Fang J and Xiang T. (2024) MEDICAL DECISION SUPPORT FOR FOOTBALL PLAYERS BASED ON MACHINE LEARNING HISTORICAL INJURY DATA. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 24 (96) pp. 479-489.
DOI: https://doi.org/10.15366/rimcafd2024.96.029

ORIGINAL

MEDICAL DECISION SUPPORT FOR FOOTBALL PLAYERS BASED ON MACHINE LEARNING HISTORICAL INJURY DATA

Jinhua Fang[1], Ting Xiang[2*]

[1] College of Physical Education and Health, Guilin Institute of Information Technology, Guilin, Guangxi 541100, China
[2] School of Outdoor Sports, Guilin University of Tourism, Guilin, Guang Xi, 541006, China
E-mail: xianglaoshi6688@163.com

Recibido 18 de octubre de 2023 **Received** October 18, 2023
Aceptado 18 de puede de 2024 **Accepted** May 18, 2024

ABSTRACT

Utilizing massive clinical data in the field of football players injuries for assisted medical decision support is the core technology and inevitable development trend of smart healthcare. However, due to the characteristics of medical data such as feature redundancy and imbalance of data sample categories, it has been difficult for traditional data mining algorithms to be directly applied in medical data research. In this paper, we propose a data-driven football players injury prediction method based on the experimental study of football players injuries occurring during the learning process of youth professional soccer training, which is based on the machine learning method of decision tree classifier. Through a semester of data statistics and experiments, the model has a high accuracy of injury prediction, which can provide early warning of youth football players injuries and support medical decision-making.

KEYWORDS: Machine Learning; Injury Data; Medical Decision Support

1. INTRODUCTION

The healthcare field generates a large amount of complex data about patients' clinical examinations, treatment reports, hospital resource management records, electronic medical records, medicines, and so on (Dash, Shakyawar, Sharma, & Kaushik, 2019). Many diseases, especially chronic diseases, have an insidious onset, long duration, and complex etiology. Traditional medical decision-making methods often make it difficult to accurately analyze and diagnose such diseases, which affects the effectiveness of treatment. It has been reported that the number of deaths due to chronic diseases in China has accounted for 86.6% of the total number of deaths (He et al., 2022), resulting in a disease burden that has accounted for 70% of the total disease burden. The large amount of medical data necessitates more effective extraction and processing methods to correctly diagnose diseases and predict the likelihood of a patient's illness or lesion via machine learning technology, which balances the knowledge of human experts in relevant fields with the data analyzing and processing potential of computers, in order to obtain the best disease diagnosis and improve the current situation of disease prevention, diagnosis, and research. Medical decision support, as the key technology of smart healthcare, blends IT, artificial intelligence, and healthcare, and has enormous benefits for disease prevention and patient care (Al-Asadi, 2018). Among the huge amount of medical data, sports injury cases are also an important part of it. Historical injury data, characterized by complex structure and large information level, are not conducive to efficient and highly accurate diagnosis by doctors. By introducing computer technology into the medical industry and establishing a medical decision support system, it can help doctors in the whole process of diagnosis, treatment, examination, and drug cost of this series of diagnosis and treatment and provide great convenience for both doctors and patients. Medical Decision Support System (DSS) analyzes and reasoning of medical data and related professional knowledge of various structures through machine learning and artificial intelligence methods, so as to assist doctors in making diagnosis decisions or risk prediction of diseases (Al-Asadi, 2018; Luu et al., 2020). DSS can be categorized into two types: fuzzy rule systems (Anooj, 2012) and machine learning-based DSS. The difference is that fuzzy rule-based systems can extract rules for information that is not suitable for quantification, such as a certain symptom that can be caused by multiple diseases and as mentioned in the literature (Gadaras & Mikhailov, 2009), fuzzy decision support systems can extract rules from human experts in related fields, which means that fuzzy decision support systems can do their job well even without a sufficient number of samples, whereas machine learning methods need to train a classification model or a prediction model on the selected samples. The machine learning approach requires training the selected classification or prediction models with samples, and then using the trained models for disease diagnosis or risk prediction. In this paper, machine learning-based medical decision support systems are introduced as the core (Claudino et al., 2019). Machine learning-based medical decision support mainly includes two aspects of disease risk prediction and intelligent diagnosis, and its applications have been widely carried out in many fields: Du et al. proposed a support vector machine algorithm based on FOA optimization parameters, which was successfully applied to medical diagnosis prediction, and the prediction effect of this algorithm was better than that of support vector machine (PSO-SVM) and support vector machine (GA-SVM) based on genetic optimization algorithm (Du, Liu, Yu, & Yan, 2017) through the comparison of the UCI dataset. Support Vector Machine (PSO-SVM) and Genetic Optimization Algorithm Support Vector Machine (GA-SVM). Seera M et al. combined neural network and classification regression tree to solve the problem of medical data classification, and the feasibility of the algorithm was concluded by comparing the quantitative metrics such as accuracy, specificity, and so on (Seera & Lim, 2014). Gorzałczany M B et al. proposed a Multi-Objective Optimization Algorithm (MOEOA) for the prediction of medical diagnosis and the prediction of medical diagnosis (Gorzałczany & Rudziński, 2017). algorithm (MOEOA) based fuzzy rule classification system, which mines simple and easy to understand rules and is suitable for use in mining various medical data (Theron, 2020). This paper presents a data-driven football players injury prediction method based on a decision tree classifier machine learning approach based on an experimental study of football players injuries learned from a soccer ball. The prediction method is based on the exercise data collected during a recent training session and predicts the risk of football players injuries that may occur during a training session in a future cycle (Karnuta et al., 2020; Ruddy et al., 2019).

2. Methodology

2.1 Principle of Decision Tree Classifier

At this stage, there are many algorithms used for healthcare data analysis and processing, and the decision tree algorithm (Soleimanian, Mohammadi, & Hakimi, 2012) is one of the more common algorithms. The main task of the decision tree is to extract rules from the data and to predict the category to which the new data belongs. When dealing with complex problems, the classification rules constructed by the decision tree are similar to a tree structure, forming the classification rules sequentially from top to bottom, so the rules are simple and easy to understand, and are often used to extract the intrinsic information of medical data. Hunt EB et al. first proposed the conceptual learning system (CLS) learning algorithm in 1966 (Diehr & Hunt, 1968). Conceptual learning system is also the first decision tree algorithm. The decision tree algorithm uses information theory knowledge to analyze and summarize all the attributes of the training samples, and finally generates a model similar to a tree structure. In the tree structure of a decision tree, the nodes represent the categories of the samples and the branches represent the classification rules. The root node of the tree is the most informative attribute in the training sample, and the middle node of the tree indicates that the subtree to which it belongs contains the most informative attribute in the training sample. The decision tree searches for a path from the root node that is most suitable for the test data to determine the category of the test data, which is based on the principle of selecting the attribute with the greatest information gain, dividing the dataset into several subsets, and then using a recursive method to classify each subset into the same category, and then finally obtaining a model for the new dataset's classification and prediction. Decision trees are generated by recursive splitting, i.e., repeated splitting of values of attributes (Wang, Kwong, Wang, & Jiang, 2014). Attributes are selected and split based on selection criteria, such as sport load, age of the athlete, number of injuries, and physical status. The basic idea of selecting any splitting criteria in the internal nodes is to make the data in the child nodes belong to a certain category (injured or non-injured). In general, the recursion stops when all training instances have corresponding categorized values, i.e., the dataset has completed the operation of injury prediction.

2.2 Decision Tree Modeling

The construction of a decision tree model includes: tree generation, selection of classification attributes, and cross-validation. The main task of cross-validation is to test the constructed classification rules and construct a decision tree with high accuracy through repeated corrections (Kim, 2009). The modeling process of the decision tree is as follows:

(1) Generate and build a tree: A decision tree generates a model that resembles the structure of a tree, including root nodes, branch nodes, and leaf nodes. A leaf node represents a category of the data sample, and a branch node represents a path to its leaf node.

(2) Attribute selection and splitting: Information Gain, Information Gain Rate and Gini Coefficient are the more common methods used for attribute splitting, in which the smaller the Gini Coefficient is, the more reasonable the rules of classification are. To select the attribute splitting based on the Gini coefficient, let C be the number of categories in the training samples and Sbe the number of samples, if the probability of the $i-th$ category appearing in S is p_i , then the Gini coefficient of S is denoted as:

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \qquad (1)$$

If the set S is divided into S_1, S_2 according to the attribute C, the sample numbers of S_1, S_2 are denoted by s_1, s_2, respectively, and the sample number of is denoted as S, the Gini coefficient of the attribute $C$is calculated as:

$$Gini(C) = \frac{s_1}{s} Gini(S_1) + \frac{s_2}{s} Gini(S_2) \qquad (2)$$

The Gini coefficient is calculated for each attribute in the set of candidate attributes that may be selected as a split attribute, and the attribute with the smallest Gini coefficient is used as the best split attribute for the partition, and the best partition Gini coefficients are set for all the candidate attributes and compared to the candidate attributes. The attribute with the smallest Gini coefficient is used as the final test attribute. This method is suitable for sample data with fewer categories and generates subsets of similar size.

(3) Cross-validation: Cross-validation is a very critical step in the construction of decision tree models, and the goodness of the extracted rules has a great relationship with cross-validation. The conventional practice is to use the training set to test the constructed decision. If the training samples are good, then the whole data set is used for testing, and the classification accuracy can be estimated to determine whether the extracted classification rules are reliable or not.

(4) Pruning: The constructed decision tree model is prone to overfitting. Pruning is often used to solve this problem. There are mainly pre-pruning and post-pruning, the former is to judge whether the nodes currently being processed are to be further classified by certain judgment criteria in the process of tree building; the latter is to let the tree "grow fully" by adopting the greedy strategy, and then prune the tree according to the classification error rate. The more commonly used pruning method for decision tree pruning is minimizing the loss function. Let the number of nodes of node G be $|G|$, and t be the leaf node of G, then the number of samples of the leaf node is N_t, and the loss function is obtained as follows:

$$C(G) = \sum_{i=1}^{|G|} N_t H_t(G) + \alpha|G| \qquad (3)$$

The first term in the formula is the loss function and the second term represents the complexity of the decision tree model. When α is determined, the decision tree pruning operation needs to take into account both the error and the complexity of the model.

3. Training Data Preparation

3.1 Data collection

A portable 10 Hz GPS, 100 Hz 3D accelerometer, 3D gyroscope, and 3D digital compass were used to collect volume and load data from the subjects. Each subject wore a tight-fitting undershirt with the receiver placed between the shoulder blades, and each player wore his or her own acquisition device during each training session.

A total of 927 training sessions were recorded over a 23-week period, and a set of exercise load metrics were extracted from the data using a software package. Twelve characteristics were extracted from each exercise data set, describing the kinematic, metabolic, and physical characteristics of each soccer participant in terms of exercise load. Information on age, weight, height and field position was also collected for each participating soccer subject. Table 1 provides a description of the functions considered.

Table 1: Characterization of training loads

EIGEN-SYMBOL	CHARACTERIZATION OF THE TRAINING LOAD
DIT	Distance of movement in meters during training
DEX	Exercise distance in excess of 40 m/s
DMC	Exercise distance in units of metabolic capacity
HML	Exercise where the athlete's metabolism exceeds 200 W/Kg
HML/M	Average dm per minute
DEXP	Exercise distance in meters with more than 25.5W/Kg less than 19.8Km/h
ACC2M	Number of times acceleration exceeds 2m/s
ACC3M	Number of times acceleration exceeds 3m/s
DCC2M	Deceleration exceeding 2m/s
DCC3M	Deceleration exceeding 3m/s
FIS	Impact forces exceeding 2g. Impact force is the sum of collision and ground impact forces in training
ROF	Ratio of DSL to velocity intensity
AGE	Age of trainee
BMI	Body Mass Index (BMI): weight (in kg) and height squared (in meters)
ROLE	Athlete's position on the field (forward, midfielder and defender)
NIP	Number of injuries prior to participation in training
CTH	Cumulative training hours
GAMES	Length of participation in official matches prior to training

During the 23-week period, the medical personnel documented all non-contact injuries. A non-contact injury was defined as a tissue injury caused by a soccer player missing at least the second day of training after the injury occurred during the next training session. This dataset had 21 non-contact injuries.

3.2 Feature extraction and dataset construction

Four training datasets were constructed based on the 12 exercise load features described in Table 1, and each training dataset included 954 data. 1. Load characterization dataset for exercise. The sports load feature dataset (WF) is created by utilizing exponentially weighted moving average (EWMA) for the training load data from the previous six training sessions. In this work, the EWMA of the feature PI was estimated with a span of 6 (PIWF) to account for the athlete's previous injuries as well as the temporal distance of the current training session. PIWF=0 indicates that the soccer trainee has never been hurt; PIWF>0 indicates that the soccer trainee has had at least one injury in the past; and PIWF>1 indicates that the soccer trainee has suffered multiple injuries in the past. 2. Short-term/long-term exercise load ratio data set (ACWR). The likelihood of sport injury was estimated using criteria normally used in sports science, i.e., calculating the ratio between the last 6 training sessions of EWMA and the previous 28 days of EWMA. 3. MSWR stands for mean ratio of mean over standardized exercise load data set. Based on another method for estimating the likelihood of football players injuries, calculating the ratio between the mean and standard deviation of the training workload over the previous 6 days. The lower the variability of an athlete's exercise load throughout training, the greater his MSWR. 4. A dataset was built using the three feature sets stated above (WF, ACWR, and MSWR) as well as the exercise load features. This dataset provides a vector of 42 characteristics and labels that indicate if an injury occurred during the subsequent training session.

4. Data processing

First, we use a decision tree classifier to do feature selection, which reduces the dimensionality of the feature space and therefore the chance of overlap. To choose the optimal collection of features that can predict injuries in our dataset, we employ submission feature elimination and cross-validation (RFECV). We train a Decision Tree Classifier (DT) and a Random Forest Classifier (RFC) on the new training dataset generated by feature selection. Therefore, in this paper, we record the football players data of youth soccer players at the beginning of a semester and train the predictive classification continuously as the training sessions progress. Before the start of week W_i of the training program, the athletic data of week $W_1 \sim W_i$ were studied and the ability to predict injuries in week W_{i+1} was evaluated. Given that injury prediction is a binary classification problem with a positive injury class (1), we asses s classification goodness in terms of precision, recall, F1 score, and AUC. Precision is the ratio of successfully classified examples to all instances allocated to the class by the classifier. Recall represents the proportion of instances of a given category that the classifier correctly classifies, and the harmonic mean of precision and recall is used to get the F1 score. The likelihood that the classifier would categorize a randomly picked positive instance over a randomly selected negative instance (assuming that "positive" is greater than "negative") is expressed as the area under the curve. (Assuming "positive" is preferred over "negative"). AUC near 1 suggests accurate classification, while AUC near 0.5 shows random classification. We compare the predictive performance of DT and ETRFC based on four benchmark classifications. Benchmark classification B1 assigns instances to classes based on a random assignment principle. Benchmark B2 always assigns the majority of classes (i.e., non-injured classes), while the benchmark classification always assigns the minority of classes (i.e., injured classes). If the exponentially weighted mean variable P I > 0, the base classification B4 is a classifier based on classification 1 (injury), otherwise the base classification B4 is a classifier based on classification 0 (no injury). Through the feature selection task, three out of 42 feature vectors were selected: PI^{WF}, dMSWR HML, and DECWF 2.

The function PI^{WF} measures the time difference between a player's current training time and the usual training time of a player who has been injured in the past. The parameters dMSWR HML and DECWF 2 represent two training qualities, high metabolic load and quick deceleration, respectively. We discovered that 41.77% of the injuries detected by the injury classifier happened after a previously injured athlete returned to regular training and had the eigenvalues dMSWR HML and DECWF 2, which represent the mean values of metabolic load variability and sudden deceleration over the previous 6 days, respectively.

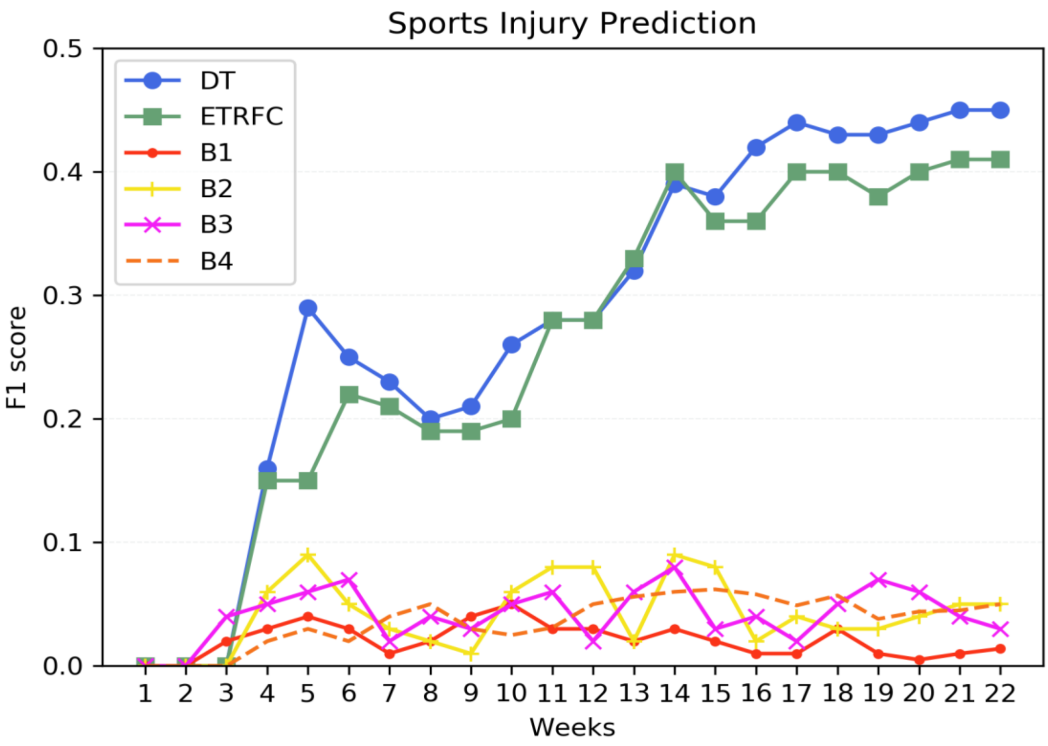

Figure 1: Prediction of football players injuries

Because of the low number of injuries at the start of the training program, the predictors performed poorly and missed several injuries. However, the classifier's predictive power improved over time, and it predicted the majority of the injuries in the second half of the training phase. The data in Figure 1 reveal that the first phase has a greater influence on the classifier's cumulative performance. This implies that attempting to forecast injuries from the start is not a viable strategy, as classification performance may be low at first due to data paucity. An initial phase of data collection is required in order to collect sufficient data, the length of which will rely on the needs and plans of young soccer player management. When the data in Figure 1 was examined, it was discovered that the classifier's performance had stabilized after 16 weeks of data collecting. As a result, beginning in week 16, the most reasonable technique may be to use the classifier for injury avoidance. From the data shown in Fig. 1, it can be seen that DT detected more than half of the injuries

(11 out of 21 injuries), and with an F1 score of 0.45, it is the classifier with the highest prediction accuracy in the figure. Comparison of the prediction metrics of the DT and ET RF C classifiers with the baseline classifiers is shown in Table 2. As can be seen from Table 2, the prediction performance of the DT over the ETRFC was much better than that of the benchmark classifications. At the end of the test cycle, DT detected 62% of the football players injuries (recall rate of 0.62) and correctly predicted 42% of the football players injuries (precision of 0.42). From the test results, the machine learning method can improve the scientific data support for the prevention of football players injuries in adolescents.

Table 2: Comparison of projected indicators

CLASSIFIER	CLASS	ACCURACY	RECALL	F1	AUC
DT	0	0.97	0.98	0.98	0.75
	1	0.42	0.62	0.45	
ETRFC	0	1.00	0.99	0.97	0.72
	1	0.35	0.57	0.43	
B4	0	0.96	0.75	0.85	0.55
	1	0.03	0.18	0.14	
B3	0	0.98	0.98	0.98	0.53
	1	0.04	0.03	0.04	
B2	0	0.98	1.00	0.99	0.51
	1	0.00	0.00	0.00	
B1	0	0.00	0.00	0.00	0.51
	1	0.01	1.00	0.06	

The classifiers performed marginally worse than the previous three selected features when using the whole feature dataset for DT, ETRFC, and baseline training, with precision, recall, F1 score, and AUC of 0.39, 0.56, 0.45, and 0.73, respectively. In order to investigate whether the position of the athlete affects the likelihood of injuries, different classifiers were applied to the three positions of the athlete (defender, midfielder and striker) and comparisons revealed much worse performance than the classifiers without differentiation. In order to investigate whether the position of the athlete affects the likelihood of injury, we used different classifiers for the three positions (defender, midfielder and striker), and found that the performance was much worse than that of the classifiers that did not differentiate between the positions, with a precision, recall, F1 score and AUC of 0.01, 0.03, 0.04 and 0.53, respectively.

5. Conclusion

A strategy for forecasting injuries in youth soccer players is proposed in this study. Our technology can be used by athletic trainers, coaches, and physical therapists to optimize training sessions, thereby preventing injuries, improving training, and lowering rehabilitation expenditures. The suggested study shows how machine learning may be utilized to handle tough football players analytical problems like injury prediction. If the dataset is enlarged to include more teams, it is possible to develop a more generalized and stable injury prediction algorithm. When dealing with a larger number of injuries, the problem can be converted from a binary classification (injury/no injury) to a multilevel classification or regression problem, where information about the kind or severity of the injury can be used to make more diverse predictions. Finally, due to the model's adaptability, the prediction method can be easily extended to forecast injuries in other professional football players or in adult athletes.

Acknowledgement

1. 2022 Guangxi University young and middle-aged teachers Research basic ability Improvement Project (Number: 2022KY1608). 2. 2023 Guangxi Higher Education Undergraduate Teaching Reform Project, (Number: 2023JGA419).

Reference

Al-Asadi, M. A. M. (2018). Decision support system for a football team management by using machine learning techniques.

Anooj, P. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences, 24*(1), 27-40.

Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports Medicine-Open, 5*, 1-12.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data, 6*(1), 1-25.

Diehr, G., & Hunt, E. B. (1968). *A comparison of memory allocation algorithms in a logical pattern recognizer*: Department of Psychology, University of Washington.

Du, J., Liu, Y., Yu, Y., & Yan, W. (2017). A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. *Algorithms, 10*(2), 57.

Gadaras, I., & Mikhailov, L. (2009). An interpretable fuzzy rule-based classification methodology for medical diagnosis. *Artificial intelligence in medicine, 47*(1), 25-41.

Gorzałczany, M. B., & Rudziński, F. (2017). Interpretable and accurate medical data classification–a multi-objective genetic-fuzzy optimization approach. *Expert systems with applications, 71*, 26-39.

He, L., La, Y., Yan, Y., Wang, Y., Cao, X., Cai, Y., . . . Feng, Q. (2022). The prevalence and burden of four major chronic diseases in the Shanxi Province of Northern China. *Frontiers in public health, 10*, 985192.

Karnuta, J. M., Luu, B. C., Haeberle, H. S., Saluan, P. M., Frangiamore, S. J., Stearns, K. L., . . . Makhni, E. C. (2020). Machine learning outperforms regression analysis to predict next-season Major League Baseball player injuries: epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000-2017. *Orthopaedic journal of sports medicine, 8*(11), 2325967120963046.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis, 53*(11), 3735-3745.

Luu, B. C., Wright, A. L., Haeberle, H. S., Karnuta, J. M., Schickendantz, M. S., Makhni, E. C., . . . Ramkumar, P. N. (2020). Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: an analysis of 2322 players from 2007 to 2017. *Orthopaedic journal of sports medicine, 8*(9), 2325967120953404.

Ruddy, J. D., Cormack, S. J., Whiteley, R., Williams, M. D., Timmins, R. G., & Opar, D. A. (2019). Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Frontiers in physiology, 10*, 829.

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert systems with applications, 41*(5), 2239-2249.

Soleimanian, F., Mohammadi, P., & Hakimi, P. (2012). Application of decision tree algorithm for data mining in healthcare operations: a case study. *Int J Comput Appl, 52*(6), 21-26.

Theron, G. F. (2020). *The use of data mining for predicting injuries in professional football players.*

Wang, R., Kwong, S., Wang, X.-Z., & Jiang, Q. (2014). Segment based decision tree induction with continuous valued attributes. *IEEE transactions on cybernetics, 45*(7), 1262-1275.