

University of Southampton

Faculty of Engineering and Physical Sciences

Electronics and Computer Science

# **Player2Vec: Using Deep Learning for Analyzing Soccer Players'**

## **Playing Style**

By

**Rishi Aluri**

7<sup>th</sup> September, 2022

**Supervisor: Professor Tim Norman**

**Second Examiner: Doctor Enrico Marchioni**

A dissertation submitted in partial fulfilment of the degree

of MSc **Artificial Intelligence**

## ABSTRACT

Soccer is a dynamic sport that requires teams to constantly improve and adapt their tactics. A major problem for Football Clubs is finding the right player who will fight their team tactics. In this research, we present a deep learning approach to characterizing a player's playing style into a Player2Vec player vector and further engineering action descriptors that can help Football Clubs in deciding the right fit player that can fit their teams depending on the team tactics. The Player2Vec vector successfully retrieves 86.7% of all players in the top-20 list from the player retrieval task and the performance exceeds the results of the previous study. The high precision of Player2Vec ensures that the player similarity analysis is accurate and provides the right players similar to a given target player. The player similarity analysis can provide Football Clubs with the opportunity to find new players who are similar to the players who are either retiring or transferring to other clubs. It can also benefit players by highlighting the key areas for development by comparing themselves to the desired player. The action descriptors can further help in player recruitment by providing unique information about the player that can be used to tactically justify if the player can fit the team. We provide a human interpretable player vector and action descriptors that can also be used for data analysis.

## **AKNOWLEDGMENTS**

I would like to express my deep gratitude to my research supervisor, professor Tim Norman, who despite of his busy schedule, took the time to supervise and guide me throughout my research stages.

I am thankful and grateful to my family for their love and support during all these years, and especially during my masters' graduation research project and also for the years to come. Special thanks to every professor that ever taught me, from my first day at school till my last.

I am also thankful and grateful to Doctor David Sumpter and the 'Friends of Tracking' GitHub community who introduced me to the world of football analytics. I am truly inspired by their work and hold a special place for them in my heart.

## **Statement of Originality**

- I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.
- I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.
- I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.
- I have acknowledged all sources, and identified any content taken from elsewhere
- I have not used any resources produced by anyone else.
- I did all the work myself, or with my allocated group, and have not helped anyone else.
- The material in the report is genuine, and I have included all my data/code/designs.
- I have not submitted any part of this work for another assessment.
- My work did not involve human participants, their cells or data, or animals.

## Table of Contents

I.	Introduction.....	10
I.	Motivations .....	10
II.	Scope of this Project .....	11
III.	Dissertation Structure.....	12
II.	Literature Review.....	13
I.	Expected Goals (xG).....	13
II.	A Spatio-Temporal Action Rating System for Soccer (STRASS) .....	13
III.	Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams	13
III.	Data Preparation.....	14
I.	Data Sources .....	14
II.	Wyscout Open-Source Dataset .....	15
III.	Simplified Representation of Dataset .....	16
I.	Soccer Player Action Description Language .....	16
II.	Selecting Relevant Action types .....	19
IV.	Feature Engineering.....	20
I.	Feature Engineering 10 action descriptors using the relevant actions .....	20
II.	Constructing Heatmaps .....	24
V.	Compress Heatmaps to build Player2Vec Player vectors .....	25
IV.	Deep Learning for Feature Extraction .....	27
I.	Dimensionality Reduction and latent feature extraction.....	27
I.	Fundamental Architecture of CAE.....	27
II.	Convolutional Autoencoder vs NMF .....	28
II.	Loss function for Convolutional Autoencoder.....	29
III.	Machine Learning Models .....	30
I.	Baseline model.....	30
II.	Software and Hardware Implementations.....	30
III.	Model Training Considerations.....	30
IV.	Experimentation.....	31
V.	MODEL27_CAE2 Diagnostic Analysis .....	32
V.	Evaluation .....	35
I.	Player Retrieval from Anonymized Event Stream Data .....	35
II.	Player Similarity Analysis.....	39

III.	Player Replacement Task .....	41
VI.	Case-Study Analysis .....	44
	Liverpool FC Champions league triumph 2018/19 season.....	44
VII.	Discussions .....	47
VIII.	Conclusions.....	49
IX.	References .....	50

## List of Figures

Figure 1 Soccer has an average of 3 goals per game .....	10
Figure 2 Wyscout Pitch Coordinates[9] .....	16
Figure 3 SPADL pitch coordinates[12] .....	17
Figure 4 The actions leading up to the goal from L. Messi. The first five actions show FC Barcelona's ball progression to the opponent half with the final pass from Coutinho leading to the shot from Messi and Barcelona scoring the goal. ....	19
Figure 5 Shot results w.r.t the distance of the shot. ....	23
Figure 6 Cross results w.r.t the distance of the crosses. ....	23
Figure 7 Pass results w.r.t the distance of the passes. ....	23
Figure 8 Pass distances values. ....	23
Figure 9 Dribble Distance Values. ....	23
Figure 10 Example of a heatmap detailing the shot playing style of Lionel Messi, winger at FC Barcelona in the 2017/2018 season. ....	24
Figure 11 Summary of the tasks performed in the Data Preparation step .....	25
Figure 12 (a) A shot heatmap of size 48x48. (b) The heatmap is inputted to a Convolutional Autoencoder that has a 'code' part in between which reduce the dimensions and creates a latent feature space of 32 dimensions (shape: 4,4,2). (c) Reconstructed shot heatmap. ....	27
Figure 13 Fundamental Architecture of Convolutional Autoencoder .....	28
Figure 14 CAE vs NMF[17] .....	29
Figure 15 Formulae of Binary Cross Entropy loss function[18] .....	29
Figure 16 Sigmoid Activation Function always outputs the values in range 0 to 1[19] .....	30
Figure 17 Architecture of CAE model 'MODEL27_CAE2' .....	32
Figure 18 MODEL27_CAE2 training and validation loss. The model stopped learning after 67 epochs on a batch size of 32. The validation loss is 2%. ....	33
Figure 19 Examples of heatmaps reconstructed from the original heatmaps. The first row shows the original representation of the heatmaps followed by the row which contains the reconstructed heatmaps of the original image. The remaining rows follow this trend. ....	34
Figure 20 Player Retrieval Task .....	36

Figure 21 Comparison of five best performing models that generate the Player2Vec player vector. The models were tested for player retrieval task using Manhattan distance. 'MODEL27_CAE2' (green) performs better than the rest of the models (dashed lines).....	37
Figure 22 Comparison of five best performing models that generate the Player2Vec player vector. The models were tested for player retrieval task using Euclidean distance. 'MODEL27_CAE2' (green) performs better than the rest of the models (dashed lines).....	38
Figure 23 (a) Shows the comparison of the ten action descriptors of Olivier Giroud and Sandro Wagner. (b) shows the comparison of the ten action descriptors of Olivier Giroud and Alexandre Lacazette. The ‘pass-forward’ and ‘pass-short’ attributes show the closer similarity of Lacazette to Giroud than Wagner, based on the tactics of Arsenal FC.....	43
Figure 24 Fabinho lifting the Champions League Trophy after playing a crucial role for Liverpool FC during the 2018/19 season campaign[33]. ....	44
Figure 25 Comparison of the ten action descriptors of Emre Can and Fabinho.....	45
Figure 26 Fabinho's passing sonar for Liverpool in the 2019/20 Premier League season[33].....	46

## List of Figures

Table 1 Wyscout Open Access Dataset .....	15
Table 2 SPADL 12 attributes.....	17
Table 3 Actions leading up to the goal by Lionel Messi.....	18
Table 4 Selecting relevant action types: Pass, Cross, Dribble and Shot .....	20
Table 5 Model hyperparameters subject to experimentation. ....	31
Table 6 Summary of Wyscout soccer ball event data .....	35
Table 7 Top-k results and mean reciprocal rank (MRR) comparison of the proposed solution to the previous study. ....	36
Table 8 Performance comparison of the five best CAE models .....	38
Table 9 Player Similarity Analysis of Messi and Dybala. ....	40
Table 10 Player Similarity Analysis of Modric and Fabregas. ....	40
Table 11 Player Similarity Analysis of Kante and Gueye.....	40
Table 12 Player Similarity Analysis of Ronaldo and Son.....	41
Table 13 Player Similarity Analysis of Ramos and Laporte. ....	41
Table 14 The names of the top 5 most similar players to the striker Olivier Giroud and the difference in the similarity in numbers. ....	42



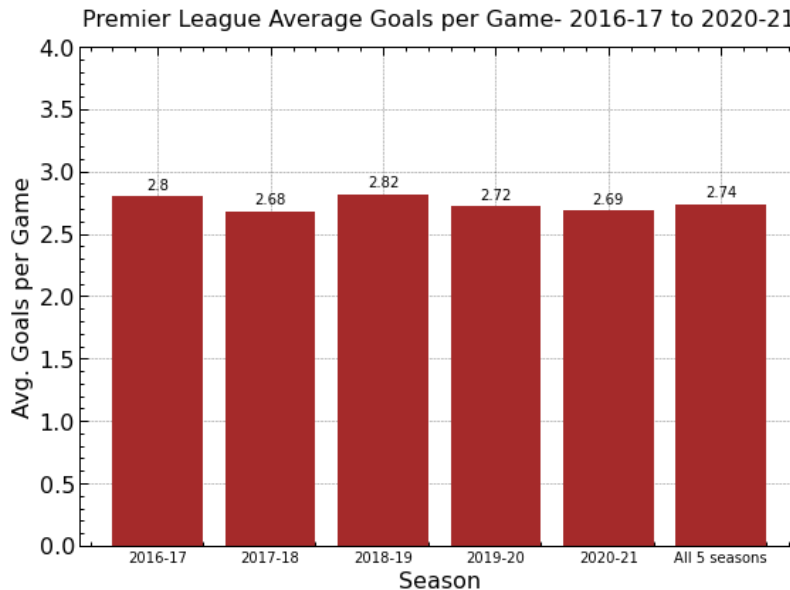
Table 15 Player similarity analysis of Olivier Giroud. Sandro Wagner is ranked the most similar player to Giroud. Lacazette is the 2nd most similar player to Giroud. ....	43
Table 16 Player similarity between Emre Can and Fabinho .....	45

## I. Introduction

*“The game of life is a lot like football. You have to tackle your problems, block your fears, and score your points when you get the opportunity.”*

*-Lewis Grizzard*

Soccer/Association football (also called football) is a simple sport that is easy to understand but nearly impossible to master. Soccer is a sport played for 90 minutes with one ball, 2 teams, 22 players (11 on each team), and an average football pitch of size 115 by 68 meters[1]. Studies have found that football is one of the biggest sports in the world, making up 43% of the sports industry. Several countries have many football teams competing in both regional and national championships. Being the biggest sport watched in the world, it plays a significant part in people's lives. The most interesting fact about Soccer is the low-scoring nature of matches even though there are more than 1 million ways of scoring a goal Figure 1.



*Figure 1 Soccer has an average of 3 goals per game*

With Football being so popular around the world, the practice of tracking and analyzing football data of players and matches both academically and in betting markets is a normal trend. The availability of football data presents a unique opportunity for the artificial intelligence (AI) and machine learning (ML) communities to develop, validate, and apply new techniques in the real world that can not only benefit the bookmakers who set the odds, and punters who place the bets but also help the teams and players in improving their performance, strategizing their tactics, and also scouting for new players.

## I. Motivations

In many professional sports, data analysis is becoming increasingly crucial. Sports teams are analyzing massive volumes of data to obtain a competitive advantage over their opponents. In

soccer, traditional match statistics are usually raw counts or fractions, such as ball possession, number of shots on target, or pass success rate. While these data are fascinating, they do not provide a comprehensive view of the game. Furthermore, they can occasionally hide crucial information. For example, the sheer quantity of shots taken by a player tells us nothing about the difficulty or quality of the efforts. Recent studies have concentrated on methods that enable a more thorough and informative examination of soccer players. One example is a popular predictive analysis metric currently used by the footballing community which is the Expected Goals (xG) metric[2].

When discussing about soccer, it is a common trend to always compare the players based on their respective playing styles. Fans often argue about the current best players in soccer and given the limited positions in the field and high similarity of play style between most players, it is quite a challenge for data analysts to characterize the players based on the playing style. From a practical point of view, it can be quite important for football clubs to characterize playing styles for the below reasons:

**Player Development Monitoring** Young players often look up to world-class soccer players and try to emulate the style portrayed by professional players. Coaches and football clubs can maintain a player vector that can help assess the style of their players and help them improve their performances by comparing them to world-class players and their player vectors. Coaches can use the player vector of the young players to track the development and assess the changes in play style to improve professional players and even young players playing for the football academy.

**Scouting** Teams can benefit from knowing each player's style of play so that in case any player decides to leave the club, the scouting team can look for players having a similar play style as the departing player. In such a way, replacements can also be found for the aging players in the team who will soon retire from the sport. Scouting for new players is a risky task as the scouted player needs to fit into the team, work well with the current team tactics and should provide value to the team. Hence, a robust system that can provide a trustworthy characterization of play style for a player can benefit the recruitment team to make correct decisions while recommending new players for the football club.

## II. Scope of this Project

Soccer is a dynamic sport with many movements and interactions among players across time and space. Hence, analyzing event stream data from soccer matches can be very challenging. To achieve satisfactory results from this project and at the same time solve some real-world challenges, a fair scope for the project has to be set. Therefore, the research objectives that will be aimed to be accomplished are as follows:

**RO1)** Players rarely perform the actions in the same location but at the same time these actions are what define the playing style of the players and hence every location where the action happens is important. Therefore, the first objective will be to devise a play-style characterizing system that intelligently generalizes the location of the actions performed by the players.

**RO2)** All event stream data of players' actions contain both discrete (ex: Team name, player name) and continuous (ex: time, minutes played, start location, end location) variables. Since most

machine learning models work on either discrete variables or continuous variables, a robust technique has to be engineered that converts the continuous variables to be discrete in nature.

**RO3)** Even after realizing the player’s play style it can be a difficult task for a scouting team to choose a replacement from the abundant number of possible candidates. The process of replacing a current player with a new player cannot only depend on the new player’s playing style but also should depend on the value the new player will potentially bring to the team. Therefore, a set of value-based features have to be discovered to help the recruitment team of a football club in the decision-making process.

The scope of this project is set to derive a complete method that can distinguish a player’s play style. Typically, a play style of a player is heavily influenced by the skills possessed by the player and the tactics employed by the team. Therefore, to work with a constant definition of playing style we will define the meaning of a player’s play style. We will employ the definition given by Decroos and Davis [3] who define it as *“A player’s playing style can be characterized by his preferred area(s) on the field to occupy and which actions he tends to perform in each of these locations”*.

Using this definition, we will work on achieving the objective of characterizing a player’s play style into a player vector, which we name as Player2Vec player vector, which is both human interpretable and suitable for data analysis tasks.

### **III. Dissertation Structure**

The dissertation is structured in a similar order in which the research was conducted. The Literature Review section is followed after the Introduction section. Section 3 - Data Preparation, explains the data gathering and preprocessing steps to create the Player2Vec player vector. Section 4 - Deep Learning for Feature Extraction, gives the architectural details about the deep learning model used for this project. Section 5 Evaluation, provides an assessment of our Player2Vec approach, followed by a Case-Study Analysis section to show how the Player2Vec approach can be used for analysis. Lastly, sections Discussions and Conclusions provide the overall summary of the project with limitations and future scopes.

## II. Literature Review

### I. Expected Goals (xG)

The paper identifies that the randomness and scarcity of goals limit the ability to properly judge current performance and predict the future performance of teams and players using goals alone [2].

In order to provide a more continuous context, the xG model allows an estimate of the expected number of goals a team ought to have based on their performance on key metrics up to the current time.

Key metrics included goals scored, shots on goal, missed shots, blocked shots, turnovers and faceoff statistics, upon which Ridge regression was used to build models of xG. The first recorded use of xG in soccer was in 2016 and is now commonly used by sporting media as a performance metric for both teams and players.

### II. A Spatio-Temporal Action Rating System for Soccer (STRASS)

STARSS is an approach for automatically rating the actions performed by soccer players, which leverages historical match data to assign a rating to the actions performed by the players in a match [4]. For a given match, the presented approach proceeds in three steps. First, the approach splits the match into phases, which are uninterrupted sequences of actions where one team is in possession of the ball. Second, it assigns a phase rating to each phase based on historical match data. The higher the assigned rating, the more likely that the phase will end with a goal. The approach focuses on the actions that contribute to the offensive output of the team. Third, the approach distributes the phase rating across the individual actions that constitute the phase.

Unlike more simple approaches for rating soccer players, this approach goes beyond shots and goals. It considers all the actions that contribute to a team's offensive output and accounts for the spatio-temporal context in which these actions were performed. Several case studies show that this approach is able to identify top-performing players in individual matches as well as throughout the course of an entire season.

### III. Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams

Decroos and Davis [3] attempt to characterize a player's playing style in an objective and data-driven manner based on analyzing event stream or play-by-play match data. The goal of their work was to summarize a player's playing style into a fixed-length player vector. They performed this by overlaying a grid on the pitch and counting how often a player performs a specific action in a given location. They cope with the challenge of dimensionality by reducing it using Non-negative Matrix Factorization. Finally, they concatenate the compressed player actions to form the player vector. This literature study is the inspiration for Player2Vec.

### III. Data Preparation

The main task performed in this section is as follows:

**Given:** A event stream of actions performed by a player.

**Aim:** Construct a fixed-size Player2Vec player vector that describes a player's playing style and is comprehensible to both machine learning algorithms and human analysts.

Coming up with a method to characterize play style can be a challenging task since we have to keep track of the spatial locations, discrete actions and a variable number of events for each player. At a high level, the approach we propose is as follows:

First, extract all the events that occur in a match from Wyscout open-source dataset. Second, convert the dataset to a simplified human understandable language that can be also used by machine learning systems. Then, select the actions that are offensive in nature and come from an open play in a match. Third, engineering 10 features, two for each action-type (shot, dribble, cross) and four for action-type: pass. These 10 action descriptors will help in player similarity analysis and also in decision-making during the player recruitment processes. Fourth, for every relevant action-type develop a heatmap of that action performed by a player over the course of a sequence of matches. This heatmap can now be vectorized but since it has high dimensionality, it will computationally expensive to use it for further analysis. Therefore, perform dimensionality reduction using a Convolutional Autoencoder (CAE) by extracting the compressed representation of a player's heatmap matrix. The compressed heatmap map is then reshaped to form a vector. Now, for each player develop a player vector '*player2vec*' by concatenating all the compressed vectors of each action type. This concludes the steps performed in this section.

The Evaluation section will assess the usage of the Player2Vec player vector. After processing the player2vec vector, a player retrieval task is performed to evaluate the accuracy of the player vector. Player2vec usage is then showcased with a player similarity analysis and a player replacement task. Also, a case study analysis is done that provides insight into Liverpool FC's player scouting method and their Champions League triumph.

#### I. Data Sources

The data sources related to soccer can be in various forms with two of them being the most popular:

1. Event data- This data captures event-by-event tracking of the on-the-ball actions performed by individual players on the football field. The event data is collected by human annotators who examine video feeds of soccer matches and describe all on-the-ball activities players execute on the field chronologically, such as passes, shots, crosses, dribbles, tackles and interceptions. The main competitors in this space are StatsBomb [5], Opta Sports [6] and Wyscout [7].
2. Tracking data- This type of data has higher granularity than Event data. Tracking data not only capture each and every action performed by the player on the ball but also notes the location of all the players on the pitch at regular intervals of time.

This study will involve only the match event data because it is fairly detailed and contains the fundamental structure of each match. Since the project involves developing individual player

vectors, match event data will provide rich information about each action performed by each player which can then be adopted and used quickly.

This project will be using event stream data from Wyscout open data source for generating the player vectors.

## II. Wyscout Open-Source Dataset

The Wyscout open access dataset [8] is by far considered as the most abundant form of match event data source currently existing for everyone to use since 2019. The data which we will be using for this project involves the 2017–2018 seasons of the following five national soccer leagues in Europe: the Spanish First Division, Italian First Division, English First Division, German First Division, and French First Division. There is also information regarding the World Cup 2018 and the European Cup 2016, both of which are national team events but not included in this project. The data includes around 1,941 matches, 3,251,294 events, and 4,299 players (Table 1).

Competition	Matches	Events	Players
Spanish first division	380	6,28,659	619
English first division	380	6,43,150	603
Italian first division	380	6,47,372	686
German first division	306	5,19,407	537
French first division	380	6,32,807	629
World cup 2018	64	1,01,759	736
European cup 2016	51	78,140	552

*Table 1 Wyscout Open Access Dataset*

Wyscout compiled and distributed the soccer logs. The data collecting technique was carried out by skilled video analyzers (the operators) using proprietary software (the tagger). To ensure the accuracy of data gathering, the tagging of events in a match was conducted by three operators, one for each team and one as the accountable supervisor of the overall match output.

Each data set is provided in JSON format (JavaScript Object Notation) for easy access and usage. There are seven different file formats offered, and six of them include context-specific data on the competitions, matches, teams, players, coaches, and referees (11 MB in total). The essential in-game data is kept in the events file type (911 MB total). The events file contains information on the 21 different types of "events" that were seen throughout the match, including acceleration, ball out, clearance, counterattack, duel, foul, goalkeeper leaving line, interception, link-up play, non-ball, offside, opportunity, own goal, pass, progressive run, set pieces, shot, save attempt, touch, touch in box, and transition. These event categories are expanded upon by 78 "subevent" categories.

The Match events concentrate on actions involving the ball and are documented for both the offensive team (the team with the ball) and their opposition (the defensive team). Around three-quarters of the events are recorded for the actions performed by the offensive teams, indicating that data gathering is skewed towards the offensive team.

The pitch coordinate system used by Wyscout is illustrated in Figure 2. The location of an event is typically identified by at least one set of coordinates (start\_x, start\_y), however many events also include two pairs, for example describing the start and end location of a pass.

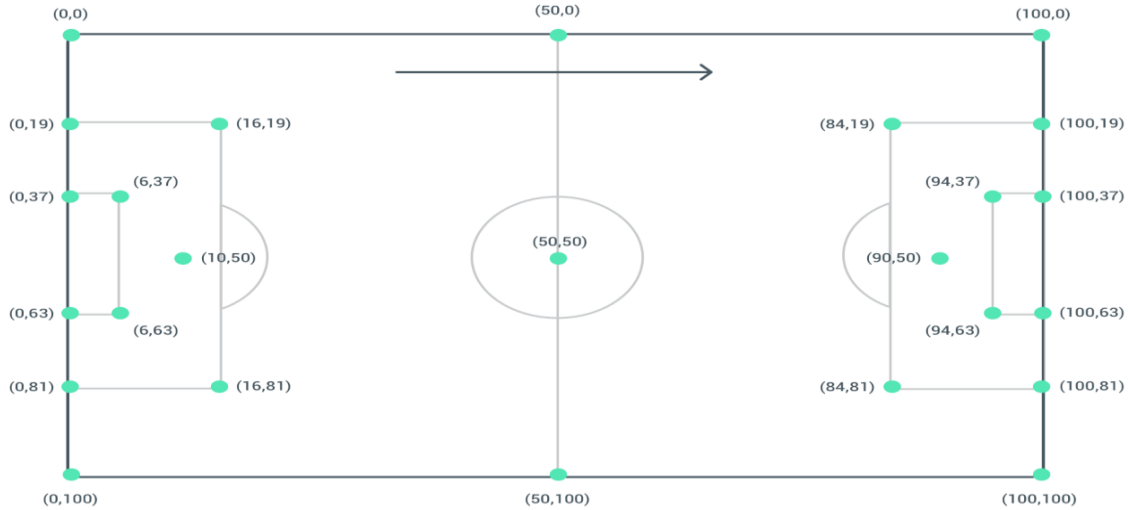


Figure 2 Wyscout Pitch Coordinates[9]

From the start of the play session until the end, a timestamp is continually supplied. Regardless of any pauses in play, the time keeps ticking. Tags for the first, second, first half of overtime, and the second half of overtime are included in the match logs.

Even though there is some amount of quality control, the actual observation error is not specified in the Wyscout documentation. Time and location are both recorded with a high degree of precision, but both are done manually. The distribution of coordinates can be empirically seen to exhibit some rounding of values at the centre x and y values.

### III. Simplified Representation of Dataset

#### I. Soccer Player Action Description Language

Although Wyscout has a simpler structure compared to Opta or StatsBomb, the player information is condensed in the events data and hence, quite difficult to extract. To simplify the process and represent the data in a more understandable format, the Wyscout data is converted to a unified representation called SPADL.

The goal of Soccer Player Action Description Language (SPADL)[10] is to combine the various event stream formats into a single language to facilitate further data analysis. SPADL was created to precisely define and explain on-pitch actions in a straightforward, comprehensive manner that is understandable by humans. The human-interpretability allows one to analyse what takes place on the field and determine whether the values ascribed to such activities are consistent with the perceptions of soccer experts. The simplicity and completeness lower the likelihood of errors and provides us with all information to express the actions in the whole context. To address the challenges posed by the variety of event stream data formats and to benefit the soccer analytics



community, Decroos, T., et al. released a Python package, called ‘*Socceraction*’ [10, 11], that can automatically convert the data of different vendors like Opta, StatsBomb and Wyscout to SPADL.

SPADL, in contrast to the formats used by commercial suppliers to express events, is a language for defining player actions. The difference is that actions are a subset of events that entail a player executing the action. For instance, a passing event counts as an action, while an event that signals the game's completion does not. SPADL represents a game as a sequence of on-the-ball actions  $[a_1, a_2, a_3, \dots, a_m]$ , where  $m$  is the total number of actions that happened in the game. Each action is a tuple of the same twelve attributes shown in Table 2:

Attribute	Description
game_id	Game ID in which the action was performed
period_id	Period ID in which the action was performed (1 - first half, 2 - second half)
seconds	Action's start time
player	Name of the player who performed the action
team	Team name of the player
start_x	x coordinate where the action started
start_y	y coordinate where the action started
end_x	x coordinate where the action ended
end_y	y coordinate where the action started
action_type	Type of the action (e.g., pass, shot, dribble)
result	Result of the action (e.g., success or fail)
bodypart	Player's body part used for the action

Table 2 SPADL 12 attributes

### Start and End Locations

In the SPADL (Figure 3), a standard field of 105 m by 68 m is used, with the origin located in the lower-left corner of the pitch.

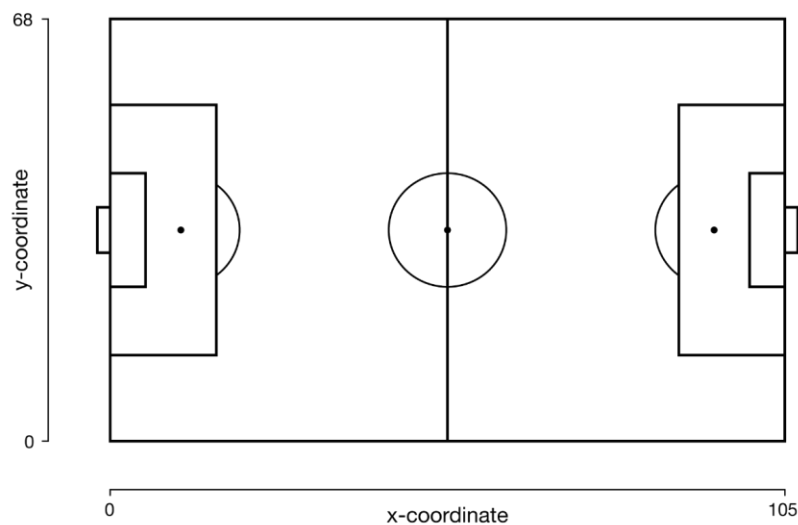


Figure 3 SPADL pitch coordinates[12]

### Action Type

The action type attribute can have 22 possible values. These are pass, cross, throw-in, crossed free kick, short free kick, crossed corner, short corner, take-on, foul, tackle, interception, shot, penalty shot, free kick shot, keeper save, keeper claim, keeper punch, keeper pick-up, clearance, bad touch, dribble and goal kick. These 22 possible values are generic enough such that similar actions have the same type and are also specific enough to accurately describe what happens on the pitch.

### Result

For the result attribute, the value ‘success’ can be assigned to an outcome given an action accomplishes its intended result, or ‘fail’ otherwise. A pass to a teammate is an example of successful action. A throw that travels over the sideline is an example of a failed activity. Some actions might have unexpected outcomes such as Offside (for passes, corners, and free kicks), own goal (for shoots), and yellow and red card penalties (for fouls).

### Body Part

There are four potential values for the body part attribute: foot, head, other, and none. A special body part head/other is used for Wyscout, which does not distinguish between the head and other body parts.

### Example

Below is an example of a series of actions that lead to a goal scored by Lionel Messi from FC Barcelona. Table 3 represents the SPADL actions and Figure 4 illustrates the actions on the pitch.

Game_Id	Period_Id	Time_Seconds	Team	Player	Start_X	Start_Y	End_X	End_Y	Action_Type	Result	Bodypart
2565845	2	2581.950521	FC Barcelona	I. Rakitić	55.65	31.28	59.85	47.6	pass	success	foot
2565845	2	2584.408674	FC Barcelona	Philippe Coutinho	59.85	47.6	75.6	54.4	dribble	success	foot
2565845	2	2586.866826	FC Barcelona	Philippe Coutinho	75.6	54.4	87.15	62.56	pass	success	foot
2565845	2	2588.847585	FC Barcelona	Jordi Alba	87.15	62.56	97.65	54.4	pass	success	foot
2565845	2	2590.559495	FC Barcelona	Philippe Coutinho	97.65	54.4	80.85	35.36	pass	success	foot
2565845	2	2593.110855	FC Barcelona	L. Messi	80.85	35.36	105	37.4	shot	success	foot

*Table 3 Actions leading up to the goal by Lionel Messi.*

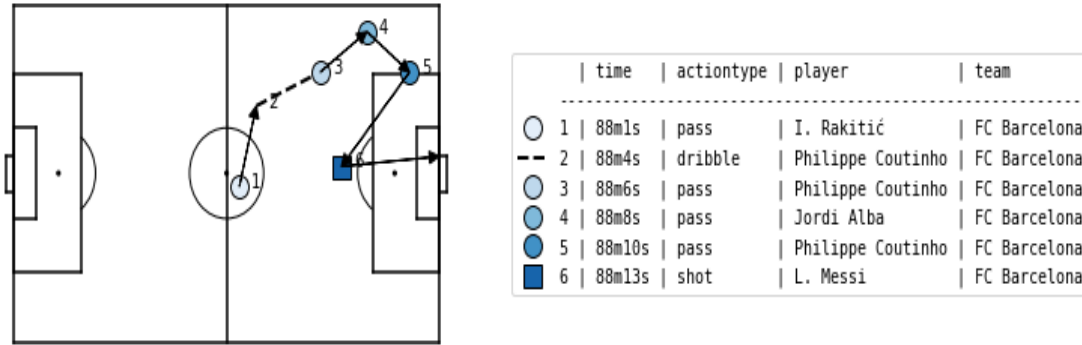


Figure 4 The actions leading up to the goal from L. Messi. The first five actions show FC Barcelona's ball progression to the opponent half with the final pass from Coutinho leading to the shot from Messi and Barcelona scoring the goal.

## II. Selecting Relevant Action types

SPADL converts the events from Wyscout data into actions of different possible types. From here on the process involves selecting relevant action types that a player performs that provide information about that particular player's style of play. The input events stream contains 19 different actions; however, we have to only consider offensive action types that were performed during the flow of a match for two main reasons:

First, placement is a key component of defending in a game, and choosing a stance to block specific actions is frequently involved. Determining defensive style is therefore by definition impossible for us to achieve without using off-the-ball location data. In addition, the majority of on-the-ball defensive actions (such as tackles and clearances) are typically made out of necessity rather than to display a player's playing style.

Second, in this project only open play actions are considered since most of the set piece actions such as free-kicks, penalties and throw-ins are done by pre-defined players or pre-defined locations (Example- wingers for throw-ins, strikers for penalties). Set pieces define the team tactics more rather than a player's style of play.

To summarize, each action type must fit two criteria to be considered relevant for characterizing playing style: it must be offensive and it must occur during open play. When applying these two criteria, the remaining relevant action types are passes, dribbles, crosses, and shots (Table 4).

Action type	Offensive	Open Play
Bad_Touch	No	Yes
Clearance	No	Yes
Corner_Crossed	Yes	No
Corner_Short	Yes	No
Cross	Yes	Yes
Dribble	Yes	Yes
Foul	No	Yes
Freekick_Crossed	Yes	No
Freekick_Short	Yes	No
Interception	No	Yes
Keeper_Claim	No	Yes
Keeper_Pick_Up	No	Yes
Keeper_Punch	No	Yes
Keeper_Save	No	Yes
Pass	Yes	Yes
Shot	Yes	Yes
Shot_Freekick	Yes	No
Tackle	No	Yes
Throw_In	Yes	No

Table 4 Selecting relevant action types: Pass, Cross, Dribble and Shot

## IV. Feature Engineering

### I. Feature Engineering 10 action descriptors using the relevant actions

Feature engineering is the process of using extensive research and domain knowledge to create new variables that extract hidden attributes from the raw data. In this project, we discover ten features that present unique characteristics of the style and value a particular player brings to the game. As we now have five relevant action-types (discussed in the previous section), each action will be used to engineer two features, in the following way:

1. **Shot Action-Type:** Shots play a crucial role in football as they result in goals and goals win matches. Players are often judged by the number of goals they have scored over a sequence of matches or a complete season. They provide a unique look at the potential of a player to be world-class. For shot action-type, the two features engineered are a) Shots-Successful (goal scored) and b) Shots-Failed (goal not scored). The reasoning behind these attributes being discovered is to answer the below questions:
  - a. How often does the player tend to shoot? (Number of shots taken over the season)
  - b. How potent is the player in front of the goal? (Number of shots getting converted to goals)
  - c. What is the ratio of the shots failed to the shots successful? (The finishing ability of a player or ability to score a goal)

2. **Dribble Action-Type:** Dribbles are the second most common action performed by players after 'pass'. Most dribbles are short adjustments to their locations to make a pass/cross/shot (or other) while few dribbles are long in distance either during a counter-attack or a player presenting a flair with the ball and moving past defenders. Hence, dribbles can be categorized into a) Short distance dribbles and b) Long distance dribbles. Figure 9 shows that the mean distance of dribble is around 15m and the box plot suggests that about 60% of dribbles are less than 15m. Therefore, Long distance dribbles are those dribbles that have a distance of more than 15m whereas short-distance dribbles are less than or equal to 15m. The reasoning behind engineering these attributes is to answer the below questions:
  - a. How good is the player in ball progression to the opponent half? (Number of dribbles attempted)
  - b. How good is the player in finding pockets of spaces (in-between defenders) on the pitch? (Number of short-distance dribbles)
  - c. How good is the player in taking on multiple defenders and going past them? (Number of long-distance dribbles)
3. **Cross Action-type:** Like shots, crosses lead to strikers either heading the ball in goal or volleying the ball into the goal. Crosses are a special type of 'pass' action that showcase a player's vision to pick out the players to cross the ball to. A good cross is always a defender's nightmare and most goals in football are scored because of a cross in the opponent's defensive line. For crosses the two features engineered are: a) Crosses Successful and b) Crosses Failed. The rationale behind this approach is to answer the below challenges:
  - a. How often does the player tend to cross? (Number of crosses taken over the season)
  - b. How dangerous is the cross played by the player? (Number of successful crosses in front of the goal, hence, increasing the chances of a shot on goal)
  - c. What is the ratio of the crosses failed to the crosses successful? (The accuracy of the player to cross the ball)
4. **Pass (pass-start and pass-end) Action-type:** Passes are the most common action performed by any player on the field. Pass action-type can single-handedly define a player's style and hence requires more robust feature engineering. The pass-start action-type suggests the location from which the passes originate, hence, the most important thing a player decides when performing a pass is to either perform a high-pass or a low-pass. High passes are those passes that travel long distances, above the ground and over the heads of players. It is performed to switch the play from one end to another and prevent losing possession of the ball. On the other hand, low passes are mostly grounded passes that are short-distance in length and quick in nature. Therefore, the two features that are derived from this understanding are a) short length pass, and b) long length pass. From the literature study, a long pass is generally measured to be a pass with a length of more than 30 yards (27m) and short pass is of length less than or equal to 30 yards (27m) (Figure 8). Further, the pass-end action-type suggests the location where the passes reach on the pitch, hence, the direction of the pass has been considered. A forward pass in a game can help in increasing the chances of a goal by progressing the ball closer to the opponent's goal whereas a backward pass suggests a more defensive approach followed by the player to keep ball possession. Therefore, the two features that are derived from this understanding are c) Forward direction pass, and d) Backward direction pass. The reasoning behind not engineering a pass success-fail ratio is that most of the passes executed in a game are successful passes and the rationale behind the above four features is to answer the below questions:

- a. Is the player a playmaker? (Number of forward passes  $>$  Number of backward passes)
- b. How good is the player's vision? (Number of long-distance passes)
- c. How good is the player in maintaining the ball possession? (Number of backward passes, Number of short-distance passes).

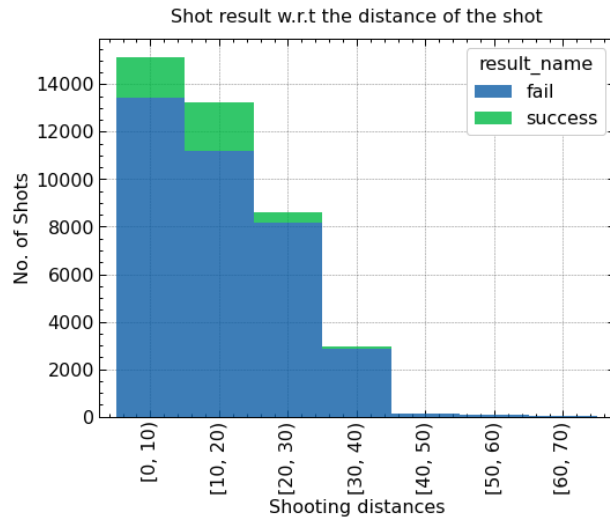


Figure 5 Shot results w.r.t the distance of the shot.

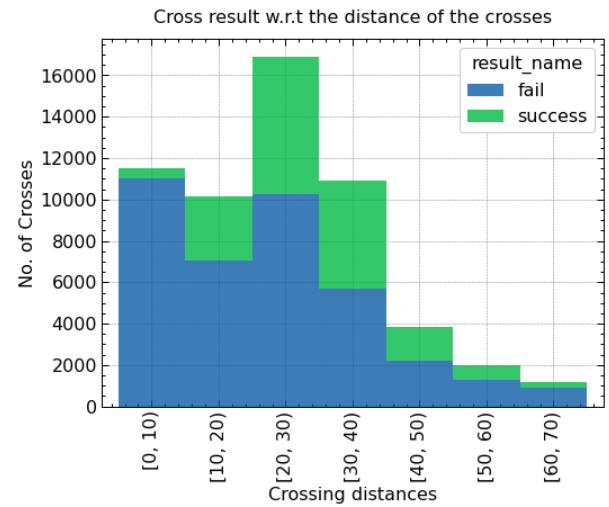


Figure 6 Cross results w.r.t the distance of the crosses.

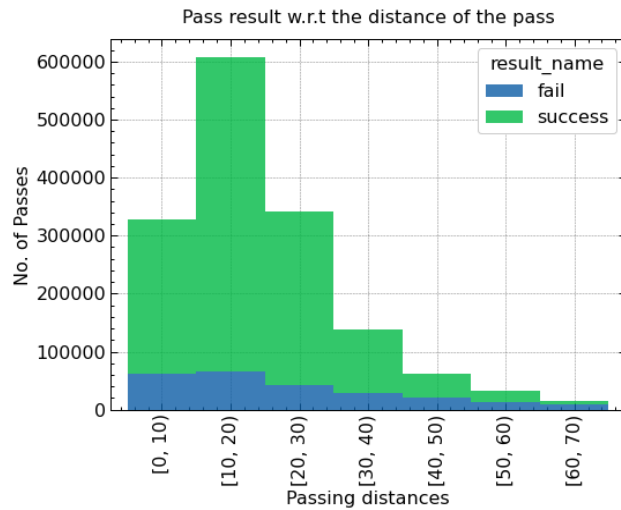


Figure 7 Pass results w.r.t the distance of the passes.

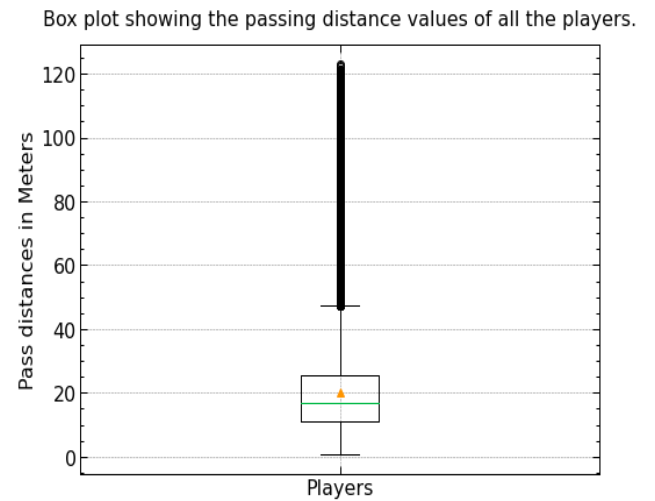


Figure 8 Pass distances values.

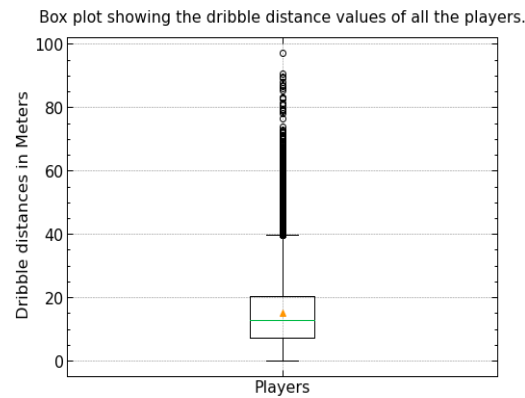


Figure 9 Dribble Distance Values.

## II. Constructing Heatmaps

A heatmap is the count of the total number of actions of action-type  $t$  performed by player  $p$  at a specific location on a spatial coordinate space i.e., the soccer pitch. For each player and action type, the following three steps are executed.

1. **Counting:** First, Overlay a grid of size  $m \times n$  on the soccer field. Next, select all the actions, of type  $t$ , performed by player  $p$  from the data set. Now, for every grid cell  $X_{i,j}$ , count the number of actions that started in that cell. As a result, we converted a variable-size collection of actions into a fixed-size matrix  $X \in Nm \times n$  holding the raw counts per cell.
2. **Normalization:** Normalize the matrix  $X$  such that each cell contains its count if player  $p$  had played only one game i.e., 90 minutes. If player  $p$  played 2000 minutes in total, then construct the normalized matrix  $X' = \frac{90}{2000}X$ . Since two players having the same identical playing styles can have different minutes of play, normalization of data is quite important. For example, if player  $p1$  played more minutes than player  $p2$ , then player  $p1$ 's matrix  $X$  will contain higher raw counts than the matrix of player  $p2$  but after normalizing the heatmaps both players' matrices will contain a count per 90mins.
3. **Smoothing:** To promote smoothness in the counts of nearby cells, a Gaussian blur is applied to matrix  $X'$ . Gaussian blur, an image blurring technique in image processing, involves convolving  $X'$  with a Gaussian function. Specifically, the value of each cell in  $X'$  is replaced by a weighted average of itself and its neighborhood, leading to the blurred matrix  $X'' \in \mathbb{R}_+^{m \times n}$ . The smoothed counts of  $X'$  enhances the spatial coherence of the locations where the actions were performed.

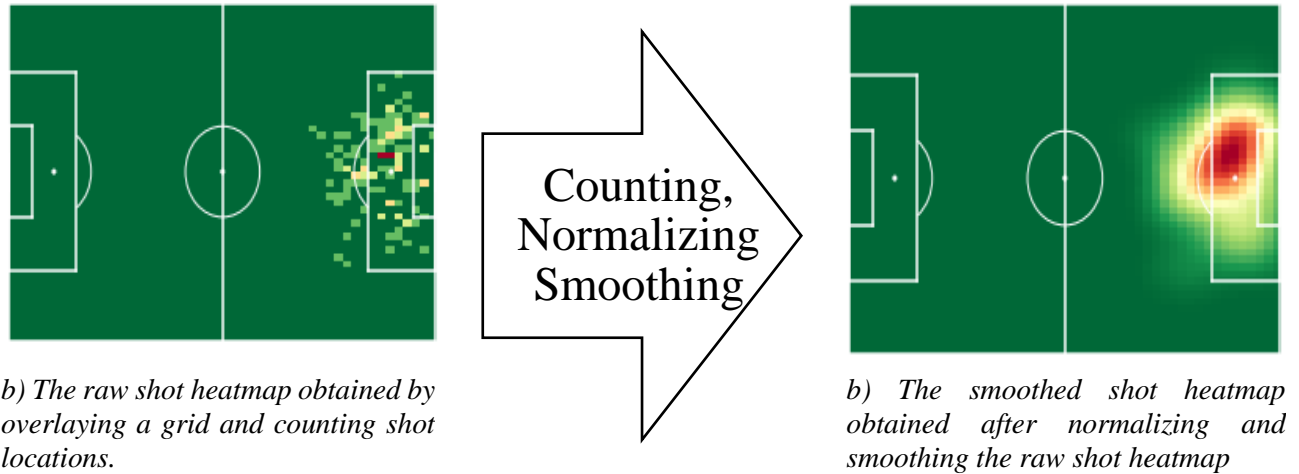


Figure 10 Example of a heatmap detailing the shot playing style of Lionel Messi, winger at FC Barcelona in the 2017/2018 season.

Since, we have four relevant action-types (shot, dribble, pass, cross) for each player, separate heatmaps have to be created for each action type.  $X''$  is the heatmap detailing where player  $p$  performs actions of type  $t$  (Figure 10 (a), (b)). For action type 'pass', we are not just interested in their start locations, but also in their end locations, hence, we construct separate heatmaps  $X_{start}''$



and  $X_{end}$ ” using the start ( $start\_x, start\_y$ ) and end ( $end\_x, end\_y$ ) locations of the ‘pass’ actions in the dataset. This divides the pass action-type into two separate actions: 1) Pass-start, and 2) Pass-end, making it a total of 5 relevant actions-types.

## V. Compress Heatmaps to build Player2Vec Player vectors

After constructing heatmaps for the actions performed by a player, we compress and flatten the players’ heatmaps with Convolutional Autoencoder (CAE). CAE has been tried and tested in reducing the dimensionality of images[13]. As a result, a CAE is trained to compress a heatmap to a latent feature representation which is a vector that represents a player’s heatmap for each action performed by that player. To achieve this the architecture of the CAE is defined to have three convolutional and three deconvolutional layers, each using a Leaky Rectified Liner Unit (Leaky-ReLU) activation function. First, we train the CAE model to recreate the original heatmap using self-supervised learning. Then, A player’s heatmap is fed into the trained CAE to extract the latent feature vector. The procedure for CAE training is discussed in full detail in the next section ‘Deep Learning for Feature Extraction’.

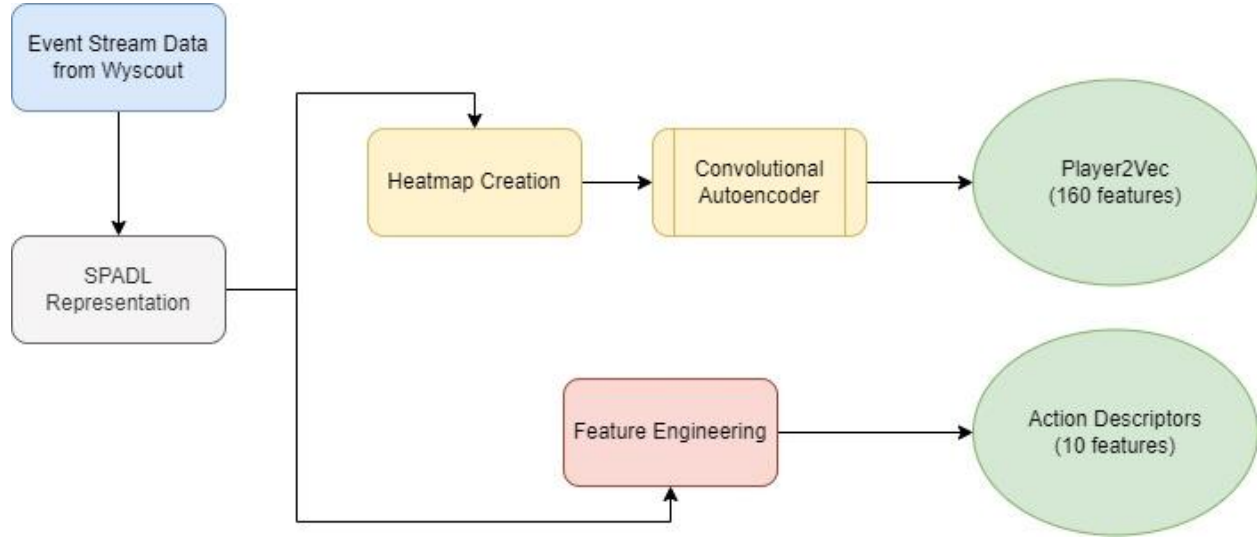


Figure 11 Summary of the tasks performed in the Data Preparation step

As a result, for a particular player  $p$  having five heatmaps ( $X''$  matrix), for each of the actions – shot, dribble, cross, pass-start, pass-end, we compress the  $X''$  matrix with shape  $48 \times 48 \times 1$  to form a matrix  $X'''$  with shape  $4 \times 4 \times 2$  (i.e., 32 dimensions). This is done by inputting the heatmap into the CAE and extracting the latent feature space (shape: 4,4,2) from the CAE. Then, by just flattening the matrix  $X'''$  we get a vector  $V$  of length 32. Hence, for each player  $p$ , we get five vectors –  $V_{shot}$ ,  $V_{dribble}$ ,  $V_{cross}$ ,  $V_{pass-start}$ , and  $V_{pass-end}$ .

Now, the Player2Vec vector  $k$  of player  $p$  is the concatenation of his compressed vectors for the relevant action types: shot, dribble, cross, and pass. The total length of a Player2Vec vector  $k$  is 160 dimensions which is equal to  $V_{shot} + V_{dribble} + V_{cross} + V_{pass-start} + V_{pass-end}$ .

Compressing the heatmaps into lesser dimensions will help in reducing the computational power to perform data analytics tasks. With the help of CAE, we reduce the dimension of heatmaps from  $2,304 \times 5$ , since a player will have 5 heatmaps for each action-type, to just 160 dimensions.

We can now assess the similarity of two players' playing styles by computing the Manhattan distance between their Player2Vec vectors. Manhattan distance works well when comparing data with high dimensionality and also because the value of each feature in each Player2Vec vector is a meaningful quantity. Unlike the Euclidean distance, which tends to unfairly punish significant differences in a few characteristics, the Manhattan distance just simply computes the sum of the absolute differences for each feature and provides a fair distance comparison.

## IV. Deep Learning for Feature Extraction

The main step of building the Player2Vec vector is the extraction of the latent features from the Convolutional Autoencoder. Figure 12 (a) shows the original heatmap is fed to the Convolutional Autoencoder. Then the CAE performs two important tasks (Figure 12 (b)).

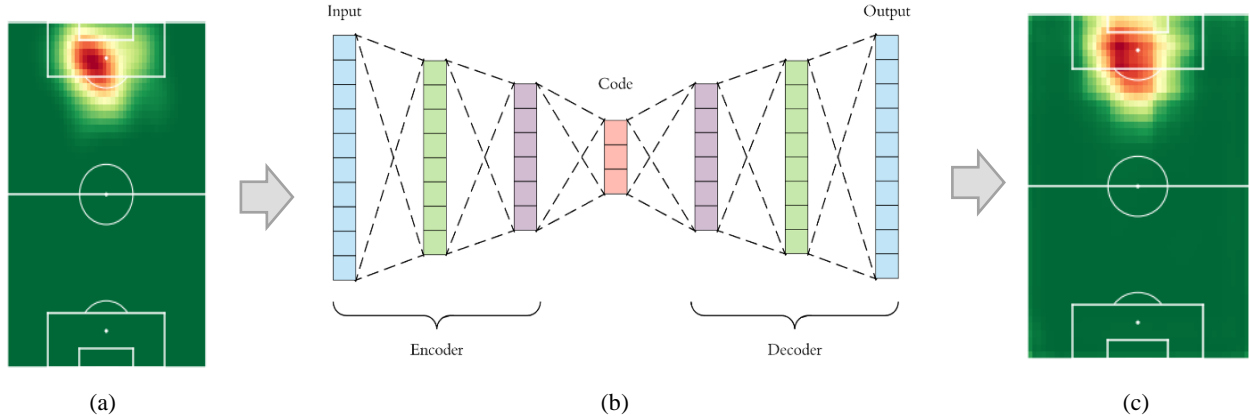


Figure 12 (a) A shot heatmap of size 48x48. (b) The heatmap is inputted to a Convolutional Autoencoder that has a 'code' part in between which reduce the dimensions and creates a latent feature space of 32 dimensions (shape: 4,4,2). (c) Reconstructed shot heatmap

First, since the original heatmap is of size 48x48 in height and width, it will be highly computationally expensive to apply any data analytics tasks on it as the flattened heatmap will be of 2,304 dimensions in length. To counter this, we use the Convolutional Autoencoder to reduce the dimensionality of the heatmap from 2,304 dimensions to 32 dimensions. The latent feature space present in the bottleneck 'code' region of the CAE contains the compressed form of the original heatmap. This compression task is performed by the 'Encoder' part of the CAE.

Second, we can then use the latent features to retrieve back the original heatmap. The reconstructed heatmap resembles the original heatmap in shape (48x48) as well as in numerical values. This reconstruction task is performed by the 'Decoder' part of the CAE.

### I. Dimensionality Reduction and latent feature extraction

When dealing with a lot of input variables, it is often necessary to reduce the dimension of the data to capture the most valuable inputs among all of those input variables. Reducing the dimensions in high dimensional data is commonly performed using PCA (Principal Component Analysis) or NMF (Non-Negative Matrix Factorization). The application of Deep learning for dimensionality reduction is a fairly new concept but shows a lot of promise.

#### I. Fundamental Architecture of CAE

“An autoencoder is an unsupervised machine learning algorithm that takes an image as input and tries to reconstruct it back using a fewer number of bits from the latent space representation” [14]. The fundamental architecture of a Convolutional Architecture contains 1) An encoder layer that is responsible for compressing, ideally, an input image into lower dimensions, 2) This latent

representation is stored in the bottleneck region. 3) The Decoder that is responsible for generating back the original input.

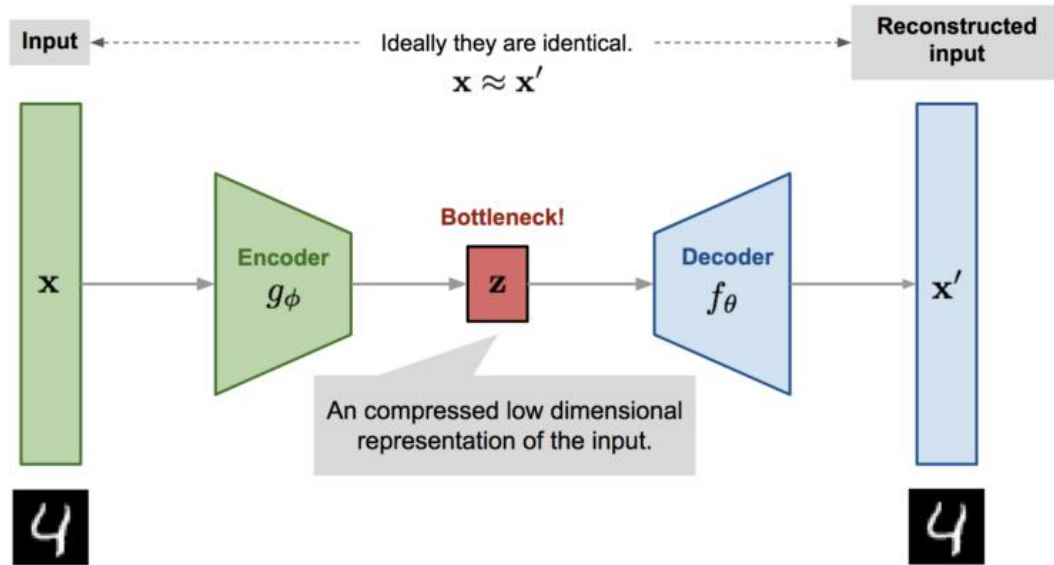


Figure 13 Fundamental Architecture of Convolutional Autoencoder

An autoencoder learns using a self-supervised learning technique which is a special instance of a supervised learning technique where the targets are generated, within the model, from the input data [15]. The autoencoder logs a reconstruction loss that it aims to reduce during every iteration. This cost function determines the efficiency of the Convolutional Autoencoder. The lower the loss the closer the reconstructed image to that of the original image.

## II. Convolutional Autoencoder vs NMF

The previous study that is inspired this project was conducted by Decroos and Davis [3] who apply Non-negative Matrix Factorization [16] as the technique for reducing the dimensions of player heatmaps. Since NMF can only learn the linear transformation of the features, to maintain linearity while performing the dimensionality reduction Decroos and Davis performed the NMF technique on individual action heatmaps. Hence, they performed the process separately for shot action, cross action, dribble action and pass action.

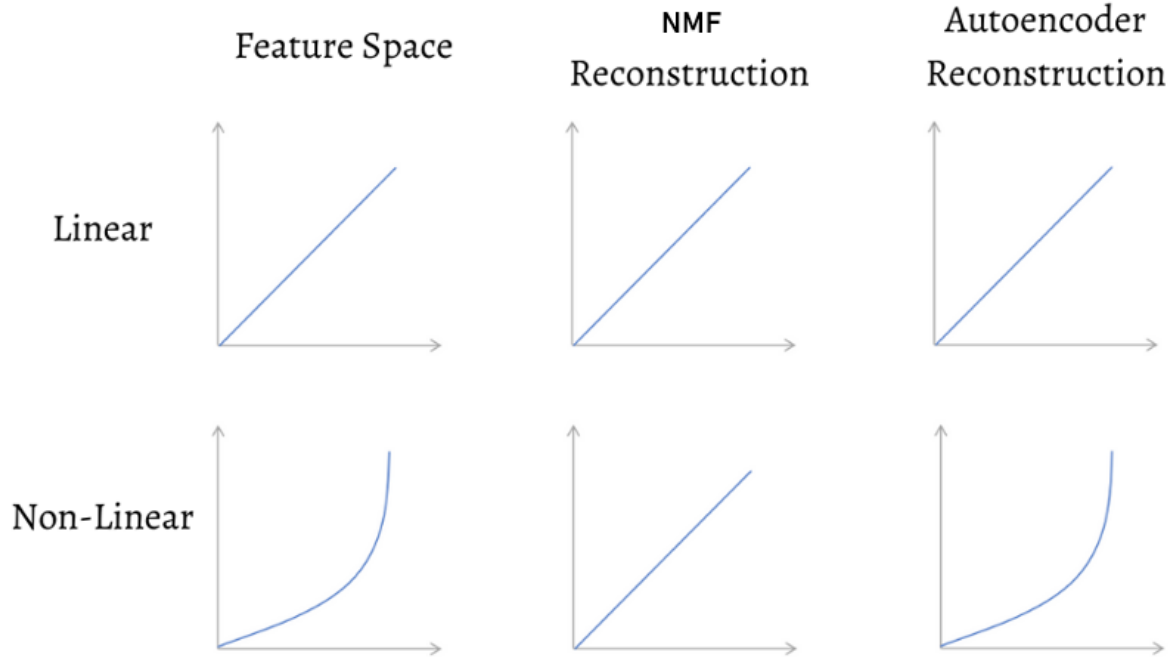


Figure 14 CAE vs NMF[17]

In such cases, Autoencoders can perform better than PCA/NMF because autoencoders are built based on neural networks, hence, can learn the non-linear transformation of the features. This turns into them having a better reconstruction ability [17]. Hence, in this project, we train the autoencoder with the heatmaps of all the action types at random. In this way, the model created is unbiased for any action type and has learned to correctly distinguish between actions and reconstruct the heatmaps accordingly.

## II. Loss function for Convolutional Autoencoder

The loss function for an autoencoder highly depends on the type of input data we want the autoencoder to adapt to. Based on the literature review, the loss function ‘Binary Cross-entropy’ works well when the values in input data are in the range 0 to 1, and since our heatmaps were normalized and smoothed using a Gaussian function to bring the values in the range 0 to 1, we chose to use Binary Cross-entropy as the loss function for our CAE models.

$$\text{Loss} = -\frac{1}{\text{output size}} \sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

Figure 15 Formulae of Binary Cross Entropy loss function[18]

As for the activation function for the last layer of the CAE models, the only activation function that compliments the binary cross entropy loss function is the Sigmoid function. The Sigmoid function is the only activation function that guarantees that independent outputs lie within the range of 0 to 1 [18, 19].

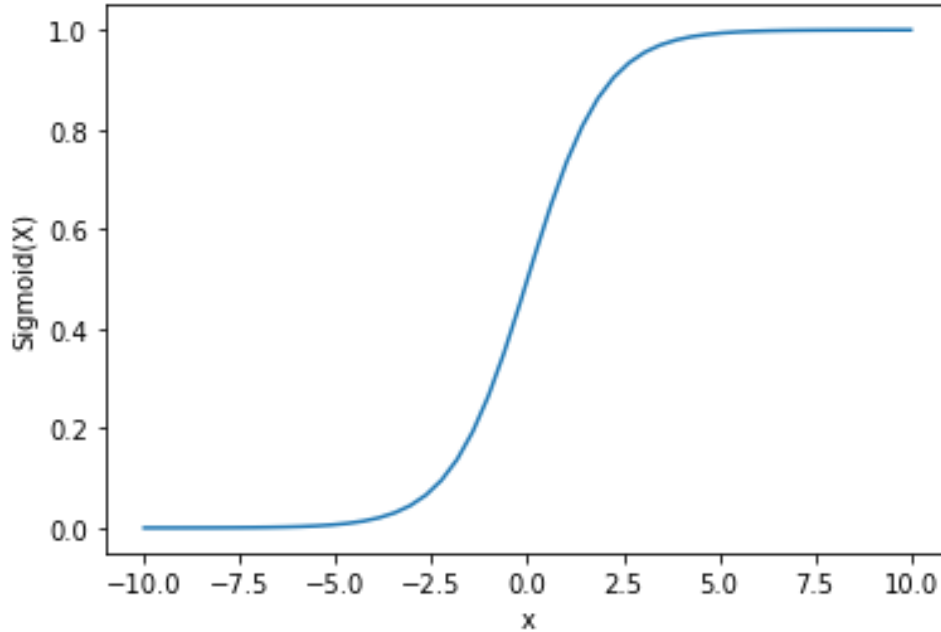


Figure 16 Sigmoid Activation Function always outputs the values in range 0 to 1[19]

Finally, we perform a visual judgement of the reconstructed heatmaps to the original heatmaps to make sure that the model is learning and distinguishing between different action types.

### III. Machine Learning Models

#### I. Baseline model

From the literature review, a baseline player vector was created using NMF for dimensionality reduction. The results of the technique on our dataset are showcased in the player retrieval task in the Evaluation section.

#### II. Software and Hardware Implementations

The CAE Models were implemented using the python language, with the TensorFlow package for deep learning modelling. Google Colab GPU was utilized to accelerate the learning process of Convolutional Autoencoder models.

#### III. Model Training Considerations

The training was conducted in batches of 8, 16, and 32. Having a batch size can improve the learning of the model [20]. Here, the batch size of 8 was learning less effectively than the batch size of 16 or 32. The batch size of 32 showed significant improvement in learning and was selected as default.

The models were tested on different optimization techniques – Adam and AdaDelta optimization functions. Although the AdaDelta optimization function showed a steady decrease in the training and testing losses, the reconstructed heatmaps did not resemble the original heatmaps. The

reconstructed heatmaps of the CAE models when tested with Adam optimizers showed promising results and hence, Adam was selected as the default optimizer technique.

Adam optimizer applies an adaptive learning rate; however, few models were fine-tuned using the TensorFlow Learning-rate Scheduler callback. The default learning rate works well in reconstructing the heatmaps, and also few fine-tuned models had a learning rate reduced by 10 times to find convergence and have better reconstruction of the original heatmaps.

The number of epochs was experimented for each model. However, an Early-Stopping Callback from TensorFlow was applied which stopped the model training if there was no learning for either 5 epochs or 10 epochs (depending on the model architecture).

#### IV. Experimentation

The model architectures created and tested were inspired from the work of Masci, et al. [21] to use Convolutional Autoencoders for hierarchical feature extraction. The design of the models tested were found to work well with three hidden layers for both encoder and decoder, containing a pair of Convolutional layers and Pooling layers. The model hyperparameters tested are provided in the Table 5.

Hyperparameter	Tested Values
Convolutional Layer number of Filters	512, 256, 128, 64, 32, 16, 8, 4
Convolutional Layers Filter sizes	(5x5), (3x3)
Activation functions	ReLU, LeakyReLU, Sigmoid
Pooling Layers	Max Pooling Layer, Average Pooling Layer
Pooling Layer size	(2x2)
Number of epochs	100, 200, 250, 500

*Table 5 Model hyperparameters subject to experimentation.*

The architecture of the best Convolutional Autoencoder model based on the visual accuracy of the reconstructed heatmaps was '*MODEL27\_CAE2*' (Figure 17). This model not only provides good reconstructed heatmaps but also gave the best results from the Player Retrieval Task which is explained in the Evaluation section.

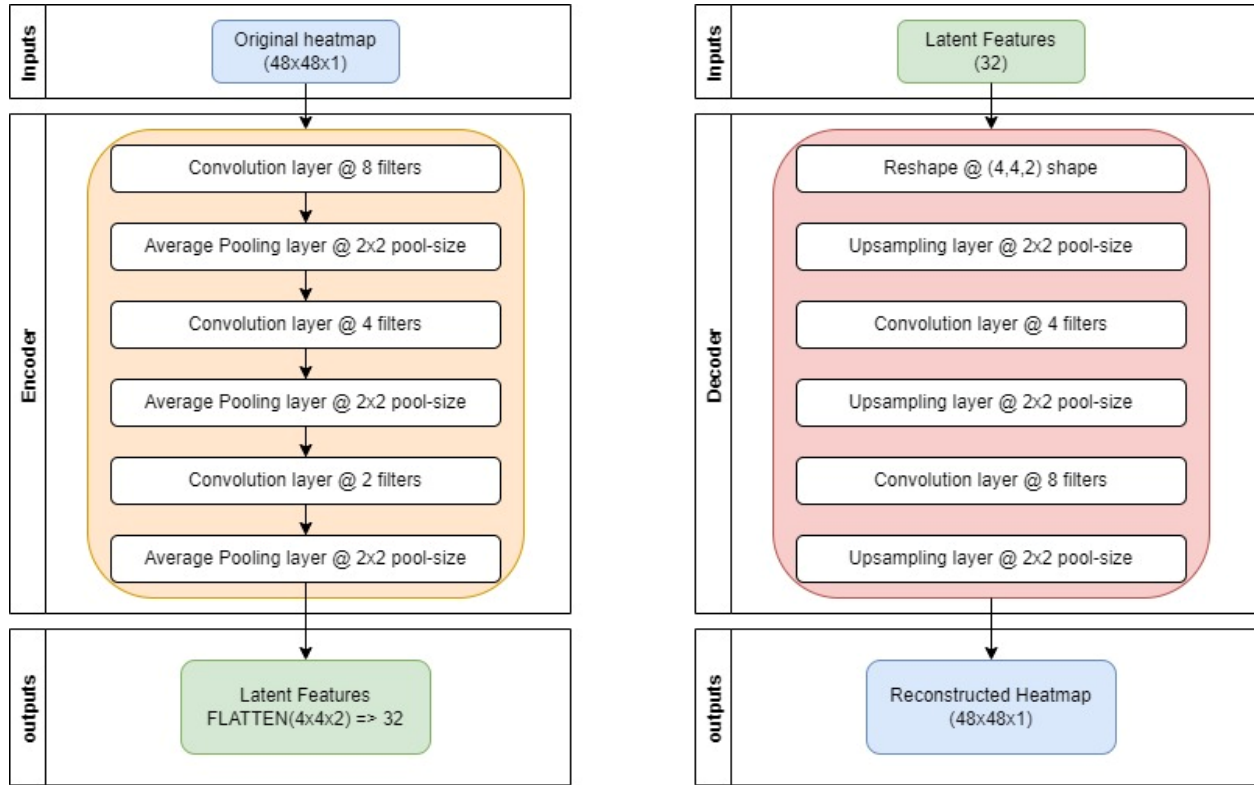


Figure 17 Architecture of CAE model 'MODEL27\_CAE2'

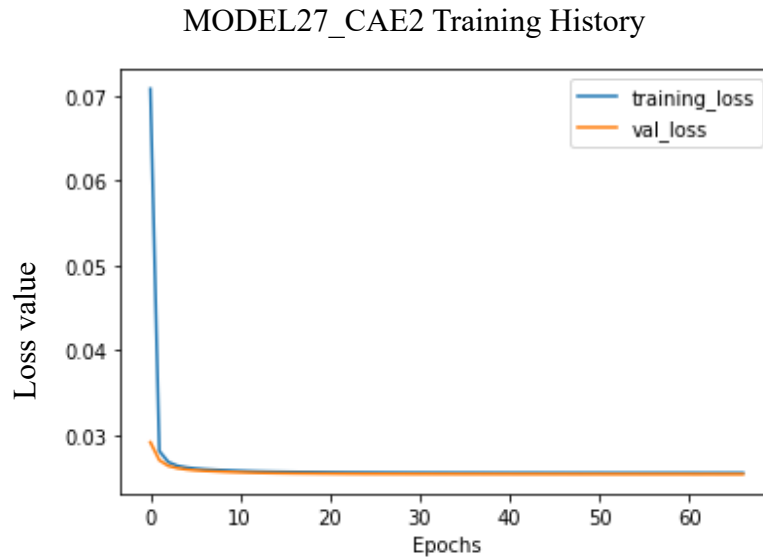
## V. MODEL27\_CAE2 Diagnostic Analysis

Based on literature review, the architecture of the 'MODEL27\_CAE2' was inspired to compose of:

1. **Input layer:** Input of the CAE was a heatmap of shape (48,48,1) where '1' is the number of channels.
2. **Encoder layer:** The inputted image is then sent to a convolutional layer with 8 filters of size (3x3), followed by an average pooling layer with pool-size (2x2). Then, a convolutional layer of 4 filters of size (3x3) followed by a similar average pooling layer. Finally, a convolutional layer of 2 filters was again followed by an average pooling layer. All the convolutional layers had 'LeakyReLU' as the activation function.
3. **Latent-Feature space:** The latent representation is just a flatten layer that converts the shape (4,4,2) into 32 dimensions.
4. **Decoder layer:** The latent features is first reshaped from 32 dimensions to (4,4,2) shape. Then using upsampling layer of size (2x2) and convolutional layer of same number of filters and sizes used in the encoder layer (but in reverse order), we recreate the heatmap with a shape of (48,48,8) where '8' is the number of channels.
5. **Output layer:** The output layer uses a convolutional layer with 'sigmoid' activation function and 1 filter to output a reconstructed heatmap of shape (48,48,1).

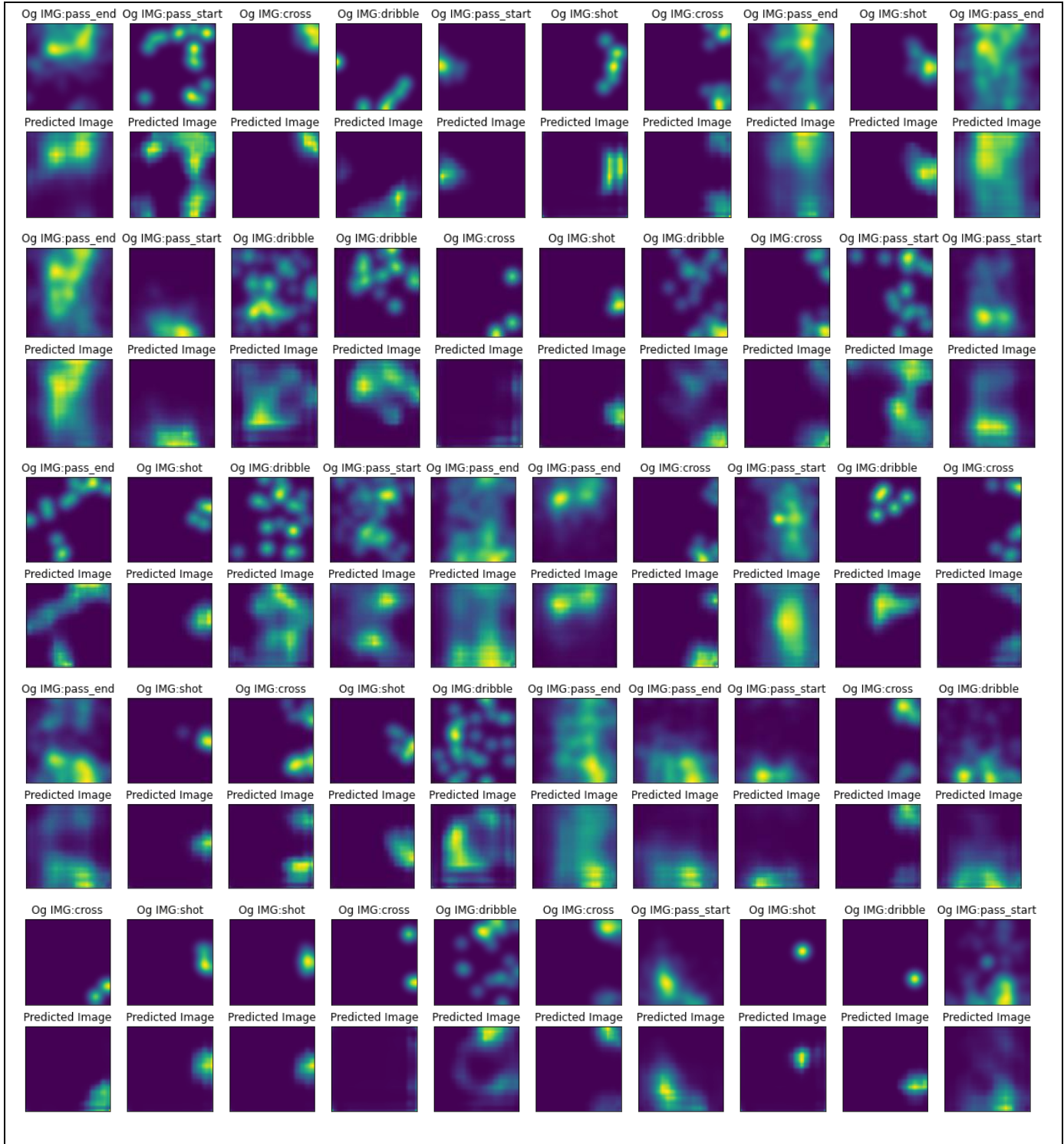


Figure 18 shows the training and testing losses for each epoch for CAE model MODEL27\_CAE2. The model was set to run for 100 epochs but stopped training after 67 epochs because of the Early-stopping callback. The model was trained on a batch size of 32. The final validation loss was reached at 2%.



*Figure 18 MODEL27\_CAE2 training and validation loss. The model stopped learning after 67 epochs on a batch size of 32. The validation loss is 2%.*

Figure 19 provides a comparison of the reconstructed heatmap stacked below the original heatmap. This visually shows the accuracy of the 'MODEL27\_CAE2' to reconstructed the heatmaps from a latent representation of only 32 dimensions.



*Figure 19 Examples of heatmaps reconstructed from the original heatmaps. The first row shows the original representation of the heatmaps followed by the row which contains the reconstructed heatmaps of the original image. The remaining rows follow this trend.*

## V. Evaluation

The lack of objective ground truth for defining a playing style makes it difficult to evaluate our proposed strategy. Therefore, we focus on addressing three key issues through our experiments: (1) Measuring the performance of the proposed strategy at the player retrieval task, (2) Providing football teams with a player similarity analysis tool to guide the players to improve their play style and provide more value to their teams, and, (3) Illustrating how our proposed method can be used for scouting and also assist in the decision-making process for player replacement task.

### Dataset

As discussed in the Data Preparation step, this project uses soccer match event data from Wyscout open access dataset[8]. First, the event data obtained from Wyscout is converted to actions using SPADL representation. Then, we filter the actions data to obtain player actions involved in the five national soccer competitions in Europe: first divisions in England, France, Germany, Italy and Spain from the 2017/2018 season. This is an adequate data range for our analysis with the final dataset having 2,612 players and 1,819,425 actions. A summary of the data is provided in Table 6.

Competition	No. of Matches	No. of Teams	No. of Players	No. of Actions
English first division	380	20	511	381652
French first division	380	20	541	371479
German first division	306	18	472	306036
Italian first division	380	20	532	391410
Spanish first division	380	20	556	368848

Table 6 Summary of Wyscout soccer ball event data

### I. Player Retrieval from Anonymized Event Stream Data

As playing style is a purely arbitrary construct with no objective basis, we would ordinarily lack any experimental techniques for tweaking these parameters. However, we can characterize playing style by retrieving players from anonymized match event stream data. To do this, first, we divide the final dataset into two sets: training and testing. We divide the dataset with the help of ‘*period\_id*’, meaning that the training set will contain the actions performed by the players in the first half of all the matches, whereas the test set will contain actions performed in the second half of all the matches. Hence, we have 930,792 (51.2%) actions in training set and 888,633 (48.8%) actions for the test set.

After dividing the final dataset into train and test sets, we will perform the preprocessing of the training set with the following parameters: (a) the heatmaps constructed with a grid size of 48x48, (b) the heatmaps then smoothed using Gaussian blur with a sigma value of 2.7 (value yielding good results found after several experimentations), (c) engineer the 10 action descriptors to use in the decision-making process (2 shot features, 2 cross features, 2 dribble features and 4 pass features), (d) extract the latent features from our proposed deep learning model (Convolutional Autoencoder), and (e) concatenate all the latent features for all the relevant actions to form the player vector. This preprocessed training set will act as a set of labelled player vectors that can be used to compare anonymized player vectors.

After constructing the labelled Player2Vec vectors from the training set, we obtain a set of anonymous actions performed by a target player from the test set and construct a Player2Vec vector based on the actions performed by the target player. Then, we compare this target player's Player2Vec to the labelled set of Player2Vec vectors and draw a top-k ranking of the most similar players to that target player (Figure 20). The quality of this ranking is the position of the unknown player in the ranking, which means that, if most players appear at the top of their own rankings, then the characterization of players in terms of their playing style has been successful and if most players do not appear at the top of their own rankings, then the proposed solution fails.

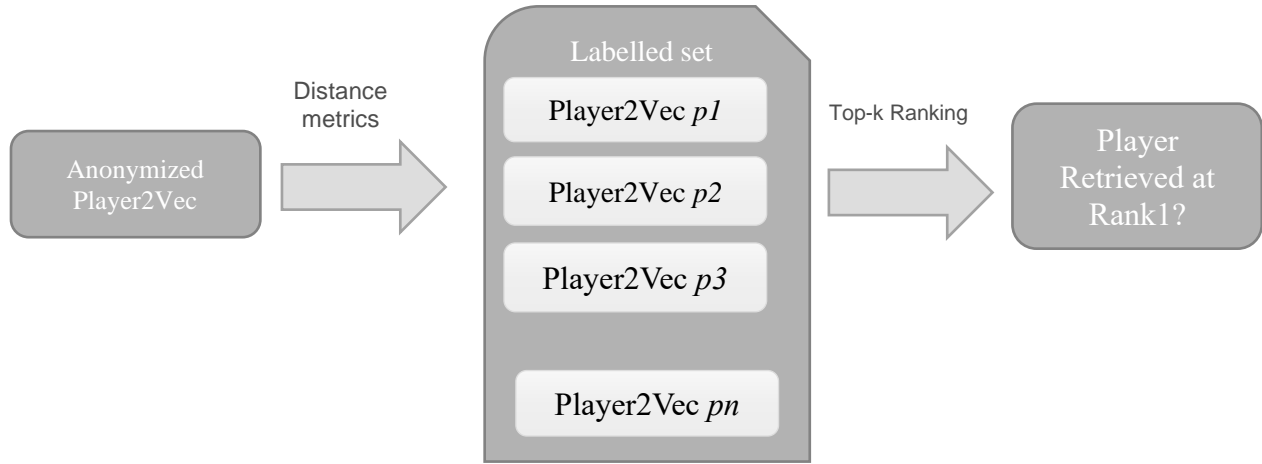


Figure 20 Player Retrieval Task

In our experiments, we further filtered our datasets to consider players that have played at least 900 minutes (10 matches). This criterion provided a total of 1,230 anonymized players which we labelled using the 1,230 players from the training set. Finally, we compared the accuracy of our task with the method projected by Decroos and Davis [3]. We found that our proposed technique of extracting latent features from a CAE model performs better than the NMF (Non-Negative Matrix Factorization) method utilized by the previous study and Table 7 provides the results showcasing the performance of both the solutions. Player2Vec was successful in retrieving 42.1% of all players with only one attempt. Player2Vec also shows improvement in the Mean Reciprocal Rank, with an increase of close to 8%.

Method	Top-1	Top-3	Top-5	Top-10	MRR
Decroos and Davis	30.3%	41.9%	48.8%	52.8%	0.469
<b>Player2Vec</b>	<b>42.1%</b>	<b>61.8%</b>	<b>69.6%</b>	<b>78.6%</b>	<b>0.546</b>

Table 7 Top-k results and mean reciprocal rank (MRR) comparison of the proposed solution to the previous study.

Table 8 shows the five best CAE models, that generate the Player2Vec, whose performance exceeds the benchmark set by Decroos and Davis's proposed method. 'Model27\_CAE2' performs the best among all the CAE models when comparing with the Manhattan distance measure. We also experimented with different distance measures and found that Euclidean distance improves

the results for model 'Model27\_CAE2' but is inconsistent in performance for other models compared to Manhattan distance metric.

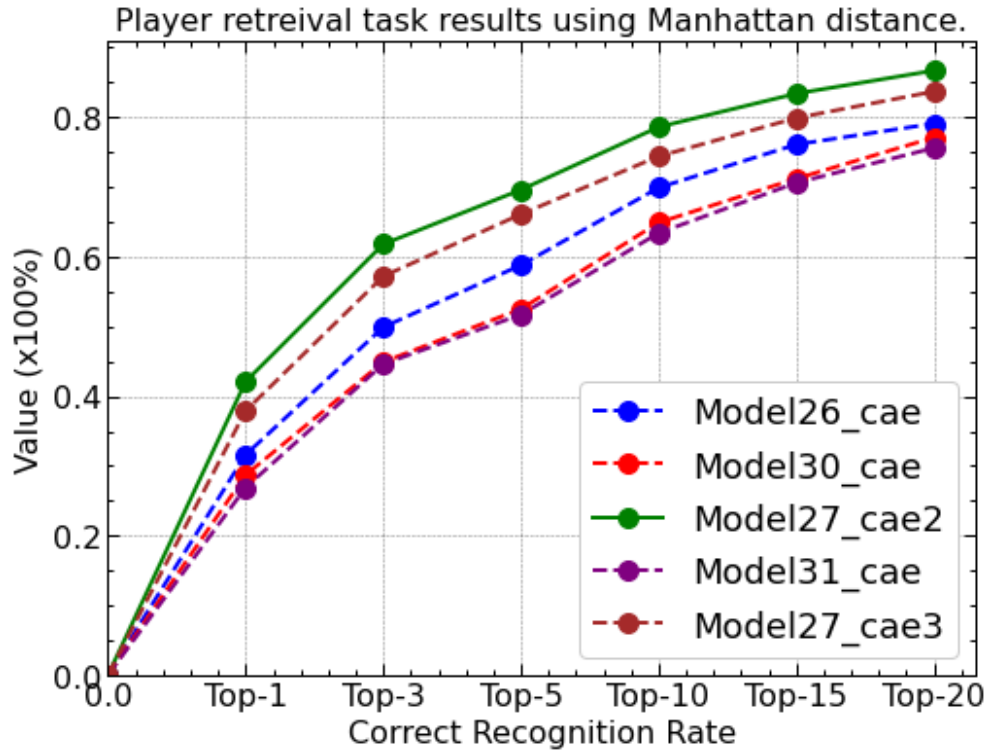


Figure 21 Comparison of five best performing models that generate the Player2Vec player vector. The models were tested for player retrieval task using Manhattan distance. 'MODEL27\_CAE2' (green) performs better than the rest of the models (dashed lines).

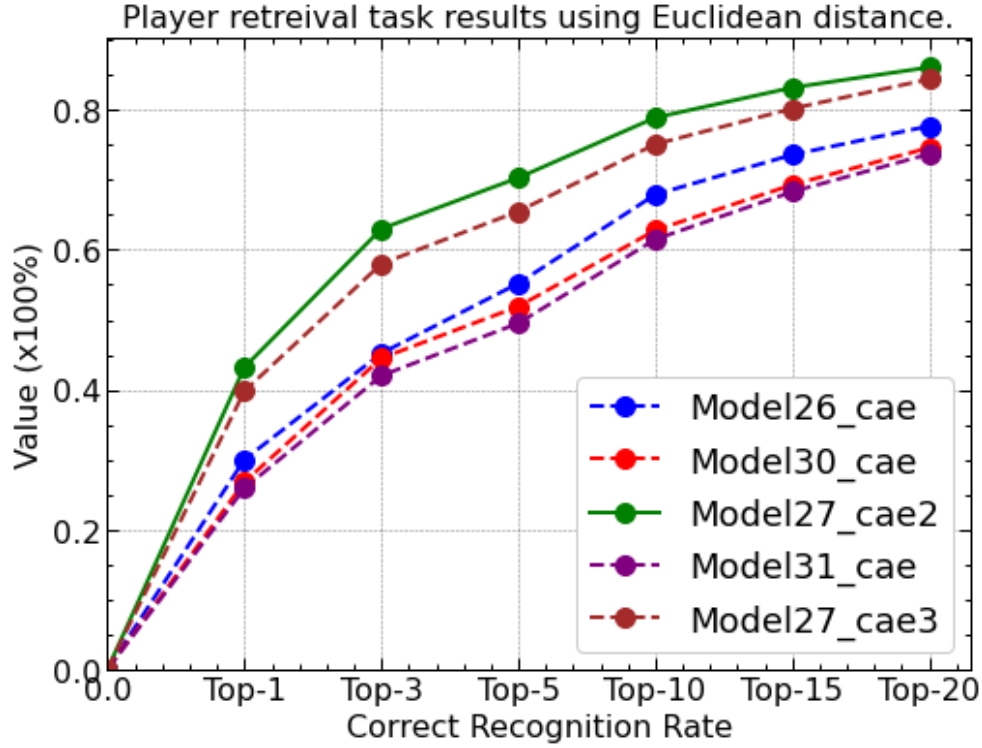


Figure 22 Comparison of five best performing models that generate the Player2Vec player vector. The models were tested for player retrieval task using Euclidean distance. 'MODEL27\_CAE2' (green) performs better than the rest of the models (dashed lines).

Lastly, Figure 21 shows the Correct Recognition Rates (CRR) for the player retrieval task when the best five CAE models, that generated the Player2Vec vectors, were tested using the Manhattan distance and Figure 22Error! Reference source not found. shows the CRR for the player retrieval task when the same five models were tested using the Euclidean distance.

Model Name	Top-1	Top-3	Top-5	Top-10	Top-15	Top-20
model26_cae	31.6%	49.9%	58.8%	69.8%	76%	79%
model30_cae	28.7%	44.9%	52.5%	64.9%	71.2%	77.1%
<b>model27_cae2</b>	<b>46.1%</b>	<b>61.8%</b>	<b>69.6%</b>	<b>78.6%</b>	<b>83.4%</b>	<b>86.7%</b>
model31_cae	26.9%	44.6%	51.7%	63.4%	70.6%	75.5%
model27_cae3	37.9%	57.2%	66.2%	74.5%	80%	83.8%

Table 8 Performance comparison of the five best CAE models

## II. Player Similarity Analysis

Soccer clubs often have to look for replacements for their aging players or the players that may leave the club during the transfer window. The recruitment team within the clubs have to work hard to replace a vacancy caused by the departing player. We propose player similarity analysis and use the Player2Vec player vector to find the top-k similar players to the target player (player departing the football club). To create the pool of data, we compute and compare the Player2Vec vectors of all the 1,441 players from our dataset that have played at least 900 minutes of match time i.e., 10 games, in the 2017/18 season of the five major soccer competitions in Europe. We investigate some popular claims in popular media about similar players and found the below results:

1. Lionel Messi is deemed to be one of the best players to touch a football ball. It has been often speculated that Paulo Dybala is a player that is similar to his fellow Argentina compatriot Lionel Messi in terms of their playing style [22, 23]. When ranked using the Player2Vec vectors, we found that Dybala was placed at rank 4<sup>th</sup> for the most similar players to Lionel Messi (Table 9).
2. Luka Modrić, a legendary player who plays for Real Madrid Football Club, is often compared to Cesc Fabregas because of his great vision and creative playmaking skills that resemble Luka Modrić [24]. When tested using our Player2Vec similarity analysis Fabregas was placed as the 2<sup>nd</sup> most similar player to Luka Modrić (Table 10).
3. Idrissa Gueye (midfielder at Everton FC as of the 2017/18 season) has been often compared N’golo Kante (midfielder at Chelsea FC as of 2017/18 season) by many journalists [25-27]. Gueye came 4<sup>th</sup> most similar player to N’golo Kante (Table 11).
4. Son Heung-Min came 4<sup>th</sup> most similar player to Cristiano Ronaldo (Table 12). Cristiano Ronaldo is one of the best attackers of this generation and Son Heung-Min has been seen to resemble the play style of that of Cristiano Ronaldo by sports media [28].
5. Aymeric Laporte has been considered to be a long-term replacement for the aging Sergio Ramos from Real Madrid FC [29]. Even Ramos acknowledges the claims of Laporte being considered similar to him in play style [30, 31]. Laporte ranked in 18<sup>th</sup> place as the most similar player to Ramos using our Player2Vec vector (Table 13). While 18<sup>th</sup> position is still good out of 1,441 players, this example does demonstrate that our approach is better at characterizing offensive playing styles than defensive styles, as defending is often more about positioning than on-the-ball actions.

Player2Vec can be used in player similarity analysis experiment and then the reconstructed heatmaps of all the relevant actions from the Player2Vec vector can be used further for the following applications:

1. For any player in their clubs, find a list of similar players to the target player.
2. Comparing the play style of a target player with their similar players and using the reconstructed heatmaps of all the relevant actions to help in identifying the areas for development in the target player.
3. Monitoring a player’s development by comparing the difference in the performance of that player throughout two or more simultaneous seasons by visualizing the reconstructed heatmaps of that player for the two or more seasons.



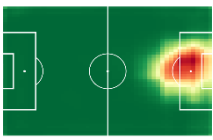
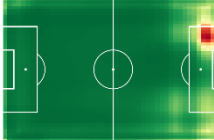

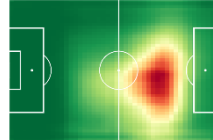
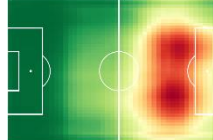
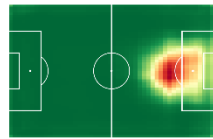

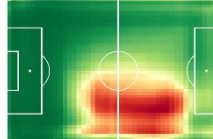
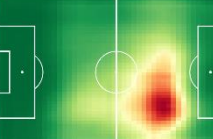

Player Name	Shot	Cross	Dribble	Pass start	Pass end
Lionel Messi					
Paulo Dybala					

Table 9 Player Similarity Analysis of Messi and Dybala.

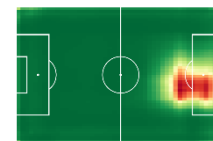
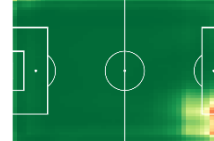
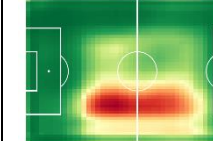
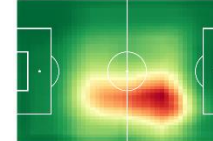
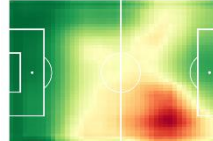
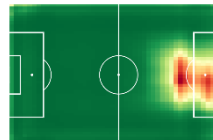
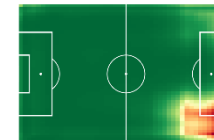

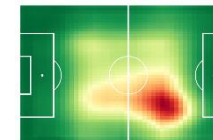

Player Name	Shot	Cross	Dribble	Pass start	Pass end
Luka Modrić					
Cesc Fabregas					

Table 10 Player Similarity Analysis of Modric and Fabregas.

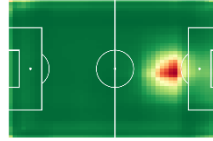
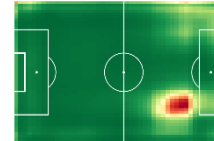
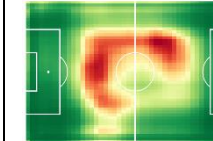
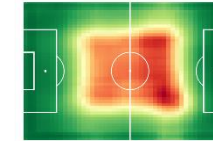
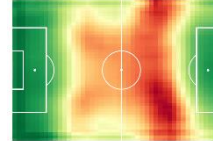
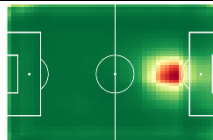
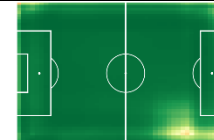


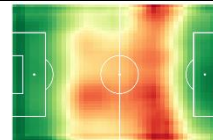
Player Name	Shot	Cross	Dribble	Pass start	Pass end
N'golo Kante					
Idrissa Gueye					

Table 11 Player Similarity Analysis of Kante and Gueye.



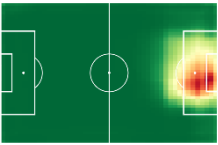
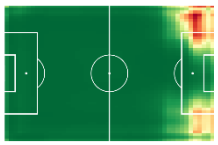
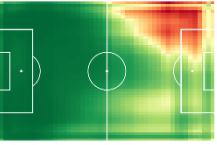
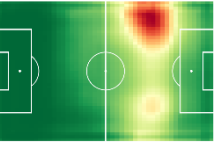
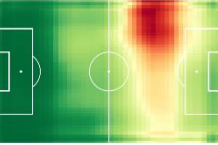
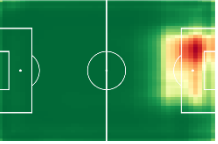
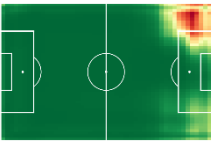
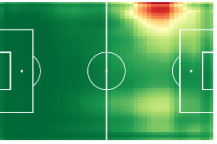
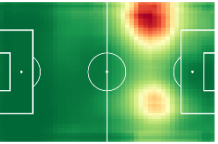
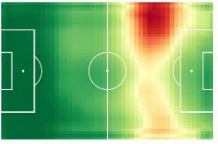
Player Name	Shot	Cross	Dribble	Pass start	Pass end
Cristiano Ronaldo					
Son Heung-Min					

Table 12 Player Similarity Analysis of Ronaldo and Son.

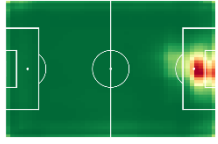
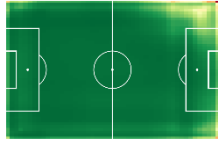
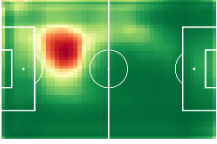
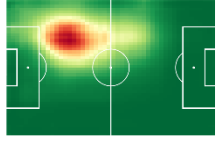
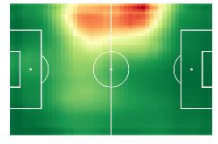
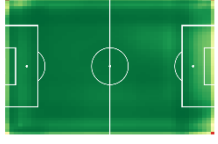

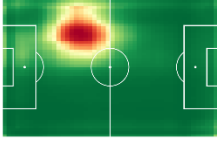
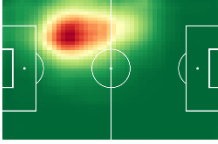

Player Name	Shot	Cross	Dribble	Pass start	Pass end
Sergio Ramos					
Aymeric Laporte					

Table 13 Player Similarity Analysis of Ramos and Laporte.

### III. Player Replacement Task

Deciding the replacement of a player in a football club can be quite a challenging task for any team. The risks involved in recruiting a new player are high, since, a lot of money is used to buy a player in the transfer market. Money well used can result in the success of the football club, but poor judgement and lack of quality research in recruiting new players can result in wasting the money invested in the transferred player if the player fails to fit in the team tactics.

To help the recruitment team of a Football club in deciding the right player to buy, we can utilize the Player2Vec player similarity analysis technique and the ten action descriptors that were engineered during the Data Preparation step. We can easily compare the performance impact of the new player being recruited with a player who currently plays at the same position in the team and make a judgement of the potential of the new player to fit the tactics of the team.

To demonstrate this utility, we can take the example of the Arsenal Football Club. Arsenal Football Club (FC) has a possessional style of playing meaning that the team concentrates on passing the ball among their players and not letting the opponents capture possession of the ball. This tactic requires its players to be good on the ball and not lose control of the ball more often. Based on this tactic criteria, the recruitment team of Arsenal FC was employed to find a striker similar to the aging 31 years old (age as of 2017) ‘Olivier Giroud’. Arsenal FC bought a new striker, in the summary of 2017, ‘Alexandre Lacazette’ from Olympiacos Lyonnais [32]. We can use the Player2Vec vector of ‘Olivier Giroud’ to identify the top five players similar to him (Table 14) and we can also visualize the play style using their heatmaps (Table 15). Table 14 shows that Alexandre Lacazette is the second most similar player to Olivier Giroud after Sandro Wagner. We can then also look at the ten action descriptors of the three players and compare Olivier Giroud to the two players (Figure 23 (a), (b)). Interestingly, we can notice the reason for Arsenal FC to decide on buying ‘Alexandre Lacazette’ over ‘Sandro Wagner’. Since Arsenal is a team that focuses on the ‘passing’ skill of the players, its players must be proficient in passing and compared to Sandro Wagner, Alexandre Lacazette shows more promising passing skills than him. The similarity of Lacazette to Giroud is more for passing attributes such as ‘pass-forward’, ‘pass-short’ than Sandro Wagner. Hence, even though Player2Vec shows that Sandro Wagner is more similar to Olivier Giroud than Alexandre Lacazette, based on the team tactics of Arsenal FC, it was more sensible to buy Alexandre Lacazette than Sandro Wagner.

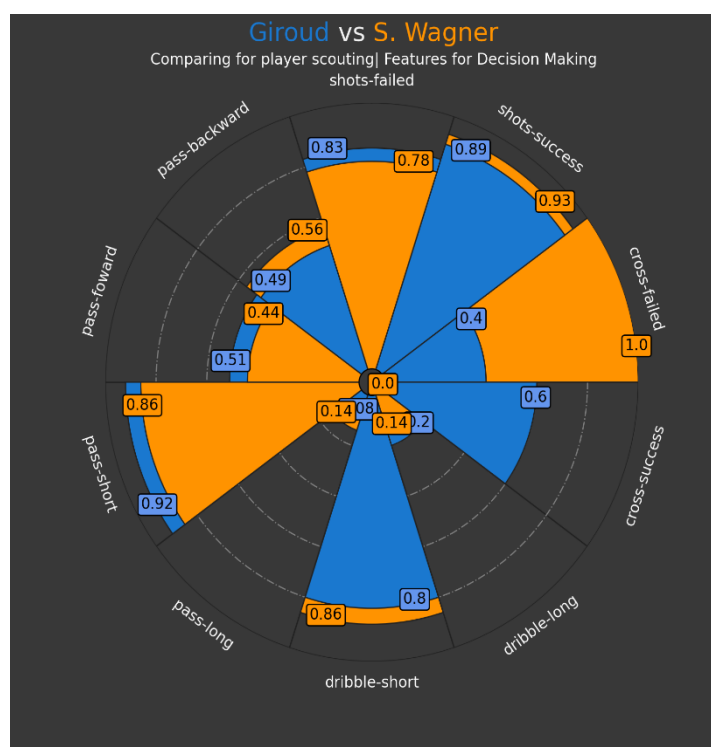
This transfer was proved as a success for Arsenal FC and ‘Alexandre Lacazette’ was not only able to fit in the team but also later on in his career went on to become the captain of the Arsenal team. A real-life scenario such as this shows how using Player2Vec and the decision-making attributes can help in deciding the right player to fit a team’s tactics.

Player Name	Difference in similarity value (error)
Sandro Wagner	0.386814
Alexandre Lacazette	0.391419
Edin Dzeko	0.394
Gabriel Jesus	0.409982
Edinson Cavani	0.419891

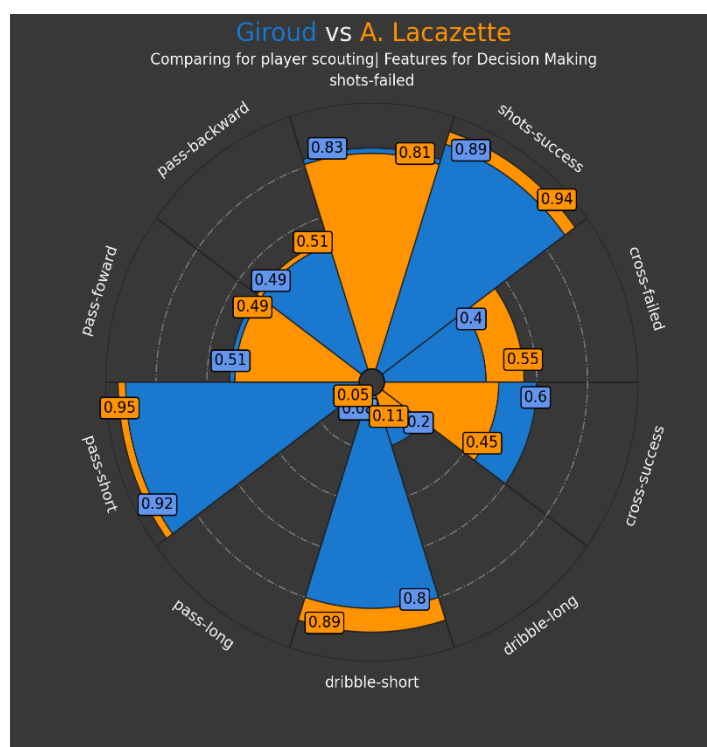
*Table 14 The names of the top 5 most similar players to the striker Olivier Giroud and the difference in the similarity in numbers.*

Player Name	Shot	Cross	Dribble	Pass_start	Pass_end
Olivier Giroud					
Sandro Wagner					
Alexandre Lacazette					

Table 15 Player similarity analysis of Olivier Giroud. Sandro Wagner is ranked the most similar player to Giroud. Lacazette is the 2nd most similar player to Giroud.



(a)



(b)

Figure 23 (a) Shows the comparison of the ten action descriptors of Olivier Giroud and Sandro Wagner. (b) shows the comparison of the ten action descriptors of Olivier Giroud and Alexandre Lacazette. The 'pass-forward' and 'pass-short' attributes show the closer similarity of Lacazette to Giroud than Wagner, based on the tactics of Arsenal FC.

## VI. Case-Study Analysis

### Liverpool FC Champions league triumph 2018/19 season

Liverpool FC is a well-known club in the football community with world-class players playing for the team. In the summer of 2018, Liverpool had to see the departure of one of their important players 'Emre Can'. Emre Can's contract with the club ended, and he decided to move to a different club 'Juventus'.



*Figure 24 Fabinho lifting the Champions League Trophy after playing a crucial role for Liverpool FC during the 2018/19 season campaign[33].*

Now, after losing Emre Can, Liverpool FC was in desperate need of a new holding midfielder. Liverpool recruited Fábio Henrique Tavares as known as 'Fabinho' from Monaco FC. Since joining Liverpool in 2018, Fabinho has proven to be a crucial part of Liverpool's tactical game. Liverpool FC was unbeaten in 21 games of the first 25 matches played by Fabinho. He was such a good holding midfielder for Liverpool FC, during the 2018/19 season, that he was considered the best holding midfielder in English Premier League for that season [33]. With a transfer so impactful as this, it will be merely impossible to consider that it was a lucky bet. Surely, Liverpool FC's recruitment team had to have thorough research to consider the players for replacing the departing player Emre Can.

### Why did Liverpool FC replace Emre Can with Fabinho?

With the quality of signings that Liverpool FC have done over the past few years, it is safe to say that the team responsible for recruiting the new players is doing a huge background check on the skills and attributes of the players playing style. Liverpool FC has a high-pressing, counter-

attacking playing style that has its players tirelessly running on the pitch for the whole match. Although the technique of Liverpool FC to find the players fit for their team is not known to everyone, we can use the Player2Vec player vectors of ‘Emre Can’ and ‘Fabinho’ to illustrate the similarity of play style between the two and also use the ten decision-making attributes to understand the possible reasoning behind Liverpool FC recruiting Fabinho as the replacement for Emre Can. Table 16 shows the similarity between the two players using the reconstructed heatmaps from both players’ Player2Vec vector.

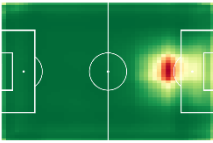
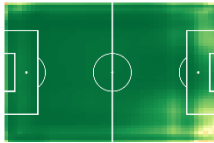
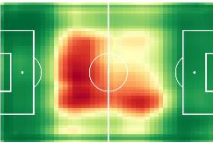
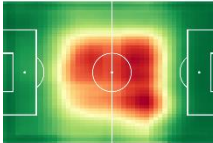
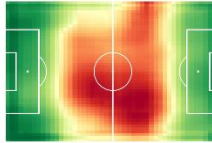
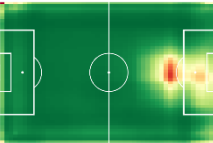
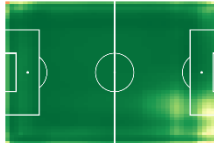
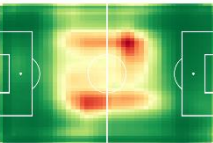
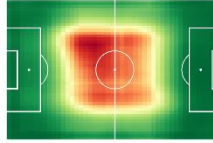
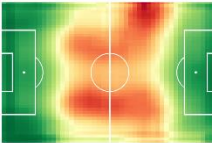
Player Name	Shot	Cross	Dribble	Pass start	Pass end
Emre Can					
Fabinho					

Table 16 Player similarity between Emre Can and Fabinho

Figure 25 shows the comparison of the decision-making attributes of the two players – Emre Can vs Fabinho. All the attributes show the closeness of value of Fabinho to the departing player Emre Can. The ‘*pass-forward*’ attribute, which explains the forward passing skill of a player, proves that the skill of Fabinho is also an effective playmaker who has a great vision of the football field to make forward ball progressive passes.

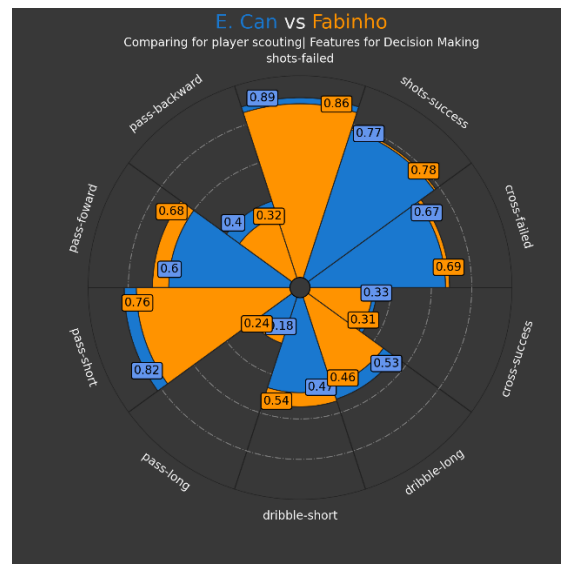


Figure 25 Comparison of the ten action descriptors of Emre Can and Fabinho.



Figure 26 shows that our claim is supported by popular sports media reporter Gary Neville of ‘Sky Sports’, who says "A lot of holding midfield players play horizontally and shuffle across but Fabinho can play vertically with his passes". Figure 26 shows the ability of Fabinho to make forward passes on the football pitch that has helped Liverpool going forward in attack. In this way, Liverpool has not only replaced Emre Can with Fabinho but also improved the team both defensively (Fabinho’s main role in midfield) and offensively.



*Figure 26 Fabinho's passing sonar for Liverpool in the 2019/20 Premier League season[33].*

This case study analysis proves how, with the help of Player2Vec player similarity analysis and the ten action descriptors, a football team can not only find players with similar play styles but also help in making reliable decisions in recruiting new players for their club.

## VII. Discussions

This project presents a novel approach that aims to characterize a player based on his playing style by considering the actions the player performs on the football pitch. One of the main contributions of the project is demonstrating the use of ten action descriptors – Shots Failed, Shots Success, Crosses Failed, Crosses Success, Long Distance Dribbles, Short Distance Dribbles, Long Distance Passes, Short Distance Passes, Forward Direction Passes and Backward Direction Passes, to help football teams in making reliable decisions for recruiting or replacing old players with new players based on their ability to fit the team and the team tactics. We also present the Player2Vec player vector that is both human-interpretable, by visualizing all the actions performed by a player through the reconstruction of the player heatmaps from the player vector, and also suitable for data analysis. From the evaluation, we find that our approach of comparing the Player2Vec vectors of players is valid in both qualitative and qualitative ways.

The player retrieval task explains that our proposed method outperforms the existing approach of Decroos and Davis [3]. We successfully retrieved 42.1% of all players with only one attempt and 86.7% of all players in the Top-20. The performance improvement of the Player2vec solution to retrieve the Top-1 list is a close to 10% increase than the previous study. We also showed improvement in the Mean Reciprocal Rank with our proposed method giving an MRR of 0.546 compared to the MMR of 0.469 in the previous study. This suggests that our Player2Vec vector efficiently captures the players' unique characteristics.

We also showed a player similarity analysis illustrating how a soccer team can use Player2Vec to search for similar players to a specific player. We provided five examples of the player similarity analysis task. First, we justified Dybala's comparison to Messi by showing the similarity in play style between the two. Both show actions in similar areas (Table 9) with a strong dribbling play style on the right-hand side of the football pitch. Second, we compare a midfielder 'Luka Modric' with another midfielder 'Cesc Fabregas' and we can identify that both are similar playmakers who have similar passing styles (Table 11). Third, we again compare two midfielders but with way different styles than the players before. From Table 11, we witness the similarity of N'golo Kante and Idrissa Gueye which justifies the opinion of the media about them being similar in play style. Fourth, we compare Cristiano Ronaldo and Son Heung-Min to see the similarity in their playing styles (Table 12). Lastly, we compare Aymeric Laporte and Sergio Ramos and explain, with Table 13, that our approach is better at characterizing offensive playing styles than defensive styles, as defending is often more about positioning than on-the-ball actions.

Further, we utilized the ten action descriptors for the decision-making process. We showed how Arsenal FC benefitted by picking up Alexandre Lacazette during the transfer period of summer 2017 and in the process, we showed how our Player2vec similarity analysis can be used in conjecture with the ten decision-making features to make reliable decisions.

However, the limitations of the proposed method exist. First, although a football team can utilize the ten action descriptors for deciding the right player for their team, we did not consider a team's tactic that a player is involved in when performing the similarity analysis. The actions taken by the player in a match can be greatly influenced by team tactics and hence can be significant in the development of the Player2Vec player vector. The recent work of Decroos, et al. [34] uses an

approach to capture both playing styles of teams and players and showed improved accuracy of a mixture model including the direction of actions.

Second, although we show the use case and reliability of the ten action descriptors, it is still just a concept that emerged from the qualitative analysis and domain knowledge. More data and research are required to statistically prove the robustness of these descriptors.

Third, the Player2Vec vector is fundamentally formed by overlaying a grid over the football field and counting the number of actions performed in each grid cell but the approach has a downside to it. The flaw is that choosing the optimum grid resolution is difficult since a coarse grid misses key variations between locations, and a finer-grained grid substantially increases data sparsity because then fewer actions happen in a single grid cell.

Lastly, the high-dimensional nature of the Player2vec vector prevents the information from being interpreted immediately. It is difficult to separate certain traits from the generated vector. For instance, because a player's shot actions are flattened and compressed in a heatmap along with the other actions, it is challenging to detect them with Player2Vec alone. Additionally, the vector's length of 32 dimensions makes it difficult for straightforward observation or comparison, necessitating the intervention of sports scientists to link the outcomes of real performance and the practical influence on players' play styles.



## VIII. Conclusions

We used deep learning to characterize a player’s play style and also engineered action descriptors to help in the process of decision-making during player recruitment and replacement tasks. We created the Player2Vec player vector by first identifying the relevant actions – passes, crosses, shots and dribbles, performed by each player on a field, then by counting, normalizing and smoothening, we developed the heatmaps of actions performed by the player. Since the created heatmap was of 48x48 in size (computationally expensive to perform any data analysis), we compressed the heatmap into an array of 32 records by reducing the dimensionality using the Convolutional Autoencoder. In the end, the Player2Vec was formed by concatenating the five action vectors which resulted in the total size of the vector being compressed from 2,304x5 (11,520) to just 160 dimensions. This Player2Vec player vector offers a complete view of a player’s playing style (within the limits of the data source), is human-interpretable and can be used in data analytics tasks.

We then discussed the improvements in the accuracy of our proposed method to the existing research and demonstrated a real-life application by performing player similarity analysis and player recruitment tasks. Lastly, we performed a case study to understand the tactical thinking that undergoes inside football clubs before recruiting new players into the club. Hence, capturing the playing style of players in soccer can be leveraged in areas such as player scouting, player development monitoring, and match preparation. We showed how to construct player vectors by transforming sets of actions from match event stream data to fixed-size Player2Vec vectors.

For future work, it would be interesting to see the performance of the Player2Vec approach for different sports such as Rugby, Cricket etc. Also, we can utilize better robust metrics such as xG [2], Expected Threat (xThreat) [35] and VAEP [11, 36] for decision-making during the recruitment process.

## IX. References

- [1] B. SPORT. "BBC SPORT | Football | Laws & Equipment | Pitch dimensions." [http://news.bbc.co.uk/sport1/hi/football/rules\\_and\\_equipment/4200666.stm](http://news.bbc.co.uk/sport1/hi/football/rules_and_equipment/4200666.stm) (accessed 10/08/2022).
- [2] H. Eggels, R. van Elk, and M. Pechenizkiy, "Expected goals in soccer: Explaining match results using predictive analytics," in *The machine learning and data mining for sports analytics workshop*, 2016, vol. 16.
- [3] T. Decroos and J. Davis, "Player vectors: Characterizing soccer players' playing style from match event streams," in *Joint European conference on machine learning and knowledge discovery in databases*, 2019: Springer, pp. 569-584.
- [4] T. Decroos, J. Van Haaren, V. Dzyuba, and J. Davis, "STARSS: a spatio-temporal action rating system for soccer," in *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*, 2017, vol. 1971: Springer, pp. 11-20.
- [5] StatsBomb. <https://statsbomb.com/> (accessed.
- [6] O. Sport. <https://www.statsperform.com/opta/> (accessed.
- [7] Wyscout. <https://wyscout.com/> (accessed.
- [8] L. Pappalardo *et al.*, "A public data set of spatio-temporal match events in soccer competitions," *Scientific data*, vol. 6, no. 1, pp. 1-15, 2019.
- [9] W. P. Coordinates. "Pitch coordinates." [https://dataglossary.wyscout.com/pitch\\_coordinates/](https://dataglossary.wyscout.com/pitch_coordinates/) (accessed.
- [10] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1851-1861.
- [11] T. Decroos and J. Davis, "Valuing on-the-ball actions in soccer: a critical comparison of XT and VAEP," in *Proceedings of the AAAI-20 Workshop on Artificial Intelligence in Team Sports*, 2020: AI in Team Sports Organising Committee, pp. 1-8.
- [12] Socceraction. "SPADL Pitch Coordinates." <https://socceraction.readthedocs.io/en/latest/documentation/SPADL.html> (accessed.
- [13] M. Maggipinto, C. Masiero, A. Beghi, and G. A. Susto, "A convolutional autoencoder approach for feature extraction in virtual metrology," *Procedia Manufacturing*, vol. 17, pp. 126-133, 2018.
- [14] S. Saikia. "Building a Convolutional Autoencoder with Keras using Conv2DTranspose." <https://medium.com/analytics-vidhya/building-a-convolutional-autoencoder-using-keras-using-conv2dtranspose-ca403c8d144e> (accessed 24/08/2022).
- [15] F. Chollet. "Building Autoencoders in Keras." <https://blog.keras.io/building-autoencoders-in-keras.html> (accessed.
- [16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.
- [17] R. Winastwan. "Autoencoders in Practice: Dimensionality Reduction and Image Denoising." <https://towardsdatascience.com/autoencoders-in-practice-dimensionality-reduction-and-image-denoising-ed9b9201e7e1> (accessed 13/08/2022).
- [18] K. Center. "Binary crossentropy." <https://peltarion.com/knowledge-center/modeling-view/build-an-ai-model/loss-functions/binary-crossentropy> (accessed 29/08/2022).

- [19] J. Verma. "The Sigmoid Activation Function - Python Implementation." <https://www.digitalocean.com/community/tutorials/sigmoid-activation-function-python> (accessed 29/08/2022).
- [20] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489*, 2017.
- [21] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International conference on artificial neural networks*, 2011: Springer, pp. 52-59.
- [22] GOAL. "Messi admits difficulties in Dybala partnership: He plays like me at Juve." <https://www.goal.com/en/news/messi-admits-difficulties-in-dybala-partnership-he-plays-like-me-/1uq96ju5zageb1sl1vez93oms3> (accessed).
- [23] R. Smith. "Is Paulo Dybala the Next Lionel Messi? 'He Can Go as High as He Likes'." <https://www.nytimes.com/2017/04/10/sports/soccer/paulo-dybala-juventus-lionel-messi-barcelona.html> (accessed).
- [24] A. Hanagudu. "6 players who can replace Luka Modric at Real Madrid." <https://www.sportskeeda.com/football/6-player-replace-luka-modric-real-madrid/3> (accessed).
- [25] S. Callaghan. "EVERTON BOSS WAS SPOT-ON WITH IDRISSA GUEYE – N'GOLO KANTE COMPARISON." <https://www.hitc.com/en-gb/2018/04/12/evertton-boss-was-spot-on-with-idrissa-gueye-ngolo-kante-comparis/> (accessed).
- [26] GOAL. "Everton boss Sam Allardyce compares Idrissa Gueye to N'Golo Kante." <https://www.goal.com/en/news/evertton-boss-sam-allardyce-compares-idrissa-gueye-to-ngolo/gddgazktcl3bl1ayeadrvalol8> (accessed).
- [27] A. Kleebauer. "Everton's Idrissa Gueye is the new N'Golo Kante - and here are the stats to prove it." <https://www.liverpoolecho.co.uk/sport/football/football-news/everttons-idrissa-gueye-new-ngolo-12965076> (accessed).
- [28] S. Ajith. "5 players who have a similar playing style to Cristiano Ronaldo." <https://www.sportskeeda.com/football/5-players-similar-playing-style-cristiano-ronaldo-mbappe-rashford> (accessed).
- [29] T. COLLINS. "4 Possible Replacements Should Real Madrid Sell Sergio Ramos." <http://bleacherreport.com/articles/2509541-4-possible-replacements-should-real-madrid-sell-sergio-ramos#slide3> (accessed).
- [30] J. Baiafonte. "Sergio Ramos Has Identified The Two Players Who Can Replace Him." <https://www.sportbible.com/football/news-transfers-sergio-ramos-has-identified-the-two-players-who-can-replace-him-20171217> (accessed).
- [31] L. Prenderville. "Sergio Ramos 'identifies Aymeric Laporte and Matthijs de Ligt as his long-term replacements' at Real Madrid." <https://www.mirror.co.uk/sport/football/transfer-news/sergio-ramos-identifies-aymeric-laporte-11710624> (accessed).
- [32] B. News. "Alexandre Lacazette joins Arsenal for club record £46.5m from Lyon." <https://www.bbc.co.uk/sport/football/40496970> (accessed).
- [33] A. Bate. "Why Fabinho is now the Premier League's best holding midfielder." <https://www.skysports.com/football/news/15117/11812068/why-fabinho-is-now-the-premier-leagues-best-holding-midfielder> (accessed).
- [34] T. Decroos, M. V. Roy, and J. Davis, "SoccerMix: representing soccer actions with mixture models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2020: Springer, pp. 459-474.

- [35] K. Singh. "Introducing Expected Threat (xT)." <https://karun.in/blog/expected-threat.html> (accessed 28/07/2022).
- [36] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "VAEP: an objective approach to valuing on-the-ball actions in soccer," in *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20*, 2020: International Joint Conferences on Artificial Intelligence Organization, pp. 4696-4700.