An educational review on machine learning: a SWOT analysis for implementing machine learning techniques in football.

Marco Beato[1*], Mohamed Hisham Jaward[2], George P. Nassis[3,4], Pedro Figueiredo[3,5], Filipe Manuel Clemente[6,7,8], Peter Krustrup[4,9]

Affiliations

1. School of Allied Health Sciences, University of Suffolk, Ipswich, United Kingdom.
2. School of School of Technology, Business and Arts, University of Suffolk, Ipswich, United Kingdom.
3. Physical Education Department, United Arab Emirates University, Al Ain, United Arab Emirates.
4. Department of Sports Science and Clinical Biomechanics, Sport and Health Sciences Cluster (SHSC), University of Southern Denmark, Denmark
5. Research Center in Sports Sciences, Health, Sciences and Human Development, CIDESD, Vila Real, Portugal.
6. Escola Superior Desporto e Lazer, Instituto Politécnico de Viana do Castelo, Rua Escola Industrial e Comercial de Nun'Álvares, 4900-347 Viana do Castelo, Portugal.
7. Gdansk University of Physical Education and Sport, 80-336 Gdańsk, Poland.
8. Sport physical activity and health research innovation and technology center (SPRINT), 4900-347 Viana do Castelo, Portugal.
9. Danish Institute for Advanced Study (DIAS), University of Southern Denmark, Odense, Denmark.

*Corresponding author

Marco Beato, School of Allied Health Sciences, University of Suffolk, Ipswich, United Kingdom, email: m.beato@uos.ac.uk

Short title/running head: A SWOT analysis on machine learning in football.

Abstract

Purpose: The abundance of data in football presents both opportunities and challenges for decision-making. Consequently, this review has two primary objectives: first, to provide practitioners with a concise overview of the characteristics of machine learning (ML) analysis; and second, to conduct a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis regarding the implementation of ML techniques in professional football clubs. This review explains the difference between artificial intelligence and ML, and the difference between ML and statistical analysis. Moreover, we summarize and explain the characteristics of ML learning approaches such as supervised learning, unsupervised learning and reinforcement learning. Finally, we present an example of SWOT analysis, which suggests some actions to be considered in applying ML techniques by the medical and sport science staff working in football. Specifically, four dimensions were presented namely the use of strengths to create opportunities and make the most of them, the use of strengths to avoid threats, work on weaknesses to take advantage of opportunities, and upgrade weaknesses to avoid threats.

Conclusion: ML analysis can be an invaluable ally for football clubs, sport science and medical departments due to its ability to analyze vast amounts of data and extract meaningful insights. Moreover, ML can enhance performance by assessing the risk of injury occurrence, physiological parameters, physical fitness, and optimizing training, recommending strategies based on opponent analysis, and identifying talent and assessing player suitability.

Key Points: Strengths, Weaknesses, Opportunities, Threats, decision-making, performance prediction, injury risk assessment, Soccer

INTRODUCTION

The decision-making process plays a critical role for practitioners working in football. Practitioners aim to optimize the training process, testing protocols, physiological parameters, physical readiness, and match strategies to increase the probability of success.[1–3] In the last decades, technology has allowed sports scientists and performance analysts to collect larger volumes of data compared to the past, [4–6] and use them in conjunction with their experience and the most relevant scientific evidence to make informed decisions. These data have been typically analyzed using visualizations and statistical methods. Nevertheless, challenges arise when determining how to effectively select variables and handle larger datasets derived from multiple sources and instruments. In recent years, artificial intelligence (AI) and machine learning (ML) have become more pervasive in football [6–8]. Although the use of AI and ML are common in our contemporary society, some confusion exists between the two terms. AI can be briefly defined as "the theory and development of computer systems able to perform tasks normally requiring human intelligence",[9] while ML refers to "the technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience (training)"[10] and it is a subset of AI. ML can find several applications in football, for example, to facilitate decision-making, performance prediction, technical and tactical pattern recognition, game activity/analytics, talent identification, and injury risk assessment.[7,11,12]

Data mining is the process of sorting through large data sets to identify patterns and relationships.[13] Through data mining and ML (which focuses on creating algorithms that can learn and predict from given data)[7] football practitioners (e.g., sports scientists and coaches) can make informed decisions to enhance physiological parameters, physical development, reduce fatigue, increase readiness, and match performance. A recent review reported that ML can be used to determine the parameters that affect (i.e., explainability, which means that a model and its output can be explained and make sense to a human being) wellness and fitness, which can be later on manipulated by football practitioners.[7] ML regression can determine the contribution of players' anthropometric characteristics to physical performance, such as sprinting and aerobic fitness.[14] Furthermore, ML can be used to assess the relationship between well-being parameters and training load and match performance. However, it showed a limited predictive capacity of such parameters to determine internal and external load.[15] ML analysis can be used for determining technical and tactical outcomes, for instance, to analyze the team pattern or the effectiveness of passing strategies.[7] ML was used to estimate players' passing skills to make predictions for the following season,[16] which coaches and performance analysts could use for scouting objectives. Moreover, multiple ML algorithms were used by Jamil et al.,[17] to classify elite and sub-elite goalkeepers (GK) in professional men's football, suggesting that a GK's ability with their feet and not necessarily their hands are what distinguishes the elite GK's from the sub-elite. Another area in which ML can be used is talent identification, which is one of the more critical challenges for football clubs. In this specific context, technical and tactical variables, together with psychological and physical variables can be assessed to determine the talent predictors that coaches need to monitor and develop.[18,19] Such information may impact the productivity (in terms of talent) of football academies and related clubs. Certainly, ML holds the promise to overcome the constraints of conventional reductionism approaches, enabling the concurrent integration of diverse data sources. It may play a pivotal role in gathering a comprehensive understanding of the game by bridging gaps across physical, physiological, technical, and tactical dimensions, while simultaneously contextualizing the information and actively pursuing integrative models. This advanced approach may accelerate analyses but also potentially heightens accuracy, thereby strengthening decision-making processes in coaching, player development, and overall team performance.

Research in the field of ML for identifying injury risks and associated factors has been steadily growing over the years, as evidenced by a recent systematic review [11]. For instance, in a study by Oliver et al.,[20] involving 355 elite youth football players, decision tree algorithms displayed an overall accuracy that was not significantly superior to statistical logistic regression in detecting injuries. However, ML (e.g., decision tree) demonstrated increased sensitivity in this context. In contrast, a study by Rommers et al.,[21] which employed extreme gradient boosting algorithms on a larger sample of 734 youth players, revealed promising results. The ML algorithm successfully identified injured players in the hold-out test sample with 85% precision, 85% recall (sensitivity), and 85% accuracy.[21] Additionally, the same study[21] achieved reasonably high accuracy in distinguishing between overuse and acute injuries based on pre-season measures. Hence, beyond predicting potential injuries, ML has the potential to categorize them effectively. This capability provides additional insights for rapidly constructing models in subsequent stages of interpretation. Furthermore, it facilitates interaction with potential injury mechanisms and factors that may influence the overall risk.[22]

To successfully implement ML in football, practitioners need to address the integration into medical, sport science, and coaching departments. A strategic management plan, anchored by a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis, is vital to evaluate the team’s internal capabilities and external possibilities concerning ML. This analysis will inform strategic decisions, leveraging strengths to harness opportunities or neutralize threats, and improving weaknesses to support ML adoption. A team's preparedness to adopt ML is crucial, as it can significantly refine their strategic approach to ML utilization, ensuring a more effective and efficient integration. This condensed strategy enables teams to navigate the complexities of ML implementation in the competitive sports environment.

In the dynamic field of football, the profusion of data creates a spectrum of possibilities and hurdles in the decision-making process. Addressing this, the review unfolds in two distinct parts: the first segment offers practitioners a streamlined synopsis of ML analysis features; the second segment presents a comprehensive SWOT analysis, assessing the practicality and impact of integrating ML methodologies within the ecosystem of professional football clubs.

MACHINE LEARNING

Difference between AI and ML

Before delving into ML in football, it is important to appreciate the evolution of ML to the modern form used to solve many real-world problems. As mentioned in the introduction, ML constitutes a subset within the broader field of AI. Modern AI gained prominence in the early 1940s and the seminal work of McCullogh and Pitts is considered as the first work on the artificial neuron (they defined a mathematical computation model similar to neural networks).[23] Various AI initiatives aim to emulate human intelligence through computational models based on artificial neurons. Consequently, AI encompasses a wide spectrum of tasks and issues, in contrast to ML, where the primary objective is the development of algorithms tailored for specific tasks. Frequently, everyday tasks can be formulated as either regression or classification problems, and ML endeavors to address these challenges systematically.

Difference between ML and statistical analysis

In numerous data science scenarios, the principal goals are inference and prediction. Inference involves creating a mathematical model of the data-generation process to formalize understanding or test hypotheses regarding system behavior. As an example in football, Zeki et al.[24] infer the neuromuscular fatigue imposed on players after a football match based on measurements such as the players' heart rate, accelerations, and distance traveled. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a football player will likely develop an injury in a future game. In a typical research project or applied setting, both inference and prediction can be of value - we want to know how the system works and what will happen next.[25]

Many methods from statistics and ML may, in principle, be used for both prediction and inference. However, statistical methods have a long-standing focus on inference, which is achieved through the creation and fitting of a probabilistic model onto the data.[26] The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a 'true' effect, and so unlikely to result from noise or disturbances. In contrast, ML emphasizes prediction, employing learning algorithms to identify patterns in complex and big datasets.[26] ML techniques prove particularly advantageous when dealing with situations where the number of input variables surpasses the number of samples, as opposed to scenarios with more samples than input variables. ML operates with minimal assumptions about data-generating systems, exhibiting efficacy even in instances where data collection lacks a meticulously controlled experimental design or involves intricate nonlinear interactions. Also, ML allows for the interdependence of data points and facilitates the identification of hidden targets/groups without needing a subjective setting while providing an error estimation.[27] However, despite achieving compelling predictive outcomes, the limited interpretability of numerous ML solutions poses challenges in directly addressing specific problems and applying them in safety-critical applications. Often, statistical methods, including hypothesis testing, are employed to validate ML outcomes, and the relative performance of ML methods is commonly compared using hypothesis testing approaches.

Classical statistics and ML diverge in terms of computational tractability as the number of variables per subject increases.[26] Classical statistical modeling, originally designed for datasets with a limited number of input variables and sample sizes considered small to moderate by contemporary standards, encounters challenges as the complexity of the relationships among numerous input variables increases. Consequently, statistical inferences become less precise and the boundary between statistical and ML approaches becomes hazier.

SUPERVISED AND UNSUPERVISED ML METHODS

Supervised learning

In supervised learning, a model is derived from a dataset that incorporates features and labels, with both entities employed during the training phase (see Table 1).[17,28,29] Once the model is trained, it predicts the label corresponding to the input features (values that a supervised model uses) when presented with unseen input (the value we want the model to predict). The supervisor oversees the learner's every move, dictating precise actions for every situation until the learner masters the mapping from situations to actions. While working under such close supervision may seem restrictive, the process is relatively straightforward – quickly recognizing patterns and replicating the supervisor's actions ensures compliance.

In supervised ML, the supervisory aspect is crucial, as it forces the model to learn parameters of the model such that the output given by the model is close to the desired output indicated by the label. In probabilistic terms, the focus is typically on estimating the conditional probability of a label given specific input features. While supervised learning represents just one paradigm among several, it predominates in the success of ML applications across various domains.[30,31] This prevalence is attributed, in part, to the fact that many pivotal tasks, such as those listed below, revolve around estimating the probability of an unknown attribute given a specific set of available data:

- Assess injury risk in elite youth football players using ML.[21]
- Classifying elite and sub-elite goalkeepers in professional men's football.[17]
- Effective injury forecasting in soccer with GPS training data and ML.[32]
- Predicting the stock price (e.g., of a Club) for the next month based on this month's financial reporting data.[29]

Despite all supervised learning problems being encapsulated by the overarching description of "predicting labels given input features," the methodology assumes diverse forms and necessitates numerous modeling decisions. These decisions hinge on considerations such as the type, size, and quantity of inputs and outputs, leading to the utilization of different models tailored for processing sequences of varying lengths and fixed-length vector representations, among other factors.

*****Table 1 here*****

Regression and classification

Perhaps the simplest supervised learning task is regression. A typical illustration of a regression problem involves predicting a player transfer market value based on various factors such as age, performance statistics, experience, etc. Goddard (2005) applied regression techniques to forecast goals scored and conceded,[33] leveraging a 25-year dataset on English league football match outcomes. The defining characteristic of a regression problem lies in the form of the target variable. When labels assume arbitrary numerical values, even within a specific interval, the problem is classified as a regression problem. The primary objective is to develop a model that produces predictions closely aligned with the actual numerical label values.

In contrast to regression, the output of a classifier takes only a finite number of values. In classification tasks, the model predicts the category (often termed a class) to which a given example belongs from a discrete set of options. For instance, automatic activity classification in sports, like jumping or running. The most basic form of classification is binary classification, where the scenario involves only two classes. While regression employs a regressor to output a numerical value, classification seeks a classifier whose output predicts the assigned class. Despite classification and regression being distinct problems, analogous models are employed to address both sets of challenges. In classification, classes are distinguished using a decision boundary, whereas in regression, efforts are directed towards minimizing the difference between training samples and the values predicted by the boundary.

Decision tree

Decision trees stand out as a widely adopted ML technique employed to establish connections between input variables, depicted within the branches and nodes of the tree, and an output value encapsulated in the leaves of the tree. The decision tree is one of the oldest and most popular techniques for supervised learning, which has been developed independently in the statistical[34] and ML[35] communities. These trees find applications in both classification problems, where they produce a category label, and regression problems, where they yield a real number as output. Various algorithms, including the well-established CART (Classification and Regression Tree, which produces only binary Trees) or ID3 (Iterative Dichotomiser 3, which produces decision trees with nodes having more than two children), are employed for fitting decision trees, employing a combination of greedy searching and pruning strategies to ensure the tree effectively fits the training data while also generalizing well to unseen input/output pairs.

A notable advantage of decision trees lies in their scalability with additional data, resilience to irrelevant features, and interpretability. The choices made at each node facilitate an understanding of the impact of each predictor variable on the ultimate outcome. Random forests operate by constructing a multitude of decision trees during training, utilizing different subsets of the dataset as the training set for each tree.[36] In classification scenarios, the final output is determined by the mode of the outputs of each decision tree, while for regression problems, the mean is computed. This approach yields a model with significantly enhanced performance compared to a single decision tree, attributed to reduced overfitting. Nevertheless, the interpretability of the model diminishes, as the decisions at the nodes of the individual trees differ.

Support vector machines (SVMs)

Support vector machines (SVMs) are ML models for classification and regression tasks [37]. In SVM models, the training data is represented as points in space, aiming to delineate distinct categories by a hyperplane (a crucial deciding boundary that partitions the input space into two or more sections) situated as far as possible from the nearest data points. New input instances are subjected to the same mapping as the training data, enabling their categorization based on their position relative to the hyperplane. In instances where the data lacks linear separability, the kernel trick is used. This is a technique employed in SVMs to transform data that is not linearly separable into a higher-dimensional feature space, where it can potentially be separated linearly.[38] Extending beyond classification, SVMs can effectively address regression problems by relying on a subset of the training data to formulate regression predictions and is commonly known as support vector regression. Advantages of using SVMs include that they are effective in high dimensional spaces, that they are memory efficient thanks to the use of a subset of training points in the decision function, and finally that they are versatile through the use of different possible kernel functions. On the other hand, using SVMs can have some disadvantages: they do not directly provide probability estimates for classification problems, and correctly optimizing the kernel function and regularization term is essential to avoid overfitting.

Neural networks

Neural networks, also known as artificial neural networks, are systems based on a collection of nodes (neurons) designed to algorithmically emulate the interconnections between neurons in the human brain.[39] Each neuron can receive signals from other neurons and transmit them to additional neurons, establishing a network of interconnections. The relationship between two neurons is facilitated by an edge or *arrow* (which, represents the weights and biases of linear transformations between the layers), characterized by a weight that signifies the significance of the input from one neuron to the output of the other. Typically, a neural networks comprises an input layer, featuring one neuron per input variable for the model, an output layer with a single neuron providing the classification or regression outcome, and several hidden layers positioned between the input and output layers, each containing a variable number of neurons. An example of the use of neural networks in team sports can be found here, Ruddy et al., developed predictive modelling of hamstring strain injuries in elite Australian footballers.[40]

The advantages of using neural networks as classification or regression models are that they usually achieve higher predictive accuracy than other techniques. However, their effectiveness is contingent upon a substantial volume of training data to optimize the model. Furthermore, neural networks lack a guarantee of convergence to a singular solution, rendering them non-deterministic. Importantly, neural networks lack interpretability due to the complexity introduced by numerous layers and neurons, making it challenging to discern the direction and magnitude of the association between each input variable and the output variable through the different weights.

Unsupervised learning

The previous sections focused on supervised learning, where a large dataset containing both features and corresponding label values is provided to the model. In this scenario, the supervised learner operates under the guidance of a highly specialized supervisor. In contrast, envisioning the opposite scenario involves working for a supervisor with ambiguous expectations. In this context, the supervisor might furnish a vast dataset and instruct the data scientist to perform some ML algorithms without providing specific guidance. This ambiguity characterizes a class of problems known as unsupervised learning, wherein the range of questions one can pose is limited only by one's creativity. One common question addressed is to find a small number of prototypes that accurately summarize the data (e.g., given a set of players' characteristics, we can group them into categories). This action is typically known as clustering. Another important and exciting recent development in unsupervised learning is the advent of deep generative models. These models aim to estimate the data density, either through explicit or implicit methods.[41,42]

Clustering

Cluster analysis (predictive or descriptive) is an approach that organizes data objects based solely on information inherent in the data describing these objects and their interrelationships.[43] The primary objective is to assemble objects within a group that exhibit similarity or relatedness while maintaining dissimilarity or unrelatedness to objects in other groups. The efficacy of clustering is contingent upon achieving homogeneity within a group and maximizing dissimilarity between groups, thereby enhancing the distinctiveness of the clustering outcomes. Cluster analysis shares commonalities with other techniques employed for partitioning data objects into groups. It can be perceived as a variant of classification, as it involves labeling objects with class (cluster) labels derived exclusively from the data. In contrast, classification is a supervised process, where new, unlabeled objects receive class labels using a model developed from objects with known class labels. Consequently, cluster analysis is considered a form of unsupervised classification. In ML, the unqualified term "classification" typically refers to the supervised classification discussed in previous sections.

There are many types of clustering techniques, but the most common approach is known as K-means. K-means is a prototype-based, partitional clustering technique striving to identify a user-specified number of clusters (K) represented by their centroids. Agglomerative Hierarchical Clustering encompasses a group of closely related techniques that yield a hierarchical clustering. It initiates by treating each point as a singleton cluster and iteratively merges the two closest clusters until a single, overarching cluster remains. Some of these techniques have a natural interpretation in terms of graph-based clustering, while others have an interpretation in terms of a prototype-based approach.

Reinforcement Learning

In methodologies of learning discussed in previous sections, predictions were made on models trained with data from a similar distribution, leading to prediction failures when the system underwent significant changes compared to its training state.[12] In such dynamic situations, we could develop an agent that interacts with an environment and takes actions, then our learning paradigm is known as reinforcement learning. This approach finds applications in diverse domains such as evaluating players performance and training,[44] including robotics, and the development of AI for video. In the recent past, deep reinforcement learning, which applies deep learning to reinforcement learning problems, has surged in popularity. Although not related to football but sport in general, notable works include the groundbreaking deep Q-network, which outperformed humans in Atari games using only visual input,[45] and the

AlphaGo program, which triumphed over the world champion in the board game Go.[46]

Reinforcement learning gives a very general statement of a problem in which an agent interacts with an environment over a series of time steps. At each time step, the agent receives some observation from the environment and must choose an action that is subsequently transmitted back to the environment via some mechanism. After each iteration, the agent receives a reward from the environment. The agent then receives a subsequent observation, and chooses a subsequent action, and so on. The behavior of a reinforcement learning agent is governed by a policy. In brief, a policy is just a function that maps from observations of the environment to actions. The goal of reinforcement learning is to produce good policies.

Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis

Earlier sections have explored the integration of ML in football, detailing and clarifying the distinct features of ML learning strategies, including supervised, unsupervised, and reinforcement learning. The forthcoming section introduces a SWOT analysis, proposing several considerations for the implementation of ML tactics by football's medical and sports science departments. It specifically outlines four strategic aspects: 1) use strengths to create opportunities and make the most of them, 2) use strengths to avoid threats, 3) work on weaknesses to take advantage of opportunities, 4) upgrade weaknesses to avoid threats. The SWOT analysis process is a valuable tool for organizations and businesses (i.e., clubs) to assess their internal and external environment. Table 2 reports some key needs for conducting a SWOT analysis.

***** Table 2 here*****

Practical tips to run a SWOT analysis in football aiming to apply ML

Before applying ML in the team, medical and sport science staff are advised to build a strategic management plan. As part of this plan, they should perform an environmental analysis, which includes scanning the internal and external factors.[47] The internal factors include analyzing the strengths and weaknesses of their team/organization.[47] The external factors analysis includes the factors outside the team/organization, the opportunities, and threats of using ML. This is called SWOT analysis and is being used in other domains too.[47] An example of a SWOT analysis for a top-level football club is presented in Figure 1. We have assumed that the club's top management has adopted ML to improve their senior squad's injury risk assessment strategy. This new approach may bring value provided the team is ready to take advantage of that opportunity (see, Figure 1).

*****Please, add here Figure 1*****

With regards to the SWOT analysis presented above, we are suggesting some actions to be considered by the medical and sport science staff working in the club. In particular and with regards to the:

- Strategic dimension 1. Use strengths to take advantage of opportunities: the supporting team staff can work with top management to convince the coaches of the competitive advantages this new approach may bring to the team. The highly skilled ML staff can work effectively on optimizing systems and building algorithms for injury risk assessment.[11]
- Strategic dimension 2. Use strength to avoid threats: the supports team staff may work on knowledge transfer to the coaches. Simultaneously, the support team staff should receive further education on technical and tactical aspects of football to better understand the game. This will help in accounting for the context when analyzing big data. In turn, this will facilitate the communication of the support team staff with the coaches and optimize knowledge implementation.[48]
- Strategic dimension 3. Upgrade weaknesses to take advantage of opportunities: implement a holistic player-centric monitoring system and consider the complexity of injury occurrence.[8,49] This will help in better interpreting the algorithms.[50]
- Strategic dimension 4. Update weaknesses to avoid threats: optimize player's monitoring and integration of ML tools with the existing systems and workflows, while working on knowledge transfer to the coaches.[50] Build a "bright spot" that will add a competitive advantage to the team.

Limitations and future directions

The implementation of ML is not without limitations or barriers. First, ML models require large amounts of high-quality data for training. In football, obtaining comprehensive and accurate data can be challenging due to variations in data collection methods, inconsistencies, and missing information. For instance, limited historical data for specific events (e.g., injuries, specific player movements) can hinder model performance. Second, ML techniques are not guaranteed to provide correct information (e.g., poor model performance, incorrect prediction and therefore, do not always enhance decision-making). Third, many ML algorithms operate as black boxes (if practitioners do not have a specific background in ML), making it difficult to understand how they arrive at specific decisions. In football, coaches and analysts need interpretable models to make informed decisions. Fourth, creating relevant features (input variables) for football-specific tasks can be complex. Deciding which player attributes, team statistics, or match context to include requires domain knowledge. Moreover, football events (e.g., goals, fouls, yellow cards) occur infrequently compared to non-events (e.g., passes, ball possession). This class imbalance affects model training and evaluation. Therefore, techniques like oversampling, undersampling, or using weighted loss functions are necessary to address this issue. Finally, football is highly context-dependent. Player actions depend on the game situation, opponent, field position, and time remaining. ML models must account for these dynamic factors.

Practical applications

ML models can analyze player data (such as physical condition, physiological parameters, match performance, and training load) to assess the likelihood of injuries. Clubs can use this information to manage player load, optimize recovery, and reduce injury risks. ML algorithms can assess player form by analyzing historical performance data. Clubs can identify players who are in peak form and make informed decisions about team selection. For scouting, ML can analyze player statistics, playing style, and potential fit with the team's tactics. It helps clubs discover talented players and make strategic signings. ML techniques can analyze opponents' playing styles, strengths, and weaknesses. Clubs can use this information to tailor their game plans, identify vulnerabilities, and exploit opponent weaknesses during matches. ML algorithms can evaluate youth players' performance metrics and potential. Clubs can identify promising talents early, nurture their development, and integrate them into the senior team. Finally, clubs that want to build a strategic management plan can use the four dimensions presented in our SWOT analysis such as the use of strengths to create opportunities and make the most of them, the use of strengths to avoid threats, work on weaknesses to take advantage of opportunities, and upgrade weaknesses to avoid threats.

Conclusion

This education review provides practitioners with a concise overview of the characteristics of

ML analysis and a guide for how to conduct a SWOT analysis regarding the implementation of ML techniques in professional football clubs. This review explains the difference between AI and ML, and the difference between ML and statistical analysis. Furthermore, we explained the characteristics of ML approaches such as supervised learning, unsupervised learning, and reinforcement learning. Finally, we presented an example of a SWOT analysis, which suggested some actions to consider when ML is implemented by medical and sport science staff in football. In conclusion, ML analysis can be an invaluable ally of football clubs and sport science and medical departments due to its ability to analyze vast amounts of data and extract meaningful insights.

References

1. Jeffries, A. C. *et al.* Development of a revised conceptual framework of physical training for use in research and practice. *Sport. Med.* **52**, 709–724 (2022).
2. Dello Iacono, A., Beato, M., Unnithan, V. B. & Shushan, T. Programming high-speed and sprint running exposure in football: beliefs and practices of more than 100 practitioners worldwide. *Int. J. Sports Physiol. Perform.* 1–16 (2023) doi:10.1123/ijspp.2023-0013.
3. Beato, M. *et al.* Rationale and practical recommendations for testing protocols in female soccer: a Narrative Review. *J. Strength Cond. Res.* (2023).
4. Scott, T. U., Scott, T. J. & Kelly, V. G. The validity and reliability of global positioning system in team sport: a brief review. *J. Strength Cond. Res.* **30**, 1470–1490 (2016).
5. Drust, B. & Green, M. Science and football: evaluating the influence of science on performance. *J. Sports Sci.* **31**, 1377–1382 (2013).
6. Rein, R. & Memmert, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springerplus* **5**, 1410 (2016).
7. Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F. & Baca, A. Machine learning application in soccer: a systematic review. *Biol. Sport* **40**, 249–263 (2023).
8. Claudino, J. G. *et al.* Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sport. Med. - Open* **5**, 28 (2019).
9. Roth, E. M. & Woods, D. D. Cognitive task analysis: an approach to knowledge acquisition for intelligent system design. in 233–264 (1989). doi:10.1016/B978-0-444-87321-7.50014-4.

10. Bhatt, G. D. & Zaveri, J. The enabling role of decision support systems in organizational learning. *Decis. Support Syst.* **32**, 297–309 (2002).
11. Nassis, G., Verhagen, E., Brito, J., Figueiredo, P. & Krustrup, P. A review of machine learning applications in soccer with an emphasis on injury risk. *Biol. Sport* **40**, 233–239 (2023).
12. Chmait, N. & Westerbeek, H. Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists. *Front. Sport. Act. Living* **3**, (2021).
13. Agarwal, S. Data mining: data mining concepts and techniques. in *2013 International Conference on Machine Intelligence and Research Advancement* 203–207 (IEEE, 2013). doi:10.1109/ICMIRA.2013.45.
14. Bongiovanni, T. *et al.* Importance of anthropometric features to predict physical performance in elite youth soccer: a machine learning approach. *Res. Sport. Med.* **29**, 213–224 (2021).
15. Campbell, P. G. *et al.* Analysing the predictive capacity and dose-response of wellness in load monitoring. *J. Sports Sci.* **39**, 1339–1347 (2021).
16. Szczepański, T. & McHale, I. Beyond completion rate: evaluating the passing ability of footballers. *J. R. Stat. Soc.* **179**, 513–533 (2016).
17. Jamil, M. *et al.* Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men’s football. *Sci. Rep.* **11**, 22703 (2021).
18. Barron, D., Ball, G., Robins, M. & Sunderland, C. Artificial neural networks and player recruitment in professional soccer. *PLoS One* **13**, e0205818 (2018).
19. García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A. & Luengo-Sánchez, S. In-game behaviour analysis of football players using machine learning techniques based on player statistics. *Int. J. Sports Sci. Coach.* **16**, 148–157 (2021).
20. Oliver, J. L. *et al.* Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *J. Sci. Med. Sport* **23**, 1044–1048 (2020).
21. Rommers, N. *et al.* A machine learning approach to assess injury risk in elite youth football players. *Med. Sci. Sports Exerc.* **52**, 1745–1751 (2020).
22. Ruddy, J. D. *et al.* Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Front. Physiol.* **10**, (2019).
23. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
24. Akyildiz, Z. *et al.* Monitoring the post-match neuromuscular fatigue of young Turkish

football players. *Sci. Rep.* **12**, 13835 (2022).
25. Houtmeyers, K. C., Jaspers, A. & Figueiredo, P. Managing the training process in elite sports: from descriptive to prescriptive data analytics. *Int. J. Sports Physiol. Perform.* **16**, 1719–1723 (2021).
26. Bzdok, D., Altman, N. & Krzywinski, M. Points of Significance: Statistics versus machine learning. *Nat. Methods* **15**, 233–234 (2018).
27. Richter, C., O'Reilly, M. & Delahunt, E. Machine learning in sports science: challenges and opportunities. *Sport. Biomech.* 1–7 (2021) doi:10.1080/14763141.2021.1910334.
28. Tian, T., Song, C., Ting, J. & Huang, H. A French-to-English machine translation model using transformer network. *Procedia Comput. Sci.* **199**, 1438–1443 (2022).
29. Bhandari, H. N. *et al.* Predicting stock market index using LSTM. *Mach. Learn. with Appl.* **9**, 100320 (2022).
30. Ettensperger, F. Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. *Qual. Quant.* **54**, 567–601 (2020).
31. Anaby-Tavor, A. *et al.* Do not have enough data? Deep learning to the rescue! *Proc. AAAI Conf. Artif. Intell.* **34**, 7383–7390 (2020).
32. Rossi, A. *et al.* Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One* **13**, e0201264 (2018).
33. Goddard, J. Regression models for forecasting goals and match results in association football. *Int. J. Forecast.* **21**, 331–340 (2005).
34. Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* **29**, 119 (1980).
35. Hunt, E. B., Marin, J. & Stone, P. J. . *Experiments in induction*. *Academic Press* (1966).
36. Gomes, H. M. *et al.* Adaptive random forests for evolving data stream classification. *Mach. Learn.* **106**, 1469–1495 (2017).
37. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
38. Murty, M. N., Raghava, R., Murty, M. N. & Raghava, R. Kernel-based SVM. Support vector machines and perceptrons: Learning, optimization, classification, and application to social networks. in 57–67 (2016).
39. Bishop, C. M. *Neural networks for pattern recognition*. (Oxford University PressOxford, 1995). doi:10.1093/oso/9780198538493.001.0001.

40. Ruddy, J. *et al.* Predictive modeling of hamstring strain injuries in elite australian footballers. *Med. Sci. Sport. Exerc.* **50**, 906–914 (2018).

41. Chen, Y. *et al.* Auto-encoding variational Bayes. *Cambridge Explor. Arts Sci.* **2**, (2024).

42. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).

43. Karim, M. R. *et al.* Deep learning-based clustering approaches for bioinformatics. *Brief. Bioinform.* **22**, 393–415 (2021).

44. Xu, Q. & He, X. Football training evaluation using machine learning and decision support system. *Soft Comput.* **26**, 10939–10946 (2022).

45. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

46. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

47. Wheelen, T., Hoffman, J. & Bamford, C. Strategic management and business policy globalization. in *Pearson* (2018).

48. Nassis, G. P. Leadership in science and medicine: can you see the gap? *Sci. Med. Footb.* **1**, 195–196 (2017).

49. Gabbett, H. T., Windt, J. & Gabbett, T. J. Cost-benefit analysis underlies training decisions in elite sport. *Br. J. Sports Med.* **50**, 1291–1292 (2016).

50. Edouard, P., Verhagen, E. & Navarro, L. Machine learning analyses can be of interest to estimate the risk of injury in sports injury and rehabilitation. *Ann. Phys. Rehabil. Med.* **65**, 101431 (2022).

Authors contribution

All authors contributed to the writing of the paper. All authors read and approved the final version.

Conflict of interest

The authors declare no conflict of interest for this paper.

Data availability statement

This manuscript does not have associated data.

Table 1. Supervised and unsupervised machine learning analysis.

Regression	Supervised ML regression is a type of predictive analysis that is used to model and analyze relationships between variables. It aims to predict a continuous target variable based on one or more independent variables. The goal is to find the best fit line or curve that minimizes the difference between predicted and actual values. This is achieved through algorithms that adjust the weights of input features to reduce error in predictions. Regression techniques are widely used in fields such as finance, medicine, and environmental science for tasks like prediction market value and estimating injury risk.	Examples of regressions analysis: *Boosting, Decision Tree, K Nearest Neighbor, Neural Network, Random Forest, Regularized linear, Support Vector Regression*
Classification	Supervised ML classification is a type of algorithm used to assign predefined labels to new data points. It works by learning from a dataset with known labels and then applying this knowledge to categorize new, unlabeled data. Common applications include sport movements analysis and medical diagnosis, where the algorithm must decide which category or class the new data belongs to based on its features.	Examples of classifications analysis: *Boosting, Decision Tree, K Nearest Neighbor, linear discriminant, Neural Network (includes Deep CNNs), Random Forest, Support Vector Machine.*
Clustering	Clustering in ML is an unsupervised learning technique used to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). It is commonly used in statistical data analysis for pattern recognition, game-tactical analysis, information retrieval, and bioinformatics. Algorithms like K-Means, Hierarchical clustering, and DBSCAN are popular methods for performing clustering tasks. The goal is to discover the inherent structure within the data, often to identify distinct subgroups without pre-labeled data or human supervision.	Examples of clustering analysis: *Density-based, Fuzzy C-Means, Hierarchical, Neighborhood-based.*

Table 2. This table reports the general characteristics of a SWOT analysis. The SWOT analysis process serves as a compass, guiding organizations toward effective strategies, risk management, and sustainable growth of a business (club).

Strategic planning	SWOT analysis helps organizations develop effective strategies by identifying their strengths, weaknesses, opportunities, and threats. It provides a comprehensive view of the current situation, enabling informed decision-making.
Self-reflection and awareness	Organizations need to understand their internal capabilities (strengths and weaknesses) and external factors (opportunities and threats). SWOT analysis encourages self-reflection and awareness, leading to better alignment with organizational goals.
Risk assessment	By evaluating potential threats (such as market changes, competition, or regulatory issues), organizations can proactively address risks. SWOT analysis allows them to prioritize risk mitigation strategies.
Resource allocation	SWOT analysis guides resource allocation. Organizations can allocate resources more effectively by capitalizing on strengths and minimizing weaknesses. It helps prioritize investments and efforts
Competitive advantage	Identifying unique strengths and opportunities allows organizations to create a competitive edge. Leveraging these advantages helps them stand out in the market.
Adaptation to change	The business landscape constantly evolves. SWOT analysis enables organizations to adapt to changes by recognizing emerging opportunities and addressing potential threats promptly
Communication and alignment	SWOT analysis fosters communication among team members, stakeholders, and leadership. It aligns everyone around a common understanding of the organization's position and future direction.

Figure 1: An example of SWOT analysis regarding the use of ML for injury risk assessment for a football team