ATHLETE INJURY PREDICTION USING MACHINE LEARNING

A Madhavi [1], K Rutvik [2] , S Yashwanth Sai[3] , T Abhinaya [4] ,V Manoj [5]

[1]Assistant Professor, Department of CSE, VNR- Vignana Jyothi Institute of Engineering andTechnology, Hyderabad, Telangana - 500090

[2]UG – Computer Science and Engineering, VNR- Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana – 500090

[3]UG – Computer Science and Engineering, VNR- Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana – 500090

[4]UG – Computer Science and Engineering, VNR- Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana – 500090

ABSTRACT

Sports injuries are prevalent and can have serious financial, psychological, and bodily repercussions. Although injuries are hard to predict, cutting-edge technologies and data-science applications might offer crucial information. Methods of Machine Learning (ML) may be applied to enhance injury prediction. This project is aimed to build a machine-learning-based model by comparing different algorithms that predicts injuries well in advance. Predicting injury beforehand would be a big assistance to the players, ultimately transforming the sports industry. Even in future tournaments, teams would be able to better plan their strategies if they were aware of the resting interval in advance.

Keywords— Training load, KNN, Logistic Regression, Random Forest, SMOTE, XGBoost Algorithm

1. Introduction

One of the most crucial elements for achieving peak performance in sports is being healthy and injury-free. In order to better understand the training loads of athletes and the frequency of injuries, researchers and practitioners from a variety of sports have been gathering data for decades. Technology and machine learning applications have grown in recent years, making it possible to predict future performance and injuries and so improve data-driven sports advice. The UEFA (Union of European Football Associations) model defines any damage to the body tissues that prevents a player from participating in sports for at least one day following the original day of occurrence is referred to as a non-contact injury. In order to foresee non-contact injuries caused by strenuous activity, we developed a hierarchical framework in our work. An exercise's burden might be internal (like heart rate) or external (like workout duration and number of leaps). For us-age in training sessions and sports, a variety of wearable gadgets with GPS are available.

Fig. 1: Types of Injuries in athletes

We estimated each athlete's burden using information gathered from questionnaire responses. In our research, machine learning will be used to anticipate injuries in athletes based on thorough training logs. A 7-year data collection of 74 top middle and long-distance runners, with a window of 3 weeks prior to the in-jury was taken into consideration allowed for the evaluation of injury prediction. By contrasting various machine learning models, a predictive system for injuries is to be built. To further classify the data, a model with high accuracy is taken into consideration. Both the load and the danger of injury can be calculated.
We give a summary of the related research on injury prediction in the next section.

2. Experimental Methods or Methodology

In this study, we develop a software tool that uses athletic data from training and competition to assess the risk of injury in upcoming contests using supervised learning techniques.
The methodology for athlete injury prediction using machine learning generally follows the following steps:

2.1 Data Collection:

The initial step is to collect data related to the athlete's injury history, training load, and any other relevant factors. The data can be collected from various sources such as wearable sensors, electronic health records, injury reports, and game statistics. Performing an initial analysis to comprehend the data is also crucial. This analysis should include locating any missing numbers, outliers, or contradictions. This will help us prepare the data for the next steps by cleaning and preprocessing it as necessary.

2.2 Data pre-processing:

The collected data needs to be cleaned, pre-processed, and transformed into an appropriate format for machine learning algorithms. This includes removing any missing values, scaling and normalizing the data, and splitting the data into training and testing sets.

2.3 Feature engineering:

Feature engineering includes choosing relevant features or variables from the data that can be used to predict injury risks. This includes identifying any patterns or trends in the data that could indicate injury risks. The weekly dataset takes into account the three weeks leading up to the injury or healthy event, and it summarizes the training load on a weekly basis. There are 22 features that each explaina week leading up to an event.
We used Chi-square feature extraction method and calculated the feature scores. Based on feature scores we selected top 10 features. A machine learning classifier is normally used on a balanced dataset to reduce bias towards the majority class, but as our dataset contains a considerable class imbalance with a large number of healthy events and relatively few injury events, this is not the case here. We used SMOTE (Synthetic Minority Oversampling Technique), a method that oversamples the minority class and generates synthetic data to balance the class distribution, to overcome this problem.

No	Feature	Range
1	Average exertion-week 1	[0.0, 1.0]
2	Maximum exertion-week 1	[0.0, 1.0]
3	Maximum training success-week 1	[0.0, 1.0]
4	Maximum recovery-Week 1	[0.0, 1.0]
5	Average exertion-week 2	[0.0, 1.0]

6	Maximum exertion-week 2	[0.0, 1.0]
7	Maximum training success-week 2	[0.0, 1.0]
8	Maximum recovery-Week 2	[0.0, 1.0]
9	Average exertion-week 3	[0.0, 1.0]
10	Maximum exertion-week 3	[0.0, 1.0]

Table 1. Features Used along with their ranges

2.4 Model selection:

The following step is to select an appropriate machine learning model based on the current issue.The models trained are KNN classifier, XGBoost classifier, Random forest and Logistic regression.

2.5 Model training:

The data needs to be trained on selected machine learning models using the selected features. The above mentioned models are trained on the training dataset on which synthetic minority oversampling technique is applied. This technique is used to support the minority classes to avoidthe bias in the data.

2.6 Model evaluation:

The trained models are evaluated on the testing data set to assess its performance. This involves assessing the accuracy, precision, recall of the model when predicting the risk of injury.

2.7 Deployment:

The model can be used to forecast the likelihood of injury in athletes after it has been trained and assessed. This entails incorporating the model into an existing system or creating a new system for predicting athlete injuries.

3. Results and Discussion

This project’s major objective was to provide the best model that predicts the athlete injury wellin advance. From the four machine learning models (KNN classifier, XGBoost classifier, Random forest and Logistic regression) which we trained XGBoost got the highest accuracy. The confusion matrix of these mod-els are as follows:

3.1 KNN Classifier

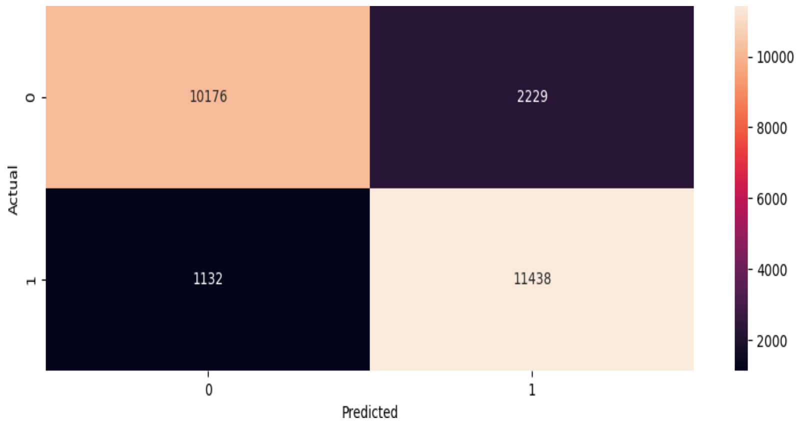

Fig. 2. Confusion matrix of the KNN Classifier

3.2 XGBoost

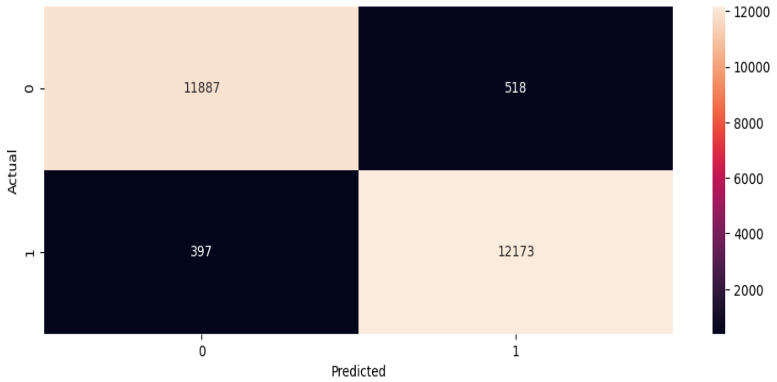

Fig. 3. Confusion matrix of the XGBoost Classifier

3.3 Logistic Regression

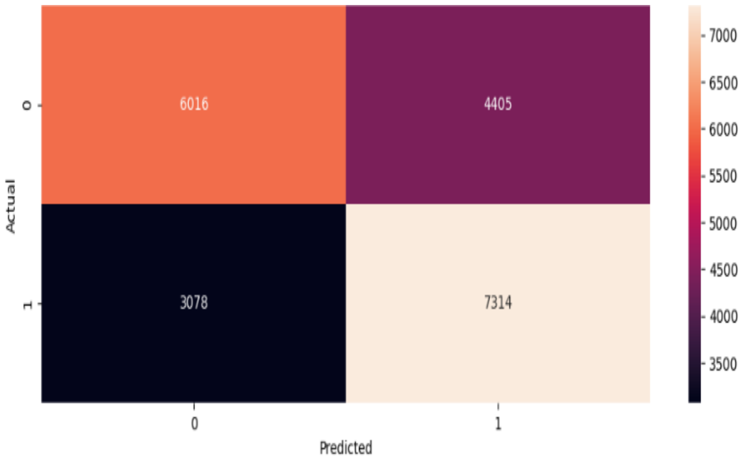

Fig. 4. Confusion matrix of the Logistic Regression

3.4 Random forest

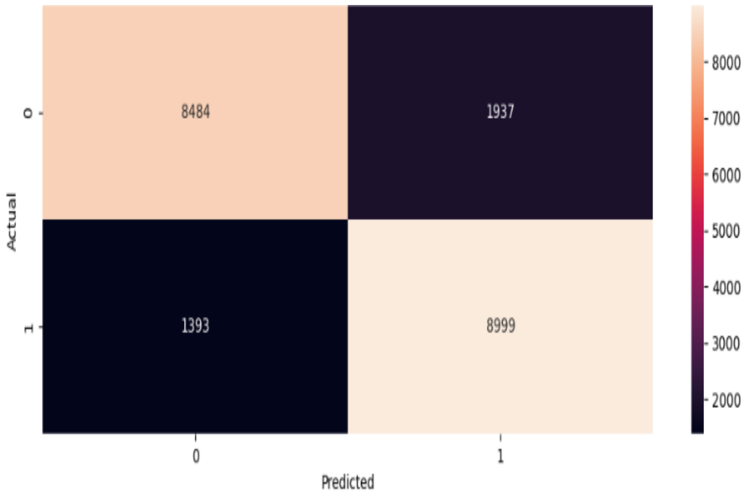

Fig. 5. Confusion matrix of the Random Forest Classifier

The accuracy of each of these models is as follows:

No	Model	Accuracy
1	KNN Classifier	0.8648248248248248
2	XGBoost Classifier	0.965005005005005
3	Logistic Regression	0.633161966078893
4	Random Forest Classifier	0.8471147840292125

Table 2. Accuracies of Machine Learning models trained

4. CONCLUSION

Our machine learning model, which is based on the XGBoost algorithm, uses training-load data from three weeks before an event to specifically train its ability to anticipate injuries. The ability to identify the harm in advance is a major bene-fit of this technology. Teams will be able to plan more effectively for upcoming competitions thanks to this. Overall, these findings show the potential benefits of applying machine learning to injury prediction and training program customization for athletes, which may offer valuable information to them.

Coaches can use the proposed system as a computerized tool to help them manage the training loads of their athletes. It therefore has the potential to introduce novel and more efficient tactics, revolutionizing data-driven guidance in the sports business. The model could be further improved by incorporating more advanced techniques such as transfer learning or ensembling multiple models. Developing different models that provide more detailed information about the injury.

References

1. A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernandez, and ` D. Medina, “Effective injuryforecasting in soccer with gps training data and machine learning,” PloS one, vol. 13, no. 7, p. e0201264, 2018.
2. N. B. Murray, T. J. Gabbett, A. D. Townshend, and P. Blanch, “Calculating acute: chronic workload ratios using exponentially weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages,” Br J Sports Med, pp. bjsports–2016, 2016.
3. G. A. Borg, “Psychophysical bases of perceived exertion,” Med sci sports exerc, vol. 14, no. 5,pp. 377–381, 1982.
4. A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernandez, and ` D. Medina, “Effective injuryforecasting in soccer with gps training data and machine learning,” PloS one, vol. 13, no. 7, p. e0201264, 2018.
5. T. J. Gabbett, “Reductions in pre-season training loads reduce training injury rates in rugby league players,” British journal of sports medicine, vol. 38, no. 6, pp. 743–749, 2004.
6. T. J. Gabbett and D. G. Jenkins, “Relationship between training load and injury in professionalrugby league players,” Journal of Science and Medicine in Sport, vol. 14, no. 3, pp. 204–209, 2011.
7. Ivan G. A. Borg, “Psychophysical bases of perceived exertion,” Med sci sports exerc, vol. 14, no. 5, pp. 377–381, 1982.
8. “Influence of training and match intensity on injuries in rugby league,” Journal of sports sciences, vol. 22, no. 5, pp. 409–417, 2004.
9. Krishna, P. R. ., & Rajarajeswari, P. . (2022). EapGAFS: Microarray Dataset for Ensemble Classification for Diseases Prediction. International Journal on Recent and Innovation Trends in Computing and Communication, 10(8), 01–15. https://doi.org/10.17762/ijritcc.v10i8.5664
10. Gabbett, “The development and application of an injury prediction model for noncontact,

soft-tissue injuries in elite collision sport athletes," The Journal of Strength & Conditioning Research, vol. 24, no. 10, pp. 2593–2603, 2010.

11. Jaspers A, Op De Beéck T, Brink MS, et al. Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? Int J Sports Physiol Perform. 2018; 13(5):625–630. doi:10.1123/ijspp.2017-0299

12. Knobbe A, Orie J, Hofman N, Burgh B, Cachucho R. Sports analytics for professional speedskating. Data Min Knowl Disc. 2017;31(6) 1872–1902. doi:10.1007/s10618-017-0512-3'

13. Peddarapu, R.K., Reddy, A.N., Pappula, S.S., Varma, P.B.D., Kumar, S.L., Yeada, S.G. (2023). Caption Suggestions Through Facial Expression Analysis. In: Senjyu, T., So–In, C., Joshi, A. (eds) Smart Trends in Computing and Communications. SMART 2023. Lecture Notes in Networks and Systems, vol 645. Springer, Singapore. https://doi.org/10.1007/978-981-99- 0769-4_61

14. Raya-Gonzalez J, Nakamura FY, Castillo D, Yanci J, Fanchini M. Determining the relationship between internal load markers and noncontact injuries in young elite soccer players.Int J Sports Physiol Perform. 2019;14(4):421–425. doi:10.1123/ijspp.2018-0466

15. C. Foster, "Monitoring training in athletes with reference to overtraining syndrome," Medicine and science in sports and exercise, vol. 30, pp. 1164–1168, 1998.

16. Alsahaf A, Azzopardi G, Ducro B, Hanenberg E, Veerkamp R, Petkov N. Estimation of muscle scores of live pigs using a kinect camera. IEEE Access. 2019;7:52238–52245. doi:10.1109/ACCESS. 2019.2910986.

17. J. D. Ruddy, A. J. Shield, N. Maniar, M. D. Williams, S. Duhig, R. G. Timmins, J. Hickey,
M. N. Bourne, and D. A. Opar, "Predictive modeling of hamstring strain injuries in elite australian footballers." Medicine and science in sports and exercise, vol. 50, no. 5, pp. 906– 914,2018.

18. R. K. Peddarapu, S. Ameena, S. Yashaswini, N. Shreshta and M. PurnaSahithi, "Customer Churn Prediction using Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1035-1040, doi: 10.1109/ICECA55336.2022.10009093.

19. Farrugia S, Ellul J, Azzopardi G. Detection of illicit accounts over the Ethereum blockchain.Expert Syst Appl. 2020; 150:113318. doi:10. 1016/j.eswa.2020.113318