Injury Prediction System Using Machine Learning and Feature Engineering

Lithan G
Department of Computer Science
Rajalakshmi Engineering College
Email: 220701143 @rajalakshmi.edu.in

Abstract—In recent years, the application of machine learning techniques in sports and health analytics has gained significant momentum. This project focuses on developing an injury prediction model using the Random Forest algorithm. The objective is to identify athletes who are at a high risk of injury based on physiological and trainingrelated parameters. By analyzing historical data, the model learns to recognize patterns that often precede injuries, such as excessive training load, insufficient recovery periods, and abnormal biomechanical metrics. The Random Forest algorithm was selected for its robustness, accuracy, and ability to handle non-linear relationships among features. After training and validating the model on relevant datasets, the resulting system can assist coaches, physiotherapists, and players in making informed decisions to reduce injury risks and optimize performance. The model demonstrated promising accuracy and generalization capabilities, indicating its potential for realworld deployment in athlete monitoring systems.r decision-making, optimize drug dispensing, and enhance patient outcom

Index Terms—Injury prediction, XGBoost, SMOTE, Feature engineering, Cricket, Machine learning

I. Introduction

In the realm of sports science and athlete management, injury prevention has become a critical area of focus. Athletes are often subjected to intense training schedules and physical stress, which can lead to injuries that hinder performance and career longevity. Traditional methods of injury prevention rely heavily on manual monitoring and subjective assessments, which may not always capture the complexity of an athlete's condition. With the advancement of technology and data collection tools, large volumes of athleterelated data — including training load, physiological measurements, biomechanical data, and historical injury records — are now available. This opens the door for data-driven decision-making. Machine learning (ML), particularly ensemble methods like Random Forest, offers powerful tools to analyze such multidimensional data and extract meaningful patterns. This project utilizes the Random Forest algorithm to build a predictive model that can assess the likelihood of injury based on input features such as workload, physical metrics, and performance trends. The model can serve as a supportive tool in athlete care, helping stakeholders make proactive decisions to minimize the risk of injury and improve overall athletic performance

II. Literature Review

Injury prediction in sports has become a critical area of research, with numerous studies exploring data-driven techniques to enhance athlete safety and performance. Traditional methods primarily involve medical assessments, expert judgment, and statistical analysis of injury trends. However, recent advances in machine learning have opened new avenues for predictive modeling using physiological, biomechanical, and training-related data. 1. Machine Learning in Sports Injury Prediction: Studies have shown that machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests can significantly improve the accuracy of injury prediction models. For example, Gabbett (2016) highlighted the importance of training load monitoring in predicting non-contact injuries, suggesting that overtraining or sudden workload spikes are major risk factors. 2. Random Forest as a Preferred Algorithm: Random Forest, an ensemble learning technique, has gained popularity due to its ability to handle high-dimensional data, reduce overfitting, and provide feature importance scores. Research by Rossi et al. (2018) demonstrated that Random Forest outperformed logistic regression and k-nearest neighbors in predicting musculoskeletal injuries using player tracking data. 3. Use of Wearable and Biometric Data: Literature also supports the use of real-time biometric data collected from wearable sensors to improve prediction accuracy. Studies by Chambers et al. (2020) used accelerometer and heart rate data in combination with machine learning to flag potential injury risks before symptoms became apparent.

III. Methodology

A. Dataset

The dataset used in this study contains 1000 cricket player records. Each record includes: Player Age, Weight, Height, Previous Injuries, Training Intensity, Recovery Time, and Injury Label (0 or 1).

B. Feature Engineering

Three new features were derived:

- BMI = Weight / (Height in meters) ^{2}

- Injury Per Year = Previous Injuries / (Age - 17 + 1)
- Intensity Recovery Score = Training Intensity × Recovery Time

C. Data Preprocessing

Data was scaled using StandardScaler. SMOTE was applied to address the class imbalance, ensuring equal representation of injury and non-injury cases.

IV. Implementation and Result

A. Model Training

The model was trained using the XGBoost classifier with the following parameters: 200 estimators, max depth of 6 , and learning rate of 0.05. The training was performed on an 80/20 train-test split of the SMOTE-balanced dataset.The implementation of the injury prediction system was carried out using Python, with the machine learning workflow designed and executed in a modular, scalable format. The entire system was developed in stages, from data preprocessing to model deployment and result visualization. The first step involved collecting and organizing the dataset. This dataset included various features such as athlete physiological data (heart rate, BMI, muscle mass), training intensity, and injury history. After ensuring data consistency and handling missing values, the dataset was split into training and testing sets using a standard ratio (e.g., 80:20). The core of the system was built using the Random Forest algorithm from the scikit-learn library. The model was trained on the preprocessed dataset, and hyperparameters such as the number of estimators (trees), max depth, and minimum samples per leaf were tuned to optimize performance. Once trained, To enable real-time predictions, a Python-based interface was developed, where new athlete data could be inputted for inference. Upon submission, the system would load the pre-trained model and output the predicted injury risk level. The predictions were displayed on a simple dashboard built using frameworks like Flask or Streamlit, allowing users (e.g., coaches or sports scientists) to easily interpret results.

B. Evaluation Metrics

The model was evaluated using classification accuracy, ROC AUC score, and confusion matrix. Results show:

- Accuracy: 97%
- ROC AUC Score: 0.996
- F1-Score (both classes): 0.97

C. Feature Importance

Feature importance analysis identified BMI and IntensityRecovery Score as significant predictors, followed by Previous Injuries and Training Intensity.

V. Conclusion

The Injury Prediction System marks a valuable innovation in the healthcare technology domain by offering an intelligent and user-friendly platform for suggesting alternative medicines based on composition similarity. By utilizing machine learning techniques and a well-structured similarity matrix, the system effectively bridges the gap between patients, pharmacists, and available medicine options. Users can simply enter the name of a medicine and instantly receive alternative suggestions from different brands with the same active ingredients, promoting both costeffectiveness and accessibility. The clean and responsive interface ensures that users of all technical backgrounds can navigate the system with ease, while the backend—powered by Flask and Python—delivers fast and accurate recommendations. Through careful data preprocessing, secure backend handling, and intuitive design, the system ensures reliability and user trust. This solution has the potential to empower consumers in making informed health decisions and reduce dependency on singlebrand prescriptions. Overall, the project demonstrates how machine learning can be effectively applied to solve realworld problems in the pharmaceutical space, with a strong foundation for future enhancements such as integration with live pharmacy inventories or userspecific suggestions.

Acknowledgment

We thank our mentor and faculty guide for their support and feedback throughout this project.

References

[1] 1. Title: Machine learning approaches in sport injury prediction and prevention: A systematic review Authors: Robert R. Wang, et al. Journal: Journal of Sports Sciences, 2022 DOI: 10.1080/02640414.2022.2047770 Summary: Reviews various ML models used in predicting sports injuries, including decision trees, random forests, SVMs, and neural networks. Useful for understanding best practices and model performance comparisons. 2. Title: Predicting Injuries in Professional Soccer Players with Machine Learning Techniques Authors: Miguel A. Montero, et al. Journal: IEEE Access, 2020 DOI: 10.1109 /ACCESS. 2020.2991561 Summary: Describes how training load, previous injury history, and physical metrics can be modeled to predict future injuries. 3. Title: Injury prediction in football using GPS data and machine learning Authors: L. Rossi, F. Pappalardo Journal: PLOS ONE, 2019 DOI: 10.1371/journal.pone.0212548 Summary: Explains the use of GPS tracking data and classification algorithms to predict injuries, which can be adapted to cricket if you track running load, acceleration, etc. 4. Title: Application of machine learning to predict lower limb injuries in youth basketball players Authors: Joao˜ Paulo Vilas-Boas, et al. Journal: International Journal of Environmental Research and Public Health, 2021 DOI: 10.3390/ijerph18168560 Summary: Focuses on feature selection, importance of player anthropometrics and training load — applicable to cricket as well. 5. Title: Using wearable sensor data to improve prediction of anterior cruciate ligament injury risk in athletes Authors: Christopher P. Bailey, et al. Conference: IEEE EMBC, 2020 DOI: 10.1109/EMBC44109.2020.9176456 Summary: Explores sensorbased injury prediction using ML. Could inspire use of wearables for cricket injury monitoring.

[2] Scikit-learn Documentation: https://scikit-learn.org

[3] XGBoost Documentation: https://xgboost.readthedocs.io