International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s, 2025
https://www.theaspd.com/ijes.php

Footyintel: Creating An AI Scout For Better Talent Recognition

Sandeep Raskar[1], Manas Thosar[2], Atharva Dandge[3], Pranav Fale[4]

[1]Department of AI and Data Science, Terna Engineering College, Mumbai University, Navi Mumbai, India, raskarsandeep@ternaengg.ac.in

[2]Department of AI and Data Science, Terna Engineering College, Mumbai University, Navi Mumbai, India, thosarmanas2122@ternaengg.ac.in

[3]Department of AI and Data Science, Terna Engineering College, Mumbai University, Navi Mumbai, India, dandgeatharva2122@ternaengg.ac.in

[4]Department of AI and Data Science, Terna Engineering College, Mumbai University, Navi Mumbai, India, falepranav2122@ternaengg.ac.in

***Abstract*–*In many ways, artificial intelligence (AI) is transforming the way talent in football is identified, and the difference it makes shows in teams scouting and recruiting. AI systems, created with a combination of machine learning, computer vision, and large language models, can analyze vast numbers of player data, ranging from performances in matches, physical measurements, and skill levels across leagues and regions. Because subjective human judgment is involved, AI is an objective and data-driven approach to scouting for promising talents in football. The other type of identification is the prediction of a player's potential on the basis of historical data and developmental trends, which will give the clubs precious information to make much better recruitment decisions. Challenges in the use of AI in football scouting include data privacy concerns, ethical issues, especially about the profiling of players, and the danger of perpetuation of biases. Despite all these, AI is a solution that may easily improve the efficiency and accuracy of talent evaluation in football.*

***Index Terms*–*LLM Training , Data-Driven, Parameters, Performance-Oriented*

I. INTRODUCTION

In the contemporary data-driven landscape, artificial intelligence (AI) is rapidly superseding traditional methodologies for identifying talent across diverse sectors, including sports and business. While AI has become indispensable in fields such as healthcare and finance, its application in talent recognition remains in its nascent stages. Historically, talent identification has been predicated on subjective assessments, resulting in inconsistent outcomes and overlooked potential. Despite advancements in AI, current talent recognition systems in soccer do not fully incorporate sophisticated data analytics, com pelling clubs to rely on conventional, often biased, scouting techniques. As soccer evolves into a data-intensive, performance oriented sport, traditional methods of scouting and talent assessment face significant challenges. Scouts and coaches typically rely on subjective observations and intuition, which, although valuable, often introduce bias and inconsistency. For instance, players' potential may be disregarded due to personal biases, geographic limitations, or insufficient scouting resources. In an increasingly globalized sport where clubs compete to identify and acquire emerging talent, there is a growing demand for more objective, data-driven solutions. AI presents a unique opportunity to address these issues by providing more accurate, consistent, and comprehensive evaluations of player performance, potential, and suitability for specific teams. AI-powered talent recognition in soccer utilizes advanced techniques such as machine learning, computer vision, and predictive analytics to examine a broad range of data, including match statistics, physical metrics, player movement patterns, and even social media activity. These systems can process and analyze vast amounts of information in real time, revealing insights that human scouts might overlook. For example, AI can evaluate a player's decision-making ability, adaptability, and tactical awareness by examining historical game footage and tracking data. By identifying key performance indicators and patterns, AI systems can forecast how a player is likely to perform in future scenarios, offering clubs a more strategic and data-supported approach to talent recruitment. As AI continues to advance, it has the potential to transform not only talent identification but also player development, team dynamics, and overall performance optimization in soccer.

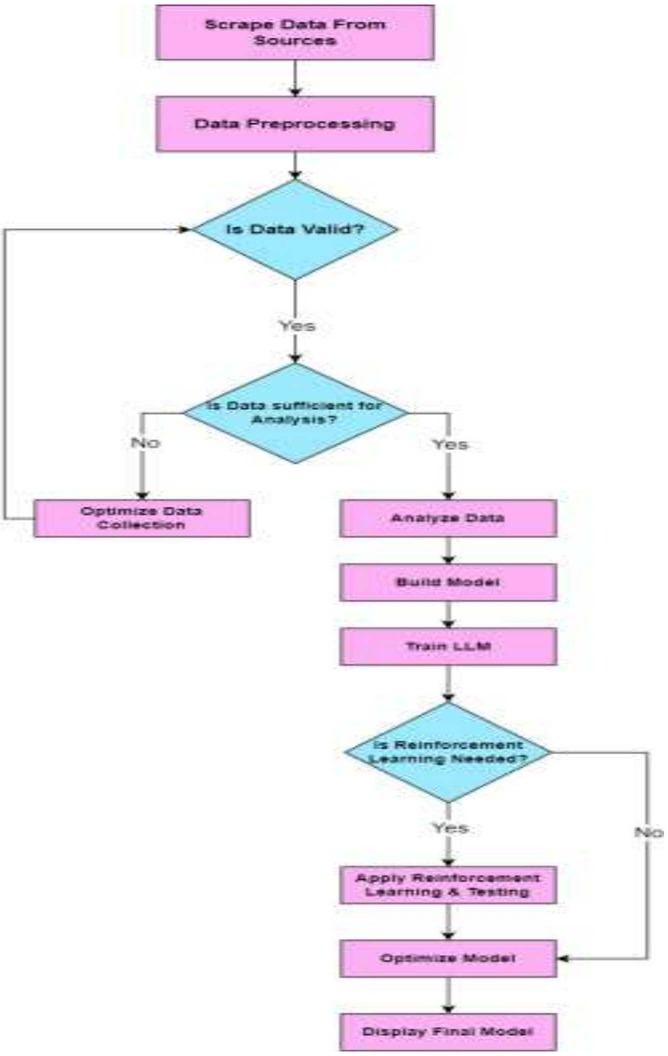
II. LITERATURE SURVEY

This literature review aims to explore the advancements in Artificial Intelligence for scouting talent in football, focusingparticularly on how data driven approaches can improve player evaluation and scouting process. Through synthesis of previous studies, the review captures the challenges and innovations involved in these areas, thus providing a base for further studies.The findings will contribute to the expanding body of knowledge on AI in sports and provide valuable insights for clubs seeking to gain a competitive advantage in player recruitment. Some studies show how to make a tactical plan, according to the player's potential. For example a paper[1] suggests the use of a tactical board, having an iterative design, to discuss the game plan. It supports the quick exploration and retrieval of relevant game situations for strategic review, saving us the manual time and efforts required.Orientation based decision making is very essential in the sports, especially in the context of football. Addressing the challenge of determining when and whom to communicate with, and how to efficiently share information among players to maximize their shared utility in a partially observable environment. By the use of techniques such as MAGIC (Multi-Agent Graph-Attention Communication) algorithm, we will have a improved accuracy in tactical and strategic decisions based on data-driven insights [2] Studies as [3] demonstrates the real time match/player's analysis. First paper demonstrates the optimization of real time video analytics and distributed computing systems.It requires players to wear devices such as watches or bands, to capture their real time performance.While the benefit of capturing the real time data is there, the devices used for the same can be expensive, leading to financial inefficiency. Also, one needs to have a high level of expertise in distributed computing to process and analyze the collected data. Some research revolves around the use of machine learning to predict the positions of players, estimate their goals for upcoming seasons/matches and determine shots taken per match, for a particular player.It compares a MLP model with several machine learning models such as SVM, random forest, etc. Machine learning might give us an optimized model, but we may face challenges such as data preprocessing, limited scope of data, etc. [4] While majority of research focuses on extracting match data, some of the research paper focuses on orientation based decision making, exploring how a player should orient himself on the field, in order to avoid any kind of injury.This paper demonstrates how pose detection techniques like OpenPose, Super resolution network and coarse orientation validation can be used to accurately determine the orientation of a player's upper torso on the field, which is crucial for understanding player behavior and performance in various game scenarios.The method achieves 96.57 percent accuracy in determining a player's orientation using pose and contextual data.This approach is very useful to simplify interpreting raw orientation data, offering valuable sports analytics insights, the over reliance on image quality can be a factor of drawback, as it may capture many images which might be blurred, low color quality, etc. [5] The value of large language models, in this case, for sports or football specifically, is seen in processing and analysis of data. That is how LLMs facilitate tactical analysis and make observations through comprehensive analysis of game footage, identification of trends, and providing insight into player and team performances. LLMs enable coaches to analyze their opponent's weaknesses and offer recommendations for game plans. LLMs will now engage fans through live commentary, specially created content, and can cover the multiple game situations simultaneously, without human intervention. These commentary generating large language models can be customized according to the audience's preference [6]. Also, we first need to have a basic level of understanding about large language models, how LLM's can be augmented using external tools so as to enhance their performance on tasks requiring reasoning, context retrieval, and complex problem solving.We can use techniques such as Retrieval-Augmented Generation (RAG) and Encoded Context Augmentation to make our LLM more efficient and optimized. Using external tools and retrievers, LLMs excel across diverse tasks, including mathematical reasoning, natural language understanding, and problem-solving [7]. But there is a challenge of managing interactions between the LLM, tools, and retrievers, which is a complex process, needing careful coordination to ensure efficient and accurate task execution.

There are studies revolving around predicting the outcome of the football match, using multi layer perceptrons(MLP's). And then compare them to other algorithms such as naive bayes, SVM, etc. We can get a comprehensive set of features that will help us in capturing the nuances on which the outcome of the match may depend [8].We can train our large language model using these methods, but first we need to provide the appropriate prompt to the model, so that it will give us accurate result.We can employ prompting strategies, such as p1,p2,p3,p4. They demonstrate how the model will provide us with a certain answer based on the prompt that we provide to it [9]. They identify the regression causes and help us in future guidance, so that we can further increase the efficiency of our model.Also we can scale our model to create valuable insights from business intelligence and analytics in competitive sports. The use of BI and A in sports can lead to better performance for teams and athletes [10].

III. METHODOLOGY

We are going to perform a comprehensive manner on the development of an AI-based talent scout for football in this study. First, data collection on player performances is going to be carried out through various truthful and reliable sources, such as FBREF, StatsBomb, and SofaScore focusing on shot maps, ball receipts, and general playstyle. Historical data from players across leagues will be collected so that their abilities can be gauged. After gathering the data, the pre-processing phase will remove null values, noise, and outliers, which can deteriorate data quality. We have a few tools from the pandas library that we might employ: Cleaning the dataset, shot distribution, and intensity By using the data visualization techniques like heatmaps and shot maps, we will represent them. Analysis of this visualization and drawing out insights regarding the performance will be of focus. The a comprehensive procedure will ultimately result in the training of our model with this meticulously curated and processed dataset, enabling us to assess and forecast player potential with greater efficacy compared to conventional scouting techniques.



A. Data Collection

To prepare the dataset for training the model, we will first need to collect the data from various resources. For our project, we need the historical data of individual players in different leagues, to assess their performance. We will observe the player's percentile metrics, shotmap and their ball receipts and understand their playstyle further.These are the performance metrics we will look for while collecting data from the web. To collect our dataset, we will scrape data from websites such as FBREF, Statsbomb and SofaScore. From FBREF, we can obtain the data about the particular player's performance in a particular season of a league and from SofaScore, we can get the data about an individual player's performance in a specific match. To prepare the dataset for our project, from FBREF, we collected the data of Florian Wirtz, a German Attacking Midfielder playing for German Club Bayer Leverkusen and the German National Team. From FBREF we have analyzed his performance of the past 365 days when compared to other players in Europe's Top 5 Leagues. We have also chosen the Bundesliga match between Bayer Leverkusen and Borussia Monchengladbach held on ¨ 24th August 2024 to analyze the shots at goal attempted by him. Furthermore, we will also analyze his shotmap for EURO's 2024, available through StatsBomb free data to understand his shooting abilities while playing for the national Team. We will also use Statsbomb data of

B. Data Preprocessing

After collecting the dataset, we cannot readily use it, as it may contain noise, dirty data, outlier points, etc. which can cause hindrance in our model, while performing on unseen data. Firstly, we need to get rid of this dirty data, in order to increase our dataset quality. We will remove null spaces and merge rows, so that we can neatly view our data. Firstly, we will scrape the team data from FBREF, of the champions league 2022-23 season. The team-wise collected data from FBREF is viewed using the pandas library. Further, we will clean all of the null space in our data using the Dropout feature of pandas.

C. Data Visualisation

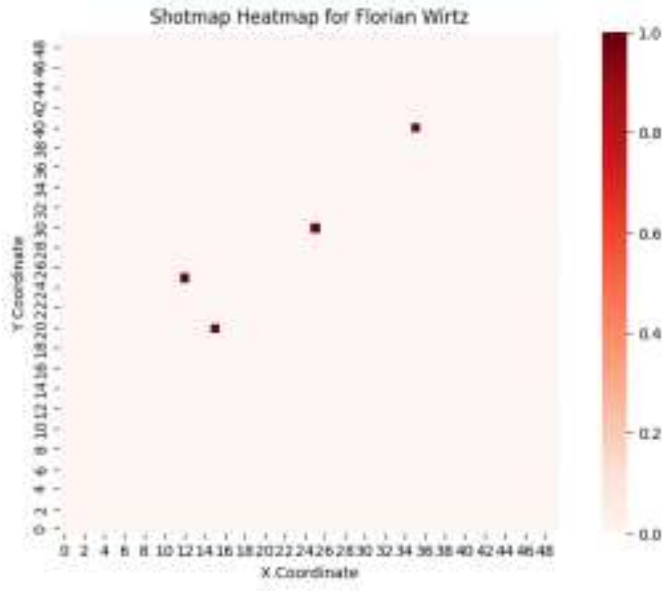


Fig. 2. shot map heatmap of Florian Wirtz

Figure 1 illustrates this graphical interpretation in a shot map heatmap of Florian Wirtz, with all information on the distribution and effectiveness of shot attempts in a two dimensional space across a particular match or set of matches. The x-axis will represent the horizontal (X) coordinate representing the shot locations, and the y-axis will represent the vertical (Y) coordinate on the field, indicating a spatial configuration such as that of a football pitch. Gradient of color in the heatmap-all the lighter shades point towards deeper ones as being more intense or with more shot taking places. Darker red squares indicate heavier shots and so a lighter area means lower or no shots at all.

In this specific visualization, four different locations on shots are observed, represented by a dark square with varying degrees of intensity.

The data points scattered out across the x and y coordinates range from about 10 to 38 on the axes. Again, the color intensity suggests that these are single isolated events; each position that exists represents one attempt at a rare or unique shot rather than a broad concentration of attempts based in the same general area.

The color bar at the right side of the plot in the heatmap shows that intensity ranges from 0.0 to 1.0, and the value 1.0 shows the highest frequency or intensity found from the data set. With the graph that is already in place, there is no shot location on the 1.0 intensity meaning that the attempts are moderately spread out across the field in different places. The heatmap may then be a useful tool in understanding the spatial trends related to shot-taking behavior of Florian Wirtz, therefore, hence providing information about his preferred shooting areas or patterns in offense. It may also be a starting point for further tactical analyses or performance appraisals of players concerning shot taking and attempts.

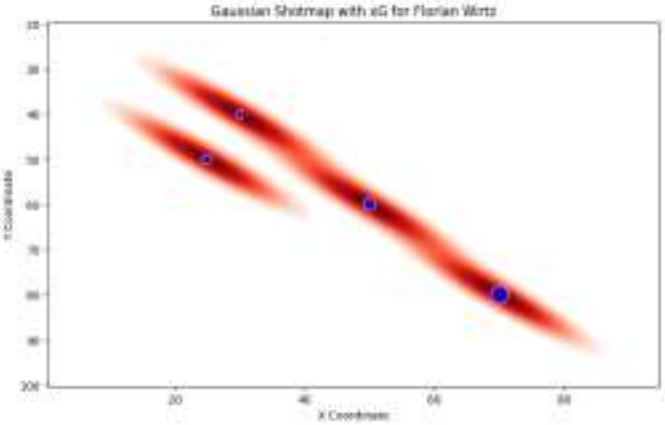
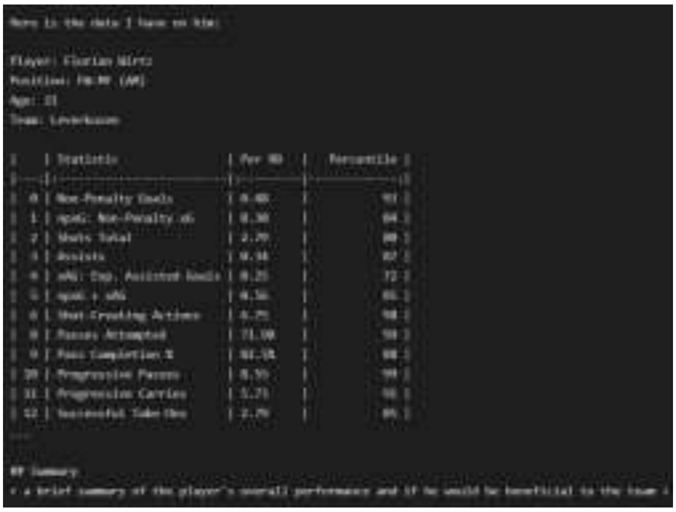


Fig. 3. Gaussian xG visualisation plot

The shot map above is an aggregate visualization of all the shots taken by any player in the entire UEFA Euros 2024 tournament. It gives the most important information regarding spatial patterns in shot taking, tendencies in shot distribution across zones and outcome distributions among others. The shot density can be represented strongly within the internal and external regions of the field, which are mostly central in nature since it is significant between the penalty spot and the 6-yard box. This concentration of shot attempts aligns with the context of modern football strategies: to concentrate solely on high-class opportunities, reasoning that shots in these areas have the highest xG values. High concentration already shows that most teams seriously direct most of their attacking efforts toward creating scoring chances in these zones. Success and xG Correlation: The green markers, which are goals, are concentrated mainly in the middle section inside the penalty box, which supports the model of xG that interprets the higher chances of converting a shot from close proximity to the goal. These regions show good tactical efficiency as the conversion of shots is maximized because of the good position it is taken from. Dispersion of Non-Scoring Shots: White dots representing unsuccessful attempts to hit the target are spread out more diffusely. Most of them are still within the penalty area, but their distribution extends beyond the box and even at the fringe of the penalty arc boundary. This spread seems to suggest that players attempt shots from distances or angles considerably more acute than those in the squares. In almost all cases, they are less likely to score from those areas, but again it suggests how teams have had to create different attacking opportunities through perhaps tactical situations like a really tight defensive block or sheer time constraints. xG Variation and Shot Quality The circles represent different expected goals (xG) values for every shot. Larger circles indicate higher chances of xG, which bunch closer to the goal, especially within the 6-yard box. High-quality shots are those closer to the goal: close-range finishes, one-on-one opportunities with the goalkeeper, or shots after defensive errors. Conversely, more spread out and frequent smaller circles are located in non-box regions, indicating speculative or long-range efforts. Spatial Trends and Tactical Implications So, if goals are spread out both across and just beyond the penalty area, it may be that teams are jumbling short and long efforts into the mix, likely sensitive to game context, defensive structures, and overall quality of the given opportunities. The still great concentration of goals in the green circles near the goal means, once again, that while speculative long-range attempts are indeed made, most successful efforts are in more controlled, high-xG positions within the box.



```
Here is the data I have on him:

Player: Florian Wirtz
Position: FW-MF (AM)
Age: 21
Team: Leverkusen

|    | Statistic                  | Per 90  |   Percentile |
|---:|:---------------------------|:--------|-------------:|
|  0 | Non-Penalty Goals          | 0.48    |           93 |
|  1 | npxG: Non-Penalty xG       | 0.30    |           84 |
|  2 | Shots Total                | 2.79    |           88 |
|  3 | Assists                    | 0.34    |           87 |
|  4 | xAG: Exp. Assisted Goals   | 0.25    |           72 |
|  5 | npxG + xAG                 | 0.56    |           85 |
|  6 | Shot-Creating Actions      | 6.75    |           98 |
|  8 | Passes Attempted           | 71.90   |           99 |
|  9 | Pass Completion %          | 82.1%   |           88 |
| 10 | Progressive Passes         | 8.55    |           99 |
| 11 | Progressive Carries        | 5.73    |           91 |
| 12 | Successful Take-Ons        | 2.79    |           85 |
---

## Summary
< a brief summary of the player's overall performance and if he would be beneficial to the team >
```

Fig. 4. Data scraped from FBREF

The following visualization is a comprehensive scouting report, built from data from FBREF, which compares Wirtz's performances on several key metrics with other attacking midfielders and wingers competing in the top five European leagues. Important Indicator: Attacking Output: Wirtz's non-penalty goals per 90 minutes sits at 0.48, ranking him at the 93rd percentile, while his non-penalty expected goals sit at 0.30, ranking him at the 84th percentile, so once again, a high chance of goal-scoring along with actual output. Creative Input: The forecasted assists (xAG) at 0.25 puts him in the 72nd percentile, and 0.34 assists per 90 minutes stands at the 87th percentile. But what is most impressive about Wirtz is his stunning 6.75 shot-creating actions per 90 minutes, which put him in the 98th percentile. Passing and Progression: His passing metrics are phenomenal, averaging 71.9 passes per 90 minutes (99th percentile), in tandem with 8.55 progressive passes per 90 minutes (99th percentile) and 5.73 progressive carries per 90 minutes (91st percentile). These numbers highlight his role as a singular transitional player that has always forwarded the ball upfield. Defensive Contributions: Although an attacking player above all, his interceptive contribution is at the 71st percentile, while his tackling contribution (36th percentile) and clearances (3rd percentile) were far lower, which is to be expected for a player playing in this role. This report confirms Wirtz's elite ability to score and create, making him one of the best players in the league from the position,
with his shortcomings emerging mainly at the back.

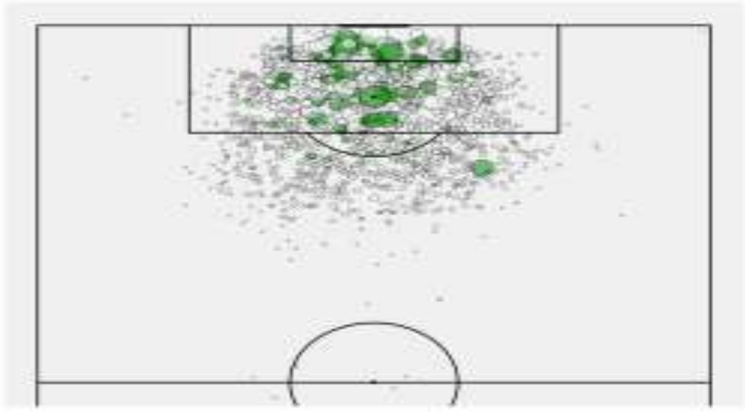


Fig. 5. Aggregated Analysis of Shot Locations for All Players in UEFA Euros 2024

The shot map above provides an aggregate visual representation of shots taken by all players throughout the UEFA Euros 2024 tournament. This comprehensive dataset reveals crucial insights into shot-taking patterns, spatial tendencies, and outcome distributions across different zones of the pitch.

Shot Density in High-Probability Areas A significant concentration of shots is observed in the central areas inside and just outside the penalty box, particularly between the penalty spot and the 6-yard box. This dense cluster of shot attempts aligns with modern football tactics that prioritize high-quality chances, as these areas typically present the highest expected goal (xG) values. The heavy concentration suggests that most teams focus their attacking efforts on generating scoring opportunities in these optimal zones. Success and xG Correlation, The green-colored markers, which denote goals, are primarily concentrated within the central area inside the penalty box, corroborating the xG model that suggests higher chances of scoring from close proximity to the goal. These areas reflect tactical efficiency, where shot conversion rates are maximized due to the advantageous positions from which they are taken.

Dispersion of Non-Scoring Attempts, The white circles, indicative of shots that did not result in goals, show a wider distribution. While many of these unsuccessful attempts are still located within the penalty area, their presence extends to areas outside the box and even near the outer edges of the penalty arc. This spread suggests that players frequently attempt shots from longer ranges or more acute angles, where the probability of scoring is generally lower. Despite the low success rates from these areas, they reflect teams' efforts to create varied attacking opportunities, perhaps due to tactical situations such as tight defensive blocks or time constraints in the game. xG Variation and Shot Quality, The varying sizes of the circles correspond to different expected goals (xG) values for each shot. Larger circles represent higher xG chances, which tend to cluster near the goal, particularly inside the 6-yard box. These shots are of higher quality, reflecting situations such as close-range finishes, one-on-one opportunities with the goalkeeper, or shots taken after defensive errors. Conversely, smaller circles (lower xG shots) are more dispersed and frequent in regions outside the box, indicating speculative or long-range efforts.

Spatial Trends and Tactical Implications, The spread of shots along the penalty box and just beyond its boundaries suggests that teams utilize a mix of short- and long-range efforts, likely depending on game context, defensive setups, and the quality of chances created. The concentrated area of goals (green circles) near the goal emphasizes that while speculative long-range shots are attempted, the bulk of successful efforts arise from more controlled, high-xG positions within the box.

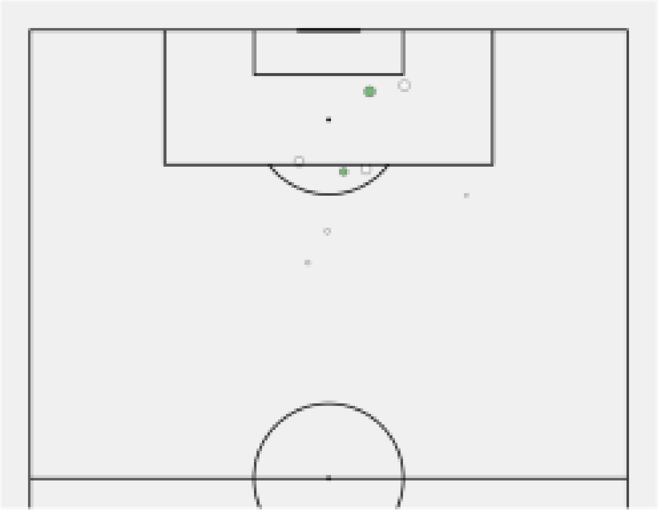


Fig. 6. Analysis of Shot Locations for Florian Wirtz (Germany) in UEFA Euros 2024

The shot map above provides a visual representation of all shots taken by Florian Wirtz during the UEFA Euros 2024 tournament. The positions on the field, the size of the markers, and their colors offer insight into shot outcomes, expected goals (xG), and shot locations.

Wirtz's shot selection predominantly takes place within high-value zones near the opponent's goal, particularly in central areas just outside the 6-yard box. Of the shots visualized:

High-Concentration Central Shots, Several shots are clustered in the central area of the penalty box, specifically in the region between the penalty spot and the goal. This positioning indicates a tactical focus on exploiting high-probability scoring areas, in line with modern football's emphasis on maximizing xG by prioritizing central shooting zones. Goals, The green-colored markers, indicative of goals, are notably concentrated within the 6-yard box and just inside the penalty area. This positioning confirms that Wirtz's suc-cessful attempts came from extremely advantageous positions, highlighting his ability to capitalize on high-xG opportunities created from close proximity to the goal. Unsuccessful Shots, The white markers represent shots that did not result in goals. These are more dispersed, extending from within the penalty area to just outside the box. The presence of unsuccessful attempts from these slightly more distal locations suggests a moderate risk in shot selection when compared to the precision demonstrated within the 6- yard box. This distribution may reflect situational demands, such as forced shots under pressure or attempts from set-piece plays. xG Influence, The varying sizes of the markers correlate with expected goals (xG) values, indicating the statistical likelihood of each shot resulting in a goal. Larger markers, particularly near the goal, reflect higher xG values, underscoring the tactical importance of positioning and the correlation between proximity to the goal and shot efficacy.

Overall, Wirtz's shot map emphasizes a high level of efficiency in shot selection, focusing on central zones where xG values are maximized. His ability to convert from such advantageous positions reflects both individual skill and Germany's offensive structure designed to create high-quality chances for key attacking players like Wirtz.

IV. CONCLUSION

The conclusion outlines how AI gives a more objective and data-driven approach to the process of scouting talent in football as compared to the more traditional ones. AI assists the coaches with better recruitment choices, analyzing large datasets, and predicting players' potential, notwithstanding data privacy and biases, revolutionizing the talent identification process in football.

REFERENCES

[1] D. Seebacher, T. Polk, H. Janetzko, D. A. Keim, T. Schreck, and M. Stein, ”Investigating the Sketch Plan: A Novel Way of Identifying Tactical Behavior in Massive Soccer Datasets,” in *IEEE Transactions on Visualization and Computer Graphics*, vol. , no. 08, pp. 645-87, 2021, doi: 10.1109/TVCG.2021.3134814.

[2] Z. Pu, et al., ”Orientation and Decision-Making for Soccer Based on Sports Analytics and AI: A Systematic Review,” in *IEEE/CAA Journal of Automatica Sinica*, vol. 08, no. 07, pp. 127-754, 2023, doi: 10.1109/JAS.2023.123807.

[3] D. Jha, et al., ”Video Analytics in Elite Soccer: A Distributed Computing Perspective,” in *IEEE International Symposium on Signal and Information Processing (SAM)*, 2022, pp. 63-217, doi: 10.1109/SAM53842.2022.9827827.

[4] S. Gupta and A. D. Bavani, ”Leveraging Machine Learning for Sports Analytics: An Application to Football,” International Journal of Advanced Computer Science and Applications (IJACSA)”, vol. 12, no. 9, 2021, doi: 10.14569/IJACSA.2021.0120906.

[5] A. Arbues-Sang ´ uesa, A. Mart ¨ ´ın, J. Fernandez, C. Rodr ´ ´ıguez, G. Haro and C. Ballester, ”Always Look on The Bright Side of The Field: Merging Pose and Contextual Data To Estimate Orientation Of Soccer Players,” 2020, pp. 1506-1510, doi: 10.1109/ICIP40778.2020.9190639

[6] A. Cook and O. Karakus ̧, ”LLM-Commentator: Novel Fine-tuning Strategies of Large Language Models for Automatic Commentary Generation Using Football Event Data,” in *Knowledge-Based Systems*, vol. 888, pp. 106-316, 2024, doi: 10.1016/j.knosys.2024.112219.

[7] Kojima, E., Gu, S., Roberts, A., Mi, P., Zhou, X., Hashimoto, T. (2023). Large Language Models Are Zero-Shot Reasoners. arXiv preprint. DOI: 10.48550/arXiv.2307.06435

[8] S. L. Burla and K. N. Parthasarathy, ”Prediction of Football Match Outcomes Using Deep Learning Models", " International Journal of Recent Technology and Engineering (IJRTE)", vol. 8, no. 5, 2020, doi: 10.35940/ijrte.E3112.018520.

[9] A. Roy and K. Veeramachaneni, ”Why Is My Prompt Getting Worse? Rethinking Regression Testing for Evolving LLM APIs,” arXiv preprint arXiv:2308.08545, 2023, doi: 10.48550/arXiv.2308.08545.

[10] O. Caya and A. Bourdon, ”A Framework of Value Creation from Business Intelligence and Analytics in Competitive Sports,” 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016, pp. 521-32, doi: 10.1109/HICSS.2016.136. keywords: Business Intelligence;Analytics;Competitive Sports;Value Creation.