Short Survey in Machine Learning for Soccer Analytics

Pedro Amadu *

Posted Date: 2 October 2024

doi: 10.20944/preprints202410.0178.v1

Keywords: supervised learning; unsupervised learning; reinforcement learning; principal component analysi; soccer analytics

Article

Short Survey in Machine Learning for Soccer Analytics

Pedro Amadu

Independent Researcher; pedroamadu@hotmail.com

Abstract: We investigate soccer analytics from supervised learning, unsupervised learning, and reinforcement learning perspectives. With the increasing availability of player tracking data and event logs, machine learning techniques have become essential for uncovering patterns in player and team performance. In this paper, we examine how supervised learning models are applied to tasks such as match outcome prediction and player rating systems, while unsupervised learning is utilized for player clustering, tactical analysis, and the discovery of hidden patterns in game data. Reinforcement learning, on the other hand, plays a key role in optimizing decision-making during matches by learning optimal strategies and tactics through trial and error. By providing a comprehensive overview of these approaches, we aim to highlight the transformative potential of machine learning in modern soccer analytics and how it continues to shape the sport. We also provide summary of other soccer analytics research in this work.

Keywords: supervised learning; unsupervised learning; reinforcement learning; principal component analysi; soccer analytics

1. Introduction

Modern soccer, sometimes referred to as football, requires advanced analytics. The sport has evolved far beyond traditional statistics such as goals scored, shots, or possession percentages. In today's game, data-driven insights are indispensable for teams, coaches, and analysts to optimize performance, improve tactical decision-making, and gain a competitive edge. With the rise of tracking technologies and the availability of large datasets capturing every movement of players and the ball, soccer analytics has shifted from merely descriptive metrics to complex models that evaluate player actions, predict outcomes, and simulate match strategies. Leveraging machine learning and artificial intelligence (AI) has enabled deeper insights into player behaviors, tactical formations, and team dynamics, revolutionizing how matches are analyzed and strategies are developed.

The integration of analytics into soccer has had a profound impact at both the professional and grassroots levels, influencing decisions on player recruitment, in-game tactics, injury prevention, and fan engagement. Clubs now use data to identify undervalued players, predict injury risks, and tailor training regimens. Machine learning, in particular, has opened new frontiers in predictive modeling, including the development of models that assess the probability of scoring, analyze pressing patterns, and even predict match outcomes. The application of supervised, unsupervised, and reinforcement learning has allowed teams to uncover hidden patterns in player performance and team strategy, offering unprecedented insights into the world's most popular sport."

This extended section can serve as a strong foundation for explaining the growing role of machine learning and data analytics in soccer. We are going to look at this topic from supervised learning, unsupervised learning and reinforcement learning perspective.

2. Literature Review

Soccer analytics, as an academic and professional field, has undergone significant evolution since its early days. Initially, the analysis of soccer data revolved around basic metrics like goals, assists, and possession statistics. These rudimentary forms of analytics were largely descriptive, used primarily by commentators and fans to understand general trends in match performance. However, with the advent of more advanced technologies and data collection methods, soccer analytics has grown into a highly sophisticated field that informs decisions at every level of the sport, from tactical adjustments during matches to long-term player recruitment strategies.

The early history of soccer analytics can be traced back to the mid-20th century, when Charles Reep, an English accountant, is credited as one of the pioneers in the field. Reep manually collected data on pass sequences and shot attempts, leading to his controversial conclusion that long-ball tactics were more effective in generating goals than short-passing strategies [1]. His work laid the foundation for using data to support tactical decisions, though his findings were criticized for oversimplifying the complexity of the game.

A major milestone in soccer analytics came with the rise of expected goals (xG) models. Expected goals quantify the quality of a shot by calculating the probability that the shot will result in a goal, based on factors like distance from the goal, angle of the shot, and the position of defenders. This metric was a significant improvement over raw shot counts, providing a more nuanced understanding of shot quality. xG models were popularized in the early 2000s by various sports analysts and eventually became a staple of modern soccer analytics [2]. xG is now widely used by clubs, analysts, and fans to evaluate team performance more accurately.

The introduction of tracking data revolutionized soccer analytics further. In the early 2010s, advancements in camera and sensor technologies allowed for the collection of player and ball movement data at granular levels of detail. Companies like Opta, StatsBomb, and STATS developed tools that track the movements of players and the ball multiple times per second throughout a match, providing new opportunities for tactical analysis [3]. This data enabled the development of new metrics, such as player heat maps, which show the areas of the field where a player is most active, and pass networks, which illustrate the flow of passes between players. These developments helped teams optimize their tactical approaches by understanding how their players interact with one another and move around the pitch.

The introduction of machine learning models marked another leap in soccer analytics. Data analysts began applying techniques from fields like computer science and artificial intelligence to develop predictive models that could forecast match outcomes or player performance. For example, machine learning algorithms have been used to predict the likelihood of injuries based on player workload and match intensity [4]. Teams now use such models to manage player rotation and prevent injuries, enhancing player longevity and overall team performance. [5] emphasizes the calibration on starting lineup formations with advanced graph neural network method.

In parallel to these advances in data collection and predictive modeling, the rise of event data also played a crucial role in shaping modern soccer analytics. Event data focuses on individual actions in a match, such as passes, tackles, and shots, providing a deeper layer of analysis that moves beyond simple counting stats. Event-based analytics help analysts quantify player contributions that may not result in goals or assists but are crucial to team success. For example, on-ball value (OBV) metrics assign a value to each action a player performs during a match, allowing clubs to assess player contributions in a more detailed and comprehensive way [6].

Today, soccer analytics is an integral part of the sport, with many clubs employing entire departments dedicated to data science. Major teams like Liverpool FC, Manchester City, and FC Barcelona are known for their sophisticated analytics departments, which use data to inform everything from recruitment strategies to in-game tactical decisions [7]. The shift from basic statistics to advanced machine learning-driven models represents a broader trend toward data-driven decision-making in soccer, reflecting similar transformations occurring in other sports like basketball and baseball [8].

The historical evolution of soccer analytics can be viewed as a progression from basic descriptive statistics to sophisticated models powered by advanced data collection and machine learning techniques. This transformation has fundamentally changed how teams, analysts, and fans understand and engage with the sport. As data availability and computational power continue to increase, soccer analytics will likely play an even more prominent role in shaping the future of the sport.

3. Method on Improving Soccer Analytics Using Machine Learning Techniques

Supervised learning, unsupervised learning, and reinforcement learning form the foundational paradigms of machine learning, each providing unique advantages when applied to soccer analytics. Their applicability ranges from predicting match outcomes to uncovering hidden patterns in player performance and even guiding real-time decision-making during matches. Below is an elaboration of these three learning paradigms, with examples from soccer analytics and relevant literature.

3.1. Supervised Learning in Soccer Analytics

Supervised learning refers to training machine learning models on labeled datasets, where the input data is paired with corresponding outputs. In soccer analytics, this paradigm is widely used for predictive modeling tasks, such as predicting match outcomes, player performance, or injury risks. A typical supervised learning task in soccer could involve predicting whether a team will win, lose, or draw based on historical match data (such as goals, possession, shots on target). Algorithms commonly employed for such tasks include Logistic Regression, Random Forests, Support Vector Machines (SVM), and Neural Networks. For example, [9] demonstrated the application of Random Forest and Gradient Boosting Machines in predicting the outcomes of soccer matches using historical match data from the English Premier League.

Supervised learning is also used for player performance prediction. By leveraging historical player statistics, models can forecast future performance, enabling clubs to make data-driven decisions about player selection and transfers. For instance, authors like [10] applied machine learning models to predict player performance metrics such as passing accuracy and goal-scoring potential based on match data. In injury prediction, supervised learning models use physiological data, such as player workload and movement patterns, to predict injury risks. This allows teams to manage player fitness and reduce the likelihood of injury.

3.2. Unsupervised Learning in Soccer Analytics

Unsupervised learning involves using algorithms to detect patterns in data without predefined labels. This paradigm is especially useful in clustering and pattern recognition tasks in soccer, where the aim is to identify underlying structures in the data that may not be immediately apparent. For example, clustering algorithms like k-means and hierarchical clustering can be employed to group players based on their playing styles or tactical roles. This can provide coaches and analysts with deeper insights into player performance that go beyond simple statistics. One such application is clustering players based on their movement data from tracking systems to identify which players are more aggressive in their positioning or which play a more defensive role.

Another common application is Principal Component Analysis (PCA), which helps reduce the dimensionality of soccer data while retaining the most critical features. This technique is often used to simplify complex datasets like tracking data, where a large number of variables (e.g., player positions at every second) can make analysis computationally expensive. By applying unsupervised learning, analysts can focus on the most important aspects of the game, such as key passes, sprints, or defensive movements that influence match outcomes.

A relevant case study on unsupervised learning in soccer analytics can be found in [6], who applied clustering techniques to player movement data to identify unique tactical formations and playing styles. The insights derived from such analyses are invaluable for opponents preparing their strategies or for teams looking to refine their tactical approach.

PCA works by transforming the original set of variables (such as player positions) into a smaller set of principal components, which are linear combinations of the original variables. These principal components capture the maximum variance in the data, meaning that the most critical patterns are retained while the less significant variations are discarded. Essentially, PCA identifies the underlying structure in the data and focuses only on the most important aspects, reducing the number of dimensions without losing essential information.

For example, in a soccer match, PCA might help reduce the complexity of tracking data by identifying the primary movements that impact the game—such as key passes, sprints, or defensive positioning—while ignoring redundant or less important movements. By doing so, PCA allows analysts to concentrate on the factors that are most likely to influence match outcomes, such as critical passes that lead to goal-scoring opportunities, player sprints during counter-attacks, or defensive movements that block opponents' progress.

This dimensionality reduction not only simplifies the analysis process but also makes models more interpretable and efficient. For instance, after applying PCA, analysts may find that just a few key components—such as changes in player positions during attacks or defensive formations—account for the majority of important tactical insights. This streamlined dataset can then be used for further analysis or modeling, such as predicting match outcomes or evaluating team performance. Therefore, PCA enables teams to extract valuable insights from large, complex datasets without being overwhelmed by the sheer volume of information.

PCA simplifies the analysis of high-dimensional soccer data by reducing it to its most significant components, allowing analysts to focus on the critical actions that influence the game, such as key passes or defensive movements. This makes the analysis more computationally efficient and easier to interpret, leading to better decision-making in both tactical and strategic planning.

3.3. Reinforcement Learning in Soccer Analytics

Reinforcement learning (RL) is a learning paradigm that revolves around decision-making in dynamic environments, where an agent learns to make decisions by interacting with the environment and receiving feedback in the form of rewards or penalties. In soccer analytics, RL is particularly suitable for optimizing decision-making processes, such as determining optimal player positioning during matches or formulating in-game strategies. Unlike supervised learning, which requires labeled data, reinforcement learning focuses on learning optimal policies through trial and error, making it well-suited for tasks where the outcome is uncertain, and decisions need to adapt in real-time.

One practical application of RL in soccer is in robotic soccer simulations, where agents (representing players) learn to cooperate, pass, and score goals in a simulated environment. These simulations have contributed to the development of intelligent systems capable of mimicking human decision-making during soccer matches. RL has also been used to model game strategies, where algorithms learn the most effective tactics based on historical match data. An example is the work by [11], who utilized reinforcement learning to optimize soccer match strategies by learning from historical game outcomes.

Beyond robotic simulations, reinforcement learning can be applied to tactical decision-making in real games. Coaches can use RL-based systems to suggest optimal substitutions or formation changes during a match based on the current state of play. By continually updating its knowledge of the game, a reinforcement learning system can adjust its recommendations in real-time, offering a competitive edge to teams that integrate this technology into their game strategy.

In soccer analytics, RL is particularly useful for tasks such as optimizing tactical strategies, predicting player movements, and enhancing team formations. For instance, RL can be applied to model the decisions a coach or player might make during a match, such as when to make a substitution, change formation, or initiate a counter-attack. The model learns by trial and error, exploring different strategies and gradually improving based on the reward structure, such as scoring a goal or preventing an opponent from scoring.

Markov Decision Processes (MDPs)

MDPs form the mathematical framework behind many reinforcement learning models. In soccer, an MDP is a way of representing the game as a series of states (such as player positions, possession of the ball, and score) and actions (such as passing, shooting, or tackling). The agent (a player or team) takes an action in each state, and the environment transitions to a new state based on that action, with rewards or penalties given depending on the outcome. In soccer, the reward might be a goal or successful possession retention, while penalties might occur for losing possession or conceding a goal. MDPs are widely used in soccer simulations to model game dynamics and optimize in-game decision-making.

Q-Learning

Q-Learning is a model-free RL algorithm that seeks to learn the value of actions in different states by updating a Q-value, which represents the expected future reward for taking a specific action in a given state. This is especially useful in soccer for modeling tactical decision-making, where players or a team must decide the best course of action at each step of the game. For instance, Q-Learning could be applied to model optimal passing strategies, where the algorithm learns the best pass to make based on the position of the teammates and opponents, the current score, and the time left in the match. Over time, the Q-values for different passes are updated to reflect their success or failure, enabling the model to recommend optimal passing decisions.

Deep Q-Networks (DQN)

A DQN extends Q-Learning by using deep neural networks to approximate the Q-values, making it scalable to more complex environments, such as those in soccer. In soccer, where the state space (combinations of player positions, ball movements, etc.) is vast, traditional Q-Learning may struggle. DQNs address this by using neural networks to handle the complexity, allowing for more sophisticated strategy development. For example, DQNs could be used to optimize a team's formation during different phases of play—defense, midfield transitions, or attacking sequences—by learning the most effective positioning and movements to maximize the chances of scoring or preventing goals.

Policy Gradient Methods

Policy gradient methods are a class of RL algorithms where the agent directly learns the policy (i.e., the probability distribution over actions) rather than learning Q-values. In soccer, this could be useful for modeling continuous actions, such as the exact angle or force to apply when taking a shot. Policy gradient methods are often used in robotic soccer simulations, where players (robots) need to learn how to execute precise movements, such as dribbling or shooting, by receiving continuous feedback from the environment. These methods are also applied in tactical decision-making, where the model learns to select actions (e.g., pressing or counter-attacking) that maximize long-term rewards like winning games.

Proximal Policy Optimization (PPO)

PPO is a popular policy gradient method known for its stability and efficiency. It's often used in complex decision-making environments like soccer, where agents need to balance exploration (trying new strategies) and exploitation (using known strategies that work). PPO has been applied in robotic soccer and simulated environments where agents must learn optimal in-game strategies, such as player positioning during attacks or defense. PPO is particularly useful in continuous control tasks, such as deciding when to switch from defense to attack based on game dynamics.

Monte Carlo Tree Search (MCTS)

MCTS is a search algorithm that has been used in combination with RL, particularly in games like chess and Go, and can also be applied to soccer. MCTS explores possible future states by simulating different sequences of actions, which is useful for strategic planning in soccer matches. For example, MCTS could be applied to explore different passing sequences or tactical formations during key moments of a match, such as deciding whether to press high or sit back in defense when the opposing team is in possession.

doi:10.20944/preprints202410.0178.v1

Applications of RL in Soccer Analytics

- Tactical Decision-Making: RL models can optimize in-game decisions such as player positioning, pressing strategies, and substitutions. For example, RL can help coaches identify the best time to substitute a player or switch formations based on real-time match conditions [11].
- Player Movement and Behavior: By learning from historical tracking data, RL models can predict and optimize player movements, suggesting the best positioning for players during transitions between attack and defense.
- Simulated Soccer Games: RL has been widely used in robotic soccer, where agents (representing players) learn to cooperate, pass, and score goals. These simulations help develop algorithms for autonomous decision-making in soccer.
- Opponent Strategy Modeling: RL can be used to model opponent behavior and adapt strategies accordingly. For instance, teams can use RL to predict how an opposing team might react to specific actions, such as a high press or counter-attack, and adjust their strategy in real time.

Reinforcement learning in soccer analytics opens up a wide array of possibilities for optimizing tactical decision-making, player positioning, and strategy development. By applying models like MDPs, Q-Learning, DQNs, and policy gradient methods, analysts can develop systems that learn through interaction with the game environment, making them well-suited for real-time, dynamic decision-making tasks. As RL continues to evolve, its application in soccer will likely deepen, providing teams with increasingly sophisticated tools for optimizing performance both on and off the pitch.

4. Conclusions

Supervised, unsupervised, and reinforcement learning offer distinct but complementary approaches to solving complex problems in soccer analytics. Supervised learning enables predictive modeling of match outcomes, player performance, and injuries, while unsupervised learning uncovers hidden structures and clusters in data, offering deeper tactical insights. Reinforcement learning, on the other hand, excels in optimizing real-time decision-making, paving the way for intelligent soccer strategies. As machine learning continues to evolve, these paradigms will likely play an increasingly central role in soccer analytics, driving the sport toward more data-driven and efficient practices.

References

1. Richard Pollard. Charles reep (1904-2002): Pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, 20(10):853–855, 2002.
2. Alexander Rathke. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(2):514–529, 2017.
3. László Gyarmati, Haewoon Kwak, and Pablo Rodriguez. Qpass: A merit-based evaluation of soccer passes. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1906–1915, 2014.
4. Alessandro Rossi, Luca Pappalardo, Paolo Cintia, F Marcello Iaia, Javier Fernandez, and David Medina. Machine learning in player workload monitoring and injury risk reduction. *Proceedings of the 12th MIT Sloan Sports Analytics Conference*, 2018.
5. Zeyu Wang, Yue Zhu, Zichao Li, Zhuoyue Wang, Hao Qin, and Xinqi Liu. Graph neural network recommendation system for football formation. *Applied Science and Biotechnology Journal for Advanced Research*, 3(3):33–39, 2024.
6. Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861, 2019.
7. Aaron Bialkowski, Patrick Lucey, Peter Carr, Iain Matthews, and Sridha Sridharan. Large-scale analysis of soccer matches using spatiotemporal tracking data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 725–730. IEEE, 2014.

8. Patrick Lucey, Aaron Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. Representing and discovering adversarial team behaviors using player roles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2706–2713. IEEE, 2013.
9. Daniel Berrar, Pedro Lopes, and James Davis. Machine learning for predicting the outcome of professional soccer matches. *Statistical Modelling*, 19(1):55–77, 2019.
10. Jón Steinar Gudmundsson and Mark Horton. Spatio-temporal analysis of team sports–a survey. *arXiv preprint arXiv:1707.04506*, 2017.
11. Yang Liu, Shaobo Jiang, Zhaoyu Han, and Xueyao Wang. A deep reinforcement learning approach for the simulation and optimization of soccer strategies. *Applied Intelligence*, 50(10):3021–3033, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.