LINE DETECTION PROJECT

**STUDENT INFORMATION**

2059935

u825020

Daan Jan Grotenhuis

Cognitive Science and Artificial Intelligence Masters

**EXTERNAL SUPERVISORS**

Ryan Stepfner, FORWARD FOOTBALL B.V.

**ACADEMIC SUPERVISOR**

Dr. Eriko Fukuda

Dr. Ifigeneia Mavridou

**DATE**

August 21th, 2025

**LINE DETECTION PROJECT**

**Daan Jan Grotenhuis**

**Research Section**

1. **Internship Goals**

The goal of my internship is to create a line system that labels successful passes in a football game based on the player's position and team formation. This system will identify the starting and ending lines and zones for each pass. The football pitch is divided into three distinct zones: Defensive Zone, Middle Zone, and Attacking Zone. These zones are static and depend on the pitch size, which ranges from 100 to 105 meters in length and 64 to 68 meters in width. Players are positioned within three lines: Defensive Line, Middle Line, and Attacking Line. The company Forward Football, where the internship is being conducted informed that there are 2 approaches to solve this problem. One being a fixed line system where the lines are fixed purely on the role of a player, so a centre-back (CB) will always be considered to be part of the defensive line, a central midfielder (CM) will always be considered to be part of the midfield line, regardless of their or their teammates positions on the field. Whilst this is the simplest approach the company noted that due to the time constraints this will be the minimum results they expect. The other approach, which is also the preferred method, is creating a dynamic line system, meaning they change throughout the game depending on player movements and tactics. The challenge of this project lies in the dynamic nature of football lines. While the initial formation of the team might indicate the general position of each player, it’s common for players to shift between lines during the match. For example, a defender might join the midfield line or an attacking player might drop back into the defensive line, depending on the playing style and tactical decisions. Therefore, my task is to create a system that can identify and label the starting and ending lines of a pass, based on the player's position at the time of the pass, even as those lines change dynamically.

The desired output for this project is an enhancement to a dataset that includes the parameters of the pass: the starting and ending zones, as well as the starting and ending lines. To achieve this, I will classify each pass instance by detecting the dynamic lines and zones where the pass begins and ends, accounting for player movement throughout the game. This will involve both fixed definitions (based on the specific defensive, midfield, or attacking players) and dynamic definitions (based on the player's position in the moment).

The table below shows an example of the desired labels to the corresponding image showing the pass instance:

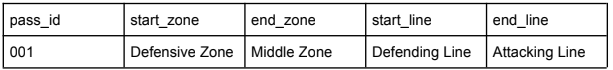
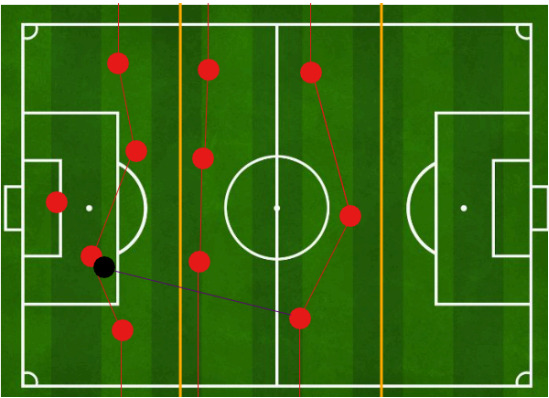
Table 1: Example of the output displaying the desired parameters to enhance the dataset

Image 1: Example of a snapshot in a game where the defensive, midfield and attack lines (red) are displayed (same order in image as in description) and the pass of the ball (black)

As shown in the image a pass (in black) was made from 1 player to another. The zone the pass was played in was in the first 3rd of the pitch (the defensive zone) and the player receiving the pass stood in the second 3rd of the pitch (the midfield zone). This is then noted as shown in table 1 as these are 2 of the 4 desired output labels. Additionally, the pass was made from the player in the back (defensive line) and ending with a player in the front (attacking line), this is one again shown in table 1.

This task is important as the label of the starting line and ending line are requirements of the FIFA standard to have in positional xy coordinate data which is still needed for Forward Football. Additionally, it can help with many different aspects of football, such as training, performance analysis, development, and more.

For training, the lines can be used to design tactical drills, such as practicing passes between lines or exploiting spaces between them. Line detection also aids in player positioning feedback, making players maintain optimal positions to create better passing lanes1. Furthermore, it provides insights into player decision-making, helping them refine choices under pressure.

For performance analysis, line detection is valuable for assessing space exploitation, such as how effectively a team is breaking defensive or midfield lines to create opportunities. It also helps in transition analysis, showing how teams switch between defense and attack and the zones where these transitions occur. Additionally, it allows for the automatic identification of key passes, showing high-impact passes from players during a match2.

For development, it can be crucial in player scouting, identifying players with exceptional vision and passing skills by tracking line-breaking passes. It can also support profiling for a specific style of play, aiding teams in developing strategies such as wide play or central overload. Furthermore, match preparation benefits from understanding an opponent's passing patterns and defensive structure, enabling counter-strategies3.

Beyond these areas, line detection can enhance broadcast visualizations, providing real-time tactical insights for audiences, and integrate seamlessly with advanced metrics like Expected Goals (XG) to deliver a more comprehensive analysis4.

1. **Methods**

The methods for this research project will use data provided by Forward Football B.V., which includes recordings from previous matches captured using their proprietary instruments. This dataset contains various labels useful for analyzing passing lines and positional structures. A key component of the dataset is the 5 Hz positional xy-coordinate data of all players and the ball. This high-frequency spatial data will serve as the foundation for identifying and classifying positional lines, such as defense, midfield, and attack. By analyzing this information, we can determine when and where a pass originates, as well as the positional line it is received in. For instance, we can identify a pass originating from the defensive line and received in the midfield line, enabling a detailed understanding of transitions during gameplay.

To implement a dynamic line classification system, I will explore different clustering algorithms to label the “line” that players belong to. Clustering algorithms are well-suited for this task because they dynamically group players based on proximity, creating a flexible system that adapts to in-game transitions. For example, a CB moving from the defensive line into the midfield line, or a central midfielder CM advancing into the attacking line, can be seamlessly reclassified based on their positional proximity to other players. This dynamic approach reflects the fluid nature of player roles during a match.

Furthermore, as there are no predefined or “true” labels for these lines in the given data, the analysis will operate in an unsupervised setting. This makes clustering algorithms particularly advantageous, as they do not rely on labeled data and instead uncover inherent patterns within the dataset5. By applying clustering algorithms, I aim to develop a robust system for real-time player line classification that mirrors the fluid dynamics of football gameplay.

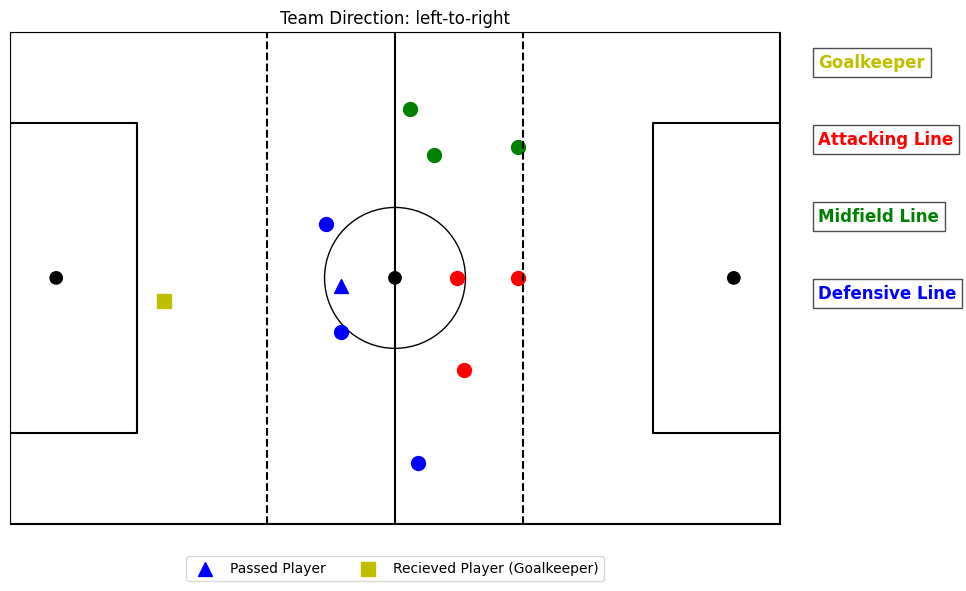
K-means clustering has proven to work with positional xy coordinate data as input for football6. Additionally, it allows me to specify the amount of clusters, which can be useful as the algorithm might create more undesired clusters, since we know there will always be 3 lines in football7. However, K-means is a more “simple” model and can be lackluster when players are in transition of being in one line to another8.

Gaussian Mixture Model (GMM) is shown to be effective for clustering football events and players as demonstrated by Spencer et al. (2017) in the paper titled “Measuring Player Density in Australian Rules Football Using Gaussian Mixture Models” where player density is Australian football were clustering using GMM. Furthermore, GMM assumes that the data comes from a mixture of Gaussian distributions, allowing for soft clustering. This could be useful when there are overlaps between lines (e.g., a player who roams between defense and midfield)9. Additionally, similarly to K-means it allows to specify three components, which is important to this project as mentioned before.

Agglomerative Clustering with Distance Constraints (ACDC) is a model that allows for constraints to maintain a continuous line separation10. For example, you can ensure that clusters maintain a north-south direction (e.g., restricting the clustering to vertical bands that we expect on a football field). Which is especially useful in football data as players are more clustered in vertical lines and not purely based on their relative proximity to each other. Similarly to the other discussed models agglomerative clustering has been used for tasks in football before with great results, such as the paper “Cluster Analysis on Football Teams Performance Data” by Kalt et al. (2024) in which the application of hierarchical clustering, including agglomerative clustering, to classify football players based on performance data. The use of this model might be more complex to implement but can provide more accurate clusters to the given task.

Other more advanced models such as Density-based spatial clustering of applications with noise (DBSCAN) were also considered. While DBSCAN shows great potential, as DBSCAN allows for clusters of arbitrary shapes, which is useful for identifying the continuous lines of players on the field11. Additionally, DBSCAN is known for its ability to handle noise, as it can identify and separate noise (players not belonging to any specific line) from the actual clusters, ensuring cleaner clustering results12. Furthermore, DBSCAN is particularly effective for clustering spatial data, which is common in soccer where player positions are represented as coordinates on the field12. However, it will not be used since it is not good if we know the amount of clusters there should be, as DBSCAN determines the number of clusters dynamically based on density, this could cause DBSCAN to create more than 3 lines (clusters) or less than 3.

Clustering algorithms typically form clusters based on the proximity of data points to one another. For example, as shown in image 2, clustering applied to a singular passing instance in the given data results in clusters of players based purely on spatial proximity. However, these clusters fail to show the vertical line clusters we aim to create. While some algorithms, such as DBSCAN and ACDC, support arbitrary-shaped clusters, an alternative approach to achieve the desired line clusters is to introduce a weight to the x-coordinate of the data.

Image 2: Example of a snapshot in a game with K-means cluster with 3 clusters (blue, green, red) (goalkeeper being its own cluster with yellow)

The input for the clustering algorithm consists of the x and y coordinates of the team performing the pass in a given instance. By applying a weight to the x-coordinate, the emphasis shifts to the y-coordinate. Conceptually, this stretches the positions of players along the x-axis, increasing the importance of their y-coordinate proximity. Consequently, the players align more closely in vertical clusters. For instance, image 3 illustrates the impact of a weight factor of 4 (i.e., the x-coordinate multiplied by 4). The comparison between images 2 and 3 highlights the significance of this weighting. Without weighting (as in image 2), clusters reflect mere spatial proximity. In contrast, with weighting applied (image 3), the clusters form vertical lines that align better with the expected results, as the y-coordinate becomes more influential. The selection of the weight parameter will be addressed in the results section.

Given the dynamic nature of football and the variety of tactical approaches, defining precise line formations and assigning players to these lines can be inherently subjective. Ambiguities often arise when players occupy positions that could reasonably belong to more than 1 cluster. For example, some players may sit precisely between two clusters, leading one model to assign them to one cluster and another model to a different cluster. Since there is no universally "correct" clustering, even football analysts and coaches might disagree depending on their tactical perspectives, game context, or the player's movements.

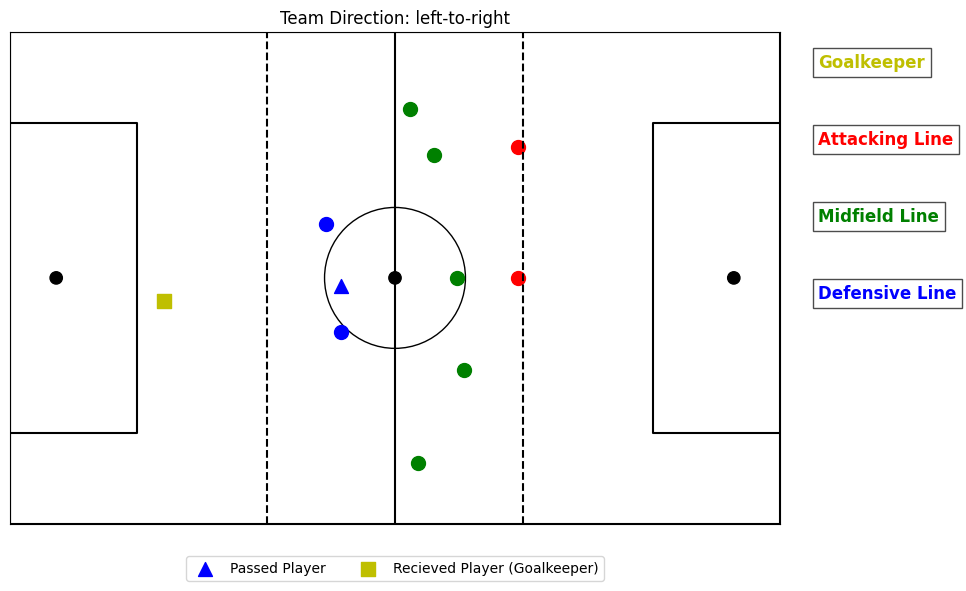


Image 3: Example of a snapshot in a game with K-means cluster with 3 clusters (blue, green, red) (goalkeeper being its own cluster with yellow) and added x-coordinate weight at weight = 4

To address this ambiguity, a voting system was implemented. This system takes results from three different clustering algorithms to resolve conflicts in cluster assignments for ambiguous passing events. For example, in certain unconventional player formations, one algorithm (e.g., GMM) might assign a player to the defensive line cluster, while another (e.g., ACDC) assigns the same player to the midfield line cluster. Image 4 shows clusters determined by the K-means algorithm, assigning the passed player (represented by a green triangle) to the midfield line. Meanwhile, image 5 shows the ACDC model assigning the same player to the defensive line. These discrepancies arise due to the inherent differences in the models and their strengths.

The voting system works by using the results of a third algorithm to act as a tiebreaker, in case of a tie. For example, in ambiguous cases where one algorithm assigns a player to one cluster and another assigns them to a different cluster (as demonstrated in image 4 & 5), the third algorithm determines the final cluster assignment. While it’s assumed that most clustering algorithms will agree in straightforward cases. However, in ambiguous scenarios, the voting mechanism provides a more robust justification for the cluster assignments. For instance, image 6 displays clusters determined by a third algorithm (GMM), where the passed player is assigned to the midfield line. With two out of three models “agreeing” that the passed player should be part of the midfield line cluster, the final decision is that the player belongs to the midfield line. Moreover, the voting system acts as a safeguard when one model performs poorly in certain passing instances due to unforeseen reasons. In such cases, the other two models compensate, ensuring more reliable results.

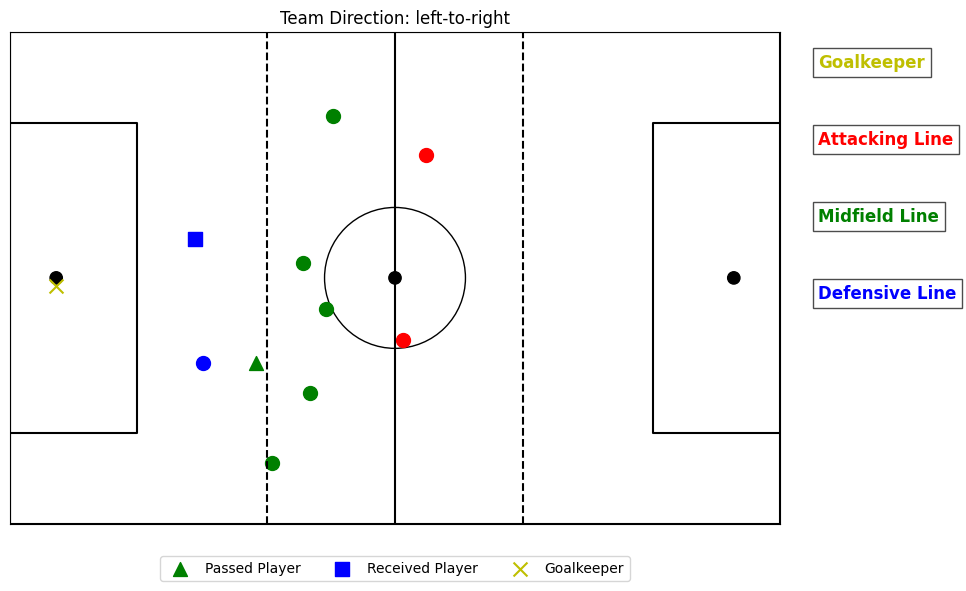
Image 4: Example of a snapshot in a game with K-means cluster with 3 clusters (blue, green, red) (goalkeeper being its own cluster with yellow) at weight = 4

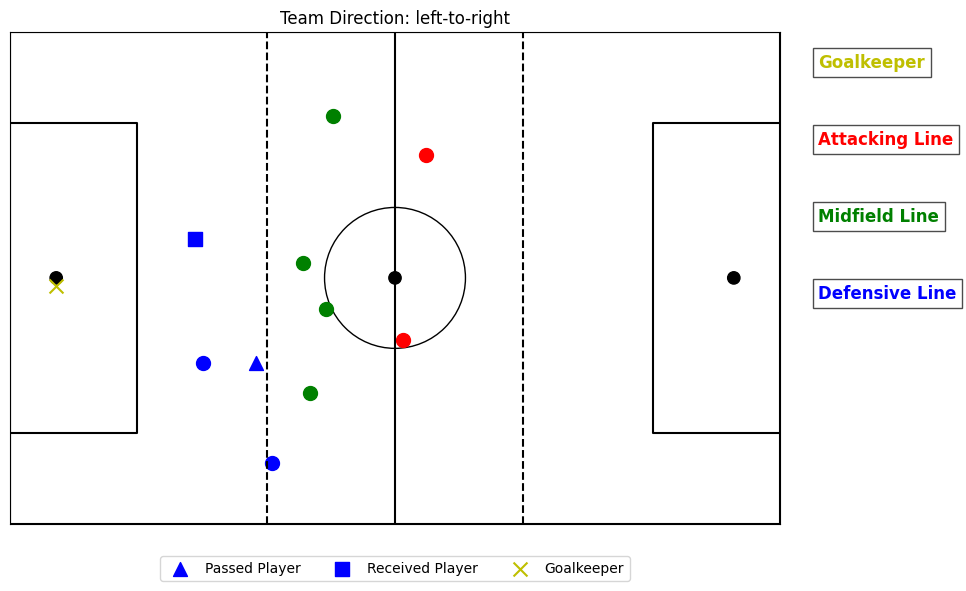
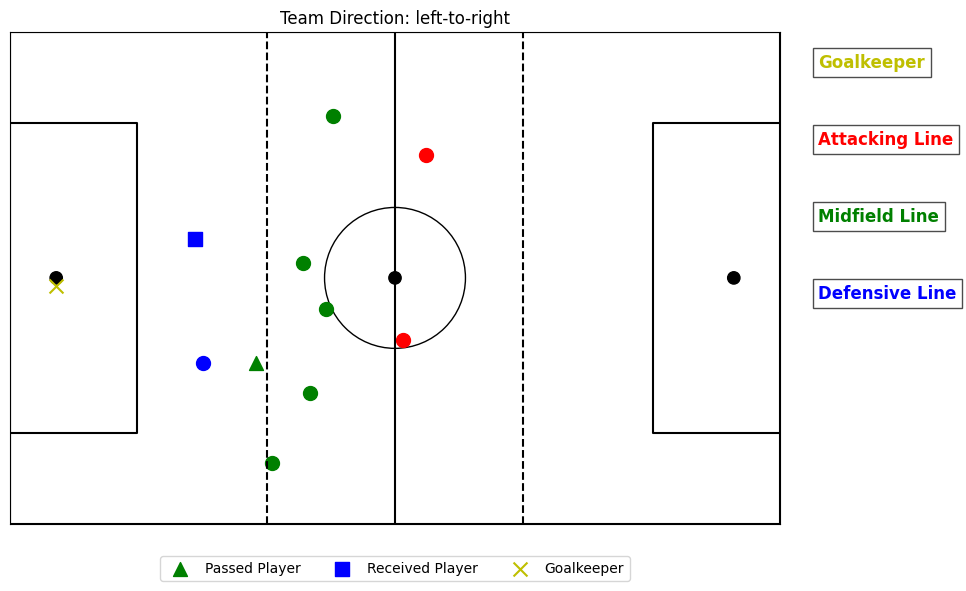
Image 5: Example of a snapshot in a game with ACDC cluster with 3 clusters (blue, green, red) (goalkeeper being its own cluster with yellow) at weight = 4

Image 6: Example of a snapshot in a game with GMM cluster with 3 clusters (blue, green, red) (goalkeeper being its own cluster with yellow) at weight = 4

Finally, to determine the starting and ending zones of each pass, the pitch is divided into three equal sections: defensive, midfield, and attacking zones. The x-coordinate of the player passing the ball is input into a function to determine which zone the pass originated from, and the same process is applied to the receiver's x-coordinate to determine the destination zone.

1. **Academic output**

The selection of the weight parameter for the models was determined by testing a range of values. However, as previously mentioned, there are no "true" labels for the line clusters, making it challenging to evaluate the accuracy of the models' results. The inherently unsupervised nature of this task means there is no consistent or objective metric to measure the models' performance. While the goal was to achieve more vertically aligned clusters, this does not imply that a very large weight should be chosen. Although emphasizing the y-coordinate is essential, the x-coordinate must still retain some significance in shaping the clusters due to the nature of football formations. Striking the right balance between these two dimensions is crucial for meaningful results.

There are several standard evaluation metrics for clustering that can provide insights into the quality of the clusters. The Silhouette Score is one such metric, which measures how similar a data point is to its assigned cluster compared to other clusters, focusing on both the compactness and separation of the clusters13. The score ranges from -1 (indicating misclassification) to +1 (indicating well-defined clustering). A high score, close to 1, suggests that the clusters are well-separated and that the data points within each cluster are cohesive. Conversely, a low or negative score indicates poor clustering, such as overlapping clusters or misclassified points14. In the context of this study, where we aim to identify distinct line clusters based on player positioning, a high Silhouette Score would indicate that the clustering effectively captures natural separations in line clusters.

The Davies-Bouldin Index measures the average similarity of each cluster to its most similar cluster, taking into account both compactness and separation15. Specifically, it assesses how tightly the data points within a cluster are grouped (compactness) and how distinct the clusters are from each other (separation). In this context, it’s expected the clusters to be compact, meaning that players within the same line cluster all belong to that cluster. Additionally, the clusters should be well-separated, meaning that players from different lines should be clearly distinguishable from other lines16.

However, it is important to note that these methods prioritize general clustering properties, rather than focusing on the task-specific objective of forming meaningful vertical lines in the context of football.

While the previously discussed metrics provide some insight into the quality of the clusters, they do not necessarily confirm the accuracy of the models' results. To address this, in addition to the standard metrics, a solution was proposed during a meeting with my supervisor: manual inspection of several known passing instances where players are positioned between clusters, since the main errors and discrepancies between models typically arise in situations where players are positioned between clusters or in unconventional formations. The domain knowledge of my supervisor will play a crucial role in determining whether the clusters align with the expected player line.

To determine the optimal weight, I tested several values, focusing on two specific scenarios where the weight has the greatest impact. The first scenario involved players being very close to one another, and the second involved players being spread far apart along the y-axis. In the first case, it was essential to ensure that the clusters remained vertical and logically consistent, even when players were in close proximity, meaning their x-coordinates would not differ significantly. In the second case, I wanted to ensure that the x-coordinate was scaled appropriately so that players, though spread out along the y-axis, would still form distinct vertical clusters. Image 7 shows the results of six different weight values using the K-means algorithm in a scenario where players are close together. The image demonstrates that smaller weights, such as 1 and 2, give insufficient weight to the x-coordinate, causing the clusters to lose their vertical alignment. The lowest optimal weight, which produces the desired vertical clusters while retaining some proximity information between teammates, was found to be 3.

Image 8 presents the results of six different weight values using the K-means algorithm in a scenario where the players are spread far apart along the y-axis. Similar to the previous image, the two lowest weights fail to produce vertical clusters and can be disregarded. The weight of 3, which yielded good results in the previous example, shows that when the midfield line is spread across a larger distance, the player positioned at the top of the field, who should be part of the midfield line, is instead incorrectly clustered in the defensive line. Increasing the weight to 4 addresses this issue, as it compensates for the spread of the midfield line and correctly assigns players to the midfield cluster. Thus, the lowest weight that successfully produces vertical clusters in this scenario is determined to be 4. Based on these two extreme examples and results, the decision was made to set the weight of the x-coordinate to 4.

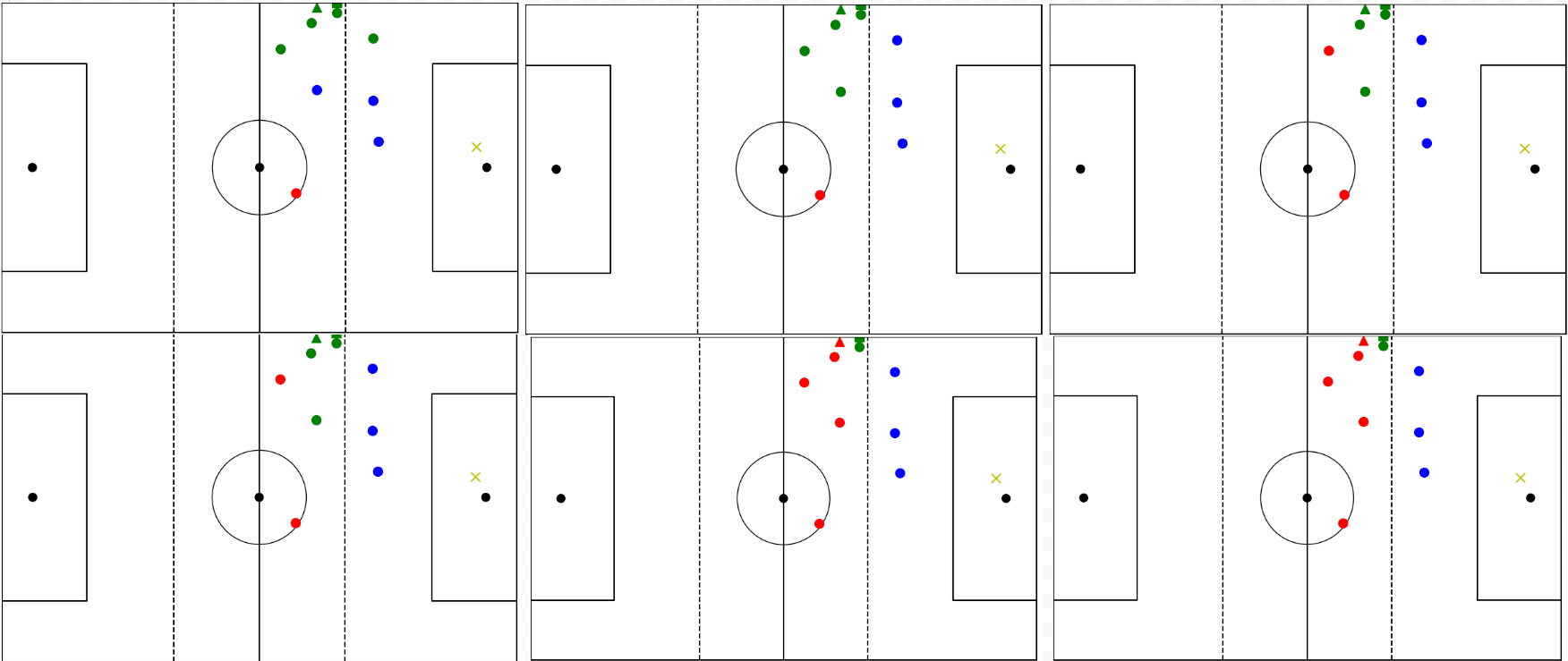
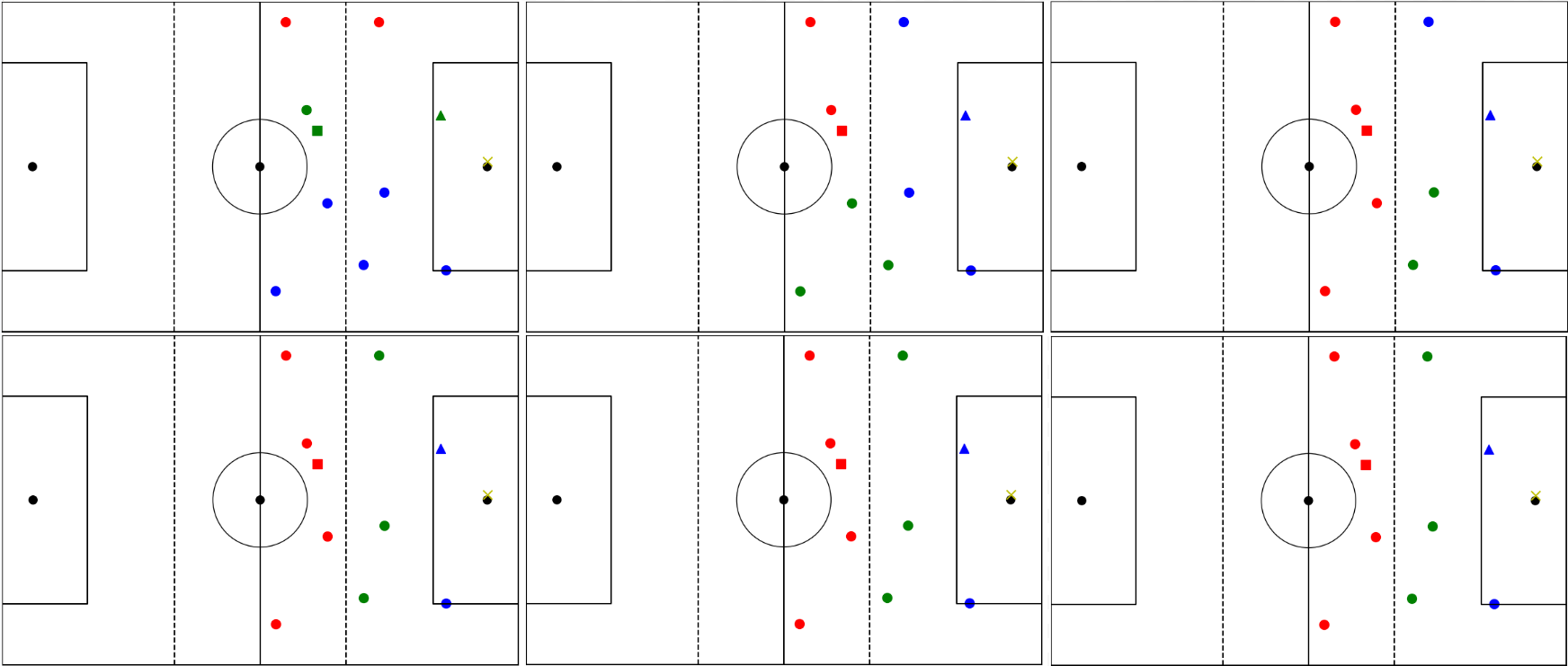
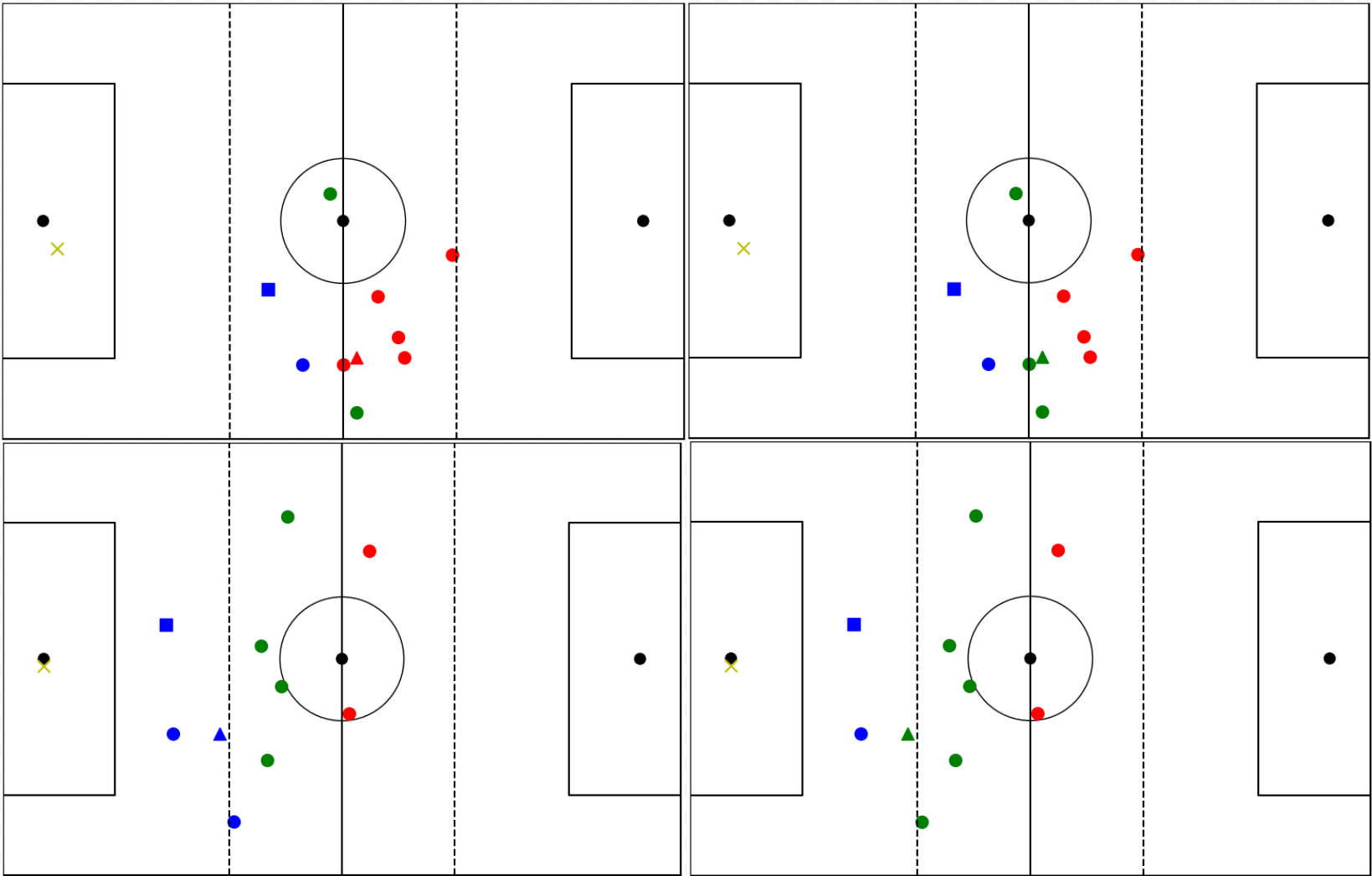
Image 7: Results of K-means with weight {1,2,3,4,5,6} respectively from top left to bottom right

Image 8: Results of K-means with weight {1,2,3,4,5,6} respectively from top left to bottom right

As previously mentioned, the final output after setting the weights and applying the model voting system will be manually checked, particularly for passing instances where players are positioned between clusters or in unconventional positions that may cause clustering challenges. The evaluation of "simple" passing instances, such as those shown in image 2 and image 3, is straightforward due to their simplicity and assumed to be correct. Image 9 presents two examples where the models initially misclassified players, but these misclassifications were corrected through the voting system. On the left side of the image, two instances are shown where the passed player was incorrectly labeled by a model. However, thanks to the voting system, these misclassifications were corrected, and the updated, corrected results are displayed on the right side of the image.

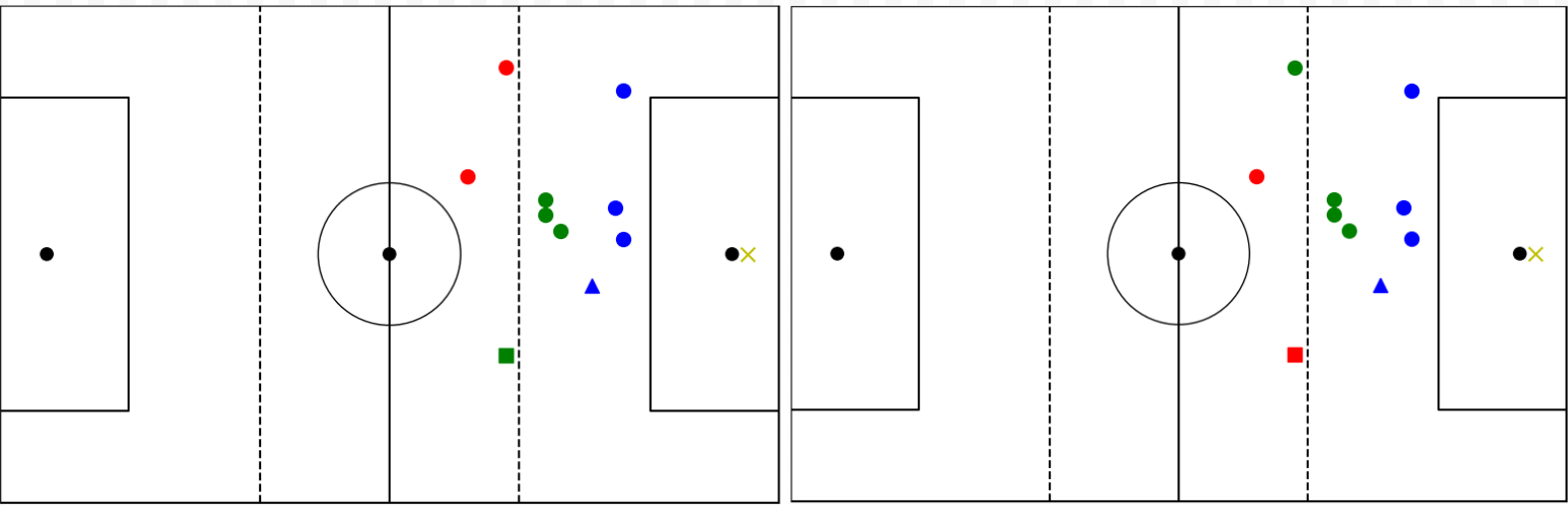
Image 9: Results of wrong cluster results on left, correct models after voting on right 

Additionally, when examining the clustering metric results shown in Table 3, the clusters in the top-left, which we know are incorrectly clustered, give a low silhouette score. This indicates poor cluster cohesion and separation, which aligns with our observations. The DBI index for these clusters is 1.17, which is not particularly low, suggesting that the clusters are not well separated. In comparison, the correctly clustered counterpart, shown in the top-right, demonstrates a significantly higher silhouette score, indicating better cohesion and separation. Additionally, a lower DBI score is observed, signifying better cluster separation and compactness. For the second set of images shown in Image 9 (bottom-left and bottom-right), nearly identical silhouette scores are recorded. However, the DBI index differs slightly, with the “correct” image (bottom-right) displaying a marginally lower DBI score, indicating slightly better cluster separation and compactness.

| **Image** | **Top-Left** | **Top-Right** | **Bottom-Left** | **Bottom-Right** |
| --- | --- | --- | --- | --- |
| **Silhouette Score** | 0.093 | 0.368 | 0.434 | 0.466 |
| **DBI Index** | 1.17 | 0.936 | 0.615 | 0.522 |

Table 3: Silhouette score and SBI index of clusters presented in image 9

These results show that for big mistakes the clustering metric does indeed reflect the poor clusters with bad scores and good clusters with good scores. It should be noted that for the passing instances in which it was known that players are positioned between clusters or in unconventional positions which might be difficult to cluster, there is practically no significance difference in scores. This further underlines the difficulty of evaluating the results of the models in these difficult passing instances. Furthermore, there were no instances where individual models performed better in terms of the silhouette or DBI scores compared to the final voted results. For instance, the passing instance displayed on the right side of Image 10 was manually labeled as correct by my supervisor and intuitively makes sense. However, when inspecting the clustering metric scores, the incorrectly clustered results shown on the left side of Image 10 present a slightly lower DBI score of 0.771 compared to the 0.783 of the "correct" results, suggesting that the "incorrect" results have better cluster separation and compactness. Similarly, with the silhouette scores, the incorrect results have a higher score of 0.461 compared to 0.414, again indicating better cluster structure and cohesion.

Image 10: Results of wrong cluster results on left, correct models after voting on right 

These results demonstrate that for significant errors, the clustering metrics effectively reflect the quality of the clusters. However, for passing instances where players are positioned between clusters or in unconventional positions, which may be challenging to cluster, there is no significant difference in the scores. This highlights the difficulty of evaluating the results of the models in these complex passing instances, especially in the quantitative sense. Which is a key reason why manual inspection of the results were conducted. Additionally, no instances were found where individual models outperformed the final voted results in terms of silhouette or DBI scores.

Additional manual checks were conducted for both setting the weight parameters and evaluating the overall results. However, to avoid redundancy and clutter, only the most significant examples that demonstrate notable differences were presented in this section. The results for the zones are always correct, as they are determined by directly analyzing the coordinates of the players.

1. **Discussion**

The primary discussion point is the need for better methods to evaluate the results. While it was manageable to manually check 20% of the data with a single dataset, this process is time-consuming. Additionally, standard clustering metrics have been shown to not always correlate with better results, mainly due to the specific nature of the clusters required for this task. A simple and effective solution would be to use already labeled data that has been validated by organizations such as FIFA. By comparing the results of these models to FIFA’s dataset, we can assess whether they produce similar outcomes. As FIFA is the largest governing body in football, overseeing events like the World Cup, it is assumed to have the most accurate and high-quality data.

Furthermore, the model could be improved by incorporating the dynamics of the game, particularly the events leading up to the pass. For example, by examining the change in position over the 4 seconds prior to the pass, it could be possible to determine the direction of movement for players positioned between clusters. In such cases, players could be assigned to the cluster they are moving toward, which would improve the clustering results. This approach would enhance the accuracy of the individual models. Additionally, this temporal aspect could extend beyond player positions to include other factors, such as the state of the game (e.g., whether it is a counter-attack), further refining the labeling of each player.

In the case of GMM, players can belong to more than one zone, with a probability assigned to each zone (soft assignment). This feature could be particularly useful when, for example, a right-back (RB) is transitioning from the defensive line to the midfield. During the middle of the transition, the player might be labeled as part of the midfield, and when transitioning back, the same player might still be labeled as midfield. By using the probability assignments, we can set a threshold that allows the player to be assigned to their "native" line until a certain point. This approach is especially beneficial in situations where tactics involve high-standing wingbacks, who function almost like midfielders but are still considered part of the defensive line. While this method may require significant tuning to get the desired results, it provides a useful solution for situations where it is unclear to which cluster a player should belong. However, the time required for tuning the probability thresholds might be a limitation.

Additionally, it is important to consider the use of potentially more advanced algorithms. While the current data and clusters are relatively simple and not high in dimensionality, more complex algorithms could offer more effective ways to cluster the players. Exploring these algorithms may provide better results and more accurate clustering, especially in cases where the simple models might struggle to capture nuances in player positioning.

An alternative approach would be to consider dynamic weights rather than fixed ones. As demonstrated in certain instances where players were either very close or very far from each other, it is possible to implement a system where the weight adjusts based on the distance between players. This dynamic weight system could enhance the model’s ability to better account for varying player distances. For example, if a model struggles to cluster players correctly, simply adjusting the weight might allow the players to be clustered accurately.

The inclusion of more models could be beneficial, especially with the existing voting system in place. By incorporating more complex models and increasing the number of votes, it is expected that the average accuracy of the results will improve. However, this leads back to the challenge of not being able to directly measure accuracy, as it remains uncertain whether adding more models will necessarily lead to better results without being able to prove it.

Overall, the results were very good when manually inspecting the clusters. While it was challenging to directly prove the accuracy through traditional methods such as accuracy or other standard metrics, many instances were straightforward. The primary difficulty, however, stemmed from the more complex cases, as demonstrated in the examples displayed in the results section. But these also proved to be effectively tackled by the clustering models and voting system.

### a

1. **Reference**
2. Zhou, D., Chen, G., & Xu, F. (2022). Application of deep learning technology in strength training of football players and field line detection of football robots. Frontiers in Neurorobotics, 16, 867028. https://doi.org/10.3389/fnbot.2022.867028
3. Catapult. (n.d.). Performance analysis in football. Catapult. Retrieved from https://www.catapult.com/blog/performance-analysis-in-football
4. AnalyiSport. (n.d.). Performance analysis in football. AnalyiSport. Retrieved from https://analyisport.com/performance-analysis-in-football/
5. Hudl. (n.d.). Performance analysis in football. Hudl. Retrieved from https://www.hudl.com/blog/performance-analysis-in-football
6. Future Machine Learning. (n.d.). Role of clustering algorithms in unsupervised learning. Future Machine Learning. Retrieved from https://futuremachinelearning.org/role-of-clustering-algorithms-in-unsupervised-learning/
7. Aaronm830. (2021). K-means clustering of progressive passes for football analysis. GitHub. Retrieved from https://github.com/aaronm830/K-Means-Clustering-of-Progressive-Passes-for-Football-Analysis
8. Analytics Vidhya. (2021). K-mean: Getting the optimal number of clusters. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/
9. Springer. (2022). Lecture notes in computer science: Vol. 13624. In F. Rossi & A. Farioli (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2022 (pp. 65-75). Springer. https://doi.org/10.1007/978-981-97-2136-8\_6
10. MathWorks. (n.d.). Cluster Gaussian mixture data using soft clustering. MathWorks. Retrieved from https://www.mathworks.com/help/stats/cluster-gaussian-mixture-data-using-soft-clustering.html
11. Springer. (2020). Lecture notes in computer science: Vol. 12345. In E. Hüllermeier & M. Studer (Eds.), Advances in Computational Intelligence (pp. 275-284). Springer. https://doi.org/10.1007/978-3-030-61705-9\_35
12. Stats Perform. (n.d.). Identifying patterns in build-up play using clustering. Stats Perform. Retrieved from https://www.statsperform.com/resource/identifying-patterns-in-build-up-play-using-clustering/
13. Chiebukar. (2020). Footballers' ability clustering. GitHub. Retrieved from https://github.com/Chiebukar/Footballers\_Ability\_Clustering
14. Scikit-learn. (n.d.). Silhouette score. Scikit-learn Documentation. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette\_score.html
15. Future Machine Learning. (n.d.). Understanding silhouette score: A key metric for clustering. Future Machine Learning. Retrieved from https://futuremachinelearning.org/understanding-silhouette-score-a-key-metric-for-clustering/
16. Scikit-learn. (n.d.). Davies-Bouldin score. Scikit-learn Documentation. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies\_bouldin\_score.html
17. PyShark. (n.d.). Davies-Bouldin index for k-means clustering evaluation in Python. PyShark. Retrieved from https://pyshark.com/davies-bouldin-index-for-k-means-clustering-evaluation-in-python/

Hristovski, R. (2017). Unpredictability in competitive environments. In Advances in Sport and Exercise Psychology Measurement (pp. 173-187). ResearchGate. Retrieved from <https://www.researchgate.net/profile/Robert-Hristovski/publication/323700817_Unpredictability_in_Competitive_Environments/links/5aa6786c458515b3e5c134ed/Unpredictability-in-Competitive-Environments.pdf#page=173>

Uppsala University. (2024). Unsupervised learning for clustering football players' movements: DIVA portal. Retrieved from https://uu.diva-portal.org/smash/get/diva2:1833447/FULLTEXT01.pdf

**Personal Development Section**

1. **Personal Evaluation**

Throughout my internship, I focused on creating a dynamic line clustering model. In addition to this project, I set several personal development goals and learned many new skills. I learned about how companies store data in cloud services like AWS, how to extract the necessary data for specific tasks, and how to navigate these processes in a professional environment. I gained hands-on experience with clustering algorithms, particularly learning how to use models beyond the simple k-means that I was familiar with. Applying these algorithms to football positional data in new scenarios helped me refine my understanding and approach to machine learning techniques. The internship also improved my coding skills. Some of the challenges I faced were related to writing the necessary code to achieve specific goals, which improved my problem-solving ability and programming techniques. I also learned how to communicate my problems and solutions more effectively with my supervisor, which streamlined our collaboration. Time management was another key area of improvement. By implementing logs and a weekly schedule, I was able to track progress and ensure that the project stayed on track. This organizational tool greatly enhanced my ability to deliver the project on time. Moreover, my domain knowledge in football analytics deepened as I explored how data can support various aspects of the game, from training to performance analysis. This internship showed me how simple data could be transformed into insightful analytical tools, providing valuable perspectives on football strategy and player performance. Finally, critical thinking played a central role in solving the challenges I encountered. Overall, this internship allowed me to grow both technically and personally, equipping me with valuable skills for future projects.

Something that went well during the internship was the dynamic line detection system, which was the main task of the internship, performed excellently. The system met the company's requirements and delivered precise results with no errors. I was able to process the data without excluding any significant portions (except for missing data). Initially, I faced challenges with including goalkeepers in the clustering algorithm, as they didn’t fit well within any team line. However, after adjustments, the system was able to handle the goalkeeper's data without causing problems such as clustering them with defenders or creating an unnecessary cluster for them. Communication with my supervisor was smooth throughout the internship. He understood my needs and provided clear guidance, which helped avoid any major roadblocks.

Something that didn't go well during the internship is the fact that the model is not fully generalizable, as it cannot seamlessly work with new match datasets right out of the box. One of the key challenges was identifying the goalkeeper, as this label was missing from the dataset. To address this, I had to manually identify the goalkeeper based on their position on the field, which was time-consuming. If the dataset had included labels for the goalkeeper, it would have been much easier to implement the model. Additionally, determining the pitch size from the data itself was required, as the x and y coordinates of players helped deduce the pitch boundaries. Since pitch sizes can vary, this created potential inaccuracies when defining zones. Another challenge was that incorporating goalkeepers into the model took significantly more time than anticipated. Although they make up a small portion of the data, their inclusion created a distinct scenario that required additional effort to adjust the model. Despite these setbacks, the final model worked well, although these aspects did extend the development timeline.

Obstacles I faced over the course of the internship were the primary challenges I had dealing with the goalkeepers, as explained earlier. To address this, I first developed the system without considering the goalkeeper, which allowed me to create the general pipeline for 90% of the data. Once this was completed, I revisited the goalkeeper issue and incorporated special handling rules, such as creating conditional statements for when the goalkeeper was involved in passing or receiving the ball. This approach ensured that I could include goalkeepers in the results without affecting the rest of the system. My supervisor's help was invaluable during this process. His prior experience with similar projects and knowledge of the data helped me navigate the complexities of working with goalkeepers and other data-specific challenges. Another significant obstacle was evaluating the models, as the lack of true labels made it difficult to assess the accuracy of the results. Traditional metrics like silhouette scores and DBI were not very helpful in the complex scenarios where accurate clustering was crucial. To resolve this, I worked closely with my supervisor to manually review instances that were particularly challenging to cluster correctly. While we couldn't check every single difficult instance, we thoroughly reviewed a large portion of the data and were confident that our manual checks would generalize to the rest of the dataset.

My main learning objectives were to explore the use of coordinate data, enhance my Python skills, learn how to work with AWS and databases, and gain experience in a practical workplace setting. The use of coordinate data was a new area for me, and throughout my internship, I gained hands-on experience with positional data, particularly in the context of football. I explored clustering algorithms and learned how to handle data in a dynamic way. I feel more confident in my ability to work with coordinate data and plan to continue experimenting with different models to improve my approach. I was able to take my Python coding skills to the next level, particularly in the context of machine learning and data processing. I had to solve problems that required thinking outside of the box and adapting my coding techniques to different scenarios, such as working with clustering algorithms like GMM. I plan to continue enhancing my coding skills by continuing the use of Python for all future projects. I had limited prior experience with cloud platforms like AWS, but during the internship, I learned how to interact with databases in a different format. This knowledge will be valuable as I look to expand my expertise in cloud-based systems. To continue working on this, I plan to take online courses and tutorials on AWS and cloud computing to gain more proficiency. I find this especially useful as this internship showed me how companies will rely on cloud platforms like AWS, and if I possess the skills to work the AWS I feel like I can become a more desirable candidate when looking for a permanent workplace. Working at Forward Football B.V. allowed me to gain real-world experience and develop my communication skills. Interacting with colleagues and discussing technical concepts with my supervisor and colleagues helped me refine how I explain complex issues in simpler terms. Going forward, I plan to focus on improving my teamwork and communication skills further, especially in a corporate environment. My domain knowledge in football, while not easily measurable, was significantly enhanced. A key moment for me was when my supervisor explained how understanding passing lanes could be used to improve training and player development, particularly by measuring pass accuracy. This insight helped me connect my technical work to its real-world applications. To deepen my knowledge, I plan to stay updated with research in the field, as this will allow me to apply both my technical and domain knowledge more effectively and continue work in the field of AI and football.

My main realization of a new talent, which I’m pleased with, is that I often struggle with procrastination and laziness. However, during this internship, I had the opportunity to work on tasks that I genuinely enjoyed. As a result, I found myself being more engaged and focused, and I didn’t experience the same tendencies to procrastinate. In fact, I often felt ahead of my peers when we discussed our internship experiences

1. **Optimization points**

Possible optimization points that can be of use for the organization based on what I have learned during the internship is having additional labels in the data could have helped resolve some of the issues I encountered, such as identifying the goalkeeper. With access to these 'true' labels, the task would have been more straightforward. However, given the nature of the task and the data, it is difficult to establish a definitive 'truth' regarding what is correct. That said, having a small labeled dataset, manually annotated by experts, would have been valuable for quantitatively assessing the effectiveness of my models and results, using accuracy as a metric instead of relying solely on clustering evaluation metrics

1. **Future career perspective**

For future career perspectives some knowledge and skills I was able to use is the knowledge gained from my machine learning courses was particularly useful, as it allowed me to understand how clustering algorithms function, which made it easier to identify mistakes and come up with solutions, such as implementing a weighted x-coordinate. Additionally, the extensive coding experience I gained over the past five years at Tilburg University was invaluable. It helped me efficiently tackle tasks and create functions that ensured I could obtain the correct results

My experience with the connection between my studies and practice in the internship was smooth. Since we had already covered clustering algorithms in our studies, it was beneficial to know what to expect and how to work with methods like k-means. A lot of the coding was based on the general knowledge I gained during my studies, but, as with most coding tasks, it involved a combination of prior experience and problem-solving to determine how to approach and solve each challenge

For my future career perspectives I’ve gotten some great insight during this internship. I’ve always known that I wanted to work in a corporate setting, focusing on AI and data. Additionally, I wanted to work in a field that I am passionate about and have significant domain knowledge in, specifically football and gaming. However, knowing you want to work in a particular field or position is different from understanding the day-to-day tasks and challenges you’ll encounter. This internship gave me valuable insight into the type of work available in this specific field, and I truly enjoyed it.

This internship definitely contributed to my future career perspectives. I always knew I wanted to work in a field like this, but I wasn't sure about the specific tasks or projects I'd be involved in. This internship gave me a clear picture of the kind of work I’d be doing daily, and I thoroughly enjoyed it. It also sparked my curiosity about the other possibilities within AI and football. This experience confirmed my desire to continue in this field, not only because I enjoyed it, but also because my domain knowledge played a crucial role in the project, and I see a growing demand for this type of work.

**Appendix**

**Organization Details**

The organization where I will be conducting my master’s fall external internship is named “Forward Football B.V.”. The Address of Forward Football B.V. is Van Marwijk, Kooystraat 10-A, 1114 AG Amsterdam-Duivendrecht. The website of Forward Football B.V. is “<https://forward.football/>”. The time span of the external internship will be 420 hours, starting officially from 25 September until 15 of January, the working hours per week is up until this moment not decided however it will likely be equally spread across all the weeks contained in the external internship period.

Forward Football B.V. is a predominantly football focused company, however, they also provide products for basketball. They provide many different services and products targeted for football, these products/services range from the LPM player tracking to the Tech.Football. These products can be used in a wide variety of ways but the main use would predominantly lie in the use to track players progress and collect data for further analysis during training and matches.

The supervisor within Football Forward B.V. will be Ryan Stepfner. His contact information is “<https://www.linkedin.com/in/ryan-stepfner/>” and “[ryan@forward.football](mailto:ryan@forward.football)”. My supervisors background experience and education as listed on his linkedin are “Data Scientist, Full Time at Forward Football”, “Data Analyst, Full Time at WithYouWithMe”, “Masters of Data Science and Computational Intelligence Specialization at the University of Newcastle”, “Bachelor of Computer Science at Western Sydney University”. Additionally his expertise as listed on his linkedin lie in “Data Science, Machine Learning, Data Analysis, Python, R”.

The work environment will mostly be independent, as I’m the only person working on this project. However, I do have the help of my supervisor (Ryan) who is also the data scientist I will mainly have contact with and can ask for feedback or suggestions when I face challenges. As we have agreed to me coming to the office once per week on Thursday I expect the main communication and feedback to be done on that day, this allows me to get some more help and work towards progressing some of my goals and personal development during this time.

**Research Description**

The research of this external internship will be about line detection. The objective of the

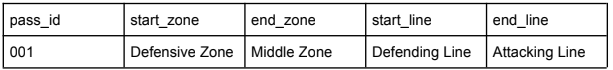
project will be to “detect the zones and lines where a pass begins and ends”. It will involve identifying the starting and ending zones and lines for each pass. The desired output is an enhancement to the dataset, adding parameters such as:

Table 1: Example of the output showing the desired parameters to enhance the dataset

The football pitch is divided into three distinct zones: Defending Zone, Middle Zone, Attacking Zone. These zones are static and depend on the pitch size, which varies from 100 to 105 meters in length and 64 to 68 meters in width. Players are organized into three lines: Defensive Line, Middle Line, Attacking Line. These lines can change as players move during the game and can be defined in two ways:

● Fixed: Based on units (e.g., a defense line consisting of defenders like (LeftBack)LB, (CenterBack)CB, (RightBack)RB,

etc.). This is the simplest solution.

● Dynamic: Based on the player's position, irrespective of the unit. This is the preferred

approach.

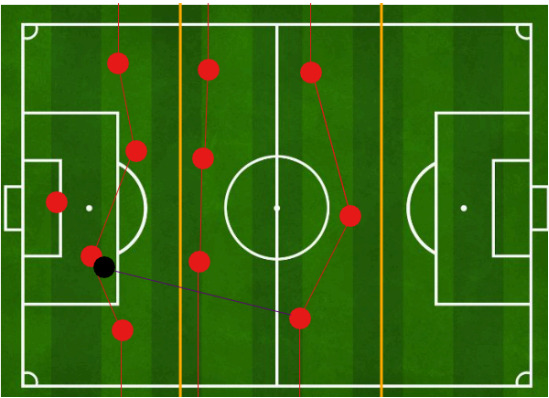
The image below shows these lines.

Image 1: Example of a snapshot in a game where the defensive, midfield and attack lines (red) are displayed (same order in image as in description)

The current plan involves displaying this data and adding a filter to sort and view passes by zones and lines with options to toggle visible lines. Such as;



Image 2: Example of passing line (yellow) displayed on a match representation alongside defensive, midfield and attacking lines (blue) being represented

The methods for this research project will include using the data provided by Forward Football B.V., the data include recorded data from previous matches recorded with their instruments. This data will include different labels that can be used to determine the passing lines and positional lines, the most important part of the data is the 5hz positional xy-coordinate data of all the players and the football. This will allow for the classification of the positional lines (defense, midfield, attack), which can then determine when a pass is given and from which line and received in which line.

The schedule of the research project will be created in a flexible manner, as this allows for changes when intermediate deadlines are not met or exceeded earlier then intended. However a rough draft of the schedule of look like the following:

* september 25 = start
* mid october = complete literary research for scientific methodology, collect all data (cleaned and loaded, if needed, ready for use), select machine learning models for implementation
* mid november = get early results for feedback
* mid/end december = start writing portfolio
* 15 january = end

The academic learning objectives that I wish to complete during this internship include:

* Gain experience in cleaning, preparing and handling positional xy-coordinate datasets of players and the ball
* Learn handling missing or noisy data and optimizing data in machine learning
* Develop an algorithm to detect football match events (e.g., passes, positional lines) from real-world data
* Create a dynamic line detection system based on player positions instead of fixed predetermined formations
* Improve collaboration skills by working independently with feedback from my supervisor

**Personal Development**

As I did my bachelor thesis in the domain of football and AI, I specifically looked for an internship that would allow me to further my personal development in both the domain of football and AI. Football Forward B.V. allows me to further develop knowledge and skills in both these domains.

First, the use of coordinate data is a new endeavor for me. Using this in combination with machine learning will allow me to develop the skills needed to become knowledgeable on how to use this type of data in future tasks. Additionally, as of the writing of this report I am unsure of which model to use for this specific task, however, the use of a model that I have not used yet could be a reality, this will allow me to further my toolbelt in the model I can use and have used. Both these points will also allow me to develop my python skills to the next step, as this is the first practical task done for a company. Furthermore, the use of AWS and databases in a new format will be another skill to learn, as I personally have never use this before

Additionally I believe that experience in a practical workplace like the office of Forward Football B.V. will allow me to learn social skills, learning to communicate in a different environment and with a different wanted result. Previously I would usually have to communicate with different group members to coordinate how to work together as a group and finish a project. Now however, I will need to properly communicate with my supervisor to understand what the product is that I need to create, this will be done by initially going to the office once per week for the first month, as both me and the supervisor agreed that the first month is crucial as I will developing my plans and models, and open and clear communication is crucial as this will ensure the provided output product will be as the supervisor wished for. Domain knowledge on football is difficult to quantify and measure, however, an example of a way to further my domain knowledge in football came after my first meeting with my supervisor, who explained the practical reasons why the passing lanes that I’m trying to quantify can help achieve the FIFA standard and also help football trainers find players that need further personal development, with passing lanes specifically the expected pass accuracy and actual pass accuracy are 2 ways in which the passing ability can be measured.

Meeting and activities with the supervisor that are planned for currently only coming to the office once per week, as this allows me to ask direct questions and have the supervisor give me examples or explain why certain things are done in a certain way.

**References**

Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 338-349.

Forcher, L., Altmann, S., Forcher, L., Jekauc, D., & Kempe, M. (2022). The use of player tracking data to analyze defensive play in professional soccer-A scoping review. *International Journal of Sports Science & Coaching*, *17*(6), 1567-1592.

Zejun, W. A. N. G., & Songhui, Y. O. U. (2021). Soccer tactical analysis methods and development tendency based on positional data under the background of big data. *Journal of Shanghai University of Sport*, *45*(9), 60-69.

Moura, F. A., van Emmerik, R. E., Santana, J. E., Martins, L. E. B., Barros, R. M. L. D., & Cunha, S. A. (2016). Coordination analysis of players’ distribution in football using cross-correlation and vector coding techniques. *Journal of sports sciences*, *34*(24), 2224-2232.

Low, B., Coutinho, D., Gonçalves, B., Rein, R., Memmert, D., & Sampaio, J. (2020). A systematic review of collective tactical behaviours in football using positional data. *Sports Medicine*, *50*, 343-385.

Steiner, S., Rauh, S., Rumo, M., Sonderegger, K., & Seiler, R. (2018). Using position data to estimate effects of perceptual features of play on passing decisions in soccer. *Current Issues in Sport Science (CISS)*, *3*, 009-009.

Malqui, J. L. S., Romero, N. M. L., Garcia, R., Alemdar, H., & Comba, J. L. (2019). How do soccer teams coordinate consecutive passes? A visual analytics system for analysing the complexity of passing sequences using soccer flow motifs. *Computers & Graphics*, *84*, 122-133.

Ahn, S., Ryu, D. W., & Lee, S. (2020). A machine learning-based approach for spatial estimation using the spatial features of coordinate information. *ISPRS International Journal of Geo-Information*, *9*(10), 587.