Annals of Data Science
https://doi.org/10.1007/s40745-021-00323-2

An Assessment of Football Through the Lens of Data Science

Poojan Thakkar[1] **· Manan Shah**[2]

Received: 28 January 2020 / Revised: 4 February 2021 / Accepted: 17 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract
The rise of Data Science and related fields of Big Data, Machine Learning, and Deep Learning has transformed the industrial landscape. The areas of sports and sports analytics are no exception. While to the layman, its influence may not be evident, but they have changed the way various sports are played up to different degrees. Hence, in recent times, sports institutions and clubs have given increased importance to such research that will ultimately help them have a competitive edge over rivals. The effects of these institutions incorporating these researches into their ways of competing have had impacts on and off the playing field. These effects aren't only in terms of physiological enhancements of the athletes, but also socio-political and economic impacts as well. Out of the various sports implementing these techniques, we will focus on the effects mentioned above of Data Science on Football ("Soccer" in the USA). The following is a detailed review of the concepts as mentioned earlier.

Keywords Data science · Sports · Football · Assessment

1 Introduction

The advent of the new millennium, particularly the late 2000s and early 2010s saw incredible advancements in fields such as Data Science, Machine Learning and Artificial Intelligence, areas which previously faced a decline in interest and investments due to a lack of infrastructure and developments [1–3]. Advancements in computer hardware, increasing use and dependence on Integrated Circuits [4], high-performance computing, cluster computing, parallel computing, and cloud computing [5] took place in the late twentieth century and

✉ Manan Shah
manan.shah@spt.pdpu.ac.in

[1] Department of Information and Technology Engineering, Indus University, Ahmedabad, Gujarat, India

[2] Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, Gujarat, India

Published online: 01 March 2021

early twenty-first century. These advancementsmade it possible to analyse data sets increasingly complex in nature, having the size of hundreds of gigabytes and terabytes [6–8]. As a result, various companies and organisations across a broad spectrum of industries have been adopting machine learning and data science techniques to firmly capitalise this increasingly data-driven society [1, 2, 9]. Retail companies such as Walmart have collaborated with tech companies such as HP and adopted a data-driven approach to provide a better overall shopping experience to their consumers [10]. Although this means a significant investment on the part of these companies, up to hundreds of millions of dollars, [10], the return on investment in terms of productivity and overall growth is noteworthy [11]. According to a study on Data-driven decision making by Economist Erik Brynjolfsson and his colleagues from MIT and Penn's Wharton School [12] the more data-driven a firm is, the more productive and profitable the firm is. While the use of data to make decisions is nothing new with techniques such as Multiple Criteria Linear programming and other data mining techniques being utilised [13–15], the explosion of the internet enhanced and emphasised its importance. The paper devised a measure called DDD(Data-Driven Decision) that rates firms on how strongly they implement data-based techniques in their analytics and decision making. As a result of using this, it was seen that one standard deviation higher on the DDD scale meant a 4–6% increase in productivity. Also, a higher DDD rating meant an increase in factors such as return on assets, market value, return on equity and others [6, 7, 16].

To the uninitiated, the terms Artificial Intelligence, Machine Learning and Data Science may seem quite vague and hackneyed [17–19]. Let us break down what these terms mean and where they come in to play into the larger scheme of things [20]. Due to the vastness of the field of data science, there are many definitions among the academia of what data science is or what the role of a data scientist is. The general description that most seem to agree upon is that Data Science aims to use various tools, algorithms, data systems and processes to gather actionable inputs from generated data [19]. Under the umbrella term of Data Science, methods such as Data Warehousing, Data Mining, Big Data, Mathematics, Statistics, Predictive Analysis for Business Intelligence and Machine Learning [21]. Big Data is a term that was coined to encapsulate methods and processes that were used to deal with data sets having the size of hundreds of terabytes and petabytes [22]. The need for Big Data techniques arose as traditional data management techniques fell short due to the volume of the data, from which not all of it was structured or usable. An IBM report of 2011 revealed that a vast majority of the data generated in the world was generated in the two years leading up to that year (IBM 2011. What is big data?). Big data can be separated from traditional data management by the 5Vs: Volume, Variety, Value, Velocity and Veracity [23–25]. As Big Data encompasses all types of data, i.e., structured, semi-structured, unstructured and data sets of all sizes, it can be used by all enterprises to their advantage to expand their businesses rapidly [1, 2]. The government and policymakers can also utilise it as Big Data provides a systematic approach of dealing with a variety of problems. The transition of companies and enterprises from conventional practices to Big Data practices may be a tedious one as it may include a massive overhaul of personnel and software/hardware

equipment [26, 27]. But it must be viewed as a one-time investment with an enormous return on interest.

A variety of domains and industries and domains use the techniques mentioned above. The predictive analysis which is the outcome of the above-outlined fields of studies such as Data Science and more specifically, Big Data, Machine Learning, and Artificial Intelligence is useful in industries such as Healthcare, Retail, Manufacturing, Cybersecurity, Banking, and numerous other areas. One particular domain that has been revolutionisedis the field of Sports and analytics [28]. The initial breakthrough in implementing decisions based on analysis of data, rather than conventional thinking in sports, was achieved by the Oakland Athletics baseball team in 2002. The team, competing in the Major League Baseball (MLB), under their General Manager, adopted a data-driven, sabermetric and evidence-based playing system. This approach helped them overcome factors such as a lower budget compared to many other teams in the league, to find success in the league [29]. Michael Lewis well documents their journey in his book "Moneyball: The Art of Winning an Unfair Game". Although this was one of the first instances where science and technology were at the heart of decision making, it was certainly not the first where people tried to use science and analysis to gain a tactical advantage in sports. The first notable time science and statistics were used in sports was in 1912 by American sports journalist Hugh Fullerton. Hugh analysed the ball hit success probability by dividing the baseball field into various zones in his essay [30], which was published in the American Magazine. This practice heralded an era of innovation where different sports started utilising in-game analytics, further hand notation systems, formation analysis, tracking devices, zonal analysis, analysis of player movement, fitness levels, overall physiological analysis et cetera. But all these systems were limited by the technology of their time. As a result, various human biases and preconceived notions came into existence depending upon who conducted these studies. An example of this is the famous British football analyst Charles Reep, who was one of the first analysts in English Football. He analysed more than 2500 games by 1968 and published his findings in the Journal of the Royal Statistical Society [31]. Although his results were highly influential in English Football, Reep became a figure of controversy. This was because various subsequent studies found that his findings related to the optimal number of passes leading up to a goal, and the best way to score a goal might not be accurate for all cases [32], signifying either bias or flawed analysis. Hence technological advancements, especially computational advancements, represent an opportunity to prevent these biases and faulty analytics in sports.

While developments in this field have taken place concerning various sports such as Basketball [33], cricket, baseball, this paper aims to evaluate multiple developments in the game of football (soccer) that came to be a result of the interdisciplinary field of data science. We will take a look at how algebra, statistics, data mining, data analysis, concepts of machine learning such as regression, computational power advancements, sophisticated equipment et cetera are used here. We will do this by taking a detailed look at how these concepts have influenced decision making in sports (Sect. 2), how these concepts are utilised in the scope of the sport (Sect. 3), the future challenges that may arise and may represent an opportunity to fine-tune analysis further, and the concluding section.

2 Impacts of Data Science and Analytics in Sports

In terms of a lasting impact, Charles Reep's study is the right candidate for consideration. Charles Reep using the analysis of 2500 football (soccer) games by 1968, came to a few conclusions. The first being that it took only three or fewer passes to score 80 per cent of the goals. He also concluded that 10:1 was the shots to goal conversion ratio. This study made him a proponent of the long ball system, where the ball is kicked over a long distance, to bypass many players, decrease the number of passes and decrease the amount of time. Although later studies raised various doubts over this method, i.e., misinterpretation generalisation of findings [34], Charles Reep'sresearch was at the forefront of football analysis in England, and hugely influenced English football philosophy. So much so, that England was known for its long ball play, something that can be noticed in lower-tiered leagues in Britain to this day [32].

An essential aspect of impacts is the financial aspect. Taking advantage of the resources at disposal to analyse a game from a statistical and sabermetric viewpoint, can help a team with less economic power have a significant edge over its wealthier rivals. The perfect example of this is the Baseball team Oakland Athletics, who in the early 2000s, adopted a sabermetric approach in playing the game. Despite having one of the lowest spending budgets and wage bills, the team was able to churn out positive results, in doing so, defeating some of the wealthiest teams in the sport, such as the New York Yankees. What helped them was taking a statistical and data-driven approach, which led them to sign undervalued but highly effective players and changing their style of play accordingly. This approach saw the team reaching the playoffs of the MLB in 2002, 2003 [29]. The success of this approach meant that many organisations hired full-timesabermetric analysts to implement this approach. Teams such as the New York Mets successfully emulated this approach, leading them to the playoffs in 2015. This model significantly impacted other sports around the globe as well.

In terms of using real-time analytics, a prime example is the German National Football team. In the 2014 Football World Cup, the German squad utilized SAP's Match Insights to improve their performance. By using SAP HANA's database to record the passing and movement of each player, a data set of hundreds of millions of points was produced. The real-time analysis of this data using SAP's platform helped the players and the coaching staff analyse their shortcomings, weaknesses and helped them strengthen these areas [35]. This approach proved to be tremendously successful as the German team went on to win the 2014 World Cup. Another excellent example where data analysis helped sports teams formulate their strategies in real-time is baseball. Due to the introduction of softwares in Baseball that track every factor of performance such as trajectory, velocity, distance travelled, batter's preferred side, favourite waiting position [32], there was an increase in defensive formation changes. This increase was doubled every season since 2011 [36]. Hence, this meant that teams analysed the situation of the gameplay by play, ball by ball.

In addition to analysing tactical factors, sporting factors, physiological factors, one aspect that can drastically improve using data analysis is the positional data

analysis. In the 2016 Olympics held at Rio De Janeiro, Laura Ludwig and Kira Walkenhorst won the gold medal in beach volleyball. They did this by collaborating with Dr Daniel Link of TU Munich, who used positional data analysis from various tracking devices and conventional video analysis to come up offensive tactics for numerous rallies and made changes to their style of play [32]

Data Science can also be used to analyse other factors, such as social and economic factors. According to [37], the positive performances of Liverpool Football Club's player Mohamed Salah was linked with a decrease in prejudice and hate crimes towards the player's religion. The paper undertook a detailed analysis and used placebo groups as well to analyse this effect. In terms of economic impacts, various studies have been done to examine the impact of sports on the economy. One such study, [38], examines the effect of Singapore hosting the Singapore Grand Prix F1 event in terms of tourism and other economic benefits.

3 Linking Data Science and Sports

Constantinou et al. [39] used a Bayesian Network Model called Pi-Football, to forecast the results of Association Football matches. A Bayesian network is a probabilistic graphical model that represents a set of variables by a directed graph, by taking into account their conditional dependencies, hence representing an accurate joint probability distribution. The exciting prospect here is that this paper utilises various essential but overlooked factors to provide a revised subjective forecast. The model takes into account the strength of the two competing teams to generate an initial objective forecast. The actualestimate is then madeafter it takes into account factors such as team form, fatigue and psychological impact. The paper uses the Ranked Probability Score (RPS) to measure the accuracy of the forecasts [40]. The paper also expands upon the current literature that exists in the field of making profitability by placing bets on bookmaking odds [41]. The model presented by the paper may prove particularly useful for the following reasons:

(1) The ability to take into account the uncertainty resulting from the subjective factors taken into consideration.
(2) The ability to appropriately weight past/recent actions of a team.
(3) Ease of using the model for leagues other than the one in the study.
(4) To analyse the performance of the model in terms of accuracy as well as profitability.

Moura et al. [42] Analysed various football game-related statistics from the data obtained from the group stage matches of the 2006 World Cup held in Germany. The paper used multivariate techniques to accurately represent the 14 variables of the study (Shots, Shots on goal, Goal scored, Fouls committed, Yellow Cards et cetera) in a matrix and obtained theirrespective eigenvalues and vectors. Based on the collected data, cluster analysis was performed to divide the teams into two groups using the k-means method. To quantify the degree of separation among various data points

in the cluster, a Silhouette Coefficient that took into account the number of clusters and distance between data points of the clusters was used. A positive value of this coefficient showed correct classification and a negative value represented poor classification. [43], suggest that a cluster having a Silhouette Coefficient value between 0.51 and 0.7 has a proper structure. Here the coefficient values for both the clusters were 0.54 and 0.55. Further, the results showed that more than 70% of the winning teams were correctly classified into the same group, and 67.8% of the drawing and losing teams grouped. Based on this classification, one can get a general idea of what playing tactics are generally associated with a winning team. This finding also seems to be in line with [44], which suggests that successful teams maintain a higher possession rate.

Gama et al. [45] Provides an interesting method to analyse the importance of a player. The paper describes football teams as small-world networks, comparing the players as various nodes in a network, and interactions among players as the weight of nodes. Thirty games in the Portuguese Premier League in the 2010/11 season wereanalysed by the authors to do this, in which 7583 offensive actions took place. The paper introduces various coefficients to quantify inter-player and intra-player interactions. Coefficients such as the scaled connectivity co-efficient and the Clustering coefficient give the measure of interconnectivity and cooperation among teammates. A topological dependency co-efficient that quantifies how often two teammates collaborate during an offensive action in the game provides an idea of the interdependencies and hierarchies in a team. A network density analysis gives the notion of overall teamwork and the general interaction in the team. This paper provides a framework that can be utilised by football analysts to decide the importance of a player to the side and what can be done to improve the overall interaction between the players of the team.

Hirotsu et al. [46], expand upon their previous paper [47], which utilises a Markov model to identify the optimal time to make tactical changes and substitutions.Hence they provide a framework to determine various characteristics of teams in the Premier League in the season 1999/2000, using the data available from Opta Index (2000). AMarkov model that utilised the stochastic transitions in parameters such as offensive actions and defensive actions and other parameters such as home advantage was used to generate a data table. This data was then applied to a generalised linear model, the aim of which was to build a hierarchy of the various factors in the study. Using the values and hierarchy obtained from the model, it was quite clear what factors separated the successful teams from other teams. For example, from the results obtained from the linear model, it was found that Manchester United was the best at retaining possession of the ball. Also, various coefficients for the factors in the study were obtained, providing further insight. For example, Coventry City, had the highest "scorehome" value, a coefficient which showed the propensity of a team to score more goals at home in comparison to away matches. Hence, this paper provides an interesting insight by evaluating the strengths and weaknesses of various teams and ultimately highlighting what successful teams do correctly.

Rotshtein et al. [48] gives a method to make predictions on the outcomes of football matches based on a fuzzy network model generated by identifying a non-linear dependence on a fuzzy knowledge base [49, 50] A genetic algorithm is applied to

this model, after which it is tuned using a neural network. An initial knowledge base is generated using information such as past results of the teams involved, previous encounters between the two sides themselves, the margin of win or loss, i.e., significant victory/loss, small victory/loss, draw. Interdependencies are identified among these factors, and they are represented using a knowledge matrix. Using an approximator, a generalised prediction model involving a nonlinear function is obtained, which encapsulates all the involved factor variables and their values. To optimise this function and the model itself, a genetic algorithm is applied. Further, for tuning the model and learning purposes, a neural network is used. To train the model, data of 1056 matches from 1994 to 2001 was used. To test this model, the results of 350 games from 1991 to 1993 were used. The results obtained after applying genetic algorithm were quite efficient, with extreme case results(huge margin victory/loss) showing an accuracy of over 90%, and non-extreme case results showing an accuracy of over 80%. This result seemed to be an improvement over [51], which used a similar model involving fuzzy logic, genetic algorithm and neural networks.

Tsakonas et al. [52], provides an interesting take regarding the financial and overall contribution of a player to his/her respective football club, by showing a direct correlation of the presence of players in the team with an increase/decrease in points and financial profit/loss. Here a regression model that uses regularised adjusted plus, minus rating to rate the various involved factors such as home advantage, goals scored et cetera is an improvement over regression models as proposed in Winston 2009. This regression model is used to predict match outcomes as the direct involvement of a player based on the given factors. A further Monte Carlo simulation [53], then provides the final points tally of the season of the participating teams. This Monte Carlo simulation of the 2014/15 season was also used to provide statistics related to the economic gain obtained by each side involved. By analysing the data collected from the regression model and Monte Carlo simulation, impressive results were obtained. For example, it was found that the signing of the Spanish midfielder CescFabregas by the Chelsea Football Club(winners that season), caused an increase of 3.4 points in their overall points tally, giving them an advantage over title competitors Manchester City that season. It also meant an increase in revenue by GBP 2.1 million. This research also furthers the idea that each player has his/her marquee value in terms of adding to the squad, the ability to attract a crowd and increase revenue [54]. Hence, this research may prove valuable for top teams, for whom 1–2 point may be the difference in winning the title and 2nd place. It may also be useful for teams fighting relegation, for whom 1–2 points may mean the difference between maintaining their spot in one of the best leagues in the world or playing in a lower division.

Fairchild et al. [55] is another paper that utilised a regression model. Here, the authors use regression to obtain the probability of a shot taken, resulting in a goal. Data from 99 MLS games, having 1115 shots, from the 2016 season formed the basis of this study. The model depends on factors such as the location of the ball (x,y coordinates), distance from the goal, shot angle, type of shot, type of assist. This process is similar to [56], where authors similarly tried to estimate the probability of a shot taken ending up as a goal. In the current paper, the authors also try to obtain the value of expected goals based on the probability of the shot ending up in goal, using the Poisson binomial

distribution. Using the obtained results, spatial analysis of the points from where the teams took their shots was done. It was found that certain spots from where shots were taken had a higher probability of ending up as a goal. Also, using the concept of fractal dimensions in analysing spatial data points, it was found that teams whose shot charts show a smaller fractal dimensionality, were more offensive in comparison to others. In a sport where about 70% of the games have one of the two teams scoring less than two goals, and the average value of goals scored per match is 3 [57], this can be used to evaluate movement among players. It can also help teams formulate tactics to score more goals and assess the efficiency of teams in a general sense.

Rein et al. [58] provide a Big Data overview of football. In the paper, the authors talk about analytics in sports and how current advancements in Big Data have enhanced our understanding of the game. The paper talks about how fundamental tactics are to football. Tactics are an amalgamation of various real-time interdependent factors such as player fitness levels, opposition strength, home advantage, the referee, recent form, weather and so forth. Thus, tactics are an essential part of the preparation for any game [59]. Hence, tactics being so complicated, require a proper data pipeline for storage, processing and analysis of the data, so that relevant and impactful insights may be obtained from the data [60]. This is where Big Data comes in. Big data technologies provide organisations with a framework that helps in storing data such as the psychological data, physiological data, tracking data, scout data et cetera. By using appropriate analytics technology, i.e., Machine Learning, Deep Learning, actionable intelligence can be gathered, which can then be imparted using proper visualisation and reporting techniques.

Bakker et al. [61] A unique perspective in evaluating the impact a player has on the performance of a team. The paper evaluates how cultural diversity in the roster of a team affects the performance of a team. To generate a chart of diversity in a group, an Automated Similar Judgement Program (ASJP) [61] is used. In this chart, players of different nationalities are placed further compared to players of similar regions. Also, the line connecting two players varies according to the similarity of their language. The more similar a language is, the thicker the line connecting two players. Generally, it was found that teams with higher diversity performed better in comparison to organisations with lower diversity. One factor that may affect this conclusion is the disparity in the purchasing power of different teams. The authors take that into account by factoring various control variables, the economic environment of the country, the strength of the league, ancestry of players, and many more in their research. After factoring in these, the conclusion remained the same. Teams with a higher heterogeneity in their roster fared better in comparison to teams with lesser diversity in their squad. Hence this might be an indicator for future recruiting strategies for football clubs.

4 Future Challenges and Scope of Work

Incredible work has been done to bridge the gap between the knowledge obtained from data-driven processes and how it is translated into action on the field. Despite this, there are still some grey areas where further research in these areas can help

narrow the gap even further. One area where work needs to be done is the handling of data. Tactics in football are one such area where until recently; very few studies were done due to a lack of relevant data. Although this scenario is changing due to attention being drawn to the importance of tactics in football [59, 62, 63], the problem now is the sheer amount of data generated. This problem can be attributed to breakthroughs in technologies such as player tracking devices [32, 64], various sources of data such as the online sports websites, sources that draw data from observational mediums such as the television and many more. Another problem is the absence of a framework/model which is capable ofintegrating all the factors in sport, such as the physiological data of the team, their technical competence, tactics, and skills. As these factors are hugely interdependent and show vast degrees of correlation [65, 66], it becomes challenging to come up with a theoretical model which is foolproof. Although advances in AI, ML and Deep Learning may change this scenario [67, 68], progress can still be made in this field. Another critical factor slowing down growth in this sector is the fact that applied data science in the field of sports is not collaborative, unlike industries such as healthcare, and life sciences. Studies in this sector take place as an initiative undertaken by private organisations and teams, to gain a competitive advantage over their rivals. Any breakthrough that these organisations may achieve will not be made public as rivals may analyse their approach, imitate it and may cause a loss of competitive advantage. And hence, a framework that allows transparent and collaborative research environment in other industries [69, 70] is absent here. Hence, such measures for collaboration need to be taken. Also, stakeholders in these organisations, right from the scouts, physios to the topmost level such as club owners, need to be made aware of the power of these techniques and trained accordingly. Another aspect which is often discussed by various sports coaches and veterans of sports is the ability to control your emotions under extreme conditions and not letting it affect your decision making in real-time. This aspect is often neglected in sports in favour of the development of physical and technical qualities of athletes. But in recent times, work has been carried out in Machine Learning that may help teams accurately measure and quantify this quality, even though it is considered to be abstract. For instance, [71] provide a detailed review of how Machine Learning techniques have been used to measure emotional intelligence. The paper also provides a framework that will help quantify a person's emotional intelligence with a relatively better degree of success in comparison to conventional techniques. This kind of structure may come in handy for sports teams, where athletes have to make split time decisions under tremendous pressure and constant scrutiny. Hence, integrating such a model into their functioning may be an excellent asset for the team. Another area where improvements can be made in the field of sports analytics is the presence of real-time technical analysts. It has been seen that organisations hire coaches that may help improve the team's level of competition in one aspect of the sport. But hiring technicians who can analyse the game in terms of numbers and give real-time feedback on the field that may help the manager is still not largelyprevalent in football. This is in comparison to other sports such as Formula One, where real-time feedback is provided to the driver and the crew. This may also be due to a lack of an academic framework that offers individuals with the proper balance of knowledge of the sport and the technical

tools required to analyse the game. Hence, work needs to be done to bridge the gap between academia and requirements for such degrees and individuals. While outlining various areas where improvements can be made to bridge the gap between sports and Data Science, it is also important to note that Data Science and Big Data is a continually evolving area of study. Improvements made in these fields will directly correlate to developments in any associated field. In their paper, [72], the authors provide us with a general sense of where Big Data faces challenges, by highlighting six key areas where research can be done.

5 Conclusion

With the advent of the age of data, organisations today are radically changing their approach of functioning. These organisations were swift to catch on these technological changes to maximise their profits and to gain a competitive edge over their rivals. So was the case with sports. Using smart, data-drivendecisions that came about as a result of incorporating the concepts of Data Science, Big Data, and Machine Learning into their choices, the sports teams are reaping their benefits. Athletes are getting stronger, faster, better than ever, and it is in large part due to these practices adopted by them. These practices, in which every move is scrutinised, generate large amounts of data, are quite complicated. As seen in the paper, these complex concepts require an understanding of the mathematical and computational ideas, something that the average sports fan may not possess. And while teams have tried their best to implement changes based on these concepts, a majority of the work done is done in studies undertaken by professionals in the field of computer science. Hence, the future of sports analytics lies in adopting a multi-disciplinary approach, where teams can encapsulate all the cornerstones of sports and science under one common roof, paving an exciting, foolproof and progressive path for the future of sports. And although our study has shown that significant and promising research has been done, there is a definite room for improvement. This has been highlighted in the section that shows future challenges and the scope of the study. Ultimately, we hope that future research considers these areas and these sports can successfully improve, finally giving fans what they want-excitement.

Acknowledgements The authors are grateful to Indus University and School of Technology, Pandit Deendayal Petroleum University for permission to publish this research.

Authors Contribution All the authors make substantial contributions to this manuscript. PT and MS participated in drafting the manuscript. PT wrote the main manuscript, all the authors discussed the results and implication on the manuscript at all stages.

Funding None.

Availability of Data and Material All relevant data and material are presented in the main paper.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical Approval Not applicable.

Consent for Publication Not applicable.

References

1. Parekh V, Shah D, Shah M (2020) Fatigue detection using artificial intelligence framework. Augment Hum Res. https://doi.org/10.1007/s41133-019-0023-4
2. Pandya R, Nadiadwala S, Shah R, Shah M (2020) Buildout of methodology for meticulous diagnosis of K-complex in EEG for aiding the detection of alzheimer's by artificial intelligence. Augment Hum Res. https://doi.org/10.1007/s41133-019-0021-6
3. Kundalia K, Patel Y, Shah M (2020) Multi-label movie genre detection from a movie poster using knowledge transfer learning. Augment Hum Res. https://doi.org/10.1007/s41133-019-0029-y
4. Bondyopadhyay PK (1998) Moore's law governs the silicon revolution. Proc IEEE 86:78–81. https://doi.org/10.1109/5.658761
5. Arnold U, Oberlander J, Schwarzbach B (2013) Advancements in cloud computing for logistics. Fed Conf Comput Sci Inf Syst FedCSIS 2013:1055–1062
6. Gandhi M, Kamdar J, Shah M (2020) Preprocessing of non-symmetrical images for edge detection. Augment Hum Res 5:1–10. https://doi.org/10.1007/s41133-019-0030-5
7. Patel D, Shah D, Shah M (2020) The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. Ann Data Sci. https://doi.org/10.1007/s40745-019-00239-y
8. Ahir K, Govani K, Gajera R, Shah M (2020) Application on virtual reality for enhanced education learning, military training and sports. Augment Hum Res. https://doi.org/10.1007/s41133-019-0025-2
9. Jani K, Chaudhuri M, Patel H, Shah M (2020) Machine learning in films: an approach towards automation in film censoring. J Data, Inf Manag 2:55–64. https://doi.org/10.1007/s42488-019-00016-9
10. Bryant R, Katz R, Lazowska E (2008) Big-data computing: creating revolutionary breakthroughs in commerce, science, and society in computing research initiatives for the 21st century. Comput Res Assoc
11. Tambe P (2014) Big Data Investment, Skills, and Firm Value. Manage Sci 60:1452–1469. https://doi.org/10.1287/mnsc.2014.1899
12. Mcafee A, Brynjolfsson E (2012) Spotlight on big data big data: the management revolution. Harv Bus Rev 90:1–9
13. Li J, Shi Y (2001) An integer linear programming problem with multi-criteria and multi-constraint levels: a branch-and-partition algorithm. Int Trans Oper Res 8:497–509. https://doi.org/10.1111/1475-3995.00328
14. Shi Y, Tian Y, Kou G et al (2011) Optimization based data mining: theory and applications. Springer, London
15. Olsen D, Shi Y (2006) Introduction to business data mining. McGraw-Hill/Irwin, New York
16. Sukhadia A, Upadhyay K, Gundeti M et al (2020) Optimization of smart traffic governance system using artificial intelligence. Augment Hum Res. https://doi.org/10.1007/s41133-020-00035-x
17. Jha K, Doshi A, Patel P, Shah M (2019) A comprehensive review on automation in agriculture using artificial intelligence. Artif Intell Agric 2:1–12. https://doi.org/10.1016/j.aiia.2019.05.004
18. Kakkad V, Patel M, Shah M (2019) Biometric authentication and image encryption for image security in cloud framework. Multiscale Multidiscip Model Exp Des 2:233–248. https://doi.org/10.1007/s41939-019-00049-y
19. Panchiwala S, Shah M (2020) A comprehensive study on critical security issues and challenges of the IoT world. J Data, Inf Manag 2:257–278. https://doi.org/10.1007/s42488-020-00030-2
20. Gupta A, Dengre V, Kheruwala HA, Shah M (2020) Comprehensive review of text-mining applications in finance. Financ Innov 6:1–25
21. Desai M, Shah M (2020) An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN). Clin eHealth. https://doi.org/10.1016/j.ceh.2020.11.002
22. Thakkar H, Shah V, Yagnik H, Shah M (2020) Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. Clin eHealth. https://doi.org/10.1016/j.ceh.2020.11.001

23. Ayankoya K, Calitz A, Greyling J (2014) Intrinsic relations between data science, big data, business analytics and datafication. ACM Int Conf Proceeding Ser 28-Septemb:192–198. https://doi.org/10.1145/2664591.2664619
24. Talaviya T, Shah D, Patel N et al (2020) Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artif Intell Agric 4:58–73. https://doi.org/10.1016/j.aiia.2020.04.002
25. Shah K, Patel H, Sanghvi D, Shah M (2020) A Comparative analysis of logistic regression, random forest and KNN models for the text classification. Augment Hum Res. https://doi.org/10.1007/s41133-020-00032-0
26. Naik B, Mehta A, Shah M (2020) Denouements of machine learning and multimodal diagnostic classification of Alzheimer's disease. Vis Comput Ind Biomed Art 3:1–18. https://doi.org/10.1186/s42492-020-00062-w
27. Shah D, Dixit R, Shah A et al (2020) A comprehensive analysis regarding several breakthroughs based on computer intelligence targeting various syndromes. Augment Hum Res. https://doi.org/10.1007/s41133-020-00033-z
28. Drust B, Green M (2013) Science and football: evaluating the influence of science on performance. J Sports Sci 31:1377–1382. https://doi.org/10.1080/02640414.2013.828544
29. Lewis Michael (2004) Moneyball: The Art of winning an unfair game - Michael Lewis - Google Books
30. Fullerton HS (1912) The inside game: the science of baseball. Am Mag 70:2–13
31. Reep C, Benajmin B (1968) Skill and chance in association football. J Royal Stat Soc. Ser A (General) 131(4):581–585
32. Memmert D, Rein R (2018) Match analysis, big data and tactics: current trends in elite soccer. Dtsch Z Sportmed 69:65–72. https://doi.org/10.5960/dzsm.2018.322
33. Thabtah F, Zhang L, Abdelhamid N (2019) NBA game result prediction using feature analysis and machine learning. Ann Data Sci 6:103–116. https://doi.org/10.1007/s40745-018-00189-x
34. Hughes M, Franks I (2005) Analysis of passing sequences, shots and goals in soccer. J Sports Sci 23:509–514. https://doi.org/10.1080/02640410410001716779
35. Bojanova I (2014) IT enhances football at world cup 2014. IT Prof 16:12–17. https://doi.org/10.1109/MITP.2014.54
36. ZACH HELFAND (2015) Use of defensive shifts in baseball is spreading — because it works - Los Angeles Times. https://www.latimes.com/sports/la-sp-baseball-defensive-shifts-20150719-story.html. Accessed 3 Jan 2021
37. Alrababa'h A, Marble W, Mousa S, Siegel AA (2019) Can exposure to celebrities reduce prejudice? The effect of Mohamed Salah on islamophobic behaviors and attitudes. https://doi.org/10.31235/osf.io/eq8ca
38. Henderson JC, Foo K, Lim H, Yip S (2010) Sports events and tourism: the Singapore formula one grand prix. Int J Event Festiv Manag 1:60–73. https://doi.org/10.1108/17852951011029306
39. Constantinou AC, Fenton NE, Neil M (2012) Pi-football: a bayesian network model for forecasting association football match outcomes. Knowledge-Based Syst 36:322–339. https://doi.org/10.1016/j.knosys.2012.07.008
40. Epstein ES (1969) A scoring system for probability forecasts of ranked categories on JSTOR. J Appl Meteorol 8:985–987
41. Dixon MJ, Coles SG (1997) Modelling association football scores and inefficiencies in the football betting market. J R Stat Soc Ser C Appl Stat 46:265–280. https://doi.org/10.1111/1467-9876.00065
42. Moura FA, Martins LEB, Cunha SA (2014) Analysis of football game-related statistics using multivariate techniques. J Sports Sci 32:1881–1887. https://doi.org/10.1080/02640414.2013.853130
43. Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, 99th edn. Wiley, Hoboken
44. Jones PD, James N, Mellalieu SD (2004) Possession as a performance indicator in soccer. Int J Perform Anal Sport 4:98–102. https://doi.org/10.1080/24748668.2004.11868295
45. Gama J, Passos P, Davids K et al (2014) Network analysis and intra-team activity in attacking phases of professional football. Int J Perform Anal Sport 14:692–708. https://doi.org/10.1080/24748668.2014.11868752
46. Hirotsu N, Wright M (2003) Determining the best strategy for changing the configuration of a football team. J Oper Res Soc 54:878–887. https://doi.org/10.1057/palgrave.jors.2601591

47. Hirotsu N, Wright M (2002) Using a markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. J Oper Res Soc 53:88–96. https://doi.org/10.1057/palgrave/jors/2601254
48. Rotshtein AP, Posner M, Rakityanskaya AB (2005) Football predictions based on a fuzzy model with genetic and neural tuning. Cybern Syst Anal 41:619–630. https://doi.org/10.1007/s10559-005-0098-4
49. RotshteinKatel'Nikov APDI (1998) Identification of nonlinear objects by fuzzy knowledge bases. Cybern Syst Anal 34:676–683. https://doi.org/10.1007/BF02667040
50. Rotshtein AP, Shtovba SD (2001) Fuzzy multicriteria analysis of variants with the use of paired comparisons. J Comput Syst Sci Int 40:499–503
51. Tsakonas A, Dounias G, Shtovba S, Vivdyuk V (2002) Soft computing-based result prediction of football games. Ist Int Conf Inductive Model
52. Sæbø OD, Hvattum LM (2018) Modelling the financial contribution of soccer players to their clubs. J Sport Anal 5:23–34. https://doi.org/10.3233/jsa-170235
53. Hvattum LM (2013) Analyzing information efficiency in the betting market for association football league winners. J Predict Mark 7:55–70. https://doi.org/10.5750/jpm.v7i2.614
54. Gennaro Vince (2007) Diamond dollars: The economics of winning in baseball. In: Potomac Books Inc. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=18.%09Gennaro+2007.+Diamond+Dollars%3A+The+Economics+of+Winning.+Maple+Street+Press.+1-253&btnG=#d=gs_cit&u=%2Fscholar%3Fq%3Dinfo%3AvoGYPaWVTGQJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den. Accessed 3 Jan 2021
55. Fairchild A, Pelechrinis K, Kokkodis M (2018) Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. J Sport Anal 4:165–174. https://doi.org/10.3233/jsa-170207
56. Pollard R, Ensum J, Taylor S (2004) Estimating the probability of a shot resulting in a goal: the effects of distance, angle and space. Int J Soccer Sci 2:50–55
57. Anderson Chris (2010) Comparing the best soccer leagues in the world. In: Sport. Inc. 3.1(Fall). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=1.%09Anderson%2C+C.%2C+2010%2C+Comparing+the+best+soccer+leagues+in+the+world.+Sports%2C+Inc.+3.1%28Fall%29%2C+10-12&btnG=. Accessed 3 Jan 2021
58. Rein R, Memmert D (2016) Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. Springerplus. https://doi.org/10.1186/s40064-016-3108-2
59. Yiannakos A, Armatas V (2006) Evaluation of the goal scoring patterns in European Championship in Portugal 2004. Int J Perform Anal Sport 6:178–188. https://doi.org/10.1080/24748668.2006.11868366
60. Coutts AJ (2014) Evolution of football match analysis research. J Sports Sci 32:1829–1830. https://doi.org/10.1080/02640414.2014.985450
61. Bakker D, Müller A, Velupillai V et al (2009) Adding typology to lexicostatistics: a combined approach to language classification. Linguist Typol 13:169–181. https://doi.org/10.1515/LITY.2009.009
62. González-Víllora S, Serra-Olivares J, Pastor-Vicedo JC, da Costa IT (2015) Review of the tactical evaluation tools for youth players, assessing the tactics in team sports: football. Springerplus 4:1–17. https://doi.org/10.1186/s40064-015-1462-0
63. LI Ping (2005) Tendency of Offensive Tactics of Modern Football from the 11~(th) and 12~(th) European Football Championship--《Journal of Chengdu Physical Education Institute》2005年05期. J Chengdu Phys Educ Inst
64. Lu W-L, Ting J-A, Little JJ, Murphy KP (2013) Learning to track and identify players from broadcast sports videos. IEEE Trans Pattern Anal Mach Intell 35:1704–1716
65. Júlio G (2009) Trends of tactical performance analysis in team sports: bridging the gap between research, training and competition. Rev Port Ciências do Desporto 9:81–89
66. Carling C, Bloomfield J, Nelsen L, Reilly T (2008) The role of motion analysis in elite soccer work rate data. Sport Med 38:839–862
67. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
68. Dutt-Mazumder A, Button C, Robins A, Bartlett R (2011) Neural network modelling and dynamical system theory: are they relevant to study the governing dynamics of association football players? Sport Med 41:1003–1017. https://doi.org/10.2165/11593950-000000000-00000

69. Goecks J, Nekrutenko A, Taylor J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. https://doi.org/10.1186/gb-2010-11-8-r86
70. Blankenberg D, Von Kuster G, Bouvier E et al (2014) Dissemination of scientific software with galaxy toolshed. Genome Biol 15:2–4. https://doi.org/10.1186/gb4161
71. Sharma M, Khera SN, Sharma PB (2019) Applicability of machine learning in the measurement of emotional intelligence. Ann Data Sci 6:179–187. https://doi.org/10.1007/s40745-018-00185-1
72. Xu Z, Shi Y (2015) Exploring big data analysis: fundamental scientific problems. Ann Data Sci 2:363–372. https://doi.org/10.1007/s40745-015-0063-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.