Sports Injury Prediction System using Random Forest Classifier

Akshay Shringarpure, *Ronak* Shetty, *Ajinkya* Surve, and *Amarsinh* Vidhate[1]
[1]Dept. of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai Maharashtra, India

Abstract. One of the largest growing industries in the modern-day world is the sporting industry. Currently valued at around 500 billion USD, with a growth scope of exponential potential, its ability to attract investors is incredible. And just like any other investment. It is part andparcel of the investor's fiscal responsibility to take good care of their assets. The biggest assets in the sporting industry are of course the players, and the greatest threat to said assets is injuries. We take into consideration said factors and deem it important to solve said issues, and understanding the money involved, the industry sides with us too. We seek to solve the said problemby taking into account all previous injury records and datasets of various players and predicting the kind, number, and severity of the injuries in the future. We seek to create a methodology for such prediction, which applies to all and any sports, being one of the only such models of its kind.

1. Introduction

Sports Injuries are common in competitive sports as players always have to push themselves to peak conditions and often end up causing harm to themselves or opponent players. Today, a player earns money for every minute spent on the field and a retiring injury can neither be afforded by the players nor the team management. According to a census, about 11 million injuries are sustained each year, most cases being damage to limbs and head injuries. Head injuries are the major cause of death in competitive sports. Accordingto a survey of major sports leagues in the U.S., the financial loss due to sports injuries was around $29 M in the year 2017, which includes loss of sponsorships due to failed performanceof teams and medical expenses. Our project isof great benefit for asset protection and management. Simply put, our project can save millions of dollars.

1.1. Objective

Our project can help sports organizations in evaluating players based on their history of injuries. The various clubs and team management can decide whether to sign a player or not based on the predicted data. Another medical aspect of this software is warning the players about their predicted injuriesso that they can be cautious during the game and protect themselves from serious harm or permanent damage. The team players can also strategize their game plans based on these predictions such that every player in the game can function at their highest form.

1.2. Motivation

All members of our team are big fans of sporting events of all kinds. One of the worst pains that you can feel as a fan, is when the club you support buys a player who develops an injury record and never flourishes into the greatness he or she was pegged for. Besides the monetaryloss of the club, the players themselves have to suffer terrible mental and physical torment. We seek to eliminate such trauma, making the sport even more exciting.

1.3. Organization of Paper

The remaining part of the report is organized in the following manner. The prerequisite for the study is covered in Chapter 2. It contains thesurvey of existing systems, limitations of existingsystems which constitutes the literature survey, problem statement, and scope. Chapter 3contains the project proposal which includes proposed work, methodology, and hardware andsoftware requirements.

The planning and formulation part of the project is covered in chapter 4. Chapter 5 contains the Design of the System. Chapter 6 includes the necessary implementation details, project outcomes, and results. Finally, Chapter 7 contains the Conclusion and Future Work.

2. Literature Survey

It has been observed that sports science and its techniques are benefitting sports persons a lot. A literature survey is given hereunder.

2.1. Survey of Existing System

- Liu, Guangying at. El [1] has proposed a learning-based model to predict sports injuries based on the available data from various sources. The authors have reduced attributes that have a significant impact on injury risk and have provided an algorithm based on the Random Forest method to test with real-world data. The better achievement is low error rates.

- Chen Huang, Lei Jiang [2] .Data monitoring and sports injury prediction model based on embedded system and machine learning algorithm[6]. The authors here propose

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

the use of an Artificial Neural Network (ANN), whose purpose is to develop and use early-doing ability and exercise load data to validate a hierarchical machine learning prediction system with accurate detection of player injuries.

- Bittencourt NFN, et.al [3]. Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept 2016. This paper is based on the demonstration of the implementation of a relatively complex system. This model seeks to identify the regularities and patterns of relationships amongst determinants and characterizes the formation of a pattern that arises from a complex web of afore-mentioned determinants.

- Alejandro López-Valenciano,(2018) et.al has proposed a preventive Model for Muscle Injuries. It's a novel approach based on Learning Algorithms [4]. Here, their motive was to analyze and juxtapose the functioning of multiple MachineLearning (ML) methods to pick the best performing injury risk factor model. The model focused on lower extremity muscle injuries of male football and handball athletes.

- David L. Carey, et. al (2017). had mentioned predictive modeling of training loads and injury in Australian football[5]. The data that the authors worked on, was collected from an elite Australian football club, over 3 seasons. The data of the first two seasons were used to build their injury prediction model, and the predictions were generated for the third season.

These are the most notable works in this venture that comprises, and sum up our understanding so far. Our collective research has led us to a basic, and in turn, an in-depth understanding of the problem statement, further accentuating our comprehension of the need, and lack of a solution to the problem [6,7,8].

Multiple models exist, respectively efficient in their sports, but the efficiency of these models could still be further improved, and the need for a more comprehensive model still exists, keeping in focus the algorithm used and its efficiency.

2.2. Limitations of Existing System

Existing systems are currently minimal in the application, and even when in use, have major issues. Some are very complex to implement for wide usage whereas some have less accuracy, or tend to be time-consuming. To overcome these limitations, first, we had to understand these issues and their significance. We have compared all the different methodologies implemented by fellow researchers based on their limitations as presented below.

2.3. Problem Statement

Competitive sports are a multibillion-dollar industry, and the most prized assets of said industry are the players that participate. Sportspersons undergo a lot of training and physical exertion, which causes extensive wear and tear to the body, and in accordance,

Author	Method Used	Limitations
Bittencourt NFN, Meeuwisse WH, Mendonça LD, Nettel-Aguirre A, Ocarino JM, Fonseca ST (2016) Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concept.	Complex Web of Determinants	Complex model causes increased implementation cost. Model based on too many uncertain factors
David L. Carey, Kok-Leong Ong, Rod Whiteley, Kay M. Crossley, Justin Crow, Meg E. Morris (2017) Predictive modelling of training loads and injury in Australian football	Multiple classification algorithms	SVM and logistic regression giving low accuracy for certain datasets
Alejandro López-Valenciano, Francisco Ayala, José Miguel Puerta, Mark De Ste Croix, Francisco Vera-García, Sergio Hernández-Sánchez, Iñaki Ruiz-Pérez, Gregory Myer (2018) A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms	Alternating Decision Tree	Multiple ADTrees required to achieve good accuracy and avoid underfitting.
Chen Huang, Lei Jiang (2020). Data monitoring and sports injury prediction model based on embedded system and machine learning algorithm	Artificial Neural Network (ANN)	Neural networks depend a lot on training data. This leads to the problem of over-fitting and generalization.

Fig. 1. Methodology Comparison

of injuries. Suppose, a million-dollar player gets injured during a game and cannot continue, It can costa fortune to his/her team owner, as well as affectsaid player's career drastically in case of prolonged injury. We seek to predict the injury patterns of various players after taking into account past data, and hence, predicting theseverity and duration of the injuries that a player might incur in the future.

2.4. Scope

We believe that our project is of great significance to a multi-billion-dollar industry. We primarily focus on major asset care, and as a result, our work is indispensable. What is also incredible is that our project works for multiple sports, making its appeal more universal.

3. Proposed Solution

3.1. Proposed Work

To develop a system that accurately predicts player injuries with the help of previous data and specialized algorithms, so that players and clubs can take necessary precautions.

3.2. Proposed Methodology

Injury prediction, in its nature, is a pattern of the non-linear format. What we understand from current cumulative

research, is that injury classification is to partition various patterns into groups and that, there are (or were) many different models which can be categorized into traditional time series methods and Intelligent methods and semi-recently, artificial neural network models. The drawback of such traditional systems (like linear regression, time series model), is the very first fact that we learn i.e.; Injury prediction, in its nature, is a pattern of the non-linear format. Traditional systems only consider time sequence features, and it is difficult for them to deal with the nonlinear nature of patterns. ANN, on the other hand, gives better performance with nonlinearity issues, but in most (if not all) of the NN models, the error on the training data set is good, butthey perform badly when out-of-sample data is presented to the network. As a result, after a lot of research, we have concluded that the random forest algorithm is the best as, even in comparison to Support Vector Machine (SVM) and decision trees, it gives us the most accurateresults, with comparative time delay.

4. Planning and Formulation

4.1. Schedule and flow:

The idea was conceived and worked on by the start of June-July 2021. First ensured the commencement of the research phase, which lasted for approximately a month, for the duration of which, the problem statement and current solutions were explored in explicit detail. Then began our work on the formulation and implementation of our ideology and solution. And in accordance, we could reach a model of our solution, which works as expected.

4.2. Detail Plan of Execution

The first phase of the proposal spanned 4 months. To effectively design and develop a cost-effective model, the Waterfall model was practiced.

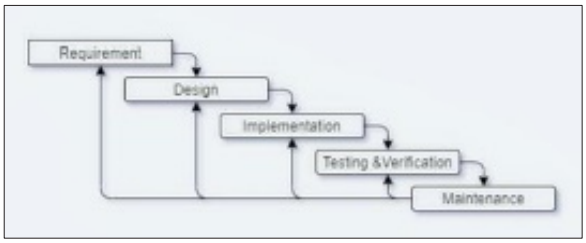

Fig. 2. Workflow diagram

5. Design of System

5.1. Flowchart:

The diagram that we have used to better our understanding and ease our ideological implementation is a flowchart. We have included the standard procedure of employing a standard machine learning algorithm which involves data preprocessing, data splitting, training and ultimately testing the model with the testing data. It summarizes the design and working of our system eloquently, as is evident by the diagram below.

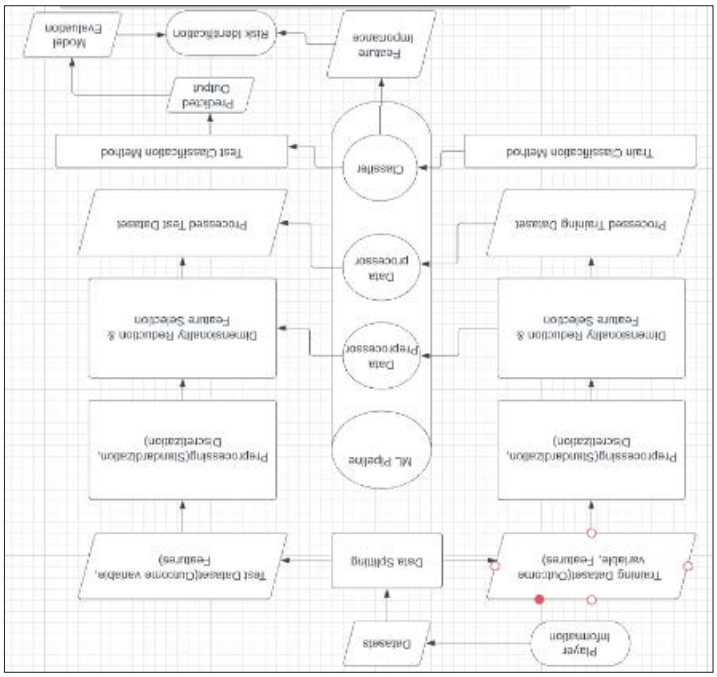

Fig.3. Flowchart of the proposed system

6. Expected Results

6.1. Implementation Details

For implementation, we have used a Jupiter notebook. In accordance, with datamanipulation and visualization, we have usedvarious python libraries like pandas, matplotlib, and seaborn. For input, we use CSV files, namely - NHL_Injury_Database.csv, NHL_Bio.csv, and NHL_Time_on_ice.csv. We then perform data cleaning and manipulation on these CSV files, combining all the data into one file, named- 'NHL_alldata.csv'. With the help of python libraries, we can then perform data visualization, like:

- Checking % of injured and non-injured players
- Checking the injury type of the players
- Checking injury count of different types
- of injury
- Checking the injury of the player
- Checking injury chances along with age

Fig 3. Loading the data from NHL_Injury_Database.csv, NHL_Bio.csv, NHL_Time_on_ice.csv

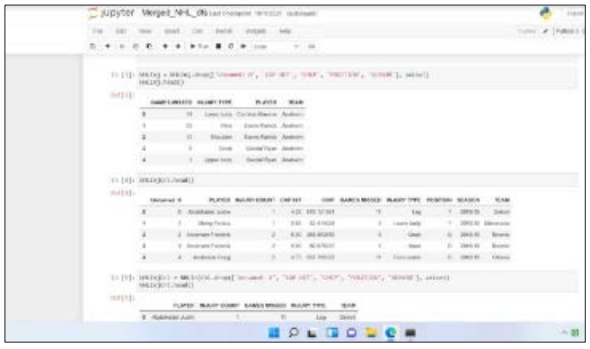

Fig. 4. Cleaning player information data

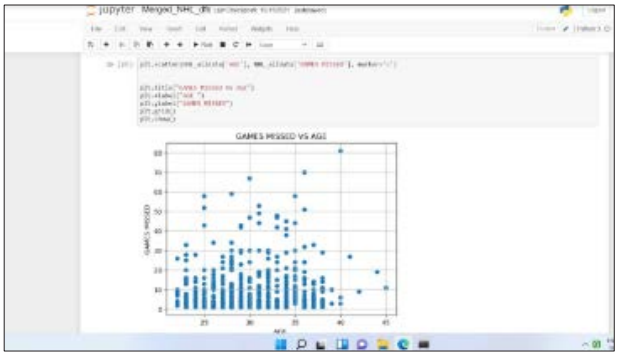

Fig. 5. Scatterplot of Games missed by age

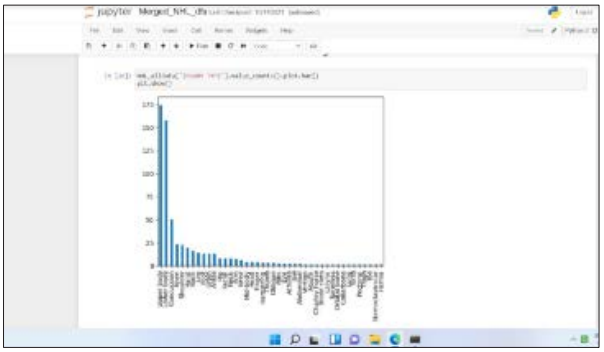

Fig. 6. No of Injuries Vs body parts

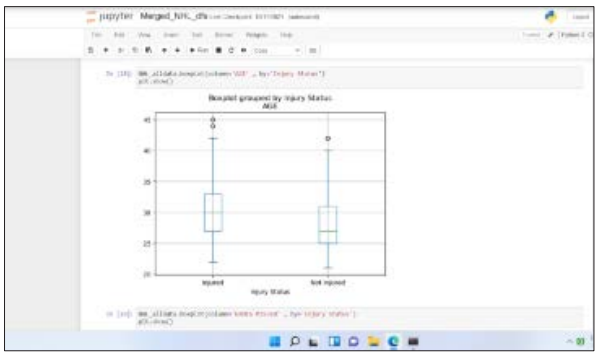

Fig. 7. No of the times players are injured (1, 2, 3)

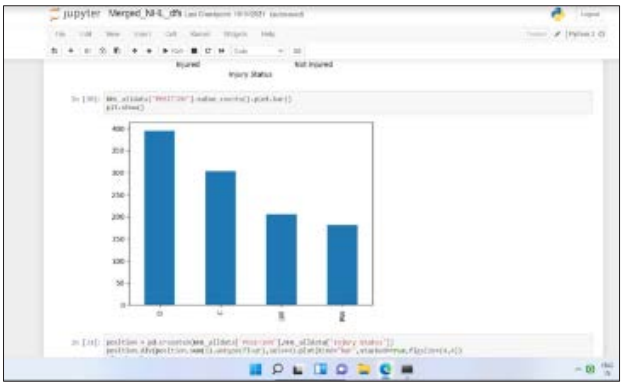

Fig. 8. Boxplot of Age Vs Games missed

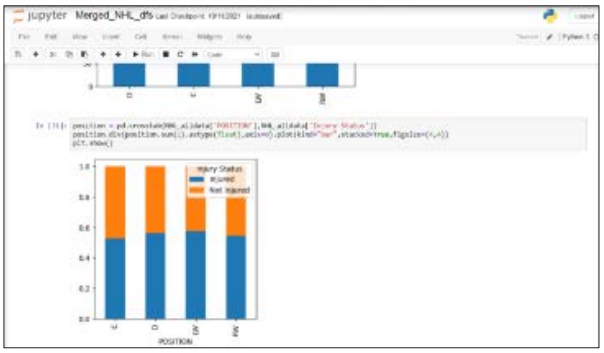

Fig. 9. Injury Status by Position of play

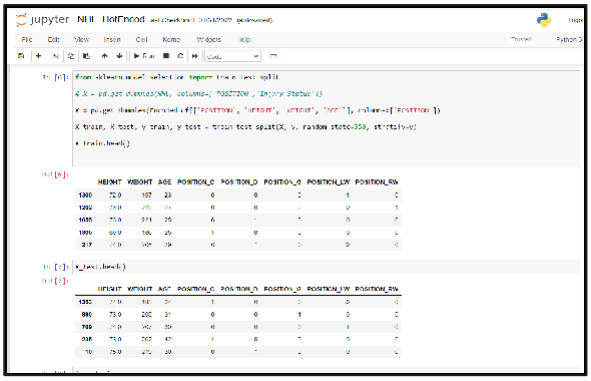

Fig. 10. Splitting of Data

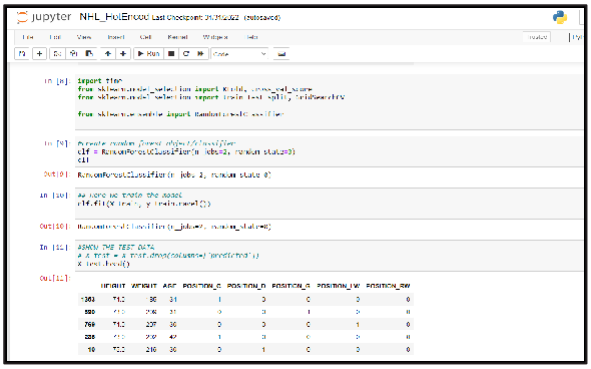

Fig. 11. Training and Testing data

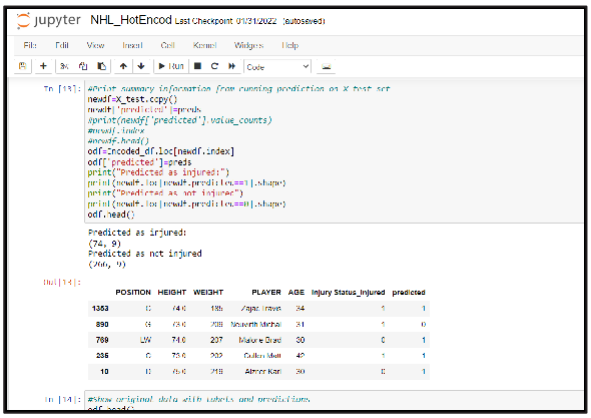

Fig. 12. Injury Prediction Results

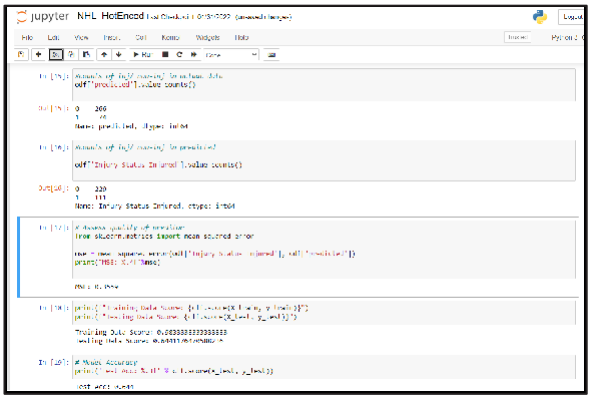

Fig. 13. Accuracy Score

6.2. Project Outcomes

As can be inferred from the data and graphs, the correlation between age and injuries is evidentialin the sense that, as age increases, so does the frequency of injuries. Along with this, it can also be inferred that players who play position d,get injured very often, in comparison to other positions. What is also very noticeable, is the correlation between weight and injuries, further adding another aspect to the prevention of injuries.

The amalgamation of various such considered criteria could potentially affect how a club looks after its players and could also determine optimum utilization of their players. As the player ages, or gains more weight, the club could plan to play him in another position, which may better suit him, and help the club get the most out of him through consistent use.

Our model is designed to potentially shape the transfer policies of clubs, in the sensethat clubs can determine what players are worth the money, and could perform consistently in the long-term and whether they should bet on older talents or not.

Our model could also help players elongate their careers, as they too will grasp a better understanding of their ideal positions, weight, and overall understanding of their injuries. They could plan their careers by our predictions, maximizing their efficiency and prowess in their sport. This model provides insightful readings regarding a player's fitness, bankability, and vulnerabilities.

7. Conclusion and Future Work

An ideology was put in motion with the inauguralwork of our project. The ability to predict injuries,and assess and determine patterns in said prediction is a novel concept.

In addition, our approach is further newer, and hence has a lot of potential for further research. Our current approach gives the most accurate results available, and can truly change the modern sports-entertainment industry.

Our project could change the multi-million-dollar sporting industry as we know it. But we are very well aware that our idea and its execution are in a state of infancy, and a lot ofwork has to be done for the full potential of the idea to be realized. A very raw concept of it has already been displayed to you last semester. Weseek to further strengthen our idea with the help of algorithmic hybridization. Our base working model uses the Random-Forest algorithm, which wedeemed to best fit us as it provided the best accuracy and was comparatively quick. But given the severity of our project, we decided to work on the suggestions put forth by our mentors, and enact hybridization. This will severely improve the accuracy and speed of our application, and in turn, improve efficiency.

Throughout our research, which is still ongoing, we have come across multiple algorithms, out of which we have shortlisted some of them such as linear regression, logistic regression, naive Bayes, and linear discriminant analysis. We might also add a more user-friendlyGUI to increase the reach and ease the use of our project.

References

1. Wang Jian Bai, Hwa Sun, Guangying Liu, Hongmei Li "A learning-based system for predicting sports injuries," MATEC Web of Conferences **189**, 10008 (2018)

2. Huang, Chen and Jiang, Lei, “Data monitoring and sports injuryprediction model based on embedded system and machine learning algorithm, " MAM, **81,** 103654 (2020)

3. Bittencourt, N F N and Meeuwisse, W H and Mendon, “Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition,” British Journal of Sports Medicine, **50**, 21, 1309-1314, (2016).

4. López-Valenciano A, Ayala F at. el , “A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms,” Med Sci Sports Exerc., **50**(5):915-927, (2018)

5. David L. Carey, Kok-Leong Ong, Rod Whiteley, Kay M. Crossley, Justin Crow, Meg E. Morris "Predictive modeling of training loads and injury in Australian football", IJCSS,**17**, 49-66,(2018).

6. Song, Hesheng and Han, Xiu-Ying and Montenegro-Marin, Carlos and Krishnamoorthy, Sujatha, "Secure prediction and assessment of sports injuries using deep learning-based convolutional neural network," JAIHC,**12,**1-12,(2021)

7. Brooks JH, Fuller CW, Kemp SP, Reddin DB. Incidence, risk, and prevention of hamstring muscle injuries in professional rugby union. Am J Sports Med.,**34(8)**:1297-306(2006)

8. Gregory Ornon, Jean-Luc Ziltener, Daniel Fritschy, Jacques Menetrey, "Epidemiology of injuries in professional ice hockey: a prospective study over seven years" *J EXP ORTOP* **7,** 87 (2020)