■ RESEARCH ARTICLE

Predictive Analysis of Future Injury Using Machine Learning

WooMin Matthew Jeon

Asia Pacific International School (APIS), Seoul, 01874, South Korea; shinego345@gmail.com

ABSTRACT: We performed predictive analysis on the athlete's physical injury by leveraging multiple machine learning algorithms with the historical features of the athlete's injury. Injury is a significant concern in professional sports. Preventing physical injury is beneficial to sustain the athlete's performance and to extend their career. Recent advances in computing technology have made significant progress in injury prevention. Unfortunately, such an application is not easy. The span of the player's physical conditions is extensive, and most importantly, the area is intensively engaged in the medical regime. Acquiring athletes' injury information is strongly restricted due to personal privacy. We hypothesized that synthetic data would be a feasible tool to elucidate the recent methodological applicability for injury prediction if the athletes' physical condition is classified with their performance. Given this assumption, we evaluated the models with various metrics and inspected which features are more important for the likelihood of future injuries. Our result shows that training intensity is the most important feature, and the average accuracy is about 0.5 regardless of the models used. Since the main goal of this study is to illustrate the capability of prediction using machine learning models, we demonstrated the whole analysis procedure, including the evaluation of results.

KEYWORDS: Biomedical and Health Sciences, Sport Injury, Machine-Learning Aid, Injury Prediction.

■ Introduction

Injury prediction is a trending topic in competitive professional sports[1] Injuries are common but can have physical, psychological, and financial impacts on the athlete's mental health and performance.[2] The prediction of its occurrence or occurrence frequency plays a crucial role in enhancing the safety and performance of athletes.[3] Various complicated risk factors are associated with injury prediction, so simple modeling cannot be implemented alone. An individual athlete's clinical conditions also broadly vary with the type of sport.[4] In addition, disclosing the players' clinical information is unfavorable, undermining the prediction's accuracy. If multi-dimensional datasets such as biomechanics, environmental conditions, and historical injury records are provided, we can make accurate predictions. However, secured predictive models should be established to identify high-risk scenarios and individuals more prone to injuries.[5] This "foreseeing ability" is a matter of preventive measures, from which specific training programs or real-time monitoring can be facilitated to prevent injury.

The application of machine learning (ML) methods, a branch of artificial intelligence (AI), is widely adopted to improve injury prediction.[6] ML offers several distinct advantages when it comes to predicting injuries. Firstly, it can analyze large and complicated data much more effectively than conventual statistical approaches by discovering unknown patterns and relationships of the variable that might not be apparent. This capability allows for the more accurate identification of risk factors and early warning signs about specific types of injuries. Secondly, machine learning models can adapt and improve over time as they get more data and learn from new observations, making them increasingly effective and precise in predicting future incidents. Thirdly, these predictive models can be applied across various domains, from sports and healthcare to industrial settings, providing specific insights and interventions to prevent injuries before they occur. Interestingly, deep learning is also broadly facilitated for desirable outcomes. While ML highly relies on algorithms to process data and make predictions, deep learning uses artificial neural networks to predict from learning from its errors.[7] Since deep learning requires much more datasets than ML, it has more computation power and can avoid overfitting.[8] However, the most impactful feature of the prediction performance of deep learning is unknown, so different metrics need to be applied for highly accurate prediction.[7]

Several reviews[7,8,10] characterized the application features of machine learning (ML) and deep learning (DL) to sports injuries. Of course, various factors influence the outcomes: sport type, the way of exercising, players' physical performance, injury nature, and so on. Unfortunately, the expected advantages are still debated, and the acquired accuracy likely remains below expectations. As mentioned, determining the high-risk factors for a solid model could be challenging since the injury is in the medical regime. Without validation of the athlete's physical information like muscle development, exercise intensity, or chronic illness, extracting the risk factors for reliable level prediction is difficult. In the worst scenario, a superficial model may be established based on unvalidated factors, such as inappropriate evaluation of an athlete's physical characteristics. This tendency could worsen further if multi-dimensional datasets of athletes' bio information are not provided. Therefore, evaluating the methodological feasibility of ML and DL before practical application is recommended. We hypothesized that well-classified injury data would be enough to evaluate the methodological applicability of the computing aid analysis.

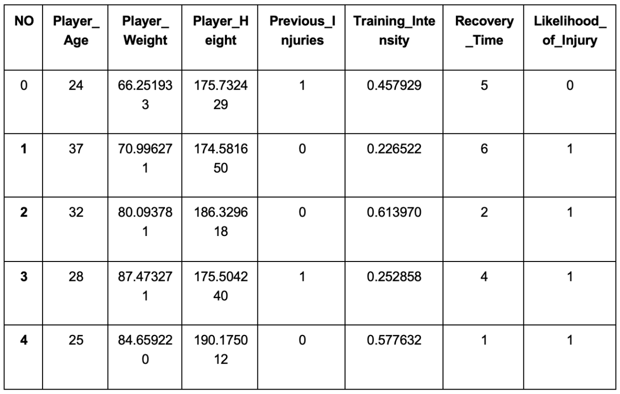
DOI: 10.36838/v7i7.8

The term "well-classified" means that it should reflect an athlete's physical performance injury history, training intensity, recovery time, and the likelihood of the injury. These attributes should be independent of probing their contributions with the prediction models. We found a synthetic dataset to satisfy these conditions and then attempted to evaluate the predictive performance for the likelihood of future injuries using machine learning algorithms.

■ Methods

Data:

Table 1: The raw dataset structure with the attribute. A part of the athletes' injury data was applied in this study. This raw data describes the likelihood of injury for the athlete related to biometric information.



NO	Player_Age	Player_Weight	Player_Height	Previous_Injuries	Training_Intensity	Recovery_Time	Likelihood_of_Injury
0	24	66.251933	175.732429	1	0.457929	5	0
1	37	70.996271	174.581650	0	0.226522	6	1
2	32	80.093781	186.329618	0	0.613970	2	1
3	28	87.473271	175.504240	1	0.252858	4	1
4	25	84.659220	190.175012	0	0.577632	1	1

Analysis:

Our analysis aims to predict the "Likelihood of Injury" based on given historical information called features. The prediction target and features are as follows:

- Target: Likelihood of Injury
- Features: Age, Weight, Height, Previous Injury, Training Intensity, Recovery Time

The dataset is synthetic and pre-processed, so no noisy points or outliers exist. However, after mounting the data on our process notebook, we further preprocessed the datasets to clarify the application of the machine learning analysis procedure. We split the dataset into a 75:25 ratio for machine learning model training and test, and then evaluated the 25% data for the prediction. The analysis procedure is depicted in Figure 1. For each model evaluation, the divided dataset was trained and tested again.

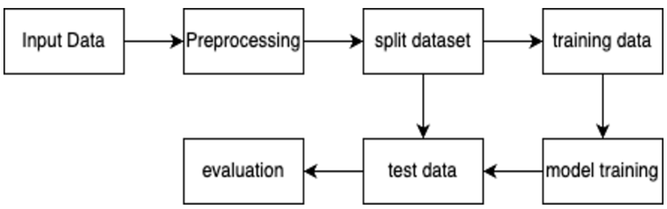


Figure 1: The procedure of analysis is depicted. The arrows indicated the flow of analysis. This sequence was utilized in the Python process applied.

First, all features were visualized to see if there were any strange data points in the distributions. Then, the correlation of features was inspected by plotting a correlation heatmap. This procedure helped us to determine if there are any features we can drop due to high correlation. If a strong correlation between the features appears, the machine learning models can learn the same information from the best-related feature. Figure 2 shows the Pearson correlation coefficient[12] among defined features and the variables for the target. The Pearson coefficient number is a measure that indicates the correlated degree of the two variables, as displayed in a color bar.

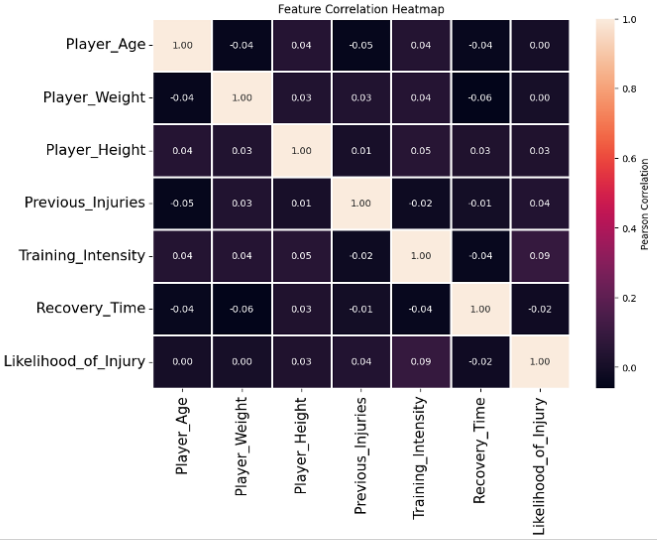


Figure 2: Correlation heatmap among features and the target (Likelihood of Injury). The values are Pearson correlation coefficients. The lighter the colors, the higher the correlation, as the color scale bar describes. Most values are less than 0.10, indicating no strong correlations between the attributes.

As shown in Figure 2, most coefficient values are lower than 0.05, implying that no highly correlated features exist, as expected. Hence, all features were applied for our analysis. We split the input data into two different sets: train and test sample sets. The training sample was implemented to train the models, while the test sample was applied to evaluate the models. This allowed us to minimize the evaluation bias since we were not using the same samples for training and testing. The splitting procedure used train_test_split from sklearn.[13] The ratio of train and test samples was set at 75:25 in stratified manners, meaning that the samples maintained the proportion of target portions in each sample. We tested the following 7 supervised models to see if there are any outperforming models:[7]

- Nearest Neighbors (NN)
- Linear SVM (LSVM)
- Naive Bayes (NB)
- Decision Tree (DT)
- Random Forest (RF)
- AdaBoost
- MLP

We selected the above models among others[7,8] because of the high-performance rate of multiple machine-learning algorithms. Some models were dropped due to technical difficulties. The default parameters are mainly used, as suggested in the example. Optimizing model parameters was challenging for this study. We propose such optimization as a future study. The confusion matrices for all models are presented in Figure 3. The values in the confusion matrices are close to 0.5 overall,

DOI: 10.36838/v7i7.8

which means that the model did not learn much the features. This is expected because we did not see a high correlation between features and the target, as shown in Figure 2.

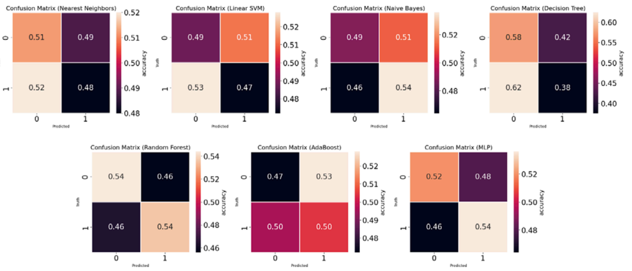


Figure 3: Confusion matrices for all models applied. The y-axis is truth information, and the x-axis is predicted values. The cell values are normalized. The values are mostly 0.50 ± 0.1-0.7, indicating 50% with 1-7% variation.

Results and Discussion

Table 2 summarizes the evaluation results using various metrics for all models applied in this study. Like the confusion matrices, the overall scores are around 0.5 with a standard deviation of ±0.025-0.58, regardless of the metrics. It also confirms our observation in confusion matrices: the models did not learn from the features. The Random Forest model shows better performance compared to others. Indeed, this trend is expected since ensemble models generally perform better. However, if we tune the model correctly, the AdaBoost and MLP would perform better than we currently see. However, tuning the model is out of the scope of this study and leaves it for our future study. The decision tree model shows poor performance compared to other models. We supposed that the training data is highly unbalanced and biased. Overall, Random Forest, MLP, and Naive Bayes perform slightly better, indicating slightly above 0.5 values.

Table 2: Summary of evaluation from the various metrics for all models applied in this study. These values completely reflect the acquired points from the confusion matrices.

Name	Accuracy	Precision	Recall	F1-Score	AUC
Nearest Neighbors	0.496	0.496	0.480	0.488	0.496
Linear SVM	0.480	0.480	0.472	0.476	0.480
Naive Bayes	0.512	0.511	0.536	0.523	0.512
Decision Tree	0.476	0.470	0.376	0.418	0.476
Random Forest	0.540	0.540	0.536	0.538	0.540
AdaBoost	0.488	0.488	0.504	0.496	0.488
MLP	0.528	0.522	0.536	0.5327	0.528
Std (each)	0.025	0.025	0.058	0.042	0.025

Figure 4 shows the importance of features for the two applied models (DT and RF), presenting the prioritization of which features are more critical for the model evaluation. This plotting was only available in those two models. Interestingly, the order of feature importance of both models is the same. The "Training Intensity" is the most utilized feature for both models' predictions. Then, the athlete's players' demographics (age, weight, height) are followed. The "Previous_Injuries" in DT is zero, meaning the feature was not used for learning. In contrast, RF shows that "Previous_Injuries" occupies a 0.025 rate.

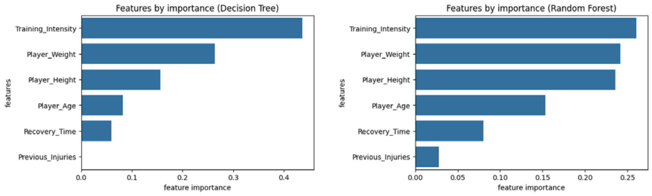


Figure 4: Feature importance of Decision Tree (DT) and Random Forest (RF). The y-axis shows the feature names, and the x-axis shows the importance score. The training intensity significantly contributes to the model prediction, but both models show a similar trend for the attributes.

Conclusion

In this study, we conducted a prediction analysis of the athlete's injury occurrence by applying various machine-learning models to synthetic data. The dataset is well-classified and strongly reflects the athlete's physical performance, including injury history, training intensity, recovery time, and the likelihood of the injury. Although the applied dataset is synthesized, its structural feature satisfied our hypothesis. We compared the performance with the possibility of injury for detailed analysis using various supervised models. The evaluation results did not appear to be well-performed. All prediction evaluated values are around 0.5, which is near 50%. We suppose this is due to the intrinsic characteristics of the synthetic data itself. The features applied for our evaluation are not highly correlated with the predicting variables, so the machine learning models could not learn valuable information from the features. This intrinsic feature is the exact characteristic we expected when choosing the synthetic data. Our study aimed to evaluate the capability of the methodological application to the athlete's injury prediction rather than examining how well machine learning aids injury prediction in predicting the injury likelihood. We found that the injury occurrence was highly relevant to the training intensity, as shown in the Pearson coefficient (Figure 2) and the contribution level plot (Figure 4). Nevertheless, the acquired accuracy of 50% does not mean machine learning aid prediction is not impractical. The review studies[7,8] exhibit that the overall performance accuracy ranges from 0.5 - 0.9 varying with factors such as type of sport, training intensity, model training method, location of injury, etc. However, the reviews addressed that implementing machining learning to injury prediction is challenging but an enabling tool to produce outstanding predictive projections from many sports-related datasets. Such technical applications could accelerate cost savings in the professional sports business because they can play a crucial role in enhancing the safety and performance of athletes. Considering this aspect, our study is meaningful because we showed how prediction analysis is generally done with various machine learning models, and this procedure can be directly applicable to actual data with similar features. From our modeling, we could prioritize the contribution of the individual features for prediction. However, the dataset should be uniformly balanced and classified to get acceptable injury prediction accuracy. This means that data collection must be

DOI: 10.36838/v7i7.8

carried out effectively using novel approaches. For example, it should be acquired by monitoring athletes' performance with highly sensitive player-worn sensors and video footage or by tracking individual athletes' biometrics with professional medical equipment and medical practitioners. This validation of the dataset is critical to making accurate predictions. Research on predicting injury occurrence is significantly committed to individual athletes' variability and appropriate training history. Time-dependent factors like fatigue, recovery rate, and training history also determine the prospective injury risk efficiency. Also, environmental factors like exercising gear and equipment, the opponent team, playing conditions, and team dynamics are considerable parameters. In this study, these effective factors were not taken into account in the original data, so our approach was not fully satisfactory. We speculate that these factors are the major contributors to the increase in uncertainty in our study. Importantly, various studies have claimed to make the correct decision on the risk factors predicted for injury occurrence.[6] However, the actual injury prediction's capability seems challenging unless the factors are compromised. Thus, as stated, we cannot determine which factors can lead to uncertainty in our approach to better strategic performance, even if we consider the intrinsic nature of our applied dataset.

Notably, at this moment, we cannot clarify how balanced data leads to reliable injury prediction at the level of our technical approach. As mentioned above, the prediction highly depends on various risk factors. We plan to investigate this issue for our future work, which will also further evaluate the different models by tuning the parameters. This proposed work would provide a better answer for the key uncertainty factors in the prospective injury predictive analysis.

Here, we can briefly narrate the characteristics of the methodological features of the applied models based on the literature that studied sports injury prediction (the script for this analysis is shared via the link 14): The details of the individual models refer to the reference.[8]

- Nearest Neighbors: easy to apply, but limited with data size and may be less accurate
- Linear SVM: as an ensemble model, applicable for high dimensional data
- Naive Bayes: simple probabilistic supervised classification with high accuracy
- Decision Tree: reasonable accuracy but limited with high dimensional data
- Random Forest: better performance accuracy but limited with high dimensional data
- AdaBoost: much better accuracy and possible with high dimensional data, compared to decision tree and random forest
- MLP: as a type of neural network, high accuracy with the capability of high dimensional data

Acknowledgments

Thanks to Dr. Z Sung (STEM Program Mentor) for helping to set up the central concept and for thoughtful technical comments.

References

1. Seow, D., Graham, I., Massey, A., Prediction models for musculos keletal injuries in professional sporting activities: A systematic review, Translational Sports Medicine, 09 July **2020**
2. Bahr, R., Krosshaug, T., "Understanding injury mechanisms: a key component of preventing injuries in sport", *British Journal of Sports Medicine* **2005**;39:324-329.
3. Caroline Finch, A new framework for research leading to sports i njury prevention", Journal of Science and Medicine in Sport, Vol 9, **2006**, Issue 1–2, pp 3-9.
4. Verhagen, E.A.L.M., Stralen, M.M., Mechelen, V., Behaviour, W. , the Key Factor for Sports Injury Prevention. **2010**, *Sports Med 40*, pp 899–906.
5. Clifton, D. R., Grooms, D. R., Hertel, J., Onate, J. A., Predicting I njury: Challenges in Prospective Injury Risk Factor Identification. J Athl Train. **2016** Aug, 51(8), pp 658-661.
6. Amendolara, A., Pfister, D., Settelmayer, M., Shah, M., Wu, V., D onnelly, S., Johnston, B., Peterson, R., Sant, D., Kriak, J., Bills, K., An Overview of Machine Learning Applications in Sports Injury Prediction", Cureus. **2023** Sep 28;15(9), e46170.
7. Van Eetvelde, H., Mendonça, L.D., Ley, C., et al. Machine learn ing methods in sport injury prediction and prevention: a systematic review", **2021**, *J EXP ORTOP 8*, 27.
8. Koosha Sharifani and Mahyar Amini, Machine Learning and De ep Learning: A Review of Methods and Applications, World Info rmation Technology and Engineering Journal, **2023**, Vol 10, Issue 07, pp. 3897-3904.
9. Leija Wang, "Injury Prediction in Sports: A Survey on Machine L earning Methods", The National High School Journal of Science, May 18, **2024**.
10. Ye, X., Huang, Y., Bai, Z., Wang, Y., "A novel approach for sports injury risk prediction: based on time-series image encoding and d eep learning", Front. Physiol., 17 December **2023**, Sec. Computati onal Physiology and Medicine, Volume 14.
11. https://www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset
12. "Pearson's Correlation Coefficient: A Comprehensive Overview" , Complete Dissertation by Statistics solution, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/
13. Scikit-Learn, Machine Learning in Python, "https://scikit-learn.org/stable/"
14. https://colab.research.google.com/drive/19D-8N8afmqgmHJVMyT4BX9M-qt7fzLik?usp=sharing

Author

Woomin Matthew Jeon is a senior at Asia Pacific International School, interested in biology for a pre-med program to pursue a track for medical school. He hopes to build his career as a medical doctor in the field of medical technology.

DOI: 10.36838/v7i7.8