastructure improvements like better field designs and real-time monitoring systems. Additionally, these predictions support personalized rehabilitation programs, ultimately leading to safer playing environments, enhanced performance, and prolonged athletic careers.

II. LITERATURE REVIEW :

Thesis on Predictive injury Modelling of Football Injuries:
The goal of this thesis is to investigate the potential of predictive modelling for football injuries. This work was conducted in close collaboration with Tottenham Hotspur and the PGA European Tour. In this review, three main investigations were conducted

Predicting the recovery time of football injuries using UEFAinjuryrecordings:
For this investigation, three datasets of UEFA injury recordings were analyzed using different machine learning algorithms to build a predictive model.

Predicting injuries in professional football using exposure records:

The relationship between exposure (in training hours and match hours) in professional football athletes and injury incidence was studied. The primary task was to predict the number of days a player can train before getting injured.

Predicting intrinsic injury incidence using in-training GPSmeasurements:

A significant percentage of football injuries can be attributed to overtraining and fatigue, which can be detected using GPS. This research aims to predict when an injury is most likely to take place for different players of the THFC team using GPS data gathered during their training sessions.

A Narrative Review of Different Statistical Approaches

Injuries are a common occurrence in team sports and can have significant financial, physical, and psychological consequences for

A Narrative Review of Different Statistical Approaches

Injuries are a common occurrence in team sports and can have significant financial, physical, and psychological consequences for athletes and their sporting organizations. There are several methods available to identify injury risk factors, but choosing the right method is crucial, as incorrect statistical approaches can lead to inaccurate inferences and poor decision-making

This narrative review aims to:

Outline commonly implemented methods for determining injury risk.

Encourage researchers to carefully consider the different types of variables examined in relation to injury risk and how the analyses of these variables are interpreted.

Describe advances in statistical modeling and the current evidence relating to predicting injuries in sports.

Dataset Details:

The dataset comprises two primary components: the Injury Record File and the Playlist File. The Injury Record File includes detailed information about injuries sustained by players during regular-season games over two consecutive seasons. Key fields include Player Key, a unique identifier for each player; Game ID, a unique identifier for each game; Play Key, a unique identifier for each play; and Player Metrics, which encompass speed, rotational movement, temperature, and hydration level. Additionally, the file contains Injury Details that specify the nature and severity of injuries, enabling researchers to correlate injuries with specific player actions and physiological states. The Playlist File documents the context of each play within the dataset, indexed by Player Key, Game ID, and Play Key. Key fields in this file include Roster Position, which denotes the player's assigned role during the game; Stadium Type, classified as indoor or outdoor; Field Type, detailing whether the surface is natural grass, artificial turf, or hybrid; Weather Conditions, including temperature, humidity, and precipitation; Play Type, such as pass, rush, or kick; Player Position, specifying the player's location and role in the play; and Position Group, which provides a broader classification of player roles such as offense or defense.

Data Integration The Player Key, Game ID, and Play Key fields serve as relational identifiers linking the Injury Record File and Playlist File. This linkage enables a granular analysis of injury events within their gameplay and environmental context.

that enhance monitoring by leveraging frequent transaction data. These systems analyze product trajectories through the supply chain to detect anomalies and flag counterfeit behavior, ensuring a more reliable form of product authentication.

Data Processing

First we will collect data from the different players and data from smartwatches which contains the different data such as the speed of the player 1 and player 2 , rotation speed ,movement, direction and other data such as the temperature and other things

The dataset comprises two primary components: the Injury Record File and the Playlist File. The Injury Record File includes detailed information about injuries sustained by players during regular-season games over two consecutive seasons. Key fields include Player Key, a unique identifier for each player; Game ID, a unique identifier for each game; Play Key, a unique identifier for each play; and Player Metrics, which encompass speed, rotational movement, temperature, and hydration level. Additionally, the file contains Injury Details that specify the nature and severity of injuries, enabling researchers to correlate injuries with specific player actions and physiological states. The Playlist File documents the context of each play within the dataset, indexed by Player Key, Game ID, and Play Key. Key fields in this file include Roster Position, which denotes the player's assigned role during the game; Stadium Type, classified as indoor or outdoor; Field Type, detailing whether the surface is natural grass, artificial turf, or hybrid; Weather Conditions, including temperature, humidity, and precipitation; Play Type, such as pass, rush, or kick; Player Position, specifying the player's location and role in the play; and Position Group, which provides a broader classification of player roles such as offense or defense.

Smartwatches and GPS devices are used to track the real-time location and activities of players, providing metrics such as speed, acceleration, distance covered, and movement intensity during training sessions and games. Additionally, physiological data, including heart rate variability, hydration levels, and fatigue metrics, are recorded to evaluate player performance and potential stressors. Information from injury logs maintained by sports organizations and medical teams is also included, detailing the type of injury, affected body part, treatment methods, surgery details, recovery times, and prescribed rest days. Furthermore, the dataset includes player demographic and health data collected through medical screenings and player profiles. These details encompass age, weight, height, BMI, existing injuries, medical conditions, fitness levels, sleep patterns, and stress levels. By integrating these diverse data points, researchers can assess whether a player is injured and identify the contributing factors.

Problem Statements

Task 1: This project aims to develop a machine learning model that analyzes relevant factors such as player performance metrics, physiological data, training load, and historical injury records to predict the probability of injury.
Task 2 : The goal is to create a reliable system that aids in early intervention and decision-making to enhance player safety and performance by graphs.
Task 3 : what is level of injury and whether a players both are harmed or only player injured and tells or its is minor injuris.

The dataset incorporates data from multiple sources to ensure a holistic understanding of player injuries. Smartwatches and GPS devices are used to track the real-time location and activities of players, providing metrics such as speed, acceleration, distance covered, and movement intensity during training sessions and games. Additionally, physiological data, including heart rate variability, hydration levels, and fatigue metrics, are recorded to evaluate player performance and potential stressors. Information from injury logs maintained by sports organizations and medical teams is also included, detailing the type of injury, affected body part, treatment methods, surgery details, recovery times, and prescribed rest days.

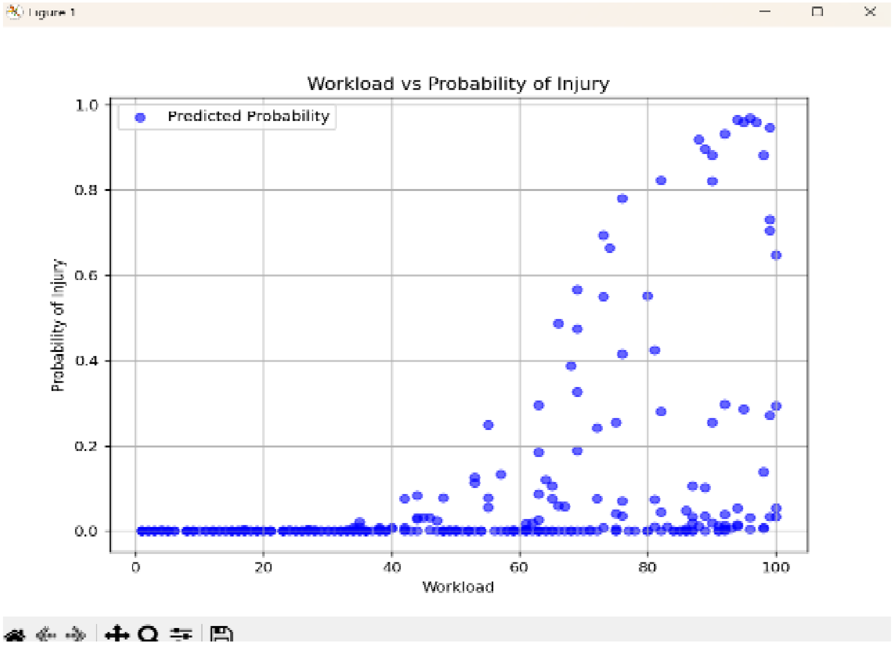
Furthermore, the dataset includes player demographic and health data collected through medical screenings and player profiles. These details encompass age, weight, height, BMI, existing injuries, medical conditions, fitness levels, sleep patterns, and stress levels. By integrating these diverse data points, researchers can assess whether a player is injured and identify the contributing factors.

III.Methodology :

This process involves comparing the data collected during injury states with non-injury states to identify key indicators of potential injuries. By analyzing the correlation map, we can pinpoint features most strongly associated with injuries, such as irregular joint angles, rapid changes in rotation speed, or abnormal pulse rate fluctuations. These insights allow for the creation of predictive models that classify a player's condition in real-time.
Machine learning algorithms are often employed to enhance accuracy, utilizing the labeled dataset to train models that can detect subtle patterns indicative of injury. Over time, these models improve as they are exposed to more data, making them more reliable for early detection.
This approach not only helps identify injuries but also supports preventive measures by flagging high-risk situations. For example, deviations in a player's movement or physiological data could trigger alerts, enabling immediate intervention to avoid further harm. Ultimately,

--Final features selected for injury prediction are :

- Speed of both players
- Rotation speed
- Temperature
- Pulse rate



The methodology for injury prediction using machine learning involves the systematic collection, preprocessing, analysis, and modeling of data gathered from wearable devices like smartwatches. The core of this research revolves around leveraging machine learning algorithms to process key metrics such as speed, collision impact, body temperature, and rotation speed. These parameters are critical indicators of physical stress and injury risk. Data from two

players' smartwatches were In addition The collected data underwent preprocessing to handle missing values, normalize scales, and identify outliers that could skew the predictions. Feature engineering was conducted to extract meaningful insights, such as sudden spikes in collision impact or temperature anomalies. Machine learning models, including classification algorithms like Random Forest and Gradient Boosting, were employed to predict the likelihood of injuries (binary classification), estimate recovery periods (regression), and identify body parts most prone to injury (multi-label classification). The models were trained and validated using appropriate datasets, ensuring robustness and accuracy.
Finally, the models were evaluated based on their performance metrics, such as accuracy, precision, recall, and F1 score. Real-time smartwatch data enabled dynamic analysis and prediction, allowing for immediate interventions in high-risk scenarios. By integrating these predictions into decision-making frameworks, the methodology demonstrates a practical, scalable approach to injury prevention. This systematic process highlights the potential of combining wearable technology, data science, and machine learning to revolutionize sports injury management and player safety strategie.

Result

— **Logistic Regression**: In our dataset after binary conversion of non-numeric attributes, the number of attributes became high, and the data becomes highly sparse and multi-dimensional as a result of which logistic regression is failing to converge. Also, logistic regression fails to optimally fit the data as the data is not linearly separable.

— **Decision Tree**: Accuracy and precision found using the decision tree classifier was better as compared to LR and NB as it was a decision-based problem, and the Decision Trees provided a clear indication of which fields were important for classification.

— **XGBoost**: The best accuracy and precision for the dataset was found using the XGBoost, which is an ensemble learning technique. It is because ensemble-based learning techniques produce better results than a decision tree.

—LinearRegression:
We used Linear Regression as a baseline for our regression problem. The mean squared error was highest for this model. Since the data is highly dimensional, a linear regressor does not produce satisfactory results.

-- Support Vector Regressor (SVR) Overview
Essentially, it creates a single optimal hyperplane based on the support vectors, thus producing a better result compared to Linear Regression. However, the problem is better addressed by ensemble learning classifiers.

—RandomForestRegressor:
This ensemble learning algorithm helped us in achieving the best results that gave minimum error.

IV.Conclusion :

Injury prediction using machine learning offers significant advancements in the realm of sports science, providing a data-driven approach to enhance athlete safety and performance. By analyzing key metrics collected from wearable devices such as smartwatches, including speed, collision impact, temperature, and rotation speed, this research showcases how machine learning can be effectively employed to predict potential injuries. The use of these real-time variables allows for the identification of patterns and trends that may indicate an increased risk of injury, enabling timely interventions that can prevent serious harm.

This study highlights the value of predictive models in sports, as they not only predict the likelihood of injuries but also contribute to estimating recovery periods and identifying body parts most vulnerable to strain or damage. The findings suggest that combining data from wearables with machine learning algorithms can provide valuable insights into an athlete's physical condition and injury susceptibility.

Furthermore, the approach can be expanded to create personalized injury prevention programs tailored to an athlete's unique physiological and performance data, ultimately leading to more effective and individualized care. As wearable technology continues to evolve, the potential for real-time monitoring and predictive analytics in sports will further improve, offering new opportunities for enhancing both player safety and overall team performance.

In conclusion, the application of machine learning to injury prediction represents a paradigm shift in how we approach injury prevention in sports. It not only empowers coaches, medical professionals, and players to make informed decisions but also contributes to the broader goal of optimizing athlete health and career longevity.

Individual Contribution:

1. **Data Collection and Organizing: Shravan**
2. **Literature Review: Vinay**
3. **Exploratory Analysis: Srujan, Shravan ,Vinay, Vivek**
4. **Creating Learning Models:**

Injury Prediction (graphs Classification Task): Vinay , Srujan

Estimating level of injury (Regression Task): Shravan ,Vivek

Predicting the Injury type(Multi-Label Classification Task): Vinay ,Srujan ,Vivek

5. Analyzing Accuracy Among All codes and Reasoning for the Best Output:

Data Classification Problem : Shravan

Graphs Problem: Srujan

Writing code for prediction : Vinay ,Vivek

V.REFERENCES :

1. Kampakis, S. (2016). *Predictive modelling of football injuries*. Doctoral thesis, UCL.
2. Joshua D. Ruddy, Stuart J. Cormack. *Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches*. [NCBI]
3. *A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players*. [ResearchGate]
4. Akobeng, A. K. (2007a). *Understanding diagnostic tests 1: sensitivity, specificity and predictive values*. Acta Paediatr, 96, 338–341. https://doi.org/10.1111/j.1651-2227.2006.00180.x [PubMed]
5. Altman, N., Krzywinski, M. (2015). *Points of significance: association, correlation and causation*. Nat. Methods, 12, 899–900. https://doi.org/10.1038/nmeth.3587 [PubMed]
6. Tianqi Chen, Carlos Guestrin (2002). *XGBoost: A Scalable Tree Boosting System*.
7. R. Bekkerman. *The present and the future of the KDD Cup competition: An outsider's perspective*.