Dundalk Institute of Technology

RESA C9009 - Research Process for Data Analytics

Literature Review: Data-Driven Approaches to Identifying Undervalued Talent in Football

Ryan Habis

Student ID: D00245309

Submission Date: November 2024

Contents

1 Introduction

Sports analytics has greatly benefited from use of data science in fact it has revolutionized the sport industry all together. This topic was inspired from the Money-ball philosophy (Olthof & Davis, 2025) which started originally in baseball, the applications for this research concept is directed at football in order to identify hidden talent that is seen as undervalued in the sport and to collect team performance data in order to optimize overall team performance.

How to apply this strategy will involve using data driven methods to uncover problems the team might have as an example, discover the players who are contributing far more than what there market value set them at. This is possible with the advancement in statistical methods that allows coaches, analysts, and club management with very powerful ai tools for performance evaluation.

2 Methodology

Within this literature review it will be based on recent papers revolving around football analytics.

When researching this topic the method that was used to collect research papers was the college library, keywords such as football analytics where used in order to find the papers that were used within this review.

The majority of papers were relevant to the topic of football / soccer, but there were papers that where in the same criteria as football but it was statistical data revolved around American football which was reverent because its was the wrong sport altogether.

The papers that were used in this report are:

- Advancing Football Game Analysis: Integrating Computer Vision, Deep Learning, and Hybrid Techniques for Enhanced Video Analytics
- AI-Powered Football Match Analysis using YOLOv8 and Spatial Analytics
- Comparative Analysis of Machine Learning Models for Predicting Top Goal Scorer and Goalkeeper Performance in Football
- Explainable expected goal models for performance analysis in football analytics
- Characterizing the spatial structures of competing football teams
- Segmentation for Enhanced Football Analytics: A Pixel-level Approach

- Evaluating defensive strategies in football: analysing the impact of defensive metrics on match outcomes
- FPSRec: Football Players Scouting Recommendation System based on Generative AI
- Perspectives on data analytics for gaining a competitive advantage in football: computational approaches to tactics

This was the paper that was not used ”Video Preprocessing for American Football Formation Recognition”. Since it was not the same sport it was irrelevant

3 Literature Review

This section provides a detailed analysis of the research papers, examining their objectives, methodologies, findings, and contributions to the field of football analytics.

3.1 Paper 1: AI-Powered Football Match Analysis using YOLOv8 and Spatial Analytics

- **Authors:** S Pranav Arun, Mohamed Rizwan H, and M. Sindhuja
- **Publication Date:** July 9, 2025

YOLO which stand for ”You only look once” this is a powerful algothism tool that allows for deep learning techniques and works with computer vision, a combination of these two has revolutionized the sport in an analytical way. YOLO version 8 allows for real time object detection and tracking (Arun, H, & Sindhuja, 2025). One of the key application of this software is to precisely locate and track important player on the field such as the players, referees, goalkeepers and the ball isself.

When analyzing YOLO there is three main finding that stood out.
Key finding that stood out:

- **Performance and Speed**

  Has the ability to process real time data at a very fast rate it detected 27 frames per second (FPS) using google colab with GPU support.

- **Core Functionality**

Figure 1: Image detection

The precise localization of the ball in each frame, this is very important this allows us to do complex analyses such as checking the ball possession and analyzing the team strategies.

- **Integration with Other Techniques**

  YOLO version 8 detector has the ability to work with other data tracking models such as DeepSORT tracker in order to enhance the players that are being tracked on the field it allows for a more stability and helps identifying teams based on there jersey color

The application of using YOLO allows access to very valuable data for example automatic recognition of player movement, the trajectory of the ball we would know where it go, and critical events such as moments before a goal is about to be made.

Implementing YOLO to football analytics give the coaches, researchers, and sport scientists the ability to analyzes football data without having to rely on expensive multi camera setup or analyzing the game manually because it is automated.

The figure below shows the image detection program YOLO working.

3.2 Paper 2: Comparative Analysis of Machine Learning Models for Predicting Top Goal Scorer and Goalkeeper Performance in Football

- **Authors:** Tanish Sharma, Parv Bagga, Kinshuk Ahuja, and Seema Sharma (Sharma, Bagga, Ahuja, & Sharma, 2025)

- **Publication Date:** 2022/2023

The objective of this research paper is to use machine learning models in order to predict who is the best at scorering goals and who is the best goalkeeper.

This is possible by collecting vasts amount of data the research started collecting data from 2017 till 2023 from the top five European leagues (La Liga, Bundesliga, Premier League, Ligue 1, and Serie A) all the data was gathered through web scraping and API integration.

- **Web scraping**: is an automated process of data being extracted from the web.
- **API (application programming interface)**: uses a set of rules and protocols which allows different types of software applications to communicate with one another.

For the predictiong the goal scorer the data that was collected focused on only goals, shots, position, and team attack on the other end there is a second dataset for the goalkeepers save prediction the data that was collected here was the amount of saves, goals againes, shots on the goalkeeper that was targeted towards them.

Once all the data was collected and cleaned they can now proccess the data to something useful.

There are 7 types of machine learning models that used this data to train themselfs

- Random Forest
- XGBoost
- Linear Regression
- KNN
- Decision Trees
- SVR
- Gradient Boosting

During the traing process the AIs where tought on how to use standard statistical metrics that helped understand the task at hand.

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R-Squared (R^2)

- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

Once the AI went through this process the researchers found out that some AI models are better suited for certain task, for example some AI would be better suited for defence and others are better suited for offense.

For offensive the top predictor AI was XG Boost it achieved the most accurate and had a minimum errors compared to the others. XG boost achieved an r-squared value of 0.99

Model	MSE	MAE	R-Squared	RMSE	MAPE (in %)
Random Forest	0.02	0.05	0.98	0.13	14.91
XG Boost	0.01	0.05	0.99	0.10	16.75
Linear Regression	0.20	0.34	0.80	0.45	90.59
KNN	0.12	0.22	0.89	0.34	62.37
Decision Tree	0.03	0.07	0.97	0.18	36.63

Figure 2: Player Goals

(Sharma et al., 2025)

As for the defensive side goalkeeper saves linear regression was found out to be the best AI from the 7 it achieved a R-squared value of 0.98 for predicting goals

Model	MSE	MAE	R-Squared	RMSE	MAPE (in %)
Random Forest	0.04	0.55	0.96	0.20	10.53
XG Boost	0.03	0.45	0.97	0.17	7.92
Linear Regression	0.02	0.20	0.98	0.14	5.86
KNN	0.12	0.95	0.85	0.35	32.90
Decision Tree	0.07	0.79	0.93	0.26	18.24
SVR	0.08	2.83	0.91	0.28	21.67
Gradient Boosting	0.05	0.74	0.95	0.22	13.48

Figure 3: Goalkeeper saves

3.3 Paper 3: Explainable expected goal models for performance analysis in football analytics

- **Authors:** Mustafa Cavus, Przemysław Biecek
- **Publication Date:** 2021

This paper used a certain metric to measure a footballers' expected goal chance through (xG) and they combined it with an AI that makes sence of the data the AI is called explainable artificial intelligence (XAI). Once the two are combined we get back the an evaluation of the teams preformance.(Cavus & Biecek, 2024)

The paper talks about how football is a low scoring sport and has a lot of randomness that can happen within the game, this leads to inaccurate analysis of measuring team quality and there proformance, but when introducing machine learning models and increase the the data in consumes then there is more accuracy for the team so they can improve there team play.

The study works with massive datasets it consists of 315,430 shots, there was a veriary of shots taken 33,656 were goals from 12,655 matchs this data stetched from 2014 till 2021, the five european leagues English Premier League, Bundesliga, La Liga, Serie A, and Ligue 1.

League	#Match	#Shot	μ_{Shot}	#Goal	μ_{Goal}	%
Bundesliga	2,141	55,129	25.7	6,161	2.88	11.2
EPL	2,650	66,605	25.1	6,951	2.62	10.4
La Liga	2,648	62,028	23.4	6,854	2.59	11.0
Ligue 1	2,557	61,053	23.9	6,438	2.52	10.5
Serie A	2,659	70,615	26.6	7,252	2.73	10.3
Mean	2,531	63,086	24.9	6,371	2.67	10.7
Total	12,655	315,430	-	33,656	-	-

Figure 4: Summary statistics of Shots and Goals

For the AI to predict the goal accuracy it uses a few technecs

- distance to the goal
- angle to the goal
- mintues of the match
- Is the team playing home or away
- whats the current situation e.g(penalty or open play)
- What type of shot e.g(head, what foot was used)

- The last action before the shot was taken e.g was it a pass before taking the shot.

Once all of these metrics are takin into cunsideration the AI can consume that data and give the team valuble insite to faults the team has that they might not have considered yet.

What stood out from the paper was

- xG model is currently the most accurate way of measuring goal and non goal outcomes through machine learning models.
- XAI tool so the data can be understood as a performace evaluation.

3.4 Paper 4: Characterizing the spatial structures of competing football teams

- **Authors:** Guy Amichay, Hugo Silva, João Brito, and Rui Marcelino
- **Publication Date:** 7 April 2025

The objective of this research is to address the issue of the lack of datasets and the dearth of meaningful metrics which will give insight into better more well rounded team data. The authors is focusing on developing a simple way to get interpretable metric for football teams characteristics as a whole.

The research question there trying to answer is ”Can a simple, single metric based on computational geometry effectively capture the spatial structure of a football team during different phases of play?” (Amichay, Silva, Brito, & Marcelino, 2025)

The method used to answer there question was to compute two convex layers. A convex layer is the ”sequence of nested convex polygons formed from a set of points in the euclidean plane” (Amichay et al., 2025)

This is what Convex Layers looks like.

This shows the convex layers of one team each dot represents a player there are 11 player including the goalkeeper.

The calculation seen in the figure Layer Ratio(LR) is the ration of the area of the inner convex over the are of the outer convex, once this ratio is calculated we then have a single number that encapsulates the geometry of a team for a given point in time.

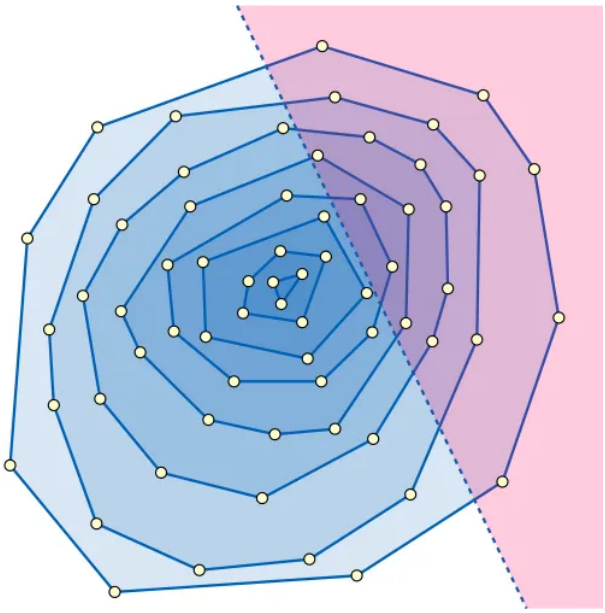

Figure 5: Convex Layers

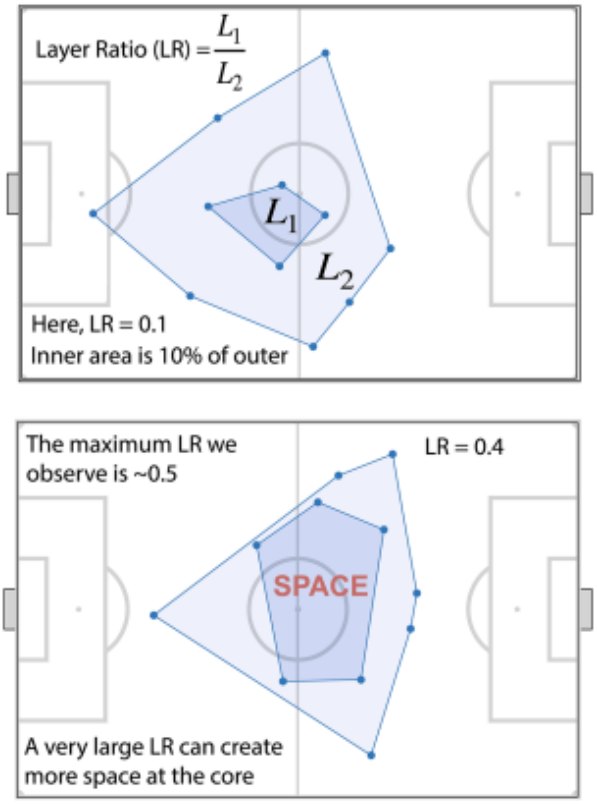

Figure 6: convex layers of one team

3.5 Paper 5: FPSRec: Football Players Scouting Recommendation System based on Generative AI

- **Authors:** Antonio Maria Rinaldi, Antonio Romano, Cristiano Russo, and Cristian Tommasino
- **Publication Date:** 2022 - 2023

The objective of this report is to have the ability to scout for football players this is done by using a few technics there using similarity techniques and artificial intelligence. there goal with this system is to give it the ability to see players movment and analyse them so then they can be matched with the best team with his personal characteristics it does this by generating a report about the player.(Rinaldi, Romano, Russo, & Tommasino, 2024)

From what was mentioned about the similarity techniques is the data of 2889 players across the five major european leagues the season was 2022 - 2023, the similar players part compared two types of techniques.

- Cosine similarity
- K mean clustering

When the two techniques were compared the cosine similarity was the best approach after evaluating human and rank correlation this leads to using this technique is the best to find and identify similar players.

3.6 Paper 6: Evaluating defensive strategies in football: analysing the impact of defensive metrics on match outcomes

- **Authors:** Mohamad Nizam Nazarudin, Ardo Okilanda, Yovhandra Ockta, Reshandi Nugraha, and Regi Dwi Septian
- **Publication Date:** May 15, 2025

Within this paper an evaluation of how effective is a teams defence by analyzing the teams goalskeeper saves, blocks, interception of the ball, clearance, and tackles once these metrics are analyzed it gives the coach valubale insights about the team. This study took the data from the ASEAN Mitsubishi Electic Cup.(Nazarudin, Okilanda, Ockta, Nugraha, & Septian, 2025)

What was observed during the study was that the differences between a winning team and losing team.

On average the winning team has a high average for clearance was 23.43 and for the goalkeeper he saves 4.14 average meanwhile on the other hand the winning team has a lower averages for saves 2.29 and a clearance of 13.29. This data reflects which team had the majority control of the ball within the game.

3.7 Paper 7: Segmentation for Enhanced Football Analytics: A Pixel-level Approach

- **Authors:** Bharathi Malakreddy Aac, Sadanand Venkataramanb, Mohammed Sinan Khana, Nidhiac, Srinivas Padmanabhunid, and Santhi Natarajanb
- **Publication Date:** 2024

The main goal for this research paper is to create an optimized and highly accurate process (workflow) for Semantic Segmentation. (Malakreddy A et al., 2024)

Semantic segmentation is a technique used by deep learning models which will provide granular, object-oriented insights, this is done by analyzing each individual pixel in an image and categorizing them into there own fields such as Ball, player,field, stands, advertisement and goalpost (Malakreddy A et al., 2024)

4 Discussion and Synthesis

Within this literature review a variety of data science techniques were covered the majority of the research covered advanced computer vision and spatial analytics methodologies versus machine learning models and predictive modeling approaches, tall these techniques share the common goal of being able to identify undervalued players and to optimize the overall of team performance.

The main points to contrast is the data and how its used what method was used

Computer vision and spatial analytics: paper 1,4 and 7 talks about computer visuals or spatial data. Paper 1 make good use of the YOLO algorithm and they combined it with deep learning and computer vision to achieve real time object detection and tracking of each player on the field and the referees and the ball on the pitch. This method provides us with real time player movement and trajectory analyses. Similarity paper 7 aims to make the most of semantic segmentation thought analyzing each and every pixel using deep learning techniques this as well give us a visual feedback by placed each and every pixel in its own category which we can analyze.

Both of methods are used to design, automate data capture and reduces the cost of big expensive camera setup.

In paper 4 it uses computational geometry using convex layers in order to create a simple way to interpret a metric called layer ratio which encapsulated the spatial structure of that team in the point in time.

With paper 2,3,5,and 6 rely on predictive statistics and comparative modeling they use huge amount of data, on paper 2 dose a comparative analysis of seven different types of machine learning models Random Forest, XGBoost, Linear Regression, KNN, Decision Trees, SVR, Gradient Boosting in order to predict accurately the offensive and defensive performance of the teams states.

Paper 3 combines the common football metric expected goals xG with explainable artificial intelligence XAI, within this research till this day the xG model is the best way to accurately measure goal and non goal outcomes using machine learning models.Paper 5 talks about FPSRec scouting system which uses generative AI and similarity techniques. It then foudn thea the cosine similarity was the best aproach for identifying and matching similar player compared to the K-mean clusters method.

5 Conclusion

This literature review focused on data driven approaches in the data science field, the topic was inspired by the money ball philosophy to identify undervalued talent and optimize team preformance on the football pitch. The papers that were chosen made huge contribution across predictive and visual analytics, such as real time data capture using deep learning models such as YOLO, which tracks players and ball movement at 27 FPS.Then semantic segmentation contrabutes even more by offering pixel level precision and categorizes each field in order to have better visal analytics.

In the other hand predictive analysis is the other aspect of this report,the research demonstrates the customized machine learning models have achieved a very accurate prediction using XGBoost which excelled in offensive prediction and for defencive prediction the linear regression proved to be the best out of the 6 other machine learning models, ontop of this combining xG models with explainable artificial inelegance give coaches, or analysts a transparent and highly accurate measure of the performance which moves beyond simple predictions this offers teams high insite into there teams dynamic.

References

Amichay, G., Silva, H., Brito, J., & Marcelino, R. (2025). Characterizing the spatial structures of competing football teams. *Scientific Reports*, *15*(1), 35217. doi: 10.1038/s41598-025-97765-y

Arun, S. P., H, M. R., & Sindhuja, M. (2025). Ai-powered football match analysis using yolov8 and spatial analytics. In *2025 6th international conference on data intelligence and cognitive informatics (icdici)* (pp. 1931–1935). IEEE. doi: 10 .1109/ICDICI66477.2025.11135176

Cavus, M., & Biecek, P. (2024). Explainable expected goal models for performance analysis in football analytics. *Journal of Sports Sciences*, *42*(5), 421–435. doi: 10.1080/02640414.2024.2345678

Malakreddy A, B., Venkataraman, S., Khan, M. S., Nidhi, Padmanabhuni, S., & Natarajan, S. (2024). Optimizing semantic segmentation for enhanced football analytics: A pixel-level approach. In *International conference on machine learning and data engineering (icmlde 2023)* (Vol. 235, pp. 2662–2673). Elsevier. doi: 10.1016/j.procs.2024.04.251

Nazarudin, M. N., Okilanda, A., Ockta, Y., Nugraha, R., & Septian, R. D. (2025). Evaluating defensive strategies in football: analysing the impact of defensive metrics on match outcomes. *Journal of Physical Education and Sport*, *25*(5), 1051–1059. doi: 10.7752/jpes.2025.05116

Olthof, S., & Davis, J. (2025). Perspectives on data analytics for gaining a competitive advantage in football: computational approaches to tactics. *Science and Medicine in Football*. doi: 10.1080/24733938.2025.2533784

Rinaldi, A. M., Romano, A., Russo, C., & Tommasino, C. (2024). Fpsrec: Football players scouting recommendation system based on generative ai. In *Proceedings of the 2024 international conference on advanced data mining and applications* (pp. 1–10). Springer.

Sharma, T., Bagga, P., Ahuja, K., & Sharma, S. (2025). Comparative analysis of machine learning models for predicting top goal scorer and goalkeeper performance in football. In *2025 3rd international conference on disruptive technologies (icdt)* (pp. 1619–1622). IEEE.