Neurocomputing 483 (2022) 59–71

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Explainable AI techniques with application to NBA gameplay prediction

Yuanchen Wang [a], Weibo Liu [b,*], Xiaohui Liu [b]

[a] Birkbeck School of Business Economics and Informatics, Birkbeck College, University of London, London WC1E 7HX, United Kingdom
[b] Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, United Kingdom

A R T I C L E I N F O

Article history:
Received 15 November 2021
Revised 16 January 2022
Accepted 30 January 2022
Available online 4 February 2022

Keywords:
Data science
Explainable artificial intelligence
NBA
Clustering
Regression

A B S T R A C T

In this paper, an explainable artificial intelligence (AI) technique is employed to analyze the match style and gameplay of the national basketball association (NBA). A descriptive analysis on the evolution of the NBA gameplay is conducted by using clustering and principal component analysis. Supervised-learning based AI models (including the random forest and the feed-forward neural network) are applied to produce accurate predictions on NBA outcomes at a season-by-season and a month-by-month basis. To evaluate the interpretability of the established AI models, an explainable AI algorithm is utilized to deduce and assess the precise reasoning behind the model prediction based on the local interpretable model-agnostic explanation method. To illustrate its application potential, the method is applied to the open-source NBA data from 1980 to 2019. Experimental results demonstrate the effectiveness of the introduced explainable AI algorithm on predicting NBA outcomes with interpretation.

1. Introduction

Sports science is a discipline that investigates how human bodies work in exercise, and how sports activities improve players' health and performance from cellular to whole-body perspectives [3,5]. The past few years have witnessed recurring research interests in the prediction and optimization of sports performance. Due to its remarkable capabilities, artificial intelligence (AI) has shown exciting application prospects in many branches of sports science, e.g., injury risk assessment [8,18,28,43] and tactical decision making [2,41].

For AI-based injury risk assessment, one of the first attempts has been made in [43] towards the diagnosis problem of sports injuries, where the decision tree and the Bayesian classification methods have been used to extract the available knowledge for diagnosing injuries. In [18], the decision tree classifier has been leveraged to build a robust predictive model to identify athletes at the high/low risk of muscle injuries based on the pre-season screening data. In [28], a variety of supervised learning algorithms (including the naive Bayes, logistic regression, random forest, and support vector machine) have been utilized to predict the incidence of the hamstring strain injury in elite Australian footballers.

Using AI techniques for decision making has also been a research hotspot [1,9–11,13,31]. For example, convolutional neural networks (NNs) and recurrent NNs have been deployed in [41] to model sequences of player actions in rugby union, where the established prediction model could provide tactical decision supports to coaches. The advantage of such a model is that it incorporates the data of field locations to improve the modelling accuracy so as to provide better tactical decisions. It should be pointed out that the performance of the prediction model is greatly limited by the availability of the data. For instance, the position data of the target player is generally utilized for the sequence modelling, whereas the absent data (e.g. position data of other players), if present, would have also helped improve the modelling accuracy.

Due to its worldwide prevalence and enormous financial value, the national basketball association (NBA) has spared no effort in improving training methods, boosting sports performances, and predicting sports outcomes. This gives rise to widespread NBA applications of AI algorithms which include, but are not limited to, the NNs, naive Bayes, decision tree, support vector machine and random forest [4,7,12,16,17,21,35,38,46]. For example, NNs have been used in [12] to analyze the sports data of 890 basketball games and predict the outcomes of basketball games. Among a list of factors (e.g. two-point shots, three-point shots and defensive rebounds) that affect sports performance, it has been found that defensive rebounds and two-point shots are two important factors for winning games. In [17], the statistics coming from 620 NBA games has been collected and adopted to train four types of NNs (i.e. the feed-forward, probabilistic, radial basis and generalized regression NNs), whose prediction results on winning teams have proven to be more accurate than those made by basketball experts.

* Corresponding author.
E-mail addresses: Yuanchen.Wang@googlemail.com (Y. Wang), Weibo.Liu2@brunel.ac.uk (W. Liu), Xiaohui.Liu@brunel.ac.uk (X. Liu).

https://doi.org/10.1016/j.neucom.2022.01.098
0925-2312/

In [21], the naive Bayes classifier has been employed to predict winners of the NBA matches.

By resorting to the NN, naive Bayes and decision tree, a set of prediction models have been built to forecast outcomes of NBA games based on the historic data [35]. Through the quantitative comparison among various prediction models using different features, the influential features (such as defensive rebounds, three-point percentage and free throws made) have been explicitly identified to have a great impact on the prediction performance. Based on NBA teams' performance indicators during the regular season, the logistic regression approach has been employed in [7] to predict the team's probability of participating in the playoff, where the team's points (per game) and its opponent's field goal percentage have been found to be the most two important factors in accurately predicting the participation probability. In [16], a two-step prediction approach has been presented to forecast the winning probabilities of NBA teams. Given the average inefficiency ratio, the predictive result on the winning probability is generated by integrating all players' performances and referring to the correlation between the sports performance and the winning probability.

Notice that much of the existing work on sports science revolves around predictions at a micro level such as an individual person or a specific play by using AI techniques. In this case, a seemingly natural idea is to conduct AI predictions at a macro level (i.e. predicting game or even season outcomes), which would bring greater applicability and entertainment value to sports audiences. Nevertheless, a key absence in these literatures is *explainable AI*, which refers to the interpretability, analysis and validation of individual model predictions in the context of the NBA [6,23–25,30]. Though defensive rebounds and two-point shots have been concluded as the two most important factors in [7,12,35], this only gives the reader a generalized viewpoint on what features could impact a team's ability to win as opposed to a tailored viewpoint to pinpoint what the exact key features are behind each individual prediction. For example, the top ranked feature in global feature importance may not necessarily be the top ranked feature for an individual data-point. In this context, it seems natural to develop an explainable AI algorithm for NBA gameplay prediction and the corresponding interpretive analysis.

The motivation of this paper is based upon the claim that open-source material incorporating both (a) predicting NBA outcomes (through open-source data) using NNs and (b) accurately interpreting NN predictions using explainable AI techniques is rare. Consequently, there is a need to provide a descriptive analysis to identify, understand and depict the evolution of the NBA gameplay from the 1980s to the current modern era from a statistical standpoint.

Motivated by above discussions, the purpose of this paper is to launch a comprehensive investigation of existing open-source materials relating to the hybrid of sports/NBA analytics and explainable AI techniques in sports science. The main contributions can be summarised into the following three aspects: 1) an explainable AI algorithm is employed to accurately predict macro-level (seasonal/game level) NBA outcomes and interpret individual predictions for the first time; 2) a descriptive analysis of the NBA gameplay is conducted, which provides an overview of the NBA gameplay from the 1980s to the current modern era in a statistical standpoint; and 3) the developed method is applied to real-world NBA datasets with satisfactory experimental results.

The rest of this paper is organized in the following manner. In Section 2, the utilized methods for the descriptive analysis of NBA gameplay and NBA match outcome prediction are introduced. Data preparation and data pre-processing are presented in Section 3. Experimental results and discussions are illustrated in Section 4. In Section 5, conclusions are drawn, and future directions are pointed out.

2. Methods

The main purpose of this paper can be summarized into two aspects: 1) a descriptive analysis of the NBA gameplay from the 1980s to the current modern era; and 2) the success strategy of a given team in the modern NBA. To be specific, the success of a given NBA team is defined by 1) a high win ratio per month; and 2) the given NBA team making the post-season playoffs. In this section, the utilized data mining and explainable AI methods for the aforementioned tasks are discussed separately.

2.1. Descriptive analysis of the NBA gameplay

In this paper, a descriptive analysis is conducted to identify, understand and depict the evolution of the NBA gameplay from 1980 to 2019 from a statistical standpoint. To be specific, trend analysis, clustering analysis and visualization are adopted. Details of the utilized methods are summarized as follows:

1) Trend Analysis:
   a. Produce line graphs depicting year on the × axis and attribute mean on the y axis, e.g. investigate mean 3 points made per season in NBA matches and how it changes over the forty-year period. Given the dataset is standardized, multiple attribute trends are plotted on the same graph for reliable comparison purposes.
2) Clustering
   a. Deploy the K-means and DBSCAN clustering algorithms so as to group the data points into distinct clusters where each cluster represents a period that epitomizes a distinct NBA game style (e.g. late 1990s style basketball at the zenith of Michael Jordan's careers vs late 2010s 3-point based style dominated by the Golden State Warriors).
   b. Use the silhouette score to evaluate the quality of the clustering results. The silhouette score is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample, where the silhouette coefficient of a sample is (b - a)/max(a, b).
3) Principal component analysis (PCA) and Visualization
   a. Depict the clustering results and trend analysis at a macro level to visualize how the NBA gameplay has changed.
   b. Apply PCA to reduce the dimensionality of the dataset and use the "proportion of variance explained by × principal components" metric to assess the reliability of a 2D/3D representation of the dataset.
   c. Visualize the obtained dataset (processed by using PCA) and use different colors to identify the cluster label. Highlight the year of the data-point from trend analysis for each data-point in the 2D/3D diagram with different colors.
   d. Plot arrows to project each original attribute as a labelled vector in the new 2D/3D principal component axes.

2.2. Predicting NBA outcomes with interpretability analysis

As stated previously, the success of a given NBA team is defined by two ways: 1) a high win ratio per month for a given team (regression problem); and b) an NBA team making the post-season playoffs (classification problem). For predicting the NBA outcomes, random forest and feedforward NN algorithms are employed to build the predictive model, while the Local Interpretable Model-agnostic Explanations (LIMEs), proposed in [26] is an explainable AI method, is adopted in this paper to interpret the predictions.

2.2.1. Win ratio prediction

After conducting data preparation and data pre-processing (which will be described in the next section), the following two steps are undertaken to carry out the predictions as well as the interpretability element of the predictions:

1) Train a random forest regressor and a feed-forward NN to predict the win ratio column based on all other features in the dataset.
   a. Implement the following procedures for assessing each model's predictive power:
      i. Set the 90/10 ratio for training/testing data.
      ii. Leave 10 out Cross Validation (Leave One Out Cross Validation unrealistic here due to computational constraints).

   b. Tune the hyperparameters of the random forest regressor and the feed-forward NN.
   c. Use R^2 (proportion of variance of win ratio column explained by other independent variables) and mean squared error (MSE) as the performance assessment metric.
   d. All predictor attributes are standardized to train the NN model but not the random forest since the splitting of values in decision tree-based models are not prone to scales.
2) Employ the LIME algorithm for interpreting the predictions obtained by the NN. In the simulation, the top 10 attributes are displayed by LIME.

NBA team post-season playoffs classification

For the classification problem, the aforementioned AI methods (random forest and feedforward NN) as well as the LIME method are employed, where the only major difference is the implementation of a classification model as opposed to a regression model (e.g. random forest classifier used instead of random forest regressor, F1 used as a testing metric instead of MSE, etc.). F1 is the harmonic mean of the precision and recall, which is used to measure the accuracy of a test.

3. Experiment

In this paper, the NBA gameplay data are extracted from online open source websites. With the data source obtained, data scraping is taken place. Data exploration and data wrangling are then carried out to justify the reliability and validity of the subsequent modelling. The employed data processing method can be concluded into the following 3 steps:

1) Data scraped from the chosen open source site (http://basketball.realgm.com/).
2) Data joining, wrangling, and aggregation on the various datasets obtained.
3) Remediation for any data quality defects in the final datasets.

Each step of this process, along with nuances (e.g. data quality issues) and how they are either remediated or taken into consideration when critically evaluating results, are fully documented in the following two subsections.

3.1. Descriptive analysis of the NBA gameplay

For data analysis of NBA gameplay, the subsequent tasks include 1) producing unsupervised learning based data exploration (Clustering and PCA); and 2) visualising (utilising PCA reduced 2D/3D datasets) the descriptive analysis results.

3.1.1. Data preparation

To gather the data with web scraping, the NBA gameplay data at a seasonal level for each team from the 1980 season to the latest 2019 season is collected. In the experiment, each data point is one NBA team's average statistics across multiple attributes for one season. The data for the regular season gameplays contains a table which outlines the season averages for key gameplay attributes for each team at each year.

3.1.2. Data pre-processing

It should be noted that the raw data includes not only the season-wide averages for team totals but also that of the opponent totals. As there might be correlation between team totals and opponent totals (i.e. if the statistics for team totals are good, one would expect the statistics for opponent totals is poor), the dataset is filtered to opponent totals per team only, for each NBA season.

The raw dataset includes 1055 data points, where the attributes "Year", "Games played" and "Team name" are removed. Games played each year will generally be static unless due to the occasional NBA lockout. "Year" is a variable which cannot be used to explore the relationship between year and gameplay attributes. Team name is removed from the data as it adds no informative statistical value. There are no missing values, and all remaining attributes are of type numerical, eliminating the need for dummy variable creation.

For the descriptive analysis of the NBA gameplay, clustering analysis, trend analysis using line graphs, and PCA are employed. The data standardisation process (i.e. removing the mean and scaling to unit variance) is applied to ensure reliability and interpretability of the utilized methodologies. Particularly, data standardization ensures the unbiasedness as all the attributes are compared and analyzed on the same metric scale.

3.2. Predicting NBA outcomes with interpretability analysis

To predict and understand the strategies a given team should undertake to achieve *success* in the modern NBA, the subsequent procedures are carried out, which includes model training, model prediction on unseen data, and model interpretation.

3.2.1. Win ratio prediction

3.2.1.1. Data preparation. To predict the win ratio of a team for a given month, the gameplay attributes at a monthly level is required. To obtain the win information, the date and result for each game played by each team is scraped and processed. As the data format of the scraped datasets is inconsistent, a simple format clean-up approach using substring and split methods is utilized. It should be noted that the gameplay results for October and April are removed due to tacit strategy purposes. After calculating the win ratios for each team per month, the gameplay attributes are retrieved to predict the win ratios.

3.2.1.2. Data pre-processing. Data pre-processing is employed to predict the win ratio of a team for a given month. In this experiment, feature selection, data profiling, data imputation and data cleaning are utilized to pre-process the raw data.

1. Feature selection

The attributes of the raw dataset are included in the website (https://basketball.realgm.com/info/glossary). In the online data source, there is no team-wide averages per month data. To manually obtain the team-wide averages, the player averages per month information is employed for data analysis. The "PER 36" of the NBA statistic is used to holistically evaluate a player's performance as opposed to purely their raw numbers. "PER 36" essentially extrapolates the raw numbers based on the assumption that the player plays for 36 min in a game, which could contribute to the evaluation of a player's impact and retrieve a team-wide average each month.

Upon initial exploratory model training and testing, the fields "PER" and "eDIFF" are removed from the dataset as predictors for the following reasons, deduced by basketball domain knowledge:

- PER is a player efficiency rating created by John Hillinger, which incorporates all a player's contributions in a game down to one rating. PER is a useful number for a player/team to acknowledge how well their all-round contribution is in a game. However, it is not easily interpretable since it cannot simply be decomposed into its individual components for an analyst to analyze. Secondly, given PER encompasses multiple if not all aspects of gameplay into one score, having team PER average as an attribute in the dataset to predict win ratio would introduce a considerable self-fulfillment or bias in predictions since a high team PER average would certainly have an extreme positive impact on a team's win ratio.
- eDIFF is the difference between the offensive and defensive ratings of a player. Since offensive and defensive ratings are already in the dataset, there is less need to have eDIFF since the existence of an obvious correlation. Furthermore, the utilization of eDIFF could suppress the predictive powers of offensive and defensive ratings due to their correlation.

The data source contains "advanced statistics" as opposed to the standard set of statistics such as Field Goals Made, Three Points Made etc. These statistics are mostly game minutes agnostic as they include mainly percentages and ratings. The two processed datasets are checked for potential duplicates (none present) and subsequently merged horizontally using month, year, player, and team as the join key. According to domain knowledge, a filter is set to remove data points where the number of average minutes played is less than 10, or the number of games played is less than 2. After that, the data aggregation is employed for obtaining month-wide averages per team so as to reduce bias in the deployed model.

2. Data profiling

Data profiling is adopted to better understand the dataset and identify issues such as missing values. Specifically, the following five components in the raw dataset are computed for each column:

1. % of column populated (i.e. non nulls);
2. Number of unique/distinct values;
3. Data type;
4. Count (length of data-frame);
5. % of column being 0.

The general profile of the dataset is illustrated in Fig. 1. All the attributes used for prediction are numerical values, which removes the need for dummy variables. The only issue is the case of null values in the dataset. Though most of these nulls are not caused by the data quality issue since they are justified under a basketball context, nulls are a blocker to modelling. The following four solutions are shortlisted to solve this problem:

1. Remove rows which contain a null in any field.
2. Remove columns which contain a null in any row.
3. Replace the null field with a surrogate mean/median of all populated values for the column.
4. Replace the null field with a predicted value based on the populated fields of the row.

3. Data imputation

Given the essence of data volume as an important factor in this paper, especially when there is only so much NBA data available for analysis from the 1980s onwards, the imputation-based methods are used to deal with null values. It should be pointed out that certain viable imputation-based methods may not be appropriate for data imputation in the context of the NBA. With mean value imputation, a significant bias would exist as the values for the null fields are imputed based on the mean of all populated fields, which would highly generalise and skew the imputed value. For example,

	Populated	Unique	Dtype	Count	Pct of Zeros in Populated Fields
Player	100.000000	2214	object	50079	0.000000
Team	100.000000	30	object	50079	0.000000
GP	100.000000	19	float64	50079	0.000000
MIN	100.000000	353	float64	50079	0.000000
FGM	100.000000	135	float64	50079	0.093852
FGA	100.000000	256	float64	50079	0.007987
FG%	100.000000	581	float64	50079	0.093852
3PM	100.000000	69	float64	50079	38.881758
3PA	100.000000	139	float64	50079	22.152998
3P%	100.000000	469	float64	50079	38.861758
FTM	100.000000	110	float64	50079	2.028795
FTA	100.000000	144	float64	50079	1.577508
FT%	100.000000	652	float64	50079	2.028795
TOV	100.000000	79	float64	50079	0.519180
PF	100.000000	115	float64	50079	0.101839
ORB	100.000000	99	float64	50079	2.326324
DRB	100.000000	143	float64	50079	0.103836
REB	100.000000	197	float64	50079	0.043931
AST	100.000000	155	float64	50079	0.812716
STL	100.000000	60	float64	50079	2.226482
BLK	100.000000	82	float64	50079	12.887638
PTS	100.000000	347	float64	50079	0.047924
Month	100.000000	5	object	50079	0.000000
Year	100.000000	34	float64	50079	0.000000
TS%	100.000000	673	float64	50079	0.047924
eFG%	100.000000	622	float64	50079	0.093852
Total S %	100.000000	1946	float64	50079	0.047924
ORB%	97.673676	252	float64	50079	0.000000
DRB%	99.890104	397	float64	50079	0.000000
TRB%	99.956069	293	float64	50079	0.000000
AST%	99.127379	581	float64	50079	0.000000
TOV%	99.480820	427	float64	50079	0.000000
STL%	97.773518	81	float64	50079	0.000000
BLK%	87.114359	139	float64	50079	0.000000
USG%	99.998003	377	float64	50079	0.000000
PPR	98.246767	353	float64	50079	0.000000
PPS	100.000000	36	float64	50079	0.051918
ORtg	99.998003	1083	float64	50079	0.000000
DRtg	100.000000	455	float64	50079	0.000000
eDiff	99.998003	1121	float64	50079	0.343464
FIC	99.970047	2696	float64	50079	0.000000
PER	99.992013	428	float64	50079	0.000000

Fig. 1. The win ratio dataset.

when a player has a record of zero blocks, which can be highly possible that the player is an offensive based guard, and we impute his block percentage based on the mean of all populated block percentages (including defensive based players with high number of blocks per game). As such, the player may be assigned a higher block percentage than the ground truth. In this case, a specific imputation method based on predictive modelling is developed to deal with null values. The procedure of the developed imputation method by predictive modelling is shown as follows:

1. Delete the arbitrary fields (such as player name and team names).
2. Sort the column names which contain null values by the number of null values from largest to smallest via a list.
3. For each field in the null-field list from step 1, train a random forest regressor, since all the fields are of numerical nature, to predict the value of the null-field list in instances where the field in the null-field list is populated, based on the values of all other populated fields, using mean-value imputation if necessary if any of the other fields are null (cases of this should be minimal).
4. Use the trained random forest regressor model from step 3 to predict the field in the null-field list in stances where it is null.
5. Repeat from step 3 onwards for all other fields in the null-field list.

The developed imputation method gives a sense of assurance, compared with using mean value imputation methods, that the values for null fields are imputed using similar data-points (e.g. a similar performing player) where the field to impute is not null. It should be noted that the developed imputation approach also guarantees that there is no data loss since no rows or columns are removed.

4. Data cleaning

Data cleaning is applied to ensure consistency in representations of NBA seasons. A season spans over two years and the data source sometimes illustrates the season by the year the season ends, but at other times by the actual year where the game has taken place (in the case of per-game results). The subsequent step is to aggregate the dataset to a monthly average level to predict a team's monthly win ratio. Finally, prior to the modeling part, the dataset illustrating the monthly gameplay attribute average is joined up with the dataset depicting the win ratios per team each month, based on the team and date as the join key. As a result, a complete null-free dataset is obtained to predict win ratios based on a rich set of gameplay attributes.

3.2.2. NBA team post-season playoffs classification

3.2.2.1. Data preparation. Data is retrieved from the regular season average section in the data source instead (but includes both standard and advanced statistics), since whether a team makes the playoffs or not is based on a per-season basis without manual aggregation. It should be noticed that because whether or not a team makes the playoffs is not explicitly stated within the retrieved regular season average data, a separate identification process of determining whether a team makes the playoffs is conducted. The identification process is carried out through a "try/except" function on a HTTP GET request specifically for playoffs data. If the GET request succeeds, it verifies playoffs statistics are available for the given team that season, hence confirming that the team made the playoffs that season.

3.2.2.2. Data pre-processing. The data pre-processing of the classification problem is the same as that of the prediction problem as mentioned in Section 3.3.1. For the classification problem, we have a 1055 row dataset portraying each team's average seasonal gameplay attributes and whether the team made the NBA playoffs for each season since 1980.

4. Results and discussions

In this section, the testing methods, metrics, and experimental results are presented. The testing part involves 1) assessing the model outcomes from a quantitative perspective; and 2) evaluating the model from the domain specific perspective. As mentioned previously, domain expertise in NBA/basketball is crucial for validating the reliability of a machine learning model's output. Experimental results of the descriptive analysis for NBA gameplay are first shown where the testing process involves the comparison of various clustering algorithms, the granularity of the clusters produced, the proportion of variance explained by PCA, and visual inspection of the descriptive visualisations produced. Additionally, the success prediction of a given NBA team with interpretability analysis is demonstrated. To be specific, the validity and accuracy of the prediction results are analyzed, and the model interpretability using LIME is also presented.

4.1. Results of NBA gameplay descriptive analysis

In the simulation, the epsilon value of the DBSCAN algorithm is set up as the maximum distance between two samples for one to be considered as in the neighbourhood of the other, which is the most crucial parameter. With various values of epsilon in the range of $0 < x < 5$ considered in the simulation, the epsilon value of 2.4 is the optimal one in terms of the silhouette score and number of clusters formed. The corresponding silhouette score is 0.135, and the number of samples (or total weight) in a neighbourhood for a point to be considered as a core point is 10.

There are two major critical observations of the DBSCAN clustering results: 1) the silhouette score is low (a low or negative silhouette score indicates poor clustering); and 2) many samples (Table 1) lie in the anomalous "noisy" category of label −1. In contrast, when epsilon is changed to a higher value, such as 4.5, a much higher silhouette score (the value of 0.475) is obtained but with an undesired clustering breakdown. The min sample value is 10. The clustering results are provided in Table 2.

The clustering breakdown in Table 2 is undesirable in this paper to identify distinct clusters of NBA dynasties since there are virtually no clusters identified, despite the higher silhouette score. This indicates that the data potentially may not conform to density-based clustering. To further explore the data distribution, the K-means clustering algorithm is employed. The cluster size and silhouette scores of K-means clustering are displayed in Table 3, where all hyperparameters are left as default.

It can be seen that the K-means clustering outcome is far from optimal in terms of the silhouette scores. However, it is clear that the performance of the K-means clustering algorithm is better than that of the DBSCAN algorithm in terms of silhouette scores, sug-

Table 1
DBSCAN clustering results with epsilon of 2.4 (* the cluster label of −1 in DBSCAN refers to a data-point being a "noisy" sample).

Cluster	Cluster Size
0	156
1	572
−1*	327

Table 2
DBSCAN clustering results with epsilon of 4.5 (* the cluster label of −1 in DBSCAN refers to a data-point being a "noisy" sample).

Cluster	Cluster Size
0	1054
−1*	1

Table 3
K-means clustering results.

Cluster:	Silhouette Score:
2	0.317
3	0.213
4	0.170
5	0.133
6	0.122
7	0.120
8	0.115
9	0.119

Table 4
The clustering results for 5 most frequently appearing NBA seasons.

Cluster Size:	Label:	Cluster Size:	5 Most frequently appearing NBA seasons:
2	0	715	**1990, 1989, 1991, 1986, 1987.**
	1	340	**2008, 2005, 2006, 2007, 2004.**
3	0	420	**2006, 2003, 2004, 2005, 1999.**
	1	330	**1990, 1989, 1988, 1986, 1991.**
	2	305	**2019, 2018, 2016, 2015, 2014.**

gesting that a density-based clustering approach through DBSCAN is probably unsuitable on this dataset. Thus, the K-means clustering output is used for the subsequent implementations in PCA and visualisation. Using the cluster sizes with the top silhouette scores, a further breakdown is displayed in Table 4.

It can be observed that certain similar seasons appear frequently within the same cluster, and the most prevalent NBA seasons in each cluster are from vastly different time periods (e.g. early 1990s vs late 2010s), signalling that the NBA gameplay has evolved over the past four decades. To explore further, the results of PCA and visualisation ensue.

Choosing a suitable number of principal components aims to balance the trade-off between the number of principal components and the proportion of variance explained by the principal components. In fact, the sole purpose of PCA is to reduce dimensionality

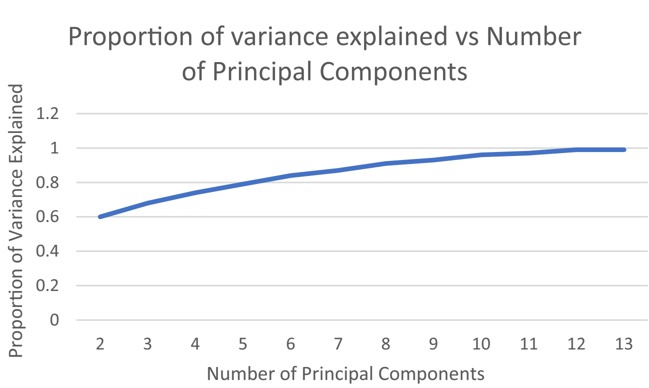

Fig. 2. The proportion of variance explained vs the number of principal components.

down to either 2 or 3 for visualisation purposes. As such, the proportion of variance explained by 2 and 3 principal components must suffice for reliable visualization. A scree plot of number of principal components vs proportion of variance explained by the principal components is depicted in Fig. 2.

In Fig. 2, the proportion of variance explained by 2 and 3 principal components is 0.6 and 0.68, respectively. The results can be seen as a reasonable but subjective approximation of the information in the dataset. Due to this approximation, the ensuing diagrams act as purely a reference for visualisation and not as the golden standard for describing the evolution of the NBA gameplay. For K-means with 2 and 3 clusters, the subsequent images depict the following:

- 2D representation of the clusters (each cluster identified through a distinct colour) with the projection of each original feature as a labelled vector on the principal component axes.
- 3D representation of the clusters (each cluster identified through a distinct colour).
- 3D representation with each data-point coloured according to its corresponding NBA season/year.

Figs. 3-5 demonstrate that: 1) there is a difference in gameplay between the various clusters identified; and 2) there is a gradual shift in gameplay as we alter the timeframe. Focusing on the clustering to begin with, the tables below exhibit the key descriptive statistics for selected attributes within each cluster, highlighting the major discrepancies in gameplay attributes between each cluster.

The clustering visuals and their accompanying breakdown in Tables 5 and 6 suggest two to three potential NBA dynasties, each with a distinct basketball style.

To visually inspect individual attributes which may be contributing to this evolution, Fig. 6 depicts the trend analysis of a select few gameplay attributes over the years (essentially interpreting Fig. 5). As all values have been standardised, all ten attributes are displayed upon one graph. The trend is quite intriguing from the perspective of an NBA fan. A few observations, along with some possible hypothesises for each phenomenon, are obtained which include the following:

- 3-point attempts and 3 points made have continuously increased over the four decades;
- Defensive rebounds have continuously increased over the four decades;
- Offensive rebounds have continuously decreased over the four decades;
- Free throw attempts and free throws made have continuously decreased over the four decades.

From the above statistics, the style of the NBA is truly different depending on the chosen year. A notable observation is that that a successful player in 1989 may not be able to automatically translate his game into success in 2019.

4.2. Results of NBA team's success strategy using explainable AI

4.2.1. Win ratio prediction

In this experiment, random forest and feed-forward NN algorithms are employed for the win prediction of a given NBA team. The hyperparameters of the two models, after hyperparameter tuning, are given as follows:

Random Forest:

- Max depth: 20
- Split criterion: MSE

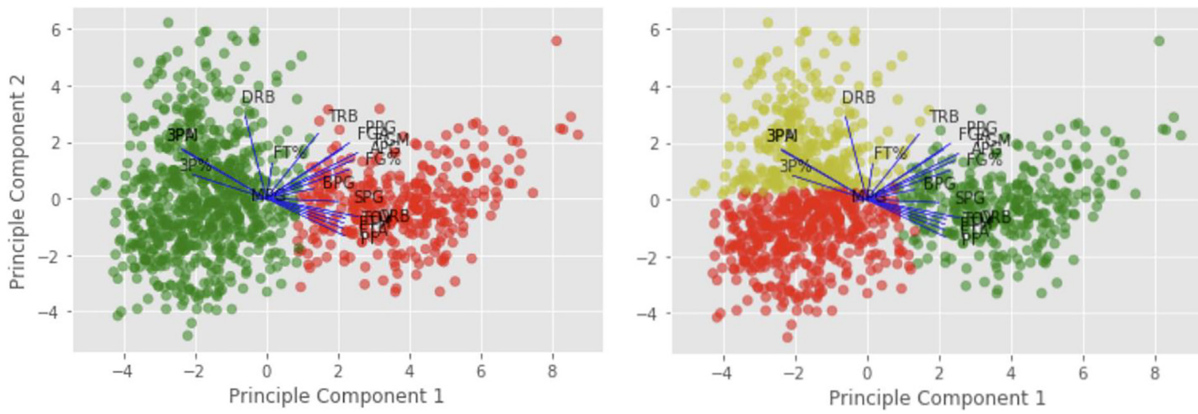

Fig. 3. 2D representation of the clusters (left is the 2 clusters' case and right is the 3 clusters' case).

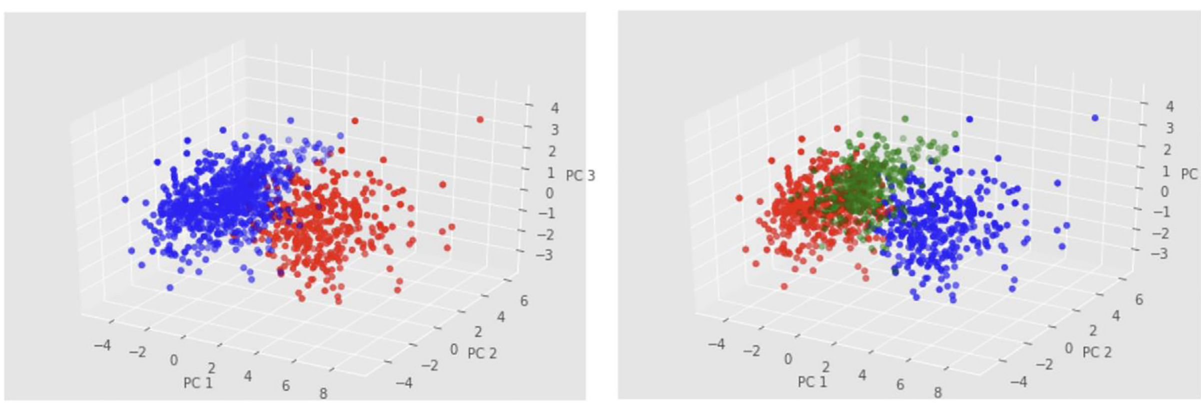

Fig. 4. 3D representation of the clusters (left is the 2 clusters' case and right is the 3 clusters' case).

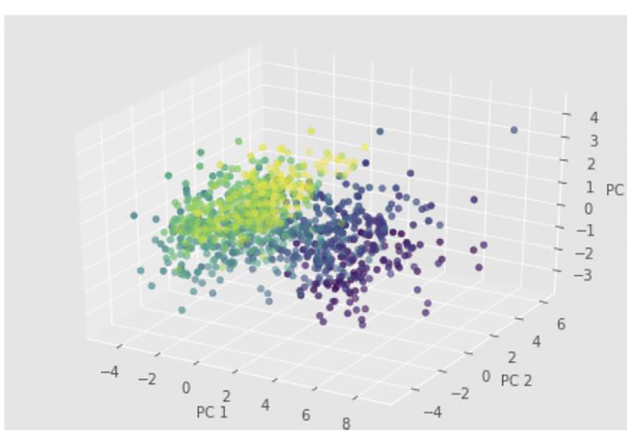

Fig. 5. 3D representation with each data-point coloured according to its corresponding NBA season/year (darker colours representing a more recent NBA season).

- Number of estimators: 200
- Maximum features: sqrt
- Out of bound score: True

Feed-forward NN:

- Number of hidden layers: 4
- Number of nodes on each layer (in order from front to end): 4, 3, 2, 1
- Model type: Sequential feed-forward NN with dense layers
- Activation type: ReLU for all layers
- Optimizer type: Adam
- Epochs: 40
- Loss: Mean Squared Error
- Validation split within the training set: 0.2
- Early stopping patience: 5

Experimental results from the random forest as well as the feed-forward NN to validate the monthly win ratio predictions are exhibited in the following table:

As shown in Table 7 that the predictive power of the feed-forward NN is superior over the random forest model based on the higher R^2 value and lower MSE value. Whilst the Leave-10-Out-Cross-Validation method is unfeasible for the NN model due to computation constraints, one would expect the metrics to exceed the Train/Test split value given a larger training set for the model to learn from (as proven in the random forest instance). An R^2 value of minimum 0.77 in accurately predicting a team's win ratio for a given month by a NN seems subjectively acceptable, which indicates a significant improvement over 0.65 in the case of a random forest model.

To test the interpretability of each method, LIME, the explainable AI method, is utilized to derive local feature importance. To test the effectiveness of local model interpretation, three specific samples are taken into consideration for verification:

1. A phenomenally successful month with a 0.94 win ratio in the Lebron James-led Miami Heat of Feb 2013;
2. An average month with a 0.5 win ratio by the New York Knicks of February 1992;

Table 5
Clustering results when N = 2.

N = 2 Clustering approach:													
Cluster 0	**FGM**	**3PM**	**3PA**	**3P%**	**FTM**	**FTA**	**TOV**	**PF**	**ORB**	**APG**	**SPG**	**PPG**	**Year**
mean	**42.4**	**1.7**	**5.4**	**0.3**	**21.7**	**28.7**	**17.0**	**24.2**	**14.4**	**25.5**	**8.8**	**108.2**	**1987.5**
std	2.4	1.2	3.3	0.0	2.2	2.8	1.9	1.8	1.0	1.8	0.9	4.8	4.8
min	34.9	0.3	1.5	0.1	15.8	20.5	11.8	19.5	11.9	21.2	6.3	97.5	1980.0
25%	40.8	0.7	2.6	0.3	20.2	26.9	15.5	22.9	13.7	24.3	8.2	104.9	1984.0
50%	42.2	1.5	4.8	0.3	21.7	28.9	17.0	24.2	14.3	25.5	8.7	107.5	1987.0
75%	43.7	2.4	7.1	0.3	23.2	30.5	18.2	25.4	15.0	26.5	9.4	110.9	1991.0
max	52.0	7.1	18.7	0.4	29.0	37.5	22.8	29.9	18.2	30.9	11.4	130.8	2010.0
Cluster 1	**FGM**	**3PM**	**3PA**	**3P%**	**FTM**	**FTA**	**TOV**	**PF**	**ORB**	**APG**	**SPG**	**PPG**	**Year**
mean	**37.0**	**6.5**	**18.2**	**0.4**	**18.4**	**24.5**	**14.7**	**21.4**	**11.5**	**21.9**	**7.7**	**98.8**	**2006.8**
std	2.3	2.0	5.4	0.0	2.0	2.7	1.2	1.7	1.2	1.9	0.8	5.8	7.4
min	30.3	1.9	6.1	0.3	13.7	18.2	11.3	16.2	8.0	16.3	5.6	83.4	1990.0
25%	35.5	5.2	14.6	0.3	17.0	22.5	13.7	20.2	10.6	20.6	7.2	95.0	2001.0
50%	37.0	6.1	17.1	0.4	18.3	24.3	14.6	21.4	11.4	21.9	7.7	98.5	2007.0
75%	38.6	7.4	20.4	0.4	19.8	26.2	15.5	22.6	12.4	23.2	8.2	102.5	2013.0
max	43.4	13.1	36.3	0.4	24.7	32.4	19.4	27.1	15.3	26.9	10.4	119.4	2019.0

Table 6
Clustering results when N = 3.

N = 3 Clustering approach:													
Cluster 0	**FGM**	**3PM**	**3PA**	**3P%**	**FTM**	**FTA**	**TOV**	**PF**	**ORB**	**APG**	**SPG**	**PPG**	**Year**
mean	**35.7**	**5.3**	**15.2**	**0.4**	**19.0**	**25.4**	**14.9**	**22.1**	**11.8**	**21.2**	**7.7**	**95.6**	**2002.5**
std	1.8	1.1	2.9	0.0	2.0	2.6	1.3	1.6	1.1	1.7	0.8	4.3	5.6
min	30.3	1.6	4.8	0.3	13.9	19.4	11.7	17.7	9.3	16.3	5.6	83.4	1990.0
25%	34.5	4.6	13.5	0.3	17.5	23.5	14.0	20.9	11.0	20.0	7.2	92.8	1998.0
50%	35.8	5.4	15.2	0.4	19.0	25.4	14.9	22.0	11.9	21.2	7.7	96.0	2003.0
75%	36.8	6.0	17.0	0.4	20.3	27.1	15.7	23.2	12.7	22.3	8.2	98.6	2006.3
max	40.9	8.5	24.1	0.4	25.0	33.6	20.6	27.1	14.7	26.3	10.4	105.6	2016.0
Cluster 1	**FGM**	**3PM**	**3PA**	**3P%**	**FTM**	**FTA**	**TOV**	**PF**	**ORB**	**APG**	**SPG**	**PPG**	**Year**
mean	**42.5**	**1.6**	**5.2**	**0.3**	**21.7**	**28.7**	**17.0**	**24.2**	**14.4**	**25.6**	**8.8**	**108.3**	**1987.2**
std	2.3	1.1	3.0	0.0	2.2	2.8	1.9	1.8	1.0	1.7	0.9	4.7	4.5
min	38.0	0.3	1.5	0.1	15.8	20.5	11.8	19.5	11.9	21.2	6.3	97.5	1980.0
25%	40.8	0.7	2.6	0.3	20.1	26.9	15.5	23.0	13.7	24.5	8.2	105.3	1983.3
50%	42.3	1.4	4.6	0.3	21.7	28.8	17.0	24.2	14.3	25.5	8.8	107.6	1987.0
75%	43.8	2.3	7.0	0.3	23.2	30.5	18.2	25.5	15.1	26.5	9.4	110.9	1991.0
max	52.0	6.8	18.7	0.4	29.0	37.5	22.8	29.9	18.2	30.9	11.4	130.8	2002.0
Cluster 2	**FGM**	**3PM**	**3PA**	**3P%**	**FTM**	**FTA**	**TOV**	**PF**	**ORB**	**APG**	**SPG**	**PPG**	**Year**
mean	**38.8**	**8.0**	**22.2**	**0.4**	**17.7**	**23.4**	**14.4**	**20.5**	**11.1**	**22.8**	**7.7**	**103.3**	**2012.4**
std	1.6	2.0	5.5	0.0	1.8	2.3	1.2	1.5	1.3	1.7	0.7	4.7	5.7
min	34.1	2.8	8.5	0.3	13.7	18.2	11.3	16.2	8.0	17.8	5.7	89.4	1993.0
25%	37.7	6.5	18.3	0.3	16.4	21.8	13.5	19.5	10.2	21.6	7.2	100.2	2010.0
50%	38.6	7.6	21.4	0.4	17.7	23.3	14.4	20.4	11.0	23.0	7.6	102.7	2014.0
75%	39.8	9.4	26.3	0.4	18.8	24.7	15.1	21.4	11.8	23.9	8.1	106.3	2017.0
max	43.4	13.1	36.3	0.4	22.8	29.8	18.6	24.7	15.3	26.9	9.9	119.4	2019.0

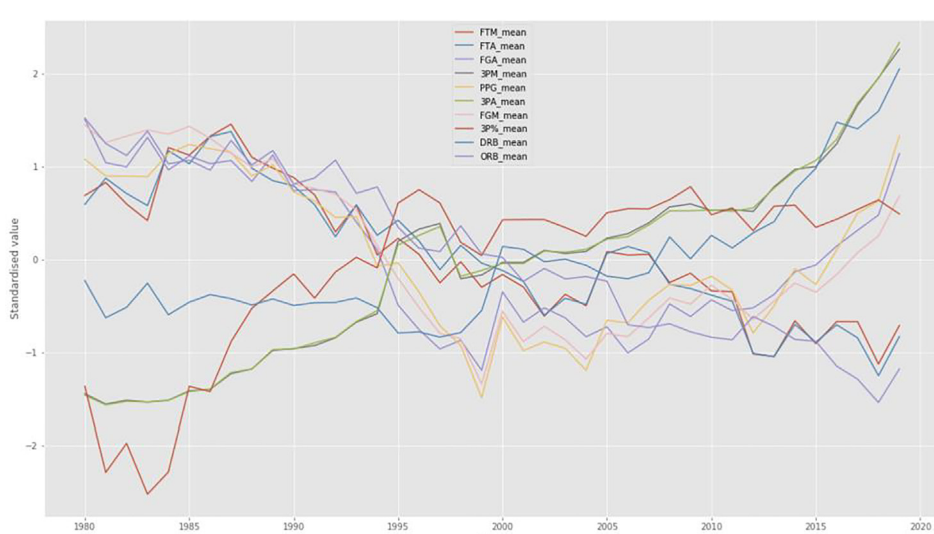

Fig. 6. Trend analysis of a select few gameplay attributes over the years.

Table 7
Experimental results of the random forest and the feed-forward NN.

Dataset length: 4718	**Random Forest Model**	**Feed-forward NN Model**
90/10 train and test split average over 10 random splits (randomness determined using seeds)	R^2: 0.65 MSE: 0.014	R^2: 0.77 MSE: 0.009
Leave 10 out cross validation	R^2 score: 0.66	NA (due to computational constraints)

3. A poor performing month with a 0.1 win ratio by the Denver Nuggets of February 2015.

In Figs. 7-9, the local feature importance and its positive/negative contribution to the predicted value are depicted for the 3 chosen samples. For demonstrative purposes, a verbal analysis of samples 1 and 3 follows, as an example of determining whether exceptionally successful and unsuccessful months can be logically explained in basketball terms.

For sample 1, the model predicts a win ratio of 0.72 compared to the actual ratio of 0.94. The following three conclusions are drawn from the experimental results.

1) The top attributes contributing to a positive value include a high floor impact (which is the mean impact each player gives in terms of assists, shot creation, and offensive rebounding), a good defensive rating mean (that limits the number of opposition points), and a good offensive firepower. From a basketball standpoint, these make complete sense and follow the general coaching principle that defence and impacting other players around the player is more crucial in winning games than purely focusing on offensive plays. The finding is highly relevant with a famous basketball phrase “Defence wins championships”.
2) The top three attributes inhibiting a high win ratio are high number of games played, high number of steals and a high assist percentage. A high number of mean steals and a high assist percentage should in theory be positive traits in an NBA game.
3) The predicted value of 0.72 is not as high as the actual win ratio of 0.92. This is a general trend among predicting test cases where the win ratio is exceptionally high because of the smaller sample size around the tails of the win ratio distribution for the model to train on.

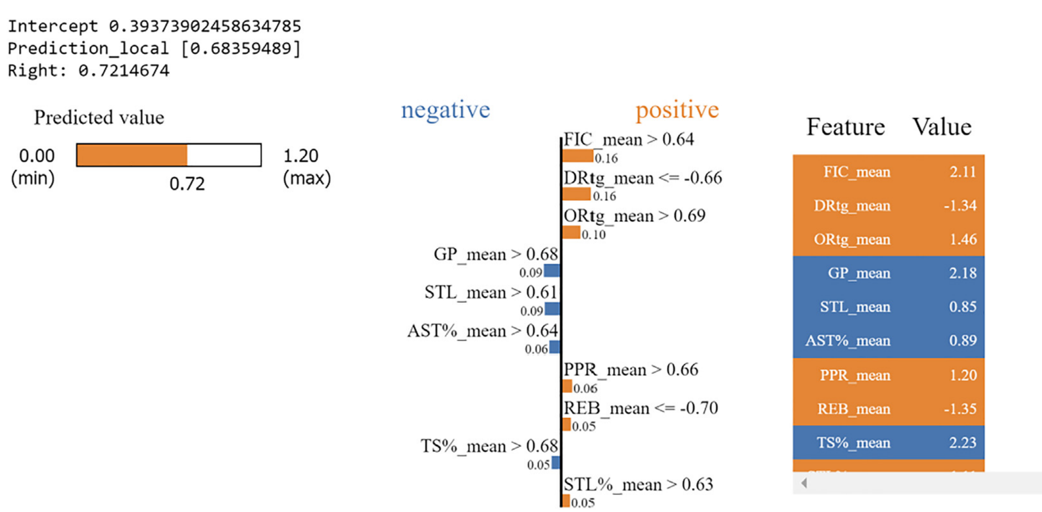

Fig. 7. LIME results of sample 1 (the 2013 case).

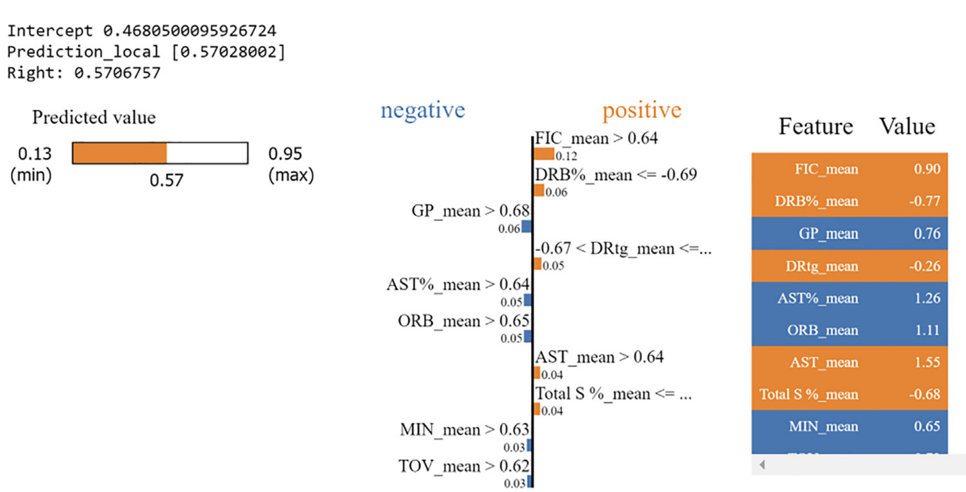

Fig. 8. LIME results of sample 2 (the 1992 case).

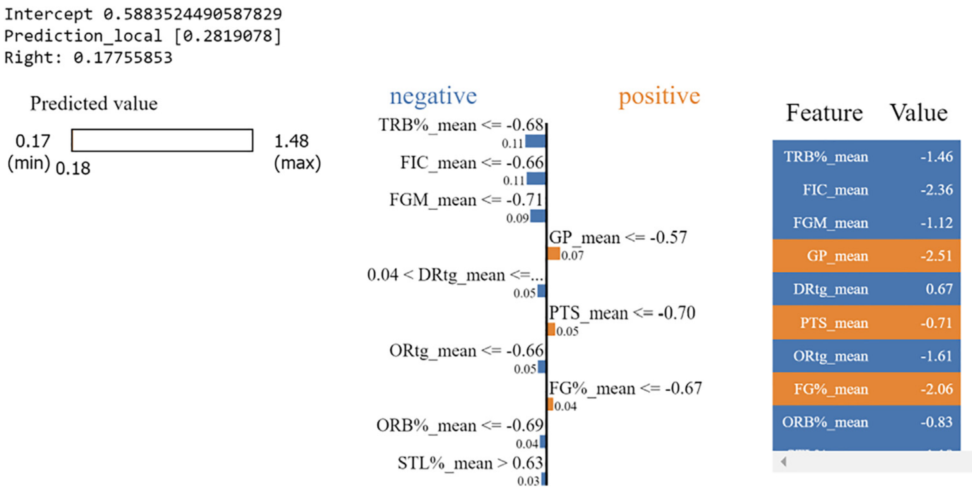

Fig. 9. LIME results of sample 3 (the 2015 case).

For sample 3, the model predicts a win ratio of 0.17 compared to the actual ratio of 0.1. According to the experimental results, the following two conclusions are outlined:

1) Low total rebounding percentage, low floor impact, and low field goals made sum up the top attributes explaining the horrendous month of basketball. Low rebounding percentages mean low second chance opportunities per possession and giving up possessions on play to the opposition. The coach in this instance should thus request extra emphasis on going after each rebound and increasing assists to improve floor impact in subsequent games.
2) Like the above scenario, the contrasting attributes seem to make little basketball sense. A low mean points scored value along with a low field percentage, which should theoretically negatively impact a team's performance, are classified as positive contributions.

It is difficult to employ the AI techniques (e.g. the random forest model and the feed-forward NN) to accurately predict the win ratio of a team for a given month, even with advanced NBA statistics. According to experimental results, a feed-forward NN predicts the win ratio of a given NBA team within a month with a respectable accuracy of 0.77 using the R^2 metric, which indicates that the feed-forward NN should generally be able to give a solid indicator of how a team will perform for a given month, given a set of NBA attributes. LIME produces rational interpretability behind each predicted value, but the contributing factors for the contrasting side, particularly the positive contributions, for each prediction sometimes make little sense in basketball terms.

4.2.2. NBA team post-season playoffs classification

In this experiment, the post parameter tuning, the optimal structure and hyperparameter values for the random forest and feed-forward NN models are given as follows:

Random Forest:

- Max depth: 25
- Split criterion: gini
- Number of estimators: 250
- Maximum features: sqrt
- Bootsrap: True

Feed-forward NN:

- Number of hidden layers: 4
- Number of nodes on each layer (in order from front to end): 5, 4, 3, 2
- Model type: Sequential feed-forward NN with dense layers
- Activation type: Linear. Softmax for final layer
- Optimizer type: Adam
- Loss function: Categorical cross entropy
- Epochs: 40
- Validation split within the training set: 0.2
- Early stopping patience: 5

The results from the various testing methodologies (including the random forest model and the feed-forward NN) to validate the monthly win ratio predictions are exhibited in the tables below:

The classification task is much easier than the corresponding win-ratio prediction task, given it is firstly a binary classification and secondly it is at a season-by-season level as opposed to a monthly level. Thirdly, from an NBA domain perspective, it is simply more complicated to predict exactly how many games a team will win in a specific month given the volatility of how teams perform game to game as opposed to if that team is generally of playoff calibre, given a set of gameplay attribute values. As shown in Table 8, we can see that the F1 score is relatively high across all metrics for both models, with the feed-forward NN having a slight advantage over the random forest model.

The three NBA team cases, the 2015–2016 Golden State Warriors, the 2011–2012 Charlotte Bobcats as well as the 2010–2011 Golden State Warriors, are employed to verify the effectiveness of local model interpretation on the NN classifier.

Table 8
Experimental results of the random forest model and the feed-forward NN.

Dataset length: 1055	**Random Forest Model**	**Feed-forward NN Model**
90/10 train and test split average over 10 random splits (randomness determined using seeds)	F1: 0.85	F1: 0.87
Leave 100 out cross validation/11 Fold Cross Validation	F1: 0.87	F1: 0.87
Leave 2 out cross validation	F1: 0.88	N/A (due to computational constraints)

The first sample is the best NBA season in history, which was obtained by the 2015–2016 Golden State Warriors:

a) 99.999% probability of making the playoffs
b) High defensive intensity, high offensive firepower, high field goal percentage (converting your shots to points) and high number of 3 points made (thanks to the Warriors' "splash brothers") contribute significantly to making the playoffs. Interestingly, where the casual NBA follower may instantly come to the conclusion that the Warriors' three pointers are core to their success, our results suggest it is more their defence (a large part down to Draymond Green presumably) which anchored their championship run, with three points contribution only 4th in our feature importance ranking.
c) Low offensive rebounding, too many field goal attempts and playing too fast counter the positive contributions but they are vastly overpowered by their positive counterparts.

The second sample is the worst NBA season in history, which was obtained by the 2011–2012 Charlotte Bobcats:

a) 99.999% probability of not making the playoffs.
b) No gameplay attributes in the top ten contributing to making the playoffs.
c) Each of the top ten attributes make logical sense in basketball terms.

The third sample is a relatively average season, which is the 2010–2011 Golden State Warriors who failed to make the playoffs:

a) 58% of not making the NBA playoffs for an average season, a correctly predicted classification.
b) Allowing too many points scored by the opposition is the core factor resulting in the lack of success. High steal percentage is, however, a substantial bright spot in the season.
c) Attributes contributing both positively and negatively make logical sense in basketball terms.

The local feature important rankings of the three cases are depicted in Figs. 10-12.

As shown in Figs. 10-12, the feed-forward NN could predict that a playoff-calibre team indeed makes the playoffs given a set of gameplay attributes (and vice versa) with high confidence. Given the relative simplicity in nature of predicting whether an NBA team makes the playoffs or not for a given season, both random forest and NN model perform well on unseen data, with the feed-forward NN having the edge with a minimum 0.88 F1 score. In this case, LIME produces rational interpretability behind each predicted value and validates the common basketball theory that emphasis on defence is the key to becoming a playoff-worthy team.

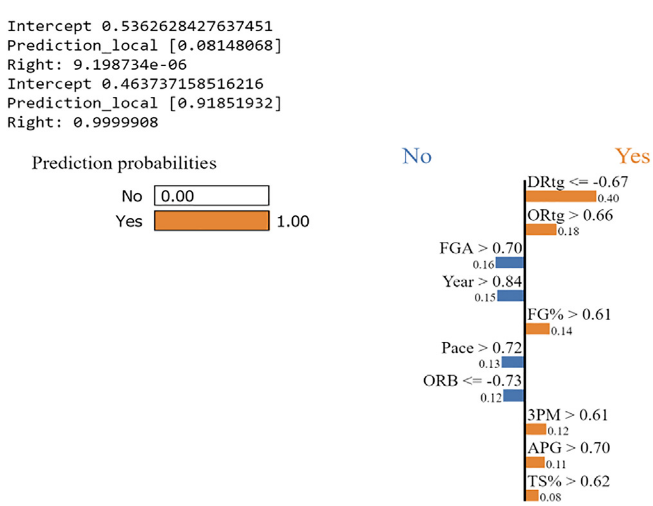

Fig. 10. LIME results of sample 1 (the 2015–2016 case).

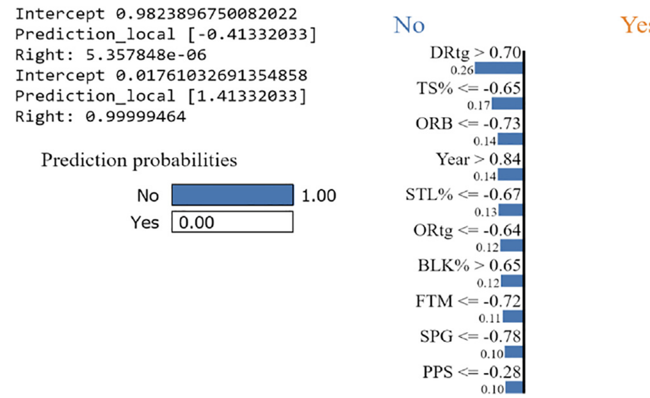

Fig. 11. LIME results of sample 2 (the 2011–2012 case).

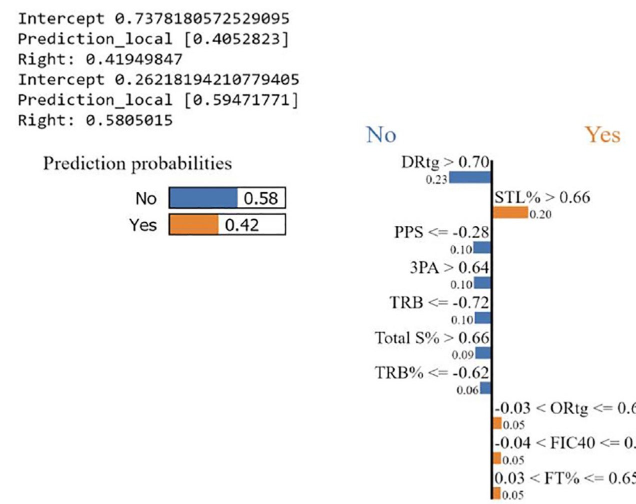

Fig. 12. LIME results of sample 3 (the 2010–2011 case).

In [21], the data mining techniques produced a probability of 0.67 in predicting a correct match winner. Furthermore, in [35], the machine learning methods which were implemented more recently in 2019 produced test set F1 scores in the range of 0.71 to 0.83 (depending on algorithm implemented) in predicting games, based on their respective datasets. With the extensive source of attributes scraped and pre-processed for use in this paper, this project's testing dataset result of 0.77 R^2 for monthly win ratio prediction (regression) and a testing dataset result of 0.88 F1 for playoff prediction (classification) trumps both results exhibited in the two most recent literatures mentioned above in predicting macro NBA outcomes. In addition to purely producing powerful predictive models, this paper also incorporates the domain of reliably interpreting each individual prediction, such that it can be appropriately leveraged by an NBA coach. This is an area not explored in the open source community based on the literature reviews undertaken.

4.3. Discussion

In this paper, the year-by-year changes in key gameplay attributes are firstly graphically illustrated and compared to back the widely-renowned claim that the NBA has changed over the past decades. Secondly, three NBA cases are analyzed by using the K-means clustering algorithm, each with a distinctive set of gameplay attributes as shown in Tables 5 and 6, again signalling gameplay evolution. Lastly, the clustering results are holistically depicted in Figs. 3-5 and the findings are demonstrated into 2D and 3D visuals, essentially summarising the findings into one diagram.

A feed-forward NN is applied to predict the win ratio per month and the seasonal playoff-making probability for a given team with respectable metrics of 0.77 R^2 and 0.87 F1. Furthermore, given the difficulties in interpreting the predictions of complex models such as NNs compared to say, random forests, LIME is employed to produce sensible reasoning behind each prediction. The established model could give an NBA fan/coach not only insights into whether the team can win games or make the playoffs, but also the corresponding reasons to understand the team's strengths and weaknesses. In addition, the LIME outcomes validate a basketball claim, which is that defence as opposed to offense is the fundamental key to success on the court.

5. Conclusion

In this paper, an overview of AI applications in sports science, particularly the NBA case, has been provided due to the lack of open-source literatures. An exploration of AI techniques in NBA has been conducted incorporating both a) accurately predicting macro-level (seasonal/game level) NBA outcomes; and b) interpreting individual predictions using an explainable-AI approach so as to validate and assess the accuracy of the interpretation. A descriptive analysis of the NBA gameplay has been carried out to analyze the evolution of the NBA gameplay from the 1980s to the current modern era from a statistical standpoint. The popular machine learning techniques (e.g. random forest and feed-forward NN) have been applied to predict the win ratio of a specific month and the seasonal playoff-making probability for a given NBA team. Experimental results have demonstrated the effectiveness of the established AI models. In addition, the explainable AI technique (the LIME method) has been successfully utilized to produce sensible reasoning of the predictions.

In the future, we aim to: 1) employ deep learning and other machine learning techniques for NBA gameplay prediction [20,36,37,39,40,42,44,46]; 2) study the robustness and generalization ability of the established machine learning models based on signal processing and system science techniques [22,34]; 3) adopt evolutionary computation algorithms (e.g. the particle swarm optimization algorithm, the genetic algorithm and the artificial bee colony algorithm) to choose the parameters of the machine learning models [14,19,27,29,32,33]; and 4) apply the developed models to other areas such as telecommunication, electrical engineering, and medical science [15,45].

CRediT authorship contribution statement

Yuanchen Wang: Conceptualization, Methodology, Software, Writing – original draft. **Weibo Liu:** Supervision, Conceptualization, Methodology, Writing – review & editing. **Xiaohui Liu:** Supervision, Methodology, Software, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author would like to express his most sincere gratitude to Dr Paul Yoo for his support especially in literature review and motivation clarification. The first author would also like to thank Dr Szabolcs Mikulas for his inaugural guidance in assessing and probing the proposed research topic.

References

[1] S. Balli and E. Ozdemir, A novel method for prediction of EuroLeague game results using hybrid feature extraction and machine learning techniques, Chaos, Solit. Fract., vol. 150, art. no. 111119, 2021.
[2] R. Beal, T. Norman and S. Ramchurn, Artificial intelligence for team sports: a survey, Knowl. Eng. Rev., 34, art. no. 28, 2019
[3] R.P. Bunker, F. Thabtah, A machine learning framework for sport result prediction, Appl. Comput. Informat. 15 (1) (2019) 27–33.
[4] C.-Y. Chen, W. Lai, H.-Y. Hsieh, W.-H. Zheng, Y.-S. Wang and J.-H. Chuang, Generating defensive plays in basketball games, In: *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Korea, Oct. 2018, pp. 1580-1588.
[5] W.-C. Chu, C. Shih, W.-Y. Chou, S.I. Ahamed, P.-A. Hsiung, Artificial intelligence of things in sports science: weight training as an example, Computer 52 (11) (2019) 52–61.
[6] C. Dindorf, J. Konradi, C. Wolf, B. Taetz, G. Bleser, J. Huthwelker, F. Werthmann, E. Bartaguiz, J. Kniepert, P. Drees and U. Betz, Classification and automated interpretation of spinal posture data using a pathology-independent classifier and explainable artificial intelligence (XAI), *Sensors*, vol. 21, no. 18, art. no. 6323, 2021
[7] B. Ergül, A.A. Yavuz, H.S. Yavuz, Classification of NBA league teams using discriminant and logistic regression analyses, Pamukkale J. Sport Sci. 5 (1) (2014) 48–60.
[8] J. C. Gustavo, D. O. Capanema, T. V. Souza, J. C. Serrão, A. C. M. Pereira, G. P. Nassis, Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review, *Sports Medicine - Open*, vol. 5, no. 1, art. no. 28, 2019.
[9] P. He, H. Li, X. Luo, M. Xing, Exponential synchronisation of linearly coupled reaction-diffusion neural networks with discrete and infinite distributed delays, Int. J. Syst. Sci. 51 (7) (2020) 1174–1187.
[10] M. Hildebrandt , J. A. Serna, Y. Ma, M. Ringsquandl, M. Joblin and V. Tresp, Reasoning on knowledge graphs with debate dynamics, In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, USA, Feb. 2020, vol. 34, no. 04, pp. 4123-4131.
[11] L. Hu, W. Naeem, E. Rajabally, G. Watson, T. Mills, Z. Bhuiyan, C. Raeburn, I. Salter, C. Pekcan, A multiobjective optimization approach for COLREGs-compliant path planning of autonomous surface vehicles verified on networked bridge simulators, IEEE Trans. Intell. Transp. Syst. 21 (3) (2020) 1167–1179.
[12] Z. Ivanković, M. Racković, B. Markoski, D. Radosav and M. Ivković, Analysis of basketball games using neural networks, In: *Proceedings of the 11th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, Hungary, Nov. 2010, pp. 251-256.
[13] K. Johnson, F. Pasquale, J. Chapman, Artificial intelligence, machine learning, and bias in finance: toward responsible innovation, Fordham Law Review 88 (2) (2019) 499–529.
[14] X. Kan, Y. Fan, Z. Fang, L.e. Cao, N.N. Xiong, D. Yang, X. Li, A novel IoT network intrusion detection approach based on Adaptive Particle swarm optimization convolutional neural network, Inf. Sci. 568 (2021) 147–162.
[15] E. Kanda, B. I. Epureanu, T. Adachi, Y. Tsuruta, K. Kikuchi, N. Kashihara, M. Abe, I. Masakane and K. Nitta, Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan, *Plos one*, vol. 15, no. 5, art. no. e0233491, 2020
[16] Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, *Omega*, vol. 98, art. no. 102123, 2021.
[17] B. Loeffelholz, E. Bednar and K. W. Bauer, Predicting NBA games using neural networks, J. Quantitat. Anal. Sports, vol. 5, no. 1, art. no. 7, 2009
[18] López-Valenciano, F. Ayala, J. M. Puerta, M. D. S. Croix, F. Vera-García, S. Hernández-Sánchez, I. Ruiz-Pérez, and G. Myer, A preventive model for muscle injuries: a novel approach based on learning algorithms, Med. Sci. Sports Exerc., vol. 50, no. 5, pp. 915-927, 2018
[19] X. Luo, Y. Yuan, S. Chen, N. Zeng and Z. Wang, Position-transitional particle swarm optimization-incorporated latent factor analysis, *IEEE Trans. Knowl. Data Eng.*, in press, doi: 10.1109/TKDE.2020.3033324
[20] M. Mafarja, S. Abdullah, A fuzzy record-to-record travel algorithm for solving rough set attribute reduction, Int. J. Syst. Sci. 46 (3) (2015) 503–512.
[21] D. Miljković, L. Gajić, A. Kovačević and Z. Konjović, The use of data mining for basketball matches outcomes prediction, In: *Proceedings of the 8th International Symposium on Intelligent Systems and Informatics*, Subotica, Serbia, Sept. 2010, pp. 309-312.
[22] B. Min, F. Xu, J. Tan, X. Wang, B. Liang, Multiple actuator-fault detectability analysis using invariant sets for discrete-time LPV systems, Int. J. Syst. Sci. 51 (16) (2020) 3451–3470.

[23] C. Oxborough, E. Cameron, A. Rao, A. Birchall, A. Townsend and C. Westermann, Explainable AI: Driving business value through greater understanding, Retrieved from PWC website: https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf, 2018.
[24] H. Park, K. Maruhashi, R. Yamaguchi, S. Imoto and S. Miyano, Global gene network exploration based on explainable artificial intelligence approach, *Plos One*, vol. 15, no. 11, art. no. e0241508, 2020
[25] K. Pelechrinis, Linnet: Probabilistic lineup evaluation through network embedding, In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Dublin, Ireland, Sept. 2018, pp. 20-36.
[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, USA, Aug. 2016, pp. 1135-1144.
[27] H.E. Rouabhia, B. Farou, Z.E. Kouahla, H. Seridi, H. Akdag, Cooperative processing based on posture change detection and trajectory estimation for unknown multi-object tracking, Int. J. Syst. Sci. 50 (13) (2019) 2539–2551.
[28] J. Ruddy, A. Shield, N. Maniar, M. Williams, S. Duhig, R. Timmins, J. Hickey, M. Bourne, D. Opar, Predictive modeling of hamstring strain injuries in elite Australian footballers, Med. Sci. Sports Exerc. 50 (5) (2018) 906–914.
[29] D. Ruiz-Fernandez, Y. Quintana-Pacheco, Heuristic method for searches on large data-sets organised using network models, Int. J. Syst. Sci. 47 (7) (2016) 1725–1733.
[30] M. Ryo, B. Angelov, S. Mammola, J.M. Kass, B.M. Benito, F. Hartig, Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models, Ecography 44 (2) (2021) 199–205.
[31] X. Schelling, J. Fernández, P. Ward, J. Fernández and S. Robertson, Decision Support System Applications for Scheduling in Professional Team Sport. The Team's Perspective, *Front. Sports Active Liv.*, vol. 3, art. no. 678489, 2021.
[32] H. Sharma, J.C. Bansal, K.V. Arya, X.S. Yang, Levy flight artificial bee colony algorithm, Int. J. Syst. Sci. 47 (11) (2016) 2652–2670.
[33] X. Shi, Z. Lin, T. Yang, X. Wang, Distributed dynamic event-triggered algorithm with minimum inter-event time for multi-agent convex optimisation, Int. J. Syst. Sci. 52 (7) (2021) 1440–1451.
[34] Z. Tang, E. Tian, Y. Wang, L. Wang, T. Yang, Nondestructive defect detection in castings by using spatial attention bilinear convolutional neural network, IEEE Trans. Ind. Inf. 17 (1) (2021) 82–89.
[35] F. Thabtah, L.i. Zhang, N. Abdelhamid, NBA game result prediction using feature analysis and machine learning, Ann. Data Sci. 6 (1) (2019) 103–116.
[36] X. Wan, Z. Fang, M. Wu, Y. Du, Automatic detection of HFOs based on singular value decomposition and improved fuzzy *c*-means clustering for localization of seizure onset zones, Neurocomputing 400 (2020) 1–10.
[37] H. Wang, S. Lu, Z. Zhou, Ramp loss for twin multi-class support vector classification, Int. J. Syst. Sci. 51 (8) (2020) 1448–1463.
[38] K. C. Wang, R. Zemel, Classifying NBA offensive plays using neural networks, In: *Proceedings of the MIT Sloan Sports Annual Conference*, Boston, MA, USA, Mar. 2016, pp. 1-9.
[39] W. Wang, Q. Yu, M. Fasli, Altering Gaussian process to Student-t process for maximum distribution construction, Int. J. Syst. Sci. 52 (4) (2021) 727–755.
[40] Y. Wang, R. Zhao, Z. Zuo, S. Guan, H. Li, Event-triggered dynamic anti-windup augmentation for saturated systems, Int. J. Syst. Sci. 52 (1) (2021) 196–216.
[41] N. Watson, S. Hendricks, T. Stewart and I. Durbach, Integrating machine learning and decision support in tactical decision-making in rugby union, J. Operat. Res. Soc., in press, doi: 10.1080/01605682.2020.1779624
[42] M. Wu, T. Wan, X. Wan, Z. Fang, Y. Du, A new localization method for epileptic seizure onset zones based on time-frequency and clustering analysis, *Pattern Recognition*, vol. 111, art. no. 107687, 2021.
[43] Zelic, I. Kononenko, N. Lavrac, V. Vuga, Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries, J. Med. Syst., vol. 21, no. 6, pp. 429-444, 1997.
[44] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, X. Liu, An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips, IEEE Trans. Nanotechnol. 18 (2019) 819–829.
[45] N. Zeng, S. Zuo, G. Zheng, Y. Ou, T. Tong, Artificial intelligence for medical image analysis of neuroImaging data, *Front. Neurosci.*, vol. 14, art. no. 480, 2020.
[46] S. Zheng, Y. Yue, J. Hobbs, Generating long-term trajectories using deep hierarchical networks, In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, Dec. 2016, pp. 1551-1559

Yuanchen Wang received his BSc degree in Business Mathematics and Statistics from the Statistics Department of the London School of Economics and Political Science, U.K, in 2016. In 2020, he received his MSc degree in Data Science (with distinction) from Birkbeck College, University of London, U.K., which was obtained whilst he was in full time employment as a data scientist at PwC in London. He began his industry career as an AI consultant and data scientist at Accenture AI and PwC, before moving to Revolut in the FinTech sector in early 2021. He possesses over five years' industry experience of applying advanced data analytics solutions for various multi-million-pound use cases within the financial services sector. His academic research interests include big data analysis, explainable AI, deep learning techniques, and the tangible applications of machine learning with the banking, Fintech, and asset/wealth management industries.

Weibo Liu received the B.S. degree in electrical engineering from the Department of Electrical Engineering & Electronics, University of Liverpool, Liverpool, U.K, in 2015, and the Ph.D. degree in artificial intelligence in 2021 from the Department of Computer Science, Brunel University London, Uxbridge, U.K. He is currently a lecturer at Brunel University London, Uxbridge, U.K. His research interests include big data analysis and deep learning techniques. He serves as an Associate Editor for the Journal of Ambient Intelligence and Humanized Computing and the Journal of Cognitive Computation. He is a very active reviewer for many international journals and conferences.

Xiaohui Liu received the B.Eng. degree in computing from Hohai University in 1982, and the Ph.D. degree in computer science from Heriot-Watt University, U.K., in 1988. He is currently a Professor of computing at Brunel University London, U.K., where he conducts interdisciplinary research involving artificial intelligence, dynamic systems, image and signal processing, and statistics, particularly for applications in biology, engineering and medicine. Professor Liu serves on editorial boards of four computing journals, founded the biennial international conference series on IDA in 1995, and has given numerous invited talks in bioinformatics, data mining and statistics conferences.