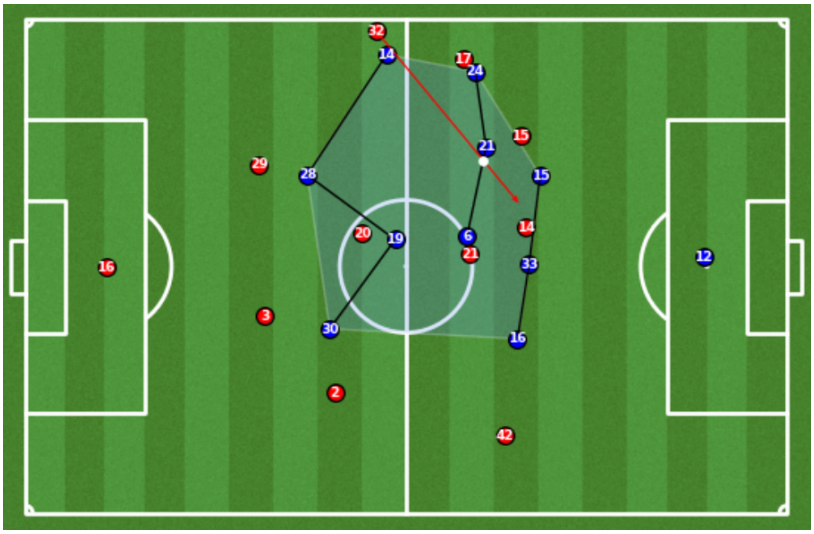
Beyond Pass Completion: Automating the Detection of Line-Breaking Passes and Introducing the 'Pass Advantage Score' in Association Football

submitted in partial fulfillment for the degree of master of science

Vladislav Manolov
14529157

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 30.06.2023



	UvA Supervisor	External Supervisor
Title, Name	Hongyun Liu	Ramon Dop
Affiliation	University of Amsterdam	AZ Alkmaar
Email	h.liu@uva.nl	r.dop@az.nl

ABSTRACT

Albeit their frequency, strategic importance, and significant contribution to goals in football, passes are subject to rudimentary post-match analysis and manual classification by data analysts. In particular, there is currently no existing system for the automated identification of line-breaking passes tailored to the needs of individual football clubs. This thesis introduces a pioneering system to fill this gap, automating line-breaking pass detection in association football. Specifically, it proposes an improved way for the professional Dutch football club, AZ Alkmaar, to enhance its assessment of line-breaking passes. The system incorporates synchronized tracking and event data into an unsupervised learning model, which uses the spatial arrangement of defending players to form distinct formation lines. Passes that intersect a formation line and simultaneously meet established domain-specific criteria are categorized as line-breaking. The findings of this study confirm the hypothesis that line-breaking passes lead to a higher goal probability (3.78%) compared to non-line-breaking passes (1.93%), subsequently significantly influencing game outcomes. The research further introduces the "Pass Advantage Score" (PAS) metric, which is calculated based on the zone in which a pass is received and the proximity of the ball receiver to the nearest defender. Thus, the study not only enhances understanding of key game dynamics but also proposes innovative tools for detailed player performance assessment and facilitating the process of post-match analysis.

KEYWORDS

football, line-breaking passes, clustering, formation lines

1 INTRODUCTION

The technological advancements in recent years have led to the inevitable 'modernization' of football, particularly in the domain of post-match analysis with the acquisition of positional and event data becoming ubiquitous in various professional leagues worldwide. Until recently, qualitative analysis on the basis of visual observations and interpretations was the standard for studying events and patterns that have transpired in football games. With the widespread adoption of event data, such as the number of passes or shots made by a team, the development of football statistics has significantly advanced, providing comprehensive summaries of games. However, traditional analysis dashboards quantify the highly sophisticated and disparate events, such as passes, in a binary pass completion metric (successful or not) that offer no insights into the type of passes being made during the game [3, 35]. This implies that regardless of whether the ball is passed back to the keeper to maintain possession or forward between the opposition lines, it is only rudimentary classified as successful or not. The latter are referred to as line-breaking or penetrative passes and are a major contributor in the build-up to a goal as a result of their offensive nature. In the scope of this thesis, a line-breaking pass is one that is successfully executed longitudinally up the field to a supporting teammate located between the opposition's formation lines.

The importance of progressive passes in football was eloquently articulated by Pep Guardiola, who is widely regarded as one of the most prominent coaches of his generation: *"You have to pass the ball with a clear intention, with the aim of making it into the opposition's goal."* [24]. With passes accounting for approximately 70% of all on-the-ball actions during a game of football [8], the statistical oversimplification of classifying passes simply as successful or not could result in a misleading understanding of a team's performance and neglect the sophisticated tactical patterns that have led up to them.

In reality, the free-flowing nature of the game and the often loosely-defined roles of the players themselves result in a continuous transition between formations and player roles. Furthermore, a team's spatial configuration and a player's deviation from their typical position on the pitch could be caused by a variety of unforeseen factors, such as the balance of play, the scoreline or the playing style of the opposition. As a result, when conducting a post-match analysis, managers and football analysts need to rely on video replays to manually examine the spatial positioning of the players on the field. This process could be time-consuming and unreliable since the complexity of football events allows for various interpretations of the same situation.

With players' movements being continuously tracked by multiple cameras in recent times, the aforementioned process can be automated through the utilization of the accumulated positional data and its appropriate visualization. This way we can obtain a thorough understanding of a player's positioning throughout any stage of the game and the overall team shape. However, to date, the use of a combination of tracking and event data for automatically classifying passes as line-breaking is lacking, particularly in the Dutch Eredivisie.

The goal of this thesis is to facilitate the process of post-match analysis by providing an additional layer of information toward a more comprehensive understanding of a team's passing performance through the automatic detection of line-breaking passes. In particular, the study's aim is to construct a 2D landscape of penetrative passing opportunities for a player in possession of the ball, emerging from the spatial alignment of opposing players in various attacking subphases in association football.

Considering the aforementioned research gap, this thesis offers the following contributions in the domain of post-match analysis:

(1) Synchronization of passes: The start and end locations of the pass from the manually annotated event data are synced with the spatial arrangements of the players of both teams from the auto-generated tracking data on a per-frame basis.

(2) Classifying players into formation lines: The players of the team not in possession of the ball are dynamically classified into formation lines by means of constrained K-Means clustering, subsequently allowing for an improvement in the interpretation of passing opportunities between the lines in various game scenarios.

(3) Automatic detection of line-breaking passes: Successful passes that are made behind an opposition line and such that advance the ball by at least 10% of the remaining pitch length between the start location of the event and the out-of-possession team's goal (referred to as the *'passable area'* in subsequent sections) are detected and visualized.

1.1 Research Question

Following the aforementioned problem that is to be addressed within this thesis, the following research question has been formed: *To what extent can the detection of line-breaking passes in association football be improved by the utilization of tracking and event data from football matches?*

1.2 Sub-Questions

(SRQ1) How can clustering techniques be applied to tracking data of spatio-temporal nature to classify football players into formation lines?
(SRQ2) How can auto-generated tracking data be synchronized with manually annotated event data for the purpose of improving the accuracy of pass locations?
(SRQ3) How can the determination of the intersection point between the ball's trajectory and a formation line be effectively used to identify whether the defensive, midfield, or attacking formation line of the out-of-possession team has been broken?
(SRQ4) What are the measures for evaluating the efficiency of line-breaking passes, and how can these data insights be utilized to assess individual players' passing performance?

2 RELATED WORK

The extensive proliferation of tracking and event data in football analytics in recent years has allowed research into a variety of data-driven problems to be conducted such as automatic classification of team formations, goal-scoring probability models, pitch control, dynamic analysis of team strategy, and expected passes [2, 6, 7, 13, 22, 27, 28, 30].

Albeit the growing interest in the usage of tracking and event data in the aforementioned domains, research on the automated classification of various pass types is still narrow and often limited to the utilization of event data only. Previous studies on passing in association football have predominantly focused on assessing their importance and benefit for the team through annotative analysis of passing patterns [9, 26]. Consequently, novel metrics have been proposed, including the risk of a pass, which represents the probability of a player executing a pass, and the reward of a pass, signifying the likelihood of it leading to a goal-scoring opportunity [25]. More recently, success probability models were adopted for the purpose of more precisely evaluating the passing ability of individual players. Such models offer an additional layer of information on top of conventional statistical methods which traditionally regard pass completion rate as the sole indicator of a player's skill [20, 34]. From a football analysis standpoint, such advancements facilitate the identification of teams' characteristic playing patterns [14, 17, 19]. Previous research in association football has also explored the concept of dominant regions, bearing a resemblance to a weighted Voronoi diagram, as delineated by Taki and Hasegawa [15].

The paucity of research on the topic of interest is particularly apparent in the Dutch Eredivisie. To date, research has been predominantly conducted on understanding team shapes through analysis of adopted formations in various stages of a game [7]. Consequently, this thesis focuses on dynamically classifying players into formation lines based on their spatial arrangement on the field and subsequently automatically detecting penetrative passes between those lines.

2.1 Clustering for Formation Analysis

In 2016, Bialkowski et al. labelled the constantly changing player positions as "the biggest issue" in dynamic football analysis [5]. Dynamic analysis of football templates is often achieved through the use of unsupervised learning techniques, such as hierarchical clustering [5, 23, 25, 28]. In the work of Bialkowski et al., the research team proposes the use of an expectation maximisation (EM) algorithm for the purpose of facilitating large-scale analysis on the topic of team formations in each match-half - a procedure similar to k-means clustering [11]. In football, K-means clustering works by partitioning the 10 outfield players into n clusters (usually 3 or 4), representing the number of formation lines. Each player is assigned to the cluster with the nearest mean. Contrary to the oversimplification caused by classifying only one formation per match half, this paper provides a more dynamic insight into the formational transitions of football teams by classifying formations at more frequent intervals.

Reliance on unsupervised learning in the domain of interest is also evident in the work of Shaw and Glickman who adopted a hierarchical agglomerative clustering approach in alliance with the Wasserstein metric for identifying each team's preferred offensive and defensive formations [16, 28]. This way the relative positioning of each team's players in and out of possession of the ball over successive time intervals can be measured.

Analysis of adjacency relationships of player's Voronoi regions is also present in modern-day formation analysis [10]. Narizuka and Yamazaki defined formations as an adjacency matrix of Delaunay triangulation, subsequently allowing them to use the produced Voronoi regions to manually identify player roles [23]. Those regions were then divided into clusters by means of hierarchical clustering. State-of-the-art methods often develop sophisticated pipelines and rely on intricate heuristic rules to interconnect the instances across frames and classify various formational patterns.

2.2 Role Labelling

Role labelling in the domain of automative football analysis has been proposed as a method to understand the roles adopted by individual players on the basis of their spatial arrangement on the field. Lucey et al. propose a method for discovering adversarial group behavior in a continuous domain through the use of a spatio-temporal basis model [18]. Large-scale analysis into the behaviour of dynamic subgroups has been conducted to examine the spatial distribution of a team during attacks that have led to a goal [12]. Work has also been carried out into individual role classification when defending corners [4]. Bialkowski et al. assigned corresponding labels to football players at each frame of the tracking data [5].

2.3 Line-breaking passes

A highly probable reason for the paucity of research into the domain of line-breaking passes in association football likely results from the lack of a universally accepted definition of what constitutes such a

pass. This lack of consensus becomes evident when examining the definitions provided by two major global football data providers, StatsBomb and StatsPerform [21, 32]. Both companies agree that such passes need to be successful and break an opposing team's formation line. StatsPerform expands on that definition by claiming that the length of a pass shall be at least 10 meters and made in the direction of the opposition's goal. Moreover, it specifies that the origin and endpoint of the pass need to be at least five meters away from the intersection point and extend at least two meters beyond the deepest defensive player in the formation. State-of-the-art work on the topic of penetrative passes in football, exemplified by Sotudeh's study [29], disregards passes occurring within the first third of the defense area and the final quarter of the pitch for the purpose of only sampling only penetrative passes of high value.

3 METHODOLOGY

Sub-research question 1 (SRQ1) is addressed by employing constrained K-Means clustering to classify players into formation lines. This classification is based on the spatial distribution of players along the x-axis at the start of a passing event. This emphasis on the x-axis is grounded in widely recognized principles in football, which typically classify players into formation lines as viewed from an overhead perspective. Other unsupervised clustering methods, such as hierarchical agglomerative clustering and density-based spatial clustering (DBSCAN), were considered and implemented. However, these were ultimately not chosen due to their significant time complexity *(refer to section 4.3.5)*.

The intricacies of the Second Research Question (SRQ2), are handled by aligning the start timestamps of events from the event data and their precise location on the field from the tracking data. The synchronization is performed as the event data is prone to human error. These inaccuracies can be a consequence of various factors: an annotator's misinterpretation, lack of detail, or even oversight. Such discrepancies could adversely impact the analysis, introducing noise and potentially leading to misguided conclusions. The synchronization strategy adopted in this thesis is a continuation of the work by Anzer and Bauer [1], whose research was primarily centered on identifying the start time of a shot for a goal-scoring probability model. By synchronizing the auto-generated tracking data with the event data, we are effectively cross-verifying the information, and this cross-validation aids in minimizing the impact of such errors. This way the accuracy and reliability of the dataset are considerably improved, subsequently allowing for more precise and credible analysis to be performed, which enhances the overall integrity of the research.

SRQ3 is addressed by seeking an intersection point between the path of the ball trajectory and the formation line. If an intersection point is found, the pass is considered line-breaking, provided it adheres to other pre-defined criteria. Based on the distance of the intersected line to the defending team's goal, we can determine whether the pass penetrated a defensive, midfield, or an offensive line. Finally, the results are visualized in a comprehensible format, allowing for professional-level analysis to be performed.

The fourth sub-research question (SRQ4) is tackled through the introduction of a novel metric in the domain of post-match analysis - "Pass Advantage Score" (PAS). It is computed by determining whether the final position of line-breaking passes - notably, those breaking the defending team's midfield line - falls within a strategically significant dynamic hot-zone area. In the context of this thesis, a dynamic hot-zone area is defined as the region between the pass recipient and the goal of the defending team, where no opposition player is within a 5-meter radius of the pass recipient. The justificaiton behind the choice of choosing passes that penetrate the midfield line lies in the fact that the midfield line can be considered as the logical boundary, splitting the area of the field occupied by the defending team into attacking and defensive halves. A pass that crosses this line has also penetrated the initial layers of the opposition's defense, thus indicating an advanced offensive maneuver. By paying particular attention to such passes, we ensure that our newly introduced metric, primarily accounts for game-changing passes that have the potential to alter the dynamic of the match.

To guarantee the sequential execution of the aforementioned procedural methods, a unidirectional pipeline is created, as depicted in Figure 1.

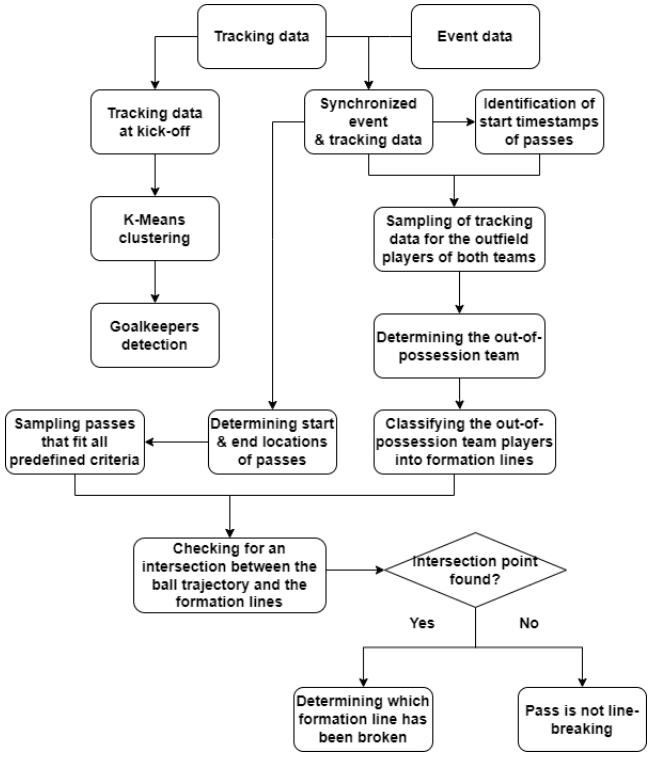


Figure 1: Implementation Pipeline

3.1 Feature Engineering

Feature engineering is executed to enhance dataset dimensionality incorporating features such as pass length, pass angle, the available passable area, the ball advancement on the field of play towards either one of the goals, and the distance of the pass recipient to the nearest defender of the opposition. The features engineered specifically for this research, along with their corresponding formulae,

are outlined in Table 1. Detailed discussions about these features are offered in the subsequent sections.

Table 1: Engineered Features and Formulae

Feature	Formula
Pass Length	$\sqrt{(end_x - start_x)^2 + (end_y - start_y)^2}$
Pass Angle	$\arccos\left(\frac{\sum_{i=1}^{N}(\mathbf{goal}_i - \mathbf{start}_i)\cdot(\mathbf{end}_i - \mathbf{start}_i)}{\|\mathbf{goal}-\mathbf{start}\|\cdot\|\mathbf{end}-\mathbf{start}\|}\right)$
Passable Area	$D_{pa} = \sqrt{(x_g - x_{ps})^2 + (y_g - y_{ps})^2}$
Ball advancement to the right goal	$\sqrt{(x_{rg} - start_x)^2 + (y_{rg} - start_y)^2} - \sqrt{(x_{rg} - end_x)^2 + (y_{rg} - end_y)^2}$
Ball advancement to the left goal	$\sqrt{(x_{lg} - start_x)^2 + (y_{lg} - start_y)^2} - \sqrt{(x_{lg} - end_x)^2 + (y_{lg} - end_y)^2}$
Distance to nearest defender	$\min \sqrt{(x_{\text{defender}} - end_x)^2 + (y_{\text{defender}} - end_y)^2}$
Starting Side	$\frac{1}{N}\sum_{i=1}^{N} x_i$
Ball status (Alive)	if $0 \leq x_{\text{ball}} \leq 105$ and $0 \leq y_{\text{ball}} \leq 68$

4 EXPERIMENTAL SETUP

In the system development lifecycle of this thesis, Python was chosen as the principal programming language. Notably, libraries like *mplsoccer* were leveraged for executing domain-specific analyses.

4.1 Tracking Data

Domain-specific data of spatio-temporal nature also referred to as *tracking data*, is obtained using optical tracking techniques. The tracking data is accumulated in the form of center-of-mass coordinates for all players on the field and the ball through Electronic Performance & Tracking Systems (EPTS). Those systems monitor the movement of the players of both teams and that of the ball at 25 frames per second for the duration of the entire game. For the purpose of this thesis, the Cyronhego TRACAB system is used for collecting the required spatio-temporal data. Apart from the spatial arrangements of the players and the ball along the x- and y- axes on the field of play, the tracking data also contains information about the frame, the jersey numbers of the footballers, and the team in possession of the ball at a given frame. In addition, the positional data contains a third dimension — the height of the ball relative to the pitch surface, which is however disregarded as it doesn't impact the model's ability to detect line-breaking passes.

Tracking the exact position of each entity on the field apart from the referee at 25 frames per second results in the accumulation of vast quantities of data on a game-to-game basis. On average, over a 90-minute long game of football, there exist approximately 135,000 records of positional data for each player on the field. Therefore, over an entire game the total number of rows of data can be calculated as $D = P * L + L$, where D denotes the total number rows of data, P denotes the number of players on the field when the final whistle is blown, and L denotes the length of the game in frames. Assuming the game finishes with all 22 players on the field and with no added time, each game offers around 3,105,000 rows of data, in which the ball data is also included.

4.2 Event Data

To facilitate the automatic detection and analysis of line-breaking passes in the Dutch Eredivisie, *event data*, comprising all events that have transpired during a game, needs to be utilized. The systematic collection of event data allows for a football game to be seen as an ordered sequence of events [33]. This data is provided by StatsBomb and obtained through the StatsBomb API. In contrast to the *tracking data* which has near-perfect accuracy as it is automatically generated via cameras and uses computer vision algorithms for processing it, the event data is annotated manually by human operators and hence is believed to be less accurate. This discrepancy in the accuracy of events is mainly due to three reasons:

(1) The data providers of the tracking and event data are different.

(2) The event data is annotated manually by humans and depends on the annotator's accuracy.

(3) Multiple people can be responsible for the annotation of events over a season, resulting in tagging inconsistencies.

The misalignment between the two datasets is particularly apparent when considering the players' positions on the pitch (often represented via a digital coordinate system), and the timestamps (which are typically subject to delays due to human reaction time). As a result, the event data only provides a rough estimate of the exact frame at which an event took place. The structure of the *event data* is depicted in Table 2, while the various type of events used in this research are shown in Table 3.

Table 2: The attributes of the event data used in this research

Event Attributes	Description
CompetitionID	the unique identified of the competition of which the game is part
MatchID	the unique identifier of a football game
ID	the unique identified of the event
Period	the period/match half in which an event takes place
TypeName	the name of a football event
PlayerID	the unique identifier of the player executing the event
StartX	the manually annotated start location of the event along the x-axis
StartY	the manually annotated start location of the event along the y-axis
EndX	the manually annotated end location of the event along the x-axis
EndY	the manually annotated end location of the event along the y-axis
IsOffCamera	indicating if the event occurred while the camera was off
RelatedEvents	a comma-separated list of the IDs of related events
IsSuccessful	indicating whether an event has been successfully executed
IsOpenPlay	indicating whether the event has occurred from active open play

Table 3: Types of events

Event Type	Description
Ball Receipt	the act of a player successfully receiving the ball
Carry	the action of a player moving the ball while maintaining control of possession
Pass	the action of transferring the ball to another player
Half Start	indicating the start of a football half
Half End	indicating the end of a football half

Given the domain of interest, other events such as shots, corners, or aerial duels are not going to be accounted for with only passes being used.

In order for post-match analysis in the presence of two distinct teams on the field to be facilitated, the tracking data is partitioned on an individual team-basis with the ball being considered as an

independent entity of its own. As discussed already in the preceding section, over 3 million positional data points are expected to be accumulated on average over a single game, regardless of whether the ball is in play or not. Therefore, in order for considerable computational resources to be spared, any sort of computations are only performed in moments when the ball is in play. This requires the implementation of an additional feature of boolean type - "ball status" indicating whether the ball is in play at the specified timestamp. This way, the frames that convey information when the ball is in play account for only 60% of all frames, subsequently resulting in considerable overhead reduction. The ball status is conditioned by checking the x and y coordinates of the ball at each frame against the pitch dimensions. In addition, the ball status is assigned a corresponding value when the play has been stopped as a result of a substitution or a set piece such as a penalty or a corner kick. However, since the *tracking data* is initially obtained in *tracab* format with values ranging from -5250 to 5250 along the x- axis and -3400 to 3400 along the y- axis, it first needs to be standardized to the adopted football pitch dimensions in the Dutch Eredivisie of 105 x 68 meters in a form similar to a two-dimensional Cartesian coordinate system. The *event data* is also initially obtained from StatsBomb in a unique format of 120 x 80 units. To ensure alignment between both data sources, the *event data* is also standardized to the pitch dimensions of 105 x 68 meters.

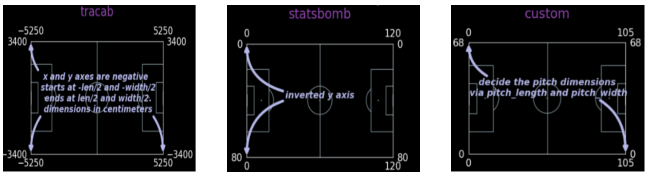


Figure 2: Pitch Dimensions: The event and tracking data are converted to a custom format of 105 x 68 which represents standard football pitch dimensions

4.3 Data Pre-Processing

4.3.1 Tracking & Event Data Synchronization.

The misalignment between the tracking and event data presents a major challenge in cases where the exact start and end locations of events need to be used. Due to the fast-flowing nature of the game and the fine margins between a player being in an offside or an onside position at the time a pass is made, best effort shall be done to align the event data with the tracking data. As discussed in *section 4.2* a shift of a second in the timestamp of the event or few meters in terms of its origin location could have a considerable impact on the formational patterns of the defending team and the calculating of metrics such as 'expected goals' (xG) or 'expected on-the-ball value' (xOBV). Synchronization between the two data sources is achieved by obtaining the *tracking data* from the first frames in each match half and the kick-off tags from the *event data*. This is then followed, by adding a range of values from -100 to +100 frames before the start and after the end of a passing event respectively. Then, the Euclidean distance between each player and the ball over that period is calculated. In the ensuing step, the distances for each frame offset are aggregated. The offset with the smallest cumulative distance is then selected, ultimately aligning the data with the precise moment of kick-off. The synchronization is evaluated by computing short animations (typically from 100 frames before the event's commencement to 100 frames after its end). These animations depict the spatial arrangements of the players of both teams and the ball on a field with dimensions of 105 x 68 meters. The start of an event is marked with an X that's shown for less than a second as part of the animation. The synchronization of the *tracking* and *event* data enables us to visualize the events and the spatial arrangements of the players on the field simultaneously by using complementary information from both datasets irrespective of the event-/tracking-data provider.

4.3.2 Determining end locations of passes.

Despite achieving synchronization of the start frames of events in accordance with the *tracking data*, reliance on manually annotated data for determining end locations of passes has also demonstrated a degree of unpredictability in the results. This unpredictability largely stems from inconsistencies in the tagging process and the potential for human error inherent in the *event data*, the reliability of which is dependent on the annotators' proficiency.

To mitigate these issues, instead of solely relying on the *event data* to determine the end locations of passes on the field of play, best effort was made to incorporate tracking data extensively. An intriguing aspect of the data is the sequence of events following successful passes. In some cases, a successful pass event is followed by a "Ball Receipt", which is then followed by a "Carry" event (*see Table 2*). Alternatively, in other instances, successful passes are directly succeeded by "Carry" events. These inconsistencies in event tagging are likely due to the absence of uniform tagging principles among data analysts in association football. To address this, the event ID of a "Carry" event was cross-verified with the IDs of related events of the passing events, and a common ball receipt ID was sought across the two events. This process enables the use of the start location of a "Ball Receipt" event as the end location of a pass. In cases where no "Ball Receipt" events were present, the start location of a "carry" event was used. This approach ensures better alignment of passing events, improving overall data quality and consistency. Additional features such as the length of a pass, and its progressiveness towards the out-of-possession team's goal have also been implemented, however, the motivation behind that and the specifics behind their implementation are covered in the subsequent section of this paper.

4.3.3 Goalkeepers Exclusion.

In association football, the formations of both teams comprise the outfield players, implying that the positioning of the two goalkeepers on the field doesn't influence the adopted formational patterns. In order for the outfield players to be sampled, the goalkeepers of both teams are excluded from the clustering pipeline. This is achieved by classifying the players of both teams at the first frame of the tracking data for each game into 3 distinct cluster groups. All data about the players that are in a cluster of their own (the goalkeepers) is then excluded from future clustering computations. *Figure 3* outlines the distinct cluster groups at kick-off and clearly shows that the goalkeepers of both teams are in independent groups of their own. This procedure is applied to every game that is being analyzed.

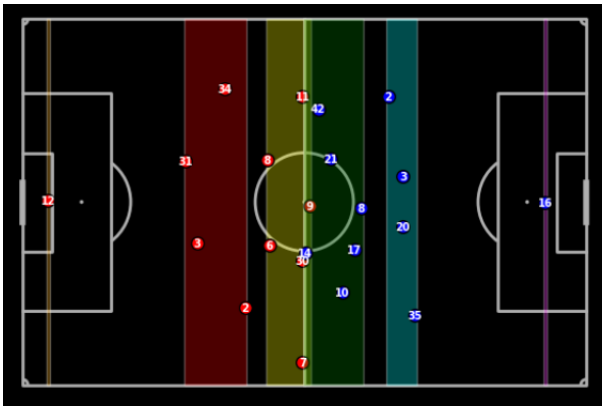


Figure 3: Cluster groups of outfield players

4.3.4 *Model Characteristics.*

The model developed in this paper aims to identify a greater number of line-breaking passes by rectifying inaccuracies inherent to the approach implemented by StatsPerform [21].

As outlined in *section 2.3*, there exists a strict stipulation by StatsPerform regarding a minimum pass length of 10 meters. This definition is believed to be suboptimal since passes made in close proximity to the opposition area are highly unlikely to exceed 10 meters in length in the direction of the goal. This would imply that shorter, potentially line-breaking passes are not being accounted for and hence improperly disregarded. In the scope of this thesis, that drawback is being addressed by calculating the remaining pitch area towards the defending team's goal and sampling passes that advance the ball by at least 10% of that area - also referred to as the *'passable area'*.

In their definition of line-breaking passes, StatsBomb classifies passes made behind the defensive line also as line-breaking although they don't necessarily intersect a defensive line and can be simply regarded as through balls being made down the flanks. In a spatial context, through balls down the flanks are usually made in spaces where there are no defensive 'lines' to break through. They are more about exploiting the space behind the defense rather than penetrating through defensive lines. Moreover, line-breaking passes often lead to immediate goal-scoring opportunities as they are usually played in central areas close to the goal. Through balls down the flanks, however, often require an additional pass (cross) into the box after reaching their target, which might not be as direct a threat to the goal. Therefore, while such passes can be a very effective attacking strategy, they do not possess the same characteristics as line-breaking passes.

This is achieved by sampling only passes, during which the ball traverses longitudinally at least 10% of the *'passable'* pitch area towards the opponent's goal. In the scope of this thesis, the passable area is defined as the geometric region towards the out-of-possession team's goal where a pass can be theoretically received by a player from the same team as the player making the pass. To measure the extent of the passable area, the Euclidean distance from the initial location of a pass to the opponent's goal is computed as follows:

$$D_{\text{pa}} = \sqrt{(x_{\text{g}} - x_{\text{ps}})^2 + (y_{\text{g}} - y_{\text{ps}})^2}$$

where D_{pa} is the passable area, x_{g} and y_{g} are the locations of the out-of-possession team's goal along the x and y axes respectively, while x_{ps} and y_{ps} are the start locations of the pass along the two axes. The length of a pass, determined by calculating the Euclidean distance between its start and end locations, is then evaluated to ascertain whether it covers at least 10% of the passable area. Such a calculation aids in sampling successful passes that meet the line-breaking requirements. In the context of this thesis, these specific instances are referred to as potentially line-breaking passes. On average, these constitute approximately 10% to 15% of all passes executed during a match.

The pass length is retrieved by calculating the Euclidean distance between the start and end locations of a pass, while the ball advancement is measured by first comparing the distances from the start and end locations of a pass to the goal of the team not in possession. Then, the end distance is subtracted from the start distance.

Importantly, the model also filters for passes that are strategically directed towards the goal of the team not in possession, within a pre-defined angular boundary. The critical aspect of the selection process lies in the calculation of the angle towards the goal, which is determined as the angle between two specific vectors. The first vector extends from the initial location of the pass to its end location, representing the ball's trajectory. The second vector points from the initial location of the pass to the midpoint of the defending team's goal, thereby indicating the most direct path towards the opponent's goal. This enables the selection of passes based on their threat level and offensive nature, which is quantified by the angle, θ, between these vectors. Specifically, only passes where $0^\circ \leq \theta \leq 50^\circ$ are considered, ensuring that the model captures passes that are likely to considerably disrupt the formational pattern of the opposition.

In order for a pass to be classified as line-breaking the following criteria shall be fulfilled:

- The pass needs to be successful, implying that the ball recipient needs to be on the same team as the player making the pass
- The pass needs to be executed from open play, meaning that passes originating from set pieces such as free-kicks, goal kicks and corner kicks are disregarded as the spatial arrangement of defending team in such instances is believed to be highly unpredictable and offer dependant on specific coaching instructions, that vary from team to team.
- The ball needs to be advanced longitudinally by at least 10% of the passable area toward the out-of-possession team's goal
- The angle of the ball trajectory towards the defending team's goal shall be within the predefined boundaries
- The ball trajectory shall intersect at least one formation line of the out-of-possession team

4.3.5 *Formation Lines.*

The absence of labelled data and the presence of numerical features requires the use of unsupervised learning techniques to assign

players to formation lines. This approach enables us to classify players into clusters based on their spatial arrangement on the field. In football, the most common formations, such as 4-3-3 or 4-2-3-1 use mainly 3 or 4 different formations lines respectively. However, the dynamic nature of the game often results in complicated 4-line formations to be simplified to 3-line ones during the game. The optimal number of clusters is determined via the Elbow method by calculating the sum of squared distances between points and their respective cluster centroids, as illustrated in *figure 4*.

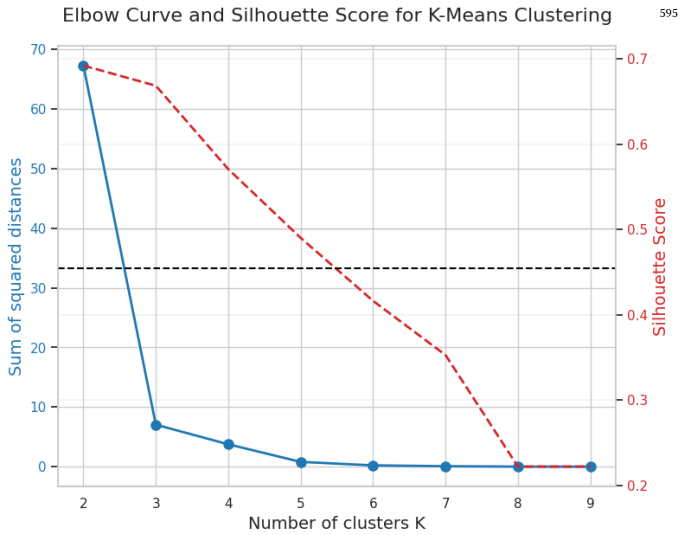


Figure 4: Elbow Method for optimal number of clusters

By plotting the explained variation as a function of the number of clusters, we can choose the elbow of the curve as the number of clusters to be used. Following preceding considerations regarding the lack of a goalkeeper's role in a team's formation, the Elbow method is only performed on the 10 outfield players. The results indicate that on average during a whole game of football, 3 clusters are the optimal number that needs to be chosen for performing K-means clustering on the outfield players of the out-of-possession team. However, the application of a generic K-Means clustering algorithm on the domain-specific problem is not optimal, since it doesn't have problem-specific constraints incorporated, such as the maximum cluster size. The maximum size of a cluster has been limited to 5 players since established football formations (such as 5-3-2, 5-4-1, and 4-5-1) consist of a maximum of 5 players in a line. The incorporation of such constraints ensures the model complies with real-world principles in the domain of interest. Not including a capacity constraint, there's a risk that a formation line might end up with a disproportionately high number of players, subsequently limiting the interpretability of the situation from an analyst standpoint.

After classifying the players into distinct cluster groups, they are interconnected, thus forming distinct formation lines. This is then followed by examining whether the ball's path intersects any of the formation lines. Should this be the case, and assuming the passes have already satisfied previous conditions, they are classified as line-breaking. By cross-examining the pass trajectory with the formation lines, we are able to identify all intersection points which have been depicted with a white dot. These intersection points are assessed based on their relative distance from the out-of-possession team's goal. By calculating the relative distance of each formation line and the intersection points to the out-of-possession team's goal we are able to determine which formation line is being intersected by the ball's path. By doing so we are able to determine whether the defense, midfield, or offensive line is being broken. The spatial arrangements of the players of AZ Alkmaar (red) and SBV Vitesse (blue) at the moment a pass is being made by an SBV Vitesse player are illustrated in *figure 5*.

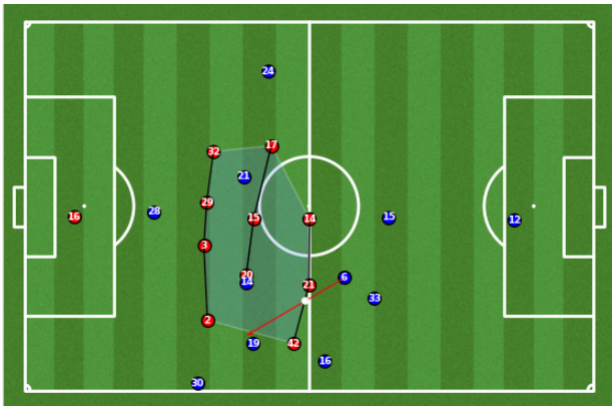


Figure 5: Graphic depiction of a line-breaking pass

4.4 Evaluation Metrics

Evaluating the quality of an unsupervised learning model can be challenging due to the lack of ground truth data. Therefore, metrics such as the Silhouette score and the Davies-Bouldin index are employed to gauge the effectiveness of the clustering algorithm.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad \textit{Silhouette Score}$$

where *a(i)* is the mean distance between a sample and all other points in the same cluster group, and *b(i)* is the mean distance between a sample and all other points in the next nearest cluster group.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d_{ij}} \right) \qquad \textit{Davies-Bouldin Index}$$

where n is the number of clusters, i and j are the cluster indices, a_i is the average distance between each point in cluster i and the centroid of that cluster, a_j is the average distance between each point in cluster j and the centroid of that cluster, and d_{ij} is the distance between the centroids of clusters i and j.

The first metric is the Silhouette Score, *s(i)*, which measures the similarity within clusters. It ranges from -1, implying incorrect cluster assignment, to 1, suggesting clear differentiation between clusters. The latter would indicate a robust correlation between the players and their respective cluster groups.

Additionally, we utilize the Davies-Bouldin Index *(DBI)* as a second measure. This index gauges both inter-cluster separation and intra-cluster dispersion, serving as a valuable tool in assessing the partitioning quality of our clusters. A lower Davies-Bouldin Index signifies a more desirable partitioning outcome.

The comprehensive results derived from the aforementioned metrics will be elaborated upon in the subsequent Results section.

5 RESULTS

In preliminary system iterations, the classification of players into formation lines was performed at every 10 frames of the tracking data. Considering that a single computation takes on average 1 second the projected time requirement for processing a single 90-minute football game would be approximately 225 minutes. This process has been optimized in the final system design by performing the K-Means clustering exclusively between the frames marking the start and end of a potentially line-breaking pass. This tactical modification has enabled us to considerably reduce the computational workload. Subsequently, the processing time has decreased considerably from the initial 225 minutes to an average of approximately 2.5 minutes, albeit with some variability based on the frequency of line-breaking passes during any given match. This signifies a substantial improvement in the efficiency of the system, making it more viable for real-world applications.

The silhouette score, measuring similarity within clusters, yielded an average of **0.81** across 34 Dutch Eredivisie matches in which AZ Alkmaar participated in the 2022/2023 season. Since its value ranges between -1 (implying clusters are assigned incorrectly) to 1 (meaning that clusters are clearly distinguished) the achieved score indicates a strong correlation between players and their cluster groups. On the other hand, the Davies-Bouldin index is a measure of inter-cluster separation and intra-cluster dispersion. A lower Davies-Bouldin index indicates better partitioning of clusters. An average score of **0.19** across all processed games suggests that the clusters are well-separated and less dispersed (values in the range between 0 and 1 that are closer to 0 suggest low within-cluster distances). The combined insights from these metrics support the model's proficiency in classifying players into formation lines based on their spatial arrangements, demonstrating the robustness and high quality of the clustering outcome. The violin plot in *figure 6* displays the importance of line-breaking passes in football by comparing the xOBV of line-breaking and non-line-breaking passes.

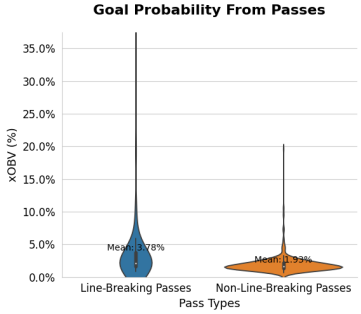


Figure 6: Comparison of goal probabilities using xOBV

The 'expected on-the-ball value' metric measures the probability of scoring a goal within a possession chain after the successful execution of an event. It can be seen that penetrative passes result in higher xOBV values.

Following the precise detection of line-breaking passes across multiple games in the Dutch Eredivisie, sufficient data has been gathered to conduct an analysis from a domain-specific standpoint. In the context of player recruitment and analysis, a major benefit of the detection of line-breaking passes is that the data can be used to help identify individual players with exceptional vision and passing skills. By using supplementary information from a database that includes the names of players and their current teams in the league, valuable insights are obtained regarding the top performers in creating line-breaking passes as shown in *figure 7*.

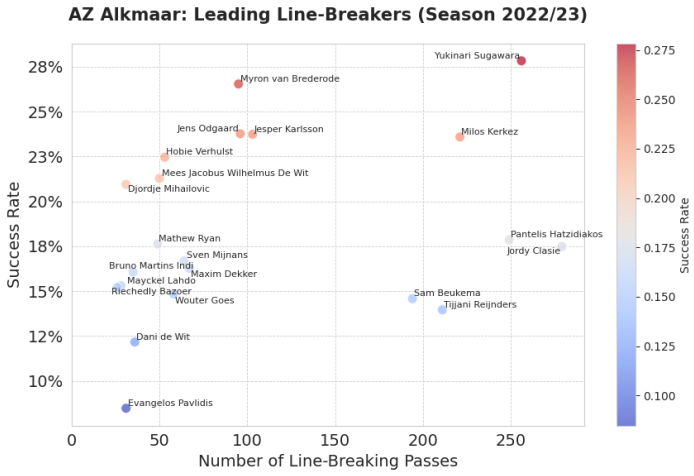


Figure 7: AZ Alkmaar: Top Line-Breakers (2022/23 Season)

This data can then be used by football scouts for the purpose of evaluating a player's ability to unlock tight defensive structures, create scoring opportunities, and exploit gaps in the opponent's spatial arrangement on the field.

The violin plot in *figure 8* offers a comparison between the median PAS values of passes that fit the aforementioned criteria and all other passes made during a game. This visualization supports the hypothesis that passes received within a dynamic hot-zone area typically exert more pressure on the defending team and increase the likelihood of generating a goal-scoring opportunity. This increase in threat level is quantifiably indicated by the higher median PAS of such passes. Moreover, this plot provides a direct response to the fourth sub-research question (SRQ4). It demonstrates the utility and efficacy of the proposed PAS metric in evaluating the strategic impact of line-breaking passes. Thus, it not only serves as an affirmation of the chosen methodological approach but also underscores the substantial potential of PAS as an analytical tool in future football match analyses.

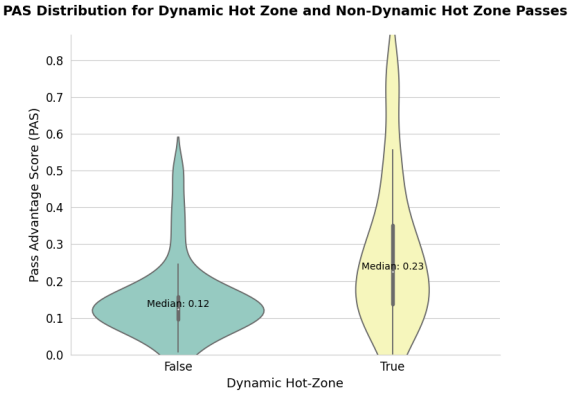


Figure 8: Introducing the Pass Advantage Score metric

6 DISCUSSION

The work outlined in this paper presents a novel approach to automatically detecting line-breaking passes in association football. To date, research on the domain of interest is scarce, particularly in the Dutch Eredivisie. In addition, no other system has been developed by AZ Alkmaar, which could have served as a baseline in terms of evaluating the performance. Therefore, all aspects of the system development lifecycle from the pre-processing of the data to the application of the machine learning model and the subsequent analysis of the results is novel. A notable contribution of this paper into the domain of sports analytics is that it offers a considerably less computationally-exhaustive procedure for detecting formational patterns in comparison to other state-of-the-art works in the domain of football analysis by performing the procedure only at the frame of the *tracking data* marking the start time of a pass. Since, on average, the number of passes that occur during a game of football is usually in the hundreds, the computational workload is drastically decreased with the model only being executed on those instances, rather than the over 150 000 frames of *tracking data*.

The limitations of this research mainly lie in the assumptions made for determining what makes a pass line-breaking. Due to the lack of a unanimous definition of such events in association football, the current criteria for such events has been established with the aim to improve current definitions and address some of their limitations. The current model detects line-breaking passes by calculating the remaining passable area from the start location of the pass to the out-of-possession team's goal and only samples cases where an advancement of at least 10% in that area has been detected. This allows for a greater number of passes to be detected especially in the final third of the pitch. In addition, sideway passes that break an opposition's formation line but pose no threat as they are not directed towards the opponent's goal are being disregarded as their on-the-ball value is not considered significant enough. This methodology, however, doesn't guarantee that all line-breaking passes are being accounted for. Although, in the scope of this thesis, best effort has been made to limit the reliance on event data, which has allowed for accurate representations of football events to be made. However, in future system iterations, depending on the nature of the problem addressed, it might not be possible to use an alternative data source. Therefore, the relatively low quality of the *event data* may present a major impediment as the granularity and accuracy of the time annotation are never 100% accurate. Particularly, the lack of unanimous principles in match annotation and the inconsistent nature of human errors makes it difficult to predict cases in which there could be imprecise annotations. In the future, it would be a sensible approach to develop a pipeline that automatically syncs the end locations of events with the *tracking data* irrespective of the problem being analyzed. A potential future mitigation strategy on a greater scale could be establishing uniform criteria when annotating football events, thus diminishing the likelihood of tagging inconsistencies. Furthermore, reliance on manually annotated event data might become scarce in the future as novel approaches of labeling events happening on the field of play are beginning to emerge. For instance, at the 2022 World Cup in Qatar, each ball was equipped with an inertial measurement unit (IMU) sensor, located in its center. Its purpose was to transmit ball data to the video operation room at a frequency of 500 times per second, enabling highly accurate detection of every ball event [31]. However, in many professional leagues worldwide such technologies are yet to be adopted, including in the Dutch Eredivisie, hence human operators are still being used to manually annotate various football events.

Another limitation of the project is related to the adopted football field dimensions. According to FIFA stipulations, football pitches are allowed to have different dimensions (between 90 and 120 meters in length and between 45 and 90 meters in width). This flexibility could pose a threat to the performance of the machine learning model as inconsistencies in football field sizes could result in an inability to draw meaningful conclusions from the data. As of the time of writing this thesis, all football grounds in the Dutch Eredivisie use a standard pitch size of 105 x 68 meters. However, if the problem is to be applied on football games played outside of the Netherlands or if the regulations are to change in the future, the *tracking* and *event* locations will need to be standardized accordingly. The model's performance could be negatively affected by a third risk factor: game disruptions. These may include unsportsmanlike behaviors such as pitch invasions or the use of flares by fans. The smoke generated from flares could obscure the camera's vision, subsequently compromising the quality of the data. Moreover, these interruptions could lead to the referee temporarily suspending the game, which would further disrupt the accuracy of the tracking data.

Finally, finding an appropriate benchmark for conducting a comparative analysis of the results of the thesis with state-of-the-art work in the field also presents a risk. The majority of the research papers on the topic of interest don't provide accuracy metrics such as precision and recall, due to the lack of ground truth data.

7 CONCLUSION

In conclusion, this thesis has demonstrated a novel way of automating the detection of line-breaking passes in association football. This research fills a significant gap in sports analytics, as line-breaking passes have received limited attention in previous studies, albeit their considerable contribution in the lead-up to dangerous situations and goals. Through the combination of tracking data of spatio-temporal context and event data from various football games, the project has proven to facilitate the process of post-match

analysis. The application of unsupervised learning methods in combination with physics-based features has enabled the identification of intersection points between the ball trajectory and the defending team's formation lines, irrespective of the spatial arrangements of the players or the teams involved. The incorporation of supplementary information such as player profiles and team data, has allowed us to gain a deeper understanding of the players who excel in line-breaking passes and their potential impact on team performance. This knowledge can prove invaluable for football scouts and talent evaluators, enabling them to make more informed decisions when identifying and recruiting players with above-standard passing abilities. Future work on the classification of line-breaking passes in association could benefit from more auto-generated event data which would be expected to considerably reduce the time and effort spent performing synchronization between the two data sources, while also providing higher-accuracy data. However, reliance on manually annotated event data is expected to remain the standard in the upcoming years until more sophisticated data-gathering systems become integrated into professional leagues worldwide. Overall, the insights gained through this thesis can contribute to enhancing team performance, improving recruitment strategies, and ultimately, shaping the future of the game.

REFERENCES

[1] Gabriel Anzer and Pascal Bauer. 2021. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living* (2021), 53.

[2] Gabriel Anzer and Pascal Bauer. 2022. Expected passes: Determining the difficulty of a pass in football (soccer) using spatio-temporal data. *Data mining and knowledge discovery* 36, 1 (2022), 295–317.

[3] Guillermo Martinez Arastey. 2018. PERFORMANCE INDICATORS IN FOOTBALL. *SPORT PERFORMANCE ANALYSIS* (2018).

[4] Pascal Bauer, Gabriel Anzer, and Joshua Wyatt Smith. 2022. Individual role classification for players defending corners in football (soccer). *Journal of Quantitative Analysis in Sports* 18, 2 (2022), 147–160.

[5] Alina Bialkowski, Patrick Lucey, Peter Carr, Iain Matthews, Sridha Sridharan, and Clinton Fookes. 2016. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2596–2605.

[6] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. 2014. Large-scale analysis of soccer matches using spatiotemporal tracking data. (2014), 725–730.

[7] Tom Boomstra. 2022. Towards automatically classifying football formations for video analysis. (2022).

[8] Lotte Bransen. 2018. One pass is not the other: a data-driven overview of the most effective passers in European football. (2018).

[9] Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a data-driven player ranking in soccer using predictive model weights. (2016), 49–55.

[10] Fabio Giuliano Caetano, Sylvio Barbon Junior, Ricardo da Silva Torres, Sergio Augusto Cunha, Paulo Régis Caron Ruffino, Luiz Eduardo Barreto Martins, and Felipe Arruda Moura. 2021. Football player dominant region determined by a novel model based on instantaneous kinematics variables. *Scientific Reports* 11, 1 (2021), 18209.

[11] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* 39, 1 (1977), 1–22.

[12] Floris R Goes, Michel S Brink, Marije T Elferink-Gemser, Matthias Kempe, and Koen APM Lemmink. 2021. The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences* 39, 5 (2021), 523–532.

[13] Joachim Gudmundsson and Thomas Wolle. 2012. Football analysis using spatio-temporal tools. (2012), 566–569.

[14] Laszlo Gyarmati and Xavier Anguera. 2015. Automatic extraction of the passing strategies of soccer teams. *arXiv preprint arXiv:1508.02171* (2015).

[15] Michael Horton, Joachim Gudmundsson, Sanjay Chawla, and Joël Estephan. 2015. Automated classification of passing in football. In *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II*. Springer, 319–330.

[16] Antonio Irpino and Rosanna Verde. 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In *Data science and classification*. Springer, 185–192.

[17] Patrick Lucey, Alina Bialkowski, Peter Carr, Eric Foote, and Iain Matthews. 2012. Characterizing multi-agent team behavior from partial team tracings: Evidence from the english premier league. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1387–1393.

[18] Patrick Lucey, Alina Bialkowski, Peter Carr, Stuart Morgan, Iain Matthews, and Yaser Sheikh. 2013. Representing and discovering adversarial team behaviors using player roles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2706–2713.

[19] Patrick Lucey, Dean Oliver, Peter Carr, Joe Roth, and Iain Matthews. 2013. Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1366–1374.

[20] Ian G McHale and Samuel D Relton. 2018. Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research* 268, 1 (2018), 339–347.

[21] Kuba Michalczyk. [n. d.]. *How impactful are line-breaking passes?*

[22] Eric Müller-Budack, Jonas Theiner, Robert Rein, and Ralph Ewerth. 2019. " Does 4-4-2 exist?"– An Analytics Approach to Understand and Classify Football Team Formations in Single Match Situations. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*. 25–33.

[23] Takuma Narizuka and Yoshihiro Yamazaki. 2019. Clustering algorithm for formations in football games. *Scientific reports* 9, 1 (2019), 13172.

[24] Martí Perarnau. 2014. *Pep Confidential: Inside Pep Guardiola's First Season at Bayern Munich*. Birlinn.

[25] Paul Power, Hector Ruiz, Xinyu Wei, and Patrick Lucey. 2017. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. (2017), 1605–1613.

[26] Charles Reep and Bernard Benjamin. 1968. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)* 131, 4 (1968), 581–585.

[27] Laurence Shaw. 2022. A brief history of statistics in soccer: Why actual goals remain king in predicting who will win? (2022). https://phys.org/news/2022-12-history-statistics-soccer-actual-goals.html

[28] Laurie Shaw and Mark Glickman. 2019. Dynamic analysis of team strategy in professional football. *Barca sports analytics summit* 13 (2019).

[29] Hadi Sotudeh. [n. d.]. *Potential Penetrative Pass(P3)*.

[30] William Spearman. 2018. Beyond expected goals. (2018), 1–17.

[31] Sky Sports. 2022. World Cup 2022: Qatar tournament to feature semi-automated offside technology with ball sensors and cameras. (2022).

[32] StatsBomb. 2022. *StatsBomb Launch New 360 Metrics: Line-Breaking Passes and Ball Receipts In Space*.

[33] et al Stein, Manuel. 2019. From Movement to Events: Improving Soccer Match Annotations. (2019).

[34] Łukasz Szczepański and Ian McHale. 2016. Beyond completion rate: evaluating the passing ability of footballers. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (2016), 513–533.

[35] Paul Worsfold and Kirstin Macbeth. 2009. The reliability of television broadcasting statistics in soccer. *International Journal of Performance Analysis in Sport* 9, 3 (2009), 344–353.